Methods in Molecular Biology 1764

Springer Protocols

Joseph A. Marsh *Editor*

Protein Complex Assembly

Methods and Protocols

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Herts, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Protein Complex Assembly

Methods and Protocols

Edited by

Joseph A. Marsh

MRC Human Genetics Unit, Institute of Genetics & Molecular Medicine,
University of Edinburgh, Edinburgh, UK

Humana Press

Editor
Joseph A. Marsh
MRC Human Genetics Unit
Institute of Genetics & Molecular Medicine
University of Edinburgh
Edinburgh, UK

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-7758-1 ISBN 978-1-4939-7759-8 (eBook)
https://doi.org/10.1007/978-1-4939-7759-8

Library of Congress Control Number: 2018935276

Printed on acid-free paper

This Humana Press imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

Most proteins can interact with other proteins or other copies of themselves and assemble into heteromeric or homomeric protein complexes. While the importance of protein-protein interactions is widely recognized, higher-order quaternary structure (i.e., the way the different subunits of a complex are arranged with respect to each other) and the processes by which protein complexes assemble have often been neglected. Protein complex assembly is fundamental to biology, both because of the many critical functions performed by protein complexes and because the assembly mechanism itself is often crucial for biological regulation. Therefore, understanding how proteins assemble into complexes and the diverse quaternary structures they can form is key to understanding many biological processes at a molecular level. Fortunately, in recent years, there has been an explosion in the methods available for characterizing protein complexes, both experimentally and computationally.

The first part of this volume focuses on a variety of different experimental approaches for characterizing protein complex structure, dynamics, and assembly. The topics covered include cryo-electron microscopy (both single particle and cryo-tomography), NMR (both liquid and solid state), deep sequencing linked to saturation mutagenesis and yeast surface display, in vivo cross-linking, mass spectrometry (including native, hydrogen-deuterium exchange, cross-linking, and coupled to biochemical fractionation), super-resolution and (immuno)fluorescence microscopy, and protein complex expression and purification.

The second part of this volume is concerned with computational strategies for studying protein complexes. Chapters are focused on methods for inferring quaternary structure from crystallographic data, the use of online databases, identifying complexes from proteomics data, simulations of protein assembly pathways, macromolecular docking, and modelling assemblies using diverse experimental and evolutionary restraints.

Overall, it is hoped that the breadth and depth of coverage will make this volume a great aid for any researcher studying protein complexes. Moreover, it will help to emphasize the importance of being familiar with the multiple experimental techniques and computational methods needed to obtain the comprehensive picture of protein complex structure, dynamics, and assembly afforded by the emerging field of integrative structural biology.

Edinburgh, UK *Joseph A. Marsh*

Contents

Contributors

Logan S. Ahlstrom · *Department of Molecular, Cellular, and Developmental Biology, Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI, USA; Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI, USA*

Sonja-Verena Albers · *Molecular Biology of Archaea, Institute of Biology II, University of Freiburg, Freiburg, Germany*

Ingemar André · *Department of Biochemistry and Structural Biology, Center for Molecular Protein Science, Lund University, Lund, Sweden*

Jessica Andreani · *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette Cedex, France*

Karine Pradeau Aubreton · *IGBMC, Illkirch, France*

Gary Bader · *Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada*

Emeline Barbet-Massin · *Munich Center for Integrated Protein Science (CIPS-M) at Department Chemie, Technische Universität München (TUM), Garching, Germany; Dynamic Biosensors, Planegg, Germany*

James C. A. Bardwell · *Department of Molecular, Cellular, and Developmental Biology, Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI, USA; Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI, USA*

Paul A. Bates · *Biomolecular Modelling Laboratory, The Francis Crick Institute, London, UK*

Julien Batisse · *IGBMC, Illkirch, France*

Roland Beckmann · *Department of Biochemistry and Center for Integrated Protein Science Munich (CiPSM), Gene Center, Ludwig-Maximilians-Universität München, Munich, Germany*

Morgan Beeby · *Department of Life Sciences, Imperial College of London, London, UK*

Stephen J. Benkovic · *Department of Chemistry, The Pennsylvania State University, University Park, PA, USA*

L. Therese Bergendahl · *MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK*

Martin Blackledge · *Univ. Grenoble Alpes, CEA, CNRS, IBS, Grenoble, France*

Elisabetta Boeri Erba · *Institut de Biologie Structurale (IBS), Université de Grenoble Alpes, CEA, CNRS, Grenoble, France*

Adam Byron · *Cancer Research UK Edinburgh Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK*

Raphael A. G. Chaleil · *Biomolecular Modelling Laboratory, The Francis Crick Institute, London, UK*

Paushali Chaudhury · *Molecular Biology of Archaea, Institute of Biology II, University of Freiburg, Freiburg, Germany*

Jiawen Chen · *Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, NY, USA*
William Dearnaley · *Virginia Tech Carilion Research Institute, Roanoke, VA, USA*
Peter Dedecker · *Department of Chemistry, University of Leuven, Heverlee, Belgium*
Sucharita Dey · *Department of Structural Biology, Weizmann Institute of Science, Rehovot, Israel*
Madeline J. Dukes · *Protochips, Inc., Applications Science, Raleigh, NC, USA*
Sylvia Eiler · *IGBMC, Illkirch, France*
Andrew Emili · *Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada; Departments of Biology and Biochemistry, Boston University, Toronto, ON, Canada*
John R. Engen · *Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA*
Olivera Francetic · *Biochemistry of Macromolecular Interactions Unit, Department of Structural Biology and Chemistry, CNRS UMR3528, Institut Pasteur, Paris Cedex 15, France*
Joachim Frank · *Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA; Department of Biological Sciences, Columbia University, New York, NY, USA*
Ziao Fu · *Integrated Program in Cellular, Molecular and Biomolecular Studies, Columbia University College of Physicians and Surgeons, New York, NY, USA*
Brian L. Gilmore · *Virginia Tech Carilion Research Institute, Roanoke, VA, USA*
Thomas G. Gligoris · *Department of Biochemistry, University of Oxford, Oxford, UK*
Florian Goebels · *Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada*
Raphael Guerois · *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette Cedex, France*
Michal Hammel · *Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
Fabienne Hans · *Institut de Biologie Structurale (IBS), Université de Grenoble Alpes, CEA, CNRS, Grenoble, France*
Louie D. Henderson · *Department of Life Sciences, Imperial College of London, London, UK*
Fabian Hertel · *Department of Chemistry, University of Leuven, Heverlee, Belgium*
Philipp Hoess · *Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany*
Scott Horowitz · *Department of Chemistry and Biochemistry, Knoebel Institute for Healthy Aging, University of Denver, Denver, CO, USA*
Lucas Hu · *Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada*
Roxana E. Iacob · *Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA*
Malene Ringkjøbing Jensen · *Univ. Grenoble Alpes, CEA, CNRS, IBS, Grenoble, France*
Sandip Kaledhonkar · *Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA*

Nir Kalisman · *Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, Hebrew University of Jerusalem, Jerusalem, Israel*
Deborah F. Kelly · *Virginia Tech Carilion Research Institute, Roanoke, VA, USA; Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA*
Brent A. Kochert · *Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA; Center for Drug Discovery, Northeastern University, Boston, MA, USA*
Emmanuel D. Levy · *Department of Structural Biology, Weizmann Institute of Science, Rehovot, Israel*
Nicolas Levy · *IGBMC, Illkirch, France*
Yanping Liang · *Virginia Tech Carilion Research Institute, Roanoke, VA, USA*
Areli Luna-Rico · *Biochemistry of Macromolecular Interactions Unit, Department of Structural Biology and Chemistry, CNRS UMR3528, Institut Pasteur, Paris Cedex 15, France*
Benoit Maillot · *IGBMC, Illkirch, France*
Alexandros Makriyannis · *Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA; Center for Drug Discovery, Northeastern University, Boston, MA, USA*
Bridget C. Marcinkowski · *Virginia Tech Carilion Research Institute, Roanoke, VA, USA*
Joseph A. Marsh · *MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK*
John F. Marshall · *Barts Cancer Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK*
Angelica V. Medina-Cucurella · *Department of Chemical Engineering and Materials Science, Michigan State University, East Lansing, MI, USA*
Birgit H. M. Meldal · *European Molecular Biology Laboratories, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, Cambridgeshire, UK*
Gary C. H. Mo · *Department of Pharmacology, University of California at San Diego, La Jolla, CA, USA*
Iain H. Moal · *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK*
Markus Mund · *Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany*
Joanna Musial · *Gene Center, Department of Biochemistry and Center for Integrated Protein Science Munich (CiPSM), Ludwig-Maximilians-Universität München, Munich, Germany*
Aravindan Arun Nadaradjane · *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette Cedex, France*
Oyindamola Oladosu · *IGBMC, Illkirch, France*
Mizar F. Oliva · *Institut de Biologie Structurale (IBS), Université de Grenoble Alpes, CEA, CNRS, Grenoble, France*
Sandra Orchard · *European Molecular Biology Laboratories, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, Cambridgeshire, UK*
Jose Luis Ortega-Roldan · *School of Biosciences, University of Kent, Canterbury, UK*

Anthony M. Pedley · *Department of Chemistry, The Pennsylvania State University, University Park, PA, USA*
Jan-Michael Peters · *Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Vienna, Austria*
Carlo Petosa · *Institut de Biologie Structurale (IBS), Université de Grenoble Alpes, CEA, CNRS, Grenoble, France*
Shannon J. L. Pinnington · *Advanced Imaging Resource, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK*
Reza Pourhaghighi · *Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada*
Bernd Reif · *Munich Center for Integrated Protein Science (CIPS-M) at Department Chemie, Technische Universität München (TUM), Garching, Germany; Deutsches Forschungszentrum für Gesundheit und Umwelt, Helmholtz-Zentrum München (HMGU), Neuherberg, Germany*
Manuel Reitberger · *Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany*
Jonas Ries · *Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany*
Marc Ruff · *IGBMC, Illkirch, France*
Loïc Salmon · *Centre de RMN à Très Hauts Champs, Institut des Sciences Analytiques, UMR 5280, CNRS, ENS Lyon, UCB Lyon 1, Université de Lyon, Villeurbanne, France*
Johanna C. Scheinost · *Department of Biochemistry, University of Oxford, Oxford, UK*
Dina Schneidman-Duhovny · *School of Computer Science and Engineering, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel*
Luca Signor · *Institut de Biologie Structurale (IBS), Université de Grenoble Alpes, CEA, CNRS, Grenoble, France*
Moriya Slavin · *Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, Hebrew University of Jerusalem, Jerusalem, Israel*
Eli van der Sluis · *Gene Center, Department of Biochemistry and Center for Integrated Protein Science Munich (CiPSM), Ludwig-Maximilians-Universität München, Munich, Germany; Department of Bionanoscience, Faculty of Applied Sciences, TU Delft, Delft, The Netherlands*
Jenny-Lee Thomassin · *Biochemistry of Macromolecular Interactions Unit, Department of Structural Biology and Chemistry, CNRS UMR3528, Institut Pasteur, Paris Cedex 15, France*
Patrick Tripp · *Molecular Biology of Archaea, Institute of Biology II, University of Freiburg, Freiburg, Germany; Spemann Graduate School of Biology and Medicine (SGBM), University of Freiburg, Freiburg, Germany*
A. Cameron Varano · *Virginia Tech Carilion Research Institute, Roanoke, VA, USA; Translational Biology, Medicine, and Health Graduate Program, Virginia Tech, Blacksburg, VA, USA*
Thomas E. Wales · *Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA*
Florian Weissmann · *Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Vienna, Austria*
Jonathan N. Wells · *MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK*

Ann P. Wheeler · *Advanced Imaging Resource, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK*
Howard White · *Physiological Sciences, Eastern Virginia Medical School, Norfolk, VA, USA*
Timothy A. Whitehead · *Department of Chemical Engineering and Materials Science, Michigan State University, East Lansing, MI, USA; Department of Biosystems and Agricultural Engineering, Michigan State University, East Lansing, MI, USA*
Yinghao Wu · Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, NY, USA
Jin Zhang · *Department of Pharmacology, University of California at San Diego, La Jolla, CA, USA*

Part I

Experimental Methods

Chapter 1

Experimental Characterization of Protein Complex Structure, Dynamics, and Assembly

Jonathan N. Wells and Joseph A. Marsh

Abstract

Experimental methods for the characterization of protein complexes have been instrumental in achieving our current understanding of the protein universe and continue to progress with each year that passes. In this chapter, we review some of the most important tools and techniques in the field, covering the important points in X-ray crystallography, cryo-electron microscopy, NMR spectroscopy, and mass spectrometry. Novel developments are making it possible to study large protein complexes at near-atomic resolutions, and we also now have the ability to study the dynamics and assembly pathways of protein complexes across a range of sizes.

Key words X-ray crystallography, Cryo-electron microscopy, NMR, Mass spectrometry, Super-resolution microscopy, Quaternary structure

1 Introduction

The tendency of proteins to form complexes and the functional implications of this behavior have been recognized since the earliest days of molecular biology. Though it is unclear who was the first to explicitly note their existence, it seems likely that interest in protein complexes arose in tandem with investigations into the nature of viruses. In 1935, W. M. Stanley reported the isolation of "a crystalline material which has the properties of Tobacco-mosaic virus" (TMV) and demonstrated that this material was predominantly composed of protein [1]. However, it is not obvious whether he understood the implications of finding such a structure for proteins beyond those comprising the TMV capsid. Nonetheless, this period in time marks a turning point for the field of biology, and over the next few decades, much of the groundwork was laid for our current understanding of protein structure.

As if to usher in the era, 1944 saw the publication of Erwin Schrödinger's classic book, *What is Life?* [2], which inspired a number of scientists, particularly physicists, to try their hand at

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_1,

biology. Among these were names such as Francis Crick, James Watson, and Maurice Wilkins, best known for their discovery in 1953 of the structure of DNA. Also familiar with the book, though not so enamored with it [3], was Max Perutz, who was at the time working on hemoglobin. By this point, it was clear that many proteins were multimeric assemblies, and by 1955 the TMV capsid had been described as a self-assembling homomer comprised of several thousand identical subunits [4]. All that remained for the study of proteins and their complexes to begin in earnest was the production of the first structural models. This feat was achieved before the end of the decade by John Kendrew and Max Perutz, first with monomeric myoglobin and shortly thereafter tetrameric hemoglobin [5]. In solving these first near-atomic resolution structures, they opened up the door to the new field of structural biology.

Following these first postwar forays into the characterization of protein complexes, technology improved rapidly, and during this period structural biology was one of the most productive fields in all of science. X-ray crystallography in particular deserves special mention, having led to no fewer than 14 Nobel Prizes since 1914. Of these Nobels, uncovering the structure of the ribosome—a huge complex consisting of dozens of protein and rRNA subunits—is perhaps the crowning achievement [6–8].

However, while X-ray crystallography was in its heyday, other fields were not silent. A classic molecular biology technique that appeared in the late 1980s was the yeast two-hybrid assay (Y2H) [9], in which two proteins of interest are fused to a DNA-binding domain and a transcriptional activator domain, allowing binary interactions (or lack thereof) between the proteins to be detected by the expression of a reporter gene. This assay has been enormously successful, with the original paper having been cited nearly 7000 times since publication. Despite its age, it is still relevant today, notably through its use in a high-throughput manner to map the binary interaction landscape of *E. coli* [10], producing a map of 2334 pairwise interactions and enabling inference of many novel protein complexes in the process. However, though simple and cost-effective, there are inherent limitations to the technique: most obviously, the necessary involvement of bulky reporter domains risks disrupting or preventing subtle interactions between many proteins. As a result, approaches using mass spectrometry have largely superseded Y2H as the method of choice for quantitative studies of the interactome.

Although mass spectrometry is at least as old as X-ray crystallography, its use in the study of protein complexes was not possible until the development of soft matrix-assisted laser desorption/ionization (MALDI) and related techniques from Karas, Bachmann, Hillenkamp, and Tanaka [11]. A short while after these breakthroughs, electrospray ionization (ESI) was also

enabled for use with proteins [12]. Both MALDI and ESI are now essential tools in biology, and by coupling mass spectrometers with liquid chromatography and affinity purification, it is possible to infer the existence of protein complexes from large-scale protein interaction data.

2 X-ray Crystallography

X-ray crystallography was the first method to make the field of structural biology a reality, bringing together three separate technologies, each important in its own right. These technologies include methods for overexpression and purification of proteins, the production of powerful X-ray sources, and computational methods for solving X-ray diffraction patterns. By and large, the ways in which X-ray crystallography can be used to determine protein structure are the same for monomeric proteins and those that form complexes. There are, however, some important differences and additional difficulties that need considering in the case of complexes. Furthermore, although cryo-electron microscopy (cryo-EM) seems poised to overtake X-ray crystallography as the method of choice for solving large heteromeric structures, there have been a number of exciting developments in crystallography that look set to ensure its future for many years to come. The following section will highlight some of these advances and attempt to give a summary of the current state of the field.

2.1 Protein Expression, Purification, and Crystallization

Acquiring samples of purified protein is a requisite first step for almost all of the methods discussed in this chapter, and X-ray crystallography is no exception. A typical setup for expression of protein for crystallization involves the transformation of *E. coli* with a plasmid containing your protein of interest, usually under the control of a strong, inducible promoter [13]. For monomeric bacterial proteins, this system is simple and easy to use, but expressing heteromeric protein complexes is often considerably more challenging, particularly those of eukaryotes. The key difficulty lies in the production of sufficient quantities of pure sample, as in non-native hosts protein complex assembly is often inefficient or simply incomplete. For eukaryotic proteins, this is compounded further by the fact that most undergo alternative splicing and other post-transcriptional or posttranslational events, the machinery for which is generally lacking in bacteria.

Prior to any benchwork, however, improvements in the cellular yield of bacterial heteromers can be achieved by carefully considering the design of the expression vector in light of the assembly pathway of the protein complex in question. For example, the order of genes within protein complexes is under selection to match the assembly order of protein complexes [14]. It has been

demonstrated experimentally that taking this fact into account can markedly increase complex assembly efficiency and that yields of heteromers in their fully assembled native state can be improved by using the native operon structure in expression vectors [15, 16].

When purifying protein complexes, there is a trade-off between obtaining highly pure samples and ensuring that the intermolecular bonds between subunits are not disrupted. Although the diversity of methods for protein purification is bewilderingly high, in practice most methods suitable for protein complexes are variations on affinity purification. Here too, careful experimental design can pay dividends, and when possible, it is generally preferable to produce bait proteins that are expressed at endogenous levels. Ideally, the number of purification steps would be limited in order to retain as much protein in its native state as possible, but in practice, multiple purification steps are often required before the sample is pure enough to crystallize. Methods such as dynamic light scattering [17] can be used to assess sample purity and readiness for crystallization.

In most cases, it will be necessary to tailor the expression and purification process to the protein complex of interest. Depending on the orientation of subunits within the structure, for example, different subunits may make better or worse bait proteins, as will N- or C-terminal purification tags. Similarly, some complexes may be disrupted by the presence of metal ions, in which case other beads, e.g., those coated in calmodulin, may be more suitable. Although there has been progress toward high-throughput expression and purification pipelines [18], much of this work still relies on trial and error informed by the expertise of individual structural biologists and research technicians.

The crystallization process is still the main bottleneck in X-ray crystallography, despite having been largely automated by the development of screening robots. There have, however, been some important methodological developments in the crystallization of membrane proteins, which will also be useful for many membrane complexes. For example, an exciting new method—X-ray solvent contrast modulation—has recently been used to visualize the interaction between membrane proteins and the phospholipid bilayer [19]. However, this method does not do away with the requirement for good quality crystals, and these are still largely obtained through trial and error—beyond a few general rules of thumb, we still do not have a good understanding of how different proteins will behave under different crystallization conditions.

2.2 Diffraction Pattern Acquisition

Once suitable crystals have been obtained, image acquisition can begin. In contrast to earlier steps, enormous progress has been made in this domain since William L. Bragg first demonstrated X-ray diffraction from sodium chloride crystals in 1913 [20]. By far the most important development in the field has been that of

synchrotron X-ray sources. Synchrotrons are able to produce X-rays at far higher intensities than traditional sources and as such greatly reduce the time it takes to produce diffraction patterns. Technical properties of the beamline can also be manipulated, for example, narrowing it in order to focus on the best quality region of the crystal, thus improving the quality of the resulting diffraction pattern.

More recently, X-ray free-electron lasers (XFELs) have also begun to make an appearance in structural biology (Fig. 1). It is hard to overstate the impact that this technology will have on the field, since XFELs are capable of producing peak beam energies approximately ten orders of magnitude greater than current third-generation synchrotrons [21] and in doing so enable a radically different approach to crystallography. The principle benefit of all this additional power is that the time needed to generate a diffraction pattern is drastically reduced. A crystal in the path of such high-energy photons will be vaporized almost instantly, but since the diffraction pattern will be obtained faster than the sample is destroyed, this does not present a problem—a fact first noted by Neutze et al. [22]—giving rise to the term "diffraction before destruction." This obviously generates a need for a great many crystals in order to obtain diffraction patterns from all angles of the structure, but this too is not a major issue, since these crystals need

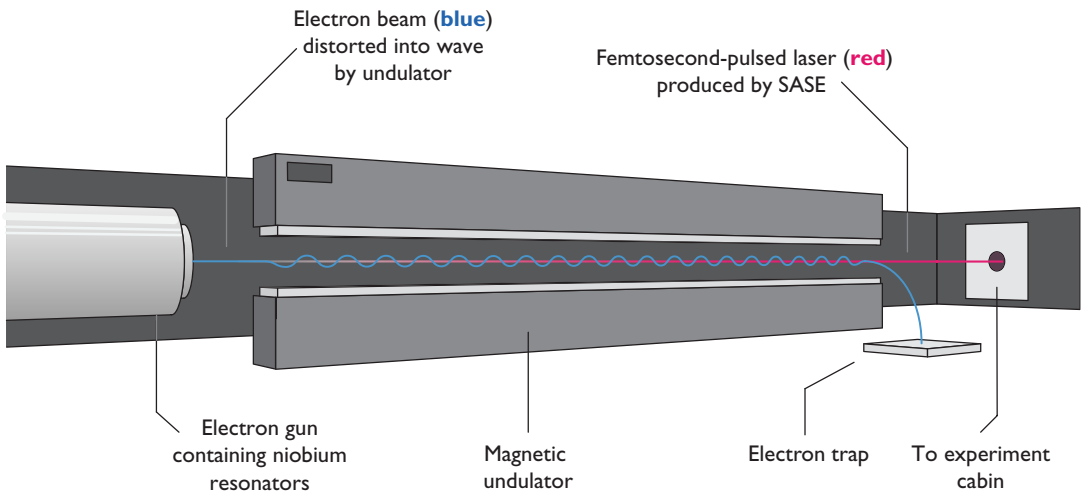


Fig. 1 X-ray free-electron lasers. An XFEL produces high-energy X-rays by a process known as self-amplified spontaneous emission (SASE). An electron bunch is accelerated close to the speed of light using superconducting niobium resonators. When this passes through the undulator, the wiggling motion induced by the magnets causes the electrons to emit photons. As these photons are traveling only slightly faster than the electrons, they interact with the electrons as they catch them up at each period in the undulator. Over the length of the undulator, this causes the electrons to bunch into very thin disks, which emit intense, synchronized flashes of X-ray laser light. These femtosecond X-ray pulses are then guided into the experiment cabin, where they encounter a stream of protein nanocrystals, producing diffraction patterns from each one

only be a few nanometers in size. In fact, since nanoscale crystals are far easier to grow, the method also circumvents the tedious trial and error process of producing the larger crystals needed for use with traditional X-ray sources.

2.3 Structure Determination

Interpretation of the crystal diffraction pattern required the solution of a long-standing challenge in the early days of X-ray crystallography, known as the phase problem [23]. The phase problem exists due to the fact that, while diffraction patterns capture the amplitude of diffracted photons from a crystal (seen as the intensity of spots on the photograph), the phase of those photons is lost in the process of image acquisition. Unfortunately, it is the phases of the diffracted photons, rather than their amplitudes, that carry the most information about the underlying crystal structure. The eventual solution of this problem by Max Perutz was the key to his and Kendrew's determination of the first protein structures.

Perutz's breakthrough came when he realized that a technique previously used for phasing crystals of much smaller molecules would also be applicable to proteins. This method, known as isomorphous replacement (IR) [24], involves soaking the crystal in a solution containing heavy metals. Crucially, the incorporation of heavy metals into the crystal does not significantly alter its structure, and as a result, the position of spots in the diffraction pattern remains almost unchanged, while subtle differences in their intensity point to the location of the heavy atoms. This provides an essential reference point for calculation of the missing X-ray phases.

For large protein complexes, polynuclear metal clusters are often used in place of individual heavy atoms because of their particularly high electron density and associated isomorphous or anomalous scattering signal [25]. This approach has recently been used to good effect in solving the structure of the notoriously difficult mediator complex [26]. However, different methods for solving the phase problem have been established in addition to IR, most notably multiple wavelength anomalous diffraction (MAD) [27]. This method operates on different principles to IR but is popular since it is limited only by the quality of the diffraction pattern provided to it.

As a consequence of the ever-expanding number of structures in the Protein Data Bank (PDB) [28] and the widespread availability of sequence data, it is often possible nowadays to avoid de novo phasing altogether. Molecular replacement by homology modeling makes use of the fact that closely related sequences generally have very similar folds and therefore can be used as a template to guide brute-force calculation of diffraction pattern phases. There are currently several programs that automate this process, for example, Phaser [29], which is available within the widely used CCP4 software suite [30].

3 Cryo-electron Microscopy

X-ray crystallography has been, and will continue to be, an enormously useful tool for investigating proteins and protein complexes. However, a recent resurgence in cryo-EM has had a transformative effect on structural biology—particularly on our ability to solve the structures of large protein complexes above 300 kDa in size. Its unique affinity for large complexes is especially convenient since these often prove prohibitively difficult to crystallize, in large part due to compositional heterogeneity of the purified samples, which cryo-EM can more easily handle. The two methods are therefore highly complementary, and indeed many structures are solved to high resolution by a combination of the two—cryo-EM for the coarse-grained structure and X-ray crystallography for atomic resolution of individual subunits. Likewise, nuclear magnetic resonance (NMR) spectroscopy also has difficulty handling large complexes and thus can be used effectively in combination with cryo-EM.

As interest in cryo-EM increases, there are signs that single-particle cryo-EM is making incursions into the size and resolution niche currently occupied by X-ray crystallography. Illustrating this, two important symbolic barriers were broken in a recent paper describing the structures of two homomeric complexes: isocitrate dehydrogenase and glutamate dehydrogenase [31]. The former weighs in at just 93 kDa and is the first single-particle cryo-EM structure of a <100 kDa complex, while the latter was resolved to 1.8 Å, breaking the <2 Å resolution barrier. As we shall see, the remarkable technological achievements displayed in this paper and several others have been driven by dramatic improvements in the two key areas of image acquisition and processing [32].

3.1 Image Acquisition in Single-Particle Cryo-EM

The first major development in cryo-EM's current flourishing came with the replacement of photographic film by digital direct electron detectors, specifically Monolithic Active Pixel Sensors (MAPS). It was not until relatively recently that digital detectors came into widespread use, as until direct electron detectors became available (not to be confused with charge-coupled devices), film was the medium that achieved the best possible detective quantum efficiencies (DQE) [33]. DQE is a measure of the signal-to-noise ratio that can be achieved relative to an ideal detector [34] and is defined as follows:

$$\mathrm{DQE} = \left(S / N_{\mathrm{in}}\right)^2 / \left(S / N_{\mathrm{out}}\right)^2$$

where S/N_{in} and S/N_{out} are the input and output signal-to-noise ratios, respectively; a DQE of 1 would imply that the detector was not responsible for any noise in the image. For reference, film has

a DQE of around 0.3, whereas the current state-of-the-art MAPS detectors achieve roughly twice that.

Ultimately, DQE is the most important factor in choosing whether to use film or digital detectors, but now that MAPS detectors have surpassed film in that regard, several other compelling advantages of digital detectors can be exploited. From a practical standpoint, they are significantly faster to use than film, since images can be viewed immediately after collection and their acquisition can be automated. They can also be used to produce high frame-rate videos, enabling them to be run in counting mode, where instead of integrating the signal produced by each incident electron across all the pixels in which a charge was registered, only the pixel with the highest charge is counted [35]. This is conceptually similar to the way in which certain microscopy techniques achieve super-resolution images, and the company Gatan has recently brought this idea to market with a dedicated super-resolution mode for their K2 Summit detector.

One exciting new technology beginning to make its presence felt is the Volta phase plate, which can be used to produce phase contrast during image acquisition. In order to be able to correctly distinguish different particles in the sample, it is important to have good contrast in the images. Unfortunately, the method by which this contrast is currently changed relies on defocusing the image slightly, and as a result, if greater contrast is required, it comes at the expense of resolution. The Volta phase plate circumvents this issue by modulating the phase directly, without affecting the focus of the image [36]. Though the principle has been understood for some time, it was not until recently that various practical issues were solved, enabling the Baumeister lab to produce a 3 Å structure of the 20S proteasome, thus matching the resolution achieved by defocus methods [37]. Most impressively, the same group has recently published a 3.2 Å structure of the 64 kDa hemoglobin molecule [38].

3.2 Image Processing and Structure Determination

A second important factor in cryo-EM's success has been the appearance of better image processing software, which has enabled researchers to get the most out of the concurrent improvements in imaging hardware. In addition to improving resolution, the emergence of electron detectors capable of producing high frame-rate videos in counting mode has a secondary benefit, in that it enabled beam-induced motion blurring in the images to be corrected computationally, a feat that was first achieved by two groups almost simultaneously in 2013 [39, 40]. Since the reduction in signal quality incurred by beam-induced movement is around fivefold if uncorrected [41], this was a highly significant breakthrough and is now the standard protocol and can be performed using the widely used RELION software [42, 43].

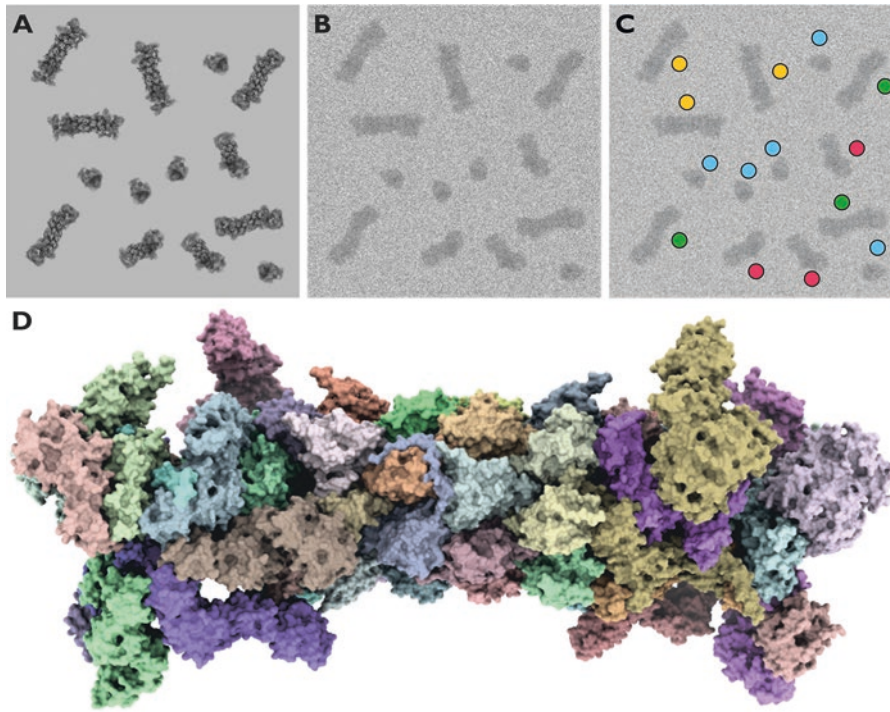


Fig. 2 Image classification in single-particle cryo-EM. (**a**) Theoretical electron micrograph of the human 26S proteasome produced by a detector with DQE = 1.0. (**b**) Image produced by detector with lower DQE, resulting in noise and phase contrast issues. (**c**) 2D image classification of proteasome particles into categories corresponding to their orientation. (**d**) Fitted 3.8 Å resolution model. Produced using K2 Summit detector and processed in RELION; PDB ID: 5T0C, EMDB ID: 8332 [127]

Computational progress has also been essential for image classification (Fig. 2). In single-particle cryo-EM, individual protein complexes are fixed in random positions and orientations in the flash-frozen sample. To determine the structure, each particle captured in the imaging process must first be categorized according to its orientation. For symmetrical structures, the number of particles required in the image is usually considerably lower, since multiple axes of symmetry effectively make many of the orientations that can be observed redundant. This has the effect of increasing the effective number of images of the particle and, conversely, makes the solution of asymmetric structures more challenging.

Dealing with structural and compositional heterogeneity is a related problem, which arises from imperfect purification or different functional states of the complex being present in the sample. Computational approaches for dealing with this arrived in 1998 with a maximum likelihood method for classifying 2D images [44]. Three-dimensional classification methods, being much more computationally intensive, did not appear till later but are now an area of active development, since at present they are one of the major

bottlenecks in structure determination [45–47]. In practice, multiple rounds of image classification and refinement are usually carried out, beginning with removal of low-quality particles, followed by 2D and 3D image classification and finishing with polishing steps.

3.3 Cryo-electron Tomography

Although single-particle cryo-EM offers good resolution without the need for crystallization, it still requires that the protein of interest be purified first, thus ruling out many of the protein complexes present in the cell, including those embedded in the cell membrane. Cryo-electron tomography offers an attractive alternative in these cases, as it allows imaging of protein complexes in their native environment, albeit with a significant reduction in the resolution achievable. By and large, the processes involved in cryo-ET are similar to those of cryo-EM, with the key difference being that one acquires images by rotating the sample through a range of different tilts, rather than relying on the protein being present in many different orientations. This tilting method is also used to produce images in electron crystallography [48].

By reconstructing the set of images produced from these different tilts, a tomogram of the structure of interest can be built. Because the exact orientation of the sample is known for each image, confounding factors such as other proteins or biological structures can be removed from the image, which would not be possible if one were to attempt single-particle cryo-EM on a non-purified sample. The downside to this approach is that the sample can only be tilted up to a point, as the effective thickness of the sample in the path of the electron beam increases with the angle of the sample. As a result, there is always a "wedge" of data missing from the set of images of a complex, which seriously limits the resolution achievable from a single structure.

However, an important development of cryo-ET is subtomogram averaging, otherwise known as single-particle tomography [49]. Here, multiple tomograms of different particles in the sample are produced and then averaged in similar fashion as for images in cryo-EM [50]. This averaging process can fill in the missing wedges in the data, provided the proteins in the sample are present in a sufficient variety of orientations [51]. Although the technique is not yet able to reliably achieve atomic resolutions, it is not far off [52], and the lure of imaging protein complexes in their natural environment will almost certainly ensure its continued development.

4 Nuclear Magnetic Resonance Spectroscopy

Many biologically important protein complexes exist in a dynamic ensemble of conformational states or contain subunits that only interact very weakly with each other. Such complexes do not lend themselves well to characterization by crystallographic or cryo-EM

methods, which can generally only resolve a single structural state at a time. NMR spectroscopy is well suited to investigating these cases as the proteins are visualized in solution, rather than crystallized or frozen. On the other hand, NMR has traditionally struggled to resolve structures beyond 30 kDa due to the fact that the relaxation of nuclear spin orientations is very efficient for large, slowly tumbling molecules. This has the effect of broadening the peaks observed in NMR spectra and, coupled with the fact that large molecules naturally produce more complex spectra than smaller ones, ensures that using NMR to study protein complexes is challenging.

4.1 Solution NMR Spectroscopy of Multi-subunit Protein Complexes

An essential tool for investigating large complexes is transverse relaxation-optimized spectroscopy (TROSY) [53], which uses constructive interference between different relaxation effects to improve the resolution of chemical shifts. Equally important is the use of deuterium labeling [54]. Like TROSY, this improves resolution by increasing the relaxation time of molecules. A further extension of these concepts is methyl TROSY, which makes use of isotopically labelled methyl groups set against a highly deuterated background. Because methyl groups produce especially intense resonances, they are easily identifiable within NMR spectra, and furthermore they are well dispersed within nearly all protein structures [55]. Using this technique, it is possible to resolve proteins with molecular weights into the low hundreds of kDa, overlapping slightly with the lower limits achievable by cryo-EM.

For yet larger protein complexes, or those with more heterogeneous structures, the complexity of the spectra itself becomes the limiting factor, rather than the spin relaxation rates. In these cases, clever use of isotope labeling can often simplify matters considerably [56]. Segmental labeling is one such example, in which isotopically labelled regions of the protein are spliced in using inteins or sortases [57]. Unsurprisingly, this is fraught with technical difficulties, but despite these the method has been used to great effect in studying large protein structures, from the 0.6 MDa ClpB disaggregase chaperone [58] to prion protein amyloid fibrils [59].

Solution NMR also provides a means for structurally characterizing complexes with a high degree of disorder, which are thus inaccessible to other techniques. By combining solution NMR measurements with other techniques, such as small-angle X-ray scattering, ensemble models of a variety of disordered or highly dynamic complexes have been constructed in recent years [60–64].

4.2 Solid-State NMR Spectroscopy

Solid-state NMR spectroscopy makes use of sample in a solid state, which is then spun rapidly inside the magnetic field, as opposed to the molecule of interest being free to tumble in solution. This is possible because of a quirk of NMR that leads to the "magic angle spinning" technique [65]: when the sample is tilted

at the magic angle θ_m relative to the external magnetic field (such that $\cos^2\theta_m = \frac{1}{3}$), the peaks on the NMR spectrum become much sharper, enabling structure to be determined. Magic angle spinning in effect mimics the natural tumbling of molecules in solution; since the rate of "tumbling" is no longer dictated by the size of the macromolecule being observed, solid-state NMR can be used to probe much larger structures (e.g., amyloid fibrils).

Solid-state NMR is also well suited to studying membrane-embedded protein complexes due to the fact that proteins in lipid bilayers are by nature oriented in the same direction. Through careful sample preparation, this natural orientation can be preserved during the course of the NMR experiment, allowing high-resolution spectra to be produced directly from the sample by aligning it at the correct angle to the external magnetic field [66]. Using both oriented sample methods and magic angle spinning, a number of impressive complexes have been solved [67–69].

5 Mass Spectrometry

5.1 Native Mass Spectrometry

The arrival of soft ionization MS techniques in the 1980s was of critical importance for the study of protein complexes, as it allowed delicate non-covalent interactions between proteins to be preserved in the gas phase, making it possible to study intact protein complexes via MS. Combined with the later development of time-of-flight mass analyzers, this became known as native MS. Because native MS does not interfere with the intermolecular bonds between protein complex subunits, it can be used to study properties such as stoichiometry, compositional heterogeneity, and dynamic processes such as assembly or disassembly.

5.1.1 Electrospray Ionization Mass Spectrometry

The ionization method of choice for native MS is currently ESI, as MALDI requires the sample of interest to be mixed with a matrix, which is then ionized using lasers. This matrix is usually formed from crystallized organic acids and as such is generally too harsh for complexes to be maintained in their native state, with a few exceptions [70]. In contrast, ESI uses the sample as is and ionizes it by passing it through a narrow glass capillary, to which a high voltage is applied, causing the charged sample to be aerosolized as it leaves the capillary. Through successive Coulomb fission events and evaporation of solvent from the sample, the ions in this mist rapidly enter the gas phase as they move toward the mass analyzer.

Another important benefit of ESI over MALDI is that it produces multiply charged protein ions with regularity [71]. This is useful when coupling ESI to tandem MS, where the protein sample is first analyzed in its native state, before being fragmented and

subject to a second round of mass analysis. Single-charge proteins produce little useful information upon fragmentation, as only a single peptide fragment will be charged, essentially wasting much of the protein. Having multiple charges per ion also reduces the corresponding m/z ratio. This is important when investigating larger proteins and protein complexes, since historically the operative range of quadrupole mass analyzers has been limited to about 4000 m/z. For this reason, time-of-flight mass analyzers have been the mainstay of native MS for many years [72], since they have good resolving power and sensitivity over a much wider range than traditional quadrupole analyzers. In 2005, however, Orbitrap analyzers became available [73], and subsequent development since then has pushed the limits of their operative mass range into the tens of thousands m/z.

Another hugely important development in ESI came with the introduction of much narrower capillaries in the electrospray devices, leading to nano-ESI [74]. Coupled with lower sample flow rates, this improves ionization efficiency substantially [75]. Equally importantly, it greatly reduces the amount of sample required for each experiment. This enables analysis of proteins, which are hard to purify in large quantities, or makes it possible to run experiments investigating dynamic processes that take place over the course of seconds to minutes.

5.1.2 Applications of Native Mass Spectrometry

Due to its low sample requirements and sensitive treatment of the intermolecular interactions, native MS is very versatile. A common and technically straightforward use of the method is simply to determine the constituent parts of a particular protein complex, which can be done via tandem MS [76, 77]. The weights of individual subunits from the complex are determined in the first round of mass analyses (MS1), with identities being inferred from fragmented peptides in the second round (MS2). From this starting point, it is then possible to generate interaction maps based on the weights of peaks corresponding to different subunit combinations, as well as getting an idea of relative binding strengths.

More interesting is the use of native MS in time-resolved studies. These include following subunit exchange processes between heat-shock proteins [78], observing conformational changes of membrane complexes upon ligand binding [79], and determining protein complex assembly and disassembly pathways [80–82]. This last example can be achieved by adding different chaotropic agents to the solution containing the intact protein complex and then observing the intermediates that are produced across different concentrations.

5.2 Cross-Linking Mass Spectrometry

Cross-linking mass spectrometry (XL-MS) uses chemical cross-linkers to provide distance constraints between different residues in a protein complex. These can either be intramolecular or intermolecular, and

as such XL-MS can be used to produce low-resolution structural information, particularly of the interfaces between different subunits. It is particularly effective when used in combination with more established structural techniques or computational modeling and has become a central part of the new, integrative approach to structural biology [83–85].

5.2.1 Chemical Cross-Linkers

The power of XL-MS comes from the availability of a wide variety of different cross-linkers that impose specific distance constraints on the interactions being probed. These can be tailored to the question at hand, with the selection of cross-linker lengths placing different constraints on the interactions that can be studied. Similarly, the biochemical specificity of these linkers can be used to look at interactions between specific functional groups. Most commonly used are homobifunctional cross-linkers that join primary amines [86], i.e., lysine residues or N-termini, with spacer arm lengths ranging from ~3 Å to ~35 Å.

Heterobifunctional linkers (in contrast to homobifunctional ones) allow different groups to be targeted, for example, joining amine to carboxyl (aspartate, glutamate, C-termini) groups. More nuanced experiments can be performed using some of the more exotic linkers that are currently being produced. Heterobifunctional photoreactive cross-linkers such as aryl azides are attractive for in vivo applications, as they are inert until photoactivation, at which point they rapidly form nonspecific cross-links with different chemical moieties in their immediate environment. Photoreactive analogs of some amino acids have also been discovered, enabling incorporation of linkers into the protein sequences themselves [87].

5.2.2 Notable Applications of XL-MS

XL-MS is ideal for looking at flexible complexes that cannot be observed using cryo-EM or X-ray crystallography. A good example of this is the family of SMC-kleisin complexes, which are essential for accurate cell division and are formed of heterodimeric, coiled-coil SMC subunits, joined by a disordered kleisin subunit to form a trimeric ring structure that entraps DNA. Several crystal structures of the various subunit interfaces (minus flexible regions) are available, but thus far XL-MS has been the only method that has had success modeling the topology of the entire complex [88]. Interestingly, cross-links between the two SMC arms suggest that, when not encircling DNA, the SMC arms are collapsed in on themselves.

A more formidable test of XL-MS comes from the ongoing effort to understand the structure of the nuclear pore complex [89]. Due to their enormous size (~120 MDa in humans, compared to ~3.5 MDa for the ribosome) and high degree of compositional variation between species, it is difficult to distinguish between subunits, many of which are paralogues of each other.

In such cases, XL-MS can provide essential information about the specific identity of different subunits and their contacts, allowing the identification of ambiguous subunits within larger cryo-EM electron density maps [90, 91].

5.3 Affinity Purification-Mass Spectrometry

In its simplest guise, AP-MS enables the identification and quantification of the interaction partners of a given protein. The general principle is as follows: a column containing beads capable of capturing your bait protein is prepared. Native cell extract (though sometimes overexpression of the protein of interest is required) is then washed over the column, leading to the capture of both the bait and proteins bound to it via co-immunoprecipitation. The eluent is generally subjected to peptide fractionation, and mass spectrometry is then used to quantify either the relative or absolute abundances of members of the purified complex. For high-throughput studies, multiple proteins are used as baits, enabling large interaction maps to be generated.

Though conceptually simple, affinity purification-mass spectrometry (AP-MS) is an enormously powerful technique, and several reviews have been written on the topic [92–94]. For the sake of brevity, the paragraphs that follow are limited to just the most important variations of a method that has been instrumental in achieving our current understanding of the protein interactome [95–98].

5.3.1 Single-Step Versus Tandem Affinity Purification

There are two approaches to affinity purification in widespread use—single-step and tandem affinity purifications (TAP) [99]. In the former, the bait protein is either expressed under completely endogenous conditions and captured using antibodies or expressed with a tag such as green fluorescent protein [100] and captured using methods appropriate to the tagging system. In contrast, TAP makes use of a unique TAP tag, which consists of a protein A domain and a calmodulin-binding peptide, linked by a Tobacco Etch Virus (TEV) protease cleavage site (Fig. 3). This tag enables a two-step purification procedure that results in stringent purification of complexes, though sometimes at the expense of weak but specific interactions.

TAP necessarily requires tagging of the bait protein, but in the single-step procedure, it is possible to avoid this if desired, in which case it is referred to as endogenous purification. There are some straightforward trade-offs to consider when deciding whether to use endogenous or tagged proteins. For non-tagged baits, there is the benefit of capturing the protein in its native state. However, this comes at the substantial cost (both in time and money) of having to raise specific antibodies against the protein in question. Furthermore, there are difficult issues associated with cross-reactivity and specificity when using antibodies, particularly in studies where multiple proteins are being targeted. Although there are methods

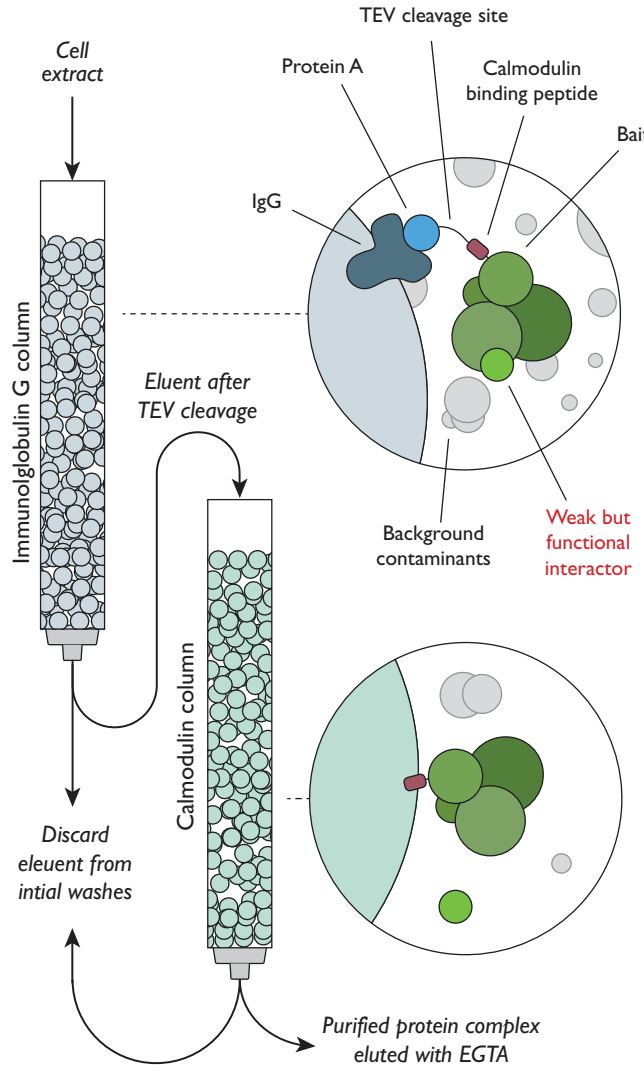


Fig. 3 Tandem affinity purification protocol. TAP differs from single-step purification procedures in that it requires two distinct washing steps. This is possible due to the TAP tag, which consists of a protein A domain linked to bait protein via a TEV cleavage site and a calmodulin-binding peptide. Cell extract is passed through the first IgG column, which captures the bait protein via the protein A domain. By adding TEV protease to the column, the bait protein and its interactors are released from the column. This eluent is then added to a second column containing calmodulin beads to which the calmodulin-binding peptide attaches. Addition of ethylene glycol tetraacetic acid causes the protein complex to be released from the beads. Due to the intense washing process that the protein complex undergoes, there is a high chance of weak interactors being removed in addition to nonspecific contaminants. Single-step procedures are more likely to retain these interactions at the expense of overall sample purity

that attempt to deal with these issues (most notably QUICK [101]), in most large-scale studies, prior tagging of bait proteins is likely to be more practical.

When using AP-MS to carry out interactome studies, there are some compelling advantages to using single-step procedures over TAP. The purpose of TAP is to remove as many possible contaminants or nonspecific interactors from the purified protein complex. This was necessary in the early days of MS, as it was not possible to quantify protein abundances, and thus contaminants in the sample would be erroneously annotated as members of the protein complex. However, there have been enormous improvements in the sensitivity of mass spectrometers since TAP was first described. With these improvements, particularly in the area of label-free quantification, it has become possible to discriminate between proteins present at biologically significant concentrations and background noise. In doing so, the importance of weak, non-obligate interactions between proteins has become more apparent [96, 102]. Since TAP removes these weak interactors, its utility is becoming increasingly restricted to situations where extremely pure protein is required, e.g., for crystallization. Therefore, single-step procedures combined with accurate quantification should generally be considered preferable to TAP for large-scale studies. Following this reasoning, a promising new technique named affinity enrichment purification has recently been described that deliberately uses only very mild washing steps [103].

5.3.2 Quantification of Protein Abundances

The emergence of MS, via both label-based and label-free methods, has had a transformative effect on the field of proteomics. A major benefit arising from the ability to quantify protein abundances is that it allows the stoichiometry of protein complexes to be determined. This is essential for distinguishing obligate interactions from transient ones and more generally for providing a complete characterization of the complex. The difficulty in using MS as a quantitative tool is that, while the location of peaks on the spectrum allows identification of peptides, peak intensity alone is not sufficient to determine peptide abundance. Label-based methods such as SILAC [104] and iTRAQ [105] (among others [106, 107]) allow for either relative or absolute quantification (through metabolic incorporation of amino acid isotopes in the case of SILAC and N-terminal isobaric tags in iTRAQ).

A significant drawback to label-based methods is their cost, which can be prohibitive. An alternative approach is label-free quantification (LFQ), which in general relies on either spectral counting [108, 109] or peak intensity-based algorithms. Spectral counting is a conceptually simple, semiquantitative approach and has been widely used (and possibly abused [110]). Intensity-based algorithms undoubtedly offer more accurate quantification; for the

interested reader, comparative analyses and reviews of several available methods are available [111, 112]. One recently developed algorithm of note that has been enthusiastically received by the community is MaxLFQ [113], which is available as part of the larger MaxQuant software package [114].

6 Conclusions

As we have seen, the study of protein complexes is currently undergoing tremendous changes brought about by the recent breakthroughs in structural biology, the emergence of mass spectrometry as a quantitative tool, and ongoing developments in computational techniques. The methods presented here offer a broad selection of those that can be used to study the physical characteristics and behavior of protein complexes, although there are some omissions, such as small-angle X-ray scattering [115] and hydrogen-deuterium exchange MS [116].

A reoccurring theme in current studies of protein complexes is the overlap between many different fields concerned with characterizing protein complexes. Many of these overlaps have had a synergistic effect on the technologies involved. This has been most obvious in cryo-EM, where hardware improvements have directly driven the development of new image processing software, but many other examples exist across structural biology and further afield. To point out a few explicitly, homology modeling has enabled much faster processing of diffraction patterns and electron density maps, improvements in purification techniques benefit essentially all of the non-computational techniques we have discussed, and many of the advances in imaging in cryo-EM will likely be transferable to XFELs.

This leaking of technologies across fields has facilitated the rise of integrative structural biology, which is becoming the most powerful approach to investigate protein complexes. Many of the most impressive structures published in the past couple of years have been the product of combinations of methods, including the transcribing mammalian pol II complex [117], the nuclear pore complex mRNA export platform [118], and the Mediator complex [119]. A second common feature of all of these papers is their focus on mechanistic descriptions of function or assembly, demonstrating a welcome move away from purely descriptive studies. Given the direction the field is moving in, early career structural biologists are advised not to be content with specializing in one method or the other [21, 120] and should endeavor to be at least familiar with most of the topics covered here.

The shift from purely descriptive studies to mechanistic ones emphasizes the fact that there is more to proteins and protein complexes than simple descriptions of structure. Of particular

importance, there is much to be gained from understanding the assembly process of protein complexes. Thanks to native MS studies, it is now well established that this occurs along ordered, thermodynamically favorable pathways, and papers on the topic have been published continually since this was first demonstrated [80, 81, 121–123].

On a larger scale, inventive use of mass spectrometry is enabling rapid improvement in our understanding of how individual protein complexes fit into the wider proteome. In a standout study from the group of Matthias Mann [96], the proteome of HeLa cells was quantified in such a way as to accurately capture interaction stoichiometries and global cellular abundances. Although not unexpected, the results from this work clearly demonstrate that the large majority of interactions, though important, are fairly weak. In contrast, stable complexes formed from interactions with stoichiometric ratios on the order of 1:1 are significantly less common but nonetheless highly connected through these weaker interactions.

The long-term objective of the techniques discussed in this review is to give a complete and unified understanding of the cellular proteome, in both its constituent parts and its behavior at scale. The progress made toward this aim would scarcely have been imaginable to the researchers who first began studying proteins in the 1950s, and there is no reason to suspect that the next 50 years will not see even greater progress. In many fields, there are novel technologies that will be revolutionary in years to come—perhaps nowhere more so than with the development of XFELs and serial femtosecond X-ray crystallography—and will be fascinating to see the new studies that this technology enables. Alternatively, cryo-EM may continue along its current trajectory to overtake crystallography as the go-to method in structural biology.

In the field of MS, although there are no obviously disruptive technologies on the horizon, continuing improvements in the sensitivity and accuracy of detectors are assured. Algorithmic development in MS is another area in which improvements are needed. Currently, there are seemingly intractable issues with peptide discrimination and quantification that need addressing, as evidenced by the first serious attempts to map the human proteome [91, 124, 125]. Nonetheless, the inherent versatility of the method across different cellular scales ensures the field's relevance in the decades to come, particularly as we move toward single-cell biology [126].

Acknowledgment

J.M. is supported by a Medical Research Council Career Development Award (MR/M02122X/1).

References

1. Stanley WM (1935) Isolation of a crystalline protein possessing the properties of tobacco-mosaic virus. Science 81:644–645
2. Schrödinger E (1947) What is life? The physical aspect of the living cell. Cambridge University Press, Cambridge
3. Dronamraju KR (1999) Erwin Schrödinger and the origins of molecular biology. Genetics 153:1071–1076
4. Fraenkel-Conrat H, Williams RC (1955) Reconstitution of active tobacco mosaic virus from its inactive protein and nucleic acid components. Proc Natl Acad Sci U S A 41:690–698. https://doi.org/10.1073/pnas.41.10.690
5. Perutz MF, Rossmann MG, Cullis AF et al (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature 185:416–422
6. Schluenzen F, Tocilj A, Zarivach R et al (2000) Structure of functionally activated small ribosomal subunit at 3.3Å resolution. Cell 102:615–623. https://doi.org/10.1016/S0092-8674(00)00084-2
7. Ramakrishnan V, Wimberly BT, Brodersen DE et al (2000) Structure of the 30S ribosomal subunit. Nature 407:327–339. https://doi.org/10.1038/35030006
8. Ban N, Nissen P, Hansen J et al (2000) The complete atomic structure of the large ribosomal subunit at 2.4Å resolution. Science 289:905–920. https://doi.org/10.1126/science.289.5481.905
9. Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. Nature 340:245–246. https://doi.org/10.1038/340245a0
10. Rajagopala SV, Sikorski P, Kumar A et al (2014) The binary protein-protein interaction landscape of Escherichia coli. Nat Biotechnol 32:285–290
11. Karas M, Bachmann D, Hillenkamp F (1985) Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. Anal Chem 57:2935–2939. https://doi.org/10.1021/ac00291a042
12. Fenn JB, Mann M, Meng CK et al (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246:64–71. https://doi.org/10.1126/science.2675315
13. Rosano GL, Ceccarelli EA (2014) Recombinant protein expression in Escherichia coli: advances and challenges. Front Microbiol. https://doi.org/10.3389/fmicb.2014.00172
14. Wells JN, Bergendahl LT, Marsh JA (2016) Operon gene order is optimized for ordered protein complex assembly. Cell Rep 14:679–685. https://doi.org/10.1016/j.celrep.2015.12.085
15. Shieh Y-W, Minguez P, Bork P et al (2015) Operon structure and cotranslational subunit association direct protein assembly in bacteria. Science 350:678–680. https://doi.org/10.1126/science.aac8171
16. Poulsen C, Holton S, Geerlof A et al (2010) Stoichiometric protein complex formation and over-expression using the prokaryotic native operon structure. FEBS Lett 584:669–674. https://doi.org/10.1016/j.febslet.2009.12.057
17. Ni QZ, Daviso E, Can TV et al (2013) High frequency dynamic nuclear polarization. Acc Chem Res 46:1933–1941. https://doi.org/10.1021/ar300348n
18. Jia B, Jeon CO (2016) High-throughput recombinant protein expression in Escherichia coli: current status and future perspectives. Open Biol 6:160196. https://doi.org/10.1098/rsob.160196
19. Norimatsu Y, Hasegawa K, Shimizu N, Toyoshima C (2017) Protein–phospholipid interplay revealed with crystals of a calcium pump. Nature 545:193–198. https://doi.org/10.1038/nature22357
20. Bragg WH, Bragg WL (1913) The reflection of X-rays by crystals. Proc R Soc Math Phys Eng Sci 88:428–438. https://doi.org/10.1098/rspa.1913.0040
21. Shi Y (2014) A glimpse of structural biology through X-ray crystallography. Cell 159:995–1014. https://doi.org/10.1016/j.cell.2014.10.051
22. Neutze R, Wouts R, van der Spoel D et al (2000) Potential for biomolecular imaging with femtosecond X-ray pulses. Nature 406:752–757. https://doi.org/10.1038/35021099
23. Taylor G (2003) The phase problem. Acta Crystallogr D Biol Crystallogr 59:1881–1890. https://doi.org/10.1107/S0907444903017815
24. Robertson JM (1936) An X-ray study of the phthalocyanines. Part II. Quantitative structure determination of the metal-free compound. J Chem Soc:1195–1209
25. Dauter Z (2005) Use of polynuclear metal clusters in protein crystallography. Comptes Rendus Chim 8:1808–1814. https://doi.org/10.1016/j.crci.2005.02.032
26. Nozawa K, Schneider TR, Cramer P (2017) Core Mediator structure at 3.4 Å extends

model of transcription initiation complex. Nature 545:248–251. https://doi.org/10.1038/nature22328
27. Hendrickson W (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science 254:51–58. https://doi.org/10.1126/science.1925561
28. Berman HM, Westbrook J, Feng Z et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242
29. McCoy AJ, Grosse-Kunstleve RW, Adams PD et al (2007) Phaser crystallographic software. J Appl Crystallogr 40:658–674. https://doi.org/10.1107/S0021889807021206
30. Winn MD, Ballard CC, Cowtan KD et al (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr 67:235–242. https://doi.org/10.1107/S0907444910045749
31. Merk A, Bartesaghi A, Banerjee S et al (2016) Breaking cryo-EM resolution barriers to facilitate drug discovery. Cell 165:1698–1707. https://doi.org/10.1016/j.cell.2016.05.040
32. Bai X, McMullan G, Scheres SH (2015) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40:49–57. https://doi.org/10.1016/j.tibs.2014.10.005
33. McMullan G, Chen S, Henderson R, Faruqi AR (2009) Detective quantum efficiency of electron area detectors in electron microscopy. Ultramicroscopy 109:1126–1143. https://doi.org/10.1016/j.ultramic.2009.04.002
34. Dainty JC, Shaw R (1975) Image science, principles, analysis and evaluation of photographic type imaging processes. Academic, London
35. McMullan G, Clark AT, Turchetta R, Faruqi AR (2009) Enhanced imaging in low dose electron microscopy using electron counting. Ultramicroscopy 109:1411–1416. https://doi.org/10.1016/j.ultramic.2009.07.004
36. Danev R, Buijsse B, Khoshouei M et al (2014) Volta potential phase plate for in-focus phase contrast transmission electron microscopy. Proc Natl Acad Sci U S A 111:15635–15640. https://doi.org/10.1073/pnas.1418377111
37. Danev R, Baumeister W (2016) Cryo-EM single particle analysis with the Volta phase plate. elife 5:1–14. https://doi.org/10.7554/eLife.13046
38. Khoshouei M, Radjainia M, Baumeister W, Danev R (2017) Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. Nat Commun 8:16099. https://doi.org/10.1038/ncomms16099
39. Bai X, Fernandez IS, McMullan G, Scheres SH (2013) Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. elife. https://doi.org/10.7554/eLife.00461
40. Li X, Mooney P, Zheng S et al (2013) Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat Methods 10:584–590
41. Henderson R, Glaeser RM (1985) Quantitative analysis of image contrast in electron micrographs of beam-sensitive crystals. Ultramicroscopy 16:139–150. https://doi.org/10.1016/0304-3991(85)90069-5
42. Scheres SHW (2012) RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol 180:519–530. https://doi.org/10.1016/j.jsb.2012.09.006
43. Scheres SHW (2014) Beam-induced motion correction for sub-megadalton cryo-EM particles. elife 3:e03665
44. Sigworth FJ (1998) A maximum-likelihood approach to single-particle image refinement. J Struct Biol 122:328–339. https://doi.org/10.1006/jsbi.1998.4014
45. Scheres SHW, Gao H, Valle M et al (2007) Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nat Methods 4:27–29. https://doi.org/10.1038/nmeth992
46. Lyumkis D, Brilot AF, Theobald DL, Grigorieff N (2013) Likelihood-based classification of cryo-EM images using FREALIGN. J Struct Biol 183:377–388. https://doi.org/10.1016/j.jsb.2013.07.005
47. Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA (2017) cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nat Methods 14:290–296. https://doi.org/10.1038/nmeth.4169
48. Wisedchaisri G, Reichow SL, Gonen T (2011) Advances in structural and functional analysis of membrane proteins by electron crystallography. Structure 19:1381–1393. https://doi.org/10.1016/j.str.2011.09.001
49. Galaz-Montoya JG, Ludtke SJ (2017) The advent of structural biology in situ by single particle cryo-electron tomography. Biophys Rep 3:17–35. https://doi.org/10.1007/s41048-017-0040-0
50. Bharat TAM, Scheres SHW (2016) Resolving macromolecular structures from electron cryotomography data using subtomogram averaging in RELION. Nat Protoc 11:2054–2065. https://doi.org/10.1038/nprot.2016.124
51. Leschziner AE, Nogales E (2006) The orthogonal tilt reconstruction method: an approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. J Struct Biol 153:284–299. https://doi.org/10.1016/j.jsb.2005.10.012
52. Schur FKM, Obr M, Hagen WJH et al (2016) An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation.

Science 353:506–508. https://doi.org/10.1126/science.aaf9620
53. Pervushin K, Riek R, Wider G, Wüthrich K (1997) Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci U S A 94:12366–12371. https://doi.org/10.1073/pnas.94.23.12366
54. Sattler M, Fesik SW (1996) Use of deuterium labeling in NMR: overcoming a sizeable problem. Structure 4:1245–1249. https://doi.org/10.1016/S0969-2126(96)00133-5
55. Ollerenshaw JE, Tugarinov V, Kay LE (2003) Methyl TROSY: explanation and experimental verification. Magn Reson Chem 41:843–852. https://doi.org/10.1002/mrc.1256
56. Zhang H, van Ingen H (2016) Isotope-labeling strategies for solution NMR studies of macromolecular assemblies. Curr Opin Struct Biol 38:75–82. https://doi.org/10.1016/j.sbi.2016.05.008
57. Liu D, Xu R, Cowburn D (2009) Segmental isotopic labeling of proteins for nuclear magnetic resonance. Methods Enzymol 462:151–175
58. Rosenzweig R, Farber P, Velyvis A et al (2015) ClpB N-terminal domain plays a regulatory role in protein disaggregation. Proc Natl Acad Sci U S A 112:e6872. https://doi.org/10.1073/pnas.1512783112
59. Frederick KK, Michaelis VK, Caporini MA et al (2017) Combining DNP NMR with segmental and specific labeling to study a yeast prion protein strain that is not parallel in-register. Proc Natl Acad Sci U S A 114:3642–3647. https://doi.org/10.1073/pnas.1619051114
60. Wells M, Tidow H, Rutherford TJ et al (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. Proc Natl Acad Sci U S A 105:5762–5767. https://doi.org/10.1073/pnas.0801353105
61. Marsh JA, Dancheck B, Ragusa MJ et al (2010) Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. Structure 18:1094–1103. https://doi.org/10.1016/j.str.2010.05.015
62. Mittag T, Marsh J, Grishaev A et al (2010) Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. Structure 18:494–506. https://doi.org/10.1016/j.str.2010.01.020
63. Marsh JA, Teichmann SA, Forman-Kay JD (2012) Probing the diverse landscape of protein flexibility and binding. Curr Opin Struct Biol 22:643–650. https://doi.org/10.1016/j.sbi.2012.08.008
64. Bozoky Z, Krzeminski M, Muhandiram R et al (2013) Regulatory R region of the CFTR chloride channel is a dynamic integrator of phospho-dependent intra- and intermolecular interactions. Proc Natl Acad Sci U S A 110:E4427–E4436. https://doi.org/10.1073/pnas.1315104110
65. Andrew ER, Bradbury A, Eades RG (1958) Nuclear magnetic resonance spectra from a crystal rotated at high speed. Nature 182:1659–1659. https://doi.org/10.1038/1821659a0
66. Hansen SK, Bertelsen K, Paaske B et al (2015) Solid-state NMR methods for oriented membrane proteins. Prog Nucl Magn Reson Spectrosc 88–89:48–85. https://doi.org/10.1016/j.pnmrs.2015.05.001
67. Loquet A, Sgourakis NG, Gupta R et al (2012) Atomic model of the type III secretion system needle. Nature 486:276–279. https://doi.org/10.1038/nature11079
68. Kaplan M, Cukkemane A, van Zundert GCP et al (2015) Probing a cell-embedded megadalton protein complex by DNP-supported solid-state NMR. Nat Methods 12:5–9. https://doi.org/10.1038/nmeth.3406
69. Huang C, Kalodimos CG (2017) Structures of large protein complexes determined by nuclear magnetic resonance spectroscopy. Annu Rev Biophys 46:317–336. https://doi.org/10.1146/annurev-biophys-070816-033701
70. Song H, Hanlon N, Brown NR et al (2001) Phosphoprotein-protein interactions revealed by the crystal structure of kinase-associated phosphatase in complex with phosphoCDK2. Mol Cell 7:615–626
71. Krusemark CJ, Frey BL, Belshaw PJ, Smith LM (2009) Modifying the charge state distribution of proteins in electrospray ionization mass spectrometry by chemical derivatization. J Am Soc Mass Spectrom 20:1617–1625. https://doi.org/10.1016/j.jasms.2009.04.017
72. Radionova A, Filippov I, Derrick PJ (2016) In pursuit of resolution in time-of-flight mass spectrometry: a historical perspective. Mass Spectrom Rev 35:738–757. https://doi.org/10.1002/mas.21470
73. Hu Q, Noll RJ, Li H et al (2005) The Orbitrap: a new mass spectrometer. J Mass Spectrom 40:430–443. https://doi.org/10.1002/jms.856
74. Wilm MS, Mann M (1994) Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? Int J Mass Spectrom Ion Process 136:167–180. https://doi.org/10.1016/0168-1176(94)04024-9
75. El-Faramawy A, Siu KWM, Thomson BA (2005) Efficiency of nano-electrospray ionization. J Am Soc Mass Spectrom 16:1702–1707. https://doi.org/10.1016/j.jasms.2005.06.011

76. Sobott F, Hernández H, McCammon MG et al (2002) A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. Anal Chem 74:1402–1407. https://doi.org/10.1021/ac0110552
77. Hernandez H, Robinson CV (2007) Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. Nat Protoc 2:715–726. https://doi.org/10.1038/nprot.2007.73
78. Sobott F, Benesch JLP, Vierling E, Robinson CV (2002) Subunit exchange of multimeric protein complexes. J Biol Chem 277:38921–38929. https://doi.org/10.1074/jbc.M206060200
79. Laganowsky A, Reading E, Allison TM et al (2014) Membrane proteins bind lipids selectively to modulate their structure and function. Nature 510:172–175. https://doi.org/10.1038/nature13419
80. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA (2008) Assembly reflects evolution of protein complexes. Nature 453:1262–1265. https://doi.org/10.1038/nature06942
81. Marsh JA, Hernández H, Hall Z et al (2013) Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell 153:461–470
82. Ahnert SE, Marsh JA, Hernández H et al (2015) Principles of assembly reveal a periodic table of protein complexes. Science 350:aaa2245. https://doi.org/10.1126/science.aaa2245
83. Stengel F, Aebersold R, Robinson CV (2012) Joining forces: integrating proteomics and cross-linking with the mass spectrometry of intact complexes. Mol Cell Proteomics 11:R111.014027–R111.014027. https://doi.org/10.1074/mcp.R111.014027
84. Ward AB, Sali A, Wilson IA (2013) Integrative structural biology. Science 339:913–915. https://doi.org/10.1126/science.1228565
85. van den Bedem H, Fraser JS (2015) Integrative, dynamic structural biology at atomic resolution - it's about time. Nat Methods 12:307–318. https://doi.org/10.1038/nmeth.3324
86. Leitner A, Faini M, Stengel F, Aebersold R (2016) Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. Trends Biochem Sci 41:20–32. https://doi.org/10.1016/j.tibs.2015.10.008
87. Suchanek M, Radzikowska A, Thiele C (2005) Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells. Nat Methods 2:261–268. https://doi.org/10.1038/nmeth752
88. Barysz H, Kim JH, Chen ZA et al (2015) Three-dimensional topology of the SMC2/SMC4 subcomplex from chicken condensin I revealed by cross-linking and molecular modelling. Open Biol 5:150005. https://doi.org/10.1098/rsob.150005
89. Beck M, Hurt E (2016) The nuclear pore complex: understanding its function through structural insight. Nat Rev Mol Cell Biol. https://doi.org/10.1038/nrm.2016.147
90. Bui KH, von Appen A, DiGuilio AL et al (2013) Integrated structural analysis of the human nuclear pore complex scaffold. Cell 155:1233–1243. https://doi.org/10.1016/j.cell.2013.10.055
91. Shi Y, Fernandez-Martinez J, Tjioe E et al (2014) Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. Mol Cell Proteomics 13:2927–2943. https://doi.org/10.1074/mcp.M114.041673
92. Oeffinger M (2012) Two steps forward-one step back: advances in affinity purification mass spectrometry of macromolecular complexes. Proteomics 12:1591–1608. https://doi.org/10.1002/pmic.201100509
93. Morris JH, Knudsen GM, Verschueren E et al (2014) Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions. Nat Protoc 9:2539–2554. https://doi.org/10.1038/nprot.2014.164
94. Aebersold R, Mann M (2016) Mass-spectrometric exploration of proteome structure and function. Nature 537:347–355. https://doi.org/10.1038/nature19949
95. Malovannaya A, Lanz RB, Jung SY et al (2011) Analysis of the human endogenous coregulator complexome. Cell 145:787–799. https://doi.org/10.1016/j.cell.2011.05.006
96. Hein MY, Hubner NC, Poser I et al (2015) A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell 163:712–723. https://doi.org/10.1016/j.cell.2015.09.053
97. Huttlin EL, Ting L, Bruckner RJ et al (2015) The BioPlex network: a systematic exploration of the human interactome. Cell 162:425–440. https://doi.org/10.1016/j.cell.2015.06.043
98. Wan C, Borgeson B, Phanse S et al (2015) Panorama of ancient metazoan macromolecular complexes. Nature 525:339–344. https://doi.org/10.1038/nature14877
99. Rigaut G, Shevchenko A, Rutz B et al (1999) A generic protein purification method for protein complex characterization and proteome

exploration. Nat Biotechnol 17:1030–1032. https://doi.org/10.1038/13732
100. Hubner NC, Bird AW, Cox J et al (2010) Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. J Cell Biol 189:739–754. https://doi.org/10.1083/jcb.200911091
101. Selbach M, Mann M (2006) Protein interaction screening by quantitative immunoprecipitation combined with knockdown (QUICK). Nat Methods 3:981–983. https://doi.org/10.1038/nmeth972
102. Perkins JR, Diboun I, Dessailly BH et al (2010) Transient protein-protein interactions: structural, functional, and network properties. Structure 18:1233–1243. https://doi.org/10.1016/j.str.2010.08.007
103. Keilhauer EC, Hein MY, Mann M (2015) Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). Mol Cell Proteomics 14:120–135. https://doi.org/10.1074/mcp.M114.041012
104. Ong S-E, Blagoev B, Kratchmarova I et al (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1(5):376–386. https://doi.org/10.1074/mcp.M200025-MCP200
105. Ross PL (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 3:1154–1169. https://doi.org/10.1074/mcp.M400129-MCP200
106. Gygi SP, Rist B, Gerber SA et al (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17:994–999. https://doi.org/10.1038/13690
107. Thompson A, Schäfer J, Kuhn K et al (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 75:1895–1904. https://doi.org/10.1021/ac0262560
108. Liu H, Sadygov RG, Yates JR (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76:4193–4201. https://doi.org/10.1021/ac0498563
109. Zybailov B, Coleman MK, Florens L, Washburn MP (2005) Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. Anal Chem 77:6218–6224. https://doi.org/10.1021/ac050846r
110. Lundgren DH, Hwang S-I, Wu L, Han DK (2010) Role of spectral counting in quantitative proteomics. Expert Rev Proteomics 7:39–53. https://doi.org/10.1586/epr.09.69
111. Nahnsen S, Bielow C, Reinert K, Kohlbacher O (2013) Tools for label-free peptide quantification. Mol Cell Proteomics 12:549–556. https://doi.org/10.1074/mcp.R112.025163
112. Fabre B, Lambour T, Bouyssié D et al (2014) Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. EuPA Open Proteom 4:82–86. https://doi.org/10.1016/j.euprot.2014.06.001
113. Cox J, Hein MY, Luber CA et al (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol Cell Proteomics 13:2513–2526. https://doi.org/10.1074/mcp.M113.031591
114. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26:1367–1372. https://doi.org/10.1038/nbt.1511
115. Mertens HDT, Svergun DI (2010) Structural characterization of proteins and complexes using small-angle X-ray solution scattering. J Struct Biol 172:128–141. https://doi.org/10.1016/j.jsb.2010.06.012
116. Zhang Z, Vachet RW (2015) Kinetics of protein complex dissociation studied by hydrogen/deuterium exchange and mass spectrometry. Anal Chem 87:11777–11783. https://doi.org/10.1021/acs.analchem.5b03123
117. Bernecky C, Herzog F, Baumeister W et al (2016) Structure of transcribing mammalian RNA polymerase II. Nature 529:551–554. https://doi.org/10.1038/nature16482
118. Fernandez-Martinez J, Kim SJ, Shi Y et al (2016) Structure and function of the nuclear pore complex cytoplasmic mRNA export platform. Cell 167:1215–1228.e25. https://doi.org/10.1016/j.cell.2016.10.028
119. Tsai K-L, Yu X, Gopalan S et al (2017) Mediator structure and rearrangements required for holoenzyme formation. Nature 544:196–201. https://doi.org/10.1038/nature21393
120. Cassiday L (2014) Structural biology: more than a crystallographer. Nature 505:711–713. https://doi.org/10.1038/nj7485-711a
121. Appolaire A, Girard E, Colombo M et al (2014) Small-angle neutron scattering reveals the assembly mode and oligomeric architec-

ture of TET, a large, dodecameric aminopeptidase. Acta Crystallogr D Biol Crystallogr 70:2983–2993. https://doi.org/10.1107/S1399004714018446
122. Macek P, Kerfah R, Erba EB et al (2017) Unraveling self-assembly pathways of the 468-kDa proteolytic machine TET2. Sci Adv 3:e1601601
123. Mallik S, Kundu S (2017) Coevolutionary constraints in the sequence-space of macromolecular complexes reflect their self-assembly pathways. Proteins 85:1183–1189. https://doi.org/10.1002/prot.25292
124. Wilhelm M, Schlegl J, Hahne H et al (2014) Mass-spectrometry-based draft of the human proteome. Nature 509:582–587. https://doi.org/10.1038/nature13319
125. Ezkurdia I, Vázquez J, Valencia A, Tress M (2014) Analyzing the first drafts of the human proteome. J Proteome Res 13:3854–3855. https://doi.org/10.1021/pr500572z
126. Macaulay IC, Ponting CP, Voet T (2017) Single-cell multiomics: multiple measurements from single cells. Trends Genet 33:155–168. https://doi.org/10.1016/j.tig.2016.12.003
127. Chen S, Wu J, Lu Y et al (2016) Structural basis for dynamic regulation of the human 26S proteasome. Proc Natl Acad Sci U S A 113:12991–12996. https://doi.org/10.1073/pnas.1614614113

Chapter 2

High-Throughput Electron Cryo-tomography of Protein Complexes and Their Assembly

Louie D. Henderson and Morgan Beeby

Abstract

Electron cryo-tomography and subtomogram averaging enable visualization of protein complexes in situ, in three dimensions, in a near-native frozen-hydrated state to nanometer resolutions. To achieve this, intact cells are vitrified and imaged over a range of tilts within an electron microscope. These images can subsequently be reconstructed into a three-dimensional volume representation of the sample cell. Because complexes are visualized in situ, crucial insights into their mechanism, assembly process, and dynamic interactions with other proteins become possible. To illustrate the electron cryo-tomography workflow for visualizing protein complexes in situ, we describe our workflow of preparing samples, imaging, and image processing using Leginon for data collection, IMOD for image reconstruction, and PEET for subtomogram averaging.

Key words Electron cryo-tomography, Subtomogram averaging, Molecular machines, Protein self-assembly, Structural biology

1 Introduction

Understanding the assembly and mechanism of protein machines lies at the heart of contemporary molecular biology. This aim, however, is challenging, given that proteins function in the highly complex and dynamic context of the cell interior or (even more challenging) within membranes. In recent years a range of techniques have enable us to move on from the reductionist approaches that involved the purification of proteins necessary during the development of molecular biology.

Electron cryo-tomography (ECT) is a variant of electron cryo-microscopy which enables the in vivo observation and structure determination of macromolecular protein complexes in situ, in three dimensions, in a near-native frozen-hydrated state [1]. The method involves vitrification of cells by flash-freezing them on an electron microscopy grid to preserve them in a near-native state. The grid is then inserted into an electron microscope and imaged

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764,
https://doi.org/10.1007/978-1-4939-7759-8_2,

repeatedly while being rotated around an axis. This results in a series of 2D projections which can be used to computationally reconstruct the 3D volume of the sample, enabling 3D views into whole cells, including the soluble and membrane-embedded protein complexes contained within them [2]. ECT therefore provides information on the in situ context of protein complexes imaged.

In addition to in situ context, however, identical subtomograms can be extracted and averaged in a method known as *subtomogram averaging*—effectively, "in situ structural biology" [3]. Averaging boosts the signal-to-noise ratio of aligned particles, enabling discernment of higher-resolution features, and has enabled determination to nanometer resolution the structures of diverse soluble and membrane-bound protein complexes in situ [4, 5], enabling insights into assembly pathways, structure/function relationships, and protein complex evolution. Because subtomogram averaging can be performed on particles in their native context, it is particularly powerful for imaging fragile or transient assembly intermediates of large protein complexes [6–8]. Advances in methodology have enabled resolutions below 10 Å to be determined, bridging the gap between atomic and cellular scales [3, 9, 10]. These advances are a result of technological and methodological improvements such as identification of optimally thin samples [10] (thinner samples produce higher-quality images), a dose symmetric tilt scheme [11] that optimizes the distribution of sample-damaging electron dosage, and post-processing software developments such as CTF correction to correct for imaging artifacts resulting from aberrations in the optics of contemporary electron microscopes [12].

ECT and subtomogram averaging are uniquely powerful due to their versatility and ability to span the divide between structural and cellular biologies. Any given research question using ECT will therefore need to determine the optimal contributions of resolution and context. Here we describe the workflow that we have developed that strike a balance between resolution and context to address questions on the bacterial flagellar motor. The flagellar motor is a large protein complex integral to two membranes, making ECT an ideal method by which to study it. Many species assemble a single motor per cell, however, necessitating the streamlined and high-throughput data acquisition pipeline that we discuss below.

1.1 Identifying and Preparing a Suitable Specimen for ECT

Determining whether ECT will be appropriate for studying a particular protein complex depends on the size, location, and abundance of the complex, and thickness of the surrounding cell. The complex needs to be sufficiently large to be identifiable within tomograms. The location within the cell (the cell periphery will be thinner than other sites) and abundance of complexes (multiple

copies of a complex per tomogram) will determine the number and quality of particles imaged for averaging.

It is desirable to obtain the thinnest possible sample for ECT as these produce images with the highest signal-to-noise ratio. Samples thicker than ~500 nm result in excessive inelastic or multiple scattering [13]. Thick samples can be dealt with by a number of means, including switching to a naturally thinner species that contains the structure of interest or genetic, mechanical, or enzymatic thinning of samples [4]. Changing to a thinner species may often be the least invasive approach. In the case of bacteria, traditional model organisms such as *E. coli* or *Salmonella enterica* have widths >1 μm; alternate strains such as *E. coli* B/r H266 [14], or species such as *Campylobacter jejuni* or *Borrelia burgdorferi* [6, 8], have cell widths closer to 500 nm. Furthermore, species with polarly localized structures of interest are more tractable due to the thinner nature of the pole over the range of a tilt series. Genetic manipulation, such as the overexpression of FtsZ or deletion of the MinCDE system in bacteria, can be used to produce minicells that are often considerably thinner than 500 nm, which when combined with centrifugation enrichment schemes can provide optimal targets for ECT [15]; alternatively, some species become thinner in different growth media. A specimen can also be mechanically thinned by FIB milling, a process in which a vitrified sample is milled by an ion beam to thin sections above and below the area of interest, to produce thin lamellae of material of interest. This may be particularly important for larger eukaryotic cells [16], although FIB milling currently remains a low-throughput approach. Finally, treatment with enzymes such as lysozyme has been used to gently deflate whole cells immediately prior to imaging [17]. The protocols described below are based on our model organism, *C. jejuni*.

2 Materials

Electron cryo-tomography requires an extensive range of wet-lab, electron microscopy, and computational equipment. In brief, you will need access to:

1. Your specimen prepared in an appropriate way to express the structures you wish to image, applying previously mentioned thinning approaches as appropriate. Optimal growth media used for cell cultivation must be nonviscous to enable wicking of excess growth medium when blotting.
2. A contemporary electron microscope configured for cryopreserved specimen data collection (*see* **Note 1**), with associated tools and consumables.
3. This protocol describes data acquisition using Leginon. Version 3.0 or above must be installed (*see* **Note 2**). The Leginon

server must be installed on a computer that has a network connection to the microscopy control computer with the Leginon client installed. A variety of other commercial and free software is also available with a range of advantages [18, 19].

4. Depending on the amount of data and processing, at least one Linux workstation and an appropriate backup solution (*see* **Note 3**). Subtomogram averaging is CPU intensive so it is advantageous to have access to a multi-processor cluster. Our subtomogram averaging software of choice, PEET (*see* below), also enables use of excess unused computational power from other workstations on the network using passwordless SSH access. This is an economical approach suitable for ad hoc PEET usage in the absence of cluster access.
5. Data reconstruction and subtomogram averaging described here requires a recent version of IMOD [20] and PEET [21] installed.

3 Methods

3.1 Sample Vitrification and Grid Preparation by Blotting and Plunge-Freezing

The core philosophy to be observed during vitrification is to ensure the frozen specimen does not come into contact with atmospheric water vapor and does not warm up.

1. To prevent fiducial clumping, prepare gold fiducial pellets by mixing 133 μL of 10-nm-diameter gold fiducials with 33 μL 5% BSA. Vortex thoroughly for 10 s and centrifuge at high speed (>20,000 × *g*) for 15+ min. Immediately after centrifugation, remove the supernatant while being careful not to agitate the loose pellet. Keep fiducial pellets refrigerated until needed.
2. Set up your vitrification device according to standard procedures dependent on model, and configure it using the desired blotting parameters (*see* **Note 4**). If the vitrification system has a humidifier system, be sure it is set up and filled with sufficient water to humidify the sample chamber (*see* **Note 5**).
3. Select electron microscopy grids best suited to your specimen (*see* **Notes 6** and **7**). Set these aside until needed.
4. Resuspend cells of interest to a density that will provide an even cell distribution on the grid (*see* **Notes 8** and **9**).
5. Transfer sample and gold fiducial pellet on ice to site of vitrification, maintaining ice incubation throughout duration (*see* **Note 10**).
6. Begin the process of cooling down the outer reservoir of the vitrification device's cryocup with liquid nitrogen (*see* **Note 11**). Fill the central well with liquid nitrogen initially to aid

cooling but prevent further filling after initial boil off; this prevents nitrogen contaminating the cryogen.

7. While waiting for the cryocup to cool, glow discharge your grids. This step will require sample-dependent optimization (*see* **Note 12**).
8. Once the cryocup is cooled, within a fume hood, fill the central chamber with liquefied cryogen by directing the gas jet against the cold wall of the central chamber (*see* **Note 13**). Care should be taken to prevent contamination of the inner chamber with nitrogen before filling with cryogen. Tissue paper can be used to wick away any contaminating nitrogen which would otherwise bubble in the warmer central chamber cryogen. It is important from this point to maintain the nitrogen reservoirs surrounding the cryogen to prevent warming and to ensure a cushion of evaporating super-cooled nitrogen gas envelopes and protects grids that you will freeze in subsequent steps.
9. Place blotting filter paper to the blotting pads in the vitrification device. If a pathogen is being vitrified, add a disk of aluminum foil between the blotting paper and the blotting pad.
10. While the vitrification device chamber achieves 100% humidity, combine 30 μL of sample with the pre-prepared gold fiducial pellet (*see* **Note 14**).
11. Transfer precooled labeled grid boxes into the box slot immersed in liquid nitrogen in the outer reservoir of the cryocup. Ensure that the lid is loosened and positioned so that the first slot is exposed for grid transfer (*see* **Note 15**).
12. Grasp one of the glow discharge grids with vitrification device tweezers and slide the tweezers onto the end of the vitrification device rod (*see* **Note 16**).
13. When prompted, apply a small volume (typically 3 μL, although this can be adjusted to optimize vitreous ice quality) of sample mixed with fiducial markers to the grid.
14. Trigger the blotting and plunge-freezing process. This will be fully automated on most commercial systems.
15. Once the sample has been vitrified, top up the outer reservoir of the cryocup with liquid nitrogen and carefully move the vitrified grid into the desired slot within the labeled grid box. During this process, be careful to keep the grid under cryogen while also preventing contact with other surfaces wherever possible. Nevertheless, a cushion of evaporating super-cooled nitrogen gas envelopes and protects objects immediately above the surface of the liquid nitrogen, meaning grid transfer is safe if conducted rapidly.
16. Repeat the vitrification process until a sufficient numbers of grids have been frozen (*see* **Note 17**). Once completed, rotate

the lid of the grid box to the closed position using precooled tools to avoid warming the grid, and tighten the lid shut so that grid slots are no longer exposed. Transfer the grid box into long-term storage. In practice, frozen grids can be stored for months or years.

17. Record grids appropriately (*see* **Note 15**).
18. Leave the cryocup to thaw within a fume hood and empty the humidifier to avoid fungal growth.

3.2 Acquire Tomograms of the Vitrified Specimen

1. Ensure the microscope optics are aligned to a satisfactory standard.
2. Run the Leginon client on the main microscope control computer and (if applicable) a separate camera control computer. Run the Leginon server and establish a connection to the Leginon client(s) (*see* **Note 18**).
3. Establish the Leginon session with appropriate responses to dialogue boxes and launch the Leginon tomography application. The microscope/camera control client(s) should appear automatically at this point in their respective fields.
4. Establish a range of microscope preset configurations by importing appropriate presets from a previous session or creating a series of presets settings appropriate for your sample to be used at various steps in the data collection process (*see* **Note 19**). For tomography, the presets required are:
   (a) *gr* (grid atlas overview preset): low magnification for constructing a grid atlas of the entire grid
   (b) *sq* (square overview preset): low magnification for inspection of a square and targeting holes
   (c) *hl* (hole overview preset): medium magnification for holes and tomography targeting
   (d) *preview* (preview preset): high magnification with high defocus for previewing and assessing potential tomography targets
   (e) *fc* (focus preset): high magnification at focus for automatically setting focus for data acquisition.
   (f) *tomo* (tomography acquisition preset): data collection preset at desired defocus and magnification appropriate to biological question
5. Refine presets, even if imported, as instabilities in the microscope may mean previous settings are no longer optimal. Send each preset to the microscope in turn, center the beam, and update the preset in Leginon to the newly centered beam. Repeat until the presets do not change when switching between them.

6. Within the tomography node, open the settings dialogue box, go to the advanced settings, and change the tilt scheme and total electron dose as appropriate for your sample.
7. Close the column valves, retrieve the grid from storage, and load it into the microscope.
8. Perform a preliminary assessment of the ice quality and to see if the grid will contain sufficient targets for data collection (Fig. 1) (*see* **Note 20**). If the grid is unlikely to be satisfactory, load another grid.
9. Find an area of the grid that is free of contamination and is not broken. Use it to set eucentric height (*see* **Note 21**).
10. Send the *focus* preset to the microscope, establish focus, and update the *focus* preset so as to be at focus.
11. Identify an object that is distinguishable at all magnifications. In the presets manager, change to the *tomo* preset and click on the image shift alignment icon to align image shifts across magnifications (*see* **Note 22**).
12. Change to *tomo* preset and insert the objective aperture; within a blank square, take a dose measurement from the presets manager to calculate exposure times at different tilts.
13. Go to the *grid targeting* node, and calculate and acquire a grid atlas.
14. In the *square targeting* node, place a reference point in an empty square, and select one or more grid squares to acquire *sq* magnification images (*see* **Note 23**). Submit the targets for acquisition.
15. Insert the objective aperture and select hole targets within the *hole targeting* node. This number will vary dependent on the number of potential targets per hole target and the biological question being asked. Be sure to select a focus point in each square image. Submit the target queue when completed (*see* **Note 24**).
16. Within the tomography targeting node, select tomography targets (*see* **Note 25**) until a number sufficient for data collection time has been reached (*see* **Note 26**). Again, be sure to place a focus point within each hole image. Submit the target queue when finished.
17. Allow tilt series to run while maintaining nitrogen levels within the cryo-holder and cold trap to enable continue data collection.

3.3 Tomographic Reconstruction Using IMOD and Particle Picking

1. Using the image processing package IMOD, align the tilt series to produce tomographic reconstructions (*see* **Note 27**).
2. Open each tomogram file and note the presence of structures of interest: using the slicer window within IMOD, select the *Model* radio button and create contours for each structure

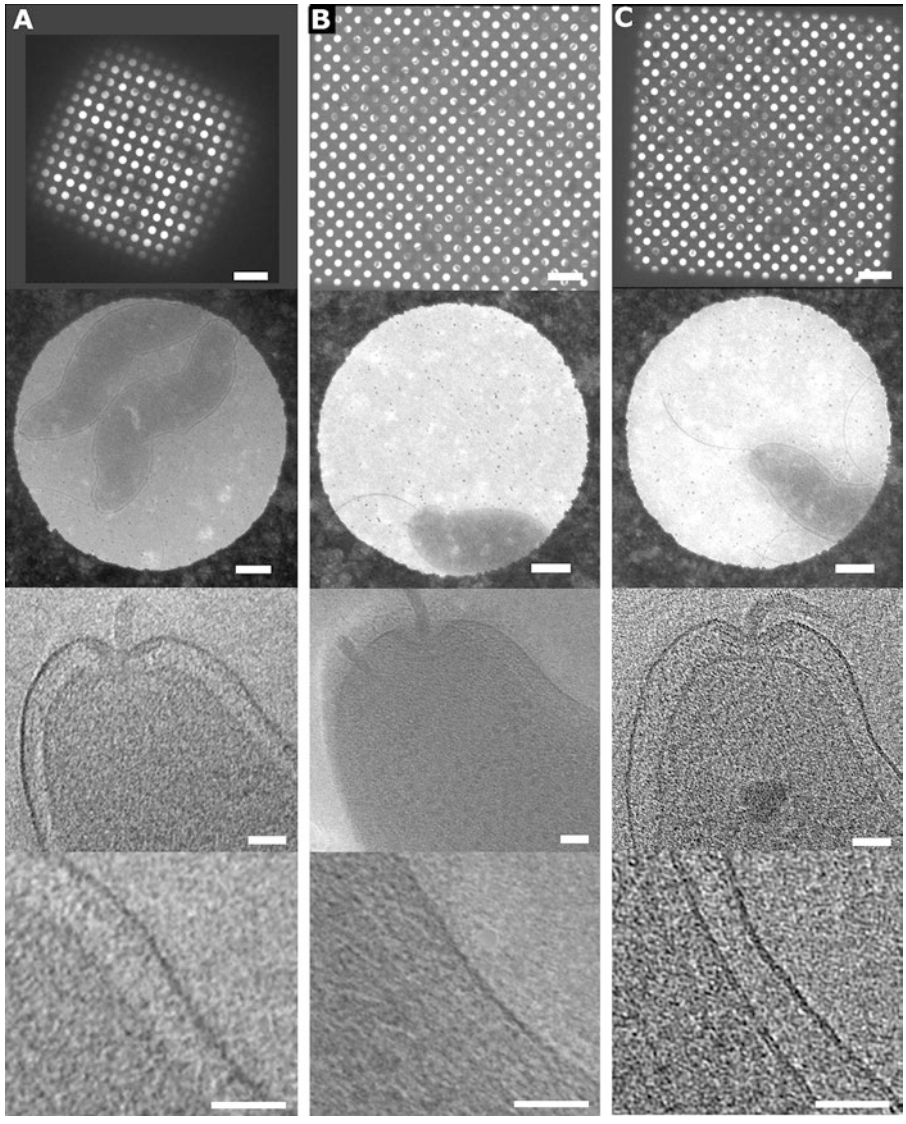


Fig. 1 Determining optimal vitreous ice thickness. Representative electron cryo-microscopy images of ***Campylobacter jejuni*** illustrating the effect of: (**a**) overly thick ice, (**b**) overly thin ice from overblotting, and (**c**) optimal ice on data collection at different magnifications. Rows from top to bottom of panels **a**, **b**, **c**: (1) 120× magnification "square image," 10 μm scale. (2) 3500× magnification "hole image," 300 nm scale. (3) 25,000× magnification slice through tomographic reconstruction, 50 nm scale. (4) 2× zoomed subset of panel 3 showing membrane leaflet resolution, 50 nm scale. Voxel depth in Å of insert 3: (**a**) 16.56, (**b**) 8.28, (**c**) 8.122

found within the tomograms. This is best done by consistent placement of two contour points in a manner that bisects the structure vertically through the central slice of the structure (*see* **Note 28**).

3. Save each model as an IMOD model file and store in an appropriate location.

3.4 Subtomogram Averaging Using PEET

1. At this point, subtomogram averaging can be performed in any suitable software. We will outline the subtomogram averaging methodology in the PEET package within IMOD; however, the package Relion may be more suitable for high-resolution reconstructions (*see* **Note 29**).
2. Load the eTomo PEET module. Within the setup tab, input each tomogram, model tilt range file until all data has been entered (*see* **Note 30**). If your dataset has rotational symmetry, randomly rotate each particle around the Y-axis so as avoid aligning the missing wedge of data.
3. Under the Volume size (Voxels) section, enter a size that fits the size of the structure being averaged, with some margins for a mask and particle rotation.
4. Under the Masking section, select an appropriate mask to encapsulate your protein complex with a surrounding envelope. Masks can be spherical, cylindrical, or custom MRC shapes dependent on what is appropriate for the structure being averaged.
5. Navigate to the *Run* tab and in the iteration table, input an appropriate iteration scheme (*see* **Note 31**).
6. Once settings have been updated, indicate the cluster or number of CPUs to use in the Parallel Processing tab and Run the program. Wait for the average to compute. Results can be used to inform iterative refinement of the volume size, masking, and iteration scheme.

4 Notes

1. Ideally data should be collected on a 200 or 300 kV microscope equipped with a LaB_6 or FEG electron source and a direct electron detector camera. For subnanometer reconstructions, a FEG electron source and direct electron detector become essential; for thicker samples 300 kV acceleration voltage becomes essential.
2. We use Leginon due to its robustness and versatility that enable consistent high-throughput tomographic data collection [22]. When optimized as outlined during the methods section, we are able to robustly collect multiple tilt series per hour, overnight and over multiple days, with little user input other than in maintaining the liquid nitrogen levels. The tilt series acquisition algorithm is taken from UCSF Tomo [19] which may also be used.
3. It is strongly advised to have a system for data management. We use an in-house MySQL database to keep track of datasets and particles within those datasets. Other systems are also available [23].

4. We use the following blot settings on an FEI Vitrobot Mk. 4, although settings will need to be optimized on different vitrification devices: blot time = 5.0 s, wait time = 60 s, drain time = 1.0 s, blot force = 3, blot total = 1. We use 90–100% humidity with humidity switched off during blotting.
5. It is important to ensure the humidifier is regularly emptied and dried to prevent fungal growth.
6. Different grid types provide advantages when performing ECT. Larger holes will enable larger cells to fit in a hole but may be less stable. All-gold grids minimize specimen movement and also have the advantage of providing a reliably even layer of ice during vitrification while also being more stable under the electron beam [24]. However, the gold layer surrounding holes is not transparent to the electron beam and, as such, makes it more difficult to assess ice quality due to higher contrast levels compared to the cells. This also prevents CTF correction due to preventing taking images above and below the target, which are used for CTF estimation. Carbon/copper grids provide a substrate surrounding the holes, which is transparent to the electron beam and enables visualization of Thon rings. This enables both easier assessment of ice quality and CTF correction via the aforementioned imaging scheme.
7. At this point, if using all-gold grids, pierce a hole in the center of each grid using the tip of the tweezers to ensure there is a hole at all magnifications to measure unobstructed beam flux. This will enable tuning of the microscope beam during imaging. In practice carbon grids, being more fragile will almost always already have at least one broken square.
8. Cell density is a key factor for collecting tomographic data. If the density is too high, the cells may clump together and obstruct structures during tilts. If the density is too low, the cells may be too sparse to pick sufficient numbers of targets. Furthermore, the ice envelope surrounding the cells may be altered by higher densities of surrounding cells, leading to heterogeneous ice thickness dependent on grid locale.
9. When vitrifying cells, structures of interest are best preserved by only removing cells from incubation immediately prior to vitrification, which can prevent other metabolic processes, removing or degrading them. This is also aided by maintaining cells on ice throughout the duration.
10. Using wide aperture tips when pipetting cells can preserve surface structures and appendages that might be otherwise sheared off. Also, when resuspending pelleted cells, gentle resuspension with pipettes can aid in preserving structures.
11. Filtering liquid nitrogen through paper towels or cloth can help remove water crystals from atmospheric moisture that may have contaminated the nitrogen.

12. The glow discharge length and high tension can be optimized depending on the cell type being imaged; alternate schemes can lead to different cell positions within the grid as well as levels of cell aggregation. Glow discharging grids should be performed shortly before vitrification to prevent recharging. This can often be done concurrently to cooling down the freezing dewar.
13. Cryogen can be ethane (the standard used by many labs) or ethane-propane mixture (63% propane, 37% ethane [25]). Ethane-propane has a number of advantages, most importantly that it does not freeze at liquid nitrogen temperatures as pure ethane does. Because ethane-propane is retained at a set temperature, thermal expansion is not an issue. In the case of ethane-propane, therefore, fill the central dewar until the cryogen forms a bulging meniscus from the top of the central chamber. With pure ethane, fill to a few millimeters below the top of the central chamber to allow for expansion during freezing. With any of these cryogens, appropriate care must be taken as they are flammable and explosive. Note that a custom all-metal cryocup is needed for ethane-propane to ensure full cooling of the mix.
14. When combining the sample and the fiducials, mix by pipetting to ensure an even distribution of cells and fiducials on the grid. If delicate surface structures are the target, wide aperture tips can be used to prevent shearing.
15. A unified naming system is recommended to keep track of the contents of grid boxes. In our lab we store ten grid boxes in a 50 mL tube and multiple tubes in a slot in a dewar and have multiple dewars. For recording, we name each grid box according to its location and write this name on the top, side, and bottom of the grid box. As an example, grid box name 1.3A.7 refers to dewar 1, slot 3, 50 mL tube "A," and grid box 7 (of 10 in tube 1.3A). The slot in the grid box can then be specified with another digit, e.g., grid 3 in that box could be unambiguously designated at 1.3A.7.3. Commercial systems are also available.
16. Tap the tweezer edge against a surface to ensure the grid is properly grasped by the tweezers. Care should be taken not to bend grids when manipulating them prior to vitrification which can lead to broken support film and uneven blotting resulting in steep ice thickness gradients.
17. Contemporary vitrification approaches result in considerable ice quality variability between grids. We usually freeze three grids per sample to account for variation. In the future, alternative, more reproducible methods are anticipated [26].
18. If Leginon does not initially detect microscope control clients, a network connection may not be established between the

main and support computers. This can be fixed by checking the network connection and closing and reopening clients.

19. The defocus value set in the "tomo" node will influence the contrast and resolution of the final data collected. The closer to focus the setting, the higher the resolution of the data that will be contained within the tomograms; however, this will also cause lower contrast. Optimization of this defocus setting can be done within the data collection session by taking test tomograms, assessing the data quality and adjusting the focus up and down accordingly. Examples of settings used to image flagella motors in *C. jejuni* are as follows:
    (a) Grid (*gr*) – magnification (56), defocus (0), spot size (8), intensity (1), exposure time (500 ms), dimension (1024×1024), binning (2×2)
    (b) Square (*sq*) – magnification (120), defocus (0), spot size (8), intensity (1), exposure time (1000 ms), dimension (2048×2048), binning (2×2)
    (c) Hole (*hl*) – magnification (3500), defocus (−100 μm), spot size (6), intensity (0.67), exposure time (1000 ms), dimension (2048×2048), binning (2×2)
    (d) Preview (*preview*) – magnification (25,000), defocus (−10 μm), spot size (4), intensity (0.5), exposure time (286 ms), dimension (2048×2048), binning (2×2)
    (e) Focus (*fc*) – magnification (25,000), defocus (0), spot size (4), intensity (0.5), exposure time (286 ms), dimension (2048×2048), binning (2×2)
    (f) Tomography (*tomo*) – magnification (25,000), defocus (−3.5 μm), spot size (4), intensity (0.5), exposure time (286 ms), dimension (2048×2048), binning (2×2)
20. Assessing ice quality for data collection can be difficult depending on the grid type. The optimal thickness for cells can be seen when the cytoplasm of the cell is darker than the surrounding ice in the hole. Overly thick ice will provide little contrast between the two, and overly thin ice will show holes within the ice where it is receding from the center. Overly thin ice can lead to dehydrated or flattened cells evident in the indistinctness of membranes in projection and tomographic reconstruction. Gold/gold grids are typically harder to assess for ice quality due to the heavy contrast of gold altering overall contrast; therefore, they can only be truly assessed at higher magnifications (e.g., at magnification of data collection). Present specific grid assessment can be done as follows:
    (a) *gr* – Grid itself is not completely black/an area to pick targets can be seen.

(b) *sq* – Edges of squares are not too thickly edged by ice/sharp due to thin ice.

(c) ***hl*** – Cells can be seen within the holes and cytoplasm appears to be a darker gray than the surrounding ice.

(d) *tomo* – Cell membranes are distinctly seen and structures of interest are present and not blurry.

21. We use the Leginon automated eucentric procedure in the Tomo Z Focus node to perform this step automatically.
22. Crystals of ice contamination are usually considered detrimental on a grid but can be very useful for image shift alignments.
23. It is often prudent to collect a single *sq*, *hl*, and *tomo* target and reconstruct the tomogram to ensure settings are as desired before queuing a larger dataset for collection.
24. The user does not need to step through every *hole targeting* image before submitting the hole target queue. Rather, the *hole targeting* images can be returned to later if desired.
25. Optimal targets for data collection are those where the portion of the cell containing the structure of interest is projecting directly into the center of the hole. Surrounding cells/debris can obscure structures during tilts, leading to lower quality data. Furthermore, when using gold/gold grids, targets closer than 100 nm to the edge of holes can be obscured by artifacts induced by the high-contrast gold, and they should not be collected.
26. Depending on the settings chosen within Leginon, it is possible to collect multiple tilt series an hour on a side entry F20 microscope. Therefore, approximate calculations of length of data collection time and the necessary number of targets should be calculated prior to picking square targets.
27. Tomogram production can be automated via a number of available implementations [23, 27–29], or automation can be developed using shell or Python scripting.
28. When picking targets, saving model files from heavily binned data can aid in the process. Each tomogram can take up to 2 min to load; hence, further binning can be applied to speed up this step, increasing throughput of target collection in large datasets. You will need to then scale the models upward to fit the unbinned datasets afterward.
29. Relion provides the advantage of a maximum likelihood alignment approach, and integrated contrast transfer function (CTF) correction, improving the tomographic data quality [30].
30. Handling subtomogram data manually can be onerous. Data entry from databases or spreadsheets can easily be scripted, and data reconstruction pipelines have been published [29] which

enable the efficient reconstruction of tomographic data through the generation of files which automatically associate the requisite data files together for use in subtomogram averaging programs. Automation is essential for analyzing subtomogram data collected from high-throughput tomography.

31. We typically use an iteration scheme such as this:
    (a) First iteration: to generate an initial model, perform an initial alignment or manually picked particles with no rotational or translational search, and average all volumes. This will generate a first rough alignment with coarse features sufficient for iterative refinement.
    (b) One to three broad-search iterations: because particles are already coarsely manually aligned, a limited search is sufficient. We use a rotational search of ±18° with 6° steps, 5 voxel translational search, and a fairly aggressive low-pass filter of 0.15 with sigma 0.05, although this will need modification according to dataset. At the end of this iteration, we average the top 30% of aligned particles.
    (c) One to three narrowing searches: we use a rotational search of ±9° with 3° steps, 3 voxel translational search, and a low-pass filter of 0.2 with sigma 0.05. At the end of this iteration, we average the top 30% of aligned particles.
    (d) Ten to fifteen refinement searches: we use a rotational search of ±3° with 1° steps, 1 voxel translational search, and a low-pass filter of 0.25 with sigma 0.05. At the end of this iteration, we average the top 100% of aligned particles.
    (e) We continue iterating until refinement quality plateaus.

Acknowledgment

LH was supported by a Biotechnology and Biological Sciences Research Council postgraduate training award and Biotechnology and Biological Sciences Research Council Grant BB/L023091/1 to MB.

References

1. Asano S, Engel BD, Baumeister W (2016) In situ cryo-electron tomography: a post-reductionist approach to structural biology. J Mol Biol 428:332–343. https://doi.org/10.1016/j.jmb.2015.09.030
2. Lučić V, Rigort A, Baumeister W (2013) Cryo-electron tomography: the challenge of doing structural biology in situ. J Cell Biol 202:407–419. https://doi.org/10.1083/jcb.201304193
3. Briggs JA (2013) Structural biology in situ—the potential of subtomogram averaging. Curr Opin Struct Biol 23:261–267. https://doi.org/10.1016/j.sbi.2013.02.003
4. Beck M, Baumeister W (2016) Cryo-electron tomography: can it reveal the molecular sociology of cells in atomic detail? Trends Cell Biol 26:825–837. https://doi.org/10.1016/j.tcb.2016.08.006

5. Chen S, Beeby M, Murphy GE et al (2011) Structural diversity of bacterial flagellar motors. EMBO J 30:2972–2981. https://doi.org/10.1038/emboj.2011.186
6. Beeby M, Ribardo DA, Brennan CA et al (2016) Diverse high-torque bacterial flagellar motors assemble wider stator rings using a conserved protein scaffold. Proc Natl Acad Sci 113:E1917–E1926. https://doi.org/10.1073/pnas.1518952113
7. Chang Y-W, Rettberg LA, Treuner-Lange A et al (2016) Architecture of the type IVa pilus machine. Science 351:aad2001. https://doi.org/10.1126/science.aad2001
8. Zhao X, Zhang K, Boquoi T et al (2013) Cryoelectron tomography reveals the sequential assembly of bacterial flagella in Borrelia burgdorferi. Proc Natl Acad Sci 110(35):14390–14395. https://doi.org/10.1073/pnas.1308306110
9. Chang J, Liu X, Rochat RH et al (2012) Reconstructing virus structures from nanometer to near-atomic resolutions with cryo-electron microscopy and tomography. Adv Exp Med Biol 726:49–90. https://doi.org/10.1007/978-1-4614-0980-9_4
10. Schur FKM, Obr M, Hagen WJH et al (2016) An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation. Science 353:506–508. https://doi.org/10.1126/science.aaf9620
11. Hagen WJH, Wan W, Briggs JAG (2017) Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. J Struct Biol 197:191–198. https://doi.org/10.1016/j.jsb.2016.06.007
12. Kunz M, Frangakis AS (2017) Three-dimensional CTF correction improves the resolution of electron tomograms. J Struct Biol 197:114–122. https://doi.org/10.1016/j.jsb.2016.06.016
13. Williams DB, Carter CB (2009) Transmission electron microscopy: a textbook for materials science. Springer, Berlin
14. Woldringh CL (1976) Morphological analysis of nuclear separation and cell division during the life cycle of Escherichia coli. J Bacteriol 125:248–257
15. Farley MM, Hu B, Margolin W, Liu J (2016) Minicells, back in fashion. J Bacteriol 198:1186–1195. https://doi.org/10.1128/JB.00901-15
16. Schorb M, Gaechter L, Avinoam O et al (2017) New hardware and workflows for semi-automated correlative cryo-fluorescence and cryo-electron microscopy/tomography. J Struct Biol 197:83–93. https://doi.org/10.1016/j.jsb.2016.06.020
17. Briegel A, Ladinsky MS, Oikonomou C et al (2014) Structure of bacterial cytoplasmic chemoreceptor arrays and implications for chemotactic signaling. eLife 3:e02151. https://doi.org/10.7554/eLife.02151
18. Mastronarde DN (2005) Automated electron microscope tomography using robust prediction of specimen movements. J Struct Biol 152:36–51. https://doi.org/10.1016/j.jsb.2005.07.007
19. Zheng QS, Braunfeld MB, Sedat JW, Agard DA (2004) An improved strategy for automated electron microscopic tomography. J Struct Biol 147:91–101. https://doi.org/10.1016/j.jsb.2004.02.005
20. Kremer J, Mastronarde D, McIntosh J (1996) Computer visualization of three-dimensional image data using IMOD. J Struct Biol 116:71–76
21. Nicastro D, Schwartz C, Pierson J et al (2006) The molecular architecture of axonemes revealed by cryoelectron tomography. Science 313:944–948. https://doi.org/10.1126/science.1128618
22. Suloway C, Shi J, Cheng A et al (2009) Fully automated, sequential tilt-series acquisition with Leginon. J Struct Biol 167:11–18. https://doi.org/10.1016/j.jsb.2009.03.019
23. Ding HJ, Oikonomou CM, Jensen GJ (2015) The Caltech tomography database and automatic processing pipeline. J Struct Biol 192:279–286. https://doi.org/10.1016/j.jsb.2015.06.016
24. Russo CJ, Passmore LA (2014) Ultrastable gold substrates for electron cryomicroscopy. Science 346:1377–1380. https://doi.org/10.1126/science.1259530
25. Tivol WF, Briegel A, Jensen GJ (2008) An improved cryogen for plunge freezing. Microsc Microanal 14:375–379. https://doi.org/10.1017/S1431927608080781
26. Jain T, Sheehan P, Crum J et al (2012) Spotiton: a prototype for an integrated inkjet dispense and vitrification system for cryo-TEM. J Struct Biol 179:68–75. https://doi.org/10.1016/j.jsb.2012.04.020
27. Cao M, Takaoka A, Zhang H-B, Nishi R (2011) An automatic method of detecting and tracking fiducial markers for alignment in electron tomography. J Electron Microsc 60:39–46. https://doi.org/10.1093/jmicro/dfq076
28. Mastronarde DN, Held SR (2017) Automated tilt series alignment and tomographic reconstruction in IMOD. J Struct Biol 197:102–113. https://doi.org/10.1016/j.jsb.2016.07.011

29. Morado DR, Hu B, Liu J (2016) Using tomoauto – a protocol for high-throughput automated cryo-electron tomography. J Vis Exp:e53608. https://doi.org/10.3791/53608

30. Bharat TAM, Scheres SHW (2016) Resolving macromolecular structures from electron cryotomography data using subtomogram averaging in RELION. Nat Protoc 11:2054–2065. https://doi.org/10.1038/nprot.2016.124

Chapter 3

Preparation of Tunable Microchips to Visualize Native Protein Complexes for Single-Particle Electron Microscopy

Brian L. Gilmore, A. Cameron Varano, William Dearnaley, Yanping Liang, Bridget C. Marcinkowski, Madeline J. Dukes, and Deborah F. Kelly

Abstract

Recent advances in technology have enabled single-particle electron microscopy (EM) to rapidly progress as a preferred tool to study protein assemblies. Newly developed materials and methods present viable alternatives to traditional EM specimen preparation. Improved lipid monolayer purification reagents offer considerable flexibility, while ultrathin silicon nitride films provide superior imaging properties to the structural study of protein complexes. Here, we describe the steps for combining monolayer purification with silicon nitride microchips to create a tunable approach for the EM community.

Key words Electron microscopy, Single-particle analysis, Affinity capture, Silicon nitride, Microchips, Protein assemblies

1 Introduction

Single-particle electron microscopy (EM) is a valuable tool for investigating the structural properties of biological complexes [1]. With this technique, structural information embedded in the images is extracted to computationally build a 3D density map of the examined assemblies. One recurring challenge for single-particle methods is obtaining a homogeneous sample that facilitates downstream imaging and computational analysis. Conventional biochemical purification is often employed to help isolate protein assemblies but can sometimes fall short of the desired sample concentration and purity. A second bottleneck in preparing single-particle EM specimens is the inherent limitations introduced by traditional materials and processes. These combined shortcomings have created an opportunity for improvement in generating better EM specimens.

Lipid monolayers present a different approach for single-particle specimen preparation [2]. The amphipathic nature of the

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_3,

lipid is well suited for both the adherence to a solid EM support and the capture of protein complexes. The versatility of lipid monolayers comprised of functionalized Ni-NTA moieties enables His-tagged proteins to bind to the lipid with increased specificity [3]. This "monolayer purification" platform can be used to recruit His-tagged proteins from cell lysates, nuclear fractions, or pre-fractionated samples directly onto the support film in a single step [4–7]. The method further evolved to include non-His-tagged protein complexes when His-tagged Protein A adaptors were introduced, bridging the Ni-NTA lipid with a target antibody [8]. This modification makes the method tunable and robust, lending itself to study proteins and assemblies for which antibodies are readily available [9–11].

Carbon-coated support films are traditionally used in the preparation of biological specimens for EM imaging purposes. Amorphous carbon supports can have a great deal of variability and defects from production that can ultimately impact resolution. With the advent of in situ EM, silicon nitride (SiN) microchips provide an alternative support material that is transparent to the electron beam [12–16]. As SiN membranes can be more consistently manufactured, their very flat surface renders them an attractive tool. The hydrophobic nature of SiN membranes also provides a compatible surface for use with lipid monolayers. Pairing SiN microchips with the monolayer purification approach has shown to be a valuable tool, including recent reports describing BRCA1-transcriptional complexes [17–19].

In this chapter, we present in detail how to make a tunable microchip specimen to visualize protein complexes derived from breast cancer cells grown under stressful conditions. The protocol describes (1) the preparation of the silicon nitride microchip, (2) the proper setup and transfer of a monolayer to the microchip, (3) the procedure for creating the "tuned" EM specimen, and (4) recommendations for data collection and image processing. Image information and a representative 3D structure of BRCA1-transcriptional complexes are shown.

2 Materials

1. Glass volumetric vial with stopper (1 mL).
2. Chloroform.
3. Parafilm.
4. Buffer A: 20 mM HEPES (pH 7.2), 140 mM NaCl, 2 mM $MgCl_2$, and 2 mM $CaCl_2$.
5. Ultrapure water.
6. Hot plate.

7. 5 N sodium hydroxide solution.
8. 5 mL syringe with Luer-Lok tip.
9. 33 mm syringe filter (0.2 μm).
10. 15 mL conical tube.
11. Aluminum foil.
12. Silicon nitride microchips (Protochips).
13. Carbon fiber tweezers (Pelco).
14. HPLC-grade acetone.
15. HPLC-grade methanol.
16. Dish or beaker for solvent.
17. Whatman 1 circular filter paper (90 mm).
18. Compressed air.
19. Glass petri dish with cover (100 × 15 mm).
20. Disposable glass cell culture tubes (6 × 50 mm).
21. Hamilton 10 μL syringe.
22. Pipet.
23. BRCA1 antibody (C-20, 0.2 mg/mL, Santa Cruz Biotechnology).
24. Glass Pasteur pipet.
25. Clear PVC vacuum tubing.
26. House vacuum system.
27. Ni-NTA lipid and DLPC lipid (Avanti Polar Lipids). Prepare by weighing 1 mg of powdered lipid into a 1 mL glass volumetric vial, and add chloroform to the 1 mL mark on the vial. Cap the vial and seal with parafilm. Store in a −20 °C freezer for up to 3 months. Lipids used in the preparation of monolayers are dissolved in chloroform. Chloroform is a carcinogen. Please review the MSDS and consult with EHS for its safe use and disposal.
28. His-tagged Protein A (10 mg, 50 mg/mL, Abcam) has been optimized for IgG binding and includes a 6× His tag at its N-terminus. It was pre-diluted to a working concentration of 0.1 mg/mL (1/500) in Buffer A, aliquotted, and stored at −20 °C. Buffer A can be substituted for the user's preferred buffer.
29. Uranyl formate solution (0.75%). Boil 3 mL of ultrapure water in a 10 mL beaker on a hot plate. Using tongs, transfer the beaker to a stir plate in a chemical fume hood. Add 22.5 mg of uranyl formate and a small stirbar. Stir for 5 min. Add 4.2 μL of 5 N NaOH. Stir another 5 min. Draw the solution up into a 5 mL syringe. Filter the mixture through a 0.2 μm PVDF

filter to remove the undissolved uranyl formate and collect in a 15 mL conical tube. Cover the tube with aluminum foil. Uranyl formate solution is light sensitive. Covering the tube in aluminum foil minimizes the exposure to light. If stored in this manner, the uranyl formate solution can be kept up to 48 h without a decline in quality. Uranyl formate is toxic if inhaled or ingested. Please review the MSDS and consult with EHS for safe use and proper disposal.

30. Negatively stained BRCA1-RNAP II complexes were examined using a FEI Spirit Bio-Twin TEM equipped with a LaB_6 filament and operating at 120 kV.
31. Images are recorded using a FEI Eagle 2k HS CCD camera having a 30 μm pixel size and employing low-dose conditions (~1–5 electrons/Å^2).
32. SPIDER is a software package [20] used for 2D classification of procedures through a multivariate data analysis approach. Individual protein complexes (particles) are selected from micrographs. After selection, all particles are extracted then processed through several iterations of multi-reference alignment implementing a K-means classification routine. The parameters in each step are modulated by the user; however, the routines are standard.
33. RELION is a software package [21] used to perform reconstruction and refinement calculations using an empirical Bayesian methodology.

3 Methods

3.1 Cleaning the Microchips

Silicon nitride membranes can be as thin as 5 nm and should be handled with care to avoid fracture (*see* **Note 1**). Microchips are supplied in Gel-Paks with the membrane side facing upward (Fig. 1a). The manufacturer recommends handling the microchips at the edges with carbon fiber tip tweezers to avoid damaging the silicon frame (*see* **Note 2**). Before use, the microchips should be cleaned in a low-dust environment with HPLC-grade acetone and methanol to avoid surface contamination (*see* **Note 3**).

1. In a dish or beaker containing acetone, submerge and release the microchip (Fig. 1b). Wash by gently swirling the dish in a circular motion for 1–2 min.
2. Promptly move the microchip to a second dish containing methanol. Do not allow the microchip to dry during transfer. Swirl to wash for 1–2 min.
3. Remove the microchip from the methanol and gently wick off the excess fluid by touching the edge to Whatman 1 filter paper.

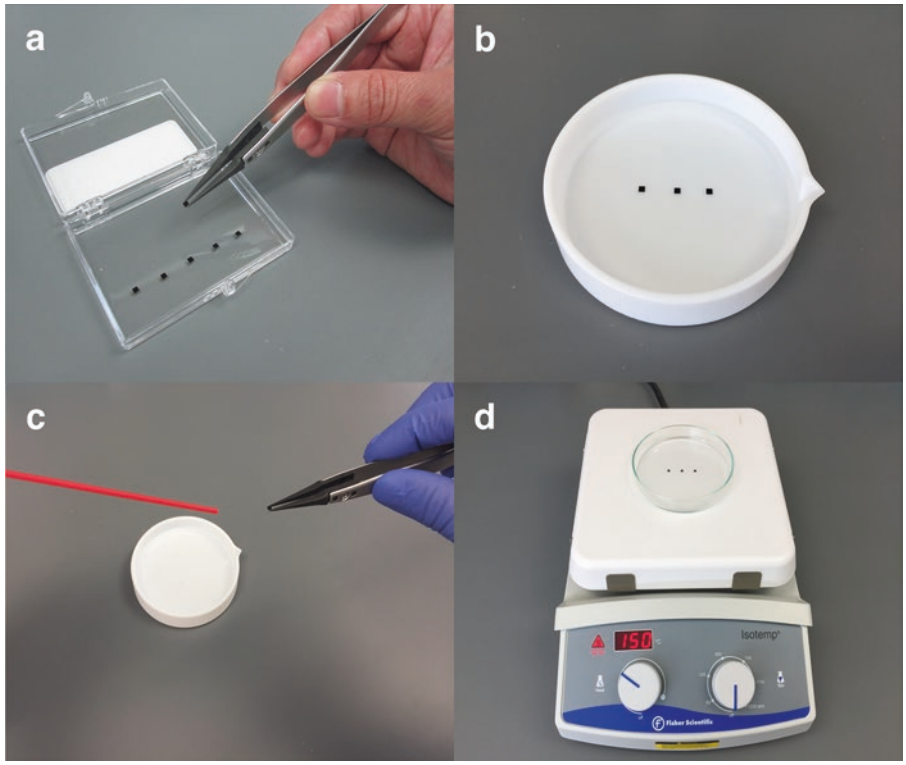


Fig. 1 The preparation of SiN microchips for single-particle EM. (**a**) The microchips are supplied in a protective Gel-Pak. Carbon fiber-tipped forceps are ideal for handling microchips. (**b**) In a shallow dish or beaker, solvents remove the protective coating and clean the microchip surface. (**c**) Compressed air (red straw) helps to dry the surface and keep it residue-free after cleaning. (**d**) Heating to 150 °C for 60 min prior to use removes the remaining moisture from the SiN membrane and enhances its hydrophobicity

4. To prevent residue or contamination from dust particles, the microchips may be dried using residue-free compressed air (Fig. 1c). With the microchip still in the tweezers, direct a gentle flow of air across the surface until it is dry (*see* **Note 4**).

3.2 Preparation of Microchip Surface

The hydrophobicity of the silicon nitride membrane can be enhanced to facilitate proper binding of the nonpolar lipid tail domains. A simple way to do this is by heating the microchip on a hot plate (Fig. 1d; *see* **Note 5**).

1. With the membrane side up, place the microchip on a clean glass petri dish or glass slide.
2. Preheat the hot plate to 150 °C. Place the dish containing the microchip onto the hot plate for 1.5 h.
3. With forceps or hot gloves, carefully remove the dish from the hot plate and place on a heat-resistant surface. Allow the microchip to cool to room temperature.

3.3 Preparation of Ni-NTA Lipid Monolayers

Individual lipid components are solubilized in chloroform, and monolayers consist of a combination of Ni-NTA (active binding) lipids and DLPC (inactive, filler) lipids. The concentration of active binding sites present in the monolayer can be adjusted by modifying the ratio of the Ni-NTA to DLPC lipids (*see* **Note 6**). The resulting lipid mixture is added over a drop of water, forming a thin monolayer film, which can then be transferred to the hydrophobic surface of a microchip (Fig. 2a; *see* **Note 7**).

1. In advance, remove solubilized lipids from storage and allow them to warm to room temperature (*see* **Note 8**).
2. Place a piece of circular Whatman 1 filter paper in a glass petri dish. Wet the filter with ultrapure water, and then place a 2-in. by 2-in. piece of parafilm on top of the wetted filter. Place the cover back on the dish.
3. Obtain three disposable glass tubes to be used for the preparation of the lipid mixture. Label one with the percentage of lipid to be used (e.g., "5% Ni-NTA") and one "rinse." The

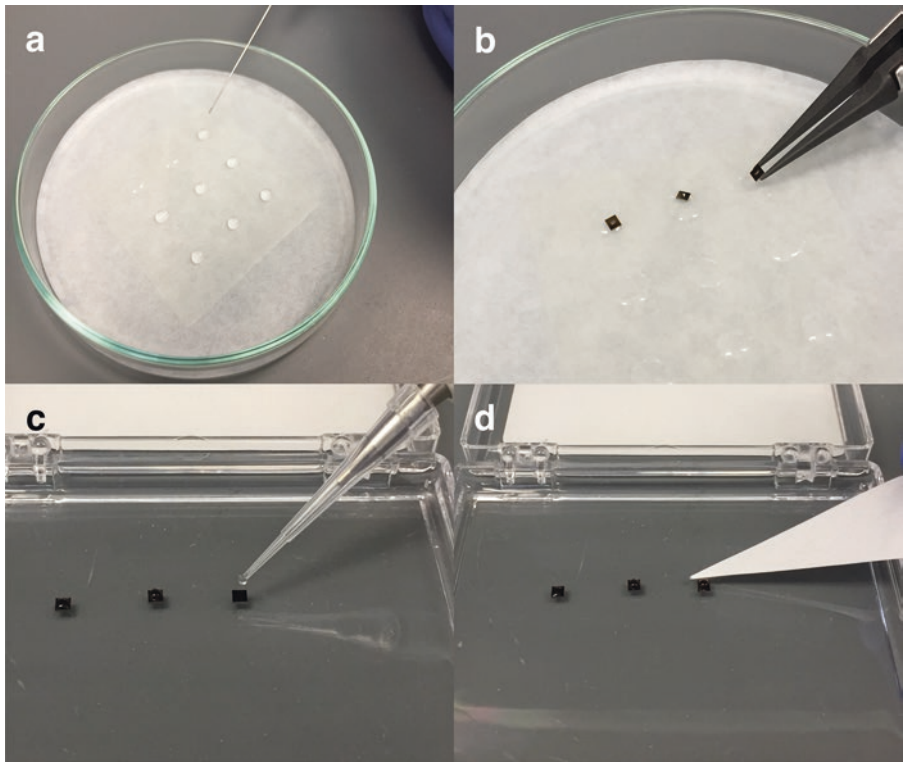


Fig. 2 Decorating tunable microchips with lipid monolayers. (**a**) The lipid mixture is applied over a water drop. Note the flattening of the drop (top left) after addition of lipid. (**b**) With the microchip inverted and the SiN membrane facing the drop, the microchip is carefully lowered onto the monolayer enabling transfer of the lipid layer to the microchip surface. (**c**) Returning the microchip to the Gel-Pak stabilizes it for adding solutions during specimen preparation. (**d**) Whatman 1 filter paper is used to gently wick away excess solutions

remaining tube will contain chloroform to be used for dilution and can be labeled "$CHCl_3$" (*see* **Note 9**). Using a Pasteur pipet, add ~0.5 mL chloroform to the "$CHCl_3$" and "rinse" tubes.

4. Rinse a Hamilton syringe by aspirating and dispensing several times with chloroform in the "rinse" tube.
5. Aspirate 10 μL of chloroform from the "$CHCl_3$" source tube and dispense into the tube labeled "5% Ni-NTA."
6. From the DLPC filler lipid source vial, aspirate and dispense 28 μL lipid into the "5% Ni-NTA" tube. Rinse the syringe.
7. From the Ni-NTA lipid source vial, aspirate 2 μL of lipid and dispense into the "5% Ni-NTA" tube. Rinse the syringe. Seal all tubes and vials with a small piece of parafilm until use to prevent evaporation (*see* **Note 10**).
8. Pipet 15 μL of ultrapure water onto the parafilm in the humid petri dish. Do this for as many microchips as is necessary for the experimentation. Separate the water drops by ~1 cm.
9. Using the Hamilton syringe, add 1 μL of the 5% Ni-NTA lipid mixture over the apex of each water drop. After addition of the lipid, the drop will flatten and spread (Fig. 2a).
10. Put the lid on the petri dish. To keep the lipid monolayers hydrated, seal them in a humid environment by wrapping parafilm around the petri dish.
11. Incubate the dish at room temperature for 10 min before placing on ice for at least 1 h (*see* **Note 11**).

3.4 Preparation of Tunable Microchip Specimen

Here we describe the transfer of the monolayer to the microchip followed by step-by-step addition of His-tagged Protein A adaptor, target antibody, and sample. Though the procedure for negative staining is described, tunable microchips also work well in cryo-EM applications [15, 17–19]. Uranyl formate stain should be prepared in advance. All filter paper used in the preparation of EM specimens is Whatman 1. Each solution will be removed from the microchip by wicking with the edge of a small piece of filter paper unless otherwise indicated. All incubations are performed at room temperature (~23 °C; *see* **Note 12**).

1. Carefully remove the parafilm sealing the glass dish containing the monolayers, and return the dish to the ice.
2. Using carbon fiber tip tweezers, place each microchip membrane side down on the surface of a lipid monolayer (Fig. 2b). The microchip typically will come to rest on the side of the water droplet. The nonpolar tail adheres to the hydrophobic surface of the microchip. Incubate for 1 min (*see* **Note 13**).

3. Gently lift the microchip off the monolayer, and place it membrane side up on a free space in a Gel-Pak (Fig. 2c). Using a pipet, add 3 μL of His-tagged Protein A and incubate for 1 min (*see* **Note 14**).
4. Carefully remove the Protein A solution by wicking it off the microchip with the edge of filter paper (Fig. 2d). Promptly add 3 μL of antibody solution (*see* **Note 15**). Incubate for 1 min.
5. Remove the antibody solution using a Hamilton syringe. Immediately add the protein sample and incubate on the microchip for 2 min at room temperature (*see* **Note 16**).
6. Remove the sample solution and immediately rinse with 3 μL ultrapure water (*see* **Note 17**).
7. Remove the water and immediately add 3 μL of 0.75% uranyl formate to wash.
8. Remove the uranyl formate wash, and immediately add another 3 μL of uranyl formate to stain the specimen. Incubate for 10–30 s (*see* **Note 18**).
9. Remove the uranyl formate wash and carefully aspirate excess stain with a Pasteur pipet connected by PVC tubing to a gentle house vacuum (*see* **Note 19**). Store the microchip on filter paper in a clean, covered glass dish or Gel-Pak to protect from debris until imaging.

3.5 TEM Image Collection

1. Allow the scope to fully evacuate the column and cool.
2. Load the microchip sample (Fig. 3a) containing tunable components (Fig. 3b) into a single-tilt EM specimen holder at room temperature. The microchip is loaded in a similar manner to copper supports. If possible, align the imaging window to the center of the specimen holder.
3. With the beam engaged, find the window on the microchip. The sample must be aligned along the axis of the beam line, in order for proper defocus to be attained. This is done at the center point of the specimen holder. The sample is now ready for imaging.
4. Once an area of good particle occupancy has been identified, digital images are acquired at a defocus value of ~−1.5 μm (Fig. 3c; *see* **Note 20**).
5. Save images in 16-bit.tiff format for downstream image processing procedures.

3.6 Data Analysis and Representative Results

1. Prior to particle selection, the original images are normalized using the standard routines in SPIDER software package. Individual complexes from the images are manually selected

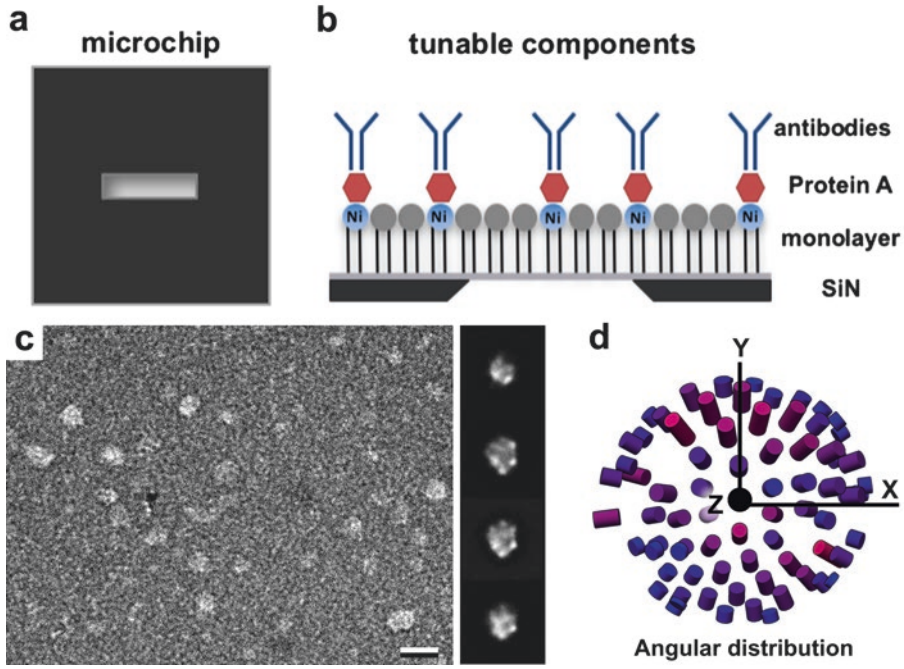


Fig. 3 EM imaging results for BRCA1-RNAP II assemblies. (**a**) The top view of a SiN microchip shows the centrally located imaging window. (**b**) Schematic to show the SiN membrane (light gray) coated with a lipid monolayer containing Ni-NTA moieties (Ni, light blue). His-tagged Protein A adaptors (red) bind to the Ni-NTA head groups and provide a docking site for IgG antibodies (dark blue) to recruit protein complexes. (**c**) A representative micrograph (left) of the BRCA1-RNAP II complexes captured on the tunable microchip. Class averages (right) were calculated using the SPIDER software package. The scale bar is 25 nm. (**d**) An angular distribution plot generated by RELION indicates particle orientations were not limited in the 3D reconstruction

using the WEB interface of the SPIDER software, employing a box size that is approximately twice the diameter of the particles of interest. Multi-reference alignment routines are implemented outputting representative 2D class averages (Fig. 3c, right panel) as previously described [19].

2. A previous map of RNA polymerase II [22] was used as a reference to reconstruct the current complexes in the RELION software package. We implemented 25 refinement iterations using an angular sampling interval of 7.5°. Other parameters input into RELION include a pixel size of 4.4 Å and a regularization parameter of $T = 4$. The angular distribution plot shows a lack of strongly preferred orientations for ~2000 particles (Fig. 3d). A representative 3D reconstruction of the BRCA1-RNAP II complex is shown (Fig. 4a, b).

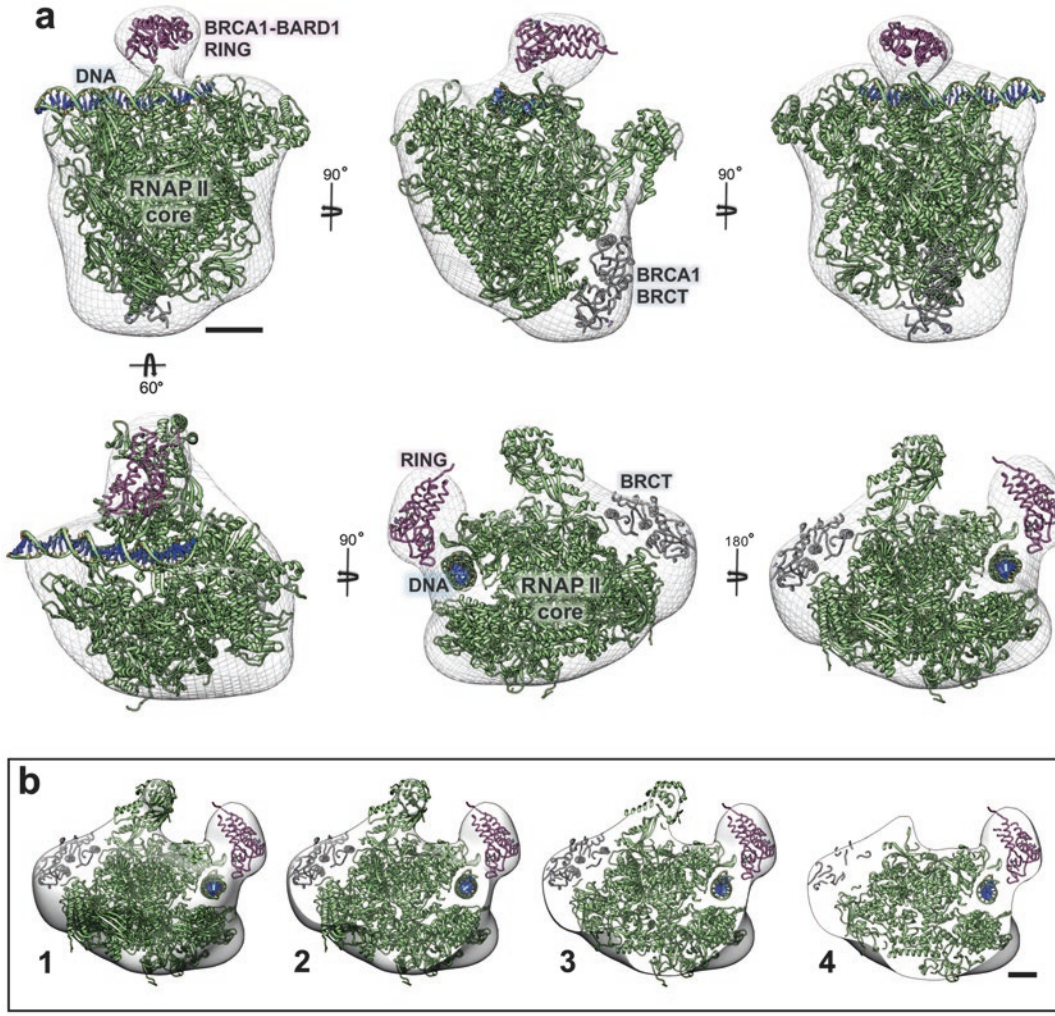


Fig. 4 3D reconstruction of BRCA1-RNAP II assemblies isolated from breast cancer cells. (**a**) The EM density map (white) is shown in different orientations with the RNAP II core and DNA (green and blue, respectively; pdbcode 5IYA, [22]) fit in the map. The BRCA1-BARD1 N-terminal RING domain (pink; pdbcode 1JM7, [23]) and C-terminal BRCT domain (gray; pdbcode 1JNX, [24]) are also shown within the density map based on a previously determined structure [19]. The scale bar is 5 nm. (**b**) Sections (1–4) through the EM density map indicate the overall fit of the atomic models with the map. The scale bar is 2.5 nm

4 Notes

1. Silicon nitride microchips suitable for TEM applications contain a viewing window composed of ultrathin, amorphous silicon nitride supported by a silicon frame. The silicon nitride films are available in a range of thicknesses with films of either 30 or 50 nm typically providing an ideal ratio of both membrane strength and electron transparency. The viewing windows may be either a flat film of constant thickness or contain

integrated features. In this study, the membrane was composed of 200 nm thick silicon nitride, into which were etched an array of microwells (10 × 10 μm square) across the surface. The depth of these wells was 170 nm, resulting in a membrane thickness of only 30 nm at the bottom of each microwell. Microchips are available with viewing windows of different sizes; however, as a general rule of thumb, the thinner the membrane thickness, the smaller the window region due to the increased fragility of the film. Suppliers of silicon nitride microchips for TEM include, among others, Protochips (used in this study), Norcada, and TEM Windows. These suppliers offer a wide range of different features and window options. Users familiar with semiconductor processing may also elect to manufacture their own microchips.

2. The carbon fiber tip tweezers are non-scratching. The microchips used in this study are diced, which provides flat edges and allows the edges to be gripped easily, without contacting the surface of the silicon nitride membrane. With a magnification device, regularly check the integrity of the silicon nitride membrane from cleaning through specimen preparation.
3. The acetone/methanol rinse steps in this procedure are employed to remove a protective photoresist coating that is applied to the surface of the microchip during the fabrication process. Note that some types of silicon nitride microchips may be supplied by the manufacturer without this protective coating. In this case, the cleaning step may be omitted, but it is still recommended as a precaution to remove contaminates or debris that may be present on the membrane surface. Cleaning the microchip should require less than 10 min.
4. Direct the airflow across, rather than perpendicular to, the microchip surface to avoid damaging the SiN membrane. The laminar airflow also prevents dust particles in the air from drying onto the membranes during solvent evaporation.
5. If the ambient conditions are relatively dry, heating the microchip may not be necessary. Otherwise, this step takes ~2.5 h.
6. The ratio of Ni-NTA to DLPC can be increased or decreased as necessary depending on the specimen preparation. DLPC filler lipid acts to spatially disperse the Ni-NTA lipid. Increase the dilution of Ni-NTA if the particle density is too high and vice versa. As a rule of thumb, 5% Ni-NTA is a good starting place for a negatively stained specimen and 25% Ni-NTA for a cryo-EM specimen.
7. Even subtle vibration can disturb the lipid interface. Prepare and store the monolayer on a vibration-free surface to ensure proper formation. Total preparation time will be ~1.5 h.
8. Minimize the time solubilized lipid is at room temperature.

9. Chloroform should be handled in a chemical fume hood. Please review the chloroform MSDS regarding its safe use and proper disposal.
10. For a 20% Ni-NTA lipid solution, increase the Ni-NTA lipid to 8 μL and decrease the DLPC lipid to 22 μL. The volume of lipids is 30 μL in a total of 40 μL. Theoretically, up to 40 monolayers can be prepared.
11. When properly stored on ice, we have found that monolayers are stable up to 24 h.
12. Preparation of a tunable microchip specimen will take ~15 min for each specimen.
13. While 1 min is the recommended time, we have found that microchips can be left on the monolayer longer without effect. To save time, all microchips can be added to the monolayers at the same time.
14. Make sure the antibody for your target is suitable for Protein A binding. There are also His-tagged versions of Protein G adaptors commercially available for antibodies with poor Protein A affinity.
15. Typically, the antibody is diluted 1/1000 (0.2–1 ng/μL) in Buffer A as a starting point. The antibody's optimal dilution in Western blotting is a reasonable guide. Buffer A can be substituted for the user's preferred buffer. For the initial experimentation, additional antibody dilutions (higher or lower) can be tested to find the optimal concentration. When troubleshooting, consider the antibody epitope including the benefits of using a polyclonal or monoclonal antibody. Polyclonal rabbit antibody against the BRCA1 C-terminus (1/1000, 0.2 ng/μL) was used for our representative data.
16. We typically use a sample concentration of 0.01–0.02 μg/μL when preparing negatively stained specimens. Sample concentration and incubation times can be determined empirically. We used 0.02 μg/μL partially purified nuclear extract from oxidatively stressed breast cancer cells as the protein source for the representative data [25].
17. The number of washes can be increased as needed for optimal staining.
18. Staining time duration should be determined empirically. Generally, incrementally decrease the staining time if the stain is too heavy and increase if too light.
19. An in-line vacuum trap flask should be present to collect uranyl formate waste. The pipet with vacuum should not come in contact with the microchip. Alternatively, the stain can be allowed to air-dry.

20. The range of defocus values generally is smaller as a result of the consistent flatness of the SiN membrane in comparison to carbon-coated supports. The same TEM conditions are used to collect images of negatively stained and ice-embedded specimens except for varying the defocus range.

Acknowledgments

This work was supported by NIH/NCI grant R01CA193578 to D.F.K.

References

1. Frank J (2009) Single-particle reconstruction of biological macromolecules in electron microscopy—30 years. Q Rev Biophys 42(3):139–158. https://doi.org/10.1017/S0033583509990059
2. Kelly DF, Dukovski D, Walz T (2010a) A practical guide to the use of monolayer purification and affinity grids. Methods Enzymol 481:83–107. https://doi.org/10.1016/S0076-6879(10)81004-3
3. Kubalek EW, Le Grice S, Brown PO (1994) Two-dimensional crystallization of histidine-tagged, HIV-1 reverse transcriptase promoted by a novel nickel-chelating lipid. J Struct Biol 113(2):117–123. https://doi.org/10.1006/jsbi.1994.1039
4. Kelly DF, Dukovski D, Walz T (2008b) Monolayer purification: a rapid method for isolating protein complexes for single-particle electron microscopy. Proc Natl Acad Sci U S A 105(12):4703–4708. https://doi.org/10.1073/pnas.0800867105
5. Kelly DF, Abeyrathne PD, Dukovski D, Walz T (2008) The Affinity Grid: a pre-fabricated EM grid for monolayer purification. J Mol Biol 382(2):423–433. https://doi.org/10.1016/j.jmb.2008.07.023
6. Kelly DF, Lake RJ, Middelkoop TC, Fan H-Y, Artavanis-Tsakonas S, Walz T (2010c) Molecular structure and dimeric organization of the notch extracellular domain as revealed by electron microscopy. PLoS One 5(5):e10532
7. Tanner JR, Degen K, Gilmore BL, Kelly DF (2012) Capturing RNA-dependent pathways for cryo-EM analysis. Comput Struct Biotechnol J 1(1):e201204003. https://doi.org/10.5936/csbj.201204003
8. Kelly DF, Dukovski D, Walz T (2010b) Strategy for the use of affinity grids to prepare non-His-tagged macromolecular complexes for single-particle electron microscopy. J Mol Biol 400(4):675–681. https://doi.org/10.1016/j.jmb.2010.05.045
9. Sharma G, Pallesen J, Das S, Grassucci R, Langlois R, Hampton CM et al (2013) Affinity grid-based cryo-EM of PKC binding to RACK1 on the ribosome. J Struct Biol 181(2):190–194. https://doi.org/10.1016/j.jsb.2012.11.006
10. Kiss G, Chen X, Brindley MA, Campbell P, Afonso CL, Ke Z et al (2014) Capturing enveloped viruses on affinity grids for downstream cryo-electron microscopy applications. Microsc Microanal 20(1):164–174. https://doi.org/10.1017/S1431927613013937
11. Guimei Y, Kunpeng L, Pengwei H, Xi J, Wen J (2016) Antibody-based affinity cryoelectron microscopy at 2.6-Å resolution. Structure 24(11):1984–1990. https://doi.org/10.1016/j.str.2016.09.008
12. Degen K, Dukes M, Tanner JR, Kelly DF (2012) The development of affinity capture devices—a nanoscale purification platform for biological in situ transmission electron microscopy. RSC Adv. https://doi.org/10.1039/c2ra01163h
13. Gilmore BL, Showalter SP, Dukes MJ, Tanner JR, Demmert AC, McDonald SM, Kelly DF (2013a) Visualizing viral assemblies in a nanoscale biosphere. Lab Chip 13(2):216–219. https://doi.org/10.1039/c2lc41008g
14. Dukes MJ, Gilmore BL, Tanner JR, McDonald SM, Kelly DF (2013) In situ TEM of biological assemblies in liquid. J Vis Exp 82:50936. https://doi.org/10.3791/50936
15. Tanner JR, Demmert AC, Dukes MJ (2013) Cryo-SiN—an alternative substrate to visualize active viral assemblies. J Analyt Mol Tech 1(1):6

16. Pohlmann ES, Patel K, Guo S, Dukes MJ, Sheng Z, Kelly DF (2015) Real-time visualization of nanoparticles interacting with glioblastoma stem cells. Nano Lett 15(4):2329–2335. https://doi.org/10.1021/nl504481k
17. Gilmore BL, Tanner JR, McKell AO, Boudreaux CE, Dukes MJ, McDonald SM, Kelly DF (2013b) Molecular surveillance of viral processes using silicon nitride membranes. Micromachines 4:90–102. https://doi.org/10.3390/mi4010090
18. Winton CE, Gilmore BL, Demmert AC, Karageorge V, Sheng Z, Kelly DF (2016) A microchip platform for structural oncology applications. NPJ Breast Cancer 2. https://doi.org/10.1038/npjbcancer.2016.16
19. Gilmore BL, Winton CE, Demmert AC, Tanner JR, Bowman S, Karageorge V et al (2015) A molecular toolkit to visualize native protein assemblies in the context of human disease. Sci Rep 5:14440. https://doi.org/10.1038/srep14440
20. Frank J, Radermacher M, Penczek P, Zhu J, Li Y, Ladjadj M, Leith A (1996) SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J Struct Biol 116(1):190–199
21. Scheres SHW (2012) A Bayesian view on cryo-EM structure determination. J Mol Biol 415(2):406–418. https://doi.org/10.1016/j.jmb.2011.11.010
22. He Y, Yan C, Fang J, Inouye C, Tjian R, Ivanov I, Nogales E (2016) Near-atomic resolution visualization of human transcription promoter opening. Nature 533:359–365. https://doi.org/10.1038/nature17970
23. Brzovic PS, Rajagopal P, Hoyt DW, King MC, Klevit RE (2001) Structure of a BRCA1-BARD1 heterodimeric RING-RING complex. Nat Struct Biol 8(10):833–837. https://doi.org/10.1038/nsb1001-833
24. Williams RS, Green R, Glover JN (2001) Crystal structure of the BRCT repeat region from the breast cancer-associated protein BRCA1. Nat Struct Biol 8:838–842. https://doi.org/10.1038/nsb1001-838
25. Gilmore BL, Liang Y, Winton CE, Patel K, Karageorge V, Varano AC, Dearnaley W, Sheng Z, Kelly DF (2017) Molecular analysis of BRCA1 in human breast cancer cells under oxidative stress. Sci Rep 7:43435. https://doi.org/10.1038/srep43435

Chapter 4

Time-Resolved Cryo-electron Microscopy Using a Microfluidic Chip

Sandip Kaledhonkar, Ziao Fu, Howard White, and Joachim Frank

Abstract

With the advent of direct electron detectors, cryo-EM has become a popular choice for molecular structure determination. Among its advantages over X-ray crystallography are (1) no need for crystals, (2) much smaller sample volumes, and (3) the ability to determine multiple structures or conformations coexisting in one sample. In principle, kinetic experiments can be done using standard cryo-EM, but mixing and freezing grids require several seconds. However, many biological processes are much faster than that time scale, and the ensuing short-lived states of the molecules cannot be captured. Here, we lay out a detailed protocol for how to capture such intermediate states on the millisecond time scale with time-resolved cryo-EM.

Key words Time-resolved, Single-particle cryo-EM, Microfluidic chip, Spraying, Short-lived intermediates heterogeneity

1 Introduction

Single-particle cryo-EM has emerged as a front-line method to determine high-resolution structures of biological molecules. In standard single-particle cryo-EM, the sample is prepared well before it is applied to the grid, and thus the molecules in the sample are in equilibrium at the time the sample is frozen. Often multiple compositions and conformations coexist, which can only be sorted out by advanced classification techniques, and which contain clues on the reaction pathways in the equilibrium. For following a reaction in a *non-equilibrium* system, some type of time-resolved technique needs to be used to follow the course of the reaction. We need to mix two or more components at a defined time and then take samples at measured intervals for cryo-EM. These samples will then have varying compositions, as the initial components decrease and the reaction product increases in concentration over time, potentially with intermediates making their appearance over part of the time.

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_4,

When the time for completion of the reaction is in the range of hours or minutes, then each sample can be easily imaged by standard cryo-EM [1]. However, many reactions are much faster, and reaction intermediates may be short-lived, over a time span that is shorter than the time required for pipetting and blotting. For time spans in the order of 20–1000 ms, a fast time-resolved (TR) cryo-EM technique using a microfluidic chip has been developed [2–6]. Here we describe a protocol for using this technique, based on experience accumulated in our lab.

2 Materials

2.1 Sample

Purified reactant 1, reactant 2, desired buffer, and Quantifoil R1.2/1.3 400 mesh cryo-EM grids (*see* **Note 1**).

2.2 Accessories to Fix Port on the Microfluidic Chip

Heat block, heat gun, glue stick, and permanent marker.

2.3 Customized Apparatus for Grid Preparation

For the preparation of grids, we have utilized a customized machine that synchronizes injection of reactants into microfluidic chips, spraying droplets on to the grid and plunging the grid into cryogen. This machine is operated under control of customized software written in Visual Basic with time-sensitive subroutines in C++. The program Howard5e, comprising a script of commands and a graphic interface previously developed, is used to control various parameters while preparing grids (Fig. 1) [7, 8]. The apparatus comprises the following parts:

1. V6 Syringe Pumps (Norgren Kloehn, Las Vegas, NV, USA).
2. 250 μL zero dead volume syringes for the V6 Syringe Pump (Norgren Kloehn, Las Vegas, NV, USA).
3. Plunger driven by the stepper motor. IM483I microstepping drivers were used to give precise computer control of plunging and blotting utilizing HT17-068 (blotting) and HT23-599 (plunging) (Applied Motor Systems Watsonville, CA) motors.
4. Two separate supplies of high humidity 90% air are required, one to maintain humidity in the chamber and the second in the air supply to the sprayer. We have utilized 10 in. canisters commonly used for under sink water purification, which are inexpensive and will withstand the necessary pressure for the supply to the sprayer. Clear canisters are preferred because of the ease in monitoring water level. The direction of air flow is opposite that used for water purification. Compressed nitrogen (40 psi) from a cylinder is used as an air supply for the sprayer. A one-way flow valve inserted between the regulator and the first can-

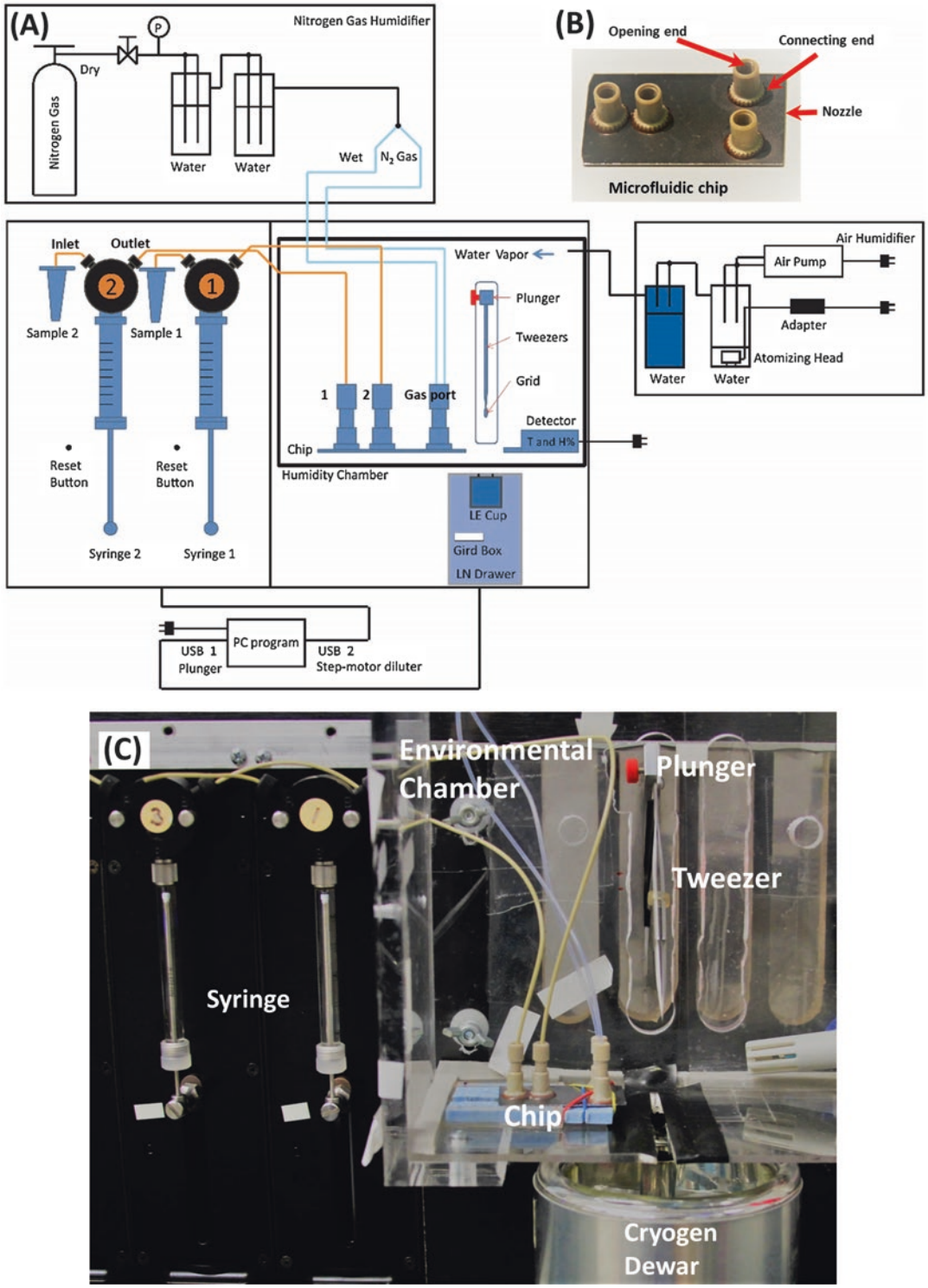


Fig. 1 Time-resolved cryo-EM apparatus. (**a**) Schematics of the time-resolved cryo-EM apparatus. (**b**) Finished microfluidic device. (**c**) Environment chamber of TR cryo-EM apparatus and microfluidic chip

ister to prevent backflow of water into the regulator when the pressure is reduced (or turned off).

5. Customized mount to fix the position of the chip.
6. Humidity control for the chamber is provided by a second pair of canisters in which the first canister contains an element from a fogger to produce high humidity and the second a sparger to absorb excess droplets, which might otherwise contaminate the grid. Air flow is produced by an aquarium pump rated 100 l/h. The humidity and temperature in the chamber is measured and the humidity maintained at 85–95% using a sensor and relay, which is used to turn the pump on and off. The temperature could also potentially be regulated but so far we have only done experiments at room temperature.
7. Cryogen dewar.
8. Microfluidic connectors, tubing.

3 Methods

3.1 Experimental Design

To study dynamic processes during a bimolecular reaction, it is imperative to know the kinetics of association and dissociation of reactants. With such knowledge, one is able to estimate the fraction of an intermediate complex at any time during the reaction. Currently, the TR cryo-EM technique has only the ability to collect data in a few discrete time points, in contrast to time-resolved spectroscopic methods where time points are in a continuous range, at multiple intervals of time resolution. Due to this limitation in TR cryo-EM, microfluidic chips with specific time point(s) need to be chosen such that the population of the reaction intermediate is maximum, so as to acquire the maximum number of particles corresponding to the reaction intermediate.

Let us consider a hypothetical biomolecular reaction in which two reactants A and B associate to form an intermediate state C, followed by conversion into the final product state D (*see* Eq. 1). The rate constants k_1 ($M^{-1}s^{-1}$) and k_2 (s^{-1})are forward reaction rate constants for the formation of the intermediate state C and the final product state D, respectively.

$$A + B \xrightarrow{k_1} C \xrightarrow{k_2} D \tag{1}$$

This hypothetical kinetic reaction is plotted as the concentration of reactants A and B versus time to determine the best time window to capture the intermediate state C (Fig. 2). The state C has the maximum yield in the time window from 50 to 300 ms, while the reaction reaches its equilibrium near 600 ms (Fig. 2). To

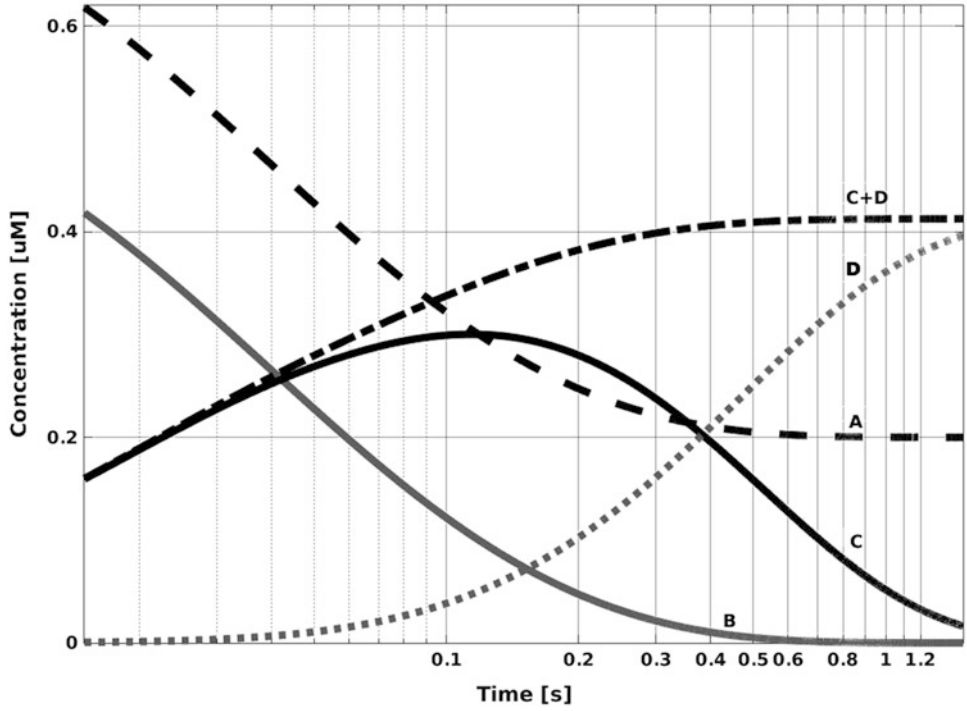


Fig. 2 The hypothetical kinetical plot for the reaction of reactants A and B yielding an intermediate state C and final product state D. An intermediate state C peaks in a window of 50–300 ms. The final product state D reaches to the maximum around 600 ms showing that the reaction has reached to equilibrium. In order to capture an intermediate state C, microfluidic chip(s) which has reaction time within 50–300 ms time scale can be employed

capture the reaction intermediate C, microfluidic chips with reaction times from 50 to 300 ms are employed.

3.2 Cryo-EM Grid Preparation

The apparatus for the preparation of sprayed cryo-grids is shown in Fig. 1. We make use of a computer-operated liquid-pumping and grid-plunging device described earlier [7]. An environmental chamber [2] monitors and controls both temperature and humidity. The microfluidic chip is fixed on the customized mount to hold it in one position. This chip mount has provisions to adjust chip positions in the horizontal plane so that the distance between the spray nozzle of the chip and the grid can be changed. The EM grid is clamped by the tweezers mounted on the plunger. The plunger is driven by a stepper motor which is program controlled [7, 8].

In all experiments, the relative humidity inside the chamber is maintained at 80–90%, and the chamber is kept at the room temperature. Compressed nitrogen gas is humidified by passing it through two consecutive water tanks. This humidified nitrogen gas is fed into the microsprayer at a manually controlled gas pressure. Once the gas flow is stable, the liquid is injected into the microsprayer chip by a syringe pump under computer control, and the liquid flow rate is set at 3 μL/s. At this point, the sprayer starts

atomizing and deposits droplets onto the grid. Finally, the grid is passed through the spray cone and plunged into liquid alkane [9]. In detail, the steps of the experiment are as follows.

3.2.1 Setting Up Time-Resolved Apparatus

The general scheme for (1) flow route of nitrogen gas to the microfluidic chip, (2) connections from the humidifier to an environmental chamber, (3) inlet and outlet connections for syringes, and (4) connections from syringes to the microfluidic chip is shown in Fig. 1a.

1. Fill the humidifier tanks with distilled water.
2. Power on the time-resolved apparatus.
3. Open the program Howard5e (*see* **Note 2**), which controls the TR apparatus.
4. Connect the inlet and outlet microfluidic tubing to each syringe, as shown in Fig. 1a.
5. Open the valve of the compressed nitrogen tank and humidify the nitrogen gas by passing it through the two consecutive water canisters as described in Subheading 2.3. Keep nitrogen gas pressure at 40 psi. The humidified nitrogen gas is divided into two paths by a splitter and connected to the two sides of the sprayer chip.
6. Reset the syringes to zero by pressing the reset button on the apparatus. Fill two 1.5-mL Eppendorf tubes with distilled deionized water (ddW) and place the inlet tubing into a 1.5-mL Eppendorf tube.
7. Sliders on the "manual" screen on the program can be used to change valve positions, load or empty syringes, and manually turn on/off other functions. Following a reset, clicking on the max button for SYRINGE1 and SYRINGE2 on the manual screen of the program will fill the syringes. Detach syringes from the pump and empty them. Draw ddW into each syringe to maximum volume, and press syringe piston slightly, so that a drop of ddW comes out of the syringe. Connect back syringes to pump. This step is necessary to remove air bubble generated by dead volume in valve and syringe.
8. Click on "Edit Run Script" under "Run Options" and change the SPRAY_VOL parameter in the parameter list at the top of the run file to D3000 which will dispense a 30 μL volume. Also check and change other parameters for plunging speed, wait time for plunging if needed (*see* **Note 3**).
9. Reset the syringe plunger to the starting position by pushing a reset button and fill the syringe using a manual slider.
10. Click "Run" button under the "Run Options" to dispense the ddW from tubing. Repeat this process a couple of times in order to hydrate tubing.

11. Mount the microfluidic chip inside the environment chamber on the holder. Connect tubing from the output of the syringes to injecting ports on the chip. Connect humidified nitrogen gas tubing to ports on the chip.
12. Change the SPRAY-VOL parameter into D1500 to inject 15 μL volume of each reactant. Each grid preparation requires 15 μL volume of each reactant.
13. Change the BEFORE-BLOT parameter to 3.5 to wait 3.5 s before plunging grid into cryogen (*see* **Note 3**). This will allow spraying stabilize by spraying for 3.5 s until the grid plunge is initiated.
14. Reset the plunger position to zero, reload the syringes, and click "Run" under "Run Options." Perform this step a couple of times to hydrate the channels of the microfluidic device.

3.2.2 Alignment

1. Perform a test run for spraying to check the cone of the spray. Adjust the horizontal distance between nozzle and plunger such that the atomizer spray covers all of the grid. Ideally, this distance is between 0.5 and 1 cm (*see* **Note 4**).
2. Put the empty liquid nitrogen dewar for transferring the grids below the environmental chamber, and check the alignment of a plunger. Check that the grid will properly enter the cryogen cup 1 cm under the surface of the ethane. Mark the position of the dewar.

3.2.3 Time-Resolved Grid Preparation

1. Load 60 μL of buffer into syringe 1 and syringe 2, and spray through the microfluidic chip.
2. Load syringe 1 with reactant 1, syringe 2 with reactant 2.
3. Turn the humidifier on, place the grid-loaded tweezer into the environmental chamber, and close the chamber.
4. Wait until humidity inside chamber reaches 80–90%.
5. Turn off the humidifier and click "Run" button to inject reactant 1 and reactant 2 into the microfluidic chip. A spray is deposited on the grid, and, under control of the program, the grid is plunged into the cryogen.
6. Transfer the grid to the grid box. Repeat **steps 4–6** to prepare additional grids (*see* **Note 5**).
7. Store grid box in liquid nitrogen dewar until data collection on microscopy.

3.2.4 Cleaning the Microfluidic Chip

1. Load the syringe with a buffer.
2. Remove the tubing 2 from port 2.
3. Inject 20 μL of a buffer into microfluidic chip 5–6 times. Reload the syringe with ddW, and inject another 5–6 times to clean the chip.

4. Disconnect the tubing 1 from port 1. Air-dry the chip and store the chip for long storage. Disconnect outlet tubing from the syringe.
5. Close the program and turn off the apparatus. Close compressed nitrogen gas tank.

3.2.5 Maintenance: Connecting Ports to Microfluidic Chip

The following steps describe how to fix ports on the microfluidic chip in case of leakage from a port(s):

1. Put the microfluidic chip in 100% ethanol for overnight, so that ports will come off easily from the microfluidic chip surface.
2. Wipe the surface of the microfluidic chip with ethanol to clean surface. Avoid wiping the area around the hole as dirt might get into the chip.
3. Place the port at the hole such that it is centered, and then mark the outer edge with the permanent marker. Mark all other holes.
4. Hold the port while inserting the 1 mL syringe's needle end into the opening end of the port.
5. Apply hot glue on port at connecting end toward the periphery. Avoid placing glue toward the center of the connecting port.
6. Place the glued connecting end of the port on the markings made on the microfluidic chip.
7. Repeat **steps 2–5** for all other connecting ports.
8. Place the microfluidic device on the heat block and wait for about 10–15 s so that glue melts and distributes evenly on the surface of the chip.
9. Align each port with respect to a center of the corresponding hole by sliding the port slowly toward the center of the hole.
10. Once every port is centered with the corresponding hole on the chip, remove the chip from heat block, and allow it to cool down for 2 min.

3.3 Data Collection and Processing

The droplets deposited on the grids prepared by mixing-spraying vary in diameter [2, 5]. The thin region of the droplets (*see* Fig. 3) needs to be identified to collect the cryo-EM micrographs [2–4]. In this laboratory, the micrographs are collected on FEI Tecnai Polara (F30) TEM with automated image collection program Leginon [10]. The micrographs are recorded with K2 Summit direct detector camera (Gatan, Inc. Pleasanton, CA) in the movie mode and frames are motion-corrected with the MotionCor2 package [11]. A representative schematic diagram of grid prepared by the TR cryo-EM apparatus and typical images acquired by cryo-EM are shown in Fig. 3. Micrographs with 1–4 μm defocus range

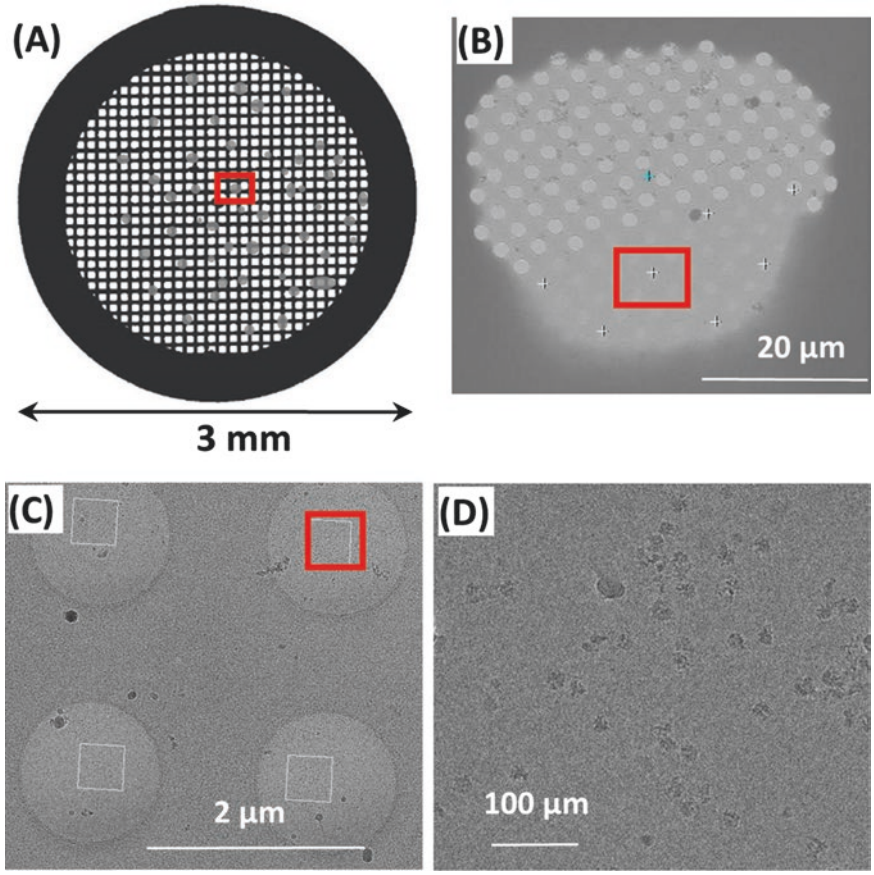


Fig. 3 Automated image collection by Leginon by targeting images from low magnification to high magnification. The images are magnified from grid view to square view, square view to hole view and hole view, and hole view to magnification at which final micrograph is acquired. The red box in each view indicates the magnified area. (**a**) Grid view: droplets of different sizes are deposited on the grid. (**b**) Square view: thin ice areas are identified from the deposited droplets on the grid. (**c**) Hole view: acquisition targets are selected at hole view. (**d**) One of such micrograph shows good particle density

are selected from all micrographs, followed by visual screening to choose good micrographs with good concentrations of particles and minimum ice contamination. At any time point, the sample deposited on the grid is heterogeneous in composition and conformation as it consists of a mixture of reactants, intermediates, and final product. The first round of 3D classification is required to sort out compositional heterogeneity, while in a second round of 3D classification conformational heterogeneity is sorted out. The details of the data processing are as follows (*see* the flowchart of the procedure in Fig. 4).

1. Perform frame alignment to correct the beam-induced motion with the MotionCor2 or unblur package [11, 12].
2. Estimate contrast transfer function (CTF) of the micrographs [13, 14].

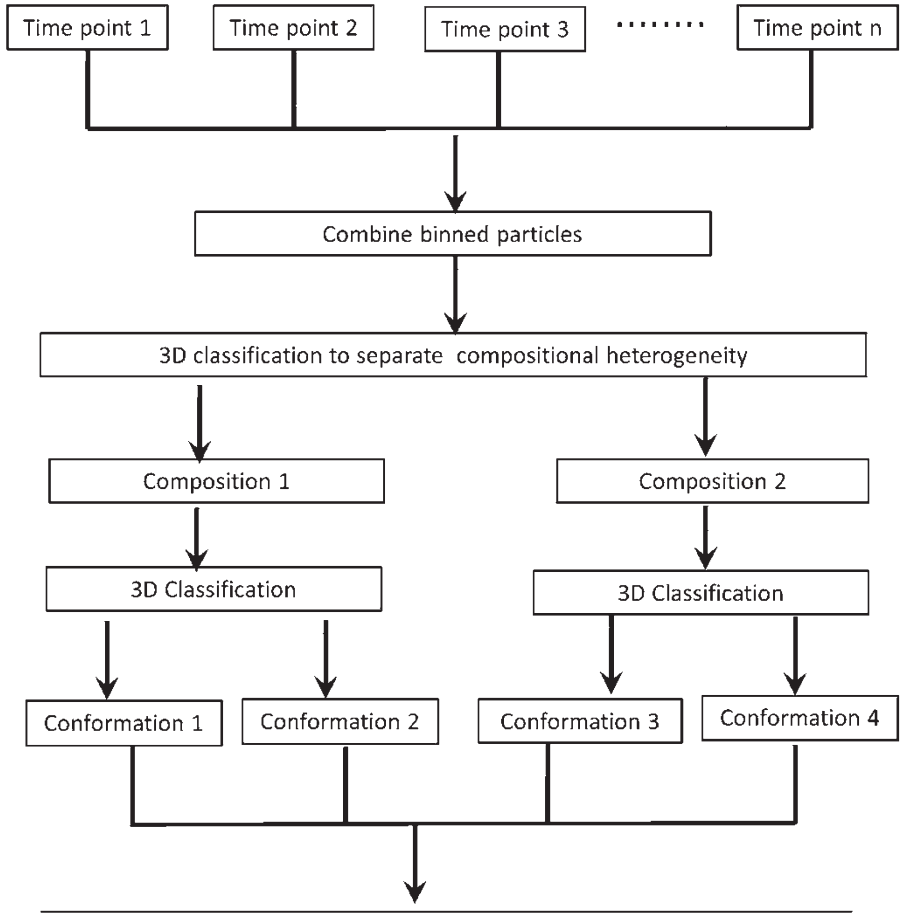


Fig. 4 Flowchart for data processing for time-resolved cryo-EM technique to trace out particles with various conformations in each time point

3. Select good micrographs from all micrographs on the basis of number of particles in the micrograph, defocus range and absence of ice contamination.
4. Particle picking: Identify particles in each micrograph and extract them with appropriate box size.
5. Perform 2D classification to separate out "junk" particles from "good" particles with RELION [15].
6. Repeat **steps 1–5** with each time point dataset.

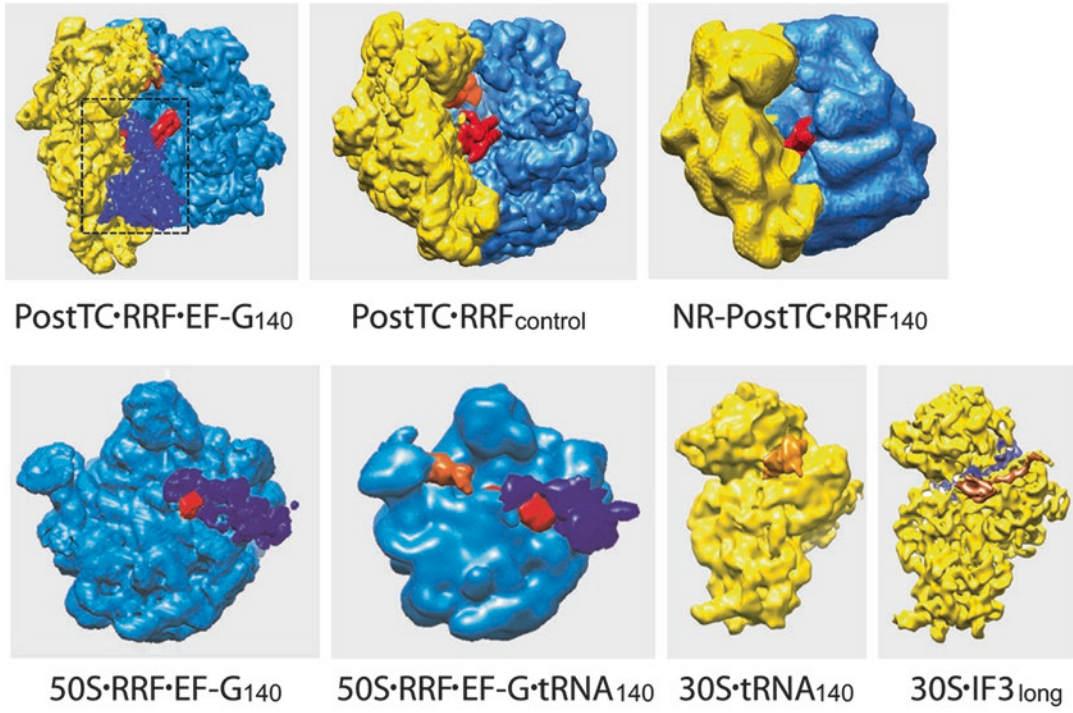


Fig. 5 Sorting out conformational compositional and conformational and heterogeneity during ribosome recycling process obtained in TR cryo-EM experiment [4]. During this experiment we obtained mainly three entities in compositional heterogeneity 30S subunit, 50S subunit, and the 70S ribosome. In terms of conformational heterogeneity, 30S subunit, 50S subunit ribosome has two conformations each, while 70S ribosome has three conformations

7. Combine the good particles from all datasets, so that 3D classification is performed with same criteria for all particles.
8. Perform 3D classification to separate out compositional heterogeneity.
9. Further 3D classification is performed on each composition to separate out conformational heterogeneity. As an example for sorting out first compositional and then conformational heterogeneity, we show different compositions and conformations obtained during the ribosome recycling process [4] (Fig. 5).
10. Determine the percentage of particles in each subpopulation with unique composition/conformation.

4 Notes

1. Grids with 400 mesh are chosen to provide more support during spray deposition and withstand nitrogen gas pressure.
2. The software is an integrative platform that controls and synchronizes actions of syringes, stepper motors, and plunger.

3. Critical parameters in the program
   (a) SPRAY-VOL$ = /AD3000 (3000 = 31.25 μL using 250 μL syringes).
   (b) SPRAY-SPEED$ = /AV300c300V300 (3.125 μL/s/syringe).
   (c) GRID-SPEED$ = 1V5000 (Plunging speed, 1V5000 = 2 m/s, 1V2500 = 1 m/s).
   (d) BEFORE-BLOT = 3.5 (wait time before plunging 3.5 s).
4. The distance between the spray nozzle and the grid is critical as a distance closer than 0.5 cm results in many droplets on grid and less spread of droplets, while a distance longer than 1 cm results in less droplets.
5. We typically load 100 μL volume at one time per reactant into the syringe which can produce six grids. Each grid preparation consumes 15 μL of volume per reactant. We do not wash the microfluidic chip with buffer after each grid preparation since syringes are filled with more than the required volume to prepare a single grid.

Acknowledgments

This research has been supported by the HHMI and NIH R01 GM29169 and GM55440 (to J.F.) and NIH AR40964 and NIH Fogarty Senior International Fellowship (to H.W.).

References

1. Fischer N, Konevega AL, Wintermeyer W, Rodnina MV, Stark H (2010) Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. Nature 466(7304):329–333. https://doi.org/10.1038/nature09206
2. Chen B, Kaledhonkar S, Sun M, Shen BX, Lu ZH, Barnard D, Lu TM, Gonzalez RL, Frank J (2015) Structural dynamics of ribosome subunit association studied by mixing-spraying time-resolved cryogenic electron microscopy. Structure 23(6):1097–1105. https://doi.org/10.1016/j.str.2015.04.007
3. Feng X, Fu Z, Kaledhonkar S, Jia Y, Shah B, Jin A, Liu Z, Sun M, Chen B, Grassucci RA, Ren Y, Jiang H, Frank J, Lin Q (2017) A fast and effective microfluidic spraying-plunging method for high-resolution single-particle cryo-EM. Structure 25(4):663–670 e663. https://doi.org/10.1016/j.str.2017.02.005
4. Fu Z, Kaledhonkar S, Borg A, Sun M, Chen B, Grassucci RA, Ehrenberg M, Frank J (2016) Key intermediates in ribosome recycling visualized by time-resolved cryoelectron microscopy. Structure 24(12):2092–2101. https://doi.org/10.1016/j.str.2016.09.014
5. Lu ZH, Shaikh TR, Barnard D, Meng X, Mohamed H, Yassin A, Mannella CA, Agrawal RK, Lu TM, Wagenknecht T (2009) Monolithic microfluidic mixing-spraying devices for time-resolved cryo-electron microscopy. J Struct Biol 168(3):388–395. https://doi.org/10.1016/j.jsb.2009.08.004
6. Shaikh TR, Yassin AS, Lu ZH, Barnard D, Meng X, Lu TM, Wagenknecht T, Agrawal RK (2014) Initial bridges between two ribosomal subunits are formed within 9.4 milliseconds, as studied by time-resolved cryo-EM. Proc Natl Acad Sci U S A 111(27):9822–9827. https://doi.org/10.1073/pnas.1406744111

7. White HD, Thirumurugan K, Walker ML, Trinick J (2003) A second generation apparatus for time-resolved electron cryo-microscopy using stepper motors and electrospray. J Struct Biol 144(1–2):246–252. https://doi.org/10.1016/j.jsb.2003.09.027
8. White HD, Walker ML, Trinick J (1998) A computer-controlled spraying-freezing apparatus for millisecond time-resolution electron cryomicroscopy. J Struct Biol 121(3):306–313. https://doi.org/10.1006/jsbi.1998.3968
9. Tivol WF, Briegel A, Jensen GJ (2008) An improved cryogen for plunge freezing. Microsc Microanal 14(5):375–379. https://doi.org/10.1017/s1431927608080781
10. Suloway C, Pulokas J, Fellmann D, Cheng A, Guerra F, Quispe J, Stagg S, Potter CS, Carragher B (2005) Automated molecular microscopy: the new Leginon system. J Struct Biol 151(1):41–60. https://doi.org/10.1016/j.jsb.2005.03.010
11. Zheng SQ, Palovcak E, Armache JP, Verba KA, Cheng YF, Agard DA (2017) MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. Nat Methods 14(4):331–332. https://doi.org/10.1038/nmeth.4193
12. Grant T, Grigorieff N (2015) Measuring the optimal exposure for single particle cryo-EM using a 2.6 angstrom reconstruction of rotavirus VP6. elife 4:e06980. https://doi.org/10.7554/eLife.06980
13. Rohou A, Grigorieff N (2015) CTFFIND4: fast and accurate defocus estimation from electron micrographs. J Struct Biol 192(2):216–221. https://doi.org/10.1016/j.jsb.2015.08.008
14. Zhang K (2016) Gctf: real-time CTF determination and correction. J Struct Biol 193(1):1–12. https://doi.org/10.1016/j.jsb.2015.11.003
15. Scheres SHW (2012) RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol 180(3):519–530. https://doi.org/10.1016/j.jsb.2012.09.006

Chapter 5

Characterizing Protein-Protein Interactions Using Solution NMR Spectroscopy

Jose Luis Ortega-Roldan, Martin Blackledge, and Malene Ringkjøbing Jensen

Abstract

In this chapter, we describe how NMR chemical shift titrations can be used to study the interaction between two proteins with emphasis on mapping the interface of the complex and determining the binding affinity from a quantitative analysis of the experimental data. In particular, we discuss the appearance of NMR spectra in different chemical exchange regimes (fast, intermediate, and slow) and how these regimes affect NMR data analysis.

Key words Protein-protein interactions, Solution NMR spectroscopy, Binding affinity, Dissociation constant, Chemical shift titration, Chemical exchange

1 Introduction

Protein-protein interactions occur with a wide range of affinities from weak complexes characterized by millimolar dissociation constants to tight complexes of picomolar affinity. While X-ray crystallography is still the most powerful technique for determining high-resolution structures of protein-protein complexes, it has its shortcomings in particular for complexes of low-to-moderate affinity or complexes displaying pervasive dynamics. Nuclear magnetic resonance (NMR) spectroscopy is a powerful complementary tool for studying interactions and is particularly suited to study complexes of low affinity [1–6]. NMR provides atomic resolution information as each NMR-active nucleus (e.g., ^{1}H, ^{15}N, or ^{13}C) of a protein experiences a distinct chemical environment and, therefore, a different resonance frequency (the so-called chemical shift).

A typical experiment for studying the interaction between a protein and a ligand (e.g., another protein, nucleic acid, or a small molecule) involves observing the perturbations of the NMR spectra of the protein upon addition of increasing amounts of the ligand. NMR signals affected by the addition of the ligand correspond to

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_5,

nuclei that change their chemical environment upon ligand binding and are therefore located in or in close proximity to the interaction site or are involved in ligand-induced structural rearrangements. In addition, by quantifying the changes in the NMR spectra (chemical shifts, linewidths, or intensities) as a function of the ligand/protein ratio, accurate information can be obtained about the complex dissociation constant and, in some cases, even about the kinetics of complex formation.

In this chapter, we demonstrate how NMR titrations can be used to study the interaction between two proteins with emphasis on mapping the interface of the complex and determining the binding affinity from a quantitative analysis of the NMR data. In particular, we discuss the appearance of NMR spectra in different chemical exchange regimes and how these regimes affect the way the NMR data should be analyzed. We use the interaction between ubiquitin and the third Src homology 3 (SH3) domain of the CD2-associated protein (CD2AP) as a model system. Obtaining the binding interface, and eventually a complete structural model, of ubiquitin in complex with various SH3 domains is important in order to understand how ubiquitin potentially regulates the interaction of SH3 domains with proline-rich ligands [7, 8].

1.1 Understanding the Appearance of NMR Spectra in Chemical Shift Titrations

In this chapter, we are considering the titration of a protein (P) with a ligand (L) according to the following equilibrium:

$$\mathrm{P} + \mathrm{L} \rightleftarrows \mathrm{PL} \tag{1}$$

The formation of the protein-ligand complex (PL) is characterized by an on-rate, k_{on}, and an off-rate, k_{off}, and the complex dissociation constant, K_D, is given by

$$K_D = \frac{k_{off}}{k_{on}} \tag{2}$$

In order to study the protein-ligand interaction by NMR, the spectrum of the protein is recorded in the absence and in the presence of increasing amounts of the ligand, and the perturbations of the protein NMR spectra are analyzed. For this purpose, two-dimensional spectra are often acquired as they offer a suitable compromise between achieving sufficient resolution for separating the protein resonances and keeping a limited total experimental acquisition time. In particular, the ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) experiment is particularly useful as it provides a "fingerprint" of the protein, where each resonance corresponds to a backbone or side chain amide group. These experiments also allow to take advantage of differential isotope labelling of the protein (^{15}N-labelled) and the ligand (unlabelled) such that only the protein is observed in the NMR experiments, while the ligand remains NMR invisible.

During the course of a chemical shift titration, the appearance of the spectra depends on the chemical shift difference (^{1}H and/or ^{15}N), $\Delta\omega$, between the free and ligand-bound protein compared to the exchange rate, k_{ex}, given by

$$k_{ex} = k_{on}[L]_{free} + k_{off} \tag{3}$$

where $[L]_{free}$ is the concentration of unbound ligand. In the fast-exchange regime ($k_{ex} \gg \Delta\omega$), a single exchange-averaged NMR resonance will be observed for each nucleus whose chemical shift, δ_{obs}, is given by the population-weighted average between the free and ligand-bound chemical shifts [9]

$$\delta_{obs} = p_B\delta_{bound} + (1 - p_B)\delta_{free} \tag{4}$$

Here, p_B is the population of the protein in the ligand-bound state. Thus, in the fast-exchange regime, a progressive change in the NMR resonance frequency is observed for increasing ligand concentration as illustrated in Fig. 1a. This exchange regime is characterized by a lack of line broadening of the NMR signals during the course of the titration or potentially by a small line broadening if an increase in the overall molecular tumbling rate occurs upon complex formation (this would normally be the case for protein-protein interactions).

In order to obtain information about the dissociation constant of the complex from the experimental chemical shift changes, we note that the population, p_B, of bound protein is related to the dissociation constant and the total concentrations of the protein ([P]) and ligand ([L]) according to

$$p_B = \frac{[P] + [L] + K_D - \sqrt{([P] + [L] + K_D)^2 - 4[P][L]}}{2[P]} \tag{5}$$

The chemical shift perturbation ($\Delta\delta$) observed during the titration experiment is then given by the following equation [10]:

$$\Delta\delta = \delta_{obs} - \delta_{free} = \frac{(\delta_{bound} - \delta_{free})}{2[P]}\left[[P] + [L] + K_D - \sqrt{([P] + [L] + K_D)^2 - 4[P][L]}\right] \tag{6}$$

Thus, in the case of fast exchange, by quantifying the chemical shift changes in the protein as a function of known concentrations of protein and ligand, the dissociation constant can be determined by analysis of the experimental data using Eq. 6. We note that this equation is only valid for complexes of 1:1 stoichiometry.

In the case of slow exchange ($k_{ex} \ll \Delta\omega$), no chemical shift perturbations will be observed directly but only modulations of the NMR signal intensities. Thus, as the ligand is added to the protein,

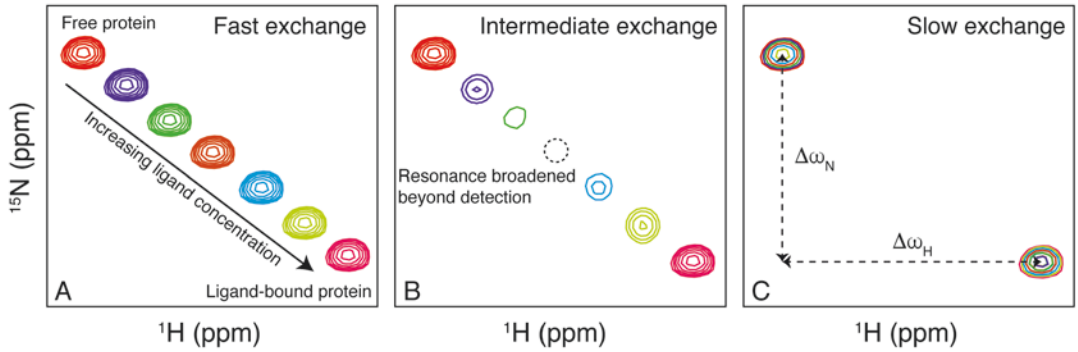


Fig. 1 Schematic representation of different exchange regimes often encountered when studying protein-ligand interactions by NMR chemical shift titrations. We assume that the protein is ^{15}N-labelled thereby allowing observation of its resonances in ^{1}H-^{15}N HSQC spectra. The ligand, which is unlabelled and therefore not observable in the NMR spectra, can be a small molecule or another protein. We assume that the protein concentration is constant and that the concentration of the ligand is gradually increased until saturation of the protein. (**a**) In fast exchange ($k_{ex} \gg \Delta\omega$), the resonances undergo a chemical shift perturbation corresponding to the population-weighted average between free and bound-state chemical shifts (Eq. 4). (**b**) In intermediate exchange ($k_{ex} \approx \Delta\omega$), the NMR signals undergo a line broadening upon addition of the ligand due to transverse relaxation rate contributions from conformational exchange (free-bound equilibrium) occurring on the micro- to millisecond time scale. The chemical shift perturbations can, in this case, no longer be described by Eq. 4. (**c**) In slow exchange ($k_{ex} \ll \Delta\omega$), no chemical shift perturbations are observed but only a modulation of the NMR signal intensities that directly correspond to the populations of free and bound protein (Eq. 7)

the resonance corresponding to the free protein will gradually disappear, and a signal corresponding to the bound protein will appear (Fig. 1c). In this case, the NMR signal intensities report directly on the populations of free (p_F) and bound protein (p_B) according to

$$p_B = 1 - p_F = \frac{I_B}{I_F + I_B} \tag{7}$$

Here, I_F and I_B are the NMR signal intensities of the resonances corresponding to the free and bound proteins, respectively, and the dissociation constant of the complex can subsequently be derived for known concentrations of protein and ligand. We note that in the slow-exchange regime, it is necessary to reassign the NMR spectra of the complex state, in order to access the chemical shift and intensity differences between the free and bound state of the protein.

In the intermediate-exchange regime ($k_{ex} \approx \Delta\omega$), a contribution to the NMR signal linewidth will be observed from the conformational exchange between free and bound forms of the protein that, in some cases, can become so large that the resonances are broadened beyond detection (Fig. 1b). Although chemical shift perturbations may be observed in this exchange regime, they do not necessarily follow Eq. 4, and the dissociation constant cannot therefore be readily derived from the chemical shift changes (if

measurable). Instead, a line shape analysis is necessary, where the linewidths of the NMR signals during the course of the titration are quantified and used to provide a measure of the dissociation constant along with the kinetic constants, k_{on} and k_{off} [11]. Alternatively, more elaborate experimental procedures can be used such as measurements of ^{15}N nuclear transverse relaxation rates at intermediate titration points (sub-stoichiometric amounts of ligand compared to protein). These experiments allow accurate quantification of the ^{15}N exchange contributions and, therefore, provide access to the population of the ligand-bound protein along with the parameters characterizing the kinetics of the complex formation [12–17]. Thus, although the analysis of the NMR spectra is not straightforward in the case of the intermediate-exchange regime, an advantage is that the NMR spectra contain additional information about the kinetic constants characterizing the underlying conformational equilibrium.

In relation to the different exchange regimes observed in NMR chemical shift titrations, it is important to understand that there is no direct link between the chemical exchange regime observed in the NMR interaction experiments and the binding affinity of the protein-ligand complex under investigation. As described above, the exchange regime (fast, intermediate, or slow) depends on the kinetics (k_{on} and k_{off}) of the equilibrium compared to the chemical shift difference between free and ligand-bound proteins (Eq. 3) and not directly on the dissociation constant that only depends on the ratio of the kinetic constants (Eq. 2). We also note that if ^{1}H-^{15}N HSQC experiments are recorded for each titration step, different exchange regimes can in principle be experienced in the ^{1}H and ^{15}N dimensions, because the ^{1}H and ^{15}N nuclei do not necessarily display the same chemical shift difference between the free and ligand-bound states.

2 Materials

1. ^{15}N-labelled and unlabelled ubiquitin.
2. ^{15}N-labelled and unlabelled third SH3 domain of the CD2-associated protein (SH3-C).
3. Buffer for NMR: 50 mM sodium phosphate buffer, 1 mM DTT at pH 6.0.
4. UV spectrophotometer for measurement of absorption at 280 nm for determination of accurate protein concentrations.
5. D_2O for addition to the NMR samples (10% v/v) for maintaining the lock signal.
6. NMR tubes (3 mm, 5 mm or Shigemi).
7. Pasteur pipette for transfer of protein samples to NMR tubes.

8. NMR spectrometer operating at a ^{1}H frequency of 600 MHz or above.
9. Software for processing NMR data (e.g., NMRPipe [18]).
10. Analysis software for NMR spectra (e.g., Sparky [19]).
11. Software for visualization of PDB files (e.g., PyMOL).

3 Methods

3.1 Protein Sample Preparation

Samples of ubiquitin and SH3-C for the NMR experiments were expressed in *Escherichia coli* (*E. coli*) Rosetta(DE3) strain with an N-terminal hexahistidine tag cleavable by the tobacco etch virus (TEV) protease [20, 21]. For proteins labelled with ^{15}N, cells were grown in M9 minimal medium with $^{15}NH_4Cl$ as the sole nitrogen source (*see* **Note 1**). The proteins were purified using a Ni-affinity column, followed by cleavage of the histidine tag and a final purification step using size-exclusion chromatography. After verification of their purity on sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), all samples were dialyzed into the same NMR buffer (*see* **Note 2**). After dialysis, the protein samples were concentrated using centrifugal filters with a molecular weight cutoff of 3 kDa, and the concentrations of the proteins were subsequently measured by UV absorbance at 280 nm using the NMR buffer as a blank measurement and the theoretical extinction coefficients derived from the primary sequence of the two proteins using the ProtParam tool (www.expasy.org) (*see* **Note 3**). Ten percent (v/v) of D_2O were added to all samples in order to maintain the lock signal and avoid drift of the magnetic field during the NMR measurements.

3.2 NMR Acquisition: General Considerations

To monitor a protein-protein interaction by NMR, an ^{1}H-^{15}N correlated spectrum such as the HSQC experiment [22] is particularly suitable as it provides a single resonance for each backbone ^{1}H-^{15}N pair, with the exception of prolines that do not contain amide protons. Acquisition of a ^{1}H-^{15}N HSQC spectrum requires initial tuning and matching of the probe as well as shimming in order to ensure a homogeneous magnetic field across the NMR sample. In addition, a number of calibrations are necessary such as the measurement of the water resonance frequency in order to achieve efficient water suppression, as well as calibration of the ^{1}H and ^{15}N pulse lengths to ensure optimal signal intensity. In order to achieve maximum resolution per total measurement time, the number of complex points and the spectral widths in both the ^{1}H and the ^{15}N dimension should be adjusted (*see* **Note 4**). We note that for larger proteins with elevated rotational correlation times, it may be an

advantage to employ ^{1}H-^{15}N transverse relaxation-optimized spectroscopy (TROSY) [23] instead of the HSQC experiment.

3.3 Chemical Shift Titrations

In order to study the interaction between ubiquitin and the SH3 domain (SH3-C) of CD2AP, chemical shift titrations were carried out. Initially, a 0.25 mM ^{15}N-labelled sample of SH3-C was titrated with unlabelled ubiquitin until reaching a [ubiquitin]/[SH3-C] ratio of 2.3, and the complementary titration was carried out by titrating a 0.25 mM ^{15}N-labelled sample of ubiquitin with unlabelled SH3-C until reaching a [SH3-C]/[ubiquitin] ratio of 3.8 (*see* **Note 5**). For each step of the two titrations, a ^{1}H-^{15}N HSQC spectrum was recorded in order to monitor chemical shift perturbations in the two proteins arising from their mutual interaction (Fig. 2a, d). Looking at the HSQC spectra recorded during the titration experiments of ubiquitin and SH3-C (Fig. 2a, d), we observe a single resonance for each amide group that change chemical shift upon addition of the unlabelled protein. In addition, no significant line broadening is observed during the course of the titration clearly showing that the ubiquitin/SH3-C interaction falls within the fast-exchange regime. We note that some resonances undergo large changes in chemical shifts corresponding to residues in or close to the interaction site, while some resonances do not display any chemical shift perturbations and, therefore, correspond to residues outside the interacting regions. In addition, the NMR signals move in a linear fashion between the first and the last titration point indicative of the absence of complex intermediates that presumably would display different chemical shifts compared to the free or fully bound states thereby leading to "curved" chemical shift perturbation profiles.

3.4 Mapping of the Binding Interface Between SH3-C and Ubiquitin

In order to analyze in more detail the interaction between ubiquitin and SH3-C, the assignment of the different NMR backbone signals is necessary. The spectral assignment of proteins is typically obtained through the use of double-labelled (^{13}C, ^{15}N) protein samples allowing the acquisition of a set of three-dimensional triple resonance experiments containing correlations of the backbone ^{15}N and $^{1}H^{N}$ chemical shifts with the corresponding intra and/or inter-residue $^{13}C'$, $^{13}C^{\beta}$, and $^{13}C^{\alpha}$ chemical shifts [24, 25]. By matching intra and inter-residue $^{13}C'$, $^{13}C^{\beta}$, and $^{13}C^{\alpha}$ chemical shifts simultaneously, a sequential walk through the backbone of the protein can be achieved thereby assigning each of the HSQC resonances to a specific amide group in the protein. Analysis of the spectra can either be done manually or through the use of more automated procedures [26].

Following the assignment of the two proteins, the chemical shift perturbations in both ubiquitin and SH3-C can be quantified by measuring the chemical shift differences between the first and last titration points for each residue. Combined ^{1}H and ^{15}N chemical shift perturbations (*see* **Note 6**) are plotted as a function of the residue number of the two proteins (Fig. 2b, e). These plots provide an

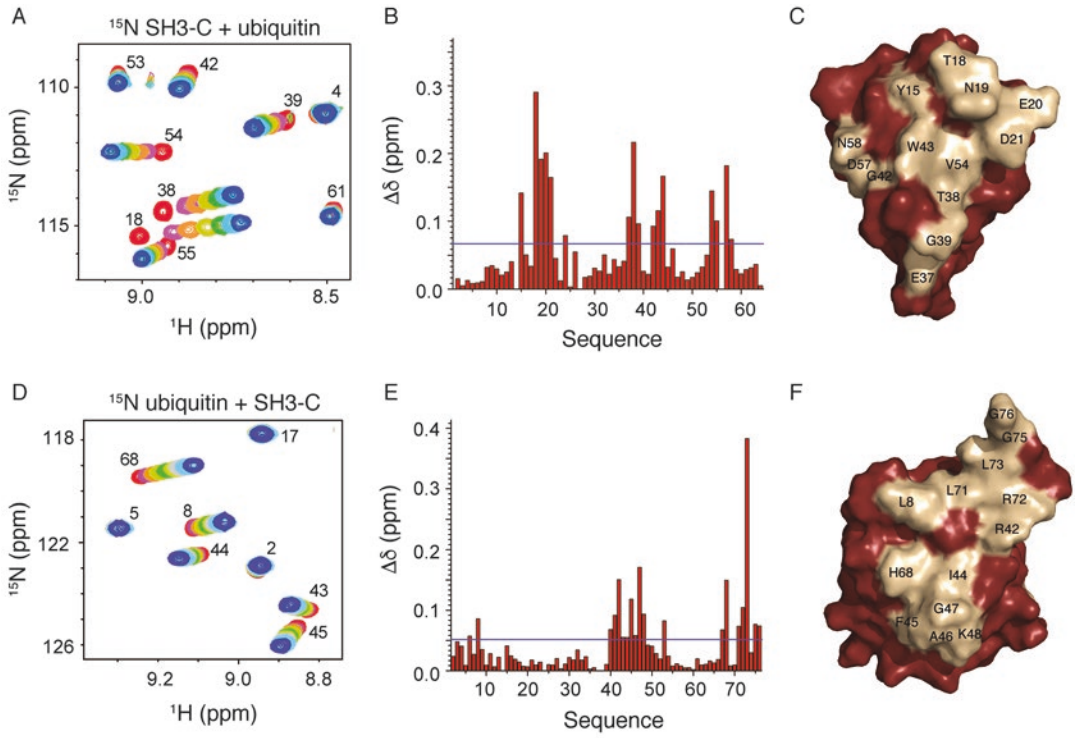


Fig. 2 Studies of the protein-protein complex of ubiquitin and the third SH3 domain of CD2AP using NMR chemical shift titration experiments. (**a**) Section of the ^{1}H-^{15}N HSQC spectrum of SH3-C showing chemical shift perturbations upon addition of increasing amounts of unlabelled ubiquitin (red, free SH3-C; magenta, orange, yellow, green, cyan, and blue, spectra recorded with increasing ubiquitin concentration). (**b**) Combined ^{1}H and ^{15}N chemical shift perturbations in SH3-C (Eq. 8) obtained between the first (red) and last (blue) titration points shown in **a**. The blue solid line corresponds to the mean chemical shift perturbation. (**c**) Surface representation of SH3-C showing the location of the residues (beige) with largest chemical shift perturbations upon addition of ubiquitin. (**d**) Section of the ^{1}H-^{15}N HSQC spectrum of ubiquitin showing chemical shift perturbations upon addition of increasing amounts of unlabelled SH3-C (red, free ubiquitin; magenta, orange, yellow, green, cyan, and blue, spectra recorded with increasing SH3-C concentration). (**e**) Combined ^{1}H and ^{15}N chemical shift perturbations in ubiquitin. (**f**) Surface representation of ubiquitin showing the location of the residues (beige) with largest chemical shift perturbations upon addition of SH3-C

immediate overview over the interacting regions of the two proteins. The residues with the largest chemical shift perturbations can be visualized on the three-dimensional structures of SH3-C and ubiquitin clearly displaying the binding interface between the two proteins (Fig. 2c, f). A structural model of the SH3-C/ubiquitin complex can subsequently be obtained from the experimental chemical shift perturbations by rigid body or flexible docking using the three-dimensional structure of the two proteins as starting models. This can, for example, be performed by the program HADDOCK that uses ambiguous interaction restraints from the chemical shift perturbations to drive the docking [27]. Distinguishing interacting and noninteracting residues can be challenging, in par-

ticular, if only small chemical shift perturbations are observed (as can be the case, e.g., for proteins interacting with small molecules) or if multiple peaks are shifting following larger structural rearrangements upon interaction. Statistical approaches can be helpful in differentiating between interacting and noninteracting residues [28]. If a set of interacting residues derived from chemical shift perturbations is to be used for docking, care must be taken to select only the interacting residues with sufficient solvent accessibility. Nevertheless, in some cases, a unique structural model cannot be obtained from this procedure due to rotational degeneracy at the interaction interface, and the collection of other distance and/or orientational restraints may be necessary such as paramagnetic relaxation enhancement (PREs) [29], residual dipolar couplings (RDCs) [21, 30–32], or nuclear relaxation rates [15].

3.5 Determination of the Dissociation Constant of the SH3-C/Ubiquitin Complex

In order to determine the dissociation constant of the SH3-C/ubiquitin complex, the dependence of the chemical shift perturbations on the total concentrations of ubiquitin and SH3-C is analyzed. As the ubiquitin/SH3-C interaction falls within the fast-exchange regime, the chemical shift changes follow Eq. 6. By a least-square analysis of the evolution of the experimental chemical shift perturbations for residues in SH3-C (Fig. 3a) and in ubiquitin (Fig. 3b) as a function of the total concentrations of the two proteins, a K_D value of 132 ± 13 μM can be derived [21]. In this case, it is an advantage to analyze multiple chemical shift perturbations simultaneously in order to increase the stability of the fitting procedure. This means that a single K_D value is obtained by analyzing the results of several residues simultane-

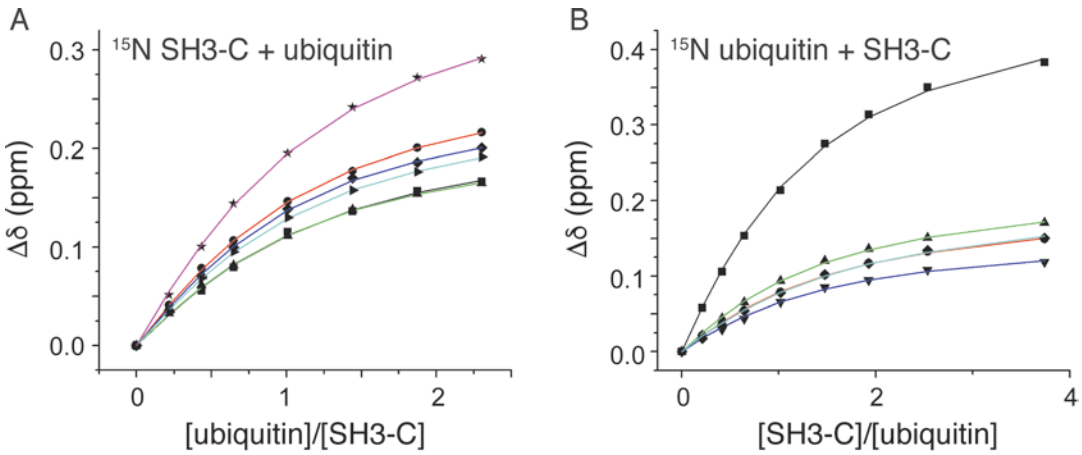


Fig. 3 Determination of the dissociation constant of the ubiquitin-SH3-C complex. (**a**) Simultaneous fit of Eq. 6 to chemical shift perturbations for selected residues in SH3-C: T18 (magenta), N19 (cyan), E20 (blue), D21 (green), T38 (red), and W44 (black). (**b**) Simultaneous fit of Eq. 6 to chemical shift perturbations for selected residues in ubiquitin: R42 (cyan), F45 (blue), G47 (green), H68 (red), and L73 (black)

ously, either from the two proteins independently or from both proteins by exploiting simultaneously the two complementary chemical shift titrations.

The obtained NMR data show that the complex formed between ubiquitin and SH3-C is of relatively low affinity. Such complexes are extremely challenging to study at high resolution by other complementary techniques such as X-ray crystallography. NMR therefore retains an important place in structural biology in order to access at atomic resolution the transient interactome that continuously plays important roles in biology.

4 Notes

1. For expression of labelled proteins in minimal media, it may be an advantage to follow the protocol proposed by Marley et al. [33], where cell mass is initially generated using unlabelled rich media followed by transfer into a smaller volume of labelled media at high cell density. After a short period of growth recovery, the cells are subsequently induced. The protocol offers a substantial reduction in isotope costs (four- to eightfold) compared to direct expression in M9 minimal media.
2. In an NMR titration experiment, it is important that the interacting proteins are dialyzed into exactly the same buffer in order to avoid chemical shift perturbations during the course of the titration arising from changing buffer conditions (e.g., changing pH or salt concentration).
3. Accurate determination of protein concentrations is essential for deriving accurate affinity constants using NMR titration experiments. In general, we can consider that a protein concentration of at least 150 μM is necessary for obtaining sufficient NMR signal intensity within a reasonable experimental time. For a chemical shift titration experiment involving two different proteins, it is usually necessary to concentrate more the unlabelled protein than the labelled protein. Thus, saturation can be achieved without excessive dilution of the labelled protein that is detected by the NMR experiment.
4. In a ^{1}H-^{15}N HSQC spectrum, the frequencies of the ^{15}N nuclei are detected in the indirect dimension, and the total measurement time is therefore determined by the number of points acquired in the ^{15}N dimension, as well as the number of transients (scans). To achieve maximum resolution in the ^{15}N dimension per total measurement time, it is therefore important to adjust the spectral width (sweep width) to the absolute minimal value that covers all NMR resonances of the protein. The fre-

quencies of the ^{1}H nuclei are measured in the direct dimension, and there are in principle no restrictions on the resolution that can be achieved (increasing the number of acquired points will increase the acquisition time for a constant sweep width). In practice, the acquisition time is limited to a few hundred milliseconds (ms) to avoid sample heating and probe damage due to the strong ^{15}N decoupling field that must be applied for the entire duration of the acquisition time in order to remove the ^{1}H-^{15}N scalar couplings in the ^{1}H dimension. Thus, in practice the ^{1}H sweep width should be set to a value that covers all signals in the ^{1}H dimension, and the number of points is subsequently limited by the acquisition time (<120 ms).

5. To achieve the most accurate determination of the dissociation constant from the chemical shift perturbations, it is necessary to approach complete saturation of the labelled protein by the unlabelled protein. In practice this is done by adding the unlabelled protein to the labelled protein until no further chemical shift perturbations are observed. In order to avoid strong dilution of the labelled protein during the course of the titration, a concentrated solution of the unlabelled protein should be used (e.g., five times the concentration of the labelled protein). Alternatively, another strategy can be applied where two samples of the ^{15}N-labelled protein are prepared, one in the absence of the unlabelled protein and one with a saturating amount of the unlabelled protein. The two samples should have the same concentration of the ^{15}N-labelled protein. Intermediate titration points can then be obtained by mixing these two samples in different ratios. The advantage of this approach is that the labelled protein will not undergo dilution during the course of the titration.
6. In NMR titrations, the combined chemical shift difference, $\Delta\delta$, for ^{1}H and ^{15}N nuclei is often reported. In order to account for the larger chemical shift dispersion in the ^{15}N dimension compared to ^{1}H, a scaling factor is applied to the chemical shift perturbations in the ^{15}N dimension according to

$$\Delta\delta = \sqrt{\left(\frac{\Delta\delta_{N}}{R_{scale}}\right)^{2} + \left(\Delta\delta_{HN}\right)^{2}} \tag{8}$$

In many studies, a scaling factor, R_{scale}, of 10 is applied corresponding to the ratio of the gyromagnetic ratios of ^{15}N and ^{1}H. In this study, we use a scaling factor of 6.5 that has been derived from experimental chemical shift distributions of ^{1}H and ^{15}N in the BioMagResBank (www.bmrb.wisc.edu) [34].

References

1. Zuiderweg ERP (2002) Mapping protein-protein interactions in solution by NMR spectroscopy. Biochemistry 41:1–7
2. Vaynberg J, Qin J (2006) Weak protein-protein interactions as probed by NMR spectroscopy. Trends Biotechnol 24:22–27
3. Takeuchi K, Wagner G (2006) NMR studies of protein interactions. Curr Opin Struct Biol 16:109–117
4. Fielding L (2007) NMR methods for the determination of protein-ligand dissociation constants. Prog Nucl Magn Reson Spec 51:219–242
5. O'Connell MR, Gamsjaeger R, Mackay JP (2009) The structural analysis of protein-protein interactions by NMR spectroscopy. Proteomics 9:5224–5232
6. Vinogradova O, Qin J (2012) NMR as a unique tool in assessment and complex determination of weak protein-protein interactions. Top Curr Chem 326:35–45
7. Stamenova SD, French ME, He Y et al (2007) Ubiquitin binds to and regulates a subset of SH3 domains. Mol Cell 25:273–284
8. Ortega Roldan JL, Casares S, Jensen MR et al (2013) Distinct ubiquitin binding modes exhibited by SH3 domains: molecular determinants and functional implications. PloS One 8:e73018
9. Williamson MP (2013) Using chemical shift perturbation to characterise ligand binding. Prog Nucl Magn Reson Spec 73:1–16
10. Jensen MR, Ortega-Roldan JL, Salmon L et al (2011) Characterizing weak protein-protein complexes by NMR residual dipolar couplings. Eur Biophys J 40:1371–1381
11. Waudby CA, Ramos A, Cabrita LD et al (2016) Two-dimensional NMR Lineshape analysis. Sci Rep 6:24826
12. Palmer AG, Kroenke CD, Loria JP (2001) Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. Methods Enzymol 339:204–238
13. Sugase K, Dyson HJ, Wright PE (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 447:1021–1025
14. Hansen DF, Vallurupalli P, Kay LE (2008) Using relaxation dispersion NMR spectroscopy to determine structures of excited, invisible protein states. J Biomol NMR 41:113–120
15. Salmon L, Ortega Roldan JL, Lescop E et al (2011) Structure, dynamics, and kinetics of weak protein-protein complexes from NMR spin relaxation measurements of titrated solutions. Angew Chem 50:3755–3759
16. Schneider R, Maurin D, Communie G et al (2015) Visualizing the molecular recognition trajectory of an intrinsically disordered protein using multinuclear relaxation dispersion NMR. J Am Chem Soc 137:1220–1229
17. Kragelj J, Palencia A, Nanao MH et al (2015) Structure and dynamics of the MKK7-JNK signaling complex. Proc Natl Acad Sci 112:3409–3414
18. Delaglio F, Grzesiek S, Vuister GW et al (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293
19. Goddard TD, Kneller DG. SPARKY 3, University of California, San Francisco
20. Ortega Roldan JL, Romero Romero ML, Ora A et al (2007) The high resolution NMR structure of the third SH3 domain of CD2AP. J Biomol NMR 39:331–336
21. Ortega Roldan JL, Jensen MR, Brutscher B et al (2009) Accurate characterization of weak macromolecular interactions by titration of NMR residual dipolar couplings: application to the CD2AP SH3-C:ubiquitin complex. Nucleic Acids Res 37:e70
22. Bodenhausen G, Ruben DJ (1980) Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett 69:185–189
23. Pervushin K, Riek R, Wider G et al (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci 94:12366–12371
24. Ikura M, Kay LE, Bax A (1990) A novel approach for sequential assignment of 1H, 13C, and 15 N spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. Biochemistry 29:4659–4667
25. Kay LE, Ikura M, Tschudin R et al (1969) (1990) three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. J Magn Reson 89:496–514
26. Jung YS, Zweckstetter M (2004) Mars – robust automatic backbone assignment of proteins. J Biomol NMR 30:11–23
27. Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: a protein-protein docking

approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737

28. Schumann FH, Riepl H, Maurer T et al (2007) Combined chemical shift changes and amino acid specific chemical shift mapping of protein-protein interactions. J Biomol NMR 39:275–289
29. Clore GM, Tang C, Iwahara J (2007) Elucidating transient macromolecular interactions using paramagnetic relaxation enhancement. Curr Opin Struct Biol 17:603–616
30. Tolman JR, Flanagan JM, Kennedy MA et al (1995) Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. Proc Natl Acad Sci 92:9279–9283
31. Tjandra N, Bax A (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science 278:1111–1114
32. Blackledge M (2005) Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. Prog Nucl Magn Reson Spec 46:23–61
33. Marley J, Lu M, Bracken C (2001) A method for efficient isotopic labeling of recombinant proteins. J Biomol NMR 20:71–75
34. Mulder FAA, Schipper D, Bott R et al (1999) Altered flexibility in the substrate-binding site of related native and engineered high-alkaline Bacillus subtilisins 1. J Mol Biol 292:111–123

Chapter 6

Reconstitution of Isotopically Labeled Ribosomal Protein L29 in the 50S Large Ribosomal Subunit for Solution-State and Solid-State NMR

Emeline Barbet-Massin, Eli van der Sluis, Joanna Musial, Roland Beckmann, and Bernd Reif

Abstract

Solid-state nuclear magnetic resonance (NMR) has recently emerged as a method of choice to study structural and dynamic properties of large biomolecular complexes at atomic resolution. Indeed, recent technological and methodological developments have enabled the study of ever more complex systems in the solid-state. However, to explore multicomponent protein complexes by NMR, specific labeling schemes need to be developed that are dependent on the biological question to be answered. We show here how to reconstitute an isotopically labeled protein within the unlabeled 50S or 70S ribosomal subunit. In particular, we focus on the 63-residue ribosomal protein L29 (~7 kDa), which is located at the exit of the tunnel of the large 50S ribosomal subunit (~1.5 MDa). The aim of this work is the preparation of a suitable sample to investigate allosteric conformational changes in a ribosomal protein that are induced by the nascent polypeptide chain and that trigger the interaction with different chaperones (e.g., trigger factor or SRP).

Key words Isotope labeling, Protein complex reconstitution, MAS solid-state NMR spectroscopy

1 Introduction

Biomacromolecular complexes are ubiquitous in nature and play fundamental roles in biological processes. The study of the architecture, dynamics, and functional properties of multicomponent complexes is often highly challenging because of their high molecular weights (above 1 MDa) but also because of their complex compositions. Recent advances in cryo-electron microscopy (cryo-EM) have allowed the determination of structures at a resolution comparable to X-ray crystallography (<3 Å) [1–3], paving the way for the structural analysis of many macromolecular complexes that had escaped crystallization so far [4]. However, knowing a structure does not necessarily mean that the dynamic processes underlying the function of such complexes are understood. Molecular

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_6,

machines such as the ribosome work by using dynamic conformational changes to carry out their functions. To unravel its mechanisms of action, structures in the presence and absence of ligands and binding partners need to be compared. Doing so can be difficult when relying exclusively on X-ray crystallography or on cryo-EM. Many dynamic complexes cannot be crystallized and both methods do not yield native dynamic features. High-resolution MAS solid-state NMR has recently emerged as a powerful technique to characterize systems that cannot be investigated by other methods. Especially for protein complexes that are too large for solution-state NMR studies, solid-state NMR represents a complementary technique to yield atomic resolution structural information [5–11]. Today, protocols for sample preparation, resonance assignment, and collection of structural restraints have been established that have allowed the determination of three-dimensional structures of biomolecular complexes in the solid state, such as fibrils [12–16], membrane-associated systems [17, 18], virus particles [19, 20], or cytoskeleton-binding proteins [21, 22]. Most importantly, NMR is able to study allosteric mechanisms at atomic resolution [23, 24] and holds the potential to quantify dynamic processes [25, 26].

In particular, we aim for the analysis of conformational changes induced by nascent polypeptide chains in essential ribosomal proteins such as L23 or L29 that are located in the large ribosomal subunit close to the exit of the tunnel. Further, we would like to understand how these conformational changes induce binding of trigger factor (TF) and/or SRP [27]. For these experiments, specific labeling schemes are necessary, in combination with approaches to reconstitute the ribosome employing unlabeled and labeled proteins to simplify spectral patterns. The approach that we take for labeling and reconstitution is outlined in Fig. 1.

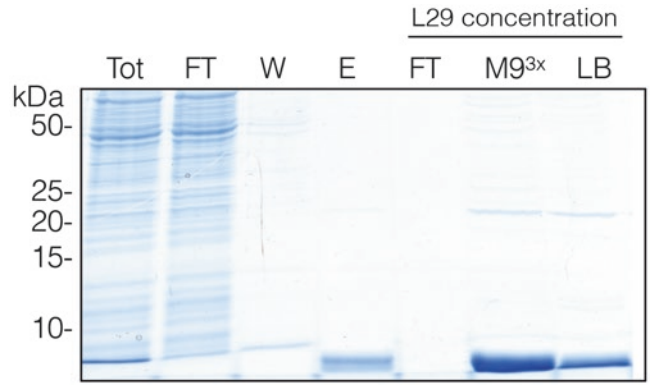


Fig. 1 SDS PAGE of purified L29. From left to right: ***Tot*** total lysate fraction before application on the Ni-NTA beads, *FT* flow-through of the applied lysate fraction, *W* wash step, *E* elution, *FT* flow-through of the concentration step, *M9^{3x}* purified and concentrated triply labeled L29 (expressed in triply labeled M9), *LB* purified and unlabeled L29 (expressed in LB) for reference (concentration: 8 μM)

2 Materials

All solutions are prepared using ultrapure autoclaved water and analytical-grade reagents. All stock solutions are filtered using 0.22 μm membrane filters to ensure a high purity of the buffers. Special care with respect to purity is in particular required for buffers employed in sucrose gradient experiments to prevent clogging of the gradient machine.

2.1 Production of u-[^{2}H,^{13}C,^{15}N, 100% H^{N}]-L29

All solutions are made in D_2O and filter-sterilized using a 0.22 μm membrane. The following method can be used for most of recombinant proteins (*see* **Note 1**):

1. LB plate of ER2566 *E. coli* cells transformed with a pET28 plasmid containing the sequence coding for His-tagged L29 cloned behind a T7 promoter and a resistance gene against kanamycin.
2. 1 L of D_2O.
3. 100 mL of 10× deuterated minimal medium (D_2O-M9 10×): 60 g/L Na_2HPO_4, 30 g/L KH_2PO_4, and 5 g/L NaCl.
4. 2 g of ^{2}H,^{13}C-glucose and 1 g of $^{15}NH_4Cl$, dissolved in D_2O.
5. 10 mL of a trace element solution (100× in D_2O): 5 g/L EDTA pH 7.5, 0.83 g/L $FeCl_3{\cdot}6H_2O$, 84 mg/L $ZnCl_2$, 13 mg/L $CuCl_2{\cdot}2H_2O$, 10 mg/L $CoCl_2{\cdot}6H_2O$, 10 mg/L H_3BO_3, and 1.6 mg/L $MnCl_2{\cdot}6H_2O$.
6. Minerals and vitamins (1× solutions in D_2O): 1 M $MgSO_4$, 1 M $CaCl_2$, 1 mg/L thiamin•HCl, and 1 mg/L biotin.
7. 1 M isopropyl β-D-1-thiogalactopyranoside (IPTG).
8. 50 mg/mL kanamycin.

2.2 Purification of u-[^{2}H,^{13}C,^{15}N, 100% H^{N}]-L29

1. Buffer "0": 50 mM HEPES/KOH pH 7.5, 500 mM NaCl, 2 mM dithiothreitol (DTT), 10% glycerol, and 20 mM imidazole.
2. Wash buffer: 50 mM HEPES/KOH pH 7.5, 500 mM NaCl, 2 mM DTT, 10% glycerol, and 50 mM imidazole.
3. Elution buffer: 50 mM HEPES/KOH pH 7.5, 100 mM NaCl, 2 mM DTT, and human rhinovirus 3C protease (HRV3C protease, 2.5 mg for one purification, cleaves off the His-tag to elute the protein from the Ni-NTA beads).
4. French press.
5. Ultracentrifuge.
6. HisPur™ Ni-NTA Superflow Agarose.
7. Gravity flow column to be filled with the Ni-NTA beads.
8. Amicon Ultra 4 mL with a membrane NMWL of 3 kDa.

2.3 Purification of 50SΔL29 Ribosome Subunits

1. Cells lacking the L29 gene [28].
2. LB medium.
3. Ribosome buffer (*see* **Note 2**): 20 mM HEPES/KOH pH 7.5, 6 mM $Mg(OAc)_2$, 30 mM NH_4Cl, and 4 mM β-mercaptoethanol (β-ME).
4. Dissociation buffer (*see* **Note 3**): 20 mM HEPES/KOH pH 7.5, 1 mM $Mg(OAc)_2$, 200 mM NH_4Cl, and 4 mM β-ME.
5. 25.7% sucrose in ribosome buffer.
6. 30% sucrose (w/v) in dissociation buffer.
7. 10% sucrose (w/v) in dissociation buffer.
8. French press.
9. Ultracentrifuge.
10. Gradient station (Biocomp Instruments).

2.4 RNC-ΔL29 Purification

1. Plasmid containing the TnaC stalling sequence with the R23F mutation (*see* **Note 4**) and an antibody recognition sequence (HA-tag) for Western blot detection.
2. ΔL29 *E. coli* cells.
3. Buffer A: 50 mM HEPES/KOH pH 7.5, 250 mM K(OAc), 25 mM $Mg(OAc)_2$, 1 mM Trp, and 1:1000 (v/v) of protease inhibitor stock solution (cOmplete Protease Inhibitor Cocktail EDTA-free (CPIC), 1 tablet dissolved in 1 mL of H_2O for stock).
4. Sucrose cushion: 40% sucrose in buffer A.
5. Buffer B: 50 mM HEPES/KOH pH 7.5, 250 mM K(OAc), 25 mM $Mg(OAc)_2$, 250 mM sucrose, and 1:1000 CPIC.
6. Buffer C: 50 mM HEPES/KOH pH 7.5, 500 mM K(OAc), 25 mM $Mg(OAc)_2$, 250 mM sucrose, and 1:1000 CPIC.
7. Buffer D: 50 mM HEPES/KOH pH 7.5, 250 mM K(OAc), 25 mM $Mg(OAc)_2$, 250 mM sucrose, 20 mM imidazole, and 1:1000 CPIC.
8. Elution buffer: 50 mM HEPES/KOH pH 7.5, 250 mM K(OAc), 25 mM $Mg(OAc)_2$, 250 mM sucrose, 150 mM imidazole, and 1:1000 CPIC.
9. Chloramphenicol 34 mg/mL.
10. Ultracentrifuge.
11. French press.
12. Talon (cobalt) affinity beads.
13. Gravity flow columns (4×).

2.5 Reconstitution

1. Reconstitution buffer: 10 mM HEPES/KOH pH 7.5, 60 mM NH_4Cl, 10 mM $Mg(OAc)_2$, and 2 mM DTT.
2. 40% sucrose in the reconstitution buffer.
3. Ribosome buffer with $MgCl_2$ instead of $(MgOAc)_2$.
4. Benchtop incubator.
5. Benchtop ultracentrifuge.
6. Bis-Tris 4–12% gradient protein polyacrylamide gel and MES SDS running buffer solution (50 mM MES, 50 mM Tris base, 1 mM EDTA, 0.1% (w/v) SDS, pH 7.3).

2.6 NMR Spectroscopy

1. Solution NMR tube.
2. NMR spectrometer (600 MHz or above) equipped with a solution-state HCN cryoprobe.
3. 1.3 mm rotor: MAS BL1.3 rotor ZrO_2 with caps.
4. 800 MHz Bruker spectrometer equipped with a solid-state 1.3 mm HCN probe.

3 Methods

3.1 Production of Isotopically Labeled L29

Perform the following steps at room temperature unless specified otherwise:

1. To prepare the deuterated growth medium, mix the 100 mL of D_2O-M9 with the 10 mL of trace elements, the fully deuterated ^{13}C-labeled glucose, the ^{15}N-labeled ammonium chloride, and 1 mL, 300 μL, 1.5 mL, and 1.5 mL of the $MgSO_4$, $CaCl_2$, thiamin, and biotin stock solutions, respectively. Complete to 1 L with D_2O.
2. Transfer colonies from the LB plate to 25 mL of the deuterated growth medium supplemented by 25 μL of kanamycin. Grow them overnight (O/N) with vigorous shaking in an incubator at 37 °C (*see* **Note 5**).
3. Inoculate the remaining 975 mL of deuterated growth medium supplemented with 975 μL of kanamycin with the 25 mL of the overnight preculture (*see* **Note 6**). Grow the cells at 37 °C in an incubator shaking at 130 rpm. Once the cell density reaches OD_{600} ~ 0.6 (*see* **Note 6**), induce protein expression by adding 1 mL of 1 M IPTG, and let expression occur O/N at 16 °C in an incubator shaking at 130 rpm.

3.2 Purification of Isotopically Labeled L29

Carry out all steps at 4 °C (on ice or in the cold room) unless specified otherwise. At each step of the purification, take out 30 μL of sample and add 10 μL of 4× SDS sample buffer to monitor the purification procedure on a SDS polyacrylamide gel.

1. Harvest the cells by centrifugation at 4,410 × *g* for 10 min at 4 °C (using a Sorvall SLC6000 rotor) and resuspend the cell pellets in about 10 mL total of buffer "0." Pass them through a small needle (~1 mm ∅) using a 10 mL syringe, and lyse them with the French press (dilute them and rinse the press with a bit of buffer "0"). To obtain a clear lysate, centrifuge the lysate for 30 min at 4 °C at 185,000 × *g* (using a Beckmann Ti45 rotor). Discard the pellet and keep the supernatant, which contains L29 (*see* **Note 7**).
2. Prewash 5 mL slurry of HisPur™ Ni-NTA (corresponding to 2.5 mL of Ni beads) by adding 20 mL of buffer "0," centrifugating at 55 × *g* for 2 min, discarding the supernatant, and repeating this two times. Keep the beads.
3. Add the clarified lysate to the prewashed beads and allow binding for 30 min with rotation.
4. Transfer the lysate and the beads into an empty gravity flow column and collect the flow-through. Wash the beads with 50 mL of wash buffer.
5. Close the bottom of the column, add 10 mL of elution buffer, close the top of the column, and allow cutting of the His-tag by the 3C protease to release L29 for 45 min while rotating.
6. Finally, collect the eluted tag-free L29 and concentrate it down to ~1 mL using the Amicon concentrator. Freeze the sample in liquid nitrogen and store it at −80 °C.
7. Verify the purification on a SDS polyacrylamide gel (Fig. 2).

3.3 Purification of Ribosomal 50S-ΔL29 Subunits

Carry out all steps at 4 °C (on ice or in the cold room) unless specified otherwise.

1. Grow the ΔL29 cells in 2 L of LB at 37 °C in an incubator shaking at 130 rpm for about 4 h. Harvest the cells by centrifugation at 4,410 × *g* for 10 min at 4 °C (using a Sorvall SLC6000 rotor) and resuspend the pellets in ribosome buffer. Pass them through a small needle using a syringe, and lyse them with the French press. To remove cell debris, centrifuge the lysate for 20 min at 4 °C at 30,600 × *g* (using a Beckmann SS34 rotor).
2. Load the clarified lysate on the 25.7% sucrose cushion in ribosome buffer: in each Ti45 centrifuge tube, put 20 mL of cushion, and then fill up the tube with lysate using a 25 mL pipette touching the wall of the tube so as not to disturb the sucrose cushion. Centrifuge O/N at 4 °C at 72,500 × *g* to pellet the ribosomes.
3. Resuspend the ribosomal pellets in ca. 1 mL of ribosome dissociation buffer each by gently shaking at 4 °C. Clarify then the resuspended ribosomes by centrifugation for 5 min at 4 °C at 8,600 × *g* in Eppendorf tubes using a tabletop centrifuge.

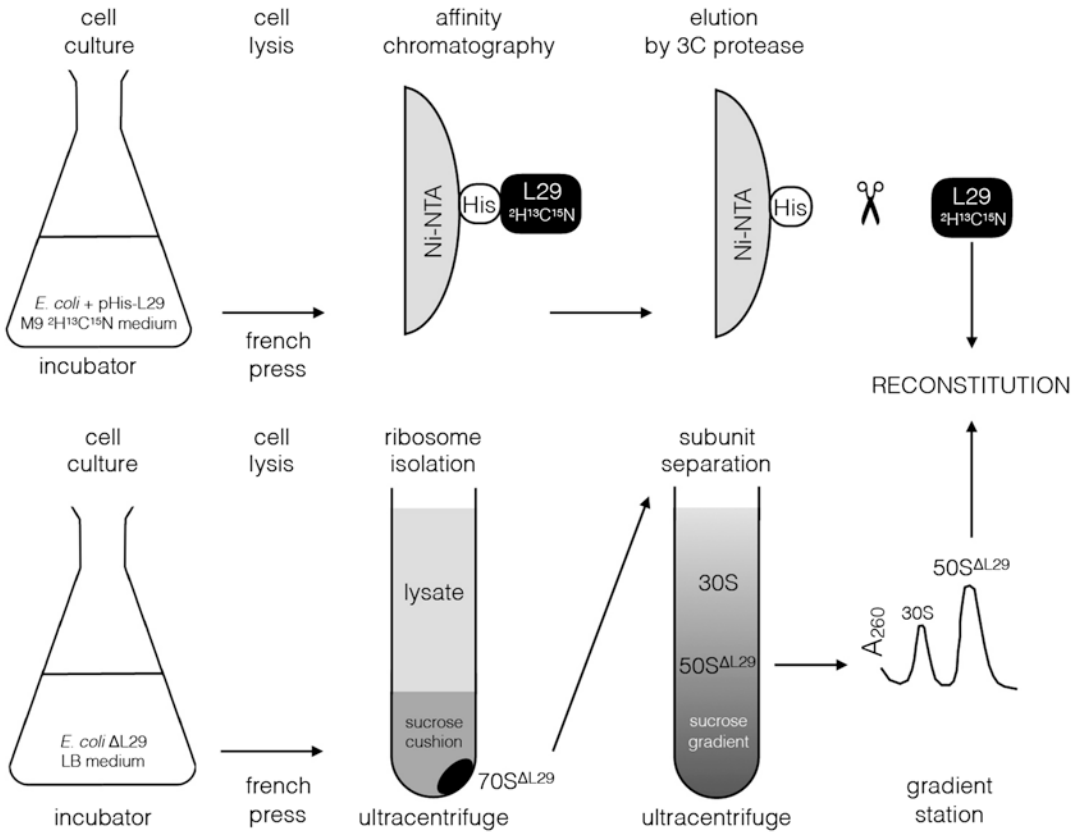


Fig. 2 Schematic representation of the two purification procedures that jointly resulted in the 50S ribosomal subunits with triply labeled L29

4. Measure the absorbance at 260 nm (A_{260}) to estimate the ribosome concentration (*see* **Note 8**).
5. Using the gradient station, prepare six 10–30% sucrose gradients in ribosome dissociation buffer: start by marking the gradient tubes in the middle according to your station and the type of caps you use (long or short); fill the tubes up to the line (a bit more) with the 10% sucrose in dissociation buffer; use a needle to bring the 30% sucrose in dissociation buffer to the bottom, and slowly fill up the bottom of the tube up to the mark while keeping the needle tip at the interface between the two sucrose solutions. Put the caps on making sure there are no air bubbles, and follow the instructions on your gradient station to make the gradients. Remove the caps carefully without disturbing the gradients.
6. On top of each gradient, load the equivalent of A_{260} = 100 of clarified ribosomes (*see* **Note 8**). Centrifuge the gradients for 17 h at 4 °C at 44,300 × *g* (using a Beckmann SW32 rotor).

7. Using the gradient station, fractionate your samples and collect the 30S and 50S peaks separately. Measure the A_{260} to estimate the 50S yield (*see* **Note 9**).
8. To pellet the 50SΔL29 ribosome subunits, dilute the sucrose at least 3× with ribosome buffer (without sucrose). Centrifuge O/N at 4 °C at 72,500 × *g* (using a Beckmann Ti45 rotor).
9. Resuspend the pellets in about 400 µL of ribosome buffer each by gently shaking at 4 °C. Measure the final A_{260}. Freeze the sample in liquid nitrogen and store it at −80 °C.

3.4 RNC-ΔL29 Purification

RNC purification is performed essentially as described previously [29]. Carry out all steps at 4 °C (on ice or in the cold room) unless specified otherwise. At each step of the purification, take out 30 µL of sample and add 10 µL of SDS lysis buffer 4× to verify the success of the purification procedure on a Western blot.

1. Transform the plasmid into *E. coli* MC4100 ΔL29 cells.
2. Grow the cells in 8 L of LB at 37 °C in an incubator shaking at 130 rpm until they reach OD_{600}~0.5, and induce expression by adding 0.2% of L-arabinose. After 1 h, add 1:1000 of the chloramphenicol stock solution to the cells to stop translation, harvest the cells by centrifugation at 4,410 × *g* for 10 min at 4 °C (using a Sorvall SLC6000 rotor), and resuspend the pellets in buffer A. Pass them through a small needle using a syringe, and lyse them with the French press. To remove cell debris, centrifuge the lysis product for 20 min at 4 °C at 30,600 × *g* (using a Beckmann SS34 rotor).
3. Load the clear lysate on the 40% sucrose cushion in buffer A: in each Ti45 centrifuge tube, put 25 mL of cushion, and then fill up the tube with lysate using a 25 mL pipette touching the wall of the tube so as not to disturb the sucrose cushion. Centrifuge O/N (~20 h) at 4 °C at 72,500 × *g* to pellet the ribosomes.
4. Resuspend the glassy pellets in ~10–20 mL total of buffer B (*see* **Note 10**).
5. Prewash 4 × 5 mL of Talon slurry (=2.5 mL beads, each in a 50 mL Falcon tube) with buffer B and equilibrate the beads with 0.1 mg/mL *E. coli* total tRNA mixture (from strain MRE600, Sigma) in buffer B for about 10 min rotating at 4 °C.
6. Divide up the resuspended pellets into four equal portions and add them to the four Falcon tubes containing the beads. Fill each tube with buffer B up to 25 mL. Allow binding by rotating for 1 h at 4 °C.
7. Transfer the content of each Falcon tube into separate empty gravity flow columns and collect the flow-through. Wash the

beads with 25 mL of buffer B each, followed by 25 mL of buffer C and 15 mL of buffer D each.

8. Elute the RNCs with 10 mL of elution buffer per column by rotating for 10 min at 4 °C.
9. Collect the elution and load it on top of 20 mL of 40% sucrose cushion in buffer A without Trp in one Ti45 tube. Centrifuge O/N (~20 h) at 4 °C at 72,500 × *g* to pellet the RNCs.
10. Resuspend the RNC pellet in buffer A without Trp to measure the OD (A_{260}) and quantify the yield (*see* **Note 9**), and then add Trp to a final concentration of 1 mM. Freeze the sample in liquid nitrogen and store it at −80 °C.
11. Perform a Western blot to verify the purification and that the RNCs are present in the sample.

3.5 Reconstitution

1. Mix 40 OD of 50SΔL29 or 41 OD of RNC-ΔL29 with 4–5 nmol or 3–4 nmol of labeled L29, respectively (*see* **Note 11**), and add reconstitution buffer to a total reaction volume of 2 mL (*see* **Note 12**). Make samples for the gel of both components of the reaction.
2. Allow the reconstitution to occur for 45 min at 25 °C in a benchtop incubator followed by O/N at 4 °C. Take a sample for the gel at the end of the reaction.
3. In a TLA110 tube, fill the bottom with 500 μL of the 40% sucrose cushion in reconstitution buffer. Then carefully add the reconstitution reaction on top of the cushion. To pellet the reconstituted 50S subunits, centrifuge for 1 h at 4 °C at 417,200 × *g* (in a Beckmann TLA110 rotor).
4. After centrifugation, remove the supernatant of the cushion, take a sample for the gel, and then freeze it in liquid nitrogen and store it at −80 °C. It contains the excess labeled L29 that can be used again after proper buffer exchange.
5. Discard the sucrose cushion and resuspend the reconstituted 50S in ca. 1 mL of ribosome buffer (*see* **Note 13**). Take a final sample for the gel, then freeze the final reconstituted sample in liquid nitrogen, and store it at −80 °C.
6. To verify the reconsitution, load the following samples on a Bis-Tris 4–12% gradient protein gel (polyacrylamide gel for good resolution in the low range of protein molecular weights) (Fig. 3). Tot, end reaction sample corresponding to 0.2 OD 50S; Sup, supernatant of the cushion (excess of L29); Pel, 0.2 OD from the pellet of the cushion (corresponding to the reconstituted 50S); 50SΔ, 0.2 OD of 50S-ΔL29; L29, L29 from the purified concentrated sample. Comparing Pel and 50SΔ, a band corresponding to L29 can be recognized (in Pel but not in 50SΔ). Tot should be the sum of Pel and 50SΔ.

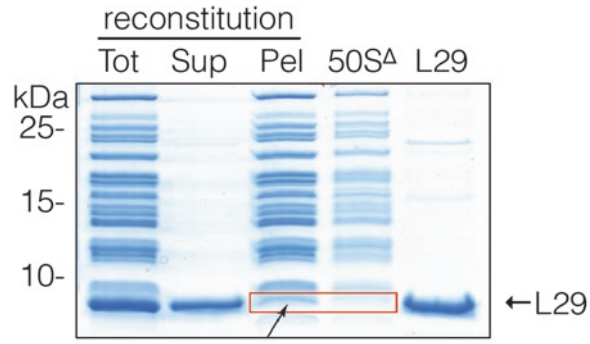


Fig. 3 Bis-Tris 4–12% gradient protein gel to monitor the reconstitution of labeled L29 into 50SΔL29 ribosome subunits. Refer to the text for a detailed explanation for each lane

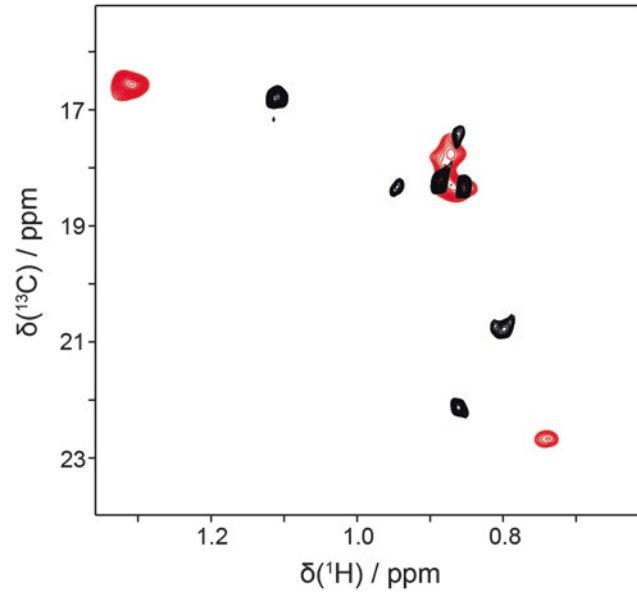


Fig. 4 Solution-state ^{1}H,^{13}C correlation spectra of selectively valine methyl protonated L29 (black) and L29 reconstituted in a protonated large ribosomal 50S subunit (red)

3.6 Solution-State NMR Experiments of the Reconstituted Sample

To test whether the reconstitution was successful, we prepared selectively valine methyl protonated L29. L29 contains four valine residues in total. Seven out of eight methyl groups are readily detectable. After reconstitution within the 50S subunit, resonances are in general broadened but still well observable (Fig. 4). For two resonances, we observe major chemical shift differences which are due to a change of the chemical environment in the context of the 50S subunit.

3.7 Solid-State NMR of the Reconstituted Sample

1. To avoid degradation of the sample, the rotor was filled immediately prior to the NMR measurement. To fill the 1.3 mm rotor, put an FKM (fluorocarbon material) insert at the bot-

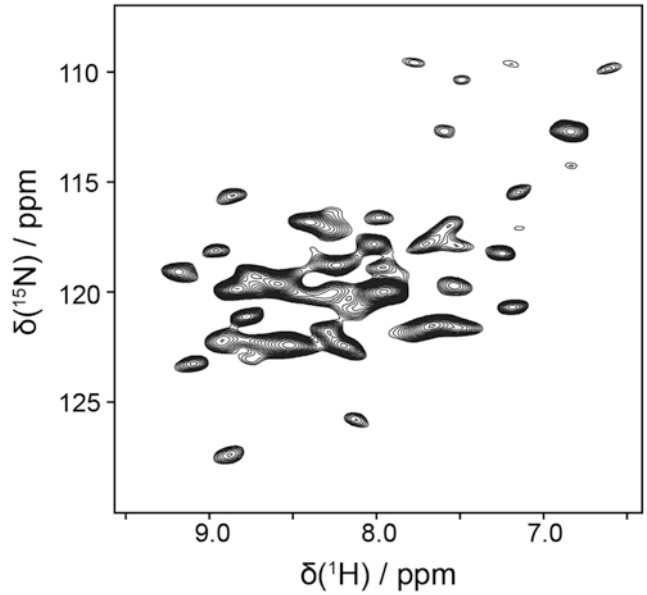


Fig. 5 MAS solid-state $^{1}H,^{15}N$ correlation spectrum of u-[$^{2}H,^{15}N$] labeled L29, reconstituted in a protonated large ribosomal 50S subunit. The experiment was performed using a 800 MHz Bruker spectrometer at a MAS rotation frequency of 60 kHz. The temperature was adjusted to yield an actual sample temperature of 10 °C. L29 contains 63 residues (MW: 7.07 kDa). Approximately 40 resolved cross peaks can be observed in the $^{1}H,^{15}N$ correlation spectrum

tom and the bottom cap. Place the rotor in a filling tool (*see* **Note 14**) and then the sample. Centrifuge O/N at 4 °C at 96,000 × *g* to pellet the 50S subunits into the rotor. After the centrifugation, take out the supernatant, put an insert on top of the sample, and close the rotor with the drive cap.

2. Perform the desired solid-state experiments at 60 kHz magic-angle spinning (MAS). As an example, Fig. 5 shows the solid-state $^{1}H,^{15}N$ correlation spectrum of triply labeled L29 reconstituted in unlabeled 50S measured at 60 kHz MAS at 10 °C on a 800 MHz spectrometer.

4 Notes

1. Triply labeled recombinant proteins can be cumbersome to produce. It may be required to slowly adapt the cells to the D_2O environment by growing them in protonated minimal medium first and then gradually increase the amount of D_2O in the medium (e.g., 50, 75, 90, 100%). In this case, grow cells in 5 mL of minimal medium containing low concentrations of D_2O, wait until they grow (reach an $OD_{600} > 1$), then inoculate another 5 mL of minimal medium containing more D_2O

from the first test, and continue like this. Since there is no effect from the labeled glucose or ammonium chloride, use unlabeled components for these trials. Once the cells are adapted and grow in D_2O, plate them to later inoculate the fully labeled medium. Note that some types of *E. coli* cells may never adapt to D_2O.

2. The composition of ribosome buffers is extremely important with regard to their stability. The pH of the buffers should be 7.5, and most importantly, the ratio monovalent/divalent cations is crucial to avoid dissociation of the two subunits. This ratio should be between 6 and 10. For the divalent cation, use Mg^{2+}, and for monovalent cation, use K^+ or NH_4^+. It is recommended to use acetate as a counterion.
3. The composition of ribosome buffers plays a huge role in stability. To dissociate 70S ribosomes into individual subunits, the dissociation buffer needs to meet two pretty harsh conditions. First, the ratio monovalent/divalent cations has to be much higher than the normal 6–10:1 ratio (around 200:1). Second, there must be a very high concentration of Cl^- anions, so use KCl or NH_4Cl but by no means K(OAc) or NH_4(OAc) to make the dissociation buffer.
4. The R23F mutant of the TnaC stalling sequence is much more efficient for stalling and gives a much higher yield of RNCs in the ΔL29 cells than the wild-type TnaC sequence (threefold higher yield, EBM et al., data unpublished).
5. In the absence of a LB plate with deuterium-adapted cells, start by growing them from a glycerol stock in 5 mL of LB + kan during the day. At the end of the day, pellet the cells to discard the LB unlabeled medium and use them for an O/N culture in deuterated minimal medium.
6. To inoculate the triply labeled minimal medium, an O/N culture that reached an OD_{600} of 1.6 or more is required. If they are not at this value, do not start the 1 L culture but wait until they are. Once the 1 L culture is inoculated, the growth will be much slower than in LB medium, so it may take 6–7 h until they reach the OD at which protein expression can be induced. Monitor the growth by regularly measuring OD_{600}.
7. This procedure is only valid if the recombinant protein ends up in the cytoplasm and not in inclusion bodies or in the membrane. It is implicit that the purification strategy is tested with protein produced in LB. What is important to note here is that all steps of production should be conducted in deuterated medium, but all steps of purifications should be made in water-based buffers to allow for the H-D exchange of amide and other exchangeable sites of the protein.
8. It is important to quantify the ribosomes at this point in order not to overload the gradients, in which case the resolution

would not be good enough to separate the 30S from the 50S subunit.

9. The absorbance at 260 nm gives the concentration of the 50S subunits according to 1OD = 36 pmol and of the 70S according to 1OD = 24 pmol.
10. To resuspend the RNC pellets, add 1–2 mL of buffer and shake for a while, then take out the resuspended RNC, and add fresh buffer again to continue the resuspension, until the pellet is fully resuspended. Resuspension is more efficient in small volumes.
11. 1.3 mm rotors for solid-state NMR with inserts can accommodate up to 2 mg of sample. This corresponds to 1.3 nmol of 50S ribosomal subunit, or 36 OD, and to 900 pmol of 70S, or 37.5 OD. To fill the rotor knowing that there are losses during the centrifugation steps, use 40 OD of 50S sample or 41 OD of 70S sample. Then add L29 in ca. threefold molar excess.
12. The total reaction volume depends on the available ultracentrifuge tubes (thin or thick walls). Make sure that this volume plus 500 μL does not exceed the maximum volume recommended for one tube at 417,200 × *g*. Preferably, use only one tube since resuspending always leads to small losses of sample.
13. Ribosome pellets should always be resuspended in small amounts of buffer, but the final volume depends on the rotor filling device. Simply avoid diluting them too much if possible.
14. Rotor filling tools are extremely useful to avoid losing precious sample. They are available at Giotto Biotech©, or they can be designed in house.

Acknowledgment

We acknowledge support from the Helmholtz-Gemeinschaft and the Deutsche Forschungsgemeinschaft (Grants Re1435 and SFB-1035, project B07). In addition, we are grateful to the Center for Integrated Protein Science Munich (CIPS-M) for the financial support. We acknowledge support from EMBO (Fellowship ALTF 52-2014) and from the European Commission (EMBOCOFUND2012, GA-2012-600394) and Marie Curie Actions.

References

1. Cheng YF (2015) Single-particle Cryo-EM at crystallographic resolution. Cell 161:450–457
2. Frank J (2017) Advances in the field of single-particle cryo-electron microscopy over the last decade. Nat Protoc 12:209–212
3. Orlov I, Myasnikov AG, Andronov L et al (2017) The integrative role of cryo electron microscopy in molecular and cellular structural biology. Biol Cell 109:81–93
4. Callaway E (2015) The revolution will not be crystallized. Nature 525:172–174

5. Mainz A, Jehle S, van Rossum BJ et al (2009) Large protein complexes with extreme rotational correlation times investigated in solution by magic-angle-spinning NMR spectroscopy. J Am Chem Soc 131:15968–15969
6. Mainz A, Bardiaux B, Kuppler F et al (2012) Structural and mechanistic implications of metal-binding in the small heat-shock protein αB-crystallin. J Biol Chem 287:1128–1138
7. Mainz A, Religa T, Sprangers R et al (2013) NMR spectroscopy of soluble protein complexes at one mega-Dalton and beyond. Angew Chem Int Ed Engl 52:8746–8751
8. Mainz A, Peschek J, Stavropoulou M et al (2015) The chaperone αB-crystallin deploys different interfaces to capture an amorphous and an amyloid client. Nat Struct Mol Biol 22:898–905
9. Barbet-Massin E, Huang C-T, Daebel V et al (2015) Site-specific solid-state NMR studies of "trigger factor" in complex with the large ribosomal subunit 50S. Angew Chem Int Ed Engl 54:4367–4369
10. Sarkar R, Mainz A, Busi B et al (2016) Immobilization of soluble protein complexes in MAS solid-state NMR: sedimentation versus viscosity. Solid State Nucl Magn Reson 76-77:7–14
11. Quinn CM, Polenova T (2017) Structural biology of supramolecular assemblies by magic-angle spinning NMR spectroscopy. Q Rev Biophys 50:1–44
12. Petkova AT, Yau W-M, Tycko R (2006) Experimental constraints on quaternary structure in Alzheimer's β-amyloid fibrils. Biochemistry 45:498–512
13. Wasmer C, Lange A, Van Melckebeke H et al (2008) Amyloid fibrils of the HET-s(218-289) prion form a beta solenoid with a triangular hydrophobic core. Science 319:1523–1526
14. Tuttle MD, Comellas G, Nieuwkoop AJ et al (2016) Solid-state NMR structure of a pathogenic fibril of full-length human alpha-synuclein. Nat Struct Mol Biol 23:409–415
15. Colvin MT, Silvers R, Ni QZ et al (2016) Atomic resolution structure of monomorphic a beta(42) amyloid fibrils. J Am Chem Soc 138:9663–9674
16. Wälti MA, Ravotti F, Arai H et al (2016) Atomic-resolution structure of a disease-relevant Aβ(1-42) amyloid fibril. Proc Natl Acad Sci U S A 113:E4976–E4984
17. Lange A, Giller K, Hornig S et al (2006) Toxin-induced conformational changes in a potassium channel revealed by solid-state NMR. Nature 440:959–962
18. Shahid SA, Bardiaux B, Franks WT et al (2012) Membrane-protein structure determination by solid-state NMR spectroscopy of microcrystals. Nat Methods 9:1212–U1119
19. Lu MM, Hou GJ, Zhang HL et al (2015) Dynamic allostery governs cyclophilin A-HIV capsid interplay. Proc Natl Acad Sci U S A 112:14617–14622
20. Andreas LB, Jaudzems K, Stanek J et al (2016) Structure of fully protonated proteins by proton-detected magic-angle spinning NMR. Proc Natl Acad Sci U S A 113:9187–9192
21. Yan S, Guo CM, Hou GJ et al (2015) Atomic-resolution structure of the CAP-Gly domain of dynactin on polymeric microtubules determined by magic angle spinning NMR spectroscopy. Proc Natl Acad Sci U S A 112:14611–14616
22. Yehl J, Kudryashova E, Reisler E et al (2017) Structural analysis of human Cofilin 2/filamentous actin assemblies: atomic-resolution insights from magic angle spinning NMR spectroscopy. Sci Rep 7:44506
23. Grutsch S, Bruschweiler S, Tollinger M (2016) NMR methods to study dynamic allostery. PLoS Comput Biol 12:e1004620
24. Olsson S, Strotz D, Vogeli B et al (2016) The dynamic basis for signal propagation in human Pin1-WW. Structure 24:1464–1475
25. Chevelkov V, Fink U, Reif B (2009) Quantitative analysis of backbone motion in proteins using MAS solid-state NMR spectroscopy. J Biomol NMR 45:197–206
26. Schanda P, Meier BH, Ernst M (2010) Quantitative analysis of protein backbone dynamics in microcrystalline ubiquitin by solid-state NMR spectroscopy. J Am Chem Soc 132:15957–15967
27. Kramer G, Boehringer D, Ban N et al (2009) The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. Nat Struct Mol Biol 16:589–597
28. Kramer G, Rauch T, Rist W et al (2002) L23 protein functions as a chaperone docking site on the ribosome. Nature 419:171–174
29. Bischoff L, Wickles S, Berninghausen O et al (2014) Visualization of a polytopic membrane protein during SecY-mediated membrane insertion. Nat Commun 5:4103

Chapter 7

Characterizing Protein-Protein Interactions Using Deep Sequencing Coupled to Yeast Surface Display

Angelica V. Medina-Cucurella and Timothy A. Whitehead

Abstract

In this chapter, we discuss a method to determine the affinity and specificity of nearly all single-point mutants for a full-length protein binder. This method combines deep sequencing, comprehensive mutagenesis, yeast surface display, and fluorescence-activated cell sorting. This approach has been used to study sequence-function relationships for protein-protein interactions. The data can be used to determine the fine conformational epitope on the protein binder.

Key words Deep sequencing, Yeast surface display, Nicking mutagenesis, FACS, Conformational epitope mapping

1 Introduction

Delineating the sequence determinants of stability, affinity, and specificity of protein-protein interactions (PPI) has been a major research goal for decades. The classical approach to study PPI sequence-function relationships has been "alanine-scanning" mutagenesis in which residues are individually mutated to alanine and assayed by assorted biophysical techniques [1, 2]. The change of binding affinity upon mutation gives a reasonable measure of the importance of the perturbed residue. However, such classical mutagenesis and screening studies are extremely labor-intensive.

Recently, transformative methods utilizing large-scale mutagenesis, surface display, and next-generation sequencing (NGS) have been developed to obtain relative binding contributions of individual residues by testing thousands of PPI mutants in a single experiment [3–8]. All methods share a similar framework: a population containing mutants of one PPI partner is prepared and cloned into a surface-display vector. The population is selected and/or screened for positive or negative binding to the other partner, and then the selected and unselected populations are deep sequenced and analyzed. Finally, the change in frequency for each

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_7,

library member is calculated and converted to a relative binding score [8]. In the method developed by our lab [7], we utilize yeast surface display (YSD) [9, 10] as it affords quantitative screening via fluorescence-activated cell sorting (FACS) [5, 6, 11]. Compared with competing methods using YSD [5, 6, 12], our approach is faster and less expensive albeit with a limited dynamic range of approximately tenfold change in binding affinity centered about the wild-type sequence. Accordingly, our method is suitable for fine maturation of PPI affinity and specificity or to determine fine conformational epitopes.

In this chapter, we provide a detailed protocol to determine relative binding affinities and conformational epitope maps for PPIs (overview in Fig. 1). We cover creation of single-site saturation mutagenesis (SSM) libraries using nicking mutagenesis [13], transformation of libraries into yeast by the method of Gietz and Woods [14], screening of the SSM YSD library using FACS, DNA preparation for sequencing on an Illumina platform, and data analysis to determine a relative binding score and conformational epitope map. Relative binding calculations and estimated errors are carried out according to methods described in Kowalsky et al. [7]. Note: we assume the end user has (1) one PPI partner successfully induced and displayed in a YSD format with the other partner biotinylated and with (2) reproducible measurement of the apparent dissociation constant using protocols as described in Chao et al. [9].

2 Materials

2.1 Yeast and Bacteria Strains and Plasmid

1. **Yeast strain**: *Saccharomyces cerevisiae* strain EBY100 is available at American Type Culture Collection and prepared to be chemically competent according to Gietz and Woods [14] (*see* a shortened protocol in Subheading 3.1.4).
2. **Bacterial strain**: *Escherichia coli* strain XL1-Blue high-efficiency electrocompetent cells are available through Agilent Technologies. Other competent cells with at least 1×10^9 transformants per μg of plasmid can be used.
3. **Yeast display vectors**: The YSD vector used, pETconNK, is freely available on Addgene (plasmid #81169) [15]. The gene of interest is inserted between NdeI and XhoI restriction sites.

2.2 Nicking Site Saturation Mutagenesis (SSM) Library Preparation

All enzymes and buffers for SSM library preparation with nicking mutagenesis are from New England Biolabs Inc. (NEB) unless noted otherwise.

2.2.1 Reagents, Media, and Plates

1. pETconNK plasmid containing gene of interest (freshly prepared from a dam^+ bacterial strain).

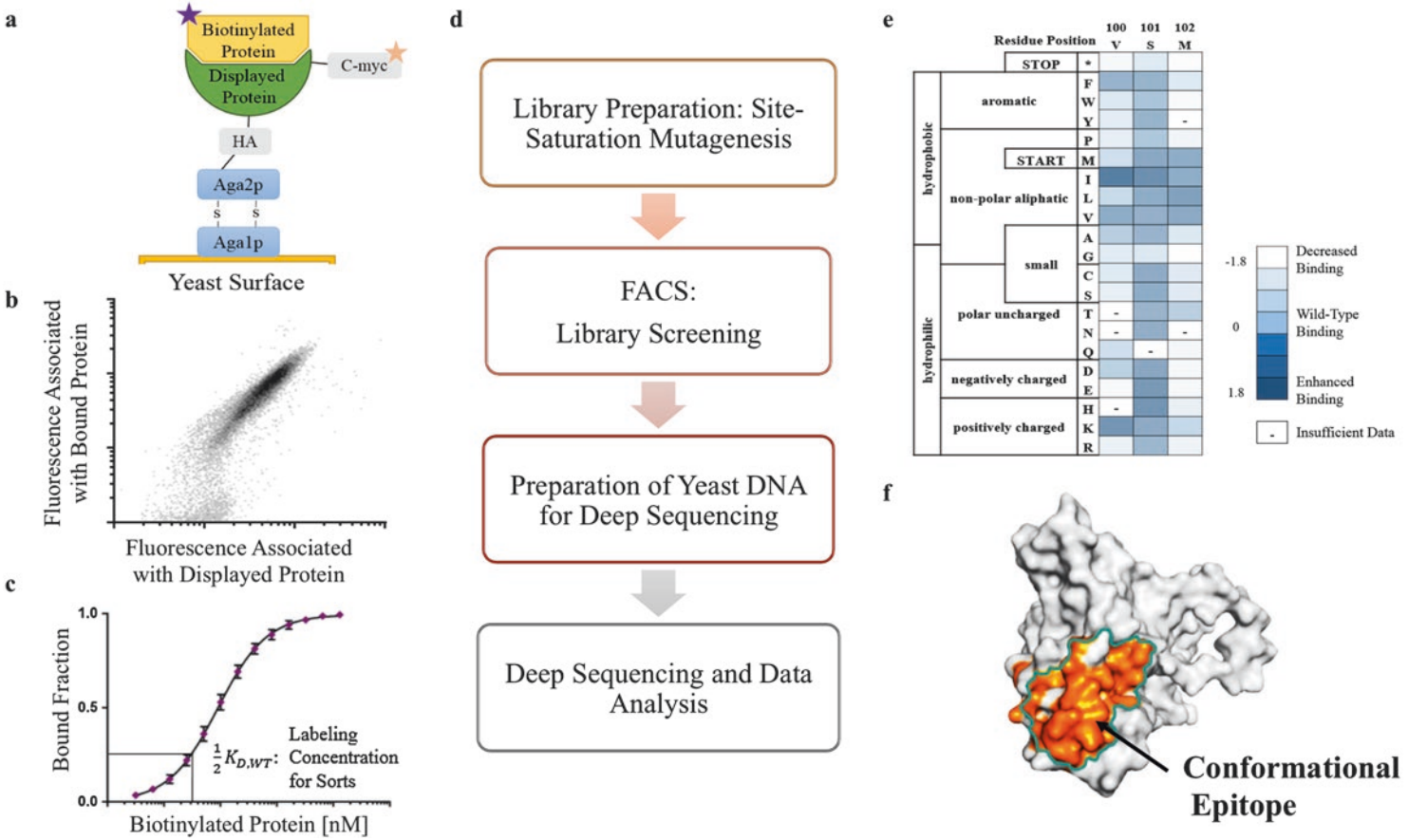


Fig. 1 A streamlined process required for PPI characterization using deep sequencing and mutagenesis analysis. (**a–c**) Requirements for the pipeline: the binding activity of two proteins is measured using yeast surface display coupled to flow cytometry, and the relative dissociation constant is determined using yeast clonal titrations. The top panel is adapted from Chao et al. [9]. (**d**) The workflow covered in this chapter to characterize protein-protein interactions. (**e, f**) Deep sequencing results can be visualized as a heatmap and used to determine the conformational epitope of one member of the interaction

2. Nuclease-free water (NFH_2O, Integrated DNA Technologies).
3. Custom mutagenic primers (*see* Subheading 3.1.1).
4. SEC_Rev primer: 5′ – CAAGTCCTCTTCAGAAATAAGCTT TTGTTC – 3′.
5. T4 polynucleotide kinase buffer.
6. 10× CutSmart Buffer.
7. 5× Phusion HF Buffer.
8. 10 mM ATP.
9. 50 mM DTT.
10. 50 mM NAD^+.
11. 10 mM dNTPs.
12. 50% v/v sterile glycerol solution using deionized H_2O.
13. **TB media**: 4.76% w/v of TB powder (premixed) and 0.8% v/v of glycerol. Sterilize by autoclaving.
14. **LB agar plates**: 2.5% w/v of LB powder (premixed) and 1.5% w/v of agar. Sterilize by autoclaving.

*Add kanamycin to a final concentration of 30 μg/mL when preparing the small plates to calculate the transformation efficiency and the large bioassay dishes for SSM libraries (*see* Subheading 3.2.1).

2.2.2 Enzymes

1. 10 U/μL T4 polynucleotide kinase.
2. 10 U/μL Nt.BbvCI.
3. 10 U/μL Nb.BbvCI.
4. 100 U/μL exonuclease III.
5. 20 U/μL exonuclease I.
6. 2 U/μL Phusion High-Fidelity DNA Polymerase.
7. 40 U/μL Taq DNA ligase.
8. 20 U/μL DpnI.

*Diluent for all enzymes required for Subheading 3.2.1 is 1× NEB CutSmart Buffer.

2.2.3 Equipment and Materials

1. Zymo Clean and Concentrator-5 kit (Zymo Research).
2. Corning square bioassay dishes, 245 mm × 245 mm × 25 mm (Sigma-Aldrich).

2.3 Chemically Competent Library Yeast Transformation

2.3.1 Yeast Solutions and Plates

1. **Growth media**: Synthetic dextrose medium supplemented with casamino acids (SDCAA): 2% w/v dextrose (D-glucose), 0.67% w/v yeast nitrogen base without amino acids (Sigma-Aldrich), 0.5% w/v Bacto casamino acids technical (BD Biosciences), 0.54% w/v Na_2HPO_4, and 0.856% w/v $Na_2HPO_4{\cdot}H_2O$. Filter sterilize. Add 1% v/v of 10,000 U/mL

penicillin-streptomycin immediately prior to growth to prevent bacterial contamination.

2. **Induction media**: Synthetic galactose medium supplemented with casamino acids (SGCAA): prepare like SDCAA but with 2% w/v of galactose instead of dextrose.
3. **SDCAA agar plate**: 0.54% w/v Na_2HPO_4, 0.856% w/v $Na_2HPO_4 \cdot H_2O$, 18.2% w/v sorbitol, and 1.5% w/v agar. Sterilize by autoclaving. 2% w/v dextrose (D-glucose), 0.67% w/v yeast nitrogen base without amino acids, 0.5% w/v Bacto casamino acids technical. Sterilize by filtrating. Add the filter-sterilized solution into the cool autoclaved mix (approximately below 50 °C) at 1:10 ratio. Store for up to 6 months at 4 °C.
4. **Yeast storage buffer**: 20% w/v glycerol, 20 mM HEPES, and 150 mM NaCl pH 7.5. Filter sterilize.

2.3.2 Reagents

1. 10 mg/mL salmon sperm DNA (Invitrogen).
2. 50% w/v polyethylene glycol (PEG), filter sterilize.
3. 1 M lithium acetate, LiOAc.

2.4 Library Screening

2.4.1 Buffers and Reagents

1. Phosphate buffered saline (PBS) at pH 7.4: 0.8 w/v NaCl, 0.02% w/v KCl, 0.144% w/v Na_2HPO_4, and 0.024% w/v KH_2PO4. Sterilize by filtrating.
2. Phosphate buffered saline with bovine serum albumin (PBS-BSA) at pH 7.4: prepare as PBS and supplemented with 0.1% w/v bovine serum albumin (BSA). Sterilize by filtrating.
3. Anti-c-myc-FITC antibody, FITC (Miltenyi Biotec).
4. Streptavidin, R-phycoerythrin conjugate, SAPE (Thermo Fisher).
5. Biotinylated PPI partner protein (*see* **Note 1**).

2.5 Deep Sequencing Preparation of Yeast DNA

2.5.1 Buffers and Reagents

1. **TE media**: 10 mM Tris-HCl at pH 8.0 and 0.1 mM EDTA.
2. 5 U/μL Zymolyase (Zymo Research).
3. 10× lambda nuclease buffer (NEB).
4. SYBR Gold Nucleic Acid Gel Stain (Thermo Fisher).
5. Agencourt AMPure XP (Beckman Coulter).
6. Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies).
7. 70% v/v ethanol.

2.5.2 Enzymes

1. 5000 U/mL lambda nuclease (NEB).

2.5.3 Equipment

2. Zymo Research Yeast Plasmid Miniprep II kit.
3. Qiagen mini-prep kit.
4. 96-well magnetic plate.

3 Methods

3.1 Library Preparation: Site Saturation Mutagenesis (SSM)

Because a protein of 250 amino acids is encoded by a 750 bp gene, separate SSM libraries are prepared for the gene of interest (Fig. 2a) to allow compatibility with 250 bp paired end (PE) Illumina MiSeq sequencing reads (*see* **Note 2** for considerations for library preparation).

3.1.1 Design of Mutagenic Oligonucleotides

SSM libraries are created using degenerative oligonucleotides containing a "NNK" codon to cover all possible point mutations, where N represents any of the A/T/G/C, and K represents T/G. Mutagenic oligos are designed to be complementary to the wild-type template sequence as determined by the orientation of the BbvCI restriction site on the pETconNK vector (Fig. 2b, c; *see* **Note 3**).

1. Design your mutagenic oligos using QuikChange Primer Design Program (www.agilent.com). Use a degenerate "NNK" codon to cover all possible 20 amino acids at each codon position.
2. Order the mutagenesis oligos on a 500 pmol DNA Plate Oligo from Integrated DNA Technologies and resuspend to 10 μM in TE, pH 8.

3.1.2 Preparation of SSM Libraries by Nicking Mutagenesis

This protocol is exactly as described in Wrenbeck et al. [16]. All reactions should be prepared on ice unless otherwise stated.

1. To phosphorylate the oligos, make a mixture comprising 5 μL of each NNK mutagenic primer.
2. Into a PCR tube, add 20 μL of the 10 μM mutagenic primer mixture, 2.4 μL of T4 polynucleotide kinase buffer, 1 μL of 10 mM ATP, and 1 μL of T4 polynucleotide kinase. Incubate the reaction mixture at 37 °C for 1 h.
3. At the same time and in a separate PCR tube, add 18 μL of NFH_2O, 3 μL of T4 polynucleotide kinase buffer, 7 μL of 100 μM SEC_Rev primer, 1 μL of 10 mM ATP, and 1 μL of T4 polynucleotide kinase. Incubate the reaction mixture at 37 °C for 1 h.
4. Store phosphorylated oligos at −20 °C.
5. The day of mutagenesis, dilute phosphorylated mutagenic primers 1:1000 and SEC_Rev primer 1:20 using NFH_2O (*see* **Note 4**).
6. For the preparation of ssDNA template strand, in a PCR tube, add 0.76 pmol of dsDNA plasmid (approximately 2–3 μg), 2 μL of 10× CutSmart Buffer, 1 μL of 1:10 diluted exonuclease III (final concentration of 10 U/μL), 1 μL of Nt.BbvCI, 1 μL of exonuclease I, and NFH_2O to 20 μL final volume.

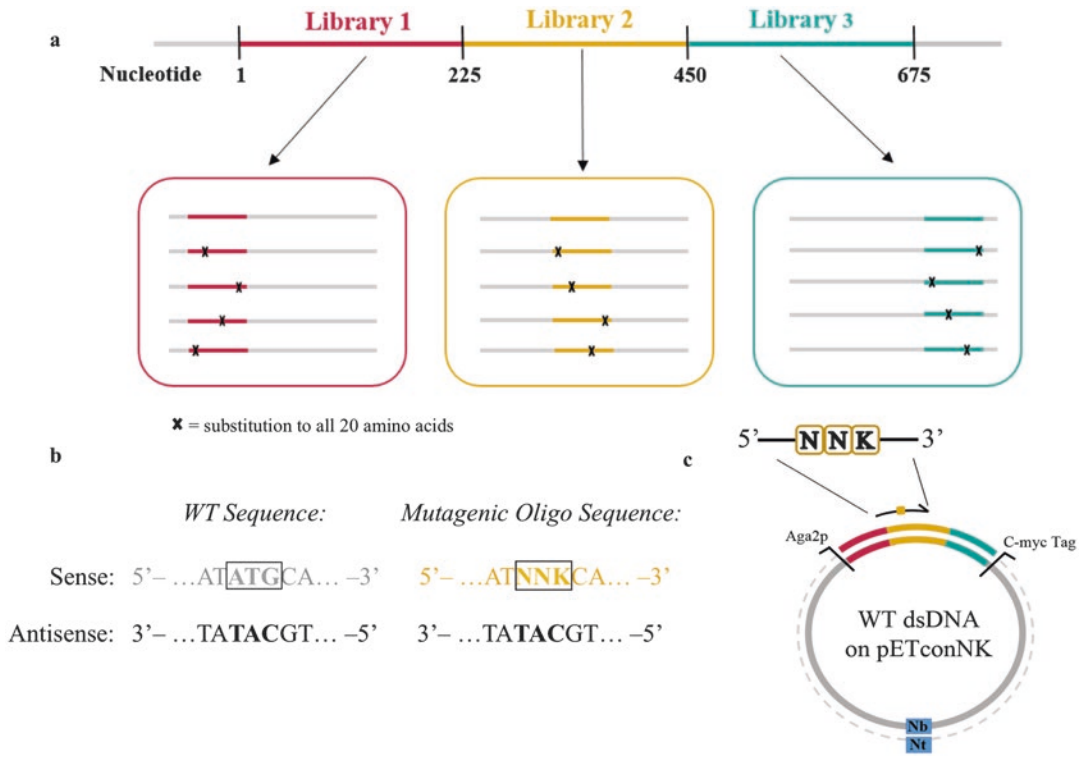


Fig. 2 Essential considerations needed for preparing site saturation mutagenesis (SSM) libraries. (**a**) The gene of interest is segmented in multiple libraries containing contiguous sections of 200–250 bp. Here, sections of 225 bp are shown for compatibility with 250 bp PE Illumina MiSeq sequencing. (**b**) Each mutagenic oligo contains an "NNK" codon to cover all possible 20 amino acids. (**c**) An Nt.BbvCI restriction enzyme (Nt) is used to create a nick on the sense strand. Mutagenic oligos are designed to be complementary to the antisense ssDNA template

7. Place the tube in a preheated (37 °C) thermal cycle with the following program: 60 min at 37 °C, 20 min at 80 °C, and hold at 4 °C.
8. To proceed with the comprehensive codon mutagenesis on the first strand, in each PCR tube, add 26.7 μL NFH_2O, 20 μL of 5× Phusion HF Buffer, 4.3 μL 1:1000 diluted phosphorylated mutagenic oligos, 20 μL of 50 mM DTT, 1 μL of 50 mM NAD^+, 2 μL of 10 mM dNTPs, 5 μL of *Taq* DNA ligase, and 1 μL of Phusion High-Fidelity DNA Polymerase. Mix the tube content briefly.
9. Place the tube into a preheated (98 °C) thermal cycler with the following program: 2 min at 98 °C, 15× cycles of 30 s at 98 °C, 45 s at 55 °C and 7 min at 72 °C, followed by a final incubation at 45 °C for 20 min, and hold at 4 °C. Add additional 4.3 μL oligo at the beginning of cycles 6 and 11.

10. Purify each reaction using a Zymo Clean and Concentrate kit to a final volume of 15 μL using NFH_2O according to the manufacturer's instructions.
11. To degrade the template strand, transfer 14 μL of the purified DNA product to a new PCR tube, and add 2 μL of 10× CutSmart Buffer, 2 μL of 1:50 diluted exonuclease III (final concentration of 2 U/μL), 1 μL of 1:10 Nb.BbvCI (final concentration of 1 U/μL), and 1 μL of exonuclease I.
12. Place the reaction tube in a preheated (37 °C) thermal cycle with the following program: 60 min at 37 °C, 20 min at 80 °C, and hold at 4 °C.
13. To synthesize the second mutagenic strand, add 27.7 μL NFH_2O, 20 μL of 5× Phusion HF Buffer, 3.3 μL of 1:20 diluted phosphorylated SEC_REV primer, 20 μL of 50 mM DTT, 1 μL of 50 mM NAD^+, 2 μL of 10 mM dNTPs, 5 μL of *Taq* DNA ligase, and 1 μL of Phusion High-Fidelity DNA Polymerase to the same reaction mixture. Mix the tube content briefly.
14. Place the tube in a preheated (98 °C) thermal cycler with the following program: 30 s at 98 °C, 45 s at 55 °C, 10 min at 72 °C, 20 min at 45 °C, and hold at 4 °C.
15. Add 2 μL of DpnI into each reaction tube, and incubate the reaction for 60 min at 37 °C to degrade methylated and hemimethylated wild-type DNA.
16. Purify each reaction using a Zymo Clean and Concentrate kit to a final volume of 6 μL using NFH_2O according to the manufacturer's instructions.
17. Transform the entire 6 μL reaction product into *E. coli* XL1-Blue following standard electrocompetent transformation protocol [17].
18. After recovery, bring the final volume of the transformation to 2–2.5 mL with additional sterile media (TB media).
19. Prepare six tenfold serial dilutions and plate 10 μL of each. To calculate the transformation efficiency, the next day count the section that contains between 10 and 100 colonies. It is important to obtain at least 99.9% coverage of the theoretical diversity of the library (*see* **Note 5**).
20. Spread the remaining cells onto the prepared large bioassay dishes.
21. Place in a 37 °C humidity-controlled incubator overnight when bioassay dishes have dried.

3.1.3 Extraction of dsDNA SSM Library Plasmid

1. On the next day, scrape the large plates using between 5 and 10 mL TB media, and collect the cells in a 50 mL centrifuge tube.

2. Vortex the cell suspension, and extract the library plasmid DNA of a 1 mL aliquot of the cell suspension using a Qiagen mini-prep kit. Additional mini-preps can be done if large amounts of library DNA are required.
3. Store the rest of the cells at −80 °C by resuspending the pellet in 3 mL of 50% v/v glycerol.

3.1.4 Chemically Competent Library Yeast Transformation

Competent yeast can be prepared up to 6 months ahead of time.

1. Grow the EBY100 cells in 500 mL YPD to an OD_{600} of 1.2 and then harvest at 4000 × *g* for 5 min.
2. Resuspend the pelleted cells in 250 mL sterile H_2O and repellet.
3. Resuspend the pelleted cells in 10 mL of 100 mM LiOAc and repellet.
4. Resuspend in 3.5 mL of 100 mM LiOAc, and then add 1.5 mL of 50% v/v glycerol and the mixture vortexed.
5. Prepare aliquots of 210 µL of cells to a tube, and store at −80 °C. Do not snap-freeze cells.
6. Boil 30 µL of salmon sperm DNA at 97 °C for 10 min.
7. Add 720 µL of 50% PEG, 108 µL of 1 M LiOAc, and 30 µL of boiled salmon sperm DNA to 210 µL of chemically competent EBY100 cells.
8. Vortex hard until there is a uniform mixture.
9. Add 5 µg of library plasmid to the mixture and vortex briefly.
10. Incubate the mixture at 30 °C for 30 min.
11. Heat shock the cells by incubating at 42 °C for 20 min.
12. Pellet the cells by spinning at 18,000 × *g* for 30 s.
13. Resuspend the cells pellet in 1 mL of SDCAA media, and let stand for 5 min.
14. Prepare six tenfold serial dilutions from the suspension, and plate on SDCAA plates using 10 µL of each. Incubate for 2–3 days at 30 °C to calculate transformation efficiency (*see* **Note 5**).
15. Add the remaining culture into 100 mL of SDCAA media. Grow for 30 h at 30 °C and 250 rpm.
16. On the next day, resuspend the cell culture at OD_{600} = 1 in 50 mL of SDCAA media.
17. Grow overnight at 30 °C and 250 rpm.
18. Prepare multiple cells stocks by pelleting, resuspending in yeast storage buffer to an OD_{600} = 1, and storing in 1 mL aliquots (approximately 1×10^7 cells) at −80 °C. Do not snap-freeze cells (*see* **Note 6**).

3.2 Library Screening

3.2.1 Preparation of Labeling Reactions

1. For each PPI partner to analyze, thaw a 1 mL aliquot as prepared on previous section, spin down at 2500 × *g* for 3 min, and remove the supernatant.
2. Resuspend the pellet in 1 mL SDCAA media, and grow for 4–6 h at 30 °C and 250 rpm.
3. Spin down the cells at 2500 × *g* for 3 min, and reinoculate at OD_{600} = 1.0 in 1 mL of SGCAA media. Induce overnight using the predetermined induction conditions (*see* **Note 7**).
4. Spin down the cells at 2500 × *g* for 3 min, wash with 1 mL of ice-cold PBS-BSA, and spin down again.
5. Resuspend the cells in ice-cold PBS-BSA at an OD_{600} = 2.0.
6. In PBS-BSA, label 1 mL (2×10^7) cells with the biotinylated protein at half of the apparent dissociation constant, and incubate at room temperature for 30 min using a tabletop mixer. Vary the total reaction volume to ensure that the number of biotinylated protein is at least tenfold higher than the PPI partner that is displayed on the yeast cell surface. For example, assuming a 10:1 partner/displayed protein ratio at a typical PPI apparent dissociation constant of 10 nM, 2×10^7 cells (1 mL) should be labeled with 5 nM biotinylated partner protein (half of the apparent dissociation constant). The total reaction volume is calculated following Eq. 1. Thus, label 1 mL of cells 2305 μL of PBS-BSA with 16.6 μL of 1 μM partner stock solution in

$$\begin{aligned}\text{Total reaction volume} &= \frac{10^6\,\mu L}{5\,\text{nmol partner}} \\ &\times \frac{10\,\text{nmol partner}}{1\,\text{nmol displayed protein}} \\ &\times \frac{1\,\text{nmol displayed protein}}{6.02\times10^{14}\,\text{protein}} \\ &\times \frac{5\times10^4\,\text{protein}}{\text{cells}} \times 2\times10^7\,\text{cells} \\ &= 3322\,\mu L\,\text{of total reaction volume}\end{aligned} \tag{1}$$

7. Spin down at 2500 × *g* for 5 min, wash the pellet with 5 mL of PBS-BSA and spin down, and remove supernatant again. In this and subsequent steps, PBS-BSA should be ice-cold, the tabletop centrifuge should be refrigerated, and all tubes should be kept on ice and protected from light.
8. Label cells with 60 μL of FITC, 50 μL of SAPE, and 1.89 mL of PBS-BSA, vortex briefly, and incubate the labeled cells on ice for 10 min.
9. Repeat **step 7**.
10. Leave the cell pellet on ice until ready to sort.

3.2.2 Sorting Conditions Set-Up

1. Set Gate1, Gate2, and Gate3 on your cell sorter as shown in Fig. 3.
2. Add 4 mL of ice-cold PBS-BSA to the cell pellet, mix by vortexing, and transfer to a FACS-compatible tube.
3. Obtain the reference population by sorting 240,000 cells (*see* **Note 8**) using Gate1$^+$ (Fig. 3a).
4. Obtain the displayed population by sorting 240,000 cells using Gate1$^+$/Gate2$^+$ (Fig. 3b).
5. Obtain the bound population for each PPI by sorting 240,000 cells using Gate1$^+$/Gate2$^+$/Gate3$^+$ (Fig. 3c).
6. Recover the collected cells in 5 mL of SDCAA media for approximately 30 h at 30 °C and 250 rpm.
7. Prepare cells stocks by storing 1 mL of OD_{600} = 4 cell stocks in yeast storage buffer and at −80 °C.

3.3 Deep Sequencing Preparation of Yeast DNA

3.3.1 Primer Design and Library Amplification Test

Yeast DNA is prepared for deep sequencing using a two-step PCR amplification: the first step is with a gene-specific primer set ("inner" primers), while the second step uses an invariant set of "outer" primers (Fig. 4). Inner primers are designed to be complementary to adjacent 5′ and 3′ ends of each library followed by an Illumina universal primer sequence (Fig. 4a). The following rules needs to be considered to determine these regions:

1. The length of the segment section plus the library should not be longer than 250 base pairs.

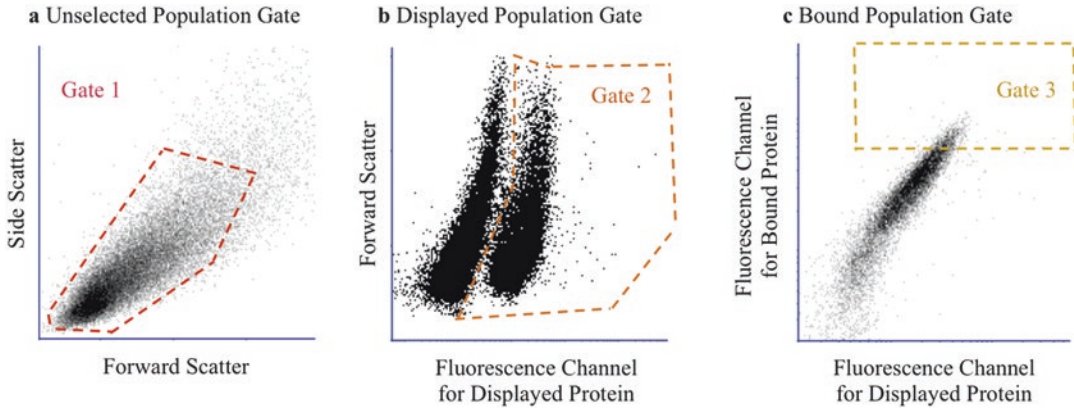


Fig. 3 Sorting gates used for library screening. Yeast SSM libraries are labeled with biotinylated complementary protein at half of the apparent dissociation constant. Next, SSM libraries are sorted using three different gates as shown: (**a**) Gate1 set with the light scatter parameters for yeast, forward scatter/side scatter; (**b**) Gate2 set on the forward scatter and the fluorescence channel for displayed protein (FITC); and (**c**) Gate3 set on the fluorescence channel for displayed protein and fluorescence channel for bound protein. Gate3 is configured to collect the top 5–10% of the bound population

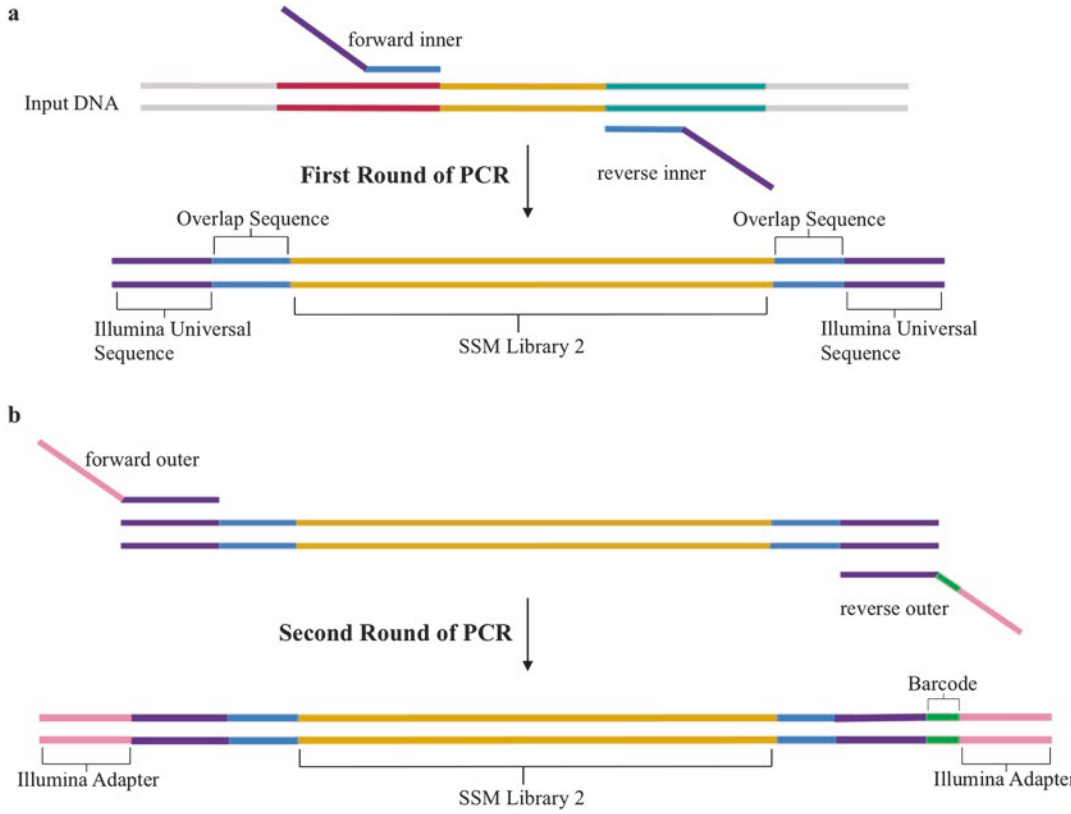


Fig. 4 PCR steps performed for deep sequencing preparation of SSM libraries. Sequential PCR reactions to amplify the genes of interest and attach the Illumina adapters are shown for SSM library 2 (gold). (**a**) After extracting the plasmid DNA from yeast cells, SSM libraries are amplified by PCR using a set of inner primers containing a segment that overlaps with the gene of interest (light blue) and the Illumina universal sequence (purple). (**b**) A second round of PCR is performed to attach the Illumina adapter sequence using a set of outer primers which contain an overlapping region to the Illumina universal sequence (purple), a unique barcode on the reverse primer (green), and Illumina adapter sequences (yellow)

2. Design the segment region to have a melting temperature of 53–56 °C using the NEB Phusion melting point calculator using Phusion High-Fidelity Polymerase.
3. Once the gene-specific sequence is designed, append the conserved primer sequence as shown in Table 1.
4. Upon receiving the inner primers, we recommend performing a PCR verification with wild-type plasmid as a template to confirm a single band of the expected size.

Further steps for yeast DNA deep sequencing preparation require the addition of universal primers to add the Illumina adapters and barcodes. Universal primers are designed using the TruSeq small RNA oligo sequences. The forward primer is the same for all preparations, while the reverse primer contains an indexing bar-

Table 1
Gene amplification and Illumina adapter primers to prepare samples for deep sequencing

Inner primers for library amplification	
Primer name	Sequence
Inner_FWD	5′-*gttcagagttctacagtccgacgatc* < segment that overlaps to sense strand > – 3′
Inner_REV	5′-*ccttggcacccgagaattcca* < segment that overlaps to antisense strand > – 3’
Outer primers to add the Illumina adapters and barcodes	
Illumina_FWD	5′-**aatgatacggcgaccaccgagatctacac***gttcagagttctacagtccgacgatc* - 3’
Illumina_REV	5′-**caagcagaagacggcatacgagat**nnnnnn**gtgactggagtt***ccttggcacccgagaattcca* - 3’
Bold, Illumina adapter; nnnnnn, indexing barcode (*see* Kowalsky et al. [18] for complete set); and italic, Illumina universal sequence	

code that allows multiplexing of samples on an Illumina lane (Fig. 4b; full sequences shown in Table 1).

3.3.2 Yeast Plasmid DNA Preparation for Deep Sequencing

1. Thaw an aliquot of the stored yeast library by hand, spin down at 2500 × *g* for 3 min, and remove the supernatant.
2. Resuspend the pellet cells in 200 μL of Solution 1 and add 5 μL of 5 U/μL of Zymolyase.
3. Incubate at 37 °C for 4 h, and mix once per hour.
4. Perform one freeze-thaw cycle in dry ice/EtOH bath and 42 °C incubation.
5. Add 200 μL of Solution 2, mix briefly, and incubate for 3–5 min at room temperature.
6. Add 400 μL of Solution 3, mix well, and centrifuge at 17,000 × *g* for 5 min.
7. Transfer the supernatant to a Qiagen mini-prep column, and spin down for 1 min at 17,000 × *g*.
8. Add 700 μL of PB buffer, and spin down for 30 s at 17,000 × *g*.
9. Add 700 μL of PE buffer, and spin down for 30 s at 17,000 × *g*.
10. Repeat **step 9**.
11. Take out supernatant, and spin down again at 17,000 × *g* for 1 min to dry the column.
12. Transfer the column to a new clean 1.5 mL microfuge tube, add 30 μL of elution buffer, and spin down for 1 min at 17,000 × *g*.
13. Reload the column with the eluate, and spin down again. Store 15 μL of eluate, and proceed with the remaining 15 μL.

14. For the purification of plasmid from the yeast preparation, in a PCR tube, add 15 μL of dsDNA plasmid, 2 μL of exonuclease I, 1 μL of lambda nuclease, and 2 μL of 10× lambda buffer.
15. Place the mixture in a preheated (30 °C) thermocycler with the following cycle: 90 min at 30 °C, 20 min at 80 °C, and hold at 4 °C.
16. Clean the PCR product following the standard procedure from Qiagen mini-prep PCR cleanup, and elute in 30 μL of TE buffer.
17. Store 15 μL of eluate, and proceed with the remaining 15 μL.
18. For the gene library amplification, in a PCR tube, add 10 μL of 5× Phusion HF Buffer, 18.5 μL of NFH_2O, 1 μL of 10 mM dNTPs, 2.5 μL of 10 μM of forward inner primer, 2.5 μL of 10 μM of reverse inner primer, 15 μL of dsDNA template, and 0.5 μL of Phusion High-Fidelity Polymerase.
19. Place the tube in a preheated (98 °C) thermocycler with the following cycle: 30 s at 98 °C, 14× cycles of 5 s at 98 °C, 15 s at 53 °C, and 15 s at 72 °C, follow by a final incubation for 10 min at 72 °C, and a hold at 4 °C.
20. Add 1.87 μL of 1:10 diluted exonuclease I.
21. Place the tube back in the thermocycler with the following cycle: 30 min at 37 °C, 5 min at 95 °C, and a hold at 4 °C.
22. In a new PCR tube, add 10 μL of 5× Phusion HF Buffer, 32.5 μL of NFH_2O, 1 μL of 10 mM dNTPs, 2.5 μL of 10 μM of forward outer primer, 2.5 μL of 10 μM of reverse outer primer, 1 μL of dsDNA template from previous PCR amplification, and 0.5 μL of Phusion High- Fidelity Polymerase.
23. Repeat the same PCR cycle used for the inner primers.
24. Run 5 μL of PCR product on 2% agarose gel, and visualize with SYBR Gold. It is important to verify that you have single clear band before proceeding (*see* **Note 9**).
25. Purify and clean the PCR product using Agencourt AMPure XP following the manufacturer's instructions for the 96-well format procedure.
26. Measure the concentration of each sample.
27. Stored the purify product at −20 °C.

3.3.3 dsDNA Quantification Using Quant-iT PicoGreen

At this point, samples are ready for deep sequencing. Follow the instructions for the Illumina MiSeq 2x250 bp submission from your sequencing facility. Usually, each Illumina MiSeq sequencing holds between 10 and 15 million reads per lane. Based on the read depth and library size, calculate the amount of reads necessary for each sample—our group uses approximately 500,000 reads per sample and multiplexes 20–30 samples per lane. Individual samples

are quantified and mixed together in a single vial. The following procedure was adopted from the Invitrogen MP 07581 manual. The final yield should be about 1–4 ng in 40 μL.

1. Allow the Quant-iT reagent to warm to room temperature while covered in foil.
2. Prepare a 200-fold dilute solution of Quant-iT into TE buffer using a foil covered culture tube. (Example: 25 μL of PicoGreen reagent into 4.975 mL of TE). This solution should be prepared and used the day of the experiment.
3. Beginning with a 50 ng/mL stock of a kit-supplied lambda DNA standard, prepare a blank and a 1:2 standard curve (0, 1.56, 3.12,…, 25 ng/mL) using the first column of a 96-well black plate.
4. In a black 96-well plate, add 2.5 μL of each sample to 97.5 μL of TE in wells.
5. Carry out extra dilutions as necessary if the concentration is too high.
6. Add 100 μL of diluted PicoGreen solution to DNA samples and standard samples, mix briefly, and incubate for 5 min at room temperature covered with foil to protect from light.
7. Measure the fluorescence of the samples (excitation ~480 nm, emission ~520 nm).
8. Subtract the fluorescence value of the reagent blank from that of each of the samples.
9. Use the corrected data to generate a standard curve of fluorescence versus DNA standard concentrations, and calculate the concentrations of each sample. In our hands, the final concentration is between 5 and 40 ng/μL.
10. Mix equivalent mass amounts of samples in a single 1.5 mL Eppendorf tube, and send to your sequencing facility.

3.4 Data Analysis

Custom scripts used in the data analysis are available at GitHub (user: JKlesmith). Sample command lines and instructions are provided at the same source.

1. Use the modified version of Enrich 0.2 software as describe in Kowalsky et al. [18] to compute the enrichment ratios of individual mutants for the DNA sequencing results from Illumina MiSeq run (Fig. 5 *see* **Note 10**). Enrich 0.2 [8] documentation is available at http://depts.washington.edu/sfields/software/enrich/docs/0.2/enrich.html. The output from Enrich 0.2 is required as input for the remaining steps. The wild-type protein sequence is also required as input for the following steps.

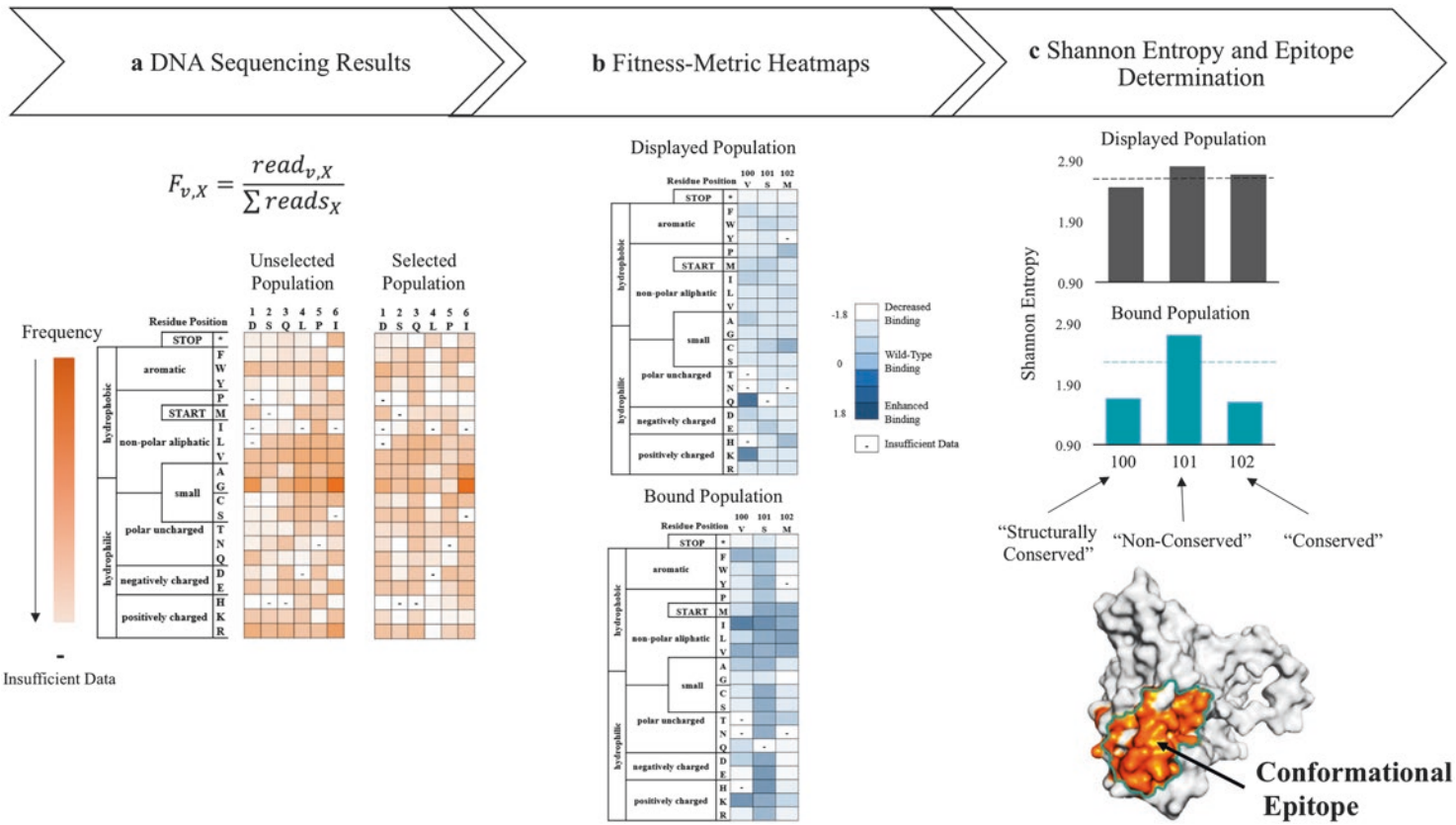


Fig. 5 Deep sequencing results and data analysis used to determine the conformational epitope. (**a**) DNA sequencing results are processed using Enrich 0.2 software [19] to calculate the frequency, $F_{v,X}$, of each point mutant, *v,* for each position, *x*, in the primary sequence. (**b**) The frequency data of each variant from different populations is transformed into heatmaps comparing the relative fluorescence of each variant in the displayed population (top) and the bound population (bottom) against the unselected population. (**c**) Heatmaps are used to calculate the Shannon entropy for each residue on the displayed (black) and bound populations (turquoise). Next, the entropy is used to determine the conserved and non-conserved positions, which allows for identification of the conformational epitope

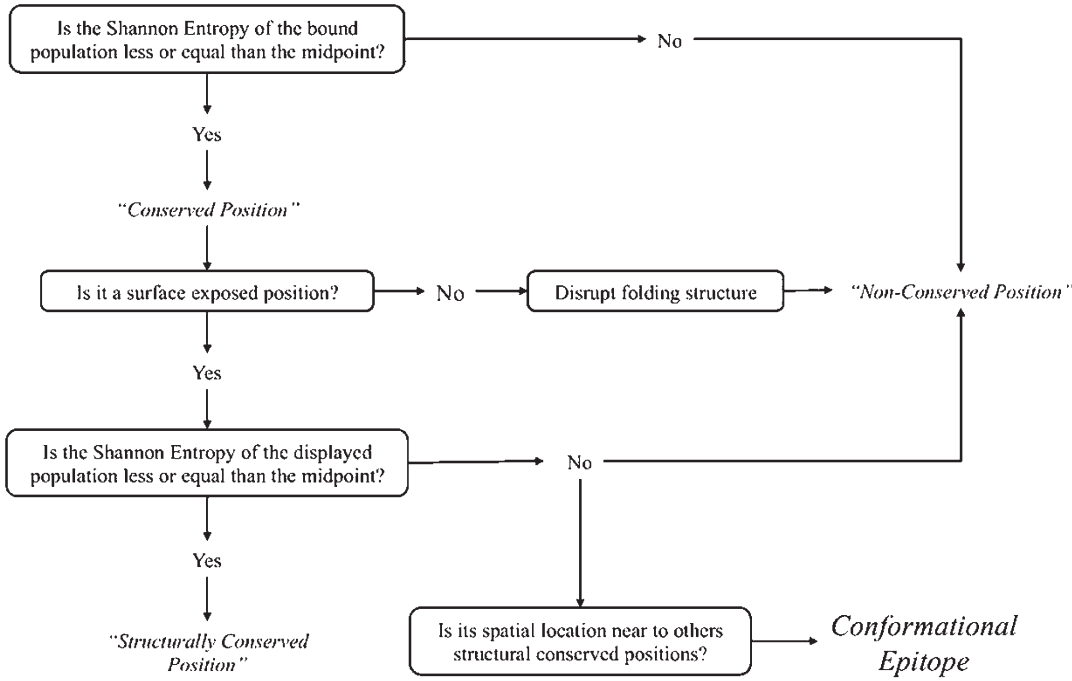


Fig. 6 Flowchart of analysis used to determine the conformational epitope

2. The relative binding of each variant on the displayed and bound population is calculated using a custom Python script called QuickNormalize.py (*see* **Note 11**). The output from this script is a .csv file that can be read by multiple programs. In our lab we use Microsoft Excel to visualize the data as heatmaps and to carry out the data analysis (*see* **Note 12**).
3. Calculate the Shannon entropy for each variant on the displayed and bound population using a custom script called FACSEntropy.py – the output file is a .csv. The entropy values are used to discriminate those residues that participate in the protein-protein interaction and to determine the conformational epitope following the cut-off analysis flowchart as shown in Fig. 6 (*also see* **Note 13**).
4. Calculate the reportable statistics using QuickStat.py script. Statistics will report the reads passing through enrich; the percentage of possible codon substitutions observed; the percent of reads with none, one, and multiple nonsynonymous mutations; and the coverage of possible single nonsynonymous mutations.

4 Notes

1. The PPI partner protein is chemically biotinylated following the instructions for EZ-Link NHS-Biotin Reagents (Thermo

Fisher). We prefer chemical biotinylation to genetically encoded biotinylation (e.g., AviTag) as the former has a higher fluorescence signal. If proteins are small, covalent labeling with multiple biotins may disrupt the structure; in such a case, we recommend genetically fusing the PPI partner to a carrier-like maltose-binding protein or an IgG Fc. Anecdotally, we have noticed cleaner results with PPI partners with monovalent interactions and for that reason recommend creating a Fab if the PPI partner is a mAb.

2. The following rules apply for preparing separate mutagenesis libraries: (1) the length of each library should be divisible by three to avoid splitting a codon; (2) the gene should be segmented into libraries with a maximum length of 225 base pairs for Illumina 250 bp paired-end sequencing (273 base pairs for 300 bp paired-end sequencing); and (3) libraries should be similar in length (±three nucleotides).
3. In some cases, the gene sequence of interest also contains a BbvCI restriction site. If the site is in the same orientation as the site on the pETconNK plasmid, continue the protocol as usual. If the BbvCI site is in the opposite direction as the site on the pETconNK plasmid, use the YSD plasmid pETCON (Addgene # 41522) as this plasmid does not contain a nicking site. The orientation of the BbvCI may not be in the same way as exists on the pETconNK plasmid. For example, if the nicking site is in the opposite direction from Fig. 2b, Nb.BbvCI (not Nt.BbvCI) should be used first to create the ssDNA wild-type template; otherwise follow the protocol as described in Subheading 3.1.2.
4. We recommend preparing phosphorylated oligos no earlier than the day before the nicking mutagenesis procedure. Avoid repeated freeze-thaw cycles.
5. For a library with ***NNK*** SSM at 75 amino acids, the theoretical library size is 2400 nucleotide variants. The percentage theoretical coverage is described by the following equation: $\%\text{coverage} = \left(1 - e^{\frac{-\text{number of transformants}}{\text{theoretical library size}}}\right) \times 100$. In the above case, 16,500 transformants will give 99.9% coverage.
6. At this point, cells could be inoculated in fresh SGCAA media for Library Screening Preparation (Subheading 3.2) or frozen aliquots can be prepared for long-term storage. We often reinoculate the cells in SGCAA media to an OD_{600} = 1 and induce at 22 °C to confirm that the mutagenesis libraries displays on the yeast surface and binds the PPI partner. We prepare between 20 and 48 aliquots for long-term storage.

7. Induction temperature should be the same as used to prepare the PPI partner in a YSD format according to Chao et al. [9]. For each new YSD protein, our lab tests induction of surface display at 18, 20, 22, and 30 °C.
8. It is important that the number of collected cells should be at least 100-fold higher than the theoretical library size to avoid complexity bottlenecks. For example, at least 240,000 cells should be collected for each sorted population for a library with ***NNK*** SSM at 75 amino acids with a theoretical library size of 2400 nucleotide variants.
9. If the correct band size was not obtained from the second PCR product, we recommend troubleshooting by running 5 μL of the first PCR product on a 2% agarose gel and visualized on SYBR-Gold to identify which PCR amplification did not work. We recommend staining the gel on SYBR Gold for at least 1 h to resolve low-intensity bands. Fewer or more cycles of each PCR could be used to improve the product.
10. Our group routinely analyzes the quality of the Illumina sequencing data using FastQC available online at http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Poor quality reads can hinder the data analysis using Enrich 0.2. The quality of the Illumina sequencing data is highest for the forward read and the first 150 bp. For issues where quality is poor on the reverse read, perform Enrich only for the forward read. We have also performed Enrich for short segments of the reads where the quality is highest.
11. The relative binding (ζ_i) for variant i is defined as

$$\zeta_i = \log_2\left(\frac{\overline{F}_i}{\overline{F}_{wt}}\right) \tag{2}$$

where $\overline{F}_i$ is the mean fluorescence of variant i and $\overline{F}_{wt}$ is the mean fluorescence of wild type. There are a number of assumptions used to calculate relative binding—*see* Kowalsky et al. [7] for further details.

12. Positions with insufficient data at more than ten substitutions should be excluded from analysis.
13. In the current experimental set-up, discriminating mutations that disrupt the interface and maintain the overall fold between those that destabilize the structure is difficult to determine, as unfolded mutants still predominantly display on the yeast surface [4, 20]. However, a recent study confirms that destabilizing mutations display with fewer copies on the yeast surface than stabilizing mutations, at least for proteins with >200 residues [15]—for small proteins destabilizing mutants appear to display at the same rate as stable mutants (T.A.W. and A.M.C.,

unpublished data). To further identify mutations that stabilize larger proteins, a FACS protocol is used with a sort gate set to collect the top 5% of the displaying population. For library screening, 2×10^6 yeast cells per mL, in PBS-BSA, are labeled with 1 μL of anti-c-myc-FITC per 2×10^5 yeast cells. The population is sorted using a gate that collects the top 5% of the displaying population. Shannon entropy obtained from this study is used to identify structurally conserved positions.

Acknowledgments

This work was supported by NSF CAREER (Award #1254238) to T.A.W. and a NIH T32 Biotechnology Training Grant (Award # T32-GM110523) to A.M.C.

References

1. Weiss GA, Watanabe CK, Zhong A et al (2000) Rapid mapping of protein functional epitopes by combinatorial alanine scanning. Proc Natl Acad Sci U S A 97:8950–8954. https://doi.org/10.1073/pnas.160252097
2. Chao G, Cochran JR, Dane Wittrup K (2004) Fine epitope mapping of anti-epidermal growth factor receptor antibodies through random mutagenesis and yeast surface display. J Mol Biol 342:539–550. https://doi.org/10.1016/j.jmb.2004.07.053
3. Fowler DM, Fields S (2014) Deep mutational scanning: a new style of protein science. Nat Methods 11:801–807. https://doi.org/10.1038/nmeth.3027
4. Whitehead TA, Chevalier A, Song Y et al (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol 30:543–548. https://doi.org/10.1038/nbt.2214
5. Van Blarcom T, Rossi A, Foletti D et al (2015) Precise and efficient antibody epitope determination through library design, yeast display and next-generation sequencing. J Mol Biol 427:1513–1534. https://doi.org/10.1016/j.jmb.2014.09.020
6. Doolan KM, Colby DW (2015) Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing. J Mol Biol 427:328–340. https://doi.org/10.1016/j.jmb.2014.10.024
7. Kowalsky CA, Faber MS, Nath A et al (2015) Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing. J Biol Chem 290:26457–26470. https://doi.org/10.1074/jbc.M115.676635
8. Fowler DM, Araya CL, Fleishman SJ et al (2010) High-resolution mapping of protein sequence-function relationships. Nat Methods 7:741–746. https://doi.org/10.1038/nMeth.1492
9. Chao G, Lau WL, Hackel BJ et al (2006) Isolating and engineering human antibodies using yeast surface display. Nat Protoc 1:755–768. https://doi.org/10.1038/nprot.2006.94
10. Van DJA, Wittrup KD (2014) Yeast surface display for antibody isolation: library construction, library screening, and affinity maturation. Methods Mol Biol 1131:151–181. https://doi.org/10.1007/978-1-62703-992-5_10
11. Mata-Fink J, Kriegsman B, Yu HX et al (2013) Rapid conformational epitope mapping of anti-gp120 antibodies with a designed mutant panel displayed on yeast. J Mol Biol 425:444–456. https://doi.org/10.1016/j.jmb.2012.11.010
12. Adams RM, Mora T, Walczak AM et al (2016) Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. Elife 5:5980–5985. https://doi.org/10.7554/eLife.23156
13. Wrenbeck EE, Klesmith JR, Stapleton JA et al (2016) Plasmid-based one-pot saturation mutagenesis. Nat Methods 13:928–930. https://doi.org/10.1038/nmeth.4029
14. Gietz RD, Schiestl RH (2007) High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc 2:31–34. https://doi.org/10.1038/nprot.2007.13

15. Klesmith JR, Bacik J-P, Wrenbeck EE et al (2017) Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. Proc Natl Acad Sci U S A 114:2265–2270. https://doi.org/10.1073/pnas.1614437114
16. Wrenbeck E, Klesmith J, Stapleton J, Whitehead T (2016) Nicking mutagenesis: comprehensive single-site saturation mutagenesis. Protoc Exch. https://doi.org/10.1038/protex.2016.061
17. Sambrook J, Russell DW (2006) Transformation of E. coli by electroporation. CSH Protoc 2006:pdb.prot3933. https://doi.org/10.1101/pdb.prot3933
18. Kowalsky CA, Klesmith JR, Stapleton JA et al (2015) High-resolution sequence-function mapping of full-length proteins. PLoS One 10:e0118193. https://doi.org/10.1371/journal.pone.0118193
19. Fowler DM, Araya CL, Gerard W, Fields S (2011) Enrich: software for analysis of protein function by enrichment and depletion of variants. Bioinformatics 27:3430–3431. https://doi.org/10.1093/bioinformatics/btr577
20. Kowalsky CA, Whitehead TA (2016) Determination of binding affinity upon mutation for type I dockerin-cohesin complexes from *Clostridium thermocellum* and *Clostridium cellulolyticum* using deep sequencing. Proteins 84:1914–1928. https://doi.org/10.1002/prot.25175

Chapter 8

Structurally Guided In Vivo Crosslinking

Johanna C. Scheinost and Thomas G. Gligoris

Abstract

The focus of modern molecular biology on protein structure and function has reached unparalleled levels. Whether interacting with nucleic acids or other proteins, protein contacts are the basis for fine-tuning all cellular processes. It is for this reason imperative that protein interactions are studied in ways that reflect actual events taking place inside living cells.

Here, we describe in detail a method that combines the residue-level resolution provided by structural biology with physiological studies inside living cells, with the overall goal of explaining the contribution of protein–protein interactions in cellular processes. We use as a powerful example our experience with the DNA exit gate interface of the chromosomal cohesin complex, and we argue that this methodology may be followed to address similar questions within any protein complex and in various model systems.

Key words Cohesin, Protein–protein interaction, Cysteine, Immunoprecipitation, Yeast, Crosslinking, BMOE, dBBr

1 Introduction

With recent progress in molecular biology, a confounding problem remains the level of agreement between in vivo (as in taking place inside live cells) and in vitro (as in biochemically reconstituted) experimentation: it is often the case that biochemical systems either do not fully recapitulate in vivo observations or even possess characteristics which are in conflict with in vivo evidence.

We describe here a method for crosslinking proteins bearing engineered cysteine residues. The thiol group of cysteine is used in order to covalently crosslink neighboring cysteine pairs of two interacting proteins using appropriate homobifunctional thiol-specific chemical reagents (crosslinkers). Crucially, these crosslinkers can be used on live cells and therefore capture crosslinking events taking place inside live budding yeast cells. Thus, this methodology can be used to study in vivo the interaction of proteins in different stages of the cell cycle during various genetic conditions or in altered physiological conditions.

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_8,

The method we describe here relies on known structural information of the protein interfaces to be studied. Whether this information is derived from X-ray crystal structures, NMR, or cryo-electron microscopy is not important; however, the level of resolution of the available structures determines the feasibility of the project: the higher the resolution, the better the chances of designing correct cysteine replacements and eventually capturing the protein interface formation by crosslinking.

The process of engineering cysteine replacements can be overviewed schematically in Fig. 1: from the available structure residues

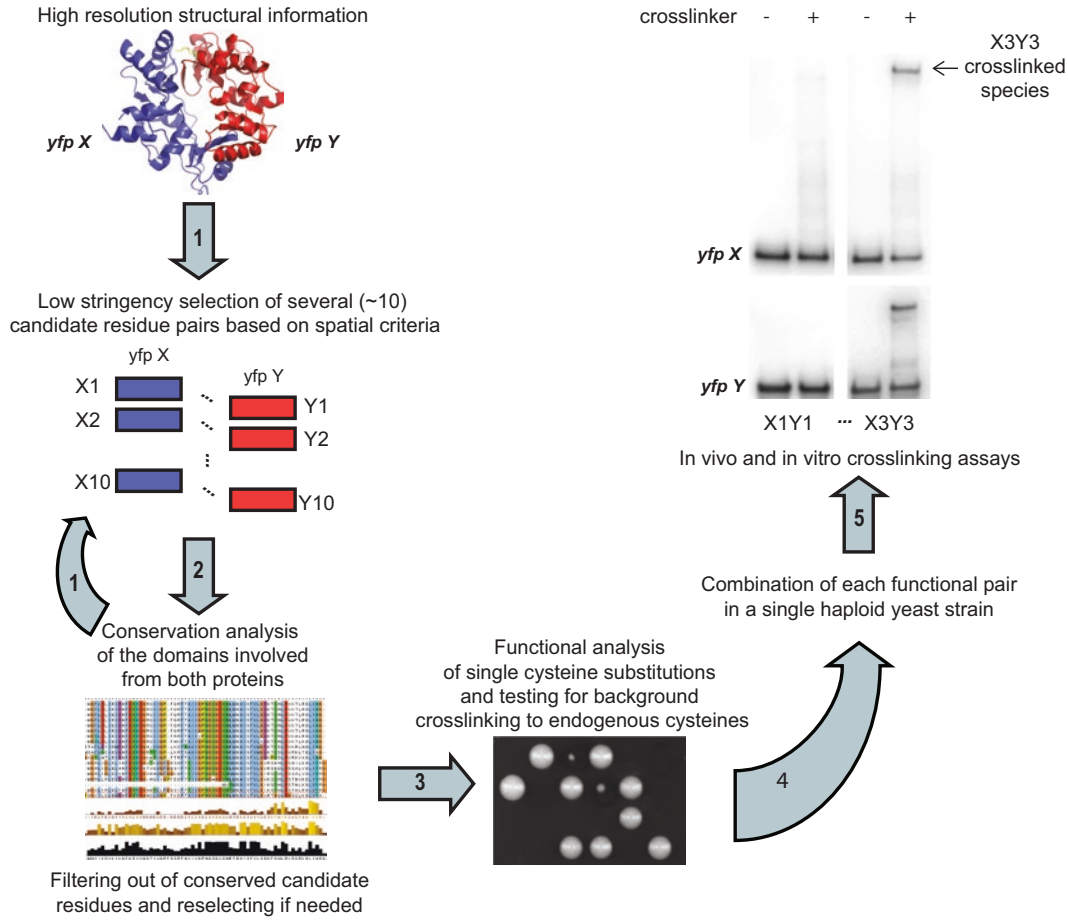


Fig. 1 Overview of the workflow. Starting with high-resolution structural data of your favorite proteins (e.g., yfpX and yfpY), a set of possible cysteine replacement pairs (X_1,Y_1 to X_{10},Y_{10} in this example) which must fit spatial criteria (*see* **Note 1**) are selected (**step 1**). Conservation analysis filters the selected pairs and if necessary more pairs are selected after rejecting conserved ones (**step 2**). The single cysteine replacements are tested for functionality and for background crosslinking using standard genetic manipulation and the protocols described in this chapter (**step 3**). Each one of the functional replacement cysteine pairs is combined eventually to a single haploid yeast strain and tested for interaction crosslinking after immunoprecipitation and finally inside living yeast cells (**step 4**)

fitting a list of simple criteria (*see* **Note 1**) are selected to be replaced by cysteine residues. Using classic yeast genetics (not to be covered in this chapter; *see* **Note 2**), yeast strains bearing single replacements are tested for adequate functionality of the mutant alleles. Next, the engineered alleles of the two interacting proteins are combined in a final yeast strain. Both replacements are tested thereafter for their combined functionality, and the resulting pairs are then tested for crosslinking both in vitro and in vivo (*see* **Note 3**).

In this chapter we list the criteria to be used to select the amino acid pairs to be replaced by cysteines, and we present the protocols to test cysteine pair crosslinking efficiency using various bifunctional thiol-reactive crosslinkers. This method has been extensively used in the characterization of the DNA exit gate interface of the *Saccharomyces cerevisiae* cohesin complex [1], and our experiments advocate its versatility in other model systems with adequate optimization.

2 Materials

Prepare all solutions using ultrapure water (18 MΩ cm at 25 °C) and analytical-grade reagents. Prepare and store all reagents at room temperature (unless indicated otherwise).

2.1 Reagents

1. PBS (phosphate buffered saline).
2. Dibromobimane (dBBr, Sigma).
3. Bismaleimidoethane (BMOE, Thermo Fisher).
4. Dimethyl sulfoxide (DMSO).
5. 1 M dithiothreitol (DTT).
6. EBX lysis buffer [2] (50 mM HEPES pH 8.0, 100 mM KCl, 2.5 mM $MgCl_2$, 0.05%, NP40, 0.25% Triton-X).
7. 10 mg/ml RNase A.
8. Roche cOmplete protease inhibitor tablets (EDTA-free).
9. Phenylmethylsulfonyl fluoride (100 mM PMS in isopropanol).
10. Acid-washed glass beads (425–600 μm).
11. Bradford solution.
12. Anti-V5 antibody (e.g., from Bio-Rad).
13. Protein G-coupled Dynabeads (Thermo Fisher).
14. 2× SDS sample buffer.
15. 3–8% Tris–acetate gels (e.g., Invitrogen).
16. Tris–acetate running buffer (Invitrogen).

2.2 Equipment

1. FastPrep-24 (MP Biomedicals) or similar bead beater.
2. 2 ml screw-cap tubes, needle (23 gauge).
3. Trans-Blot® Turbo™ Transfer System (Bio-Rad) or similar.
4. PVDF transfer packs (Bio-Rad) and further standard lab equipment.

3 Methods

3.1 In Vitro Crosslinking of Immunoprecipitated Complexes Using Dibromobimane (dBBr)

1. Grow *S. cerevisiae* strains in appropriate medium to log phase (e.g., $OD_{600\,nm}$ = 0.6, 50–100 ml).
2. Spin down 30–60 OD units (Heraeus Multifuge, 3 min, 1540×*g*, room temperature).
3. Resuspend pellet in 500–1000 μl EBX lysis buffer, freshly supplemented with RNase (0.1 mg/ml), Roche cOmplete protease inhibitors (2×), 1 mM PMSF, and 1 mM DTT or β-mercaptoethanol.
4. Add 500 μl of acid-washed glass beads.
5. Lyse cells in a FastPrep-24 for 3 × 1 min at 6 m/s. Put tubes on ice for 5 min in between FastPrep runs (unless the beating is performed in a cold room).
6. When establishing the protocol, check for complete lysis: add 3 μl lysate to a glass slide, cover with coverslip, and observe under a light microscope with 40× magnification. Cells should appear ghost-like and broken.
7. Pierce bottom of tube with hot 23-gauge needle, push tubes into 2 ml Eppendorf tubes, and spin at 2 krpm for 30 s at 4 °C in tabletop cold centrifuge (alternatively: recover lysate with a pipette; small amount of glass beads still present won't interfere with later steps).
8. Recover lysates and clear by centrifugation (20 min, 12 kg at 4 °C).
9. If desirable, determine protein concentrations, e.g., by Bradford assay, and balance protein concentrations between all samples.
10. Add antibody specific to one of the crosslinked proteins of interest/epitope tag (e.g., 3 μl anti-V5 for PK tags) to lysate for 1 h with end-over-end rotation.
11. Add 30 μl protein G-coupled Dynabeads for 1 h with end-over-end rotation (other resins may be used; however, in our experience the use of protein G coupled to magnetic beads results in very low background).
12. Wash beads with 3 × 1 ml EBX buffer without any reducing agent using a magnetic rack (add buffer, remove supernatant using magnetic rack, repeat twice, and after last wash remove remaining buffer).

13. Add 600 µl PBS (again with no reducing agent present) and resuspend beads.
14. Split the bead suspension in two (~300 µl in each).
15. Add 8 µl dBBr (stock: 1.75 mg of dBBr in 1 ml of DMSO, 5 mM in DMSO, 130 µM final) or 8 µl DMSO to suspension, and carefully shake the tube.
16. Incubate at 4 °C for 10 min.
17. Wash beads with 3 × 1 ml PBS buffer using a magnetic rack (add the buffer, and wash the beads while the tube is still on the rack by mixing two to three times; let the magnetic beads adhere to the wall of the tube, and remove the supernatant buffer; repeat the wash at least two more times).
18. After the last wash, remove the supernatant buffer using the magnetic rack; spin down without exceeding 10 kg, and remove the remaining buffer with the tube in the magnetic rack.
19. Immediately add 25 µl of PBS buffer to the beads; move tube to a standard rack.
20. Add 30 µl of 2× SDS sample buffer.
21. Elute immunoprecipitated material from beads (95 °C for 5 min).
22. Run 5–10 µl on a 3–8% Tris–acetate gel using Tris–acetate running buffer; homemade low-percentage acrylamide gels (e.g., 5 and up to 8%) are a valid option as well.
23. Blot onto PVDF membrane using Bio-Rad trans-blot turbo transfer system or similar blotting device.
24. Visualize using anti-PK antibody and anti-mouse IgG-HRP on a LI-COR Odyssey Fc or standard film.

3.2 In Vivo Crosslinking in Yeast Cells Using Bismaleimidoethane (BMOE)

All steps are to be carried out at 4 °C unless specified otherwise.

3.2.1 Crosslinking

1. Grow *S. cerevisiae* strains in appropriate medium to log phase (e.g., $OD_{600nm} = 0.6$).
2. Spin down 12 OD units (Heraeus Multifuge, 3 min, 1540×*g*, room temperature)
3. Wash the pellet in ice-cold PBS (20 ml).
4. Resuspend pellet in 1 ml ice-cold PBS and split into 2 × 600 µl in 2 ml screw-cap tubes.
5. Add 25 µl BMOE (stock: 125 mM in DMSO, 5 mM final) or 25 µl DMSO to the cell suspension, vortex briefly, and incubate for 6 min.

6. Spin down cells in benchtop refrigerated centrifuge (15 s, maximum speed).
7. Wash cells 2× with 1 ml cold PBS supplemented with 5 mM DTT.
8. Pellets may be flash frozen in liquid nitrogen and stored at −80 °C until needed.

3.2.2 Immunoprecipitation and Western Blot of Crosslinked PK-Tagged Protein

1. Resuspend pellet in 500 µl EBX lysis buffer, freshly supplemented with 0.1 mg/ml RNase, Roche cOmplete protease inhibitors (2×), and 1 mM PMSF.
2. Add 500 µl of acid-washed glass beads.
3. Lyse cells in a FastPrep-24 for 3 × 1 min at 6 m/s. Put tubes on ice for 5 min in between FastPrep runs.
4. Check for complete lysis: add 3 µl lysate to a glass slide, cover with coverslip, and observe under a light microscope with 40× magnification. Cells should appear ghost-like and broken.
5. Pierce bottom of tube with hot 23-gauge needle, push tubes into 2 ml Eppendorf tubes, and spin at 2 krpm for 30 s at 4 °C in tabletop cold centrifuge.
6. Recover lysates and clear by centrifugation (5 min, 12 kg).
7. If desirable, determine protein concentrations, e.g., by Bradford assay, and balance protein concentrations between all samples.
8. Add antibody specific to crosslinked protein of interest/epitope tag (3 µl anti-V5) to lysate for 1 h with end-over-end rotation.
9. Add 30 µl protein G-coupled Dynabeads for 1 h with end-over-end rotation (other resins may be used; however, in our hands magnetic beads give very low background).
10. Wash beads with 2 × 1 ml EBX buffer using a magnetic rack (add buffer, vortex briefly, spin down, remove supernatant using magnetic rack, repeat, spin down once more after last wash, and remove remaining liquid).
11. Elute in 50 µl 2× SDS sample buffer (95 °C for 5 min).
12. Run 5 µl on a 3–8% Tris–acetate gel (EA0375BOX, Invitrogen) using Tris–acetate running buffer; homemade low-percentage acrylamide gels (e.g., 5% and up to 8%) are a valid option as well.
13. Blot onto PVDF membrane using Bio-Rad trans-blot turbo transfer system or adequate blotting technique.
14. Visualize using anti-PK antibody and anti-mouse IgG-HRP on a LI-COR Odyssey Fc or standard film.

3.2.3 Variation for Halo-Tagged Protein

The above protocol can also be used with Halo- or SNAP-tagged proteins. Upon covalently labeling with a fluorescent ligand, the

protein can then be visualized in gel, which allows for accurate quantification of crosslinking efficiencies. The following adjustments to the protocol are necessary:

- Use 60 OD units per strain (30 OD units ± BMOE).
- Perform immunoprecipitation in the presence of TMR-Halo ligand (5 μM; Promega) or other suitable ligand.
- Elute in 40 μl of 2× sample buffer.
- Load total eluate, and separate on 5% SDS-PAGE gel (Bio-Rad Mini-Protean, using glass plates).
- After electrophoresis, keep gel sandwiched between glass plates to minimize the risk of breaking the gel (glass will not absorb the light at the wavelengths used); rinse with MilliQ water.
- Detect TMR fluorescence of gel between glass plates on a Fuji FLA-7000 instrument or similar using Cy3 presets.

3.3 General Notes Related to the Protocols

1. Use a DNAse such as benzonase (1:1000 in EBX buffer) in the lysis buffer if the protein is DNA-bound.
2. Crosslinked species may have very high MW; use a high molecular weight protein ladder such as HiMark (Thermo Fisher) and transfer for 1.5× the standard time.
3. In order to see both crosslinked and non-crosslinked species, gradient gels such as 3–8% Tris–acetate gels (Invitrogen) are recommended, but homemade low-percentage acrylamide gels (e.g. 5%) are a valid option as well.
4. Reprobe Western blot with antibody against epitope tag of the other crosslinked protein to make sure crosslinking is specific.

4 Notes

1. *Selecting the residues to be replaced with cysteines from known structures.* We use here as a pilot case our experience with the Smc3–Scc1 interface of the budding yeast (*Saccharomyces cerevisiae*) cohesin complex (Fig. 2a). Cohesin forms a tripartite ring complex which entraps within its lumen one or both of the sister chromatid DNA fibers. While the interface allowing entry of DNA into the ring is still debatable [3, 4], the interface allowing opening of the ring and its release from chromosomes is consensually agreed upon: the Smc3-kleisin (Smc3-Mcd1/Scc1) interface, which is formed by a four-helix bundle with two helices from the Smc3 coiled coil and another two helices coming from the N-terminal part of the Scc1 kleisin subunit [1].

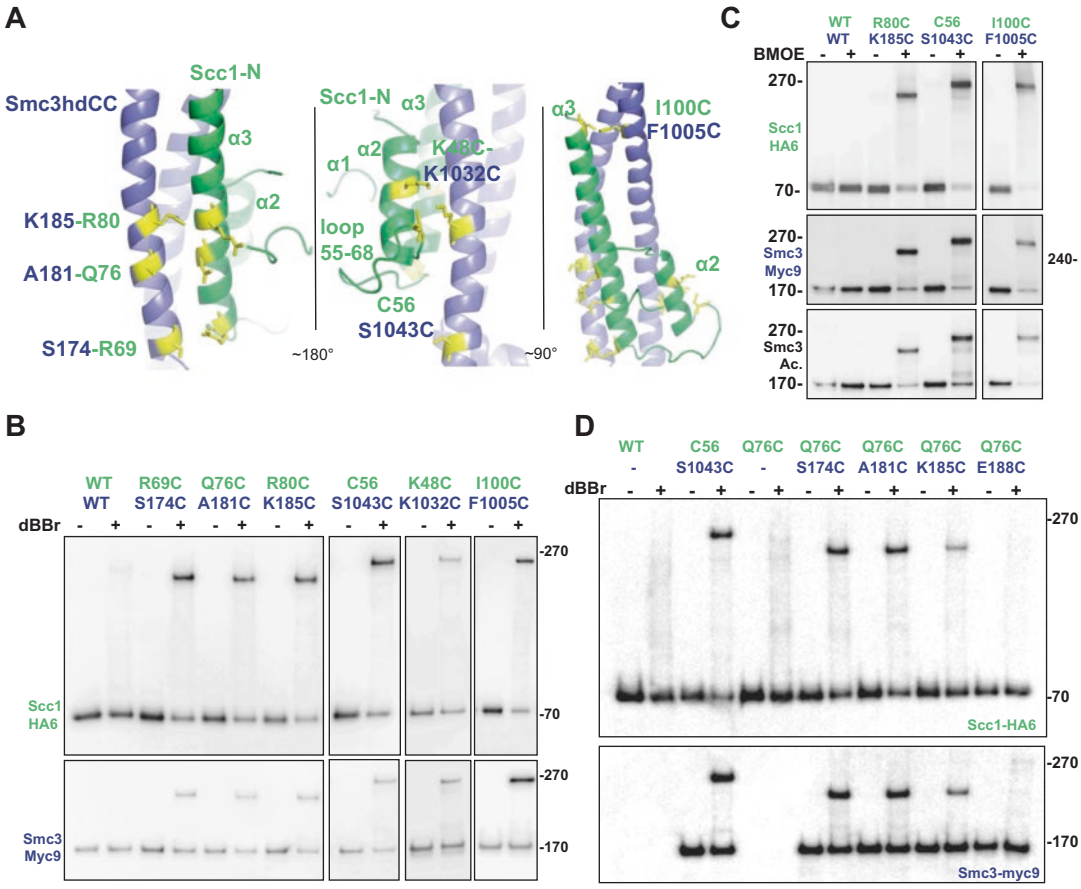


Fig. 2 Testing the Smc3-kleisin crystal structure. (**a**) Scc1-N α2 and α3 helices (green), Smc3 coiled coil (Smc3CC, blue), and substituted residues (yellow) from the crystal structure [1] of the budding yeast Smc3-Scc1 interface (pdb: 4UX3). (**b**) Thiol-specific crosslinking (dBBr) between α2 and α3 of Scc1-N and the Smc3 coiled coil (CC) after immunoprecipitation against Scc1-HA6. All mutations were functional and all observed crosslinks were dependent on a pair of cysteine substitutions. (**c**) In vivo thiol-specific crosslinking (BMOE) between α2 and α3 of Scc1-N and the Smc3 coiled coil (CC) followed by immunoprecipitation of Scc1-HA6. In all shifted bands, Smc3 is acetylated, with acetylation being a sign that functional cohesive cohesin has been crosslinked. (**d**) Specificity of the thiol-specific crosslinking assay. Immunoprecipitations against Scc1-HA6 followed by in vitro crosslinking (dBBr) and Western blotting. The Scc1 Q76C substitution crosslinks efficiently with Smc3 S174C and Smc3 A181C, less efficiently with Smc3 K185C, and not at all with the most distant cysteine (Smc3 E188C). This experiment demonstrates the residue-level specificity that may arise from cysteine–cysteine crosslinking

In order to test the topological function of the cohesin ring (i.e., the actual DNA entrapment), all three interfaces had to be engineered so that the presumed cohesin trimeric ring could be chemically shut and a yeast circular mini-chromosome entrapped within it [1, 5]. For this purpose, it was important to capture the ring formation both inside live cells and after isolating cohesin from yeast cells. Several cysteine residue replacements were

designed and tested for functionality and thiol-specific crosslinking (Fig. 2b–d), especially with regard to the DNA exit gate [1].

The replacement of the endogenous residues follows a few simple rules:

(a) Within a clusterf of evolutionary-related species, the selected residues to be replaced should not be conserved residues. Thus, sequence conservation analysis should be the first thing to be performed after the selection of ~10 (up to ~15 in extensive interaction surfaces) candidate pairs; the species to be included depend on the overall conservation analysis of the protein complex (relevant software to be used could be, e.g., the freeware Jalview or Chimera programs).

(b) Residues to be replaced should never belong to the hydrophobic or polar bonding interaction interface. The best candidates are usually less conserved residues, in register and very close proximity of the interface.

(c) Residues to be replaced may be in α-helices, β-strands, unstructured coils, and connecting loops; however, in any case priority should be given to those residues appearing less conserved.

(d) Ideal replacement residues are the ones resembling the most to the electronegative character of cysteine side chain; thus, serine residues are preferred targets for replacements. However, our experience dictates that potentially all amino acids are replaceable if their side chains do not contribute to the binding interface (*sensu stricto*).

(e) Spatial criteria: the distance between the side chains of the residues to be replaced (as measured from the most distal atom to Cα) should be between ~3 and 6 Å. This distance covers both crosslinkers used here with up to 8 Å to be bridged (*see* **Note 3**).

2. *The genetics of cysteine replacements.* An outline of the workflow is presented in Fig. 1. The protocols describing genetic manipulations can be found in various handbooks on yeast genetics, e.g., in [6]. It is important to test every cysteine replacement for functionality, with the easiest and most reliable test being the assessment of growth and temperature sensitivity following tetrad analysis (e.g., of a replaced single-copy version bearing the cysteine over the deletion of the endogenous protein). Alternatively, one could use CRISPR/Cas9 [7] in order to modify the endogenous ORF; in any case, tetrad analysis should be performed, ideally coupled with case-specific functional analysis of the cysteine alleles.

3. *Choice of crosslinkers.* There are several homobifunctional thiol-specific crosslinkers available, which covalently and irreversibly crosslink two cysteine residues in close proximity. They mainly differ in the length of the spacer, e.g.:

   (a) dBBr (dibromobimane): 5 Å.

   (b) BMOE (bismaleimidoethane): 8 Å.

   (c) BMB (1,4-bismaleimidobutane): 10.9 Å.

   (d) BMH (bismaleimidohexane): 13 Å.

With increasing bridging capacity, however, the possibility of nonspecific crosslinking to endogenous cysteines arises. It is therefore more effective to design cysteine pairs with a distance of <8 Å and use crosslinkers with shorter spacers such as dBBr and BMOE in order to achieve specific crosslinking.

Crosslinking efficiencies both in vivo and in vitro of well-designed cysteine pairs typically reached up to 70%.

References

1. Gligoris TG et al (2014) Closing the cohesin ring: structure and function of its Smc3-kleisin interface. Science 346:963–967
2. Liang C, Stillman B (1997) Persistent initiation of DNA replication and chromatin-bound MCM proteins during the cell cycle in cdc6 mutants. Genes Dev 11:3375–3386
3. Nasmyth K (2011) Cohesin: a catenase with separate entry and exit gates? Nat Cell Biol 13:1170–1177
4. Murayama Y, Uhlmann F (2015) DNA entry into and exit out of the cohesin ring by an interlocking gate mechanism. Cell 163:1628–1640
5. Haering CH, Farcas A-M, Arumugam P, Metson J, Nasmyth K (2008) The cohesin ring concatenates sister DNA molecules. Nature 454:297–301
6. Amberg DC, Burke DJ, Strathern JN (2005) Methods in yeast genetics: a Cold Spring Harbor Laboratory course manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
7. Reider Apel A et al (2017) A Cas9-based toolkit to program gene expression in *Saccharomyces cerevisiae*. Nucleic Acids Res 45:496–508

Chapter 9

Characterizing Intact Macromolecular Complexes Using Native Mass Spectrometry

Elisabetta Boeri Erba, Luca Signor, Mizar F. Oliva, Fabienne Hans, and Carlo Petosa

Abstract

Native mass spectrometry (MS) enables the characterization of macromolecular assemblies with high sensitivity. It can reveal the stoichiometry of subunits as well as their two-dimensional interaction network and provide information regarding the dynamic behavior of macromolecular complexes. Here, we describe the workflow to perform native MS experiments. In addition, we illustrate the quality control analysis of proteins using MS in denaturing conditions.

Key words Macromolecular assemblies, Native mass spectrometry, Stoichiometry, Two-dimensional map of interactions

1 Introduction

Mass spectrometry (MS) is a powerful technique which can measure the mass of molecules with high accuracy, sensitivity, resolution, and speed [1]. Since the pioneering design of the first instrument by physicist and Nobel laureate Sir J. J. Thomson in 1912, MS has seen enormous progress [2]. In particular, the development of electrospray (ESI) [3, 4] and matrix-assisted laser desorption/ionization (MALDI) in the 1980s allowed the analysis of large biomolecules (e.g., proteins, polynucleotides, and glycans) by MS [1, 2]. A few femto- or picomoles are generally sufficient for determining the mass of a macromolecular species with an error of only a few parts-per-millions (ppm), illustrating the high accuracy and sensitivity of this technique. MS analysis can also achieve high resolution and selectivity, as it allows the analyses of a heterogeneous sample.

The original version of this chapter was revised. An erratum to this chapter can be found at https://doi.org/10.1007/978-1-4939-7759-8_32

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_9,

A mass spectrometer consists of three main components: an ion source, a mass analyzer, and a detector [1]. The ionization source generates sample ions, and the mass analyzer separates ions according to their mass-to-charge ratio (m/z). Key examples of analyzers for biomolecules include the time of flight (TOF) [1, 5] and Orbitrap [6, 7]. The detector counts molecular ions derived from the sample and measures their relative abundance. Signals from the detector are transmitted to a computer to generate mass spectra, where m/z values of ions are plotted versus their intensities [1].

When MS is performed under native conditions, noncovalent interactions are preserved, and macromolecular complexes can be studied [8–10]. In particular, one can determine the stoichiometry of subunits, map the interactions between proteins within a complex, and determine the assembly pathway of subunits forming a particle [11] (Fig. 1). Native MS can also provide information regarding the dynamic behavior of macromolecular assemblies. For example, it can assess the presence of distinct oligomers and monitor changes in their equilibria determined by different buffer concentrations or pH values [12]. The subunit composition of intact complexes can be monitored as a function of time. By incubating light and heavy forms of noncovalent complexes (e.g., labeled with ^{2}H or ^{13}C and ^{15}N), the kinetics of subunit exchange can be monitored if affected by the distinct labeling [13].

To carry out a native MS experiment, one needs to (1) perform a buffer change in ammonium acetate (AmAc), (2) use appropriate ionization conditions, and (3) utilize a mass spectrometer modified to detect high m/z ranges [14–16]. Regarding ionization conditions, most native MS experiments utilize nano-ESI [17]. An ESI source generates molecular ions from a sample by applying a high voltage on a flowing liquid and generating an aerosol [18]. Typically, the diameter of nano-ESI capillaries ranges between 1 and 10 μM. Thus, the flow rate is slow (20–30 nL/min), and the required sample amount is in the picomole range. Moreover, the small size of the generated droplets allows the use of aqueous buffers without any heating [19].

Regarding the instruments for native MS experiments, modified quadrupole-TOF (Q-TOF) mass spectrometers are often used because they can analyze very large particles [20] and also can perform collision-induced dissociation experiments (see below). In addition to Q-TOFs, Orbitrap instruments with an extended mass range (EMR) are also utilized because of their high mass resolution (e.g., resolution of 25,000 at m/z 5000) [21–23].

Normally, native MS experiments start with the acquisition of spectra in MS mode (Fig. 2). This means that all generated ions are allowed to reach the detector. The acquired spectra represent an overview of all species present in the sample (aka spectra acquired in MS mode) and allow one to identify the relevant species of interest. These species are then selected, and new spectra are acquired in tandem MS mode (aka MS/MS mode) (Fig. 2). In these

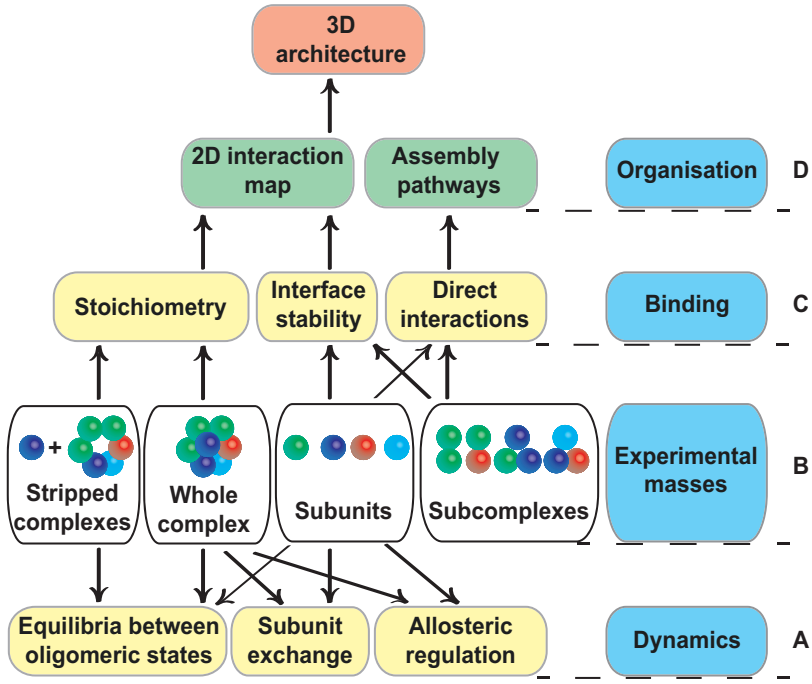


Fig. 1 Structural information about macromolecular assemblies obtained by native MS. The experimental masses of a holo-complex and its subcomplexes and individual subunits (row B) provide access to information such as dynamic properties, binding interactions (rows A and C), and three-dimensional (3D) organization of the macromolecular assembly (row D). Rows B–C: under denaturing conditions, the subunits are chromatographically separated and their masses are determined (see also Fig. 3). Using native MS conditions (e.g., utilizing ammonium acetate buffer), the measured mass of an intact complex (or subcomplex) reveals the stoichiometry of the subunits. MS/MS spectra can further confirm the stoichiometry by dissociating the complex into monomers and stripped complexes (***see*** also Fig. 2). By adding organic solvents, overlapping subcomplexes (e.g., dimers, trimers) are generated. The composition of the different subcomplexes reveals the direct interactions between the subunits [53, 54] and reports on the stability of intermolecular interfaces. Row D: combining all these data allows one to draw an accurate interaction network of a protein complex (2D map of interactions). When the MS data are combined with structural information obtained by EM or SAXS, a 3D architecture of a macromolecular assembly can be modeled. To assess the assembly pathway of macromolecular complex, individual subunits can be mixed in solution, and a mass shift is detected in case a subcomplex is formed. Row A: the dynamic behavior of complexes can be studied by native MS. For example, the presence of different oligomeric states can be monitored over time, and the change in their distribution can be assessed [12]. By incubating light and heavy isoforms of a protein (e.g., labeled with ^{13}C and ^{15}N), the subunit composition of intact complexes can be varied and monitored as a function of time [13, 56, 57]. Moreover, native MS can relatively quantify populations of assemblies containing distinct ligands, providing allosteric information [39]. Reproduced from [11] with permission from Wiley-Blackwell Publications

experiments, only a specific ion population is transmitted to the collision cell, where it collides with molecules of an inert gas such as argon. Collision-induced dissociation (CID) causes the dissociation of a specific complex (e.g., M_n) by ejecting monomers (M_1) and forming the so-called stripped complexes (M_{n-1}) [24]. This allows the confirmation of the stoichiometry and the identification of the subunits at the core and at the periphery of a macromolecu-

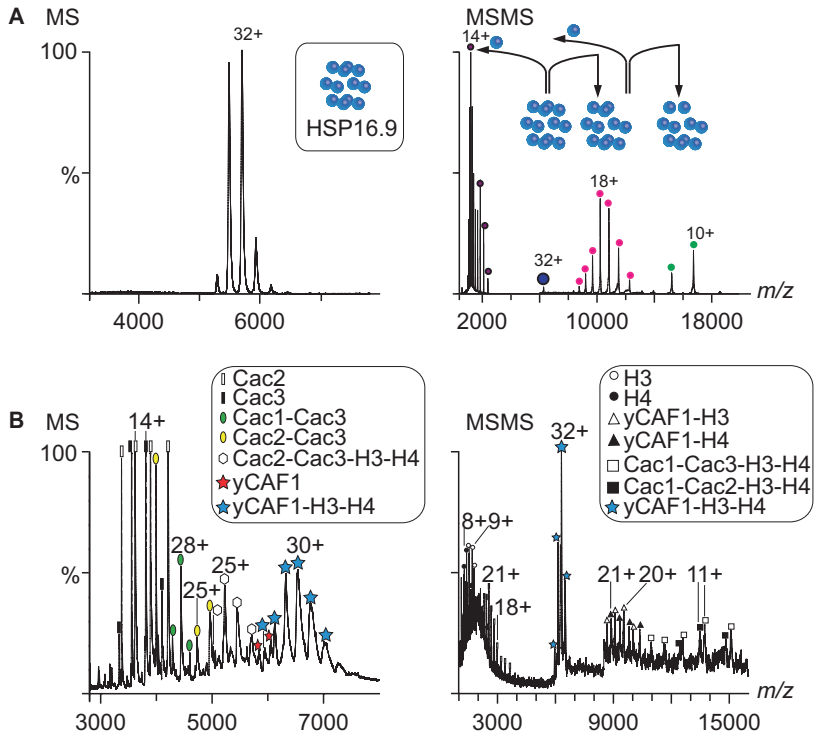


Fig. 2 MS and MS/MS spectra of a homo- and a hetero-complex. (**a**) The dodecameric HSP16.9 protein homocomplex was analyzed by native MS. Left panel: MS spectrum of intact dodecameric assembly. Right panel: MS/MS spectrum of the 32^+ ions. The protein complex (M_n, 12-mer) dissociates by ejecting highly charged monomers at low ***m/z*** (M_1, 1-mer) and generating lowly charged undecamers (M_{n-1}, 11-mer) and decamers (M_{n-2}, 10-mer) at high ***m/z*** (known as stripped complexes). Oligomeric species are indicated by a colored dot, as follows: 12-mer, blue; 11-mer, pink; 10-mer, green; and 1-mer, purple. Reproduced from [58] with permission from ACS. (**b**) The stoichiometry of yCAF1-H3-H4 complex was assessed by native MS. Left panel: MS spectrum of the heteroassembly. In addition to the intact CAF1, trimer in the unbound state (174 kDa) and bound to H3-H4 (201 kDa) (labeled as red and cyan stars, respectively), dimers such as Cac1-Cac3 (green ellipses) and Cac2-Cac3 (yellow ellipses), trimers such as Cac2-Cac3-H3-H4 (white hexagon), and single monomeric proteins (blank and white rectangles) were also detected. Most likely, these subcomplexes were present in solution and not generated in the gas phase. Right panel: MS/MS spectrum of the 32^+ ions confirmed a 1:1:1:1:1 stoichiometry of the yCAF1-H3-H4 complex. Reproduced from [59] with permission from eLife Sciences Publications

lar assembly [11]. In addition to CID, surface-induced dissociation (SID) has been also used to study soluble and membrane protein assemblies [25, 26].

Native MS analyses can be combined with a quality control step performed on proteins using MS in denaturing conditions [27] (Fig. 3). This step establishes the experimental mass of the subunits forming a complex. Thus, the heterogeneity/homogeneity of a particular protein subunit as well as the presence of isoforms, posttranslational modifications, or mutations can be assessed by comparing the experimental mass with the expected mass calcu-

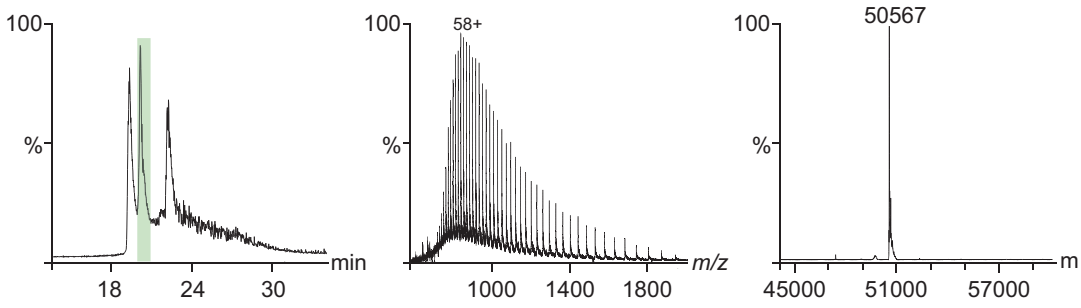


Fig. 3 The yCAF1 complex was analyzed by MS in denaturing conditions. Left panel: the subunits of yCAF1 (i.e., Cac1, Cac2, and Cac3) were chromatographically separated, yielding retention times (RTs) of 19.2, 22.1, and 20.1 min, respectively. Central panel: MS spectrum of Cac3 (RT = 20.1 min). Right panel: deconvoluted spectrum of Cac3. The average mass was calculated from the protein sequence (50,525 Da). The detected mass (50,567 Da) was higher than the expected one; this mass difference could indicate the presence of an acetylation mark (42 Da)

lated from the protein sequence. Carrying out such experiments requires one to (1) separate the protein subunits of a complex by HPLC, (2) use appropriate ionization conditions, and (3) utilize a standard mass spectrometer (e.g., ESI-TOF) to detect m/z ranges up to 3000. In our laboratory, the quality control step is normally performed using an ESI-TOF. In the case of membrane proteins, or glycosylated subunits or those larger than 100 kDa, the quality control is done using a MALDI-TOF, as described in detail elsewhere [28].

To conclude, MS represents a powerful and sensitive tool for obtaining valuable information on macromolecular complexes, including the stoichiometry, the two-dimensional map of interactions, and the assembly pathway.

2 Materials

Prepare all solutions using analytical-grade reagents and ultrapure water. Make the solutions fresh immediately prior to their use and store all reagents at 4 °C. Do not add sodium azide to reagents.

2.1 Native MS Experiments

2.1.1 Buffers and Reagents

1. Acetic acid (e.g., glacial, ≥99.85%).
2. Acetate salts of metal cations (if required for the stability of the noncovalent assemblies).
3. Acetonitrile [high-performance liquid chromatography (HPLC) grade].
4. 7.5 M ammonium acetate solution (AmAc).
5. Ammonium hydroxide solution ~10% in H_2O, for HPLC.
6. Cesium iodide (CsI) at 6 mg/mL in 50% 2-propanol.

7. DL-Dithiothreitol (DTT).
8. Ethanol (EtOH, 96.9%).
9. Formic acid (FA, HPLC grade).
10. 2-Propanol (HPLC grade).
11. Argon (purity: N50, i.e., 99.999%).
12. Nitrogen (N50).
13. Various nonionic detergents (e.g., DDM, DM) and nanodisks (in the case of membrane protein complex analyses) [29–31].

2.1.2 Equipment, Devices, Tools, and Software

1. Metal-coated capillaries (Thermo Fisher Scientific; New Objective, Inc.; or other suppliers).
2. Borosilicate glass capillaries, e.g., 1.0 mm outside diameter (OD) × 0.78 mm internal diameter (ID) or 1.2 mm OD × 0.69 mm ID, with an inner filament (e.g., Clark Electromedical Instruments) (in case you prepare coated needles in-house).
3. Capillary puller with a heated filament (e.g., Sutter Instrument Co., Model P-97) (for pulling capillaries in-house).
4. Sputter coater (e.g., Quorum Technologies, Polaron Range SC7680 coater) (for coating needles in-house).
5. Glass lidded dish with a diameter of 9 cm (for preparing needles in-house).
6. Double-sided adhesive tape (for preparing needles in-house).
7. Conductive elastomer for nanospray probe (Waters).
8. AA tweezers (Dumont).
9. Ceramic cutter.
10. 0.5–20 μL GELoader® tips (Eppendorf).
11. Optical stereomicroscope (e.g., Wild, M3Z).
12. Micro Bio-Spin® 6 chromatography columns (Bio-Rad).
13. Gel filtration columns (e.g., Superdex®, GE Healthcare).
14. Vivaspin® 0.5 mL concentrators (various MWCO exclusion limits, Sartorius).
15. Amicon Ultra® 0.5 mL centrifugal filters (Millipore).
16. Refrigerated centrifuge (e.g., with a 24-place rotor used for 1.5 and 2.0 mL tubes; Heraeus Fresco®, Thermo Scientific).
17. UV spectrophotometer.
18. A mass spectrometer (e.g., nano-electrospray ionization-quadrupole time of flight (nano-ESI-Q-TOF) instrument [14, 32] or Orbitrap [33]).
19. MassLynx® 4.0 software (Waters Corporation, Manchester, UK).
20. Massign® software package [34].

2.2 MS Experiments in Denaturing Conditions

2.2.1 Buffers and Reagents

1. Water (HPLC grade).
2. Acetonitrile (ACN), purity ≥99.9% (e.g., LC-MS Chromasolv, Fluka).
3. Methanol (HPLC).
4. Isopropanol (HPLC).
5. Mobile phase A: 0.03% trifluoroacetic acid (TFA).
6. Mobile phase B: 95% ACN, 0.03% TFA.
7. Calibration mixture (e.g., ESI-L Low Concentration Tuning Mix, Agilent Technologies. It contains ten different molecules whose *m/z* signals range from 118.08 to 2721.89).
8. Myoglobin (e.g., from horse heart, Waters API Test Solution Kit).

2.2.2 Equipment, Devices, Tools, and Software

1. Glass vials with screws.
2. C8 reverse phase chromatography cartridge for protein desalting (e.g., Agilent Technologies, Zorbax 300SB-C8, 5 μm, 300 μm ID × 5 mm length).
3. C8 reverse phase chromatography column for protein separation (e.g., Agilent Technologies, Zorbax 300SB-C8, 3.5 μm, 1.0 μm ID × 75 mm length).
4. Binary HPLC chromatography pump system (e.g., Agilent Technologies, 1100).
5. ESI-TOF mass spectrometer (e.g., Agilent Technologies, 6210).
6. Specific software to deconvolute the raw MS spectra (e.g., MassHunter Quantitative Analysis®, Agilent Technologies).
7. Software to calculate the mass of the proteins from their amino acid sequences (e.g., General Protein/Mass Analysis for Windows, GPMAW®, Lighthouse Data).

3 Methods

3.1 Native MS Experiments

3.1.1 Sample Buffer Exchange

1. The sample should be desalted and buffer exchanged into AmAc. This can be done using different approaches. For example, chromatography can be performed using a gel filtration or desalting column pre-equilibrated in AmAc. The sample can also be subjected to repeated cycles of concentration and dilution with AmAc using an ultrafiltration device (e.g., Vivaspin® concentrators or Amicon Ultra® centrifugal filters). Extensive dialysis against a solution of AmAc is also possible. However, in our experience, this method has been less effective than ultrafiltration.

2. The concentration of AmAc chosen should be compatible with the stability of the complex being studied and may need to be empirically determined. Typical concentrations range between 50 and 350 mM AmAc. After the buffer exchange, the concentration of the purified macromolecular assembly should be assessed using UV absorbance at 280 nm and appropriate extinction coefficients (*see* **Notes 1–12**).

3.1.2 Instrumentation and Instrumental Settings

1. A mass spectrometer modified to detect a high m/z range is required (e.g., up to 60,000 m/z). A nano-ESI-quadrupole time of flight (nano-ESI-Q-TOF) [14, 32] or a nano-ESI-Orbitrap can be used [33].
2. In the case of a nano-ESI-Q-TOF instrument, the following instrumental parameters can be utilized: capillary voltage = 1.2–1.3 kV, cone potential = 40 V, RF lens-1 potential = 40 V, RF lens-2 potential = 1 V, aperture-1 potential = 0 V, collision energy = 30 V, and microchannel plate (MCP) = 1900 V. Pressure in the transfer region between the ionization source and analyzer $\approx 10^{-3}$ mbar, collision cell pressure $\approx 10^{-2}$ mbar, and TOF pressure $\approx 8 \times 10^{-6}$ mbar. Starting range of acquisition is between 1000 m/z and 8000 m/z for which the TOF pusher pulse is 60 μs. For CID experiments, the collision voltage can be increased up to 400 V, and the m/z range of acquisition is modified according to the m/z of the parent ions (*see* **Notes 13** and **14**).

3.1.3 Opening and Loading of a Nano-ESI Needle

1. Using a tweezer, take a coated needle from its lidded box, and cut 1–1.5 cm from the end of the capillary (i.e., its bottom), using a ceramic cutter.
2. Using a GELoader® tip, load 2–4 μL of the buffer-exchanged sample into the needle.
3. Insert the capillary into the holder of the ionization source, and tighten the holder screw to block the needle in its position.
4. Place the capillary and its holder under an optical microscope stage, and open the tip of the needle using an AA tweezer (*see* **Notes 15–22**).

3.1.4 Instrument Calibration

1. The TOF of the instrument should be calibrated prior to each data acquisition, using an appropriate calibrant. For instance, CsI can be utilized prior to native MS experiments. A solution containing 6 mg/mL CsI in 50% isopropanol allows a calibration up to 8000 m/z. The calibration solution should be sprayed using a nano-ESI source, and a calibrant spectrum should be recorded (*see* **Note 23**).

2. Specific features of the software controlling the instrument (i.e., MassLynx) allow the calibration of the TOF. In simple terms, the software compares the measured m/z values of the calibrant with the theoretical ones. Afterward, the software calibrates the TOF to match the experimental values with those expected.
3. The experimental error expressed in parts-per-million (ppm) is normally between 5 and 30 ppm. The calibration of the TOF takes place differently according to the instrument you use. A full description of the instrument calibration should be provided by the vendor.

3.1.5 Acquisition of Sample Spectra in MS Mode

1. The capillary containing the sample is placed in its holder in front of the orifice which is the entrance of the instrument. Normally a voltage of 1.2–1.3 kV is applied and the nano-ESI process starts.
2. A specific button allows the acquisition of the spectra in MS mode. The initial instrumental setting should be adjusted to optimize the signal of the sample (*see* **Notes 24–30**).

3.1.6 Acquisition of Sample Spectra in Tandem MS Mode

1. The instrument software allows the use of a quadrupole as a filter for ions. In this setting, only the ion population of the analyzed macromolecular complex with a certain m/z (aka parent ion) can pass through the quadrupole and reach the collision chamber where the CID of the assembly takes place.
2. To dissociate a macromolecular assembly, you should increase the voltage of the collision cell and its pressure. Normally, we set the pressure to $\approx 2 \times 10^{-2}$ mbar and increase the voltage stepwise by 20–30 V until the signal of the parent ion is low and the peaks of the product ions are intense (*see* **Note 31**).

3.1.7 Data Analysis

1. Spectra provide m/z values of the species present in a sample. Data analysis should provide the masses (M) of these species and the standard deviation of the measurements (*see* **Notes 32–34**).

3.2 MS Experiments in Denaturing Conditions

1. The quality control performed using MS in denaturing conditions is very useful because in our experience the experimental mass is often different from the expected one calculated from the protein sequence. For example, this discrepancy can be due to posttranslational modifications such as phosphorylation and acetylation or to an unexpected protein truncation (Fig. 3).

3.2.1 Control and Preparation of the HPLC System

1. Before each series of measurements, the HPLC system is purged first with mobile phase A and then with mobile phase B, for the same amount of time (6 min) at the same flow rate (2.5 mL/min).
2. After the measurements, the RP-C8 cartridge and columns are equilibrated with 5% mobile phase B at a flow rate of 50 μL/min for 30 min before starting the LC run.

3.2.2 Instrument Settings

1. Data acquisition is carried out in positive ion mode, and mass spectra are recorded in the 300–3200 m/z range.
2. The following experimental settings are utilized: ESI source temperature is set at 300 °C. Nitrogen is used as drying gas (with a flow rate of 7 L/min) and as nebulizer gas (using a pressure of 10 psi). The capillary needle voltage is 4 kV. Voltages in the first part of the instrument are set as follows: the fragmentor voltage is 250 V and the skimmer one is 60 V. Spectra acquisition rate is a spectrum/s. Instrument pressure values are typically 2.33 Torr (rough vacuum) and 4.6×10^{-7} Torr (TOF vacuum) (*see* **Note 35**).

3.2.3 Calibration of the TOF

1. The calibration of the TOF takes place differently according to the instrument you use. In the case of the 6210 instrument, the procedure is semiautomated. We connect the system to the calibrant bottle; we choose the tune setting and the instrument starts its calibration.
2. In the case of the calibration molecules, the mass error normally is lower than 1 ppm. Refer to a detailed description of the TOF calibration provided by the instrument vendor.

3.2.4 Dilution of Samples and Typing Their List in Software of the HPLC System

1. Just before their analyses, the samples are diluted in 0.03% TFA to obtain a concentration of 5 μM and a volume of 20 μl. We use glass vials with screws as containers. These vials are placed on a sample loader refrigerated at 10 °C.
2. A list of samples should be typed into the appropriate software (e.g., MassHunter). The list contains the sample names, the volume injected into the mass spectrometer, and the data storage path. Details of this procedure vary according to the HPLC system. Regarding the sample list, we first analyze a solution of 5 μM of myosin to assess the performance of the HPLC system. Then, we alternate the analysis of a blank (containing only 0.03% TFA) with a sample run to avoid any "carry-over" and contamination problems.

3.2.5 LC-MS Analysis

1. Normally 4 μL of each sample (i.e., circa 20 pmol of protein) are injected into the HPLC system. The injected sample is first trapped and desalted on the RP-C8 cartridge for 3 min at a flow rate of 50 μL/min using 100% mobile phase A. Afterward, proteins are separated on the RP-C8 column using a linear gradient from 5 to 95% of mobile phase B for 15 min and subjected to ESI prior to the TOF detection of their m/z signals.
2. Before starting a new sample run, the RP-C8 cartridge is re-equilibrated for 10 min with 100% mobile phase A at a flow rate of 50 μL/min. Using the same flow for the same time, the RP-C8 column is equilibrated with 5% mobile phase B. Overall, each LC-MS run takes 30 min.

3.2.6 Data Analysis

1. The m/z values are utilized to calculate the M values. Normally a specific software is utilized to perform these calculations (*see* **Notes 36–38**).

4 Notes

1. Before MS experiments, macromolecular assemblies are purified using specific approaches such as epitope tagging and affinity purification techniques. Normally the identity of the subunits forming a complex is already known before starting a native MS experiment. For instance, proteins can be identified by MS-based proteomics [35].
2. Regarding purification buffers, we try to avoid β-mercaptoethanol and to use instead DDT or tris(2-carboxyethyl)phosphine. The use of the protease inhibitor 4-(2-aminoethyl)-benzenesulfonyl fluoride (aka Pefabloc®) can be problematic. In contrast, other protease inhibitors are acceptable. When possible, avoid phosphate-buffered saline (PBS), 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), and glycerol. A recommended buffer is tris(hydroxymethyl)aminomethane (Tris).
3. Prior to native MS, macromolecular assemblies are buffer exchanged into AmAc. This exchange can be performed using different approaches such as gel filtration. This choice is related to the noncovalent complex you aim to analyze and its concentration and volume. In our experience, Micro Bio-Spin® 6 chromatography columns, Vivaspin® concentrators, and Amicon Ultra® centrifugal filters can be used when the sample volume is limited (i.e., 20–25 μL). Based on our empirical observation, we consider dialysis as the least effective approach for performing buffer exchange prior to native MS experiments.
4. When using Micro Bio-Spin® 6 chromatography columns, invert the column sharply four times to resuspend the settled gel matrix. Remove the lid and place the column in a 2 mL microcentrifuge tube. Allow the storage buffer to drain by gravity from the gel bed for 2 min. Discard the drained buffer, and then place the column back into the microcentrifuge tube. Load 500 μL of AmAc on the column and centrifuge it at 1000 × *g* for 2 min. Repeat this wash three times and every time discard the buffer. Place the column in a clean 1.5 or 2.0 mL microcentrifuge tube. Carefully pipette the sample (whose volume is between 20 and 75 μL) on the center of the column. Centrifuge the column at 1000 × *g* for 4 min. The buffer-exchanged sample is collected in the microcentrifuge tube.

5. When utilizing Vivaspin® concentrators and Amicon Ultra® centrifugal filters, pipette 500 µL of AmAc, and centrifuge them at 9000 × *g* for 3 min. Repeat this wash three times and every time discard the buffer. Then load 200 µL of AmAc into the concentrator, next pipette the sample, and finally add a volume of buffer needed to reach a total volume of 500 µL. Centrifuge the filters at 9000 × *g* for several minutes until the required volume is obtained. Repeat this step at least three times. The buffer-exchanged sample is present in the reservoir of the concentrators (and not in the microcentrifuge tube).
6. By using low-molecular-weight cutoff (MWCO) concentrators (such as 3–10 kDa MWCO), it can take a long time (even hours) to obtain the desired volume (such as 25 µL).
7. Regarding the appropriate concentration of AmAc for buffer exchange, this is normally 1.5 times higher than the sum of salt concentrations present in the original buffer. For example, if the initial sample buffer contains 20 mM Tris and 150 mM NaCl, then 250 mM AmAc can be used. Normally the pH of AmAc is 7. If the analyzed complex is more stable in a buffer with a specific pH, the pH of AmAc solution can be adjusted using volatile reagents such as ammonia (not NaOH) or acetic acid (not hydrochloric acid).
8. Normally we use AmAc. However, other buffers have been used in the literature such as ammonium bicarbonate [36, 37], triethylammonium bicarbonate [37], triethylammonium acetate [38, 39], and ethylenediammonium diacetate [39].
9. Salts such as NaCl and KCl or related acids or bases (HCl and NaOH) should be avoided or eliminated before any ESI-MS experiments. These nonvolatile ions affect the efficiency of droplet formation and evaporation, suppressing the ionization of the analytes [40, 41].
10. Some complexes may require the presence of reducing agents (e.g., DTT) or of specific metal ions. These should be added to the AmAc solution at as low a concentration as possible. In the case of DTT, a concentration of 1–2 mM is acceptable.
11. A macromolecular complex may dissociate during buffer exchange. To avoid this problem, you may switch to a different approach for buffer exchange. For example, if you observe dissociation of the complex using gel filtration, you might try ultrafiltration to exchange the buffer. You may also increase or decrease the concentration of AmAc or modify the pH of the buffer. Another option is to perform mild cross-linking of the complex such as the graphix approach [42–44].
12. In the case of a membrane macromolecular assembly, you may add a detergent, which is compatible with MS, such as n-decyl-β-D-

maltopyranoside (DM) or n-dodecyl beta-D-maltoside (DDM) [29, 45, 46]. Bicelles and nanodisks have been also used [30].

13. Native MS investigations of macromolecular assemblies are performed using mass spectrometers under high vacuum [14]. Therefore, the detection of solution-phase species takes place in the gas phase. Thus, the relative abundance of the different species in the MS spectra is semiquantitative because there are differences in ionization efficiency, transmission, and detection between the various macromolecular complexes and subcomplexes present in a sample [15].
14. Since native MS analyses take place under high vacuum, distinct types of interactions are differently affected by the gas phase. Electrostatic interactions may be stronger in the gas phase than in solution, while hydrophobic interfaces may be weakened [15, 47]. In our experience, there are often enough interactions to maintain an assembly intact during the gas phase measurements.
15. To perform nano-ESI for native MS experiments, 1–20 μM of sample are required as a final concentration. Each sample injection requires 1–2 μL, and for a full characterization of a complex, at least 5–20 μL are required.
16. The flow of nano-ESI is approximately 20–50 nL/min and allows an excellent sensitivity without the need for a desolvation gas or for heating of the source.
17. When possible during native MS experiments, load 1–2 μL of 5 μM of the investigated complex into the nano-ESI capillary. This amount of sample normally provides a MS signal of good intensity and also lowers the possibility of nonspecific associations. These interactions may appear when a high sample concentration (such as 20–30 μM) is used [15]. They can be identified by the appearance in the spectrum of a sequence of oligomers, whose relative intensities are inversely proportional to their m/z ranges (i.e., the higher their m/z, the lower their intensity). The nonspecific interactions can be lowered by further diluting the sample.
18. Normally we purchase metal-coated needles. However, it is possible to prepare the capillaries in-house [15, 16]. To do this, use scissors to cut two 1-cm-wide bands of a double-sided adhesive tape, and attach these to the inside of a lidded glass dish. Program the capillary puller. This setup depends on the puller and on the heated filament. You should optimize heating time and temperature until the shape of the tip capillary is satisfactory. Pull a needle and place it into the lidded dish using tweezers. Repeat this step until the glass dish contains circa approximately 30 needles. Using a sputter coater, coat the needles with a metal such as gold. The coater vendor

should provide a detailed description of the coating procedure.

19. Shortening the length of a needle by cutting its bottom can facilitate the flow of the sample through the capillary because the sample flows for a shorter distance before reaching the needle tip and being ionized.
20. Cutting the tip of a needle is a critical step affecting the flow of the sample and the ionization process. Working with a long tip may cause a low flow rate but allows further trimming, in the case of a clogged needle. The internal diameter of a cut capillary is around 1–10 μm.
21. A conductive elastomer ferrule makes the electrical connection between the nano-ESI probe and the metal-coated capillary. After being used for some months, the ferrule becomes enlarged and should be substituted.
22. Normally we use metal-coated needles to perform nano-ESI. It is also possible to use chip-based ESI (e.g., TriVersa NanoMate® from Advion) [48], which allows unmanned operations by loading several samples consecutively in an automated manner (e.g., during overnight analyses).
23. CsI forms large cluster ions whose formula can be written as $[(CsI)_n\ Cs]^+$. For calibrating a m/z range larger than 8000, a concentration of CsI higher than 6 mg/mL in 50% isopropanol should be used (e.g., 10–20 mg/mL). We spray the CsI solution for 1 or 2 min because the infusion of this calibrant can easily contaminate the optics of the mass spectrometer, decreasing the instrument sensitivity.
24. We normally perform the analysis of protein and protein complexes in positive mode. This means that only positive ions are detected. However, examples of detection in negative mode have been reported in the literature [49–51]. We analyze DNA and RNA both in positive and negative modes.
25. To optimize the nano-ESI signal during native MS experiments, you can change (1) the distance between the needle and the entrance cone, (2) the voltage, (3) the backing pressure of nitrogen, (4) the pressure in the transfer region between the ionization source and analyzer, (5) the pressure inside the collision cell, and (6) the collision voltage. You can also optimize (7) the acquisition setting and (8) the range of acquisition and (9) tune the transmission quadrupole to spend more time on the m/z range of interest.
26. Normally, we utilize a voltage of 1.2 kV. However, in some cases, the capillary voltage may be increased up to 1.8 kV. Higher voltage could damage the metal coating of the needle, making it less conductive.

27. A backing pressure of nitrogen ranging between 0 and 2 bar is normally applied to start the sample flow in the capillary. This pressure is often reduced once the sample signal is stable. However, a backing pressure may be constantly required to obtain a stable signal for certain samples. When you use the backing pressure, a sample droplet should first emerge from the tip of the needle before applying the capillary voltage. After applying the voltage, the ESI process starts and no droplet can be seen.
28. The pressure in the transfer region, the one inside the collision cell, and the collision voltage should be carefully optimized for two reasons. (1) These parameters influence the degree of solvation of the macromolecular complex. They should be set to strip away residual buffer and water bound to the assembly but without dissociating the complex. If optimized, they generate spectra with ions showing narrow peak widths. (2) The three parameters also affect the focusing of ions and their transmission through the mass spectrometer [52]. When optimized, ions are transmitted with greater efficiency to generate spectra with higher intensities.
29. In case you start a native MS experiment and no ions are detected, the capillary may be blocked, and you should further cut the capillary tip. If a capillary is repeatedly clogged, you should dilute the sample or change the buffer conditions. For example, repeat the buffer exchange steps utilizing a different AmAc concentration.
30. It is possible to add a low percentage of organic solvents (e.g., EtOH, ACN) to AmAc to selectively break the weaker subunit interactions and generate overlapping subcomplexes (e.g., dimers, trimers). The composition of the different subcomplexes reveals the direct interactions between the subunits [53, 54] and provides information about the stability of intermolecular interfaces.
31. When MS/MS experiments are performed, it may be necessary to increase the pressure of the collision cell to raise the signal of the stripped complexes. A higher pressure will enhance the focusing of ions, increasing the ion population reaching the detector.
32. Regarding data analysis, its goal is to determine the masses (M) of species present in a spectrum. In simple terms, two neighboring m/z values (M/z_1 and M/z_2) are determined experimentally (x and y), and two equations are written ($M/z_1 = x$ and $M/z_2 = y$). Since $z_1 = z_2 - 1$, the equations are solved to determine M, z_1, and z_2.
33. In our laboratory, we use a software called MassLynx (Waters) to calculate M from m/z values obtained by native MS

experiments. Specifically, the program takes several combinations of neighboring m/z values to determine distinct M of a macromolecule. Using these measurements, a mean value of M and its standard deviation are calculated. The M values are normally determined from m/z values corresponding to the left edge of the peaks. These values provide the "least-adducted" M of the noncovalent complexes [55]. Indeed, the M of most macromolecular assemblies is higher than the one calculated. This is ascribable to buffer and water molecules still bound to the complex in the gas phase [15].

34. Once per week, we perform the ballast of the rugged mechanical oil-sealed pump (e.g., EM 18) by opening its gas ballast valve. This allows the pump oil to flow back into the appropriate reservoir. Once per month, we clean the entrance orifice and the cone of the mass spectrometer. After disassembling this first part of the instrument, we sweep the orifice with EtOH using a cotton swab. We sonicate the cone first in a solution containing 50% ACN and 10% FA for 10 min and then in water for the same amount of time. We leave the cone dry under the extractor hood for 10 min. Once per year, engineers of a specialized company perform the preventive maintenance of the instrument by cleaning each part of the mass spectrometer.
35. Normally we use a metal capillary to perform ESI-MS in denaturing conditions. We utilize 20 μL of 5 μM of a protein. The flow of ESI is approximately 1 μL/min and requires a desolvation gas (e.g., nitrogen) and heating of the source at 60 °C.
36. We use a software called MassHunter (Agilent Technologies, Bioconfirm v.B.07.00) to calculate M from m/z values obtained by MS experiments in denaturing conditions. This software works in a similar way to MassLynx. You can choose several settings, such as m/z and M ranges to optimize the M calculation.
37. The GPMAW software (Lighthouse Data, v.7.00b2) is utilized to calculate the theoretical masses for protein sequences. It can calculate the average mass of proteins and their monoisotopic mass. The average mass is based on the weighted average of the atomic masses of the different isotopes of each element (e.g., C = 12.011, H = 1.00794, O = 15.9994) present in a biomolecule [1]. The monoisotopic mass is calculated using the exact mass of the most abundant isotopes of each constituent element (e.g., C_{12} = 12.000000, C_{13} = 13.003355). For instance, the average mass of tryptophan is 186.213 Da, and its monoisotopic mass is 186.07932 Da. Generally, we utilize average mass of proteins because our instruments do not have sufficient resolution to resolve isotopic peaks.

38. Regarding the maintenance of the instrument, at the end of each use, the ESI needle should be rinsed several times with 50% methanol to avoid its clogging. Once per week the ESI source should be cleaned. To do that, remove the spray shield and the capillary cap, and sonicate them with 50% isopropanol for half an hour. After drying with nitrogen, these parts must be reassembled on the ESI source. The ballast of the rough pump is performed once per week.

Acknowledgment

We thank Paul Sauer, Jennifer Timm and Daniel Panne for providing the yeast CAF1 complex. We thank the members of the Viral Infection and Cancer Group at the IBS for the helpful discussion. This work used the mass spectrometry platform of the Grenoble Instruct Centre (ISBG; UMS 3518 CNRS-CEA-UJF-EMBL) with support from FRISBI (ANR-10-INSB-05-02) and GRAL (ANR-10-LABX-49-01) within the Grenoble Partnership for Structural Biology (PSB). It was financially supported by the French Infrastructure for Integrated Structural Biology Initiative and by the French National Centre for Scientific Research (CNRS).

References

1. de Hoffmann E, Stroobant V (2007) Mass spectrometry: principles and applications, 3rd edn. Wiley, New York, NY, p 502
2. Lossl P, van de Waterbeemd M, Heck AJ (2016) The diverse and expanding role of mass spectrometry in structural and molecular biology. EMBO J 35:2634–2657
3. Kebarle P, Verkerk UH (2009) Electrospray: from ions in solution to ions in the gas phase, what we know now. Mass Spectrom Rev 28:898–917
4. Konermann L, Ahadi E, Rodriguez AD, Vahidi S (2013) Unraveling the mechanism of electrospray ionization. Anal Chem 85:2–9
5. Cotter RJ (1999) Peer reviewed: the new time-of-flight mass spectrometry. Anal Chem 71:445A–451A
6. Makarov A (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal Chem 72:1156–1162
7. Perry RH, Cooks RG, Noll RJ (2008) Orbitrap mass spectrometry: instrumentation, ion motion and applications. Mass Spectrom Rev 27:661–699
8. Sharon M (2013) Biochemistry. Structural MS pulls its weight. Science 340:1059–1060
9. Leney AC, Heck AJ (2017) Native mass spectrometry: what is in the name? J Am Soc Mass Spectrom 28:5–13
10. Marx V (2016) Proteomics: taking on protein complexes. Nat Methods 13:721–727
11. Boeri Erba E, Petosa C (2015) The emerging role of native mass spectrometry in characterizing the structure and dynamics of macromolecular complexes. Protein Sci 24:1176–1192
12. Boeri Erba E, Barylyuk K, Yang Y, Zenobi R (2011) Quantifying protein-protein interactions within noncovalent complexes using electrospray ionization mass spectrometry. Anal Chem 83:9251–9259
13. Yee AW, Moulin M, Breteau N et al (2016) Impact of deuteration on the assembly kinetics of transthyretin monitored by native mass spectrometry and implications for amyloidoses. Angew Chem Int Ed Engl 55:9292–9296
14. van den Heuvel RH, van Duijn E, Mazon H et al (2006) Improving the performance of a quadrupole time-of-flight instrument for macromolecular mass spectrometry. Anal Chem 78:7473–7483
15. Hernandez H, Robinson CV (2007) Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. Nat Protoc 2:715–726

16. Kirshenbaum N, Michaelevski I, Sharon M (2010) Analyzing large protein complexes by structural mass spectrometry. J Vis Exp
17. Wilm M, Mann M (1996) Analytical properties of the nanoelectrospray ion source. Anal Chem 68:1–8
18. Fenn JB, Mann M, Meng CK et al (1989) Electrospray ionization for mass spectrometry of large biomolecules. Science 246:64–71
19. Benesch JL, Ruotolo BT, Simmons DA, Robinson CV (2007) Protein complexes in the gas phase: technology for structural genomics and proteomics. Chem Rev 107:3544–3567
20. Snijder J, Rose RJ, Veesler D et al (2013) Studying 18 MDa virus assemblies with native mass spectrometry. Angew Chem Int Ed Engl 52:4020–4023
21. Rose RJ, Damoc E, Denisov E et al (2012) High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. Nat Methods 9:1084–1086
22. van de Waterbeemd M, Snijder J, Tsvetkova IB et al (2016) Examining the heterogeneous genome content of multipartite viruses BMV and CCMV by native mass spectrometry. J Am Soc Mass Spectrom 27:1000–1009
23. Yang Y, Wang G, Song T et al (2017) Resolving the micro-heterogeneity and structural integrity of monoclonal antibodies by hybrid mass spectrometric approaches. MAbs 9:1–8
24. Kondrat FD, Struwe WB, Benesch JL (2015) Native mass spectrometry: towards high-throughput structural proteomics. Methods Mol Biol 1261:349–371
25. Quintyn RS, Zhou M, Yan J, Wysocki VH (2015) Surface-induced dissociation mass spectra as a tool for distinguishing different structural forms of gas-phase multimeric protein complexes. Anal Chem 87:11879–11886
26. Harvey SR, Liu Y, Liu W et al (2017) Surface induced dissociation as a tool to study membrane protein complexes. Chem Commun (Camb) 53:3106–3109
27. Rozen S, Tieri A, Ridner G et al (2013) Exposing the subunit diversity within protein complexes: a mass spectrometry approach. Methods 59:270–277
28. Signor L, Boeri Erba E (2013) Matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) mass spectrometric analysis of intact proteins larger than 100 kDa. J Vis Exp
29. Laganowsky A, Reading E, Hopper JT, Robinson CV (2013) Mass spectrometry of intact membrane protein complexes. Nat Protoc 8:639–651
30. Hopper JT, Yu YT, Li D et al (2013) Detergent-free mass spectrometry of membrane protein complexes. Nat Methods 10:1206–1208
31. Leney AC, McMorran LM, Radford SE, Ashcroft AE (2012) Amphipathic polymers enable the study of functional membrane proteins in the gas phase. Anal Chem 84:9841–9847
32. Sobott F, Hernandez H, McCammon MG et al (2002) A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. Anal Chem 74:1402–1407
33. Snijder J, Heck AJ (2014) Analytical approaches for size and mass analysis of large protein assemblies. Annu Rev Anal Chem (Palo Alto, Calif) 7:43–64
34. Morgner N, Robinson CV (2012) Massign: an assignment strategy for maximizing information from the mass spectra of heterogeneous protein assemblies. Anal Chem 84:2939–2948
35. Aebersold R, Mann M (2016) Mass-spectrometric exploration of proteome structure and function. Nature 537:347–355
36. Loo JA (2000) Electrospray ionization mass spectrometry: a technology for studying non-covalent macromolecular complexes. Int J Mass Spectrom Ion Process 200:175–186
37. Heck AJ, Van Den Heuvel RH (2004) Investigation of intact protein complexes by mass spectrometry. Mass Spectrom Rev 23:368–389
38. Pagel K, Hyung SJ, Ruotolo BT, Robinson CV (2010) Alternate dissociation pathways identified in charge-reduced protein complex ions. Anal Chem 82:5363–5372
39. Dyachenko A, Gruber R, Shimon L et al (2013) Allosteric mechanisms can be distinguished using structural mass spectrometry. Proc Natl Acad Sci U S A 110:7235–7239
40. King R, Bonfiglio R, Fernandez-Metzler C et al (2000) Mechanistic investigation of ionization suppression in electrospray ionization. J Am Soc Mass Spectrom 11:942–950
41. Annesley TM (2003) Ion suppression in mass spectrometry. Clin Chem 49:1041–1044
42. Kastner B, Fischer N, Golas MM et al (2008) GraFix: sample preparation for single-particle electron cryomicroscopy. Nat Methods 5:53–55
43. Boeri Erba E, Klein PA, Signor L (2015) Combining a NHS ester and glutaraldehyde improves crosslinking prior to MALDI MS analysis of intact protein complexes. J Mass Spectrom 50:1114–1119

44. Caillat C, Macheboeuf P, Wu Y et al (2015) Asymmetric ring structure of Vps4 required for ESCRT-III disassembly. Nat Commun 6:8781
45. Barrera NP, Isaacson SC, Zhou M et al (2009) Mass spectrometry of membrane transporters reveals subunit stoichiometry and interactions. Nat Methods 6:585–587
46. Barrera NP, Robinson CV (2011) Advances in the mass spectrometry of membrane proteins: from individual proteins to intact complexes. Annu Rev Biochem 80:247–271
47. Ruotolo BT, Robinson CV (2006) Aspects of native proteins are retained in vacuum. Curr Opin Chem Biol 10:402–408
48. Painter AJ, Jaya N, Basha E et al (2008) Real-time monitoring of protein complexes reveals their quaternary organization and dynamics. Chem Biol 15:246–253
49. Kelly MA, Vestling MM, Fenselau C, Smith PB (1992) Electrospray analysis of proteins: a comparison of positive-ion and negative-ion mass spectra at high and low pH. Org Mass Spectrom 27:1143–1147
50. Madler S, Barylyuk K, Boeri Erba E et al (2012) Compelling advantages of negative ion mode detection in high-mass MALDI-MS for homomeric protein complexes. J Am Soc Mass Spectrom 23:213–224
51. Allen SJ, Schwartz AM, Bush MF (2013) Effects of polarity on the structures and charge states of native-like proteins and protein complexes in the gas phase. Anal Chem 85:12055–12061
52. Chernushevich IV, Thomson BA (2004) Collisional cooling of large ions in electrospray mass spectrometry. Anal Chem 76:1754–1760
53. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA (2008) Assembly reflects evolution of protein complexes. Nature 453:1262–1265
54. Ahnert SE, Marsh JA, Hernandez H et al (2015) Principles of assembly reveal a periodic table of protein complexes. Science 350:aaa2245
55. McKay AR, Ruotolo BT, Ilag LL, Robinson CV (2006) Mass measurements of increased accuracy resolve heterogeneous populations of intact ribosomes. J Am Chem Soc 128:11433–11442
56. Keetch CA, Bromley EH, McCammon MG et al (2005) L55P transthyretin accelerates subunit exchange and leads to rapid formation of hybrid tetramers. J Biol Chem 280:41667–41674
57. Chevreux G, Atmanene C, Lopez P et al (2011) Monitoring the dynamics of monomer exchange using electrospray mass spectrometry: the case of the dimeric glucosamine-6-phosphate synthase. J Am Soc Mass Spectrom 22:431–439
58. Boeri Erba E, Ruotolo BT, Barsky D, Robinson CV (2010) Ion mobility-mass spectrometry reveals the influence of subunit packing and charge on the dissociation of multiprotein complexes. Anal Chem 82:9702–9710
59. Sauer PV, Timm J, Liu D et al (2017) Insights into the molecular architecture and histone H3-H4 deposition mechanism of yeast Chromatin assembly factor 1. elife 6:e23474

Chapter 10

Hydrogen-Deuterium Exchange Mass Spectrometry to Study Protein Complexes

Brent A. Kochert, Roxana E. Iacob, Thomas E. Wales, Alexandros Makriyannis, and John R. Engen

Abstract

Hydrogen-deuterium exchange (HDX) mass spectrometry (MS) can provide valuable information about binding, allostery, and other conformational effects of interaction in protein complexes. For protein-ligand complexes, where the ligand may be a small molecule, peptide, nucleotide, or another protein(s), a typical experiment measures HDX in the protein alone and then compares that with HDX for the protein when part of the complex. Multiple factors are critical in the design and implementation of such experiments, including thoughtful consideration of the percent protein bound, the effects of the labeling protocol on the protein complex, and the dynamic range of the analysis method. With careful planning and techniques, HDX MS analysis of protein complexes can be very informative.

Key words Dissociation constant, Binding, Deuteration, Percent bound, LC-MS, Interactions, Ligand

1 Introduction

1.1 Background

Characterizing protein complexes—including stoichiometry, binding interfaces, and allosteric effects—can play a significant role in our understanding of biology and in the discovery and development of new and improved therapeutics. There are many structural tools that can be used to study protein complexes, including, of course, X-ray crystallography, NMR spectroscopy, X-ray or neutron scattering, and cryo-electron microscopy. Mass spectrometry also has a role to play, including stoichiometry measurements and the use of hydrogen-deuterium exchange (HDX) mass spectrometry (MS) [1–3]. HDX MS can provide valuable information about binding, allostery, and other conformational effects of interaction in complexes [4–6].

HDX MS is based on labeling backbone amide hydrogens with deuterium and measuring the incorporation with a mass spectrometer [7]. Exchange between the backbone amide hydrogens and

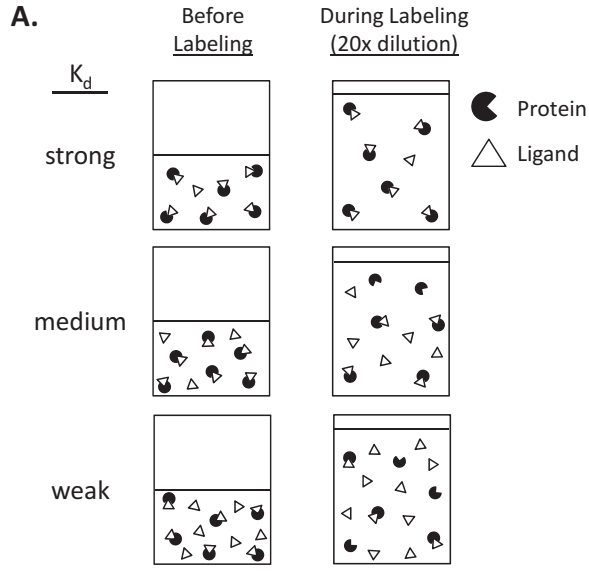
Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_10,

solvent D_2O relies on multiple factors, including temperature, pH, hydrogen bonding, and solvent exposure [8]. By controlling temperature and pH, hydrogen bonding and solvent exposure can be probed. Comparison experiments are often employed wherein two states are compared, such as ligand-bound vs. apo, mutant vs. wild type, etc. For protein-ligand complexes, a typical experiment measures HDX in the protein alone and then compares that with HDX for the protein when part of the complex. Here, ligand can mean small molecule, peptide, or other proteins.

1.2 Theory Considerations

The theory of binding should play a major role in any HDX MS experiment involving complexation. Most important is understanding binding strength, kinetics, and how these will change as a function of concentration during the dilutions that are included in a typical HDX MS experiment (Fig. 1). As described below, both theoretical parameters and known concentrations are used to calculate the percentage of bound protein molecules (referred to hereafter as % bound) at each stage of the experiment. For complexes with strong binding (small dissociation constant, K_d), all protein molecules (notched black circles in Fig. 1a) are bound to ligand (white triangles in Fig. 1a) even after the dilution as a consequence of the addition of the D_2O labeling solution. As the K_d increases (binding strength weakens), additional ligand is required to reach high % bound (Figs. 1b and 2). Further, because % bound decreases upon the labeling dilution, weak binding can be problematic, and this effect is more drastic for proteins that have a larger K_d. With a large K_d, one can try to compensate by starting with a high percentage of protein bound when the protein and ligand are equilibrating (before D_2O dilution). A typical dilution for an HDX experiment is anywhere from 10× to 20× with D_2O buffer, with higher dilution values used to force HDX to be unidirectional (*see* **Note 1**).

The consequences of various K_d's and amounts of bound/free protein and ligand are multifaceted. We illustrate a few considerations in Fig. 2. In a practical experiment, these considerations, including % bound, concentration of ligand, K_d, and ligand:protein ratio, must be incorporated into the workflow. One can calculate [9, 10] the % bound (Fig. 2a) if the concentration of ligand, the concentration of protein, and the K_d for the interaction are known (*see* **Note 2**). For all calculations and discussions of binding presented herein, we assume that there is a single binding site. Realize that the ligand may be a small molecule, a peptide, another protein, or even a membrane. If binding is strong (e.g., K_d is nM), calculation of % bound protein and % bound ligand can indicate if it will be possible to obtain HDX information for all members of the complex in one experiment or if multiple experiments are required (*see* **Note 3**).



B.

K_d	% Protein Bound Before Labeling	% Protein Bound During Labeling
0.01 μM	100.0%	100.0%
5.0 μM	98.3%	81.4%
20 μM	95.6%	55.1%

Ligand:Protein is 30 μM:10 μM

Fig. 1 Schematic of theoretical binding events before and during deuterium labeling as a function of K_d. (**a**) Strong binding (low K_d) means all protein molecules are bound to ligand, even after D_2O labeling dilution. A 20-fold dilution is shown as an example, although dilution is typically in the range 10- to 20-fold. In weak binding, more ligand must be added to force the protein molecules to be bound: even though a high percentage of protein molecules are bound before labeling, many protein molecules are not bound after dilution. (**b**) Numerical values representative of cartoons in panel a. For this example calculation, using a 20-fold dilution, the concentrations of 10 μM protein and 30 μM ligand (during labeling) were fixed and K_d was varied

Plotting % bound versus ligand:protein ratio at various K_d's (Fig. 2b)—or % bound versus K_d at various ligand:protein ratios (Fig. 2c)—illustrates that at low (e.g., 10–100 nM) K_d where the interaction between the ligand and protein is strong, there is a high percentage of bound protein molecules even when the ligand:protein ratio is small. As the K_d becomes larger (weaker binding), it becomes more and more difficult to reach a high percentage of bound protein. When comparing two K_d's at the same ligand:protein ratio (dotted vertical line in Fig. 2b), the K_d that is smaller will result in a higher percentage of bound protein.

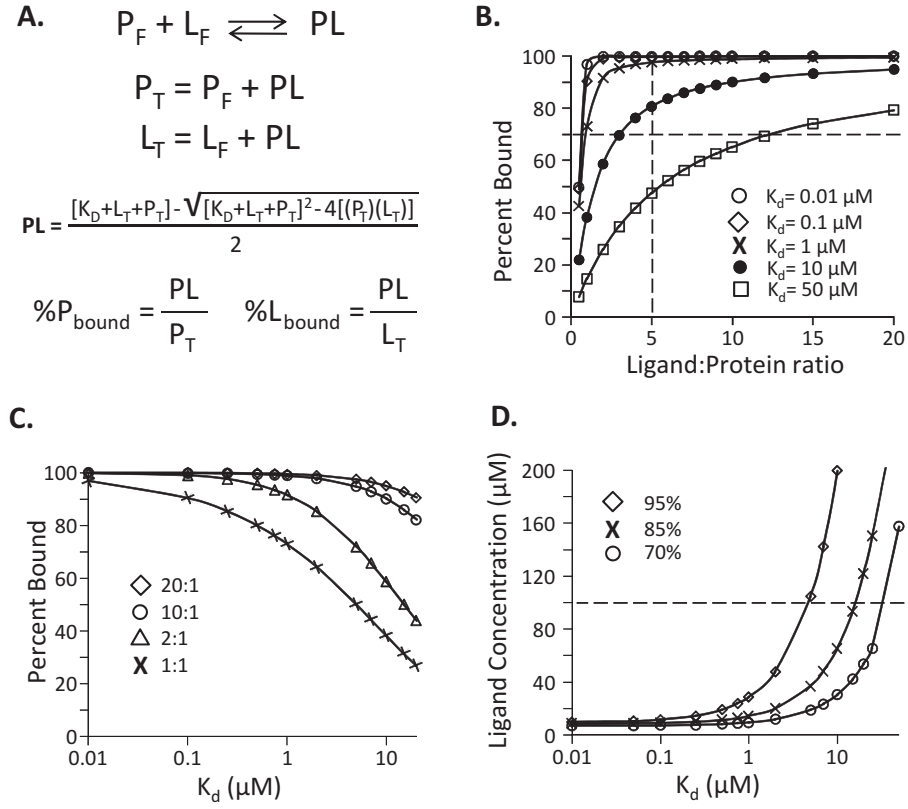


Fig. 2 Theoretical considerations prior to an experiment. (**a**) Equations used to determine the percentage of protein bound to a ligand where L_F represents free ligand concentration, P_F represents free protein concentration, PL represents protein-ligand complex concentration, P_T represents total protein concentration, L_T represents total ligand concentration, and K_d represents the dissociation constant. (**b**) Percent bound versus ligand:protein ratios at various K_d's, as shown in the legend. Ligand:protein ratio is expressed as moles:moles. At lower K_d, the same ligand:protein ratio (dotted vertical line) will yield a higher percentage of protein bound. For example, if percent bound must be greater than 70% (dotted horizontal line), a 5:1 ratio is not feasible for the highest K_d. (**c**) Percent bound versus K_d at various ligand:protein ratios (as in the legend). Ligand:protein ratio is expressed as moles:moles. (**d**) Ligand concentration (in µM) during labeling versus K_d at various percent bounds shown in the legend. For this calculation, protein concentration during labeling is held constant at 10 µM. Too much ligand (dotted horizontal line, here at 100 µM) can become problematic and interfere with detection

The higher the % bound, the better because fewer unbound ligand molecules will interfere with interpretation of the mass increases of protein molecules that are bound (*see* **Note 4**). However, not all interactions are sub-nanomolar, so it is often not possible to have such a high percent bound.

A simultaneous and very important consideration is how much ligand the method can tolerate. To force a high % bound, theoretical calculations such as Fig. 2 suggest that a large excess of ligand must be present. For example, to achieve 95% bound for a K_d of 0.1 µM, roughly 10 µM ligand is required for a 10 µM protein

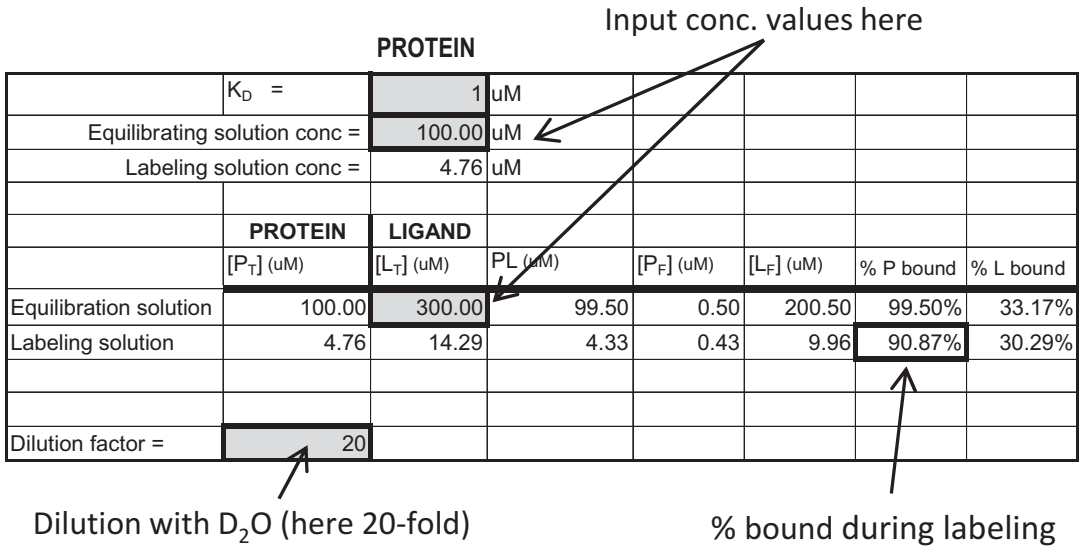


Fig. 3 Microsoft Excel (or similar) can be used to quickly calculate the % bound, as well as other values. Gray boxes are areas of required input, including K_d, concentration of protein, P_T (*see* Fig. 2a) (*Note*: at equilibration step after mixing with the ligand, not concentration of the stock material), ligand concentration, L_T (at equilibration step), and dilution factor. The value for PL is determined using the equations described in Fig. 2a (not all intermediate values are shown here). The % bound (both protein and ligand) during labeling is the desired outcome (lower right)

solution (during labeling). When the K_d is 5 μM, however, roughly 100 μM ligand is required (Fig. 2d). It may not be realistic to maintain a high % bound (e.g., 95% bound) for large K_d's (e.g., 5 μM) because too much ligand (e.g., 100 μM) will be injected into the mass spectrometer (*see* **Note 4**). In such circumstances, particularly those where a poor K_d cannot be avoided, it may be better to compromise with a lower % bound. If in the above example the % bound was 85% instead of 95%, using the same K_d's (0.1 μM and 5 μM), only 9 μM and 37 μM ligand are needed, respectively, for a 10 μM protein solution (during labeling).

The necessary theoretical calculations of the % bound, as well as other useful values, can and should be made before an actual experiment is performed. A spreadsheet (Fig. 3) is a convenient way to calculate the values, as well as investigate the feasibility of various experiments. The spreadsheet we use for making these calculations is available upon request from the authors. In this calculation, the areas of needed input are entered, and the % bound is calculated using the equations in Fig. 2a for both the equilibrium conditions and the deuterium labeling conditions. It is also important to consider the dilution factor upon adding D_2O. Significant time can be saved by planning a binding experiment in this way before carrying it out. Feasibility is determined, mixing ratios at equilibration outlined, and the impact of dilution is revealed.

2 Materials

2.1 Selected Reagents

1. Deuterium oxide (D_2O), >99.5% atom % D (Cambridge Isotope Laboratories).
2. Deuterium chloride (DCl) (Cambridge Isotope Laboratories).
3. Sodium deuteroxide (NaOD) 40 wt % 99 + atom % D (Cambridge Isotope Laboratories).
4. Porcine pepsin, lyophilized powder 3200–4500 units/mg protein (Sigma-Aldrich).

2.2 Key Buffers

As an example, for the data presented in this chapter, we used the following buffers:

1. Equilibration buffer (no deuterium): 50 mM Tris–HCl, 50 mM KCl, pH = 8.0 (in H_2O).
2. Labeling buffer (deuterium): 50 mM Tris–HCl, 50 mM KCl, pD = 8.0 (in D_2O). Recall that pD = pH (reading) + 0.4 [11].
3. Quench buffer: 150 mM KH_2PO_4 pH = 2.3 or 2.0 M guanidine HCl pH = 2.3 (if guanidine is needed).

2.3 LC-MS Materials

1. Lockmass solution: 200 fmol/μL GluFib peptide in 50% acetonitrile, 50% water, and 0.1% formic acid.
2. ACQUITY UPLC HSS T3 C18 1.8 μm 1.0 × 50 mm analytical column (Waters Corp.).
3. VanGuard BEH C18 1.7 μm guard column (Waters Corp.).
4. Pepsin column packed in-house, pepsin immobilized on POROS 20 AL beads (Life Technologies).

3 Methods

3.1 Methods: Controls

A typical schematic for an HDX MS binding experiment is shown in Fig. 4. The general idea is that the protein and the ligand are mixed together (into what is often called the equilibration solution) and allowed to sit for some period of time, at some desired temperature (often room temperature, e.g., 21 °C). This equilibration solution is diluted with D_2O and labeling begins. Labeling is allowed to proceed for predetermined time(s), quenched, and deuterium is located using LC-MS. After the mixing of protein and ligand, this protocol resembles typical continuous-labeling HDX MS experiments [12–18].

Follow the experimental protocol outlined below for the actual deuterium labeling experiment, but before that, some initial testing is required to optimize experimental parameters. These tests mimic the protocol used in the deuterium labeling experiment, but

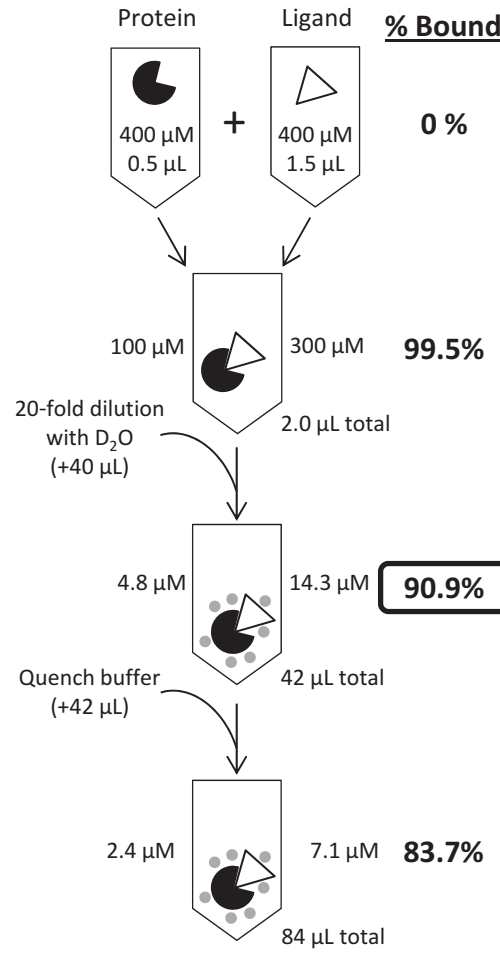


Fig. 4 Schematic showing the various steps that occur during an HDX MS experiment designed to probe ligand:protein interactions. This hypothetical experiment occurs under conditions described in Fig. 3: the dissociation constant is 1 µM, and the concentrations of the protein of interest and ligand are 300 µM and 100 µM, respectively, after being mixed together from stock solutions (both 400 µM). A 20-fold dilution with D_2O buffer is used in this scenario. The percentage bound at each step is noted on the right side of the figure

instead of diluting with deuterium-containing buffer, an identical buffer made with H_2O is used. In these tests:

1. Evaluate the signal and quantities of material required by the mass spectrometer. During this step, the volumes, concentration, moles of material injected, dilution factor, etc. can be varied to determine what kind of signal can be seen for the protein of interest. Use the spreadsheet (Fig. 3) to make rational decisions about how each parameter can be optimized.
2. Obtain a preliminary peptic digestion map. By following the identical protocol (below) but using H_2O rather than D_2O,

the protein is digested under the same conditions that will be used when deuterium is present. Identify all the peptic peptides produced, and construct a preliminary coverage map (*see* **Note 5**). One is aiming for high backbone coverage and redundancy in the peptides, particularly if the ligand is a protein (*see* also **Note 4**). Methods of digestion (including online vs. offline, choice of enzyme, etc., *see* **Note 6**) and quench buffer conditions (*see* **Note 7**) can be altered and manipulated until the optimal conditions are found for the best coverage of the protein.

3. Work out all instrumental conditions with undeuterated materials. This includes determining if wash steps are needed (*see* **Note 8**) and optimizing the LC conditions (*see* **Note 9**).
4. Once the instrumental conditions are finalized, obtain the working peptide map. This map may be different from the preliminary map (**step 2**) because during adjustment of gradient, digestion conditions, and so forth, things like retention time, number of peptides, etc. may have changed.

3.2 Methods: Binding Experiment

1. Follow Fig. 4. Mix the appropriate ratios (calculated using Fig. 3) of protein and ligand together (*see* **Note 10**).
2. Allow the mixture of protein and ligand to equilibrate for some time (e.g., 1 h) at room temperature (*see* **Notes 11** and **12**). Equilibration ensures that the protein and ligand have enough time to bind to one another and increases the likelihood of a uniform population.
3. Transfer an aliquot from the equilibration solution mixture to a new tube, and add D_2O buffer to initiate the labeling (*see* also **Note 10**). A 10- to 20-fold dilution with D_2O is typical. The sample is often left at room temperature for labeling, although other labeling temperatures are sometimes used (e.g., 4 °C, 30 °C, etc., *see* **Note 11**).
4. Let the labeling reaction sit for predetermined amounts of time. These times vary according to protein and range anywhere from 5 s to 8 h to multiple days, even. A typical time course covers at least five decades, with equally spaced time points, and is most often performed in triplicate.
5. Once the labeling time has elapsed, labeling is quenched by adding quench buffer (*see* **Note 7**) to lower the pH to 2.5. A variety of quench buffers are available, and the pH of the quench buffer may need to be altered in order to ensure that the final pH will be 2.5 once everything is mixed. A 1:1 labeling reaction:quench buffer (by volume) is common. Store the quench buffer on ice to ensure it is cold.

6. Immediately digest the quenched sample, either by injecting into an online digestion system or by adding protease solution (*see* **Notes 6** and **13**). Proceed to **step 9**.
7. Repeat **steps 1–6** for protein alone, with no ligand present. This will provide a point of reference and allow for differences in deuteration between free and bound protein(s) to be determined. It is important to follow the identical dilutions and steps, so mix buffer (with no ligand) at **step 1** and proceed as if ligand(s) were present.
8. Prepare some undeuterated controls by repeating **steps 1–6**, but at **step 3**, use the identical buffer but with H_2O rather than D_2O. These samples will act as controls and are needed to calculate deuterium uptake.
9. Separate the peptides (*see* **Notes 14** and **15**) from the protease digestion, and elute them into a mass spectrometer for mass analysis (*see* **Notes 16** and **17**).

3.3 Data Interpretation

Interpreting HDX can be done in several different ways, depending on the scientific question. Ultimately, one likely wants to better understand the changes in exchange in the protein upon binding of the ligand. A typical processing and interpretation workflow is (*see* also [19, 20]) as follows:

1. Measure the average deuterium incorporation in both free and bound proteins (Fig. 5a).
2. Plot deuterium uptake as function of the deuterium labeling time (Fig. 5b).
3. Calculate the differences upon binding to the ligand. This is often done by subtracting the deuterium incorporation in the bound form from the deuterium incorporation in the apo/free form. With this subtraction equation, a negative number corresponds to decrease or protection from labeling upon ligand binding (peptide 1 in Fig. 5), while a positive number corresponds to an increase or exposure in labeling upon ligand binding (peptide 2 in Fig. 5).
4. Do not assume that something should be seen—*see* **Note 18**—and do not assume that there is always a decrease in exchange as a result of binding. Ligand binding, particularly if the ligand is another protein, could increase deuteration due to structural rearrangements in the complex and changes in hydrogen bonding patterns [21].
5. Once differences in deuteration are determined from the deuterium incorporation graphs, the differences can be mapped onto a structure, if known (Fig. 5c), and even colored according to the magnitude of the differences. The meaning of the data can then be considered (*see* also **Note 19**).

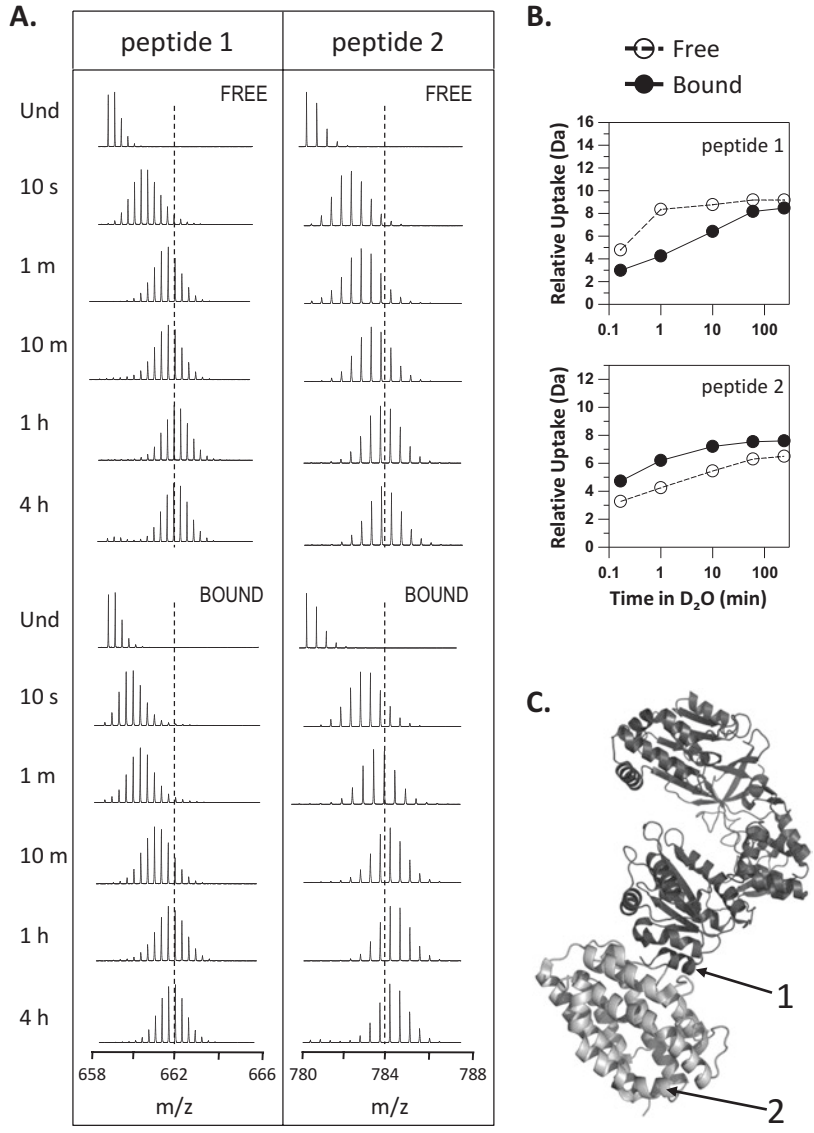


Fig. 5 Typical data workflow for an HDX MS experiment designed to probe protein-protein interactions. Two representative peptides (peptide 1 and peptide 2) have been chosen to display (**a**) mass spectra for the free (top) and bound forms (bottom), (**b**) deuterium uptake graphs for the spectra in panel a, and (**c**) the location of the peptides in the protein (indicated with arrows). Peptide 1 represents a scenario in which protection from exchange occurs upon binding. Note the lower *m*/*z* values for the isotope clusters in the bound form (dotted line provided for visual guidance). In contrast, peptide 2 undergoes deprotection from exchange upon binding (*see* also **Note 18** and **step 4** of Subheading 3.3), and more label is incorporated in the bound state

3.4 Extracting a K_d

HDX MS can, in certain circumstances, be helpful in measuring dissociation constants that are not easily obtained by other methods. There may also be scenarios in which the K_d for an interaction is not known yet HDX MS is desired. When a K_d is not known for an interaction, a series of HDX MS experiments (a titration) can be performed in which the ligand:protein ratio is varied (*see* **Note 20**). When the scientific question is "where are changes upon binding?", perhaps only a few titration points are necessary, and a K_d is not determined by HDX MS [22]. If regions of the protein are sensitive to binding, changes in deuteration will be observed in those regions during the titration (Fig. 6). Due to the amount of experiments involved, and the relative ease with which K_d's can be determined by other means, HDX MS is likely one of the last methods selected for measuring K_d.

1. First, determine if changes in HDX can be observed at the whole protein level as the ligand:protein ratio is varied (Fig. 6a). Hold the concentration of protein constant. Beware again that no evidence for change at the whole protein level does not mean there are no small changes at the peptide level, especially those that are equivalent in magnitude but opposite in sign.
2. Perform a peptide-level experiment with digestion after the quench step. Determine if changes can be observed in any peptides as the ligand:protein ratio (*see* **Note 20**) is varied (Fig. 6b).
3. Select a peptide (or several peptides) that show(s) changes in the presence of ligand and a deuterium labeling time where the changes are obvious.
4. Perform a titration, with the single labeling time selected in **step 3**. Hold the protein concentration constant while varying the ligand concentration. Typical ratios of ligand:protein could range from 1:1 to 50:1, depending on ligand, and could often include 10 or more ligand:protein ratios.
5. Interpret by plotting deuterium level in the selected peptides as a function of ligand concentration (as in [22]), or determine the % bound versus ligand concentration (as in [23]). *See* Fig. 6c.

By changing the amount of ligand in each titration point, two populations emerge: protein molecules bound to ligand and protein molecules not bound to ligand. The ratio of these two species changes as the concentration of ligand changes. Figure 6a, b show this effect for intact protein and for peptic peptides. As more and more of the protein is bound, there is less deuterium uptake, resulting in a smaller m/z value (lower mass distribution). If a lower concentration of ligand is used, less of the protein will populate the bound state, and instead begin to populate the unbound state where it is able to incorporate more deuterium leading to an increase in mass and a larger m/z value. By following the two states

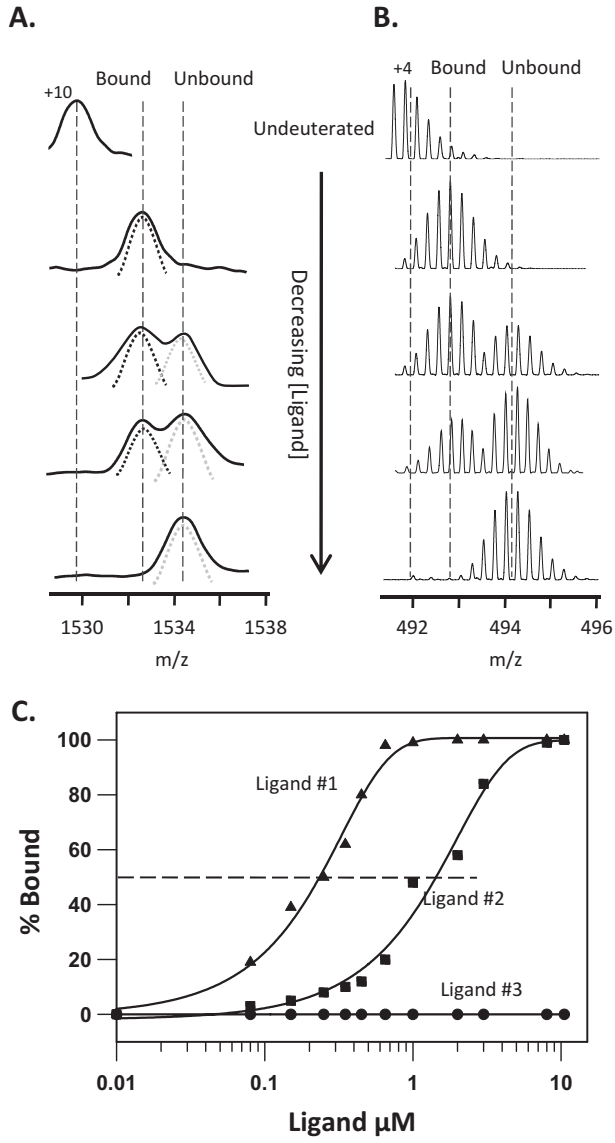


Fig. 6 Hypothetical data used to find the K_d from HDX MS. (**a**) Intact spectra (+10 charge state) of a protein at different ligand concentrations (high at top, low at bottom). Such spectra are acquired at a single deuterium labeling time point, usually when the binding effect is the largest. Increasing the concentration of ligand leads to an increase in percent bound (protected species of lower *m*/*z*). (**b**) Spectra of example of peptides from the protein in panel a, at different percentages bound. (**c**) Titration curves for three separate ligands for a protein at a constant concentration. The K_d is the concentration of ligand at which 50% of the protein is bound and is denoted by a dashed line on the graph

(bound and unbound) at a certain charge state, one can begin to track the percentage that is bound. The concentration of ligand that is required to achieve 50% bound protein is the K_d for the interaction, which may be extracted by fitting the raw % bound vs. ligand concentration values [23].

Several important caveats should be considered in a titration experiment. Not all binding interactions may lead to easy to characterize and analyze bimodal distributions such as those in Fig. 6 (*see* also **Note 18**). EX1 kinetics [24] and run-to-run carryover [25] may also display bimodal isotope patterns and may interfere with titration experiments if not dealt with (*see* **Note 8**). If the binding interface is unknown, distinguishing EX1 kinetics that are a result of a binding interface and examining the bound and unbound state may be difficult, something that can often be solved by mutagenesis [22].

4 Notes

1. Once the dilution with D_2O buffer occurs, the original binding equilibration will be disrupted. Re-equilibration may or may not be very rapid; in cases where it can be slow, re-equilibration may lead to problems for short deuterium labeling times. There could be a scenario in which a short deuterium labeling experiment is quenched before the protein and ligand are able to fully re-equilibrate, which would result in probing a different % bound population than expected.
2. The concentrations used in the percent bound equation should be those when the protein is in D_2O buffer, post-dilution. This is when the concentrations of protein and ligand matter the most as this is when the interaction is being probed/labeled with deuterium.
3. For protein-peptide and protein-protein interactions, it may be possible to obtain HDX information for all members of the complex in one experiment [22]. Several control experiments will be required: the analyte protein alone, the analyte protein bound to its ligand, and the unbound ligand. Calculations (*see* Figs. 2a and 3) can reveal how much ligand (e.g., peptide or other protein(s) of interest in a complex) will be unbound during the experiment. If a majority of the ligand remains unbound, because a high concentration is required to force binding to the analyte protein during labeling, the unbound ligand will incorporate deuterium to the same degree as completely free ligand (ligand not mixed with analyte protein). In these cases, to obtain information about what binding does to the ligand, another experiment is required where ligand concentration is fixed and the protein of interest is placed in

excess. The necessity and feasibility of such experiments can be determined from theoretical calculations. A digestion map should be generated in each case because the digestion pattern might change when the binding partner is present compared with the protein alone.

4. In the case of a weak K_d, it may be necessary to add large amounts of ligand in order to get a reasonable % binding (≥70%) during labeling. The signal(s) of the analyte protein can become obscured by the signal from high concentrations of ligand, which might not be of interest (*see* **Note 3**). This issue becomes more problematic when the ligand is a large protein (e.g., mAb, MW = 75,000 Da unique sequence), as opposed to a small molecule (e.g., inhibitor MW = 300 Da). Large protein ligands, which might not be of interest, create many peptides during digestion that may interfere with the signal coming from the analyte protein. Additional peptides generated from a protein ligand may increase ion suppression or result in spectral overlap of peptides from the analyte protein. There are several ways that these problems can be dealt with in order to obtain meaningful data. One way is to use a mass spectrometer that has ion mobility capabilities [26, 27]. Employing ion mobility will add an orthogonal dimension of separation that often helps distinguish peptides that overlap in the LC and m/z dimensions. Another possibility could be to alter the LC gradient [28]. A longer gradient will allow for a more gradual elution but can come at the cost of increased deuterium back-exchange unless the flow rate is increased. The longer peptides are in the H_2O mobile phase, the more likely it is that deuterons occupying the amide backbone begin to revert back to hydrogen. Both tactics, utilizing ion mobility and/or a longer/faster LC gradient, can increase peak capacity.
5. Software for peptide identification and measuring and processing deuterium uptake drastically improves the speed of analysis and the complexity of systems that can be studied. If using an HDX MS system from the Waters Corporation, as we do in our research, PLGS and DynamX are available for peptide identification and determining deuterium uptake, respectively. There are other software packages that are available [20].
6. Digestion can take place offline (pepsin in pure water at 1–10 mg/mL at or below pH 5) or online with pepsin immobilized on a solid support and packed into a column [29, 30]. Offline digestion can also occur with pepsin immobilized on a solid support (e.g., agarose beads such as from Pierce or immobilized in-house onto POROS beads) and the beads removed with centrifugation. Other acid proteases besides pepsin can be tested for digestion [28, 31–33].

7. The quench buffer can be changed to optimize digestion, depending on the protein. Testing the quench buffer should be done without deuterium. A good starting quench buffer is 150 mM KH_2PO_4 buffer pH 2.5 (a phosphate buffer is a good starting point as pK_a1 ~2.1). If the analyte protein will not tolerate a phosphate buffer or there are cofactors that prevent the use of phosphate, then a citrate or a glycine buffer can be used. If the analyte protein has disulfide bonds, then reducing agents such as TCEP or DTT can be added to the quench buffer. If the quench buffer contains a reducing agent such as TCEP or DTT, we recommend that the solutions are made fresh every day and never stored. If the analyte protein is resistant to digestion using the above starting points, then chaotropic agents such as guanidine hydrochloride (GdHCl) or urea can be added to partially denature the protein. Keep in mind that pepsin as a protein is also sensitive to chaotropic agents. While the concentration of GdHCl required for optimal digestion will differ among proteins, a typical quench buffer contains 2.0 M GdHCl pH = 2.3.
8. A wash step should be run, as needed, to prevent carryover from previous samples [25, 34]. Some proteins have high run-to-run carryover and others do not; carryover cannot be predicted. Carryover can happen due to incomplete elution from the pepsin column, the desalting trap, and/or the separation column. Test for carryover with a blank injection after a non-deuterated sample is injected. Apply appropriate wash steps as needed. Routine washing of the pepsin column during the peptide separation is suggested. An effective solution to wash the pepsin column contains 1.5 M GdHCl, 4% acetonitrile, and 0.8% formic acid.
9. Before using the LC system, insure that all mobile phase lines do not have any air bubbles. If air bubbles are present, purge the lines. Ensure there are no leaks in the LC system. Double check the pH of the mobile phases to ensure that acid was added (sometimes easy to forget). Use the same LC for the test digestions and the deuterium labeling experiments. It is important that the LC system stays at 0 °C even though for test digestions, there is no deuterium to back-exchange. Use the same exact experimental conditions (mobile phases, pH, gradient program, desalting time, etc.) throughout the experiment to ensure good reproducibility [28]. The LC method (i.e., gradient program) may need to be altered and tested: for larger proteins the gradient may need to be longer. It is important to remember, however, that longer LC gradients will increase back-exchange.
10. A good strategy is to, once the appropriate ratios (calculated using Fig. 3) of protein and ligand are determined, mix a batch

of protein-ligand complex in a single tube (the equilibration solution). Make enough material for all undeuterated controls as well as all labeling experiments. Pull from this single tube enough material for undeuterated controls and then, later, material for deuterium labeling. Aliquots can be withdrawn, and D_2O labeling buffer added to start the labeling reaction or D_2O buffer can be added to the single tube to start the labeling. Always add large volumes to small ones to ensure adequate and reproducible mixing.

11. Equilibration for 1 h at room temperature is a good starting point but is not necessarily optimal for all proteins. Depending on the interaction and K_d, a shorter incubation time may be tolerable; in some cases, equilibration time may need to be increased (*see* Fig. 6). Sometimes both equilibration and labeling have to be done at a lower temperature particularly if the protein is unstable at higher temperature. Note that labeling will occur at a slower rate as temperature is decreased [35].
12. When investigating small molecule binding, there are a couple of considerations to keep in mind before proceeding. The solubility of the compound is very important, and the percentage of organic solvents (e.g., DMSO) should be kept at minimum and should not exceed 5% in the equilibration solution. Higher % (i.e., >5%) of DMSO can have a destabilizing effect on protein conformation. If the compound is too hydrophobic, it may interact with the pepsin column and or trapping and separation columns and cause blockage. Small molecules can also suppress peptide signal at the point of chromatographic elution.
13. After the labeling reaction has been quenched, the samples can be analyzed by MS immediately or placed directly on dry ice in a sealed tube and frozen at −80 °C for analysis later (usually with 3–4 days). The samples can remain frozen for 7–14 days without loss of significant label at −80 °C. When the time comes for analysis, thaw rapidly to 0 °C and proceed to digestion.
14. For most HDX MS experiments, a short LC gradient (typically 6–10 min) is used in order to limit the back-exchange of the peptides. A VanGuard BEH C18 trap in conjunction with a HSS T3 (or BEH) C18 analytical column, or comparable C4 trap and column, can be used in order to separate digested peptides. It is important that everything is kept at 0 °C to ensure that back-exchange is as small as possible. Be aware that very high concentrations of ligand, especially when ligand is a protein (*see* **Note 4**), can compromise the separation efficiency due to column overload. The column size and capacity may need to be considered in such cases.

15. In cases where a membrane protein is investigated, the buffers used throughout the procedure should be free of detergents unfriendly to mass spectrometry (e.g., Triton, Tween, CHAPS); their presence will interfere with the signal in the mass spectrometer. Some detergents are less problematic such as LMNG (lauryl maltose neopentyl glycol) and DDM (*n*-dodecyl-ß-D-maltoside), at concentrations not exceeding 0.01%. A longer desalting step should be included in the LC method to allow for complete removal of the detergent.
16. The mass spectrometer should be calibrated frequently, at a minimum every day at the beginning of use.
17. Any type of mass spectrometer can be employed to acquire the data. It is desired that a higher-resolution mass spectrometer is used in order to correctly identify and distinguish charge states of peptides. *Note* that if ETD or ECD fragmentation, or ion mobility, is sought after, then a MS capable of performing that measurement should be used for analysis.
18. No differences in deuterium incorporation between the free and bound states can occur. A lack of change in the backbone amide hydrogen exchange does not necessarily mean that the protein and the ligand were not interacting. Because exchange is a function of the combined effects of solvent exposure and hydrogen bonding, there are situations where backbone amide hydrogens are not affected by binding or where more deuterium can exchange in a complex. Complexes that are driven mainly by electrostatic or hydrophobic interactions of the side chains of very structured elements, particularly when the ligand is another protein, may undergo no changes in HDX upon interaction [22].
19. If there is no crystal or NMR structure available, one should be especially careful in interpreting the HDX MS data. What has usually been measured is a difference in deuteration when bound versus free. Should a region be suspected as a "binding site" or an "epitope," it must be investigated further by other means such as mutagenesis [36]. One must determine whether changes are seen due to binding, or allostery, keeping in mind that multiple regions that appear to be protected when bound could form part of a discontinuous binding interface [37, 38].
20. When K_d is not available, a good starting point is a 10:1 ligand:protein (by moles) ratio, followed by increasing the quantity of ligand (e.g., 25:1, 50:1) if weak binding is suspected or decreasing the ligand concentration (e.g., 5:1, 2:1) if strong binding is suspected. *See* also [22].

Acknowledgments

Support from the NIH (R01 GM101135, J.R.E.) and a research collaboration with the Waters Corporation are gratefully acknowledged.

References

1. Lossl P, van de Waterbeemd M, Heck AJ (2016) The diverse and expanding role of mass spectrometry in structural and molecular biology. EMBO J 35(24):2634–2657
2. Liko I, Allison TM, Hopper JT, Robinson CV (2016) Mass spectrometry guided structural biology. Curr Opin Struct Biol 40:136–144
3. Pirrone GF, Iacob RE, Engen JR (2015) Applications of hydrogen/deuterium exchange MS from 2012 to 2014. Anal Chem 87(1):99–118
4. Jaswal SS (2013) Biological insights from hydrogen exchange mass spectrometry. Biochim Biophys Acta 1834(6):1188–1201
5. Brock A (2012) Fragmentation hydrogen exchange mass spectrometry: a review of methodology and applications. Protein Expr Purif 84(1):19–37
6. Konermann L, Pan J, Liu YH (2011) Hydrogen exchange mass spectrometry for studying protein structure and dynamics. Chem Soc Rev 40(3):1224–1234
7. Zhang Z, Smith DL (1993) Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. Protein Sci 2(4):522–531
8. Englander SW, Kallenbach NR (1983) Hydrogen exchange and structural dynamics of proteins and nucleic acids. Q Rev Biophys 16(4):521–655
9. Engen JR (1999) Analysis of unfolding and protein dynamics in the regulatory domains of hematopoietic cell kinase with hydrogen exchange and mass spectrometry. Dissertation, University of Nebraska-Lincoln
10. Mandell JG, Baerga-Ortiz A, Akashi S, Takio K, Komives EA (2001) Solvent accessibility of the thrombin-thrombomodulin interface. J Mol Biol 306(3):575–589
11. Glasoe P, Long F (1960) Use of glass electrodes to measure acidities in deuterium oxide. J Phys Chem 64:188–193
12. Engen JR, Smith DL (2000) Investigating the higher order structure of proteins. Hydrogen exchange, proteolytic fragmentation, and mass spectrometry. Methods Mol Biol 146:95–112
13. Houde D, Engen JR (2013) Conformational analysis of recombinant monoclonal antibodies with hydrogen/deuterium exchange mass spectrometry. Methods Mol Biol 988:269–289
14. Forest E, Man P (2016) Conformational dynamics and interactions of membrane proteins by hydrogen/deuterium mass spectrometry. Methods Mol Biol 1432:269–279
15. Masson GR, Burke JE, Williams RL (2016) Methods in the study of PTEN structure: x-ray crystallography and hydrogen deuterium exchange mass spectrometry. Methods Mol Biol 1388:215–230
16. Guttman M, Lee KK (2016) Isotope labeling of biomolecules: structural analysis of viruses by HDX-MS. Methods Enzymol 566:405–426
17. Mayne L (2016) Hydrogen exchange mass spectrometry. Methods Enzymol 566:335–356
18. Tsirigotaki A, Papanastasiou M, Trelle MB, Jorgensen TJ, Economou A (2017) Analysis of translocation-competent secretory proteins by HDX-MS. Methods Enzymol 586:57–83
19. Wales TE, Eggertson MJ, Engen JR (2013) Considerations in the analysis of hydrogen exchange mass spectrometry data. Methods Mol Biol 1007:263–288
20. Claesen J, Burzykowski T (2017) Computational methods and challenges in hydrogen/deuterium exchange mass spectrometry. Mass Spectrom Rev 36(5):649–667. https://doi.org/10.1002/mas.21519
21. Konermann L, Rodriguez AD, Sowole MA (2014) Type 1 and Type 2 scenarios in hydrogen exchange mass spectrometry studies on protein-ligand complexes. Analyst 139(23):6078–6087
22. Engen JR (2003) Analysis of protein complexes with hydrogen exchange and mass spectrometry. Analyst 128(6):623–628
23. Marcsisin SR, Engen JR (2010) Molecular insight into the conformational dynamics of the Elongin BC complex and its interaction with HIV-1 Vif. J Mol Biol 402(5):892–904

24. Weis DD, Wales TE, Engen JR, Hotchko M, Ten Eyck LF (2006) Identification and characterization of EX1 kinetics in H/D exchange mass spectrometry by peak width analysis. J Am Soc Mass Spectrom 17(11):1498–1509
25. Fang J, Rand KD, Beuning PJ, Engen JR (2011) False EX1 signatures caused by sample carryover during HX MS analyses. Int J Mass Spectrom 302(1–3):19–25
26. Iacob RE, Murphy JP 3rd, Engen JR (2008) Ion mobility adds an additional dimension to mass spectrometric analysis of solution-phase hydrogen/deuterium exchange. Rapid Commun Mass Spectrom 22(18):2898–2904
27. Cryar A, Groves K, Quaglia M (2017) Online hydrogen-deuterium exchange traveling wave ion mobility mass spectrometry (HDX-IM-MS): a systematic evaluation. J Am Soc Mass Spectrom 28(6):1192–1202. https://doi.org/10.1007/s13361-13017-11633-z
28. Engen JR, Wales TE (2015) Analytical aspects of hydrogen exchange mass spectrometry. Annu Rev Anal Chem (Palo Alto, Calif) 8:127–148
29. Ehring H (1999) Hydrogen exchange/electrospray ionization mass spectrometry studies of structural features of proteins and protein/protein interactions. Anal Biochem 267(2):252–259
30. Wang L, Pan H, Smith DL (2002) Hydrogen exchange-mass spectrometry: optimization of digestion conditions. Mol Cell Proteomics 1(2):132–138
31. Cravello L, Lascoux D, Forest E (2003) Use of different proteases working in acidic conditions to improve sequence coverage and resolution in hydrogen/deuterium exchange of large proteins. Rapid Commun Mass Spectrom 17(21):2387–2393
32. Ahn J, Cao MJ, Yu YQ, Engen JR (2013) Accessing the reproducibility and specificity of pepsin and other aspartic proteases. Biochim Biophys Acta 1834(6):1222–1229
33. Rey M, Yang M, Burns KM, Yu Y, Lees-Miller SP, Schriemer DC (2013) Nepenthesin from monkey cups for hydrogen/deuterium exchange mass spectrometry. Mol Cell Proteomics 12(2):464–472
34. Majumdar R, Manikwar P, Hickey JM, Arora J, Middaugh CR, Volkin DB, Weis DD (2012) Minimizing carry-over in an online pepsin digestion system used for the H/D exchange mass spectrometric analysis of an IgG1 monoclonal antibody. J Am Soc Mass Spectrom 23(12):2140–2148
35. Engen, JR, Wales, TE, Shi, X (2011) Hydrogen exchange mass spectrometry for conformational analysis of proteins. In: Meyers RA, editor. Encyclopedia of analytical chemistry. Wiley, New York. https://doi.org/10.1002/9780470027318.a9780470029201
36. Ahn J, Engen JR (2013) The use of hydrogen/deuterium exchange mass spectrometry in epitope mapping. Chemistry Today 31(1):25–28
37. Komives EA (2005) Protein-protein interaction dynamics by amide H/H-2 exchange mass spectrometry. Int J Mass Spectrom 240(3):285–290
38. Malito E, Faleri A, Lo Surdo P, Veggi D, Maruggi G, Grassi E, Cartocci E, Bertoldi I, Genovese A, Santini L, Romagnoli G, Borgogni E, Brier S, Lo Passo C, Domina M, Castellino F, Felici F, van der Veen S, Johnson S, Lea SM, Tang CM, Pizza M, Savino S, Norais N, Rappuoli R, Bottomley MJ, Masignani V (2013) Defining a protective epitope on factor H binding protein, a key meningococcal virulence factor and vaccine antigen. Proc Natl Acad Sci U S A 110(9):3304–3309

Chapter 11

Structural Analysis of Protein Complexes by Cross-Linking and Mass Spectrometry

Moriya Slavin and Nir Kalisman

Abstract

Cross-linking and mass spectrometry is used more and more for the structural analysis of large proteins and protein complexes. Although essentially a low-resolution method, it avoids the main drawbacks of established structural techniques. Particularly, it is largely insensitive to the inherent flexibility of the studied complexes and is applied under native conditions. It is also applicable to nearly every structural system. Therefore, cross-linking and mass spectrometry is the method of choice for elucidating the general architecture of protein complexes. Advances in instrumentation, techniques, and software now allow every lab that is working with proteins to apply the approach without much difficulty. The most specialized step in the workflow, the mass spectrometry measurement, can be done in most facilities that are performing standard proteomics. We detail here a step-by-step protocol of how to successfully apply the approach in collaboration with the mass spectrometry facility in your institution.

Key words Structural biology, Molecular machines, Protein architecture

1 Introduction

Structural studies of large protein complexes are very challenging. Established techniques such as crystallography or cryo-electron microscopy are often not applicable to such complexes because of their inherent flexibility. Additionally, large complexes might be unstable and break into heterogeneous mixtures under the preparation conditions of these approaches. In recent years, the method of cross-linking and mass spectrometry (abbreviated as XL-MS or CX-MS) has emerged as an attractive alternative that largely avoid such issues [1]. In XL-MS, the protein complex is incubated with a short bifunctional cross-linking reagent under native conditions (Fig. 1). The cross-linking reaction creates stable covalent links between protein side chains that are structurally close within the context of the intact complex. The proteins are then denatured and digested into peptides, some of which are still cross-linked in pairs. Through the "peptide-sequencing" capability of the mass

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764,
https://doi.org/10.1007/978-1-4939-7759-8_11,

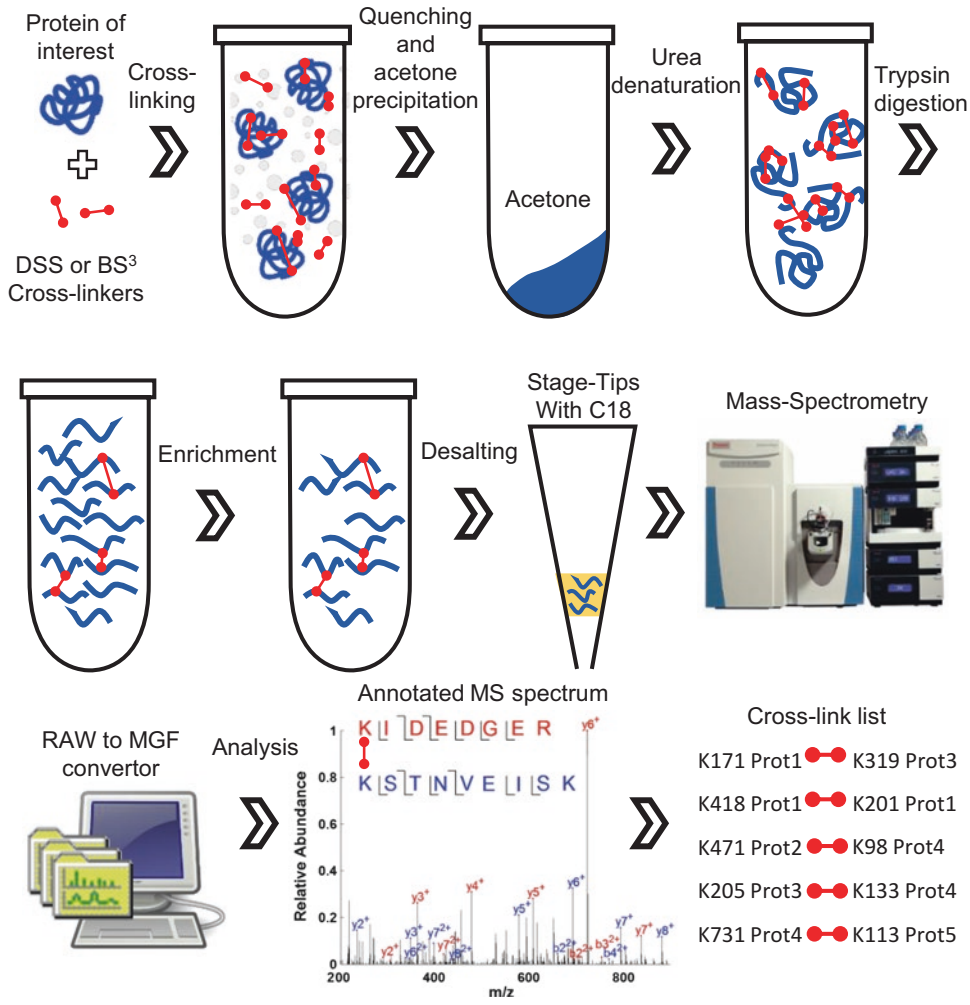


Fig. 1 The XL-MS workflow. The protein complex is incubated with a short bifunctional cross-linking reagent under native conditions. Subsequent steps extract the cross-linked protein, denature it, and digest it with trypsin. The resulting peptide digest is enriched for the cross-linked peptides by size exclusion chromatography. The sample is desalted and measured in the mass spectrometer. The mass spectrometer output files are processed and searched for spectra that report on cross-linked peptide pairs. An example of such spectrum shows the overlapping fragmentation series from two peptides, which indicates that they were cross-linked into a single ion. The result of the analysis is a list of cross-links that fully details all the pairs of residues (protein names and residue numbers) that underwent cross-linking

spectrometer, the sequences of two cross-linked peptides can be found, thus identifying the exact pair of residues that underwent cross-linking. This "residue resolution" is one of the main advantages of XL-MS. A typical XL-MS application on a protein complex will result in hundreds of cross-link identifications, which in turn are converted into connectivity maps and distance constraints. This is a very rich structural resource that elucidates the architecture of the protein complex.

Cross-linking and mass spectrometry was a key methodology in recent structural works that covered such diverse systems as chaperones, transcription, or photosynthesis [2–5]. In these works, the cross-links were mainly used to report on the interactions between different protein subunits within a complex. In many cases, the cross-link data were combined with other structural information sources. For example, excellent synergy occurs between cross-links and electron densities obtained by cryo-electron microscopy. The electron density gives the global envelope and shape of the particle, while the cross-links constrain locally the possible arrangements of the subunits within that envelope.

The advances in instrumentation and software now allow any lab to easily apply XL-MS on its protein complex of interest. The wide applicability, inherent simplicity, and low cost of the XL-MS workflow make it the approach of choice for the initial structural probing of most systems. The most specialized step in the workflow is the mass spectrometry measurement, which requires an expensive mass spectrometry system. Fortunately, these days many research institutions have a mass spectrometry facility to which samples are submitted for proteomics analysis. Since both XL-MS and standard proteomics are run on the same instruments, we assume that most labs could measure their cross-linked samples at their local MS facility. Accordingly, we wrote this protocol to be used by labs that have the standard molecular biology equipment but not their in-house mass spectrometer.

This protocol describes cross-linking with two specific cross-linking reagents—BS^3 (bis-sulfosuccinimidyl suberate) and DSS (disuccinimidyl suberate). These are very popular regents that were used in a large fraction of the recent publications utilizing XL-MS. Their popularity arises from several factors. First, they specifically cross-link primary amines, which on proteins means the side chains of lysine residues and the N-termini. This specificity greatly simplifies the analysis of the cross-link data. Second, they are highly reactive at the physiological pH of 7.5. Finally, they strike a good balance of being short enough to convey meaningful structural information and yet sufficiently long to ensure cross-linking. We therefore highly recommend their use if compatible with the specific system.

2 Materials

2.1 Cross-Linking and Digestion

1. The DSS or BS^3 cross-linking reagents. Both DSS and BS^3 are moisture sensitive. Their powders should be stored in a desiccator at 4 °C. DSS can be dissolved in DMSO to a concentration of 100 mM and then divided into aliquots and stored at −80 °C for many months. BS^3 solution should be prepared fresh immediately before use. BS^3 can be dissolved in PBS or HEPES buffers to a concentration of 100 mM.

2. Buffers compatible with cross-linking. Make sure that your sample is in a buffer that is compatible with the cross-linking reaction. If buffer is not compatible, then buffer exchange must precede the cross-linking. For both BS3 and DSS, the pH of the buffer must be above 7.0 and below 8.5. The buffer components must not contain primary amines that would react with those regents. HEPES (50 mM) or PBS (1×) is a good buffer, while Tris is not because of its primary amine moiety.
3. 1 M ammonium bicarbonate solution for quenching.
4. Acetone—HPLC grade.
5. Urea buffer: 8 M urea solution in DDW. Urea buffer should be prepared fresh before the MS preparation.
6. Urea/DTT: 8 M urea buffer with 10 mM DTT.
7. Iodoacetamide (IAA): Prepare a 0.5 M solution of IAA in urea buffer. IAA is light and moisture sensitive. Stored desiccated at 4 °C. Prepare solution just before use and perform reaction in the dark.
8. Digestion buffer: 25 mM Tris–HCl, pH 8, 10% ACN.
9. Sequencing grade trypsin: Vendors commonly sell lyophilized trypsin. 20 μg of lyophilized trypsin powder should be reconstituted in 40 μl of 50 mM acetic acid, divided into aliquots of 1–2 μg trypsin in each, and stored in −80 °C.
10. Acidifying: 5% trifluoroacetic acid (TFA) in DDW.

2.2 Desalting on Stage Tips

1. C18 resin: Empore solid phase extraction octadecyl (C18) 47 mm diameter disks (from 3M).
2. 200 μl pipette tips, 2 ml and 1.5 ml collection tubes.
3. All solvents in this section should be at least HPLC grade and preferably MS grade.
4. Wetting solution: 50% water, 50% ACN, 0.1% TFA.
5. Washing solution: 0.1% TFA in water.
6. Elution solution: 25% water, 75% acetonitrile, 0.1% formic acid.
7. MS reconstitution solution: 97% water, 3% acetonitrile, 0.1% formic acid.

2.3 Size Exclusion Chromatography (SEC)

1. SEC buffer: 70% water, 30% acetonitrile, 0.1% TFA—all HPLC grade.
2. Sep-Pak C18 cartridges (*see* **Note 1**).
3. Superdex Peptide PC 3.2/30 column (GE Systems) and HPLC system.

3 Methods

3.1 Cross-Linking the Protein Sample

3.1.1 Cross-Linking with the BS^3 Reagent

1. Make sure that your sample is in a buffer that is compatible with the cross-linking reaction.
2. Prepare a concentrated solution of 100 mM BS^3 in buffer, and add it to the protein sample for the desired final concentration of BS^3. Mix well. It is recommended to use the minimal BS^3 concentration that would still provide sufficient inter-subunit cross-linking. Previous studies [2, 5] have shown the optimal concentration of cross-linker to be 1–3 mM for systems of purified protein complexes. For very concentrated samples (such as lysates), you can go up to 10 mM to ensure the cross-linking reagent is not dwindled by the sample.
3. Incubate the protein sample with the cross-linker for 45 min at 30 °C or for 90 min on ice.
4. Add 30 mM ammonium bicarbonate to quench any unreacted BS^3 in the sample and incubate for additional 15 min.
5. At this stage, native conditions do not matter anymore as cross-linking has been completed. This is a good stopping point and sample can be frozen and processed at a later time.

3.2 Cross-Linking with the DSS Reagent

1. The cross-linking protocol for DSS is essentially the same as that for BS^3. Both DSS and BS^3 have the same cross-linking reactivity toward primary amines and the same 8-carbon spacer. Unlike BS^3, DSS is membrane permeable and particularly suited for in vivo cross-linking. Because DSS is not water soluble, it must be dissolved in DMSO before being added to the sample (*see* **Note 2**).
2. Add the DSS solution in DMSO to the protein sample for the desired final concentration of DSS. Mix well. The addition of DSS will cause the solution to turn opaque because of DMSO micellization. Incubate under shaking to ensure the mixing of the DSS-DMSO solution with your sample.

3.3 Sample Preparation for Mass Spectrometry

3.3.1 Acetone Precipitation of Proteins

1. Place your sample in an acetone-resistant tube (e.g., 1.5 ml Eppendorf tubes). Estimate your sample volume and add at least fivefold of acetone to it. Mix well (*see* **Note 3**).
2. Cool the sample in −80 °C for 1 h.
3. Centrifuge for 10 min at 14,000 × *g*, preferably under maximal cooling.
4. Discard the supernatant without disturbing the pellet. Do not dry the pellet or it would be very difficult to resolubilize. Small amounts of residual acetone will evaporate in subsequent steps.

3.3.2 Protein Denaturation, Reduction, Alkylation, and Digestion

1. Add 20 µl of urea/DTT solution to the protein pellet. Pipette vigorously until all the pellet is resuspended (*see* **Note 4**). Incubate at room temperature for 30 min. This step denatures the proteins and breaks any disulfide bonds between cysteine residues.
2. Add 2.2 µl of iodoacetamide to final concentration of 25 mM, and incubate at 37 °C for 30 min in the dark. This step modifies the side chains of cysteine residues so that they cannot reform disulfide bonds.
3. Dilute the sample by tenfold with 220 µl of digestion buffer. Add the trypsin and mix well. Use a trypsin-to-protein mass ratio of 1:50–100. Incubate the trypsin digestion for 12–18 h at 37 °C while shaking (600 rpm).
4. Stop the trypsin digestion by acidifying the sample. Adding 5% TFA solution to 1/10 the sample volume will reduce the pH to 4.
5. Place the sample in a speed-vac for 20 min to evaporate the acetonitrile.

3.4 Desalting on Stage Tips

Salts and urea must be removed from the sample before it is injected to the mass spectrometer. To that aim, the peptides are loaded onto C18 resin and bound to it through hydrophobic interactions. The salts do not bind and are washed away. The peptides are then eluted from the C18 with acetonitrile. Here, we detail an inexpensive and efficient way to desalt the sample by using tips and a centrifuge [6].

3.4.1 Stage Tips Preparation

1. With av blunt bore needle, cut a plug with a diameter of 1 mm through a disk of Empore C18 resin (Fig. 2a). Squeeze the resin plug into the bottom of a standard 200 µl pipette tip. More plugs can be added to increase the peptide loading capacity of the tip. The loading capacity of a tip with three plugs is about 7 µg of peptides.
2. Punch a hole in the cap of 2 ml tube, and place the tip through the hole so that it is held halfway. Close the tube with the cap and place in a centrifuge. The tip is now kept suspended above the bottom of the tube. Any solvent on top of the resin will be eventually forced through it by centrifugation.

3.4.2 Desalting

1. Add 50 µl of wetting solution onto the tip and centrifuge at 1500 × *g* for 2 min.
2. Add 100 µl of washing solution and centrifuge at 1500 × *g* for 2 min. Repeat. Empty the 2 ml collection tube.
3. Add the acidified peptide digest on top of the resin and centrifuge at 1500 × *g* for 2 min.
4. Add 100 µl of washing solution and centrifuge at 1500 × *g* for 2 min. Repeat. In the last wash, make sure all the solution on

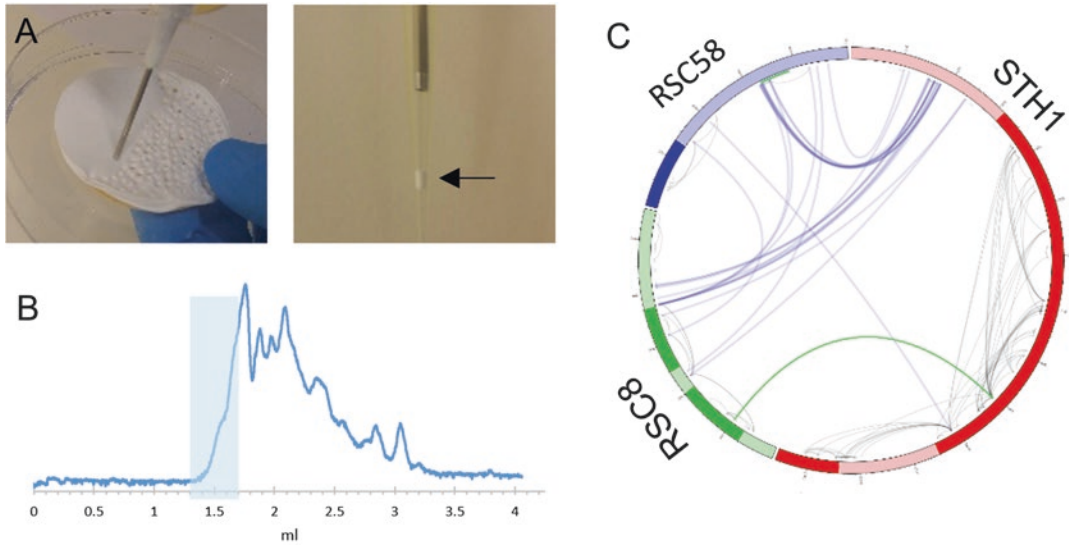


Fig. 2 Steps in the workflow. (**a**) Preparation of the desalting stage tips. A small plug of C18 resin is cut by a blunt bore needle (left) and then pushed to the bottom of a 200 μl pipette tip (arrow on right). (**b**) Chromatogram of SEC enrichment of cross-linked peptides from a large protein complex. The shaded area mark the fractions in which most of the cross-links are later identified. (**c**) CX-Circos is a powerful visualization tool for cross-link data. Here, the cross-links (arcs) between three proteins in a complex are plotted. Specific domains in the proteins and their cross-links can be colored differently

top of the resin washed through. If some wash solution is left, continue to centrifuge until cleared (*see* **Note 5**).

5. Cut off the caps from new 1.5 ml tubes. Transfer the tips and punctured caps onto the new tubes.
6. Add 20 μl of elution solution and centrifuge at 1000 × *g* for 1 min. Repeat.
7. Remove the tip and cap and place the tube with the eluted peptides in a speed-vac for 11–13 min until dry.
8. Reconstitute the peptides by adding the MS reconstitution solution to the tube and mix well. Peptides are now ready for mass spectrometry measurement. The exact volume of reconstitution depends on the amount of peptides and the settings of the chromatography system that is coupled to the mass spectrometer. The amount of peptides can be measured by A_{280} absorbance in a NanoDrop instrument. Typically, you will want to inject to the mass spectrometer about 0.5 μg of your peptides in a volume of about 3 μl.

3.5 Enrichment of Cross-Linked Peptides

Linear peptides rather than cross-linked peptides are the great majority in the trypsin digest. Therefore, enrichment of the digest for cross-linked peptides will lead to better results. While enrichment is optional, it may increase the number of identified cross-links by more than 50%. Here, we present an enrichment

method that is based on size exclusion chromatography (SEC) [7]. It utilizes the molecular weight of the cross-linked peptides, which is on average twice as that of linear peptides.

3.5.1 Buffer Exchange by Sep-Pak C18 Cartridge

1. Add 200 µl of wetting solution to the cartridge. Use a syringe in applying air pressure to the top of the cartridge, and force the solution through the resin to a waste tube.
2. Add 200 µl of washing solution and force it through the resin. Repeat.
3. Add the acidified peptide digest on top of the resin and force it through the resin.
4. Add 200 µl of washing solution and force it through the resin. Repeat.
5. Prepare new 1.5 ml tubes to collect the eluted peptides in the next step.
6. Add 150 µl of elution solution and force it through the resin to the new 1.5 ml tube. Repeat.
7. Place the tube in a speed-vac until dry.
8. Reconstitute the peptides in 20 µl of SEC buffer.

3.5.2 Size Exclusion Chromatography

1. Load the peptides in a SEC buffer onto the Superdex Peptide column at a flow rate of 50 µl/min. Collect fractions every 100 µl. A typical chromatogram is shown in Fig. 2b. Fractions collected at flow volumes of 1.3–1.6 ml are the richest in cross-linked peptides. Elution volumes of 1.6–1.8 ml also contain some cross-linked peptides (*see* **Note 6**).
2. Dry completely the relevant fractions in a speed-vac. Mix peptides with MS reconstitution solution as discussed in Subheading 3.3.2, **step 8**.

3.6 Measurement in the Mass Spectrometer

1. Mass spectrometers coupled to reverse-phase liquid chromatography (LC-MS) are now available for proteomics analysis in many research institutions. The same mass spectrometers can also measure cross-linked samples. However, even after enrichment, the cross-linked peptides are still a minor component of the total peptide content of the sample (the majority being linear peptides). Because most mass spectrometers will first measure the more abundant peptides, the standard proteomics setting is slightly changed to increase the chances of measuring cross-linked peptides. Therefore, inform your mass spectrometry facility of the following subsections when submitting the sample.
2. Cross-linking will lead to samples that are more complex than the original protein content. Longer gradients are therefore beneficial. Use a 45 min gradient for cross-linked samples of two to three proteins. Use a 90–120 min gradient for cross-linked samples of larger protein complexes.

3. Set data-dependent triggering for MS/MS fragmentation that selects for ions with a charge of +3 or higher. For more complex samples with >5 proteins, select only for ions with a charge of +4 or higher. These settings are required because cross-linked peptides typically have a charge of +4 or higher and never less than +3.
4. From our experience, the optimal fragmentation energy for cross-linked peptides during MS/MS is the same as that for linear peptides.
5. We measure our cross-linked samples on a Q-Exactive Plus mass spectrometer with the following settings: buffer A, water with 0.1% formic acid; buffer B, acetonitrile with 0.1% formic acid. Gradient rising linearly from 0% buffer B to 45% buffer B over 90 min, then rising to 80% buffer B over 5 min; full MS resolution 70,000; MS1 AGC target 1e6; MS1 maximum IT 200 ms; scan range 450–1800; dd-MS/MS resolution 35,000; MS/MS AGC target 2e5; MS2 maximum IT 300 ms; loop count top 12; isolation window 1.1; fixed first mass 130; MS2 minimum AGC target 800; charge exclusion, unassigned, 1, 2, 3, 8 > 8; peptide match off; exclude isotope on; dynamic exclusion 45 s.

3.7 Analysis of the Mass Spectrometry Data

1. Convert the mass spectrometer output files from RAW to MGF format. A good conversion tool is the Proteome Discoverer software, which is installed in MS facilities that use Thermo mass spectrometers (*see* **Note 7**).
2. Download the software package—Find_XL—at http://biol-chem.huji.ac.il/nirka/software.html. Unzip the package and find a text document with full instructions for its use in the top folder. The package also comes with example data files from a cross-linking experiment on RNA polymerase II. Use this example to test the installation and get hands-on experience.
3. The analysis requires two inputs: (1) the MGF files of the MS data. Multiple files can be used and consolidated into a single nonredundant cross-link list. (2) The sequences of the proteins in the studied complex. Follow the instructions on how to set the paths to the input files.
4. The output is a text file containing a list of identified cross-links. The cross-links are sorted by decreasing confidence scores. Therefore, only the top of the list should be considered. The question of where to cut the list depends on the false-positive rate (FPR) that you expect from the data. To determine the FPR, the analysis is concurrently running also the protein sequences in reverse. Consequently, the output list will contain entries with the "REV" annotation that are clearly

false positives. The list should be cut so that the ratio of "REV" entries divided by the total number of entries above the cut is the desired FPR.

5. An excellent visualization tool for the connectivity map implied by the cross-links (Fig. 2c) is Circos [8], which can be accessed at http://cx-circos.net/.

4 Notes

1. Up to 200 μg of peptides can be loaded onto this C18 device.
2. The freezing temperature of DMSO is 19 °C. If it is used for cross-linking on ice, the DSS-DMSO solution must be largely diluted by the sample.
3. Acetone precipitation will remove some detergents (such as Triton) but not others (such as SDS). Since detergents are mostly incompatible with mass spectrometry, they should be removed by other means or avoided altogether.
4. If the pellet is very large or hard to solubilize, you can add more urea/DTT. However, this might lead to very large volume of the sample in subsequent steps and should be avoided if possible.
5. This is a good stopping point. Peptides can be stored on the C18 resin for many months before being eluted. Add some washing solution on top of the resin to prevent it from drying and store at 4 °C.
6. SEC on narrow columns of such small volumes is challenging. Avoid dead volumes by using narrow tubing and removal of unnecessary devices in the flow path.
7. Here, we describe how to use our analysis software—"Find_XL." Other software options are "Xi" from the Rappsilber lab at http://rappsilberlab.org/rappsilber-laboratory-home-page/tools/ or "StavroX" from the Götze lab at http://www.stavrox.com/.

Acknowledgment

This work was funded by Israel Science Foundation grant 1768/15.

References

1. Walzthoeni T, Leitner A, Stengel F, Aebersold R (2013) Mass spectrometry supported determination of protein complex structure. Curr Opin Struct Biol 23:252–260
2. Kalisman N, Adams CM, Levitt M (2012) Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. PNAS 109:2884–2489

3. Robinson P, Trnka M, Bushnell D, Davis R, Mattei P, Burlingame A, Kornberg R (2016) Structure of a complete mediator-RNA polymerase II pre-initiation complex. Cell 166:1411–1422
4. Liu H, Zhang H, Niedzwiedzki D, Prado M, He G, Gross M, Blankenship R (2013) Phycobilisomes supply excitations to both photosystems in a megacomplex in cyanobacteria. Science 342:1104–1107
5. Chen ZA, Jawhari A, Fischer L, Buchen C, Tahir S, Kamenski T, Rasmussen M, Lariviere L, Bukowski-Wills JC, Nilges M, Cramer P, Rappsilber J (2010) Architecture of the RNA polymerase II–TFIIF complex revealed by cross-linking and mass spectrometry. EMBO J 29:717–726
6. Rappsilber J, Mann M, Ishihama Y (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc 2:1896–1906
7. Leitner A, Reischl R, Walzthoeni T, Herzog F, Bohn S, Förster F, Aebersold R (2012) Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography. Mol Cell Proteomics 11(3):M111.014126
8. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645

Chapter 12

Global Characterization of Protein Complexes by Biochemical Purification-Mass Spectrometry (BP/MS)

Reza Pourhaghighi and Andrew Emili

Abstract

A proteomic platform for global analysis of protein complexes and protein-protein interactions (PPIs) is described. Briefly, after comprehensive physiochemical separation of soluble protein extracts using non-denaturing ion exchange chromatography (IEX), each fraction is subjected to quantitative tandem mass spectrometry analysis.

Key words Protein-Protein Interaction, High Performance Liquid Chromatography (HPLC), Intact Protein Separation, Ion Exchange Chromatography (IEX), Nano-Liquid Chromatography Tandem Mass Spectrometry (nLC-MS/MS)

1 Introduction

Protein complexes are stable macromolecular assemblies responsible for different biochemical activities essential to cell homeostasis, growth, and proliferation. Hence, comprehensive study of composition and accurate identification of the interacting proteins in such multiprotein complexes is important and considered as a significant aspect of the cell biology.

We have recently developed a novel approach for global study of protein complexes based on the extensive complementary fractionation followed by in-depth quantitative MS profiling and strict computational filtering [1]. In this context, in order to separate and enrich native protein complexes, extremely deep biochemical fractionation is initially performed by employing multiple protein separation techniques such as ion exchange chromatography (IEX) and isoelectric focusing (IEF). Subsequently, quantitative mass spectrometry is used to identify stably associated interacting proteins that reproducibly co-elute through complex protein separations. Interactions are then scored based on the similarity and consistency of recorded protein co-fractionation patterns and supporting functional association evidence as additional constraints.

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_12,

In the present chapter, the detail experimental procedure for non-denaturing protein extraction form *E. coli* cells and native ion exchange-high performance liquid chromatography (IEX-HPLC) protein separations followed by sample preparation steps for a comprehensive quantitative mass spectrometry analysis of resulted fractions are explained. The complementary data analysis methods are described later in Chapter 25.

2 Materials

2.1 Equipment

1. Microcentrifuge with temperature control.
2. Vacuum pump with liquid trap suitable for aqueous filtrates.
3. Vortex mixer.
4. Rotatory shaker.
5. Incubating shaker.
6. Microcentrifuge tubes (1.5 mL).
7. Centrifugal microfilters.
8. HPLC-IEX column: Mixed-bed PolyCATWAX (PolyLC Inc.) 200 × 2.1 mm, 5 μm, 1000-Å (*see* **Note 1**).
9. HPLC system.
10. SpeedVac concentrator.
11. nLC-MS/MS system.

2.2 Reagents

It is essential that you consult the appropriate Material Safety Data Sheets and your institution's Environmental Health and Safety Office for proper handling and hazardous material in this protocol.

Use HPLC-grade solvents and water to prepare reagents required. Freshly prepare all the reagents.

1. Bradford reagent (commercially available).
2. Modified B-PER protein extraction buffer: Add 10% v/v glycerol, 0.5 mM DTT, 0.2 mg/mL lysozyme, 2 μL/mL DNase I, and 1× EDTA-free protease inhibitor to the commercial B-PER reagent (Thermo Scientifics) (*see* **Note 2**).
3. HPLC mobile phase-A: 10 mM Tris–HCl buffer pH 7, 5% glycerol, and 0.01% NaN_3.
4. HPLC mobile phase-B: 10 mM Tris–HCl buffer pH 7, 1.5 M NaCl, 5% glycerol, and 0.01% NaN_3.
5. Dithiothreitol (DTT) stock solution: Dissolve 7.7 mg DTT to obtain a DTT stock solution with final concentration of 0.5 M (*see* **Note 3**).

6. Iodoacetamide (IAA) stock solution: Dissolve 9.2 mg of IAA in 500 µL 50 mM NH_4CO_3, pH 8 to obtain a IAA stock solution of 0.1 M (*see* **Note 4**).
7. Sequencing grade trypsin.

3 Methods

Timing is critical throughout the protocol. Work quickly and consistently. Try to minimize the time especially before the protein sample is fractionated. Keep the sample on ice unless otherwise stated.

3.1 Soluble Protein Extraction

1. Pellet *E. coli* cells by centrifugation at 5000 × *g* for 10 min (*see* **Note 5**).
2. Add 4 mL of modified B-PER protein extraction buffer per gram of cell pellet, and mix gently until it is homogeneous.
3. Lightly mix the lysate for 30 min at 4 °C using a rotatory shaker or alternatively incubate it on ice.
4. Centrifuge lysate at 15,000 × *g* for 10 min at 4 °C to separate soluble proteins from the cell debris and insoluble proteins.
5. Gently remove the supernatant, and filter it by a 0.45 µm centrifugal filters using manufacturer's recommended time and speed.
6. Use Bradford protein assay to measure the protein concentration in lysate.

3.2 Ion Exchange Chromatography (HPLC-IEX)

A Tris–HCl pH 7 buffer system is typically suitable for IEX separation of most cell lysates. Proteins can be eluted from the IEX column with a salt (NaCl) gradient and be recovered in their biologically active forms.

1. Before running protein extract on HPLC, equilibrate the IEX column by running two blank gradients using buffers prepared. Re-equilibrate the column between gradient runs with buffer-A for 30 min.
2. Set the injection volume of the HPLC method to inject 1–1.5 mg of soluble protein extract into the IEX column.
3. The recommended flowrate for a 2.1-mm i.d. column is 0.2 mL/min.
4. Following could be used as a template gradient schedule for HPLC-IEX:
5. 100% A from 0 to 5 min, followed by a linear gradient to 15% B from 5 to 60 min, a linear gradient to 50% B from 60 to 90 min and then a gradient to 100% B from 90 to 100 min followed by an isocratic hold at 100% B until 120 min.

Tabfle 1
Summary of HPLC-IEX parameters

HPLC-IEX parameters	
Injection	1–1.5 mg
Flow rate	0.2 mL/min
Time (min)	LC gradient (%B)
0–5	0–0
5–60	0–15
60–90	15–50
90–100	50–100
100–120	100–100
Detection	280 nm
Fraction collection intervals	2 min

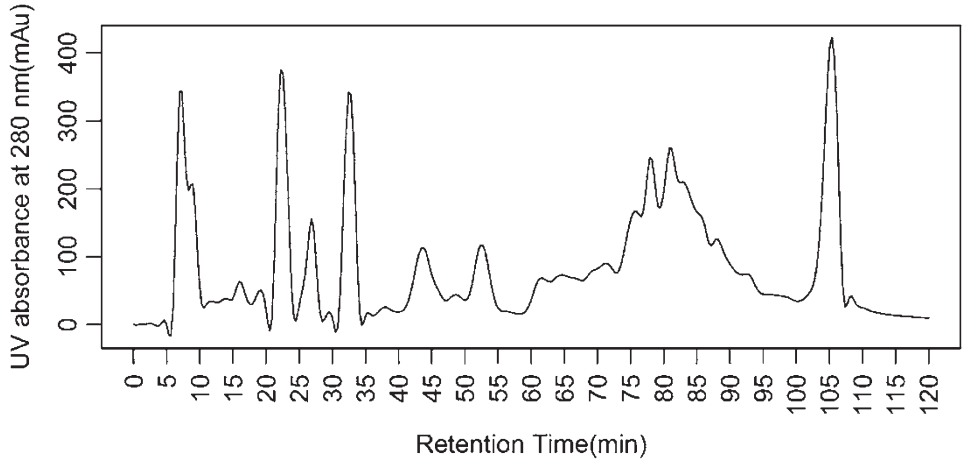


Fig. 1 Chromatogram obtained from HPLC-IEX separation of *E. coli* lysate at 280 nm. The other experimental parameters are summarized in Table 1

6. Protein elution could be monitored by absorption at 280 nm.
7. Collect the fractions with 2-min intervals.

Described HPLC-IEX parameters are summarized in Table 1. Figure 1 shows a chromatogram obtained from separation of *E. coli* lysate using abovementioned parameters.

3.3 Trichloroacetic Acid (TCA) Precipitation

1. If fractions are collected in 96-well plates, carefully transfer them into 1.5-mL tubes.
2. Precipitate the proteins by adding 10% v/v cold TCA to fractions and incubate them at 4 °C overnight.

3. Spin the samples in a microcentrifuge at 21,000 × *g*, 4 °C for 30 min.
4. Pipette out supernatant with care, leaving protein pellet intact.
5. Wash the pellet with 200 µL ice-cold acetone, and incubate it for an hour at −20 °C (*see* **Note 6**).
6. Repeat the acetone wash steps (**steps 3–5**) for one more time.
7. Spin the samples in at 21,000 × *g*, 4 °C for 30 min.
8. Remove supernatant and leave the protein pellet to air dry for about 30 min.

3.4 Trypsin Digestion

1. Dissolved the dried pellet in 90 µL 5 mM DTT, 50 mM NH_4CO_3 pH 8, and incubate the sample for 15 min at 50 °C with gentle agitation (*see* **Note 7**).
2. Bring the protein solution to room temperature, and add 10 µL 100 mM IAA to final concentration of 10 mM, and incubate the sample for 15 min at room temperature in the dark with gentle agitation.
3. Add 1 µL DTT from 0.5 M stock solution to obtain a final concentration of 5 mM in order to quench excess of IAA.
4. Add sequencing grade trypsin at 1:50 enzyme/protein ratio, and incubate the samples at 37 °C overnight with gentle agitation.
5. Quench the digestion with acidifying the solution by adding formic acid to final concentration of 1% v/v.
6. Lyophilize the peptides with vacuum centrifuge to dryness and dissolve them in 1% formic acid (*see* **Note 8**).

3.5 LC-MS/MS

1. Inject about 1 µg for samples an into the LC-MS/MS system.
2. A 60-min LC gradients as below is generally appropriate for LC-MS/MS analysis of IEX fractions: A linear gradient from 5% to 30% B from 0 to 46 min and then a gradient to 100% B from 46 to 50 min followed by an isocratic hold at 100% B until 60 min.
3. Recommended MS parameters for 60-min method in Orbitrap Q-Exactive HF are listed in Table 2.

3.6 Computational Proteomics Analysis

Search all MS/MS spectra with MaxQuant (or any available search engine) against the appropriate database. The calculated related intensity of proteins in each IEX fraction are then subjected to computational filtering and machine learning process to identify the high-confidence physical interactions among proteins which are described in detail in Chapter 25.

Table 2
Typical LC-MS/MS parameters for a 60-min run in Orbitrap Q-Exactive HF

nLC-MS/MS parameters	
Time (min)	*LC gradient (%B)*
0–46	5–30
46–50	30–100
50–60	100–100
Full-MS	
Microscans	1
Resolution	60,000
Automatic gain control target	3e6
Maximum ion time	70 ms
Number of scans	1
Scan range	300–1650 *m/z*
Dd-MS²	
Microscans	1
Resolution	15,000
Automatic gain control target	1e5
Maximum ion time	25 ms
Loop count	15
Isolation window	1.4 *m/z*
Normalized collision energy	27
Dd setting	
Charge exclusion	Unassigned, 1
Exclude isotopes	On
Dynamic exclusion	6 s

4 Notes

1. An IEX column with different dimension from what introduced here could be also used. However, experimental parameters like loading capacity and flow rate should be adjusted accordingly.
2. Add the DTT and protease inhibitor immediately before use. Keep the lysis buffer on ice.

3. DTT is susceptible to oxidation and should be prepared freshly before use. Keep the DTT solution on ice and freeze remaining at −20 °C for later use.
4. IAA is sensitive to light, and it should be freshly prepared and kept in the dark.
5. The bacterial cell pellet could be kept frozen at −80 °C. The method described in this chapter works for both fresh and frozen cell pellets.
6. Incubate the acetone bottle in −20 °C for at least 1 h before adding to sample.
7. Make sure that the pH solution is above 7.5 to avoid alkylation of lysine and histidine.
8. Using Ziptip C18 tips for further cleaning the samples is recommended but not needed.

Reference

1. Havugimana PC, Hart GT, Nepusz T et al (2012) A census of human soluble protein complexes. Cell 150(5):1068–1081

Chapter 13

Proteomic Profiling of Integrin Adhesion Complex Assembly

Adam Byron

Abstract

Cell adhesion to components of the cellular microenvironment via cell-surface adhesion receptors controls many aspects of cell behavior in a range of physiological and pathological processes. Multimolecular complexes of scaffolding and signaling proteins are recruited to the intracellular domains of adhesion receptors such as integrins, and these adhesion complexes tether the cytoskeleton to the plasma membrane and compartmentalize cellular signaling events. Integrin adhesion complexes are highly dynamic, and their assembly is tightly regulated. Comprehensive, unbiased, quantitative analyses of the composition of different adhesion complexes over the course of their formation will enable better understanding of how the dynamics of adhesion protein recruitment influence the functions of adhesion complexes in fundamental cellular processes. Here, a pipeline is detailed integrating biochemical isolation of integrin adhesion complexes during a time course, quantitative proteomic analysis of isolated adhesion complexes, and computational analysis of temporal proteomic data. This approach enables the characterization of adhesion complex composition and dynamics during complex assembly.

Key words Bioinformatics, Cell adhesion, Cell signaling, Data analysis, Hierarchical clustering, Integrins, Interaction networks, Proteomics

1 Introduction

Cells use adhesion receptors on the plasma membrane to bind molecules in their local environment. Proteins, glycoproteins, and proteoglycans form a complex milieu of structural and nonstructural extracellular matrix (ECM), soluble factors, and cell-surface counter-receptors with which cells can interact using adhesion receptors to sense their surroundings [1]. Engagement of extracellular ligands with integrin adhesion receptors triggers the formation of intracellular, integrin-associated complexes of adhesion proteins, which serve to scaffold cytoplasmic connections to the contractile cytoskeleton and transmit mechanical and biochemical signals bidirectionally across the plasma membrane [2]. The coordination and processing of these signaling cues by integrin adhe-

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764,
https://doi.org/10.1007/978-1-4939-7759-8_13,

sion complexes controls many fundamental aspects of cell behavior, including cell proliferation, migration, and differentiation. Consequently, dysregulation of adhesion scaffolding and signaling networks can contribute to various disease states [3–12].

In this chapter, methods for the proteomic characterization of temporal profiles of integrin adhesion complex assembly are detailed. The biochemical purification of adhesion complexes during a time course allows analysis of "average" snapshots of adhesion complex composition following induction by a bead-immobilized extracellular ligand. Mass spectrometry (MS)-based proteomic analysis of isolated complexes enables identification and quantification of adhesion proteins, and downstream data mining can be applied to detect and explore proteomic data features and distill adhesion protein assembly profiles. Together, this pipeline enables the proteomic profiling of the dynamics of adhesion complex assembly (Fig. 1). Results obtained using this approach serve as a valuable foundation for further in-depth validation, functional analysis, and biological experimentation.

1.1 Adhesion Complex Assembly

Integrins are the best-characterized family of cell–ECM adhesion receptors, and they interact with a range of extracellular ligands [13]. Binding of integrins to their ligands initiates the intracellular assembly of adhesion proteins into integrin-bound adhesion complexes, which can be observed by microscopy techniques as plaque-like structures known as focal contacts, focal adhesions, and fibrillar adhesions, which represent different stages of adhesion complex maturation and mechanical properties or applied external forces [14]. The assembly of adhesion complexes is tightly and dynamically regulated, but the precise interactions of adhesion proteins are not well defined spatially or temporally. Fluorescence microscopy techniques have been used to examine the sequential recruitment of tagged candidate adhesion proteins to small, early adhesion complexes [15, 16], suggesting a model of hierarchical adhesion protein recruitment (Fig. 2). Furthermore, several adhesion proteins (e.g., talin and vinculin) appear to form "pre-complexes" before associating with integrin-bound adhesion complexes [16, 17]. Indeed, the modulation of adhesion protein conformation by mechanical force and mechanosensitive protein interactions (including those of talin and vinculin) play important roles in adhesion complex assembly and coupling to the actin cytoskeleton [19–24].

Adhesion receptors other than integrins also form complexes of adhesion proteins. For example, the cell–ECM adhesion receptor syndecans—heparan sulfate proteoglycans that also act as growth factor and chemokine coreceptors—recruit adhesion signaling proteins to their short intracellular domains [25]. Syndecans can modulate integrin trafficking and integrin adhesion complex dynamics, and thus they cooperate with integrins to regulate cell

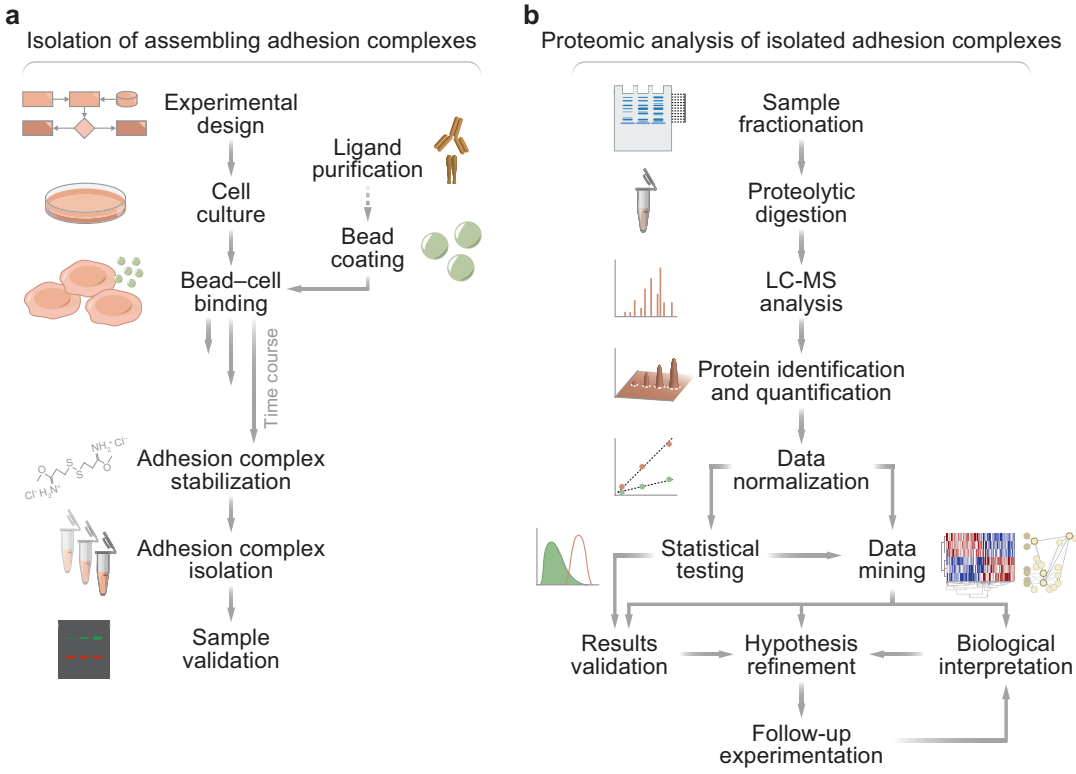


Fig. 1 A pipeline for the proteomic profiling of integrin adhesion complex assembly. (**a**) Key stages of the workflow for isolation of assembling adhesion complexes. Dashed arrow indicates optional step, depending on experimental requirements. (**b**) Key stages of the workflow for proteomic analysis of isolated adhesion complexes. In addition to cluster and interaction network analyses (detailed herein), data mining techniques that can be applied to the proteomic data include dimensionality reduction (e.g., principal component analysis, self-organizing maps, partial least squares), class prediction (e.g., random forests, support vector machines, artificial neural networks), and pathway analysis (e.g., gene set enrichment analysis, signaling pathway impact, sensitivity analysis)

adhesion and migration [26, 27]. The cell–cell adhesion receptors cadherins connect to the actin cytoskeleton via catenins, with numerous adaptor and cytoskeletal proteins recruited to cadherin-associated complexes [28], and cytoskeletal tension plays an important role in regulating the dynamics of cadherin adhesion complexes [29, 30]. Although it may be possible to analyze the assembly of adhesion complexes associated with non-integrin adhesion receptors using the methods detailed herein, the biochemical isolation protocol is optimized for the analysis of integrin adhesion complexes.

1.2 Integrin Adhesion Complex Proteomes

The sum of scaffolding and signaling proteins that localize to integrin adhesion complexes has been termed the "adhesome" [31], and the latest literature-curated adhesome database (derived from studies using multiple cell types) contains 232 proteins [3].

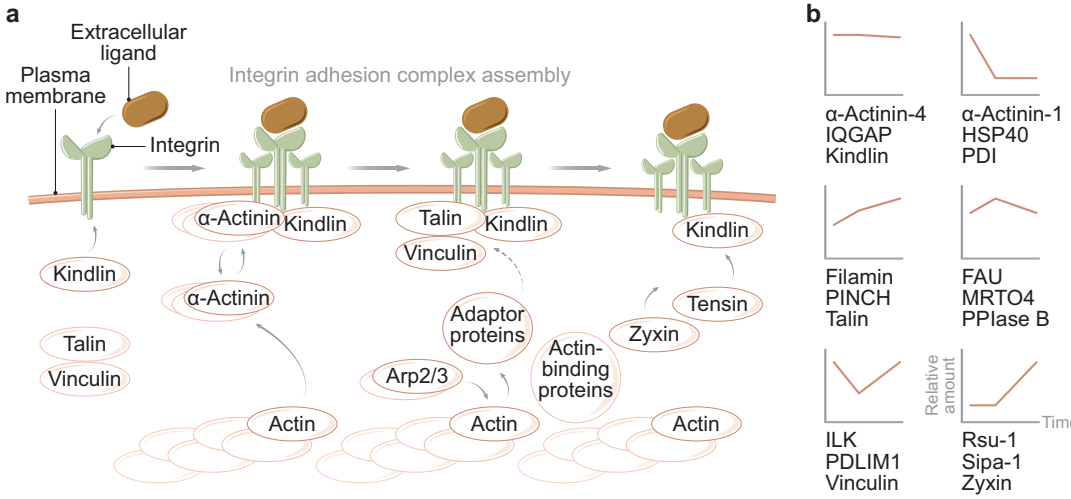


Fig. 2 Assembly dynamics of integrin adhesion complexes. (**a**) Hierarchical recruitment of adhesion proteins to ligand-bound integrin, as determined by fluorescence microscopy studies [15–17]. (**b**) Profiling of consensus adhesome protein recruitment, as determined by proteomic experiments [18]. Line profiles show trends of relative protein abundance during adhesion complex assembly, and respective examples of consensus adhesome proteins are indicated for each profile

Adhesion complexes exist as labile, plasma membrane- and cytoskeleton-linked multimolecular complexes, confounding their purification by conventional coimmunoprecipitation approaches and precluding their comprehensive, unbiased analysis until recently. The development of biochemical techniques for the isolation of ligand-induced adhesion complexes has enabled the characterization of integrin adhesion complexes by MS-based proteomics [32–34]. Initial MS analyses of integrin adhesion complexes revealed their considerable molecular complexity and diversity [35–37], suggestive of their important and multiple roles in cellular signaling and the regulation of cell behavior. Adaptation of a bead-based, steady-state adhesion complex isolation method for the proteomic analysis of time-resolved integrin adhesion complexes, as detailed herein, enabled the temporal profiling of adhesion complexes [18]. Proteomic profiling revealed distinct recruitment dynamics of proteins involved in specific functional processes, with many core adhesion proteins detected in high abundance later in adhesion complex assembly [18].

The multiple time points, ligands, and treatments that could be tested in a single experiment using the approach described herein result in data of multiple dimensions, so the interrogation and biological interpretation of such multidimensional data can be a time-consuming task. To expedite this, a range of computational techniques for data mining, such as cluster analysis and interaction network analysis [38], can be used to analyze higher-level data features, predict functional outcomes, and generate new mechanistic

hypotheses that can then be tested experimentally. Cluster analysis is a type of multivariate statistical analysis that groups together elements (proteins, transcripts, genes, etc.) with similar detection or expression profiles across similar samples in the experiment [39, 40]. Using hierarchical clustering, quantitative proteomic data can be organized naturally, in an unsupervised manner, according to the structure of the dataset, which can help identify sets of distinct profiles of protein recruitment to adhesion complexes. Network analysis uses reported and predicted interactions to add biological context to experimental observations [41, 42]. Using interaction network analysis, changes in adhesion complex composition during complex assembly can be mapped and intuitively visualized as a graph of nodes (representing proteins) and partitioned into subnetworks according to the distribution of edges (representing putative protein–protein relationships). The computational integration of multiple integrin adhesion complex proteomes generated an experimentally defined "meta-adhesome," from which a core set of 60 frequently identified integrin-associated proteins (a consensus adhesome) was established [18]. Interaction networks derived from the meta-adhesome and consensus adhesome provide more coherent representations of potential interactions between adhesion complex components and thus are useful "master" graphs from which to filter and interrogate mapped adhesion complex proteomic data more readily.

1.3 Modifying the Method

Numerous aspects of the method described herein could be adapted or extended according to project requirements or user interests. For example, the ligand for adhesion complex isolation can be selected to determine which adhesion receptors are ligated and induced to form adhesion complexes. The ECM glycoprotein fibronectin has been successfully used to isolate adhesion complexes associated primarily with α5β1 integrin (depending on the integrin expression profile of the cell line) [18, 35]. A recombinant soluble form of the cell adhesion molecule vascular cell adhesion molecule 1 (VCAM-1) has been used to isolate adhesion complexes associated with α4β1 integrin [35, 43, 44]. VCAM-1 is a cell-surface integrin counter-receptor, so it is possible that this protocol could be used for the isolation and analysis of other cell–cell adhesion complexes. Considerations such as ligand solubility, ligand affinity, and ligand orientation upon conjugation to beads will affect the suitability of a ligand for this method. In addition, monoclonal antibodies against integrins have been used to isolate adhesion complexes associated with activated β1 integrin [45], so this protocol could be readily extended to use other adhesion receptor-specific antibodies that induce adhesion complex assembly.

Owing to the large number of cells required to yield sufficient isolated protein for comprehensive MS analysis and to the mechan-

ics associated with the bead-based purification technique, the adhesion complex isolation protocol described herein has been optimized for mammalian cells maintained in suspension. The method could be applied to various cell types that can be grown in suspension or kept in suspension for the duration of the experiment. The bead-based purification technique enables rapid isolation of integrin adhesion complexes, so it is appropriate for time-resolved capture of adhesion complexes, including at early time points of complex formation. Bead incubation times can be varied to capture different stages (albeit "average" snapshots) of adhesion complex assembly. In addition, the ligand-coated beads are superparamagnetic and can be manipulated with a magnet to apply tensile forces to adhesion receptors [46], which, if scaled up accordingly, could enable the investigation of force-dependent adhesion complex assembly.

Label-free MS-based quantification of protein samples is straightforward to implement and broadly applicable to most cell systems, and it has been previously used for the quantification of bead-isolated integrin adhesion complex components [18, 35, 44, 45]. The application of MS ion intensity-based label-free quantification using a high-resolution mass spectrometer is detailed herein. This protocol could be modified to use other label-free approaches [47–49] or label-based approaches, such as metabolic labeling [50] or chemical isobaric labeling [51–53], or targeted MS techniques, such as data-independent acquisition [54–56].

Posttranslational modifications of proteins are important biochemical events that mediate and influence cellular signal transduction. Phosphorylation of serine, threonine, and tyrosine residues, for example, plays a central role in the spatial and temporal regulation of adhesion signaling [57–65]. Biochemical, proteomic, and imaging approaches have been used to measure general and site-specific protein phosphorylation induced by cell adhesion or localized to sites of adhesion [66–71]. Although this chapter focuses on the proteomic identification and quantification of, chiefly, unmodified peptides derived from isolated adhesion proteins during adhesion complex assembly, it may be possible to adapt the protocol detailed herein to incorporate analysis of posttranslational modifications of adhesion complex components. For the analysis of phosphorylation, for example, key analytical challenges include the low stoichiometry of protein phosphorylation, the low abundance of phosphorylated proteins, the highly dynamic and regulated nature of phosphorylation in cells, and the increased hydrophilicity of phosphorylated peptides (phosphopeptides), which can all confound the unbiased detection of phosphopeptides by liquid chromatography (LC)-MS techniques [72]. To overcome these challenges, users may be required to scale-up the protocol to obtain sufficient yields of phosphorylated proteins, to perform selective phosphopeptide enrichment to reduce sample complexity and the

likely excess of more abundant unmodified peptides, and to modify the LC system to increase retention of hydrophilic phosphopeptides. An alternative adhesion complex isolation methodology using integrin ligand-coated plates has been successfully adapted for the analysis of phosphorylated adhesion complex proteins, and interested users are first directed to the corresponding protocol by Robertson et al. [73]. Using this approach, MS analysis revealed that phosphorylation events in adhesion complexes can result from adhesion-induced phosphorylation of resident adhesion complex proteins or from recruitment of constitutively phosphorylated proteins to adhesion complexes [71]. Since many posttranslational modifications are transient, analyzing more completely the temporal regulation of adhesion protein modifications during adhesion complex assembly could provide tremendous insight into cell adhesion signaling mechanisms.

2 Materials

This protocol involves the use of substances that are hazardous to health. Consult the material safety data sheets and assess associated health risks prior to the commencement of work. Store, handle, and dispose hazardous substances in accordance with local and regional health and safety requirements.

2.1 Isolation of Assembling Adhesion Complexes

Owing to the time-sensitive nature of several of the protocol steps, it is recommended to prepare stock solutions and reagents in advance of adhesion complex isolation wherever possible and appropriate. For example, stock solutions of 5% (w/v) Triton X-100 and 1 M sucrose can be prepared in advance of making the cytoskeletal stabilizing (CSK) buffers and stored at 4 °C for up to 1 week. CSK buffers should be used on the day of preparation.

1. Standard laboratory attire and personal protective equipment, as required.
2. Cells of choice.
3. Appropriate cell culture medium and growth supplements, as required.
4. Cell culture vessels and other necessary plasticware and laboratory equipment for maintaining and passaging cells.
5. Sterile phosphate-buffered saline (PBS) for passaging cells.
6. Trypsin (or similar) or cell dissociation buffer, if required, for passaging adherent cells.
7. Humidified cell culture incubator for maintaining cells.
8. Hemocytometer or other cell counting device.
9. Light microscope.

10. 4.5 μm-diameter tosyl-activated superparamagnetic polystyrene beads (Dynabeads M-450).
11. Vortex mixer.
12. 1.5 mL microcentrifuge tubes.
13. 0.1 M sodium phosphate buffer (PB): 19 mM sodium phosphate monobasic, 81 mM sodium phosphate dibasic, pH 7.4.
14. Magnetic separator for 1.5 mL microcentrifuge tubes (DynaMag-2).
15. Adhesion receptor-specific ligand or antibody of choice for bead conjugation.
16. Non-adhesion receptor-specific control ligand or antibody of choice for bead conjugation.
17. Temperature-controlled shaking incubator for 1.5 mL microcentrifuge tubes (ThermoMixer C).
18. Bovine serum albumin (BSA).
19. PBS without calcium or magnesium.
20. 1 M Tris–HCl, pH 8.5 (*see* **Note 1**).
21. 2% (w/v) sodium azide, if required (*see* **Note 1**).
22. Dulbecco's modified Eagle's medium containing 25 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (DMEM-HEPES).
23. CSK buffer: 10 mM 1,4-piperazinediethanesulfonic acid, pH 6.8, 50 mM sodium chloride, 150 mM sucrose, 3 mM magnesium chloride, 1 mM manganese chloride (*see* **Note 1**).
24. CSK-Tris buffer: CSK buffer containing 20 mM Tris–HCl, pH 8.5, 10 μg/mL leupeptin, 10 μg/mL aprotinin, 0.5 mM 4-(2-aminoethyl)benzenesulfonyl fluoride hydrochloride (AEBSF), 2 mM sodium orthovanadate (*see* **Note 1**).
25. CSK+ buffer: CSK buffer containing 0.5% (w/v) Triton X-100, 10 μg/mL leupeptin, 10 μg/mL aprotinin, 0.5 mM AEBSF, 2 mM sodium orthovanadate (*see* **Note 1**).
26. Benchtop centrifuge capable of holding 50 mL centrifuge tubes.
27. Radioimmunoprecipitation assay (RIPA) buffer, if required, for whole cell lysis: 50 mM Tris–HCl, pH 8.0, 5 mM ethylenediaminetetraacetic acid (EDTA), 150 mM sodium chloride, 1% (w/v) Triton X-100, 1% (w/v) sodium deoxycholate, 0.1% (w/v) sodium dodecyl sulfate (SDS), 10 μg/mL leupeptin, 10 μg/mL aprotinin, 0.5 mM AEBSF, 2 mM sodium orthovanadate (*see* **Note 1**).
28. Total protein assay kit (Pierce BCA Protein Assay Kit), if required, for whole cell lysis.

29. 1 M manganese chloride (*see* **Note 1**).
30. 50 mL centrifuge tubes.
31. Moisture-resistant, flexible, self-sealing plastic film (Parafilm M).
32. Temperature-controlled shaking incubator for horizontal centrifuge tubes (New Brunswick Innova 44).
33. Dimethyl 3,3′-dithiobispropionimidate (DTBP) cross-linker (*see* **Note 1**).
34. 15 mL centrifuge tubes.
35. 3 mL plastic Pasteur pipettes.
36. Magnetic separator for 15 mL centrifuge tubes (DynaMag-15).
37. Microscope slides and coverslips.
38. 50% (v/v) glutaraldehyde, 0.1% (w/v) crystal violet, 10% (v/v) acetic acid, 96-well microtiter plate, 96-well-plate reader (absorbance, 570 nm), if required, for crystal violet assay (*see* **Note 1**).
39. Bioruptor Standard ultrasonication device.
40. Low-protein-binding 1.5 mL microcentrifuge tubes (Protein LoBind Tubes).
41. Reducing sample buffer (5× stock): 125 mM Tris–HCl, pH 6.8, 25% (w/v) glycerol, 10% (w/v) SDS, 20% (v/v) β-mercaptoethanol (*see* **Note 1**). Supplement reducing sample buffer (5× stock) with 0.01% (w/v) bromophenol blue, if required, for polyacrylamide gel electrophoresis (PAGE) (*see* **Note 1**).
42. Refrigerated benchtop microcentrifuge capable of holding 1.5 mL microcentrifuge tubes.
43. SDS running buffer (NuPAGE MES SDS Running Buffer).
44. 4–12% (w/v) polyacrylamide gels (10-well, 1 mm-thickness, 4–12% NuPAGE Bis-Tris Mini Gels).
45. Gel-running tank (XCell *SureLock* Mini-Cell).
46. Gel-loading pipette tips.
47. Protein standards (10–250 kDa Precision Plus Protein All Blue Prestained Protein Standards).
48. Gel-running tank power supply.
49. Gel knife.
50. Transfer buffer (NuPAGE Transfer Buffer).
51. Methanol (*see* **Note 1**).
52. Blotting pads (sponges) of the dimensions of the polyacrylamide gels.
53. Filter paper cut to the dimensions of the polyacrylamide gels.

54. Transfer membrane (0.45 μm-pore size Whatman Protran Nitrocellulose Membrane) cut to the dimensions of the polyacrylamide gels.
55. Tweezers.
56. Glass pipette.
57. Blotting module (XCell II Blot Module).
58. 0.1% (w/v) Ponceau S in 7% (v/v) trichloroacetic acid, if required, for transfer membrane staining (*see* **Note 1**).
59. Coomassie-based protein stain (InstantBlue), if required, for gel staining (*see* **Note 1**).
60. Blocking buffer (Odyssey Blocking Buffer (TBS)).
61. Tilting platform.
62. Tween 20.
63. Tris-buffered saline (TBS).
64. Adhesion protein-specific antibodies of choice validated for immunoblotting.
65. Non-adhesion protein-specific control antibodies of choice validated for immunoblotting.
66. Appropriate secondary antibodies (IRDye family secondary antibodies).
67. Immunoblotting imager (Odyssey IR imaging system).
68. Immunoblotting image analysis software (Image Studio).

2.2 Proteomic Analysis of Isolated Adhesion Complexes

Use LC-MS-grade reagents, store solvents appropriately in glass bottles, and avoid plasticware wherever possible when preparing samples for proteomic analysis to avoid potential contamination from plasticizers. Some equipment for proteomic analysis (e.g., C18 analytical column) can be constructed in-house, but commercial alternatives are available for purchase.

1. Standard laboratory attire and personal protective equipment, as required.
2. SDS running buffer (NuPAGE MES SDS Running Buffer).
3. 4–12% (w/v) polyacrylamide gels (10-well, 1 mm-thickness, 4–12% NuPAGE Bis-Tris Mini Gels).
4. 3 mL plastic Pasteur pipettes.
5. Gel-running tank (XCell *SureLock* Mini-Cell).
6. Gel-loading pipette tips.
7. Protein standards (10–250 kDa Precision Plus Protein All Blue Prestained Protein Standards).
8. Gel-running tank power supply.
9. Gel knife.

10. Coomassie-based protein stain (InstantBlue) (*see* **Note 1**).
11. Tilting platform.
12. Scalpel blades.
13. Low-protein-binding perforated 96-well microtiter plates and low-protein-binding 96-well collection plates.
14. Acetonitrile (*see* **Note 1**).
15. 25 mM ammonium bicarbonate (*see* **Note 1**).
16. Benchtop centrifuge capable of holding 96-well microtiter plates.
17. Vacuum centrifuge capable of holding 96-well microtiter plates.
18. Sample oven with adjustable temperature control capable of maintaining temperatures up to at least 56 °C.
19. Dithiothreitol (DTT) (*see* **Note 1**).
20. Iodoacetamide (*see* **Note 1**).
21. Trypsin with limited autolytic activity (sequencing-grade modified trypsin) (*see* **Note 1**).
22. Trypsin resuspension buffer.
23. Formic acid (*see* **Note 1**).
24. Octadecyl (C18)-bonded silica disks (3M Empore SPE Extraction Disks) and disk cutter (gauge 16 blunt-tip needle and 25 μL syringe plunger assembly; assembled in-house).
25. Trifluoroacetic acid (TFA) (*see* **Note 1**).
26. Low-bleed 96-well autosampler plates with sealing mats with septa.
27. Ultrahigh-performance LC system (Dionex UltiMate 3000 RSLCnano).
28. Accurate-mass, high-resolution mass spectrometer (Q Exactive Plus Hybrid Quadrupole-Orbitrap).
29. LC-MS instrument operating software (XCalibur).
30. C18 analytical column (15 cm, 75 μm-inner diameter column packed with 1.8 μm C18 particles; pulled and packed in-house).
31. Personal computer with at least midrange, modern system specifications (e.g., at least 1 GHz dual-core processor, 2 GB memory, 1 GB graphics card).
32. Web browser (Chrome).
33. Data analysis software platform for identification and quantification of proteins from raw MS data (MaxQuant [74]).
34. Peptide search engine (Andromeda [75]).
35. Software framework, if required (.NET Framework).

36. Vendor-specific raw MS data file reader library, if required (MSFileReader).
37. Spreadsheet program (Excel).
38. Data analysis software platform for statistical and bioinformatic analysis of processed MS data (Perseus [76]).
39. Java platform, if required (Java Standard Edition 9).
40. Clustering software (Cluster 3.0 [77]).
41. Clustering data visualization software (Java TreeView [78]).
42. Interaction network analysis software (Cytoscape [79]).

3 Methods

To enable proteomic profiling of integrin adhesion complex assembly, adhesion complexes are stabilized and isolated at various time points, analyzed using quantitative LC-MS, and resultant proteomic data interrogated using computational tools (Fig. 1). The complex isolation method is technically challenging, and experiments should be conducted carefully to ensure reproducible and specific protein purification. Extensive method optimization may be required for chosen parameters and project-specific systems, and all experiments should incorporate appropriate controls and quality control measures (detailed below). It is also important to validate findings from the proteomic results for a portion of the dataset using an independent experimental method, such as immunoblotting, immunocytochemistry, immunohistochemistry, or reverse-phase protein array, to enable firm conclusions to be drawn. Furthermore, these data typically represent the starting point for further analysis or hypothesis generation.

The entire pipeline for the proteomic profiling of integrin adhesion complex assembly can take several weeks to complete. For the isolation of assembling adhesion complexes, preparation of cells and beads requires 2–3 days (Subheading 3.1.1); adhesion complex purification requires 1 day (Subheading 3.1.2); and sample validation requires 2–3 days (Subheading 3.1.3). For the proteomic analysis of isolated adhesion complexes, proteolytic digestion requires 2–3 days (Subheading 3.2.1); LC-MS analysis requires 3–13 days, depending on sample number (Subheading 3.2.2); and proteomic data analysis requires 4–12 days, depending on sample number and computing power (Subheading 3.2.3).

3.1 Isolation of Assembling Adhesion Complexes

This method has been optimized for mammalian cells maintained in suspension (specifically, the K562 erythroleukemia cell line) owing to the large number of cells required and the mechanics of the bead-based isolation approach. The protocol could be applied to a range of cell types that can be grown in suspension or kept in

suspension for a short period of time. The adhesion receptor-specific ligand or antibody for adhesion complex isolation can be chosen to determine which adhesion receptors are induced to form adhesion complexes. Fibronectin, which has been successfully used to isolate adhesion complexes associated primarily with α5β1 integrin [18, 35], is—depending on the integrin expression profile of the chosen cell line—a good starting point for the ligand choice in this protocol. If available, monoclonal antibodies that inhibit adhesion receptor–ligand binding, such as those available for several integrins [80], can be used to block adhesion complex recruitment to ligand-coated beads and thus establish the specificity of the chosen adhesion receptor ligand [43]. Uncoated or BSA-blocked beads should not bind cells, so they do not make satisfactory negative controls for the assessment of adhesion complex-specific protein recruitment [43]. Instead, beads conjugated with non-adhesion receptor-specific ligands or antibodies (e.g., poly-D-lysine, anti-transferrin receptor antibody), which are able to bind cells to the same extent as adhesion receptor-specific ligands or antibodies but not induce adhesion complex formation, should be used to determine the recruitment of specific adhesion complex components.

3.1.1 Preparation of Cells and Beads

1. Culture cells using standard, sterile growth conditions appropriate for the cell type (*see* **Note 2**). Two or 3 days before adhesion complex isolation, passage cells accordingly to yield approximately 1×10^8 cells in log-phase growth per experimental condition at the time of adhesion complex isolation (*see* **Note 3**).
2. Two days before adhesion complex isolation, resuspend the stock suspension of superparamagnetic polystyrene beads (Dynabeads M-450) thoroughly by vortexing the vial for at least 30 s (*see* **Note 4**).
3. Transfer a suspension containing 5×10^7 beads per experimental condition to a 1.5 mL microcentrifuge tube (*see* **Note 5**).
4. Add the initial bead suspension volume of PB to the beads in the microcentrifuge tube.
5. Place the microcentrifuge tube in a magnetic separator (DynaMag-2) for 1 min, and discard the supernatant (*see* **Note 6**).
6. Remove the microcentrifuge tube from the magnetic separator, and wash beads twice with the initial bead suspension volume of PB.
7. Resuspend the washed beads in an appropriate volume of PB such that 15–25 μg ligand is present per 5×10^7 beads at $4–8 \times 10^8$ beads/mL (*see* **Note 7**). Remember to subtract the volume of ligand to be added (*see* **step 8**) from the volume of PB used to resuspend the beads.

8. Add ligand to a final concentration of 200 μg/mL, and mix thoroughly by vortexing (*see* **Note 8**).
9. Gently rotate or agitate beads in a temperature-controlled shaking incubator (ThermoMixer C) at 25 °C for 15 min.
10. Add 10% (w/v) BSA in PB to beads to a final BSA concentration of 0.01–0.1% (w/v) (*see* **Note 9**).
11. Continue gentle rotation or agitation of beads in the shaking incubator at 25 °C for 24 h (*see* **Note 10**).
12. Place the microcentrifuge tube in the magnetic separator for 1 min, and discard the supernatant.
13. Remove the microcentrifuge tube from the magnetic separator, and wash beads twice with 1 mL 0.1% (w/v) BSA in PBS, with gentle rotation or agitation of beads at 4 °C for 5 min.
14. Place the microcentrifuge tube in the magnetic separator for 1 min, and discard the supernatant.
15. To deactivate remaining free tosyl groups, incubate the beads in 1 mL 0.1% (w/v) BSA in 0.2 M Tris–HCl, pH 8.5, with gentle rotation or agitation of beads at 25 °C for 16 h.
16. Place the microcentrifuge tube in the magnetic separator for 1 min, and discard the supernatant.
17. Remove the microcentrifuge tube from the magnetic separator, and wash beads twice with 1 mL 0.1% (w/v) BSA in PBS, with gentle rotation or agitation of beads at 4 °C for 5 min.
18. If necessary, store coated, blocked, washed beads (optionally, in the presence of 0.02% (w/v) sodium azide) at 4 °C for up to 1 month (*see* **Note 11**). Alternatively, proceed to adhesion complex purification (*see* Subheading 3.1.2, **step 1**).

3.1.2 Adhesion Complex Purification

1. Pre-warm to 37 °C DMEM-HEPES and 0.2% (w/v) BSA in DMEM-HEPES.
2. Prepare and prechill to 4 °C CSK, CSK-Tris, and CSK+ buffers but do not supplement with protease and phosphatase inhibitors, where applicable, until immediately before use.
3. Place the microcentrifuge tube containing coated, blocked, washed beads in the magnetic separator for 1 min, and discard the supernatant.
4. Remove the microcentrifuge tube from the magnetic separator, and wash beads once with 1 mL DMEM-HEPES.
5. If using adherent cells, wash cells in cell culture vessels with PBS, then detach cells with trypsin (or similar) at 37 °C for 5 min. Decant detached cells into pre-warmed serum-containing cell culture medium to quench the trypsin.
6. For both suspension and adherent cells, collect cells by centrifugation at 200 × *g* for 5 min, and discard the supernatant.

7. Wash the cell pellet with 20 mL of pre-warmed DMEM-HEPES.
8. Count cells using a hemocytometer.
9. Collect cells by centrifugation at 200 × *g* for 5 min, and discard the supernatant (*see* **Note 12**).
10. Resuspend the cell pellet in a minimum of 10.5 mL pre-warmed 0.2% (w/v) BSA in DMEM-HEPES supplemented with 0.2 mM manganese chloride per experimental condition to give a concentration of 1×10^7 cells/mL.
11. Rest cells in a humidified cell culture incubator at 37 °C for 10 min.
12. Meanwhile, resuspend coated, blocked, washed beads (*see* Subheading 3.1.1, **step 18**) in 0.2% (w/v) BSA in DMEM-HEPES to give a concentration of 2×10^8 beads/mL.
13. For each experimental condition, transfer 1×10^8 rested cells (10 mL) to a 50 mL centrifuge tube, and add 5×10^7 beads in DMEM-HEPES to give a bead-to-cell ratio of 1:2 (5×10^7 beads to 1×10^8 cells) (*see* **Note 13**).
14. If necessary, wrap self-sealing plastic film (Parafilm M) around the closed lid of each centrifuge tube to prevent leakage.
15. Immediately rotate the bead-cell suspension in a near-horizontal orientation in a shaking incubator (New Brunswick Innova 44) at 70 rpm at 37 °C for the desired incubation time (*see* **Notes 14** and **15**).
16. Prepare a stock solution of 100 mM DTBP cross-linker in PBS immediately before use (*see* **Note 16**).
17. Add DTBP cross-linker to the bead-cell suspension to give a final concentration of 10 mM DTBP, and immediately rotate the bead-cell suspension in a shaking incubator at 70 rpm at 37 °C for 2 min (*see* **Notes 17** and **18**).
18. Add Tris–HCl, pH 8.5, to the bead-cell suspension to give a final concentration of 20 mM Tris–HCl, and incubate at room temperature for 2 min (*see* **Note 19**).
19. For each experimental condition, gently transfer the bead-cell suspension to a prechilled 15 mL centrifuge tube using a 3 mL plastic Pasteur pipette.
20. Place each centrifuge tube in a magnetic separator (DynaMag-15) on ice for 2 min, and gently remove and discard the supernatant (*see* **Note 6**).
21. Remove the centrifuge tubes from the magnetic separator, and wash bead-bound cells once with 5 mL prechilled CSK-Tris buffer using gentle pipetting.
22. Place each centrifuge tube in the magnetic separator on ice for 2 min, and gently remove and discard the supernatant.

23. Remove the centrifuge tubes from the magnetic separator, and wash bead-bound cells once with 5 mL prechilled CSK buffer using gentle pipetting.
24. Remove 5 μL bead-bound cells from each centrifuge tube, place on a microscope slide, and cover with a coverslip. As a quality control measure, assess bead-cell binding by light microscopy (*see* **Notes 20** and **21**).
25. Place each centrifuge tube in the magnetic separator on ice for 2 min, and gently remove and discard the supernatant.
26. Remove the centrifuge tubes from the magnetic separator, and resuspend bead-bound cells in 4 mL prechilled CSK+ buffer using gentle pipetting.
27. Sonicate the bead-cell suspension using an ultrasonication device (Bioruptor Standard) (*see* **Notes 22–25**).
28. Remove 5 μL beads from each centrifuge tube, place on a microscope slide, and cover with a coverslip. As a quality control measure, assess cell lysis by light microscopy (*see* **Note 26**).
29. Place each centrifuge tube in the magnetic separator on ice for 2 min, and gently remove and discard the supernatant.
30. Remove the centrifuge tubes from the magnetic separator, and wash beads four times with 5 mL prechilled CSK+ buffer using gentle pipetting.
31. Place each centrifuge tube in the magnetic separator on ice for 2 min, and gently remove and discard the supernatant.
32. Remove the centrifuge tubes from the magnetic separator, resuspend beads in 1 mL prechilled CSK+ buffer using gentle pipetting, and transfer samples to prechilled, low-protein-binding 1.5 mL microcentrifuge tubes.
33. Place each microcentrifuge tube in a magnetic separator (DynaMag-2) on ice for 2 min, and gently remove and discard the supernatant.
34. To cleave the cross-linker and elute proteins from the beads, add 150 μL 2× reducing sample buffer to the beads, and agitate beads in a temperature-controlled shaking incubator (ThermoMixer C) at 70 °C for 30 min then 95 °C for 5 min (*see* **Note 27**).
35. Place each microcentrifuge tube in the magnetic separator for 2 min, and collect and retain the supernatant.
36. If necessary, store eluted protein samples at −80 °C for up to 3 months (*see* **Note 28**). Alternatively, proceed to sample validation (*see* Subheading 3.1.3, **step 1**).

3.1.3 Sample Validation

1. Heat protein samples (*see* Subheading 3.1.2, **step 36**) at 70 °C for 10 min in preparation for denaturing, reducing SDS-PAGE. SDS-PAGE can be used in combination with immunoblotting as a quality control measure prior to in-depth proteomic analysis by LC-MS to validate the specificity of the adhesion complex purification (*see* **Note 29**).
2. Prepare 1× SDS running buffer (NuPAGE MES SDS Running Buffer) for SDS-PAGE using deionized water (*see* **Note 30**).
3. Prepare two 4–12% (w/v) polyacrylamide gels (10-well, 1 mm-thickness, 4–12% NuPAGE Bis-Tris Mini Gels) by carefully rinsing the gel cassettes with deionized water, gently removing the comb from the gel wells, and removing the tape near the bottom of the gel cassettes (*see* **Note 31**).
4. Gently rinse the gel wells twice with SDS running buffer using a 3 mL plastic Pasteur pipette.
5. Assemble the gels in the gel-running tank (XCell *SureLock* Mini-Cell) (*see* **Note 32**), and fill the gel wells with SDS running buffer, ensuring no air bubbles remain in the wells.
6. Record the desired order of protein samples, and carefully load 20 μL protein samples into the gel wells using gel-loading pipette tips, followed by 10 μg whole cell lysates (*see* **Note 12**) and 5 μL protein standards (Precision Plus Protein All Blue Prestained Protein Standards; 1:10 dilution) (*see* **Note 33**).
7. Carefully fill the upper and lower buffer chambers of the assembled gel-running tank with 200 mL and 600 mL SDS running buffer, respectively, ensuring gel wells are completely submerged.
8. Align and fit the gel-running tank lid, and connect the electrode cords to the power supply (red, positive jack; black, negative jack).
9. Turn on the power and run gels at 200 V for 40–60 min until the dye front reaches the end of the gel.
10. Turn off the power, disconnect the electrode cords from the power supply, disassemble the gel-running tank, and carefully remove the gels (*see* **Note 34**).
11. Prepare 1× transfer buffer (NuPAGE Transfer Buffer) for electrophoretic transfer using deionized water and by adding methanol to 10% (v/v) final concentration (*see* **Note 35**).
12. Soak blotting pads (sponges), filter paper, and transfer membrane in transfer buffer until they are saturated with transfer buffer, ensuring no air bubbles remain in the blotting pads and using tweezers to handle the filter paper and transfer membrane (*see* **Note 36**).

13. Carefully separate the two plates of the gel cassettes using a gel knife, and remove gel wells and bottom ridges using the gel knife.
14. With the front side of each gel facing upwards on a clean, flat surface covered with self-sealing plastic film, place a piece of presoaked filter paper on the front side of the gel, ensuring no air bubbles remain trapped (*see* **Note 37**).
15. Turn the gel over, place a piece of presoaked transfer membrane on the rear side of the gel (which is now facing upwards), and place a piece of presoaked filter paper on top of the transfer membrane, ensuring no air bubbles remain trapped (*see* **Note 37**).
16. Place two presoaked blotting pads into the base unit (cathode) of a blotting module (XCell II Blot Module), followed by the filter paper-gel-transfer membrane-filter paper stack (front side of the gel facing down into the base unit of the blotting module; transfer membrane above the gel in the stack, furthest from the cathode).
17. Place one presoaked blotting pad onto the filter paper-gel-transfer membrane-filter paper stack.
18. Place the second filter paper-gel-transfer membrane-filter paper stack into the base unit of the blotting module.
19. Place a sufficient number of presoaked blotting pads onto the second gel stack to fill the depth of the base unit of the blotting module, and then add one more presoaked blotting pad.
20. Align and fit the lid (anode) of the blotting module, and slide it into the lower buffer chamber of a clean gel-running tank (*see* **Note 38**).
21. Fill the blotting module with transfer buffer to just over the top of the gel stack.
22. Fill the lower buffer chamber with 650 mL deionized water.
23. Align and fit the gel-running tank lid, and connect the electrode cords to the power supply (red, positive jack; black, negative jack).
24. Turn on the power and transfer gels at 30 V for 60–90 min.
25. Turn off the power, disconnect the electrode cords from the power supply, disassemble the gel-running tank, and carefully remove the transfer membranes using tweezers.
26. In a clean container, incubate transfer membranes with blocking buffer (Odyssey Blocking Buffer (TBS)) on a tilting platform at room temperature for 1 h (*see* **Notes 39** and **40**).
27. Incubate transfer membranes with appropriate primary antibodies diluted in blocking buffer containing 0.1% (w/v) Tween 20 on a tilting platform at 4 °C overnight (or at room temperature for 1–4 h) (*see* **Note 41**).

28. Wash transfer membranes four times with TBS containing 0.1% (w/v) Tween 20 (TBS-T) on a tilting platform at room temperature for 5 min.
29. Incubate transfer membranes with appropriate secondary antibodies (IRDye family secondary antibodies) diluted in blocking buffer containing 0.1% (w/v) Tween 20 on a tilting platform in the dark at room temperature for 1 h (*see* **Note 42**).
30. Rinse transfer membranes with TBS-T, and then wash transfer membranes four times with TBS-T on a tilting platform in the dark at room temperature for 5 min.
31. Rinse transfer membranes with TBS.
32. Scan the transfer membranes using an immunoblotting imager (Odyssey IR imaging system) and associated image analysis software (Image Studio) (*see* **Note 43**).
33. Quantify band densities for adhesion protein-specific and non-adhesion protein-specific antibodies using the image analysis software, incorporating background subtraction, as appropriate (*see* **Note 44**).
34. Record band reference numbers generated by the image analysis software, export band density quantification (e.g., as a comma-separated values (CSV) file), and save the raw and quantified images (e.g., as 300-dpi tagged image file format (TIFF) files and Joint Photographic Experts Group (JPEG) files, respectively) (*see* **Note 45**).

3.2 Proteomic Analysis of Isolated Adhesion Complexes

Proteomic analyses of integrin adhesion complexes isolated using similar approaches to that described above have used various instrumental systems to perform LC-MS [18, 35, 44, 45]. To achieve deep proteomic coverage and accurate quantification of adhesion complex proteins, it is recommended to use state-of-the-art, fast-scanning, accurate-mass, high-resolution LC-MS instrumentation and acquisition methods. However, lower-resolution mass spectrometers can also be used successfully to perform sensitive analyses of isolated adhesion complexes. Label-free quantification is used in this method because it is straightforward to implement and broadly applicable to most cell systems, but the protocol could be modified to use isotopic labeling approaches, such as stable isotope labeling by amino acids in cell culture [81] or tandem mass tag labeling [82], to enable multiplexed sample quantification. Cluster analysis is performed herein using the freely available software packages Cluster 3.0 [77] and Java TreeView [78], but similar analyses can be performed using Perseus [76], R [83], ELKI [84], Expander [85], Genesis [86], or other software with hierarchical clustering and data visualization functions. Interaction network analysis is performed herein using the freely

available software package Cytoscape [79], but similar analyses can be performed using Graphia Professional, Gephi [87], igraph [88], or other interaction network analysis software.

3.2.1 Proteolytic Digestion

1. Perform denaturing, reducing SDS-PAGE on all isolated samples (*see* Subheading 3.1.3, **steps 1–10**), leaving blank gel wells between each sample to prevent sample contamination from any spillages (*see* **Note 46**).
2. Carefully separate the two plates of the gel cassettes using a clean gel knife (*see* **Note 47**).
3. Incubate gels in 20 mL InstantBlue in a clean, covered container on a tilting platform at room temperature for 60 min.
4. Rinse gels five times with deionized water, and image the stained gel using an imaging system capable of fluorescence or colorimetric detection (e.g., Odyssey IR imaging system), if required.
5. Excise gel lanes using clean (new) scalpel blades and cut into 30 slices of equal size (*see* **Note 48**).
6. Chop gel slices into ~1 mm^3 pieces using clean scalpel blades, and transfer pieces from one gel slice into a corresponding well of a low-protein-binding perforated 96-well microtiter plate seated in a low-protein-binding 96-well collection plate (*see* **Note 49**). Record sample and slice well positions.
7. Incubate gel pieces in each well of the perforated 96-well microtiter plate with 50 μL 50% (v/v) acetonitrile in 25 mM ammonium bicarbonate at room temperature for 30 min, and then remove liquid by centrifugation at 400 × *g* for 2 min.
8. Incubate gel pieces again with 50 μL 50% (v/v) acetonitrile in 25 mM ammonium bicarbonate at room temperature for 30 min, and then remove liquid by centrifugation at 400 × *g* for 2 min (*see* **Note 50**).
9. Dehydrate gel pieces with 50 μL acetonitrile at room temperature for 5 min, and then remove liquid by centrifugation at 400 × *g* for 2 min.
10. Dehydrate gel pieces again with 50 μL acetonitrile at room temperature for 5 min, and then remove liquid by centrifugation at 400 × *g* for 2 min (*see* **Note 51**).
11. Dry gel pieces in the perforated 96-well microtiter plate by vacuum centrifugation at room temperature for 20–30 min.
12. Meanwhile, set the sample oven temperature to 56 °C.
13. Incubate gel pieces with 50 μL 10 mM DTT in 25 mM ammonium bicarbonate at 56 °C for 1 h, and then remove liquid by centrifugation at 400 × *g* for 2 min (*see* **Note 52**).

14. Allow gel pieces to equilibrate to room temperature. Incubate gel pieces with 50 µL 55 mM iodoacetamide in 25 mM ammonium bicarbonate in the dark at room temperature for 45 min, and then remove liquid by centrifugation at 400 × *g* for 2 min (*see* **Note 52**).
15. Incubate gel pieces with 50 µL 25 mM ammonium bicarbonate at room temperature for 10 min, and then remove liquid by centrifugation at 400 × *g* for 2 min.
16. Dehydrate gel pieces with 50 µL acetonitrile at room temperature for 5 min, and then remove liquid by centrifugation at 400 × *g* for 2 min (*see* **Note 51**).
17. Incubate gel pieces again with 50 µL 25 mM ammonium bicarbonate at room temperature for 10 min, and then remove liquid by centrifugation at 400 × *g* for 2 min.
18. Dehydrate gel pieces again with 50 µL acetonitrile at room temperature for 5 min, and then remove liquid by centrifugation at 400 × *g* for 2 min.
19. Dry gel pieces in the perforated 96-well microtiter plate by vacuum centrifugation at room temperature for 20–30 min.
20. Meanwhile, set the sample oven temperature to 37 °C.
21. Remove the 96-well collection plate, and seat the perforated 96-well microtiter plate in a clean low-protein-binding 96-well collection plate for collection of peptides.
22. Prepare 125 ng/µL trypsin stock solution (100×) from lyophilized MS-grade trypsin in trypsin resuspension buffer (*see* **Note 53**).
23. Incubate gel pieces with 50 µL 1.25 ng/µL trypsin in 25 mM ammonium bicarbonate at 4 °C for 45 min (*see* **Note 54**).
24. Incubate gel pieces at 37 °C overnight (*see* **Note 55**).
25. Collect peptides by centrifugation at 400 × *g* for 3 min.
26. Incubate gel pieces with 50 µL 0.2% (v/v) formic acid in acetonitrile at room temperature for 30 min, and then collect peptides by centrifugation at 400 × *g* for 3 min.
27. Incubate gel pieces with 50 µL 50% (v/v) acetonitrile in 0.1% (v/v) formic acid at room temperature for 30 min, and then collect peptides by centrifugation at 400 × *g* for 3 min.
28. Evaporate collected peptide solutions in the 96-well collection plate by vacuum centrifugation at room temperature for 60–120 min until ~50 µL remain.
29. Meanwhile, construct stop-and-go-extraction (STAGE) tips from 200 µL pipette tips containing single, horizontal plugs of C18 material excised from C18-bonded silica disks (3M

Empore SPE Extraction Disks) using a gauge 16 blunt-tip needle and 25 μL syringe plunger assembly (*see* **Note 56**).

30. Activate the C18 by loading 30 μL 0.1% (v/v) TFA in acetonitrile on top of the plug of C18 material, and gently discharge the liquid through the plug to waste using a syringe or centrifugal force (*see* **Note 56**).
31. Equilibrate the C18 with 30 μL 2% (v/v) acetonitrile in 0.1% (v/v) TFA, and gently discharge the liquid through the plug to waste.
32. Equilibrate the C18 again with 30 μL 2% (v/v) acetonitrile in 0.1% (v/v) TFA, and gently discharge the liquid through the plug to waste.
33. Acidify peptide samples to below pH 3 with 10% (v/v) TFA, if necessary, and load on top of each plug of C18 material.
34. Incubate peptide samples with C18 material at room temperature for 30 s, and gently discharge the liquid through the plug, collecting the flow-through (*see* **Note 57**).
35. Wash the C18 with 30 μL 2% (v/v) acetonitrile in 0.1% (v/v) TFA, and gently discharge the liquid through the plug (*see* **Note 57**).
36. Wash the C18 again with 30 μL 2% (v/v) acetonitrile in 0.1% (v/v) TFA, and gently discharge the liquid through the plug (*see* **Note 57**).
37. If necessary, store washed, C18-bound peptides at −80 °C for up to 3 months. Alternatively, proceed to LC-MS analysis (*see* Subheading 3.2.2, **step 1**).

3.2.2 LC-MS Analysis

1. Elute desalted peptides from C18 material into a low-bleed 96-well autosampler plate with 40 μL 80% (v/v) acetonitrile in 0.1% (v/v) formic acid (*see* **Notes 58** and **59**).
2. Evaporate eluted peptide solutions in the 96-well autosampler plate to ~5 μL by vacuum centrifugation at room temperature for 40–80 min (*see* **Note 60**).
3. Adjust peptide solution volumes to 15 μL using 2% (v/v) acetonitrile in 0.1% (v/v) formic acid (*see* **Note 59**).
4. Load the 96-well autosampler plate into the autosampler tray of an ultrahigh-performance LC system (Dionex UltiMate 3000 RSLCnano) coupled to an accurate-mass, high-resolution mass spectrometer (Q Exactive Plus Hybrid Quadrupole-Orbitrap).
5. Launch the latest version of the LC-MS instrument operating software (XCalibur), and generate an instrument method for the LC system and the mass spectrometer.
6. Use the Sequence Setup view in XCalibur to specify the sample names, file paths, and autosampler tray sample positions of the

peptide samples to be injected and to determine how the data are to be acquired and processed (*see* **Note 61**).

7. Use XCalibur to instruct the instrument to load 5 μL peptide sample, and separate the peptides on a C18 analytical column using a gradient of 2–40% (v/v) acetonitrile in 0.1% (v/v) formic acid over 40 min at 250 nl/min. The total run time, including a wash step followed by re-equilibration, will be ~65 min (*see* **Note 62**).
8. Apply a potential difference of +2 kV to the column electrode, and operate the mass spectrometer in positive-ion data-dependent mode.
9. Acquire mass spectra (MS1) in the range of 300–2000 m/z at a resolution of 7×10^4, with one microscan and an automatic gain control target of 3×10^6 ions (*see* **Note 63**).
10. Select the ten most intense ions for fragmentation with 30% normalized collision energy, and apply a dynamic exclusion window of 30 s (*see* **Note 63**).
11. Acquire tandem mass spectra (MS2) at a resolution of 1.75×10^4, with one microscan and an automatic gain control target of 1×10^5 ions (*see* **Note 63**).
12. After MS data acquisition, back up complete raw data files.
13. Visually inspect and quality control the chromatograms and spectra from the LC-MS runs using XCalibur.
14. Proceed to proteomic data analysis (*see* Subheading 3.2.3, **step 1**).

3.2.3 Proteomic Data Analysis

1. Launch the latest version of data analysis software for identification and quantification of proteins from raw MS data (MaxQuant and its integrated peptide search engine, Andromeda), ensuring the correct versions of any dependent software frameworks (.NET Framework) and vendor-specific libraries (MSFileReader) are installed (*see* **Note 64**).
2. Load raw data files in the Raw files tab in MaxQuant. To combine data from fractionated protein samples (i.e., gel slices from one original adhesion complex sample), label the Fraction column with non-identical integers corresponding to the gel slice and label the Experiment column with one identifier per original adhesion complex sample.
3. In the Group-specific parameters tab in MaxQuant, select a multiplicity of 1 for label-free quantification, select trypsin/P as the enzyme with a maximum of two missed cleavages, and select protein N-terminal acetylation and methionine oxidation as variable modifications (*see* **Note 65**). Select LFQ (label-free quantification).

4. In the Global parameters tab in MaxQuant, load the FASTA file corresponding to the database to be searched (*see* **Note 66**). Select carbamidomethylation of cysteine as a fixed modification, and enable matching between runs.
5. Concatenate the database to be searched with a decoy database containing reversed sequences from the original database. Accept peptide and protein false-discovery rates of 1%.
6. Select as many cores as the computer system permits in the Threads box, and click the Start button to begin the analysis.
7. If necessary, after MaxQuant data processing, visually inspect and quality control the MS data in the Viewer tab in MaxQuant.
8. If necessary, open and inspect tab-delimited text (TXT) result files saved in the MaxQuant …\combined\txt folder using a spreadsheet program (Excel) or R [83].
9. Launch the latest version of data analysis software for statistical and bioinformatic analysis of processed MS data (Perseus [76]).
10. Load the proteinGroups.txt results file (saved in the MaxQuant …\combined\txt folder) using Generic matrix upload in Perseus (*see* **Note 67**). Select the columns that contain LFQ intensities as main columns, and select other relevant columns according to their data type (numerical, categorical, etc.).
11. Manually verify successful data import by navigating the matrix (left-hand) pane and the meta-data (right-hand) pane in Perseus.
12. Using Filter rows based on categorical column in Perseus, remove proteins only identified by site, reverse database hits, and potential contaminants.
13. Using Transform in Perseus, take the binary logarithm of all main columns.
14. If necessary, using Rename columns in Perseus, shorten column names to improve readability.
15. If required, using Add annotation in Perseus, annotate proteins with selected terms (e.g., gene ontology and disease associations) from an annotation file (e.g., mainPerseusAnnot.txt) saved in the Perseus …\conf\annotations folder.
16. Using Categorical annotation rows in Perseus, group samples by like experimental conditions for subsequent statistical testing.
17. Using Filter rows based on valid values in Perseus, constrain the data matrix by removing proteins that do not meet a certain threshold of identification across the samples (e.g., for an experiment with three biological replicates, discard proteins

not identified in three replicates of at least one experimental condition) (*see* **Note 68**).

18. Using Histogram in Perseus, verify that the data distributions are approximately normal.
19. Using Subtract in Perseus, normalize transformed LFQ data by subtracting column medians from each main column.
20. Using Replace missing values from normal distribution in Perseus, impute missing values with random numbers drawn from a normal distribution (*see* **Note 69**).
21. Export the matrix of normalized, imputed protein data to a tab-delimited TXT file using Generic matrix export in Perseus.
22. Using Multi scatter plot in Perseus, verify experimental quality and assess sample correlations on a grid of scatter plots for each pairwise sample comparison.
23. In Tests in Perseus, select appropriate statistical tests (e.g., ANOVA) for analysis of the dataset, controlling for multiple testing as appropriate (*see* **Note 70**). For example, identify proteins that are significantly enriched in the adhesion receptor-specific ligand-induced adhesion complexes compared to those induced by the non-adhesion receptor-specific control ligand.
24. In Clustering/PCA in Perseus, use principal component analysis to reduce the dimensionality of the dataset and thus more readily evaluate the structure of the data and identify conditions in the experiment for which adhesion complex composition is more different than for others.
25. Using Filter rows based on categorical column in Perseus, create a matrix of only statistically significantly enriched proteins by filtering out proteins not annotated with a "+" in the "significant" column.
26. Using *Z*-score in Perseus, standardize the matrix of significantly enriched proteins by calculating *Z*-scores per row.
27. Export the matrix of standardized protein data to a tab-delimited TXT file using Generic matrix export in Perseus.
28. Launch the latest version of clustering software (Cluster 3.0).
29. In a spreadsheet program or R, reformat the matrix exported from Perseus to give a single top row of headers. Delete all columns except for the standardized protein quantification for all relevant samples and protein identifiers or names, as appropriate.
30. Import the dataset into Cluster 3.0 (*see* **Notes 71** and **72**).
31. In the Hierarchical tab in Cluster 3.0, cluster proteins ("genes") and samples ("arrays"). Select the desired distance metric (e.g., Pearson correlation), and click on the desired clustering

method (e.g., complete linkage) to initiate the clustering algorithm (*see* **Note 73**).

32. Launch the latest version of clustering data visualization software (Java TreeView).
33. Open the clustered data table (CDT) file generated by Cluster 3.0 in Java TreeView.
34. Under the Settings menu in Java TreeView, edit the coloring, contrast, and cell scaling of the automatically generated heatmap visualization, as appropriate.
35. Clusters of proteins or samples of interest can be selected by clicking on the appropriate dendrogram nodes (branch points) in Java TreeView. Hover the cursor over each cell of the heatmap to reveal its value as a tooltip.
36. Under the Export menu in Java TreeView, save the heatmap and selected clusters of interest as postscript (PS) files. An example of clustered adhesion complex assembly data is shown in Fig. 3.
37. Launch the latest version of interaction network analysis software (Cytoscape).
38. In Cytoscape, import a species-wide interaction network from an appropriate public interaction database (e.g., BioGRID) or integrated interaction dataset (e.g., PINA [89]), or use a network construction app (plugin) available from the Cytoscape App Store (e.g., GeneMANIA) (*see* **Note 74**).
39. Create a view for the interaction network in the main Network View window in Cytoscape, if necessary.
40. Map identifiers (e.g., HUGO Gene Nomenclature Committee symbols) onto protein nodes using a mapping app (e.g., BridgeDb), if necessary.
41. In a spreadsheet program or R, reformat the matrix exported from Perseus to give a single top row of headers. Columns containing protein identifiers, standardized protein data, adjusted *p*-values, and other relevant statistics should be preserved. In addition, calculate rates of change of protein abundance between adjacent time points, if required.
42. Import the dataset into Cytoscape as a data table to map the proteomic data and other relevant statistics onto the protein nodes of the interaction network. A key column of protein identifiers must match the identifiers in the interaction network.
43. In the Select tab in the Control Panel in Cytoscape, add a column filter to remove proteins not detected in the MS dataset.
44. Under the File menu in Cytoscape, extract the selected nodes, their interactions, and their attributes into a new network (from selected nodes, all edges).

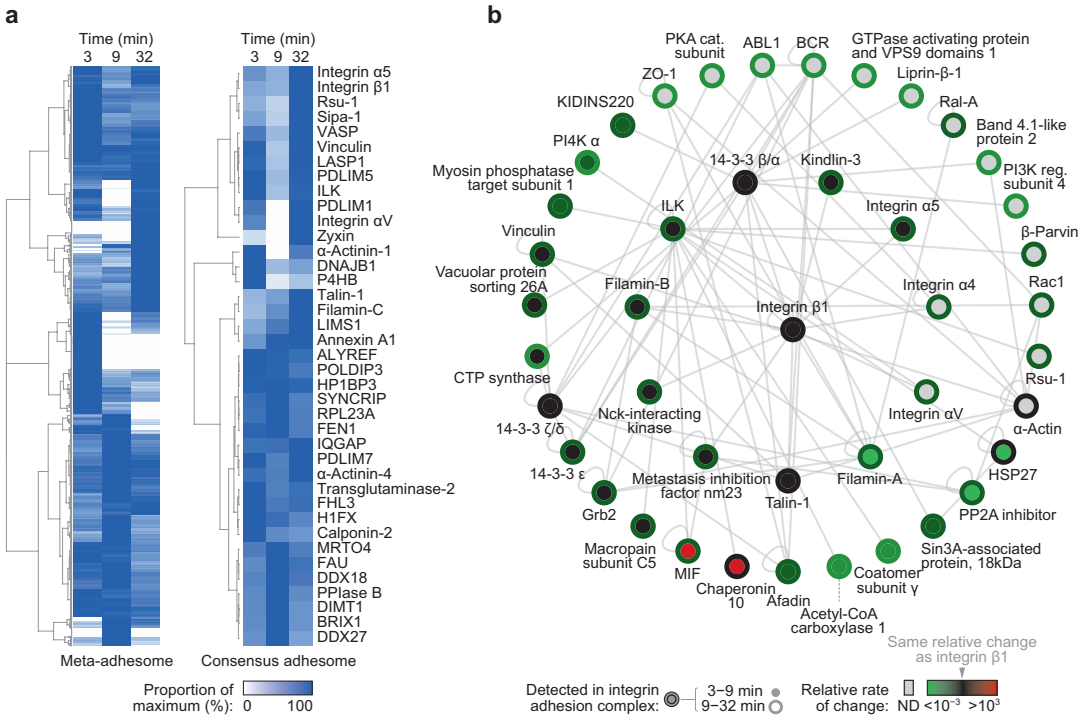


Fig. 3 Proteomic data analysis of adhesion protein recruitment dynamics. (**a**) Hierarchical cluster analysis of fibronectin-induced integrin adhesion complex formation [18]. Quantitative MS data were expressed as the proportion of maximum relative abundance in the time course for each protein and hierarchically clustered on the basis of uncentered Pearson correlation, computing distances using a complete-linkage matrix. Analysis of meta-adhesome (left) and consensus adhesome (right) proteins identified clusters of adhesion proteins with similar recruitment profiles. (**b**) Interaction network analysis of fibronectin-induced integrin adhesion complex formation. Identified proteins were mapped onto an integrated human interaction network constructed with PINA [89]. Quantitative MS data were expressed as the rate of change of abundance between adjacent time points relative to the rate of change of integrin β1 abundance. Proteins are represented as nodes (circles), with the node core colored by rate of change relative to integrin β1 from 3 to 9 min and the node border colored by rate of change relative to integrin β1 from 9 to 32 min. Reported protein interactions are represented as unweighted edges (gray connecting lines). To focus on potential integrin β1-proximal interactors, proteins more than two hops (putative direct interactions) from integrin β1 were excluded from this analysis. The network is organized with integrin β1 in the center, the inferred integrin β1 one-hop neighborhood (putative direct integrin β1 interactors) as the inner ring of nodes, and the inferred integrin β1 two-hop neighborhood as the outer ring of nodes. ND, not determined

45. In the Style tab in the Control Panel in Cytoscape, establish mapping from attribute columns (e.g., rate of change of protein abundance) to network properties (e.g., node fill color). The general properties of nodes, edges, and the entire network can also be edited.
46. Under the Layout menu in Cytoscape, arrange, scale, and rotate the new network according to requirements using manual or algorithmic options (e.g., force-directed layout) (*see* **Note 75**).

47. Select Show Graphic Details in Cytoscape to label network nodes. An example of an interaction network constructed from adhesion complex assembly data is shown in Fig. 3.
48. Select Export Network Image To File in Cytoscape to save the network view as a portable document format (PDF) file.

4 Notes

1. Sodium azide is very toxic and very dangerous for the environment; β-mercaptoethanol and glutaraldehyde are toxic, corrosive, sensitizers, and very dangerous for the environment; formic acid is toxic, corrosive, flammable, harmful, and an irritant; TFA is toxic, corrosive, harmful, an irritant, and dangerous for the environment; hydrochloric acid is toxic and corrosive; methanol is toxic and highly flammable; crystal violet is toxic, carcinogenic, an irritant, and very dangerous for the environment; manganese chloride is toxic, an irritant, and dangerous for the environment; iodoacetamide is toxic and a sensitizer; acetic acid is corrosive, flammable, and harmful; trichloroacetic acid is corrosive, is very dangerous for the environment, and has reproductive toxicity; acetonitrile is highly flammable, harmful, and an irritant; leupeptin is harmful and has reproductive toxicity; aprotinin is harmful and a sensitizer; Triton X-100 is harmful, an irritant, and dangerous for the environment; EDTA and SDS are harmful, irritants, and dangerous for the environment; sodium deoxycholate, magnesium chloride, and DTT are harmful and irritants; sodium orthovanadate and ammonium bicarbonate are harmful; trypsin is a sensitizer; bromophenol blue is an irritant and dangerous for the environment; AEBSF, DTBP, Ponceau S, and InstantBlue are irritants.
2. Test cell lines frequently to confirm the absence of mycoplasma. Consider authenticating cell lines, especially if they are listed in the database of commonly misidentified cell lines maintained by the International Cell Line Authentication Committee (http://iclac.org).
3. To yield approximately 1×10^8 K562 cells in log-phase growth per experimental condition, seed $1.0–1.5 \times 10^7$ cells in log-phase growth into fresh culture medium 72 h before adhesion complex isolation. Optimal seeding densities for other cell lines should be determined in preliminary experiments.
4. The use of 4.5 μm-diameter tosyl-activated superparamagnetic polystyrene beads (Dynabeads M-450) is described in this protocol. Beads with other chemistries (e.g., epoxy groups) and sizes (e.g., 2.8 μm diameter) are available; these may be tested

if the suggested beads do not efficiently bind the ligand or are internalized by the cells.

5. Optimal bead concentration is cell type- and ligand-dependent and should be determined in preliminary experiments but should not be less than 1×10^7 beads/mL in the experiment.
6. Other similar types of magnetic separators are available and produce comparable results with the superparamagnetic beads.
7. During ligand coating, bead concentration should be maintained at $4–8 \times 10^8$ beads/mL in a volume of at least 200 μL. As a starting point, 200 μg ligand should be used per 4×10^8 beads, but this should be optimized. For example, for four experimental conditions, resuspend 2×10^8 beads in 400 μL of PB and then add 100 μL of 1 mg/mL ligand, resulting in a final ligand concentration of 200 μg/mL and a final bead concentration of 4×10^8 beads/mL in a final volume of 500 μL.
8. Ensure ligand and buffers are free of reactive groups that will interfere with ligand coating, such as amines (e.g., in Tris).
9. In this optional step, BSA is added to aid orientation and presentation of bead-bound ligand to cells. Optimal BSA concentration should be determined in preliminary experiments. For example, to 500 μL beads, add 5 μL of 10% (w/v) BSA in PB to give ~0.1% (w/v) final BSA concentration.
10. More efficient ligand coating can be achieved at 37 °C (and incubation time can be reduced to 16 h), if the ligand is heat stable. Temperature-labile ligands can be coated at lower temperatures (e.g., 2–8 °C) but will require longer incubation times. Epoxy beads may be more appropriate for coating with temperature-labile ligands.
11. Depending on ligand stability, unused coated beads may be stored in 0.1% (w/v) BSA in PBS (optionally, containing 0.02% (w/v) sodium azide) at 4 °C for up to 1 month. Beads coated with large ECM glycoproteins such as fibronectin should be used immediately.
12. At this point, or at another suitable juncture during the protocol, a whole cell lysate should be prepared for each cell type for use as a positive control for protein expression in downstream protein analysis. In brief, wash collected cells with PBS, incubate with prechilled RIPA buffer on ice for 30 min, and separate insoluble cellular debris by centrifugation at $20{,}000 \times g$ at 4 °C for 15 min. Collect and retain the supernatant. Quantify protein concentration of clarified lysates using a total protein assay (Pierce BCA Protein Assay Kit), and store clarified lysates at −80 °C for up to 3 months.
13. Optimal bead-to-cell ratio is cell type- and ligand-dependent and should be determined in preliminary experiments. As a

starting point, a bead-to-cell ratio of 1:2 (e.g., 5×10^7 beads to 1×10^8 cells) should be used.

14. Centrifuge tubes placed in a near-horizontal orientation will ensure thorough mixing of the bead-cell suspension. Slightly raising the tops of the centrifuge tubes will help to prevent leakage. Ensure that the speed of tube rotation on the shaking platform is sufficient to thoroughly mix the bead-cell suspension but does not cause cell disruption. Some cell types may internalize the beads, in which case the rotation of the bead-cell suspension can be performed at room temperature or alternative beads can be used (*see* **Note 4**).
15. The chosen incubation times are crucial for assessing the dynamics adhesion complex assembly. At least three time points should be chosen; more time points will result in more detailed data, but the time-sensitive isolation protocol will become more challenging to coordinate and incubation times will require careful staggering. Bead-cell incubation times of 1, 7, and 30 min at 37 °C are useful starting points, but these should be optimized in preliminary experiments. Note that incubation with cross-linker adds a further 2 min to each time point. Bead-cell incubation time of at least 1 min at 37 °C is likely to be required to purify sufficient protein for useful downstream proteomic analysis, so 3 min (1 min + 2 min) is the earliest time point in adhesion complex assembly accessible using this protocol. The latest time point accessible using this protocol is determined by the point at which cells begin to internalize the beads, which can be monitored during the protocol by light microscopy, or to secrete their own ECM (2–3 h), which may interfere with specific bead-cell binding.
16. Equilibrate DTBP cross-linker to room temperature before opening the vial (~30 min) to prevent deterioration of the moisture-sensitive reagent. Use the DTBP stock solution immediately; discard if the solution becomes cloudy.
17. For a final concentration of 10 mM DTBP cross-linker, add 1.11 mL 100 mM DTBP stock solution to each 10 mL bead-cell suspension.
18. Carefully optimized cross-linking is necessary to stabilize the labile intracellular adhesion complexes while removing non-specific cellular material. The type, concentration, and time of incubation of cross-linker are critical factors for optimization. DTBP, used in this protocol, is a membrane-permeable, water-soluble, thiol-cleavable, primary amine-reactive cross-linker with an 11.9 Å spacer arm. Non-membrane-permeable cross-linker will not stabilize the intracellular adhesion complexes; non-water-soluble cross-linkers can introduce organic solvent artifacts; non-cleavable cross-linkers do not allow the release of

adhesion complex components for separation and analysis by SDS-PAGE and immunoblotting. It is essential to test the effects of cross-linking parameters on the purity of isolated adhesion complexes in preliminary experiments, using an appropriate negative control ligand or antibody to reduce or eliminate non-specifically copurifying proteins (*see* **Note 45**). Note that, for some cell types, the use of cross-linker may not be essential.

19. For a final concentration of 20 mM Tris–HCl, add 222 μL 1 M Tris–HCl, pH 8.5, to each 11.1 mL bead-cell suspension. The primary amine groups in Tris quench excess DTBP cross-linker to prevent non-specific cross-linking reactions occurring after cell lysis.
20. Bead-bound cells can be observed and quantified by light microscopy. A large proportion of beads without bound cells suggests that further optimization of bead-cell binding conditions is required. Experiments performed with non-adhesion receptor-specific control ligand or antibody can be used to ascertain the specificity of bead-cell binding. Bead-bound cell samples removed at this point in the protocol may need to be diluted tenfold to enable accurate counting.
21. As an alternative assessment of bead-cell binding, remove 10 or 20 μL bead-bound cell samples and transfer to a 96-well microtiter plate in triplicate for a crystal violet assay [90]. In brief, fix bead-bound cells with 50% (v/v) glutaraldehyde (alongside a "standard curve" of known cell number in triplicate), wash with PBS, and stain with 0.1% (w/v) crystal violet. Wash stained cells thoroughly with distilled water, destain with 10% (v/v) acetic acid, and measure absorbance at 570 nm using a 96-well-plate reader.
22. Optimal sonication times and power setting are sample- and ultrasonication device-dependent and should be determined in preliminary experiments. As a starting point, for a Bioruptor Standard, a cycle of 30 s "on" and 30 s "off" on medium power setting (200 W output power) should be repeated five times. Sonication will produce heat, especially with small sample volumes, and device water temperature should be maintained at 4 °C (e.g., by adding a small amount of crushed ice to the water bath at regular intervals). Incubate tubes on ice when not undergoing sonication, and aim for less than 30 min total sonication time for all samples.
23. As an alternative to a water bath-style ultrasonication device, a probe-based ultrasonication device, such as a Vibra-Cell, may be used. The Vibra-Cell VCX 500 operated with a 5 mm-diameter tapered microtip at 20% amplitude (500 W output power) for four 5 s pulses (with 1 min rests on ice) is a useful

starting parameter for preliminary experiments. Before each new sample, operate the Vibra-Cell for 5 s with the microtip placed in 100% (v/v) ethanol to avoid sample carry-over and then allow residual ethanol to evaporate from the microtip. When sonicating samples, immerse the probe below the surface of the sample sufficiently to prevent foaming and aerosol formation, but ensure the probe does not touch the bottom of the centrifuge tube.

24. It is recommended to avoid unnecessary exposure to the sound waves generated by ultrasonication devices. Consider using a soundproof box to surround the ultrasonication device to reduce the effects of the sound pressure generated by the device. If pregnant, exposure to the generated sound waves should be avoided.
25. If necessary for sonication, adjust the volume of bead-cell suspension per centrifuge tube according to the maximum capacity of the ultrasonication device. For example, for a 2 mL maximum capacity ultrasonication device, divide the 4 mL bead-cell suspension into two 2 mL aliquots by gently transferring to two prechilled 15 mL centrifuge tubes using a 3 mL plastic Pasteur pipette. Gently recombine samples following sonication.
26. Cell lysis can be observed and quantified by light microscopy. Only cellular debris and beads should be visible. The observation of intact cells bound to beads suggests that further optimization of cell disruption conditions is required (e.g., additional sonication cycles).
27. If protein samples are to be processed for downstream proteomic analysis by on-bead proteolytic digestion (*see* **Note 46**), omit this protein elution step and the subsequent supernatant collection step. Instead, gently wash beads twice with 1 mL prechilled PBS; transfer samples to fresh, prechilled, low-protein-binding 1.5 mL microcentrifuge tubes; and gently wash beads once with 1 mL prechilled PBS. Place each microcentrifuge tube in the magnetic separator on ice for 2 min, and gently remove and discard the supernatant.
28. Washed beads from which proteins have not been eluted (*see* **Note 27**) can also be stored at −80 °C for up to 1 month.
29. If protein samples are to be processed for downstream proteomic analysis by on-bead proteolytic digestion (*see* **Note 46**), additional sample replicates of the adhesion complex purification should be performed in parallel to enable the specificity of the adhesion complex isolation to be verified by SDS-PAGE and immunoblotting using the additional sample replicates.

30. Alternative running buffers can be used according to the required resolution of proteins on the gel. For example, MOPS SDS running buffer results in the slower running of proteins through the gel than MES SDS running buffer so is recommended for the separation of medium- to large-sized proteins.
31. Precast polyacrylamide gels with alternative compositions can be used according to the required migration profile and resolution of proteins on the gel. Alternatively, polyacrylamide gels can be cast by the user [91]. Acrylamide may be present in residual amounts on cast gels; it is toxic, carcinogenic, and mutagenic, has reproductive toxicity, and is a sensitizer, an irritant, and dangerous for the environment.
32. Ensure the gel cassettes are oriented correctly in the gel-running tank. If only one gel is used, a buffer dam must be used in place of the second gel cassette. Ensure the lever on the tension wedge is pulled forward into a locked position to seal the gel cassettes into position. It is recommended to fill the upper buffer chamber first and verify that there is no buffer leakage before filling the lower buffer chamber. Other similar gel-running tanks are available, including those that are compatible with precast polyacrylamide gels.
33. Accurate pipetting is important to load the correct volume of protein samples and to avoid spillage of samples into other gel wells. Load reducing sample buffer into any blank gel wells to achieve more uniform running of the gel.
34. Once gel cassettes have been removed from the gel-running tank, perform electrophoretic transfer immediately.
35. Alternative transfer buffers can be used according to the gel type used. Moreover, alternative electrophoretic transfer methods, such as semidry electrophoretic transfer, can be used, although some proteins in certain polyacrylamide gel compositions do not transfer as efficiently as when using wet electrophoretic transfer.
36. Polyvinylidene difluoride, which is stronger and has a higher binding capacity than nitrocellulose, can be used as an alternative transfer membrane, and transfer membranes with different pore sizes are available.
37. Thoroughly remove trapped air bubbles by gently rolling a glass pipette or other clean, smooth cylinder over the paper surface.
38. Hold the filled blotting module firmly while sliding it into the guide rails of the lower buffer chamber of the gel-running tank. Ensure the lever on the tension wedge is pulled forward into a locked position to hold the blotting module in place.

39. Prior to blocking, transfer membranes can be stained for total protein to evaluate the electrophoretic transfer and verify gel loading [92]. For example, in brief, incubate transfer membranes in 0.1% (w/v) Ponceau S in 7% (v/v) trichloroacetic acid at room temperature for 5 min, rinse with deionized water, and image the transiently stained membrane using an imaging system capable of colorimetric detection (e.g., ChemiDoc imaging systems). If all protein isolations have been performed with equal efficiency using an equal number of bead-bound cells, loading equal volumes of protein samples in the gel wells should result in very similar total staining intensities of lane positions. This also holds at different time points, as the overall staining signal from background binding of nonspecific proteins to beads will generally dominate that from the smaller number of purified adhesion receptor-specific proteins, so the total staining intensities can be used as a crude proxy for estimating relative starting protein amounts. As an alternative to staining transfer membranes, protein samples resolved by SDS-PAGE can be stained in the gels using Coomassie-based protein stains or other similar total protein stains [93]. For example, in brief, incubate gels in InstantBlue at room temperature for 60 min, rinse with deionized water, and image the stained gel using an imaging system capable of fluorescence or colorimetric detection. To estimate isolated protein yield, load a range of concentrations of whole cell lysate to create a "standard curve" of known protein amount on the same gel.
40. Clean containers to be used for immunoblotting with methanol, rinse with distilled water, rinse with isopropanol, and dry before each use. Do not use containers that have come into contact with Coomassie-based protein stains. Ensure transfer membranes are immersed in blocking buffer during blocking. Choice of blocking buffer is important for sensitive, specific immunoblotting and should be optimized in preliminary experiments. Several alternative blocking buffer formulations can be used, including casein-based blocking buffers and buffers formulated in PBS, depending on the targets of interest and the antibody detection system used. Avoid the use of detergents during blocking, as they may generate a background signal.
41. Optimal immunoblotting conditions, such as antibody dilution and incubation times, are antibody-dependent and should be determined in preliminary experiments, using manufacturers' recommendations as a starting point, where applicable. Two-color fluorescence detection, such as is enabled by the Odyssey IR imaging system, permits pairs of primary antibodies raised in different species to be multiplexed (although combining mouse and rat primary antibodies should be avoided,

where possible). Adhesion protein-specific targets for probing by immunoblotting include the adhesion receptor(s) selected by adhesion complex isolation and core adhesion proteins, such as talin, vinculin, or paxillin. Control antibodies should be used to confirm the absence of non-adhesion site organelle markers, such as syntaxin-6 (Golgi), calreticulin (endoplasmic reticulum), mitochondrial heat shock protein 70 (mitochondrion), and lamin-A/C (nucleus).

42. Use manufacturers' recommendations for secondary antibody dilutions as a starting point for optimization, where applicable. For two-color fluorescence detection, use cross-adsorbed secondary antibodies, and do not multiplex a goat-derived secondary antibody with an anti-goat secondary antibody. Moreover, alternative methods for antibody detection are available, such as chemiluminescence [94].
43. For alternative methods of antibody detection, other imaging systems (e.g., ChemiDoc imaging systems for chemiluminescence) and associated image analysis software (e.g., Image Lab Software) should be used, as appropriate. Furthermore, image analysis software such as Fiji [95] can be used to quantify detected bands from raw immunoblotting images.
44. Quantified band densities should be background corrected, which can be performed in the image analysis software, to eliminate variations in band densities from local background signal.
45. Low amounts or absence of non-adhesion site organelle markers in purified adhesion complexes (*see* **Note 41**) suggests successful reduction or elimination of potentially non-specifically copurifying non-adhesion proteins, although some background binding of non-specific proteins to beads is inevitable. Concomitant enrichment of adhesion proteins in purified adhesion complexes (*see* **Note 41**) and absence of adhesion proteins in non-adhesion receptor control isolations indicates specific isolation of ligand-induced adhesion complexes. The amount of non-specific copurification of non-adhesion proteins may change during the time course of the isolation protocol, so absence of non-adhesion site organelle markers should be confirmed by immunoblotting for each time point (loading whole cell lysate on the same gel is, therefore, an important positive immunoblotting control). The amount of non-specific recruitment of adhesion proteins may change during the time course of the isolation protocol, so non-adhesion receptor control isolations should be performed alongside ligand-induced adhesion complex purifications at each time point. Sample validation by immunoblotting is essential for protocol optimization in preliminary experiments. Poor specific adhe-

sion protein enrichment may indicate insufficient protein input, which could require protocol scale-up, further cross-linking optimization, improvement of ligand-coated bead binding (e.g., ligand choice, ligand conjugation, bead choice), different detergent conditions for cell lysis, or less stringent mechanical cell disruption. Sample validation is also important prior to large-scale LC-MS analyses of isolated samples.

46. Polyacrylamide gels of thicknesses other than 1 mm can give poorer peptide recovery after proteolytic digestion. Gels can be run until the dye front reaches the bottom of the gel for maximum protein separation, which will result in a large number of gel slices, increased sample fractionation, lower peptide complexity per run, and potentially increased depth of proteomic coverage. To reduce the number of LC-MS runs required (which is a time-consuming step and may be restricted by mass spectrometer user demands), gels can be run for a shorter period, resulting in fewer gel slices and thus quicker LC-MS data acquisition but potentially fewer peptide identifications. This trade-off is sample- and LC-MS system-dependent so should be determined in preliminary experiments. Furthermore, high-complexity samples present less of an analytical challenge when using state-of-the-art, fast-scanning, high-resolution mass spectrometers (e.g., Orbitrap Fusion Lumos Tribrid). To omit gel-based fractionation, in-gel digestion can be replaced by on-bead digestion [96, 97], which reduces sample processing (and associated opportunities for sample loss or handling errors) and substantially reduces the number of samples for LC-MS analysis.
47. Use LC-MS-grade reagents for proteomic sample processing. Perform gel handling and processing in a laminar flow hood and wear powder-free nitrile gloves to reduce sample contamination.
48. Multiple (e.g., 30) scalpel blades separated evenly and affixed firmly to a polyacrylamide gel-sized frame can be used to cut gel slices from the full length of multiple gel lanes more accurately and rapidly. Fewer gel slices can be cut if the gels are not run to completion. Only excise stained regions of the gel. Keep gels moist with deionized water during excision of gel lanes.
49. Low-profile microtiter plates perforated with small (<0.6 mm-diameter) holes at the bottom of each well will hold gel pieces and liquid, although centrifugation steps will enable liquid to be transferred through the holes to the higher volume-capacity collection plate below.
50. Ensure protein stain has been removed from gel pieces. If color remains in the gel pieces, repeat the washes with 50% (v/v) acetonitrile in 25 mM ammonium bicarbonate.

51. Empty the collection plate regularly to prevent overflow. Acetonitrile is an organic solvent and should be disposed of in accordance with institutional disposal procedures and local and regional health and safety requirements.
52. Prepare DTT and iodoacetamide solutions immediately prior to use. Keep iodoacetamide solution in the dark at 4 °C before use.
53. Alternative proteases can be used for proteolytic digestion of proteins to complement trypsin, although the efficiency of some alternative proteases may decrease for the digestion of proteins in polyacrylamide gels [98].
54. If, after incubation at 4 °C, the swelled gel pieces are not fully covered by the trypsin solution, add 25 mM ammonium bicarbonate to cover the gel pieces.
55. If necessary, perform incubation at 37 °C in a humidified chamber to prevent dehydration of gel pieces. Do not exceed 16 h for digestion with trypsin.
56. Contaminating salts and buffers that may compromise LC-MS analysis should be removed by peptide desalting using C18 solid-phase extraction. For example, STAGE tips [99] can be carefully constructed from C18-bonded silica disks (3M Empore SPE Extraction Disks) using a disk cutter (gauge 16 blunt-tip needle and 25 μL syringe plunger assembly) and 200 μL pipette tips. Alternative sample clean-up methods can be used, such as using a trap column on-line before the analytical column of the LC-MS system or using a spin column or cartridge format off-line. For more efficient solid-phase extraction of many samples, multiple STAGE tips can be carefully racked over a deep-well 96-well collection plate and processed using a centrifuge capable of holding 96-well plates.
57. Flow-through after peptide loading incubation can be reloaded on top of the plug of C18 material to increase peptide capture by the C18. Alternatively, flow-through after peptide loading incubation and peptide washes can be stored and reprocessed if the solid-phase extraction fails to capture peptides.
58. Low-bleed, "total-recovery" glass autosampler vials with screw caps with septa can be used instead of 96-well autosampler plates.
59. TFA can be used as a mobile-phase modifier for LC-MS at low pH instead of formic acid to improve chromatographic resolution and peak capacity, but TFA can suppress peptide ionization, causing a decrease in MS signal.
60. Drying time is sample-dependent; after 30–40 min vacuum centrifugation, monitor volumes of peptide solutions in collection plate wells every 10–15 min.

61. It is recommended to load and analyze a standard sample (e.g., 50 ng *Escherichia coli* whole cell lysate digested with trypsin) before running adhesion complex peptide samples to evaluate LC-MS system performance. It is also advisable to load and analyze 1 μL of one adhesion complex peptide sample before running all samples to assess sample quality (e.g., acceptable peptide elution profile, absence of contaminant polymer peaks).
62. Longer gradients can be used to increase separation of peptides, especially if gel slices contain more complex mixtures of proteins owing to shorter gel-running times, but analytical column specifications should be optimized accordingly to prevent reduction of chromatographic peak quality. Optimized 40 min gradients enable high-sensitivity analysis of peptide samples derived from moderate-complexity gel slices, and they permit the analysis of multiple samples more rapidly than do longer gradients.
63. If another LC-MS system is used, it may be necessary to adjust instrument settings accordingly.
64. MaxQuant is freely available and is compatible with several LC-MS systems. The current version runs on the Windows operating system. For large numbers of samples, MaxQuant should be run on a multicore server for more efficient analysis. Alternative MS data analysis software can be used (e.g., Proteome Discoverer) if it accepts the raw data files generated by the LC-MS system used. MS data analysis software can utilize different search algorithms (e.g., Mascot), which have distinct designs and structures, which may generate different results.
65. Additional variable modifications can be selected as required (e.g., phosphorylation of serine, threonine, or tyrosine, pyroglutamic acid conversion of N-terminal glutamine), but increasing the number of variable modifications will geometrically expand the search space of peptide masses, which is a rate-limiting step for most peptide search engines.
66. The database should be for the same species as that of the cells used for adhesion complex isolation. FASTA files not supplied with MaxQuant should be parsed using the Andromeda configuration tab in MaxQuant.
67. Any tab-delimited text file can be loaded into Perseus. The first row should contain the column names, and all other rows should contain the values.
68. The stringency of the filtering parameters will affect the number of proteins that remain in the data matrix. Valid values can be filtered across all samples, in each experimental condition (group), or in at least one experimental condition.

69. It is recommended to impute missing values, which occur frequently in LC-MS data, to facilitate downstream statistical analysis. Ideally, the distribution of random numbers used to replace missing values should represent a similar but narrower and down-shifted distribution of values compared to the distribution of valid values. Missing values can instead be imputed with a constant or not a number.
70. Parametric and non-parametric tests are available, and multiple testing correction can be applied using, for example, permutation-based false-discovery rate truncation. A threshold of artificial within-group variance ($s0$) can also be applied to give influence to the differences in protein abundances as well as the p-values for determining differentially regulated proteins.
71. If preferred, unnormalized proteomic data can be filtered, logarithm-transformed, median-centered, and normalized using Cluster 3.0.
72. Cluster 3.0 was designed for microarray data analysis, so rows refer to genes and columns refer to arrays. Upon file import, edit the job name with an indication of clustering parameters to prevent overwriting files when multiple cluster analyses are performed.
73. The choices of distance metric and linkage method affect the clustering output. Cluster 3.0 can also perform multivariate analysis using k-means clustering, self-organizing maps, and principal component analysis.
74. If a network is imported from the GeneMANIA database, the GeneMANIA app will build an interaction network from a query list of gene names corresponding to the proteins detected in the MS analysis. The construction parameters can be set to create a network consisting only of the query proteins, or the network can be expanded to integrate other predicted connected proteins. In addition to physical protein interactions, the GeneMANIA database includes co-expression, genetic interactions, and other network relationships, which can be selected as required.
75. Some layout algorithms are based on geometric arrangements, such as circles or grids, whereas others, such as force-directed layouts, use graph-theoretic properties to determine the relative positioning of nodes.

Acknowledgments

J.A. Askari, J.D. Humphries, M.J. Humphries, and other members of the Humphries Laboratory (University of Manchester) are gratefully acknowledged for the development and optimization of

the integrin adhesion complex purification protocol described herein, which was funded by the Wellcome Trust. A.B. is funded by Cancer Research UK (grant C157/A15703 to M.C. Frame, University of Edinburgh).

References

1. Hynes RO (2009) The extracellular matrix: not just pretty fibrils. Science 326(5957):1216–1219. https://doi.org/10.1126/science.1176009
2. Byron A, Morgan MR, Humphries MJ (2010) Adhesion signalling complexes. Curr Biol 20(24):R1063–R1067. https://doi.org/10.1016/j.cub.2010.10.059
3. Winograd-Katz SE, Fässler R, Geiger B, Legate KR (2014) The integrin adhesome: from genes and proteins to human disease. Nat Rev Mol Cell Biol 15(4):273–288. https://doi.org/10.1038/nrm3769
4. Larjava H, Koivisto L, Heino J, Häkkinen L (2014) Integrins in periodontal disease. Exp Cell Res 325(2):104–110. https://doi.org/10.1016/j.yexcr.2014.03.010
5. Lennon R, Randles MJ, Humphries MJ (2014) The importance of podocyte adhesion for a healthy glomerulus. Front Endocrinol 5:160. https://doi.org/10.3389/fendo.2014.00160
6. Wright DB, Meurs H, Dekkers BG (2014) Integrins: therapeutic targets in airway hyperresponsiveness and remodelling? Trends Pharmacol Sci 35(11):567–574. https://doi.org/10.1016/j.tips.2014.09.006
7. Allen S, Moran N (2015) Cell adhesion molecules: therapeutic targets for inhibition of inflammatory states. Semin Thromb Hemost 41(6):563–571. https://doi.org/10.1055/s-0035-1556588
8. Bravatà I, Allocca M, Fiorino G, Danese S (2015) Integrins and adhesion molecules as targets to treat inflammatory bowel disease. Curr Opin Pharmacol 25:67–71. https://doi.org/10.1016/j.coph.2015.11.007
9. Coelho NM, McCulloch CA (2016) Contribution of collagen adhesion receptors to tissue fibrosis. Cell Tissue Res 365(3):521–538. https://doi.org/10.1007/s00441-016-2440-8
10. Hamidi H, Pietilä M, Ivaska J (2016) The complexity of integrins in cancer and new scopes for therapeutic targeting. Br J Cancer 115(9):1017–1023. https://doi.org/10.1038/bjc.2016.312
11. Filla MS, Faralli JA, Peotter JL, Peters DM (2017) The role of integrins in glaucoma. Exp Eye Res 158:124–136. https://doi.org/10.1016/j.exer.2016.05.011
12. Finney AC, Stokes KY, Pattillo CB, Orr AW (2017) Integrin signaling in atherosclerosis. Cell Mol Life Sci 74(12):2263–2282. https://doi.org/10.1007/s00018-017-2490-4
13. Humphries JD, Byron A, Humphries MJ (2006) Integrin ligands at a glance. J Cell Sci 119(Pt 19):3901–3903. https://doi.org/10.1242/jcs.03098
14. Zamir E, Geiger B (2001) Molecular complexity and dynamics of cell-matrix adhesions. J Cell Sci 114(Pt 20):3583–3590
15. Zaidel-Bar R, Ballestrem C, Kam Z, Geiger B (2003) Early molecular events in the assembly of matrix adhesions at the leading edge of migrating cells. J Cell Sci 116(Pt 22):4605–4613. https://doi.org/10.1242/jcs.00792
16. Bachir AI, Zareno J, Moissoglu K, Plow EF, Gratton E, Horwitz AR (2014) Integrin-associated complexes form hierarchically with variable stoichiometry in nascent adhesions. Curr Biol 24(16):1845–1853. https://doi.org/10.1016/j.cub.2014.07.011
17. Hoffmann JE, Fermin Y, Stricker RL, Ickstadt K, Zamir E (2014) Symmetric exchange of multiprotein building blocks between stationary focal adhesions and the cytosol. elife 3:e02257. https://doi.org/10.7554/eLife.02257
18. Horton ER, Byron A, Askari JA, Ng DHJ, Millon-Frémillon A, Robertson J, Koper EJ, Paul NR, Warwood S, Knight D, Humphries JD, Humphries MJ (2015) Definition of a consensus integrin adhesome and its dynamics during adhesion complex assembly and disassembly. Nat Cell Biol 17(12):1577–1587. https://doi.org/10.1038/ncb3257
19. Carisey A, Tsang R, Greiner AM, Nijenhuis N, Heath N, Nazgiewicz A, Kemkemer R, Derby B, Spatz J, Ballestrem C (2013) Vinculin regulates the recruitment and release of core focal adhesion proteins in a force-dependent manner. Curr Biol 23(4):271–281. https://doi.org/10.1016/j.cub.2013.01.009
20. Iskratsch T, Yu CH, Mathur A, Liu S, Stévenin V, Dwyer J, Hone J, Ehler E, Sheetz M (2013) FHOD1 is needed for directed forces and adhesion maturation during cell spreading and migration. Dev Cell 27(5):545–559. https://doi.org/10.1016/j.devcel.2013.11.003

21. Ciobanasu C, Faivre B, Le Clainche C (2014) Actomyosin-dependent formation of the mechanosensitive talin-vinculin complex reinforces actin anchoring. Nat Commun 5:3095. https://doi.org/10.1038/ncomms4095
22. Yao M, Goult BT, Chen H, Cong P, Sheetz MP, Yan J (2014) Mechanical activation of vinculin binding to talin locks talin in an unfolded conformation. Sci Rep 4:4610. https://doi.org/10.1038/srep04610
23. Hernández-Varas P, Berge U, Lock JG, Strömblad S (2015) A plastic relationship between vinculin-mediated tension and adhesion complex area defines adhesion size and lifetime. Nat Commun 6:7524. https://doi.org/10.1038/ncomms8524
24. Austen K, Ringer P, Mehlich A, Chrostek-Grashoff A, Kluger C, Klingner C, Sabass B, Zent R, Rief M, Grashoff C (2015) Extracellular rigidity sensing by talin isoform-specific mechanical linkages. Nat Cell Biol 17(12):1597–1606. https://doi.org/10.1038/ncb3268
25. Roper JA, Williamson RC, Bass MD (2012) Syndecan and integrin interactomes: large complexes in small spaces. Curr Opin Struct Biol 22(5):583–590. https://doi.org/10.1016/j.sbi.2012.07.003
26. Bass MD, Williamson RC, Nunan RD, Humphries JD, Byron A, Morgan MR, Martin P, Humphries MJ (2011) A syndecan-4 hair trigger initiates wound healing through caveolin- and RhoG-regulated integrin endocytosis. Dev Cell 21(4):681–693. https://doi.org/10.1016/j.devcel.2011.08.007
27. Morgan MR, Hamidi H, Bass MD, Warwood S, Ballestrem C, Humphries MJ (2013) Syndecan-4 phosphorylation is a control point for integrin recycling. Dev Cell 24(5):472–485. https://doi.org/10.1016/j.devcel.2013.01.027
28. Guo Z, Neilson LJ, Zhong H, Murray PS, Zanivan S, Zaidel-Bar R (2014) E-cadherin interactome complexity and robustness resolved by quantitative proteomics. Sci Signal 7(354):rs7. https://doi.org/10.1126/scisignal.2005473
29. Miyake Y, Inoue N, Nishimura K, Kinoshita N, Hosoya H, Yonemura S (2006) Actomyosin tension is required for correct recruitment of adherens junction components and zonula occludens formation. Exp Cell Res 312(9):1637–1650. https://doi.org/10.1016/j.yexcr.2006.01.031
30. Liu Z, Tan JL, Cohen DM, Yang MT, Sniadecki NJ, Ruiz SA, Nelson CM, Chen CS (2010) Mechanical tugging force regulates the size of cell-cell junctions. Proc Natl Acad Sci U S A 107(22):9944–9949. https://doi.org/10.1073/pnas.0914547107
31. Zaidel-Bar R, Itzkovitz S, Ma'ayan A, Iyengar R, Geiger B (2007) Functional atlas of the integrin adhesome. Nat Cell Biol 9(8):858–867. https://doi.org/10.1038/ncb0807-858
32. Byron A, Humphries JD, Bass MD, Knight D, Humphries MJ (2011) Proteomic analysis of integrin adhesion complexes. Sci Signal 4(167):pt2. https://doi.org/10.1126/scisignal.2001827
33. Kuo JC, Han X, Yates JR III, Waterman CM (2012) Isolation of focal adhesion proteins for biochemical and proteomic analysis. Methods Mol Biol 757:297–323. https://doi.org/10.1007/978-1-61779-166-6_19
34. Jones MC, Humphries JD, Byron A, Millon-Frémillon A, Robertson J, Paul NR, Ng DH, Askari JA, Humphries MJ (2015) Isolation of integrin-based adhesion complexes. Curr Protoc Cell Biol 66:9.8.1–9.8.15. https://doi.org/10.1002/0471143030.cb0908s66
35. Humphries JD, Byron A, Bass MD, Craig SE, Pinney JW, Knight D, Humphries MJ (2009) Proteomic analysis of integrin-associated complexes identifies RCC2 as a dual regulator of Rac1 and Arf6. Sci Signal 2(87):ra51. https://doi.org/10.1126/scisignal.2000396
36. Schiller HB, Friedel CC, Boulegue C, Fässler R (2011) Quantitative proteomics of the integrin adhesome show a myosin II-dependent recruitment of LIM domain proteins. EMBO Rep 12(3):259–266. https://doi.org/10.1038/embor.2011.5
37. Kuo JC, Han X, Hsiao CT, Yates JR III, Waterman CM (2011) Analysis of the myosin-II-responsive focal adhesion proteome reveals a role for β-Pix in negative regulation of focal adhesion maturation. Nat Cell Biol 13(4):383–393. https://doi.org/10.1038/ncb2216
38. Byron A (2017) Clustering and network analysis of reverse phase protein array data. Methods Mol Biol 1606:171–191. https://doi.org/10.1007/978-1-4939-6990-6_12
39. Carugo O (2010) Clustering criteria and algorithms. Methods Mol Biol 609:175–196. https://doi.org/10.1007/978-1-60327-241-4_11
40. Nugent R, Meila M (2010) An overview of clustering applied to molecular biology. Methods Mol Biol 620:369–404. https://doi.org/10.1007/978-1-60761-580-4_12
41. Chen B, Fan W, Liu J, Wu FX (2014) Identifying protein complexes and functional modules—from static PPI networks to dynamic PPI networks. Brief Bioinform 15(2):177–194. https://doi.org/10.1093/bib/bbt039
42. Srihari S, Yong CH, Patil A, Wong L (2015) Methods for protein complex prediction and their contributions towards understanding the organ-

isation, function and dynamics of complexes. FEBS Lett 589(19 Pt A):2590–2602. https://doi.org/10.1016/j.febslet.2015.04.026

43. Byron A (2008) Proteomic analyses of integrin-based adhesion complexes. PhD Thesis. University of Manchester, Manchester, United Kingdom
44. Byron A, Humphries JD, Craig SE, Knight D, Humphries MJ (2012) Proteomic analysis of α4β1 integrin adhesion complexes reveals α-subunit-dependent protein recruitment. Proteomics 12(13):2107–2114. https://doi.org/10.1002/pmic.201100487
45. Byron A, Askari JA, Humphries JD, Jacquemet G, Koper EJ, Warwood S, Choi CK, Stroud MJ, Chen CS, Knight D, Humphries MJ (2015) A proteomic approach reveals integrin activation state-dependent control of microtubule cortical targeting. Nat Commun 6:6135. https://doi.org/10.1038/ncomms7135
46. Millon-Frémillon A, Aureille J, Guilluy C (2017) Analyzing cell surface adhesion remodeling in response to mechanical tension using magnetic beads. J Vis Exp 121:e55330. https://doi.org/10.3791/55330
47. Arike L, Peil L (2014) Spectral counting label-free proteomics. Methods Mol Biol 1156:213–222. https://doi.org/10.1007/978-1-4939-0685-7_14
48. Moulder R, Goo YA, Goodlett DR (2016) Label-free quantitation for clinical proteomics. Methods Mol Biol 1410:65–76. https://doi.org/10.1007/978-1-4939-3524-6_4
49. Souza GH, Guest PC, Martins-de-Souza D (2017) LC-MSE, multiplex MS/MS, ion mobility, and label-free quantitation in clinical proteomics. Methods Mol Biol 1546:57–73. https://doi.org/10.1007/978-1-4939-6730-8_4
50. Kani K (2017) Quantitative proteomics using SILAC. Methods Mol Biol 1550:171–184. https://doi.org/10.1007/978-1-4939-6747-6_13
51. Gritsenko MA, Xu Z, Liu T, Smith RD (2016) Large-scale and deep quantitative proteome profiling using isobaric labeling coupled with two-dimensional LC-MS/MS. Methods Mol Biol 1410:237–247. https://doi.org/10.1007/978-1-4939-3524-6_14
52. Núñez EV, Domont GB, Nogueira FC (2017) iTRAQ-based shotgun proteomics approach for relative protein quantification. Methods Mol Biol 1546:267–274. https://doi.org/10.1007/978-1-4939-6730-8_23
53. Zhang L, Elias JE (2017) Relative protein quantification using tandem mass tag mass spectrometry. Methods Mol Biol 1550:185–198. https://doi.org/10.1007/978-1-4939-6747-6_14
54. Holewinski RJ, Parker SJ, Matlock AD, Venkatraman V, Van Eyk JE (2016) Methods for SWATH™: data independent acquisition on TripleTOF mass spectrometers. Methods Mol Biol 1410:265–279. https://doi.org/10.1007/978-1-4939-3524-6_16
55. Röst HL, Aebersold R, Schubert OT (2017) Automated SWATH data analysis using targeted extraction of ion chromatograms. Methods Mol Biol 1550:289–307. https://doi.org/10.1007/978-1-4939-6747-6_20
56. Schilling B, Gibson BW, Hunter CL (2017) Generation of high-quality SWATH® acquisition data for label-free quantitative proteomics studies using TripleTOF® mass spectrometers. Methods Mol Biol 1550:223–233. https://doi.org/10.1007/978-1-4939-6747-6_16
57. Zaidel-Bar R, Milo R, Kam Z, Geiger B (2007) A paxillin tyrosine phosphorylation switch regulates the assembly and form of cell-matrix adhesions. J Cell Sci 120(Pt 1):137–148. https://doi.org/10.1242/jcs.03314
58. Bae YH, Mui KL, Hsu BY, Liu SL, Cretu A, Razinia Z, Xu T, Puré E, Assoian RK (2014) A FAK-Cas-Rac-lamellipodin signaling module transduces extracellular matrix stiffness into mechanosensitive cell cycling. Sci Signal 7(330):ra57. https://doi.org/10.1126/scisignal.2004838
59. Qu H, Tu Y, Guan JL, Xiao G, Wu C (2014) Kindlin-2 tyrosine phosphorylation and interaction with Src serve as a regulatable switch in the integrin outside-in signaling circuit. J Biol Chem 289(45):31001–31013. https://doi.org/10.1074/jbc.M114.580811
60. Pasapera AM, Plotnikov SV, Fischer RS, Case LB, Egelhoff TT, Waterman CM (2015) Rac1-dependent phosphorylation and focal adhesion recruitment of myosin IIA regulates migration and mechanosensing. Curr Biol 25(2):175–186. https://doi.org/10.1016/j.cub.2014.11.043
61. Wu JC, Chen YC, Kuo CT, Wenshin Yu H, Chen YQ, Chiou A, Kuo JC (2015) Focal adhesion kinase-dependent focal adhesion recruitment of SH2 domains directs SRC into focal adhesions to regulate cell adhesion and migration. Sci Rep 5:18476. https://doi.org/10.1038/srep18476
62. Lopez-Sanchez I, Kalogriopoulos N, Lo IC, Kabir F, Midde KK, Wang H, Ghosh P (2015) Focal adhesions are foci for tyrosine-based signal transduction via GIV/Girdin and G proteins. Mol Biol Cell 26(24):4313–4324. https://doi.org/10.1091/mbc.E15-07-0496
63. Horton ER, Humphries JD, Stutchbury B, Jacquemet G, Ballestrem C, Barry ST,

Humphries MJ (2016) Modulation of FAK and Src adhesion signaling occurs independently of adhesion complex composition. J Cell Biol 212(3):349–364. https://doi.org/10.1083/jcb.201508080

64. Swaminathan V, Fischer RS, Waterman CM (2016) The FAK-Arp2/3 interaction promotes leading edge advance and haptosensing by coupling nascent adhesions to lamellipodia actin. Mol Biol Cell 27(7):1085–1100. https://doi.org/10.1091/mbc.E15-08-0590
65. Stutchbury B, Atherton P, Tsang R, Wang DY, Ballestrem C (2017) Distinct focal adhesion protein modules control different aspects of mechanotransduction. J Cell Sci 130(9):1612–1624. https://doi.org/10.1242/jcs.195362
66. Kirchner J, Kam Z, Tzur G, Bershadsky AD, Geiger B (2003) Live-cell monitoring of tyrosine phosphorylation in focal adhesions following microtubule disruption. J Cell Sci 116(Pt 6):975–986. https://doi.org/10.1242/jcs.00284
67. Iyer VV, Ballestrem C, Kirchner J, Geiger B, Schaller MD (2005) Measurement of protein tyrosine phosphorylation in cell adhesion. Methods Mol Biol 294:289–302
68. Ballestrem C, Erez N, Kirchner J, Kam Z, Bershadsky A, Geiger B (2006) Molecular mapping of tyrosine-phosphorylated proteins in focal adhesions using fluorescence resonance energy transfer. J Cell Sci 119(Pt 5):866–875. https://doi.org/10.1242/jcs.02794
69. Chen Y, Lu B, Yang Q, Fearns C, Yates JR III, Lee JD (2009) Combined integrin phosphoproteomic analyses and small interfering RNA-based functional screening identify key regulators for cancer cell adhesion and migration. Cancer Res 69(8):3713–3720. https://doi.org/10.1158/0008-5472.CAN-08-2515
70. Schiller HB, Hermann MR, Polleux J, Vignaud T, Zanivan S, Friedel CC, Sun Z, Raducanu A, Gottschalk KE, Théry M, Mann M, Fässler R (2013) β1- and αv-class integrins cooperate to regulate myosin II during rigidity sensing of fibronectin-based microenvironments. Nat Cell Biol 15(6):625–636. https://doi.org/10.1038/ncb2747
71. Robertson J, Jacquemet G, Byron A, Jones MC, Warwood S, Selley JN, Knight D, Humphries JD, Humphries MJ (2015) Defining the phospho-adhesome through the phosphoproteomic analysis of integrin signalling. Nat Commun 6:6265. https://doi.org/10.1038/ncomms7265
72. Steen H, Jebanathirajah JA, Rush J, Morrice N, Kirschner MW (2006) Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements. Mol Cell Proteomics 5(1):172–181. https://doi.org/10.1074/mcp.M500135-MCP200
73. Robertson J, Humphries JD, Paul NR, Warwood S, Knight D, Byron A, Humphries MJ (2017) Characterization of the phospho-adhesome by mass spectrometry-based proteomics. Methods Mol Biol 1636:235–251. https://doi.org/10.1007/978-1-4939-7154-1_15
74. Tyanova S, Temu T, Cox J (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc 11(12):2301–2319. https://doi.org/10.1038/nprot.2016.136
75. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10(4):1794–1805. https://doi.org/10.1021/pr101065j
76. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods 13(9):731–740. https://doi.org/10.1038/nmeth.3901
77. de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20(9):1453–1454. https://doi.org/10.1093/bioinformatics/bth078
78. Saldanha AJ (2004) Java Treeview—extensible visualization of microarray data. Bioinformatics 20(17):3246–3248. https://doi.org/10.1093/bioinformatics/bth349
79. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504. https://doi.org/10.1101/gr.1239303
80. Byron A, Humphries JD, Askari JA, Craig SE, Mould AP, Humphries MJ (2009) Anti-integrin monoclonal antibodies. J Cell Sci 122(Pt 22):4009–4011. https://doi.org/10.1242/jcs.056770
81. Lau HT, Suh HW, Golkowski M, Ong SE (2014) Comparing SILAC- and stable isotope dimethyl-labeling approaches for quantitative proteomics. J Proteome Res 13(9):4164–4174. https://doi.org/10.1021/pr500630a
82. McAlister GC, Nusinow DP, Jedrychowski MP, Wühr M, Huttlin EL, Erickson BK, Rad R, Haas W, Gygi SP (2014) MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal Chem 86(14):7150–7158. https://doi.org/10.1021/ac502040v

83. R Development Core Team (2013) R: a language and environment for statistical computing. The R Foundation for Statistical Computing, Vienna, Austria
84. Achtert E, Kriegel H-P, Zimek A (2008) ELKI: a software system for evaluation of subspace clustering algorithms. Lect Notes Comput Sci 5069:580–585. https://doi.org/10.1007/978-3-540-69497-7_41
85. Sharan R, Maron-Katz A, Shamir R (2003) CLICK and EXPANDER: a system for clustering and visualizing gene expression data. Bioinformatics 19(14):1787–1799. https://doi.org/10.1093/bioinformatics/btg232
86. Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: cluster analysis of microarray data. Bioinformatics 18(1):207–208. https://doi.org/10.1093/bioinformatics/18.1.207
87. Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. Int AAAI Conf Web Soc Media. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154
88. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems: 1695. http://igraph.org
89. Wu J, Vallenius T, Ovaska K, Westermarck J, Makela TP, Hautaniemi S (2009) Integrated network analysis platform for protein-protein interactions. Nat Methods 6:75–77. https://doi.org/10.1038/nmeth.1282
90. Humphries MJ (2001) Cell-substrate adhesion assays. Curr Protoc Cell Biol Chapter 9:Unit 9.1. doi:https://doi.org/10.1002/0471143030.cb0901s00
91. Brunelle JL, Green R (2014) One-dimensional SDS-polyacrylamide gel electrophoresis (1D SDS-PAGE). Methods Enzymol 541:151–159. https://doi.org/10.1016/B978-0-12-420119-4.00012-4
92. Goldman A, Harper S, Speicher DW (2016) Detection of proteins on blot membranes. Curr Protoc Protein Sci 86:10.8.1–10.8.11. https://doi.org/10.1002/cpps.15
93. Brunelle JL, Green R (2014) Coomassie blue staining. Methods Enzymol 541:161–167. https://doi.org/10.1016/B978-0-12-420119-4.00013-6
94. Janes KA (2015) An analysis of critical factors for quantitative immunoblotting. Sci Signal 8(371):rs2. https://doi.org/10.1126/scisignal.2005966
95. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez JY, White DJ, Hartenstein V, Eliceiri K, Tomancak P, Cardona A (2012) Fiji: an open-source platform for biological-image analysis. Nat Methods 9(7):676–682. https://doi.org/10.1038/nmeth.2019
96. Hubner NC, Bird AW, Cox J, Splettstoesser B, Bandilla P, Poser I, Hyman A, Mann M (2010) Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. J Cell Biol 189(4):739–754. https://doi.org/10.1083/jcb.200911091
97. Turriziani B, Garcia-Munoz A, Pilkington R, Raso C, Kolch W, von Kriegsheim A (2014) On-beads digestion in conjunction with data-dependent mass spectrometry: a shortcut to quantitative and dynamic interaction proteomics. Biology 3(2):320–332. https://doi.org/10.3390/biology3020320
98. Giansanti P, Tsiatsiani L, Low TY, Heck AJ (2016) Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. Nat Protoc 11(5):993–1006. https://doi.org/10.1038/nprot.2016.057
99. Rappsilber J, Mann M, Ishihama Y (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat Protoc 2(8):1896–1906. https://doi.org/10.1038/nprot.2007.261

Chapter 14

Dual-Color and 3D Super-Resolution Microscopy of Multi-protein Assemblies

Philipp Hoess, Markus Mund, Manuel Reitberger, and Jonas Ries

Abstract

Breaking the resolution limit of conventional microscopy by super-resolution microscopy (SRM) led to many new biological insights into protein assemblies at the nanoscale. Here we provide detailed protocols for single-molecule localization microscopy (SMLM) to image the structure of a protein complex. As examples, we show how to acquire single- and dual-color super-resolution images of the nuclear pore complex (NPC) and dual-color 3D data on actin and paxillin in focal adhesions.

Key words Super-resolution microscopy, Single-molecule localization microscopy, PALM, STORM, Nuclear pore complex, Focal adhesions, Photoswitchable fluorescent protein

1 Introduction

In recent years, the resolution limit of fluorescence microscopy has been overcome by different approaches, all of them summed up under the term super-resolution microscopy (SRM) [1]. These methods comprise structured illumination microscopy (SIM) [2], stimulated emission depletion (STED) microscopy [3–5], and single-molecule localization microscopy (SMLM) approaches, such as (fluorescence) photo-activated localization microscopy (fPALM) [6, 7], stochastic optical reconstruction microscopy (STORM) [8], or points accumulation for imaging in nanoscale topography (PAINT) [9]. SMLM relies on the precise determination of positions of single, sparse emitters. To image individual fluorophores in densely labeled samples, their emission is separated by stochastically switching them between a dark and a fluorescent state. Hence, single fluorophores are well separated in space on individual frames and can be fit by a point spread function (PSF) model, usually a Gaussian in two-dimensional (2D) imaging. The fit returns the positions of the fluorophores with an uncertainty (localization precision) that is inversely correlated with the square root of the detected photons [10]. After post-processing, a final

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_14,

super-resolution image is reconstructed from all individual localizations. SMLM requires fluorophores that can be switched between a dark and a bright state. Mostly, photoswitchable fluorescent proteins or organic dyes in a thiol-containing buffer are used [8]. Often, UV light is used to recover fluorophores from their dark states in order to maintain a constant number of localizations over time.

To visualize structures of interest, it is necessary to label them as densely as possible. A specific spatial resolution on continuous structures can only be reached if fluorophores are at least spaced apart by half the desired resolution (Nyquist criterion) [11, 12]. For imaging of protein complexes, only a high absolute labeling efficiency allows extraction of structural information. When using organic dyes, they need to be targeted to the proteins of interest. This can be achieved with classical immunolabeling using primary and secondary antibodies (some caveats apply, *see* **Notes 2** and **10**). Besides that, the protein of interest can be genetically fused to self-labeling proteins such as the SNAP-tag [13, 14], CLIP-tag [15], or HaloTag [16]. These tags are small engineered enzymes, which covalently bind a chemical moiety that itself is coupled to a fluorescent dye [17]. Furthermore, GFP-tagged proteins can be labeled using GFP nanobodies that are coupled to a fluorophore [18]. Additionally, any small molecule that binds specifically to a structure of interest and can be conjugated with a fluorophore can be used to visualize this structure for SMLM. Different fluorescent proteins are available that can be genetically fused to the protein of interest. These fluorescent proteins have to be photoactivatable (from dark to fluorescent) or photoconvertible (between different colors, e.g., from green to red) to be suitable for SMLM. Fluorescent proteins with these characteristics include PA-GFP [19], mEOS [20, 21], PA-mKate [22], and mMaple [23]. For dual-color SMLM imaging, it is necessary that both fluorophores blink under the same conditions. This requirement can be overcome by different approaches:

1. The same reporter fluorophore is combined with different activator dyes, which are in close proximity to the reporter dye [24]. Although the emission wavelength is the same for all structures, the color can be assigned according to the wavelength used for activation.
2. In ratiometric imaging, two fluorophores that have overlapping emission spectra are excited by light of the same wavelength. Their emission light is split into two channels by a dichroic mirror. The two channels contain different fractions of the fluorescence of the two dyes. The color of the individual blinks can be retrieved by comparing their brightness in the two different channels [25].

3. Spectrally distinct synthetic dyes and photoactivatable proteins can be imaged simultaneously. The fluorophores are excited with the respective wavelengths and the emitted light is separated into two channels by a dichroic mirror.
4. If the two fluorophores require different conditions for blinking, the different colors can be acquired sequentially. For sequential imaging of fluorescent proteins, it is necessary that they are activated by light of different wavelengths or that they can be switched reversibly [26]. Besides that, it is possible to image a first color, quench the fluorophore by addition of a reducing agent or by a combination of bleaching and quenching and then post-stain a different structure [27, 28]. Moreover, in DNA-PAINT, different imager strands can be applied to the samples sequentially [29].

Three-dimensional (3D) information can be obtained by modifying the shape of the PSF depending on the axial position of the fluorophore. The simplest and most widely applied method is introduction of astigmatism to the PSF by a cylindrical lens in the detection beam path [30]. Other approaches comprise a double-helical PSF [31], a saddle-point PSF [32], bi-plane [33] and multifocal imaging [34], or 4Pi microscopy [35].

With its high spatial resolution and molecular specificity in fixed samples, SMLM is very well suited to elucidate the structure and composition of protein complexes in situ. Among others, SMLM was used to investigate the structures of focal adhesions [36], receptor organization in synapses [37], actin structures in axons [38], protein organization in the kidney glomerular basement membrane [39], and the orientation of a subcomplex within the nuclear pore complex (NPC) [40].

Here, we provide detailed protocols for different labeling and imaging strategies to study the NPC and focal adhesions in mammalian cells. We show labeling by a fusion protein of an NPC subunit with a photoconvertible protein (Fig. 1a), with a lectin that binds to the central part of the NPC (Fig. 1b), and we combine those strategies to image the NPC in dual color (Fig. 2a–c). Moreover, we demonstrate immunolabeling of paxillin together with labeling of actin by phalloidin to visualize focal adhesions in dual color and three dimensions (Fig. 2d, e). These protocols can be easily applied also to image a variety of other cellular structures.

2 Materials

All solutions are prepared with Milli-Q water and analytical grade reagents.

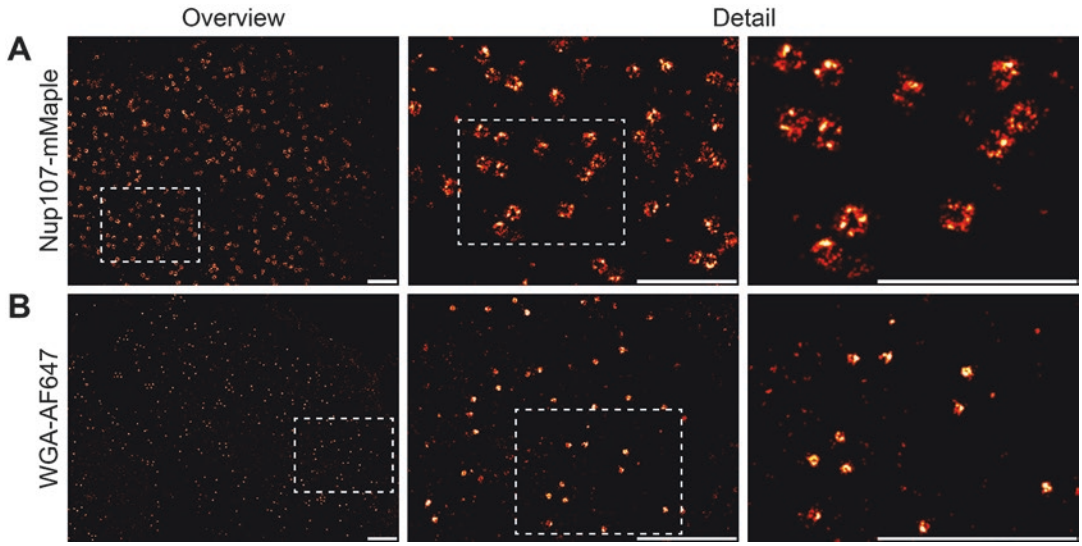


Fig. 1 Super-resolution images of the nuclear pore complex in single color. (**a**) Nup107 tagged with the photoconvertible fluorescent protein mMaple was imaged in HEK293 cells by single-molecule localization microscopy. (**b**) The center of the nuclear pore complex in HEK293 cells was stained with WGA-AF647. This lectin binds specifically to the glycosylated and disordered protein moieties that are present within the pore. Scale bars: 1 μm. For the colored version of this figure, please refer to the online version of the book chapter

2.1 Preparation of Coverslips and Bead Sample

1. Cleaning solution: 50/50 mixture of methanol and hydrochloric acid.
2. 0.1 μm TetraSpeck beads.
3. 1 M $MgCl_2$.

2.2 Sample Preparation for Staining of the Nuclear Pore Complex

1. Fixing solution: 3% [w/v] formaldehyde (FA) in PBS, pH 7.4. Prepare freshly from 2× PBS, water, and 16% [w/v] FA (*see* **Note 1**).
2. Quenching solution: 100 mM ammonium chloride in PBS. Prepare from a 1 M ammonium chloride stock solution, 2× PBS, and water. Autoclave for prolonged storage.
3. Permeabilization solution: 0.4% [v/v] Triton X-100 in PBS. Prepare from 2× PBS, water, and 100% Triton X-100.
4. Blocking solution: 2% [w/v] BSA in PBS. Prepare by dissolving BSA in PBS. Filter sterile and store at 4 °C.
5. Wheat germ agglutinin (WGA) staining solution: 0.2 μg/ml WGA-AF647 (Invitrogen, W32466) in 1% [w/v] BSA in PBS.
6. Storage solution: PBS with 0.1% [w/v] sodium azide. Dilute 10% [w/v] sodium azide in PBS.

2.3 Sample Preparation for Staining of Focal Adhesions

1. Cytoskeleton buffer (CB [41]): 10 mM MES pH 6.1, 150 mM NaCl, 5 mM EGTA, 5 mM D-glucose, 5 mM $MgCl_2$. Filter sterile; do not autoclave.

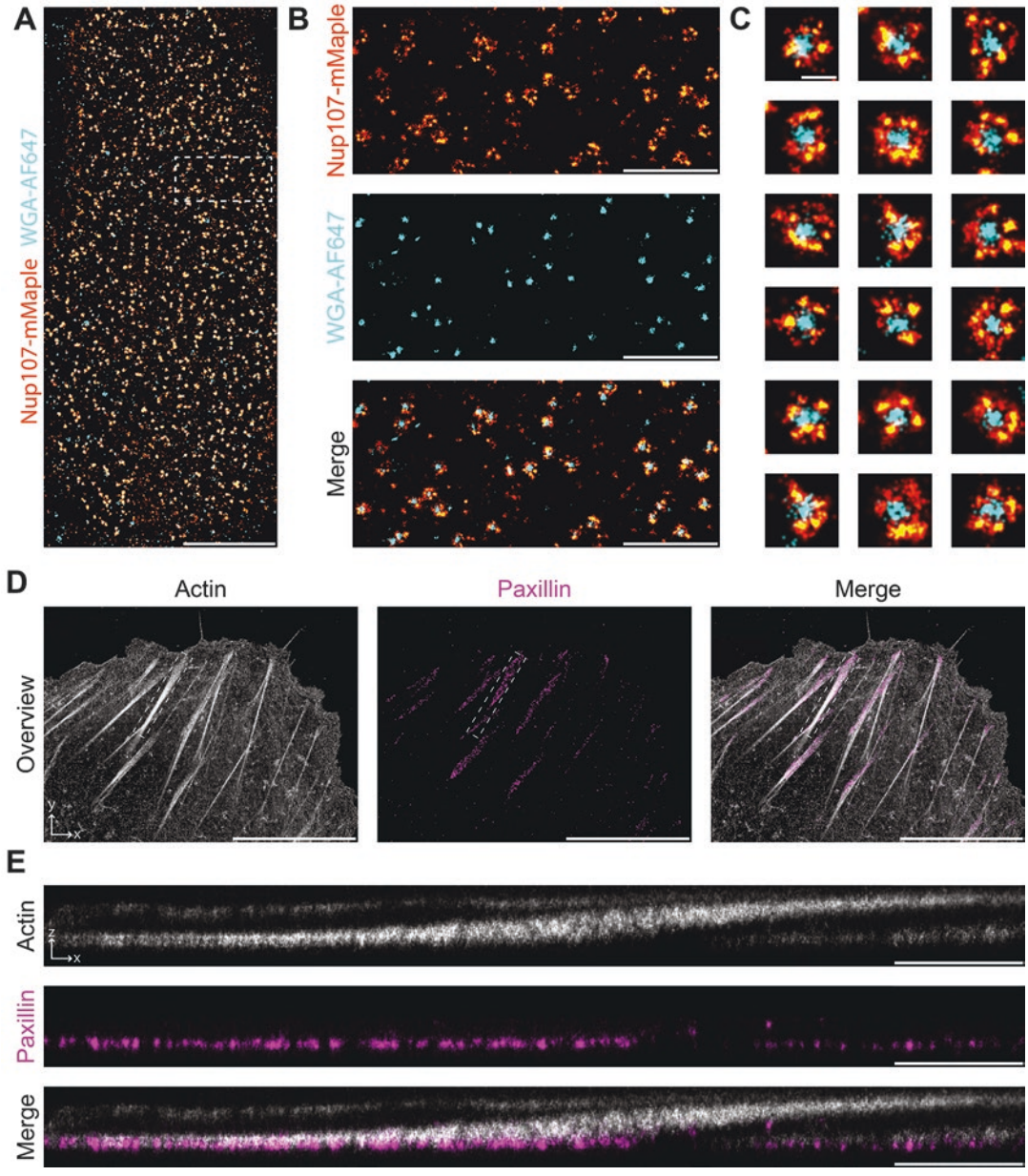


Fig. 2 Dual-color super-resolution imaging of the nuclear pore complex and focal adhesions. (**a**) Nup107-mMaple and WGA-AF647 were imaged simultaneously in two different channels separated by a 640 LP dichroic mirror. Scale bar: 5 μm. (**b**) Detailed view of NPCs in dual color as depicted in (**a**). Scale bars: 1 μm. (**c**) Individual view of selected NPCs. Scale bar: 100 nm. (**d**) The actin cytoskeleton (Phalloidin-AF647) and paxillin (antibody-staining, fluorophore: CF680) were imaged ratiometrically using a 680 LP dichroic mirror. The reconstruction of actin was rendered with a gamma factor of 0.5. Scale bars: 10 μm. (**e**) *x*–*z* projection of one focal adhesion as indicated in (**d**). Scale bars: 1 μm. For the colored version of this figure, please refer to the online version of the book chapter

2. Prefixation solution: 0.2% [v/v] glutaraldehyde (GA), 0.25% [v/v] Triton X-100 in CB. Prepare freshly from CB, Triton X-100, and 25% [v/v] GA.
3. Fixing solution: 2% [v/v] GA in CB. Prepare freshly from CB and 25% [v/v] GA.
4. Quenching solution: 0.1% [w/v] $NaBH_4$ in PBS. Prepare freshly just before needed.
5. Blocking solution: 2% [w/v] BSA in PBS. Prepare by dissolving BSA in PBS. Filter sterile and store at 4 °C.
6. Staining solutions:
   (a) Anti-paxillin primary antibody (Abcam; *see* **Note 2**): Dilution 1:300 in 2% [w/v] BSA in PBS.
   (b) Anti-rabbit secondary antibody labeled with CF680 (Sigma): Dilution 1:300 in 2% [w/v] BSA in PBS.
   (c) Phalloidin staining: Dilution 1:150 of phalloidin-AF647 (6.6 μM final concentration; Thermo Fisher Scientific) in PBS (*see* **Note 3**).

2.4 Imaging Buffers

1. For single-color imaging of mMaple: 50 mM Tris-HCl pH 8 in 95% [v/v] D_2O. Prepared by diluting a 1 M Tris-HCl pH 8 (H_2O) stock solution with D_2O (*see* **Note 4**). Filter sterile.
2. Glucose oxidase (GLOX) blinking buffer for imaging of synthetic dyes: 50 mM Tris-HCl pH 8, 10 mM NaCl, 10% [w/v] D-glucose, 35 mM 2-mercaptoethylamine (MEA), 500 μg/ml GLOX, 40 μg/ml catalase. Prepared freshly by mixing 50 μl of a 20× GLOX/catalase and 7 μl of a 5 M MEA stock with aliquoted buffer containing Tris, NaCl, and D-glucose (all components are stored at −20 °C).
3. D_2O GLOX blinking buffer for simultaneous imaging of mMaple and AF647: 50 mM Tris-HCl pH 8, 10 mM NaCl, 10% [w/v] D-glucose, 20 mM MEA, 500 μg/ml GLOX, 40 μg/ml catalase. Prepared as above but with 4 μl MEA and in 90% [v/v] D_2O.

2.5 Cell Lines

1. Imaging of the nuclear pore complex: We cloned Nup107 with an N-terminal fusion of mMaple and two miRNAs targeting the endogenous Nup107 under a Tet-inducible promoter in the plasmid pcDNA™5/FRT/TO (Thermo Fisher Scientific). Exactly one copy of the plasmid was genomically integrated using the Flp-In™ T-Rex™ 293 system (Thermo Fisher Scientific).
2. Imaging of the focal adhesions: U-2 OS cells.

2.6 Solutions for Cell Culture

1. PLL solution: 0.1% [w/v] poly-L-lysine in dH_2O.

2. Cell culture medium: Dulbecco's Modified Eagle Medium (DMEM, high glucose, w/o phenol red) supplemented with 10% [v/v] FBS, 2 mM L-glutamine, non-essential amino acids, ZellShield™ and 1 μg/ml tetracycline hydrochloride (diluted from a sterile filtered 1 mg/ml stock in water).
3. Trypsinization solution: TrypLE™ Enzyme Express.

2.7 Optical Setup

For imaging, we used a custom-built microscope. The excitation beam path consists of the laser box Toptica iChrome MLE with four lasers (wavelengths: 405 nm, 488 nm, 561 nm, and 640 nm) that are coupled in a single-mode fiber and an additional 640 nm booster laser (Toptica). The lasers are focused on the back focal plane of the objective (Nikon, NA 1.49, 60×) and the setup is operated in Epi-illumination mode. The focus is stabilized by an infrared laser which is totally internally reflected in the objective and monitored with a quadrant photodiode (QPD). Using an electric feedback loop, the objective is moved by a piezo objective positioner (Physical Instruments).

The emitted light is filtered (for AF647: 700/100; for mMaple: 613/73) and imaged onto an EM-CCD camera (Ixon Ultra, Andor). For dual-color imaging, the emitted light is laterally constricted by a slit, split by a dichroic mirror (for mMaple/AF647: 640 longpass [LP]; for AF647/CF680: 680 LP) and imaged onto different halves of the camera. Datasets containing three-dimensional (3D) information are generated by introducing astigmatism to the PSF by a cylindrical lens [30].

The microscope is controlled with LabVIEW (National Instruments) and data is acquired using Micro-Manager [42].

3 Methods

3.1 Preparation of Coverslips

1. Clean suitable glass coverslips overnight with cleaning solution. The next morning, wash the coverslips with water until the pH of the water used for washing is neutral, dry them under a laminar flow hood, and sterilize by UV illumination.
2. Before seeding of cells, coat the previously cleaned coverslips with PLL solution. Pipette 400 μl of the PLL solution onto the center of the coverslip in a 6-well plate and incubate for 2 h. Aspirate the PLL solution, wash the coverslips thoroughly with water, dry it for at least 2 h, and sterilize again by UV illumination.

3.2 Cell Culture

1. Cultivate the HEK293 cells under adherent conditions in cell culture medium at 37 °C, 5% CO_2, and 100% humidity. When reaching 80–90% confluency, trypsinize cells using trypsinization solution and split them 1:1 (approximately every 2–3 days).

2. Cultivate the U-2 OS cells as described for the HEK293 cells but leave out the tetracycline.
3. Seeding of coverslips with HEK293 cells: Seed the cells on PLL-coated coverslips at the same density as after splitting 1:1. Therefore, they will reach a confluency of about 90% after 3 days of growth. After the first day, transfer the coverslips to a new 6-well plate with fresh medium to remove surrounding cells.
4. Seeding of coverslips with U-2 OS cells: Seed the cells on washed coverslips (*see* **Note 5**) at a 1:10 dilution of a confluent plate. Grow cells for at least 1.5 days prior to fixation.

3.3 Preparation of Samples for Imaging of the Nuclear Pore Complex

All incubations are carried out at room temperature (RT) and in the dark (*see* **Note 6**); the coverslips are shaken in 6-well plates at 25 rpm during washing, fixation, and quenching, whereas the permeabilization, blocking, and staining are carried out under a humidified atmosphere.

1. Wash coverslips 3 times with 3 ml PBS shaking by hand for a few seconds to get rid of loose cells.
2. Fix cells with 2 ml of fixing solution for 30 min.
3. Wash coverslips 3 times with 4 ml PBS for 5 min.
4. Quench the fixing solution by incubating the coverslips for 15 min in 4 ml quenching solution.
5. Wash coverslips 3 times with 4 ml PBS for 5 min. When imaging only mMaple, the samples are ready for imaging after this step.
6. If not imaged or stained immediately, store coverslips in storage solution at 4 °C (*see* **Note 7**).
7. Pipette 100 µl of permeabilization solution on a new and clean Parafilm (*see* **Note 8**), and carefully put the coverslip face down on the drop (*see* **Note 9**) and incubate for 3 min.
8. Transfer the coverslips to a 6-well plate, and wash them three times with 4 ml PBS for 5 min each.
9. Block the sample with 100 µl of blocking solution face down on a new and clean Parafilm for 1 h.
10. Stain the cells with 100 µl of WGA-AF647 staining solution face down on a new and clean Parafilm for 5 min.
11. Repeat **step 8**.

3.4 Preparation of Samples for Imaging of Focal Adhesions

All incubations are carried out at RT; washing, prefixation, fixation, and quenching of the coverslips are carried out in a 6-well plate shaking at 25 rpm, whereas permeabilization, blocking, and staining are carried out under a humidified atmosphere.

1. Prefix the cells for 2 min with 2 ml of the prefixation solution. Directly transfer the coverslip into fixative without washing or removal of medium.
2. Fix the cells with 2 ml of the fixing solution for 10 min.
3. Quench the glutaraldehyde to remove autofluorescence by incubating in 2 ml of the quenching solution for 7 min.
4. Wash coverslips three times with 4 ml PBS for 5 min each until no more bubbles are formed.
5. Incubate the coverslips face down on a new and fresh Parafilm with 100 µl blocking solution for 1 h.
6. Stain the coverslips face down on a new and fresh Parafilm with the primary antibody for 1 h.
7. Wash coverslips three times with 4 ml PBS for 5 min each.
8. Stain the coverslips face down on a new and fresh Parafilm with the secondary antibody for 2 h (*see* **Note 10**).
9. Repeat **step 7**.
10. Stain the coverslips face down on a new and fresh Parafilm with phalloidin for 10 min (*see* **Note 3**).
11. Repeat **step 7**.

3.5 Preparation of Bead Sample

For dual-color imaging of mMaple together with AF647 on different parts of the camera, it is necessary to predetermine a transformation to overlay the two channels during post-processing. To do so, prepare a bead sample with TetraSpeck beads that are fluorescent in both channels.

1. Mount a coverslip in the sample holder (*see* **Note 11**).
2. Pipette 40 µl of 1 M $MgCl_2$ in the center of the coverslip (*see* **Note 12**).
3. Mix 360 µl of water with 2 µl of the TetraSpeck beads (*see* **Note 13**).
4. Pipette the water/TetraSpeck mixture on the coverslip, and mix with the $MgCl_2$ solution by pipetting up and down.

3.6 Acquisition of the Bead Sample for Channel Transformation

This is necessary for dual-color imaging of AF647 with mMaple.

1. Mount the sample holder on the microscope, and switch on the 561 nm and 640 nm laser to excite the beads in both channels.
2. Focus on the beads and adjust the tube lens of the second channel to minimize chromatic aberrations.
3. Adjust the laser powers so that the intensity of the same bead is roughly the same in both channels.
4. Acquire 50–70 individual frames that are translated in x and y.

3.7 Acquisition of the Bead Sample for 3D Calibration

1. Mount the sample holder and excite the beads with the 640 nm laser.
2. Find an area with a nice distribution of beads.
3. Insert the astigmatic lens.
4. Record a z-stack ±2 μm in 10 nm steps with an exposure time to maximize the signal but without saturating the image.

3.8 Imaging of Samples

1. Mount the coverslip in the sample holder, carefully pipette 500 μl of blinking buffer in the sample holder, clean the coverslip from the bottom with 70% [v/v] ethanol, and make sure it is not leaking.
2. Mount the sample holder on an oil immersion objective.
3. Use low laser power to find a region of interest (ROI) you want to image (*see* **Note 14**).
4. For 3D imaging, insert the astigmatic lens.
5. For imaging of mMaple in single color, set your acquisition parameters (exposure time, laser power, filter, number of frames; *see* **Note 15**), and start the acquisition. For imaging of a synthetic dye (in single or dual color), illuminate with full laser power (only 640 nm), and wait until you get single-molecule blinking. Then start the acquisition (for imaging in dual color together with mMaple, switch on the 561 nm laser before) with the desired parameters (*see* **Note 16**).
6. Start the UV activation to switch mMaple stochastically to the red state and/or to bring the synthetic dye(s) back to the bright state (*see* **Note 17**).

3.9 Post-acquisition Processing

The data analysis of the raw data was performed in SMAP (super-resolution microscopy analysis platform, unpublished, available upon request from the authors), a software framework based on MATLAB (MathWorks). The analysis can be performed in an analogous way by using, e.g., the freely available super-resolution software ThunderSTORM [43] or rapidSTORM [44].

3.9.1 Localization of Single Blinking Events

1. Peak detection is performed by wavelet-based background estimation [45], filtering, and maximum finding.
2. The detected peaks are selected based on a probabilistic threshold of $p < 0.01$. This means the peaks are with a probability of less than 1% the result of random noise.
3. Astigmatic bead stacks are fitted with an asymmetric Gaussian model to extract the size of the PSF in x and y. The z-dependence of these values is fitted with a polynomial, which is directly used in the fitting algorithm [46].
4. The individual fluorophores are localized by fitting a pixelized Gaussian function to the peaks. When fitting 3D data, the fitting algorithm uses the parameters obtained as described above to additionally determine the z-position of the fluorophores.

3.9.2 Localization of Beads to Determine the Transformation Between the mMaple and AF647 Channel

1. Based on the localizations of the beads, a projective transformation of both channels is determined that can later be used to overlay the two channels.

3.9.3 Post-processing of Ratiometric Imaging of AF647/ CF680

When imaging AF647 together with CF680, a 680 nm LP dichroic mirror is used to split the signal. In this configuration, about half of the AF647 signal is reflected by the mirror and the other half is transmitted, whereas most of the CF680 signal is transmitted.

1. The raw data is localized as for single-color data.
2. As AF647 is detected in both channels, it is possible to calculate a projective transformation for overlaying the two channels from the data itself, and it is not necessary to determine it using a bead sample.
3. The color of individual localizations is assigned using the ratio between the signals in both channels.

3.9.4 Drift Correction

1. The localizations are sorted by frame in which they were detected.
2. The sorted localizations are binned in typically ten chunks.
3. Individual super-resolution images are reconstructed for the binned data.
4. The pair-wise image cross-correlation of all images with all others is determined and fitted with a spline interpolation.
5. The lateral drift trajectory is calculated and corrected for.

3.9.5 Reconstruction of Super-Resolution Image

1. Localizations persisting over several frames and originating from one blinking event are grouped together.
2. Typically, the localizations are filtered using their following features: localization precision, size of fitted PSF, number in frames the localization is present, and number of the frame the fluorophore was detected in temporal order.
3. Individual localizations are plotted at their fitted coordinates as two-dimensional Gaussian distributions with a standard deviation that is proportional to their localization precision.
4. 3D data can be visualized by color coding the z-positions of fluorophores or by plotting a z-slice through the x–y plane (Fig. 2e).

4 Notes

1. The FA solution we are using does not contain methanol and is therefore not stabilized against oxidation. As the effective FA concentration decreases after opening an ampoule due to

oxidation, we reseal it using Parafilm and discard it if it is not used up after a few days.

2. If you have an antibody that you are already using for regular immunofluorescence microscopy, it might be necessary to optimize its concentration for using it in SMLM. In our hands, the concentration is rather on the higher end of the recommendation by the manufacturer, but it might also happen that an antibody that you are successfully using for immunofluorescence does not work for SMLM. One explanation is that the labeling density must be high enough to resolve small structures (Nyquist criterion, *see* Subheading 1).
3. Use 1:50 dilution of phalloidin-AF647 (20 μM final concentration) for single-color staining. Stain directly before imaging as phalloidin detaches from actin over time.
4. The use of heavy water is not absolutely necessary, but it has been shown to increase the number of emitted photons for photoactivatable fluorescent proteins [47].
5. For U-2 OS cells, it is not necessary to coat the coverslips with PLL as they adhere well enough without it.
6. Try to minimize light exposure of the samples to prevent premature conversion of mMaple to the red fluorescent state.
7. The samples can be stored for a few days and imaged or stained later. However, highest image quality is obtained when imaged immediately.
8. Make sure to put the clean side of the Parafilm face up. This is because any contamination in the sample that is fluorescent will affect the acquisition as the method has single-molecule sensitivity.
9. Shortly blot the coverslip on a Kimwipe before putting on the Parafilm to remove residual solution.
10. Commercially available secondary antibodies are usually labeled with several fluorophores. For dense labeling, this might lead to a too high density of activated fluorophores in the image while still leading to insufficient labeling densities. In this case, it might be advisable to label the secondary antibody yourself to obtain antibodies that only have one or two fluorophores attached.
11. As the fluorescence of the beads is very bright, it is not necessary to use the cleaned coverslips for imaging of beads.
12. The presence of divalent cations helps the TetraSpeck beads to stick to the surface of the coverslips.
13. Vortex the TetraSpeck beads before pipetting, and also vortex the water/TetraSpeck solution before pipetting it onto the coverslip.

14. When imaging mMaple, you can use its green fluorescent state to find a ROI. Be careful to not bleach mMaple because after bleaching the fluorescent protein in its green form, it cannot be converted to the red state anymore. To prevent bleaching, it is better to not look for a ROI in live mode, but take snapshots while moving around. Be aware of chromatic aberrations in your setup that lead to different focal planes for different emission wavelengths. When imaging in dual color, use the synthetic dye to find the ROI as it is way less sensitive to bleaching (and it will be converted to its dark state prior to single-molecule blinking anyway).
15. The values for these parameters have to be determined experimentally. The laser power densities are typically between 1 and 10 kW/cm^2, resulting roughly in bright state lifetimes of 20–100 ms. Adjust the exposure time so that individual blinks of molecules are visible in 1.5 frames on average. The number of frames should be sufficient to image all fluorophores before they finally bleach. Laser power should be optimized for best blinking performance of the fluorophore.
16. You can also acquire data right from the beginning and exclude the first frames where no single-molecule blinking is observed during post-processing.
17. In our software for the laser control, we are using an algorithm that counts the number of molecules every few frames and adjusts the pulse length of the pulsed UV laser to get a predefined number of blinks per frame.

Acknowledgments

We thank Ulf Matti for experimental assistance and Edward Lemke for the Flp-In™ T-Rex™ 293 cell line. This work was supported by EMBL International PhD Programme fellowships (P.H. and M.M.).

References

1. Yamanaka M, Smith NI, Fujita K (2014) Introduction to super-resolution microscopy. Microscopy (Oxf) 63:177–192. https://doi.org/10.1093/jmicro/dfu007
2. Gustafsson MG (2000) Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. J Microsc 198:82–87
3. Hell SW, Wichmann J (1994) Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. Opt Lett 19:780–782. https://doi.org/10.1364/OL.19.000780
4. Klar TA, Hell SW (1999) Subdiffraction resolution in far-field fluorescence microscopy. Opt Lett 24:954–956. https://doi.org/10.1364/OL.24.000954
5. Klar TA, Jakobs S, Dyba M et al (2000) Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. Proc Natl Acad Sci U S A 97:8206–8210

6. Betzig E, Patterson GH, Sougrat R et al (2006) Imaging intracellular fluorescent proteins at nanometer resolution. Science 313:1642–1645. https://doi.org/10.1126/science.1127344
7. Hess ST, Girirajan TPK, Mason MD (2006) Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. Biophys J 91:4258–4272. https://doi.org/10.1529/biophysj.106.091116
8. Rust MJ, Bates M, Zhuang X (2006) Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Methods 3:793–795. https://doi.org/10.1038/nmeth929
9. Sharonov A, Hochstrasser RM (2006) Wide-field subdiffraction imaging by accumulated binding of diffusing probes. Proc Natl Acad Sci U S A 103:18911–18916. https://doi.org/10.1073/pnas.0609643104
10. Thompson RE, Larson DR, Webb WW (2002) Precise nanometer localization analysis for individual fluorescent probes. Biophys J 82:2775–2783. https://doi.org/10.1016/S0006-3495(02)75618-X
11. Shroff H, White H, Betzig E (2008) Photoactivated localization microscopy (PALM) of adhesion complexes. Curr Protoc Cell Biol Chapter 4:Unit 4.21–Unit 4.27. https://doi.org/10.1002/0471143030.cb0421s41
12. Nyquist H (1928) Certain topics in telegraph transmission theory. Trans Am Inst Electr Eng 47:617–644. https://doi.org/10.1109/T-AIEE.1928.5055024
13. Keppler A, Gendreizig S, Gronemeyer T et al (2003) A general method for the covalent labeling of fusion proteins with small molecules in vivo. Nat Biotechnol 21:86–89. https://doi.org/10.1038/nbt765
14. Gronemeyer T, Chidley C, Juillerat A et al (2006) Directed evolution of O6-alkylguanine-DNA alkyltransferase for applications in protein labeling. Protein Eng Des Sel 19:309–316. https://doi.org/10.1093/protein/gzl014
15. Gautier A, Juillerat A, Heinis C et al (2008) An engineered protein tag for multiprotein labeling in living cells. Chem Biol 15:128–136. https://doi.org/10.1016/j.chembiol.2008.01.007
16. Los GV, Encell LP, McDougall MG et al (2008) HaloTag: a novel protein labeling technology for cell imaging and protein analysis. ACS Chem Biol 3:373–382. https://doi.org/10.1021/cb800025k
17. Lukinavičius G, Umezawa K, Olivier N et al (2013) A near-infrared fluorophore for live-cell super-resolution microscopy of cellular proteins. Nat Chem 5:132–139. https://doi.org/10.1038/nchem.1546
18. Ries J, Kaplan C, Platonova E et al (2012) A simple, versatile method for GFP-based super-resolution microscopy via nanobodies. Nat Methods 9:582–584. https://doi.org/10.1038/nmeth.1991
19. Patterson GH, Lippincott-Schwartz J (2002) A photoactivatable GFP for selective photolabeling of proteins and cells. Science 297:1873–1877. https://doi.org/10.1126/science.1074952
20. McKinney SA, Murphy CS, Hazelwood KL et al (2009) A bright and photostable photoconvertible fluorescent protein. Nat Methods 6:131–133. https://doi.org/10.1038/nmeth.1296
21. Zhang M, Chang H, Zhang Y et al (2012) Rational design of true monomeric and bright photoactivatable fluorescent proteins. Nat Methods 9:727–729. https://doi.org/10.1038/nmeth.2021
22. Gunewardene MS, Subach FV, Gould TJ et al (2011) Superresolution imaging of multiple fluorescent proteins with highly overlapping emission spectra in living cells. Biophys J 101:1522–1528. https://doi.org/10.1016/j.bpj.2011.07.049
23. McEvoy AL, Hoi H, Bates M et al (2012) mMaple: a photoconvertible fluorescent protein for use in multiple imaging modalities. PLoS One 7:e51314. https://doi.org/10.1371/journal.pone.0051314
24. Bates M, Huang B, Dempsey GT, Zhuang X (2007) Multicolor super-resolution imaging with photo-switchable fluorescent probes. Science 317:1749–1753. https://doi.org/10.1126/science.1146598
25. Testa I, Wurm CA, Medda R et al (2010) Multicolor fluorescence nanoscopy in fixed and living cells by exciting conventional fluorophores with a single wavelength. Biophys J 99:2686–2694. https://doi.org/10.1016/j.bpj.2010.08.012
26. Shroff H, Galbraith CG, Galbraith JA et al (2007) Dual-color superresolution imaging of genetically expressed probes within individual adhesion complexes. Proc Natl Acad Sci U S A 104:20308–20313. https://doi.org/10.1073/pnas.0710517105
27. Tam J, Cordier GA, Borbely JS et al (2014) Cross-talk-free multi-color STORM imaging using a single fluorophore. PLoS One 9:e101772. https://doi.org/10.1371/journal.pone.0101772
28. Valley CC, Liu S, Lidke DS, Lidke KA (2015) Sequential superresolution imaging of multiple

targets using a single fluorophore. PLoS One 10:e0123941. https://doi.org/10.1371/journal.pone.0123941

29. Jungmann R, Avendaño MS, Woehrstein JB et al (2014) Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and exchange-PAINT. Nat Methods 11:313–318. https://doi.org/10.1038/nmeth.2835
30. Huang B, Wang W, Bates M, Zhuang X (2008) Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. Science 319:810–813. https://doi.org/10.1126/science.1153529
31. Pavani SRP, Thompson MA, Biteen JS et al (2009) Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. Proc Natl Acad Sci U S A 106:2995–2999. https://doi.org/10.1073/pnas.0900245106
32. Shechtman Y, Sahl SJ, Backer AS, Moerner WE (2014) Optimal point spread function design for 3D imaging. Phys Rev Lett 113:133902. https://doi.org/10.1103/PhysRevLett.113.133902
33. Mlodzianoski MJ, Juette MF, Beane GL, Bewersdorf J (2009) Experimental characterization of 3D localization techniques for particle-tracking and super-resolution microscopy. Opt Express 17:8264–8277. https://doi.org/10.1364/OE.17.008264
34. Hajj B, Wisniewski J, Beheiry El M et al (2014) Whole-cell, multicolor superresolution imaging using volumetric multifocus microscopy. Proc Natl Acad Sci U S A 111:17480–17485. https://doi.org/10.1073/pnas.1412396111
35. Nagorni M, Hell SW (1998) 4Pi-confocal microscopy provides three-dimensional images of the microtubule network with 100- to 150-nm resolution. J Struct Biol 123:236–247. https://doi.org/10.1006/jsbi.1998.4037
36. Kanchanawong P, Shtengel G, Pasapera AM et al (2010) Nanoscale architecture of integrin-based cell adhesions. Nat Publ Group 468:580–584. https://doi.org/10.1038/nature09621
37. Dudok B, Barna L, Ledri M et al (2015) Cell-specific STORM super-resolution imaging reveals nanoscale organization of cannabinoid signaling. Nat Neurosci 18:75–86. https://doi.org/10.1038/nn.3892
38. Xu K, Zhong G, Zhuang X (2013) Actin, spectrin, and associated proteins form a periodic cytoskeletal structure in axons. Science 339:452–456. https://doi.org/10.1126/science.1232251
39. Suleiman H, Zhang L, Roth R et al (2013) Nanoscale protein architecture of the kidney glomerular basement membrane. elife 2:e01149. https://doi.org/10.7554/eLife.01149
40. Szymborska A, de Marco A, Daigle N et al (2013) Nuclear pore scaffold structure analyzed by super-resolution microscopy and particle averaging. Science 341:655–658. https://doi.org/10.1126/science.1240672
41. Xu K, Babcock HP, Zhuang X (2012) Dual-objective STORM reveals three-dimensional filament organization in the actin cytoskeleton. Nat Methods 9:185–188. https://doi.org/10.1038/nmeth.1841
42. Edelstein AD, Tsuchida MA, Amodaj N et al (2014) Advanced methods of microscope control using μManager software. J Biol Methods 1:10. https://doi.org/10.14440/jbm.2014.36
43. Ovesný M, Křížek P, Borkovec J et al (2014) ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. Bioinformatics 30:2389–2390. https://doi.org/10.1093/bioinformatics/btu202
44. Wolter S, Löschberger A, Holm T et al (2012) rapidSTORM: accurate, fast open-source software for localization microscopy. Nat Methods 9:1040–1041. https://doi.org/10.1038/nmeth.2224
45. Izeddin I, Boulanger J, Racine V et al (2012) Wavelet analysis for single molecule localization microscopy. Opt Express 20:2081–2095. https://doi.org/10.1364/OE.20.002081
46. Smith CS, Joseph N, Rieger B, Lidke KA (2010) Fast, single-molecule localization that achieves theoretically minimum uncertainty. Nat Methods 7:373–375. https://doi.org/10.1038/nmeth.1449
47. Ong WQ, Citron YR, Schnitzbauer J et al (2015) Heavy water: a simple solution to increasing the brightness of fluorescent proteins in super-resolution imaging. Chem Commun (Camb) 51:13451–13453. https://doi.org/10.1039/c5cc04575d

Chapter 15

Correlative 3D Structured Illumination Microscopy and Single-Molecule Localization Microscopy for Imaging Cancer Invasion

Shannon J. L. Pinnington, John F. Marshall, and Ann P. Wheeler

Abstract

Super-resolution microscopy methods enable resolution of biological molecules in their cellular or tissue context at the nanoscale. Different methods have their strengths and weaknesses. Here we present a method that enables correlative confocal, structured illumination microscopy (SIM) and single-molecule localization microscopy (SMLM) imaging of structures involved in formation of invadopodia on the same sample. This enables up to four colors to be visualized in three dimensions at a resolution of between 120 and 10 nm for SIM and SMLM, respectively.

Key words Invasion, Microscopy, Super-resolution, Cells

1 Introduction

Super-resolution imaging enables an improvement in resolution of the visualization of biological structures in their cellular or tissue context between twice and 20 times [1]. Unfortunately not all methods are created equally, with some methods such as SIM or Airyscan imaging enabling a massive improvement in 3D imaging or visualization of multiple fluorescent labels, but a two-fold resolution improvement. Others such as SMLM methods allow an incredible 20-fold improvement in imaging [2–5] but make it challenging to visualize more than one fluorescent label [6] (Fig. 1). Generally sample preparation methods have meant that a decision has to be made about which super-resolution imaging method will be used for a given sample. This can be disadvantageous where a resolution improvement beyond the 250 nm Abbe limit is required for multiple channels and one or two channels require a considerable increase in resolution. Or it may mean that one scientific question can be answered at the expense of another. In experiments using systems which can be

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_15,

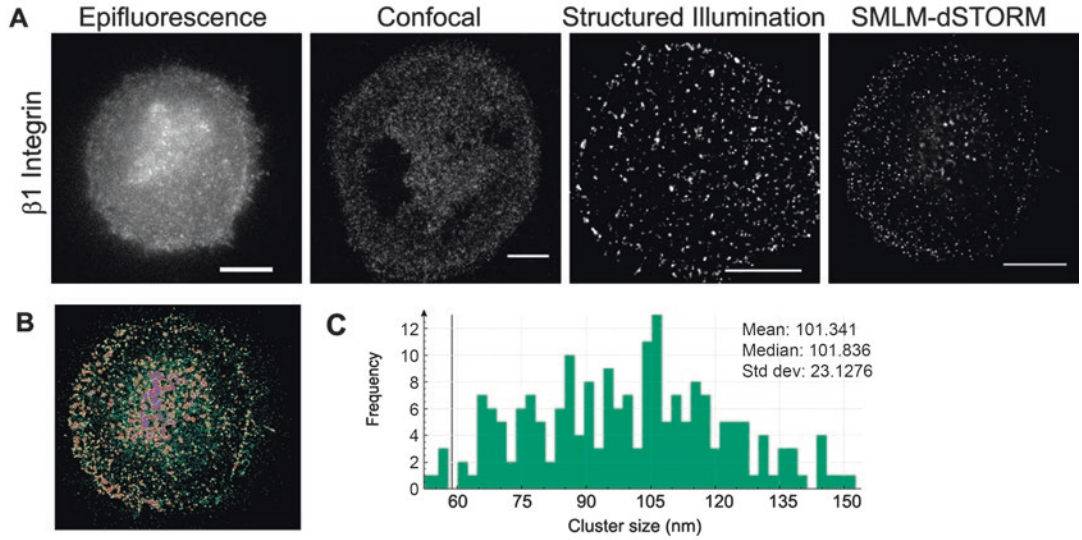


Fig. 1 Resolution improvement using super-resolution approaches. (**a**) VB6 oral squamous carcinoma cells stained using mouse anti-β6 integrin antibodies and Alexa Fluor 647-labeled FAb2 fragment donkey anti-mouse secondary antibodies as visualized by epifluorescence microscope, confocal microscopy, structured illumination microscopy, and dSTORM single-molecule localization microscopy. Bar = 10 μm. In all cases, images were acquired on a Nikon Ti2 microscope using a 100× 1.49NA objective. (**b**) Image showing cluster analysis of dSTORM data using SR-Tesseler. (**c**) Histograms showing cluster sizes of β6 integrins using the DB scan algorithm in the SR-Tesseler package

more challenging to handle, such as primary cells, organotypic cells, ES cells, and rare tissue samples, this may be less than ideal [7, 8]. Integrins are particularly challenging to image using conventional microscopy as they form small complexes that are spatially close to one another. In conventional microscopy, which is diffraction limited, the blur generated by image diffraction makes these complexes appear to be uniform staining (Fig. 1), although biochemical assays and electron microscopy analyses show this not to be the case [9]. Here we present a method for visualization of up to four structures involved in the process of invasion in transformed cell lines at twice diffraction limited resolution, with an option to allow one of these structures to be visualized to approximately 10 nm resolution with SMLM approaches, using the same sample.

Invadopodia are very small organelles (>2 μm) which are known to be involved in degradation of the extracellular matrix and have been indicated to play a key role in metastatic spread. They are known to contain metalloproteases, such as MT-MMP1 and MMP9, and can degrade the extracellular matrix (ECM) at focal points (Fig. 2a). The colocalization of the focal degradation and F-actin foci are known as invadopodia [10, 11]. Invadopodia are highly dynamic structures and turn over in minutes [12]. This means at a fixed time point, there will be active invadopodia, indi-

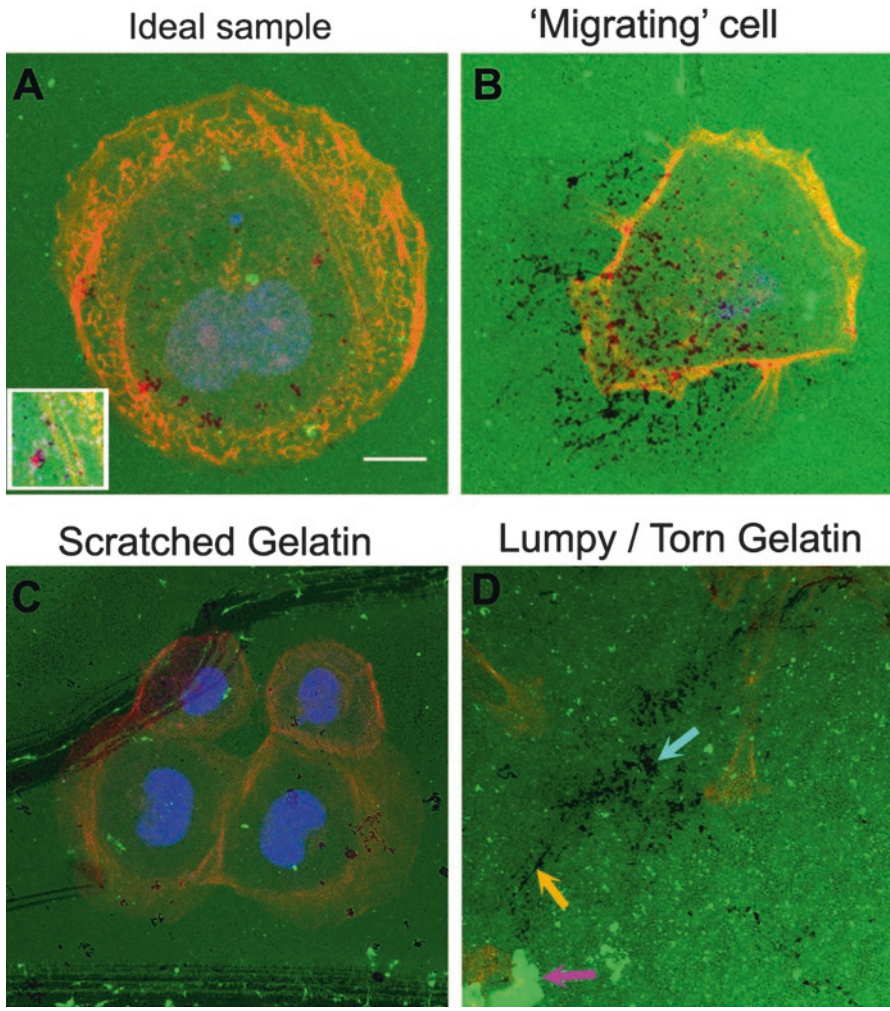


Fig. 2 Visualization of invadopodia using the gelatin degradation assay. (**a**) A VB6 oral squamous carcinoma cell degrading the gelatin extracellular matrix. Gelatin is shown in green, F-actin in red, and nuclei in blue. Punctate holes in the gelatin indicate invadopodia formation and colocalization of a dot of F-actin and indicate an active invadopodia. (**b**) A cell with a migratory morphology indicates invadopodia has formed and completely turned over. As can be seen, there are around 100 foci which are either under or immediately adjacent to the cell which suggests this cell has recently formed them. For quantification purposes, these invadopodia can be treated as belonging to the cell. (**c**) Scratched gelatin. This occurs when the substrates are mishandled during either cell seeding or fixation. Pipette or forceps mishandling can scratch the fragile gelatin surfaces leading to detachment of the substrate. (**d**) Over-degradation and shearing of gelatin. Here MBA-MB-231 breast carcinoma cells have completely degraded some of the matrix leading to formation of a very large hole—indicated by a cyan arrow. The invadopodia have coalesced into one larger hole, which would, in theory, allow the cell to move through the ECM. These structures can't be quantified as it isn't clear how many invadopodia were present. Shearing of the ECM—long thin lines indicated by an orange arrow—shows where the cell traction forces have stripped the gelatin off the glass coverslip. Either a thicker gelatin layer or a shorter incubation time of cells can reduce this problem

cated by a focus of F-actin colocalized with a hole in the ECM, and invadopodia which have turned over where only a hole in the ECM is present (Fig. 2a). Becauses of the small size of invadopodia detailed study of which proteins are localized in invadopodia, whether there are phases of protein recruitment to invadopodia, and dynamics, can be challenging. This is mostly because such studies are impeded by the diffraction limit of conventional light microscopes to 200 nm [13]. Using super-resolution, more information can be obtained. However, invadopodia are 3D structures and comprise of several elements—the focal degradation of ECM, localization of F-actin, and recruitment of other proteins such as integrins, which form small complexes.

SMLM imaging of all three of these epitopes over a 6 μm axial range with current equipment is impossible; however, using SIM this is possible, albeit that resolution will be limited to 120 nm. Choice of refraction index-matched coverslips and mounting medium containing anti-fade is important for successful SIM reconstruction. A major mismatch between the setting of the correction collar on the objective lens, the coverslip thickness, and refractive index of the mounting medium can generate spherical aberration artifacts in the final image [14]. The commercial mounting medium Vectashield can be used for SMLM imaging using the dye Alexa-647 and yields images comparable or superior to those obtained with more complex buffers, especially for 3D imaging [15]. We also find that with slight adjustment of the correction collar on our SIM system, combined with the use of high-precision coverslips, correlative SIM/SMLM/confocal imaging of our samples is enabled. For our invadopodia assay, this allows ECM degradation and the F-actin cytoskeleton to be visualized in 3D to 120 nm resolution (Fig. 3). Integrin clusters, providing they are labeled with Alexa Fluor 647 Fab fragments, can then be visualized using 2D STORM SMLM imaging to 10 nm resolution (Fig. 1a) and post hoc analysis of cluster size carried out [16] (Fig. 1b). The following method is used in our laboratory and can be adapted for other applications or proteins by adjustment of the antibodies used at the immunofluorescence stage of sample preparation. Readers must note that the protein for which the greatest resolution improvement is required must be labeled by antibodies labeled with Alexa Fluor 647 dye molecules since Alexa Fluor 647 is compatible with SIM and SMLM.

2 Materials

1. VB6 cells [17] or other cells with invasive phenotype.
2. Keratinocyte growth medium comprising: α-MEM containing 10% fetal calf serum (Gibco) supplemented with 100 IU l^{-1} penicillin, 100 μg/L streptomycin and 2.5 μg/L amphoteri-

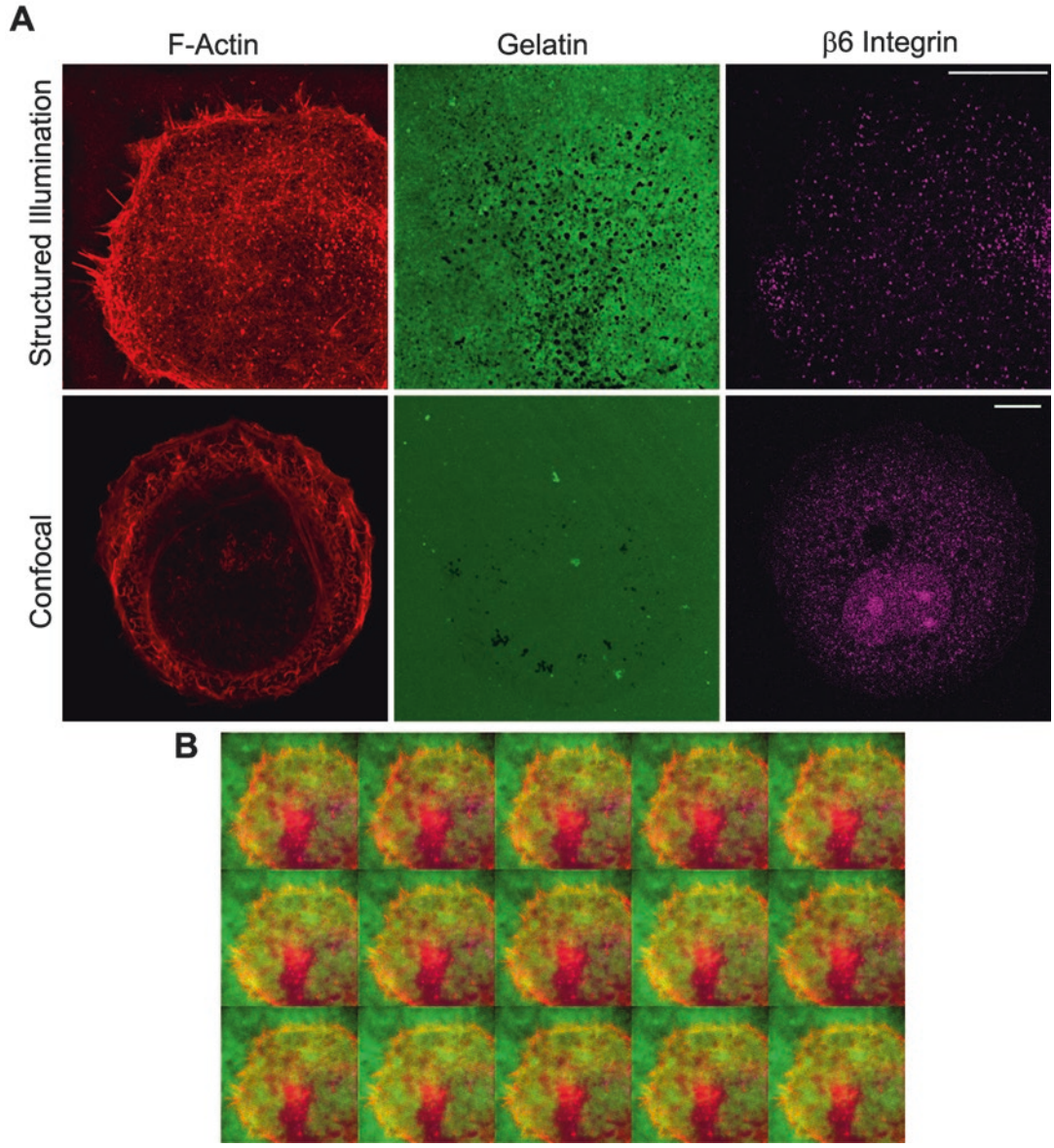


Fig. 3 (**a**) A comparison of multicolor structured illumination microscopy and confocal imaging for invadopodia formation. VB6 oral squamous carcinoma cells were seeded onto a gelatin substrate (green) for 6 h and stained for F-actin (red) and β1 integrin (magenta). The montage shows confocal and reconstructed SIM images. Bar = 10 μm. (**b**) A representative image of the raw SIM dataset used to generate the reconstruction used in (**a**)

cin B (Gibco), 1.8×10^{-4} M adenine, 5 μg/mL insulin, 1×10^{-10} M cholera toxin, 0.5 μg/mL hydrocortisone, and 10 ng/mL epidermal growth factor (Sigma).

3. High-precision number 1.5 18mm^2 glass coverslips: (ZEISS 474030-9010-000, Marienfeld Cat.No. 0107032, or round 18 mm diameter Cat.No. 0117580).

4. Gelatin from pig skin, Oregon Green® 488 conjugate (Thermo Fisher).
5. 0.5% glutaraldehyde (diluted from a 25% EM grade stock, Sigma).
6. Sodium borohydride $NaBH_4$ (Sigma-Aldrich).
7. Phosphate buffered saline (Thermo Fisher Scientific).
8. 4% paraformaldehyde (diluted from a 16% EM grade stock VWR resell for Electron Microscopy Sciences).
9. 0.2% Triton x100 (Sigma).
10. 6-well tissue culture plates.
11. Fetal bovine serum (Gibco).
12. Primary antibodies (*see* **Note 1**): anti-integrin alpha V + beta 6 antibody [10D5] (Abcam); anti-integrin beta 1 antibody [P4C10] (Novus Biologicals).
13. Alexa Fluor 568 Phalloidin (Thermo Fisher Scientific).
14. Alexa Fluor 647 Fab2 anti-mouse (Thermo Fisher Scientific).
15. DAPI (Sigma).
16. Vectashield (Vector Laboratories).
17. Parafilm.
18. Aluminum foil.
19. Nail polish.
20. Fine forceps.
21. Nikon N-SIM microscope equipped with a 100× 1.49NA objective, 405, 488, 561, and 640 nm laser lines Nikon N-STORM/Confocal system equipped with 100× 1.49NA objective and 300 mW 647 nm laser (Nikon Instruments).
22. Fiji ImageJ (www.Fiji.sc) including the following plugins: SIMcheck, ClearVolume.
23. ThunderSTORM software [18] (https://github.com/zitmen/thunderstorm).
24. SR-Tesseler [16].

3 Methods

3.1 Making Gelatin Substrates for Degradation Assay

1. Put down Parafilm using ethanol to stick flat to tissue culture hood.
2. Defrost gelatin on ice for 6 h prior to experimentation, and spin down gelatin to remove clumps (*see* **Note 2**).
3. Place 40 μL drops of gelatin onto the Parafilm.
4. Place a coverslip on top of each gelatin drop using the fine forceps.

5. Incubate for 5 min at room temperature in the dark (using foil to cover).
6. Place 40 μL drops of 0.5% PBS glutaraldehyde onto the Parafilm beside the coverslips.
7. Place the gelatin-coated coverslips onto the drops of PBS glutaraldehyde to fix gelatin to the coverslip, using fine forceps to move.
8. Incubate for 15 min at room temperature in the dark.
9. Add >0.5 mL (excess) PBS to a corresponding number of wells of the 6-well plate, and move coverslips into the PBS—gelatin side up.
10. Wash each coverslip twice with PBS.
11. Aspirate off PBS and add >0.5 mL (excess) of PBS $NaBH_4$ to each well.
12. Incubate for 3 min in an incubator in the dark (eliminates residual aldehydes).
13. Wash three times with PBS (or until no more bubbles).
14. Aspirate off PBS (*see* **Note 2**).
15. Add 2 mL cell solution (*see* **Note 3**) to each gelatin coverslip (*see* Subheading 4).
16. Plates are then incubated for 4–6 h (*see* **Note 4**) (37 °C, 5% CO_2, 100% humidified).

3.2 Indirect Immunofluorescence: Fixation, Immunolabeling, and Mounting

1. Fix cells in 2 mL 4% paraformaldehyde per coverslip, and leave for 20 min at room temperature.
2. Aspirate off paraformaldehyde and wash four times in PBS. For the fifth wash, leave in PBS for 5 min to remove residual fixative.
3. Aspirate off PBS and add 2 mL 0.2% Triton, and leave for 5 min at room temperature (not in the dark).
4. Aspirate off Triton and wash five times in PBS.
5. Block in 10% FBS for 30 min.
6. Make up solutions of both β1 and β6 primary antibodies (*see* **Note 4**) and incubate coverslips in a humidified chamber in the dark overnight at 4 °C. 100 μL primary antibody per coverslip is required.
7. Aspirate off primary antibodies and wash 5× with PBS.
8. Make up secondary antibody solution Fab AF647 mouse (1:500) (*see* **Note 1**), phalloidin 568 (1:1000), and DAPI (1:4000), and keep in the dark (wrap microfuge tube in foil) on ice; 100 μL secondary antibody per coverslip is required.
9. Incubate coverslips in secondary antibody in the dark, in a humidified chamber for 1 h at room temperature
10. Aspirate off secondary antibody solution and wash 5× in PBS.

11. Mount the stained coverslips in 20 μL Vectashield on glass slides, and ensure the coverslips are placed fairly centrally.
12. Affix coverslips to the glass slides using nail polish.
13. Label completed coverslips and leave flat in the fridge to set.
14. Immediately prior to imaging, gently wash the coverslip with double distilled water using a cotton bud wrapped in microscope lens cleaning tissue (*see* **Note 5**).

3.3 Super-Resolution Microscopy

Comparative images using standard, resolution limited imaging methods were acquired using a Nikon Ti Microscope stand and a 100× 1.49 Apo TIRF objective on the same imaging platform which the dSTORM SMLMs images were acquired. In all cases in the comparative study, the same sample was used for both standard and super-resolution imaging. The epifluorescent image was taken immediately prior to acquisition of the SMLM data of the same cell (Fig. 1a). The confocal images were acquired using a Nikon A1 scan head in Nikon Elements software. The confocal scanhead was attached to the left-hand side port of the Ti microscope and 488, 561, and 647 lasers, similar to the SIM images.

3.4 Structured Illumination Microscopy

1. 3D SIM images are acquired on a N-SIM (Nikon Instruments, UK) using a 100× 1.49NA lens and refractive index-matched immersion oil (Nikon Instruments). Samples are imaged using a Nikon Plan Apo TIRF objective (NA 1.49, oil immersion) and an Andor DU-897X-5254 camera using 405, 488, 561, and 640 nm laser lines (*see* **Note 5**).
2. To acquire SIM images, set the Z stack collection to range around a center point. Set the center point using the F-actin (568 phalloidin) channel. Set the focal plane corresponding to the bottom of the cell (*see* **Note 6**).
3. A range around the center point of 2 μm was set as this allows focused images of the gelatin degradation, actin, and invadopodia-associated proteins to be acquired.
4. For SIM acquisition, the highest laser power and shortest exposure time were selected to minimize photobleaching and speed data acquisition (*see* **Note 7**).
5. Z-step size for Z stacks was set to 0.120 μm as required by manufacturer's software. For each focal plane, 15 images (5 phases, 3 angles) were captured with the NIS-Elements software. SIM image processing, reconstruction ,and analysis were carried out using the "Stack" option in the N-SIM module of the NIS-Element Advanced Research software. In all SIM image reconstructions, the Wiener and Apodization filter parameters were kept constant. Data were saved in the .nd2 Nikon proprietary format to retain metadata.

6. For a comparator with standard resolution data, datasets were acquired using the N-SIM in wide-field mode. Here the same Z stack was acquired but the grating and phase mask required for SIM acquisition is removed from the microscope light path.
7. Images were checked for artifacts and resolution using the SIMcheck software [19].
8. Data were analyzed for number of invadopodia per cell, integrins, size of invadopodia-mediated degradation using Fiji macros and visualized for presentation in 3D using ClearVolume [20].

3.5 Single-Molecule Localization Microscopy

1. dSTORM images are acquired on an N-STORM system (Nikon Instruments, UK) using a 100× 1.49NA lens and refractive index-matched immersion oil (Nikon Instruments). Images were acquired with the sample illuminated using total internal reflection fluorescence. So only the integrins on the basal surface of the cell could be visualized (*see* **Note 8**).
2. Samples were imaged using a Nikon Plan Apo TIRF objective (NA 1.49, oil immersion) and an Andor DU-897X-5254 camera, set at EM gain 300 and with conversion settings of 3, using a 640 nm laser lines set at 300 mw power.
3. Images were "back-pumped" using the 647 laser set at 100% until single-molecule photoswitches could be visualized [6]. Datasets of 10,000 images were collected with the camera streamed at 20 Hz. Data was saved as a .tif file and preliminary analysis carried out in the N-STORM software according to manufacturers' instructions. An estimate of localization precision for the whole dataset was obtained from these analyses (10 nm).
4. Tif stacks were analyzed using ThunderSTORM [18] (*see* **Note 9**). Images were filtered using the B-Spline wavelet filter with default settings, and molecules were approximately localized using the centroid of connected components, with software default settings. Sub-pixel localizations were assigned using an integrated Gaussian model of the point spread function with a 3 pixel fitting radius and maximum likelihood fitting with an initial sigma of 1.6 assigned for fitting (*see* **Note 10**).
5. Super-resolution images were visualized using average shifted histograms.
6. Data was drift corrected, filtered with a density filter of 50 nm, and duplicate localizations removed; localizations within a 2 nm radius—likely arising from the same secondary antibody (*see* **Note 11**)—were merged. Finally localizations with an error of fitting of greater than 100 nm were excluded as they were above the threshold required for clustering.

7. Data tables were exported in .csv format and imported into SR-Tesseler [16] for clustering analysis. Analysis was carried out as described in the paper.
8. Cluster information from DB-Scan and Voronoi segmentation of clusters were exported to Microsoft Excel. Datasets of ten cells per condition were collected for cluster analysis.

4 Notes

1. Here we visualize integrins αvβ6 and β1 separately. However, any protein with a high affinity and avidity antibody with low background can be used for these assays. It is important to have the protein, which should be visualized in SMLMS labeled with an Alexa Fluor 647 secondary antibody (Figs. 1a and 3).
2. The gelatin coating of coverslips must be carried out with great care. It is essential to either use the coverslips of the size recommended or adjust the amount of gelatin used if a larger coverslip is used so that the coating of the coverslip is completely uniform (Fig. 2a). We recommend aliquotting the gelatin from a stock solution as repeated freeze-thaw cycles degrades the quality of gelling of the gelatin and can cause it to clump. No more than three freeze-thaw cycles for a gelatin aliquot is recommended. Gelatin must be slowly defrosted and centrifuged before use so precipitated gelatin and gelled "clumps" of gelatin are not used (Fig. 2b for an example). While it is challenging to completely avoid clumps, it is very difficult to visualize invadopodia in them, they create an uneven structure for the cells to spread on, and the variation in brightness from the clumps makes batch quantification of images using macros or scripts very difficult. Once the substrate is made, it can be coated with extracellular matrices such as collagen, fibronectin, or laminin or left overnight in an incubator at 37 °C and 5% CO_2 in the dark. Overnight incubation may be desirable as the gelatin degradation assay may take up to 7 h the following day.
3. Cells should be plated in their normal growth medium, and the concentration of cells for the assay should be optimized. Ideally cells should be seeded as a single cell suspension to facilitate visualization of gelatin degradation under the cell. We find concentrations of 2×10^3–2×10^4 cells/mL work best, dependent on cell type. Care must be taken when seeding and fixing cells not to accidentally tear the gelatin as this reduces the amount of quantifiable area on the sample considerably; we recommend using very fine forceps for handling and seeding cells using either a Gilson or a 5 mL pipette so the substrate is not accidentally touched (Fig. 2b—scratched substrate).

4. The length of time required for substrate degradation varies dependent on cell type. We find that aggressive cell types such as MBA-MB-231 require 4 h and head and neck squamous carcinoma cell line VB6 requires 6 h. The assay can be refined by fixing the cells at discreet time points after seeding onto gelatin, e.g., 2, 3, 4, 5, and 6 h. If the assay is left too long, the substrate will become too degraded, and it will not be possible to determine where degraded foci corresponding to invadopodia are, or cells may have migrated away from sites of degradation or may have torn the membrane (Fig. 2b—over-degraded gelatin). If insufficient time is allowed for the assay, then no invadopodia will form.
5. The SIM will need to be calibrated with test samples, such as 100 nm TetraSpeck beads (Thermo Fisher Scientific) suspended in Vectashield to minimize spherical aberrations. On Nikon systems, this is carried out by adjustment of the correction collar on the 100× objective lens. On other systems, manufacturers should advise. This calibration should be carried out by experienced microscopists only; the system manager, where available, should be approached for advice here.
6. Where possible a whole cell was imaged, when this wasn't possible due to the large size of the cell, a field of view with as much of the lamellipodia/plasma membrane visible was selected (Fig. 3).
7. SIM imaging can be time-consuming, as 15 images need to be acquired to generate one "reconstructed" super-resolution image that is used for quantification and presentation. To speed this process up, fast acquisition times were selected; we found that 30 cells per experimental condition were the minimum we could acquire to obtain statistical robust results and would recommend collection of as many cells/invadopodia as possible (Fig. 3). We recommend retaining both the "Raw" SIM image comprising of the 2 μm Z stack of 15 images and reconstructed SIM image as raw data in the study.
8. To reduce drift in SMLM experiments, the imaging system was stabilized to 25 °C, and the Perfect Focus System on our Nikon Ti microscope—which minimizes axial drift—was used. Other microscope manufacturers also have focus feedback, which reduces axial drift (e.g., Definite Focus, Zeiss). We recommend speaking to the manufacturer for advice here.
9. The camera settings were kept constant during SMLM data acquisition and were entered yielding an intensity count to photon conversion factor of 4.8.
10. The SMLM fitting settings for these data were optimized using the ground truth data included in the ThunderSTORM plugin [16]. Ground truth data is a computer-generated dataset with known x ,y, z positing which mimics the biological data, which has unknown positions. To optimize the subpixel

fitting conditions, the ground truth data is analyzed using a set of SMLM analysis parameters as described in the method and the goodness of fit determined. The goodness of fit can be described as the Jaccard index which, for SMLM data, is true-positive fits/false-positive fits + false-negative fits. Perfect fitting conditions give a Jaccard index of 1. Here a Jaccard index of 0.9 was obtained [18].

11. Fab2 antibodies are typically labeled with 3.5 dye molecules per antibody. Each dye molecule will photoswitch stochastically, and it is assumed that neighboring dye molecules do not alter this stochastic behavior in these analyses. Therefore, photoswitches within 2 nm are assumed to arise from different dye molecules on the same antibodies and are binned together for the purpose of localization of an individual protein molecule.

References

1. Galbraith CG, Galbraith JA (2011) Super-resolution microscopy at a glance. J Cell Sci 124(Pt 10):1607–1611. https://doi.org/10.1242/jcs.080085. 124/10/1607 [pii]
2. Bates M, Huang B, Dempsey GT, Zhuang X (2007) Multicolor super-resolution imaging with photo-switchable fluorescent probes. Science 317(5845):1749–1753. https://doi.org/10.1126/science.1146598. 1146598 [pii]
3. Huang B, Wang W, Bates M, Zhuang X (2008) Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. Science 319(5864):810–813. https://doi.org/10.1126/science.1153529. 1153529 [pii]
4. Betzig E, Patterson GH, Sougrat R, Lindwasser OW, Olenych S, Bonifacino JS, Davidson MW, Lippincott-Schwartz J, Hess HF (2006) Imaging intracellular fluorescent proteins at nanometer resolution. Science 313(5793):1642–1645. https://doi.org/10.1126/science.1127344. 1127344 [pii]
5. Rust MJ, Bates M, Zhuang X (2006) Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Methods 3(10):793–795. https://doi.org/10.1038/nmeth929. nmeth929 [pii]
6. van de Linde S, Löschberger A, Klein T, Heidbreder M, Wolter S, Heilemann M, Sauer M (2011) Direct stochastic optical reconstruction microscopy with standard fluorescent probes. Nat Protoc 6(7):991–1009. https://doi.org/10.1038/nprot.2011.336. nprot.2011.336 [pii]
7. Cella Zanacchi F, Lavagnino Z, Perrone Donnorso M, Del Bue A, Furia L, Faretta M, Diaspro A (2011) Live-cell 3D super-resolution imaging in thick biological samples. Nat Methods 8(12):1047–1049. https://doi.org/10.1038/nmeth.1744
8. Hosny NA, Song M, Connelly JT, Ameer-Beg S, Knight MM, Wheeler AP (2013) Super-resolution imaging strategies for cell biologists using a spinning disk microscope. PLoS One 8(10):e74604. https://doi.org/10.1371/journal.pone.0074604
9. de Rooij J, Kerstens A, Danuser G, Schwartz MA, Waterman-Storer CM (2005) Integrin-dependent actomyosin contraction regulates epithelial cell scattering. J Cell Biol 171(1):153–164. https://doi.org/10.1083/jcb.200506152. jcb.200506152 [pii]
10. Leong HS, Robertson AE, Stoletov K, Leith SJ, Chin CA, Chien AE, Hague MN, Ablack A, Carmine-Simmen K, McPherson VA, Postenka CO, Turley EA, Courtneidge SA, Chambers AF, Lewis JD (2014) Invadopodia are required for cancer cell extravasation and are a therapeutic target for metastasis. Cell Rep 8(5):1558–1570. https://doi.org/10.1016/j.celrep.2014.07.050
11. Weaver AM (2006) Invadopodia: specialized cell structures for cancer invasion. Clin Exp Metastasis 23(2):97–105. https://doi.org/10.1007/s10585-006-9014-1
12. Destaing O, Block MR, Planus E, Albiges-Rizo C (2011) Invadosome regulation by adhesion signaling. Curr Opin

Cell Biol 23(5):597–606. https://doi.org/10.1016/j.ceb.2011.04.002. S0955-0674(11)00050-0 [pii]

13. McCutchen CW (1967) Superresolution in microscopy and the Abbe resolution limit. J Opt Soc Am 57(10):1190–1192
14. Demmerle J, Innocent C, North AJ, Ball G, Müller M, Miron E, Matsuda A, Dobbie IM, Markaki Y, Schermelleh L (2017) Strategic and practical guidelines for successful structured illumination microscopy. Nat Protoc 12(5):988–1010. https://doi.org/10.1038/nprot.2017.019
15. Olivier N, Keller D, Rajan VS, Gönczy P, Manley S (2013) Simple buffers for 3D STORM microscopy. Biomed Opt Express 4(6):885–899. https://doi.org/10.1364/BOE.4.000885
16. Levet F, Hosy E, Kechkar A, Butler C, Beghin A, Choquet D, Sibarita JB (2015) SR-Tesseler: a method to segment and quantify localization-based super-resolution microscopy data. Nat Methods 12(11):1065–1071. https://doi.org/10.1038/nmeth.3579
17. Thomas GJ, Hart IR, Speight PM, Marshall JF (2002) Binding of TGF-beta1 latency-associated peptide (LAP) to alpha(v)beta6 integrin modulates behaviour of squamous carcinoma cells. Br J Cancer 87(8):859–867. https://doi.org/10.1038/sj.bjc.6600545
18. Ovesný M, Křížek P, Borkovec J, Svindrych Z, Hagen GM (2014) ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. Bioinformatics 30(16):2389–2390. https://doi.org/10.1093/bioinformatics/btu202
19. Ball G, Demmerle J, Kaufmann R, Davis I, Dobbie IM, Schermelleh L (2015) SIMcheck: a toolbox for successful super-resolution structured illumination microscopy. Sci Rep 5:15915. https://doi.org/10.1038/srep15915
20. Royer LA, Weigert M, Günther U, Maghelli N, Jug F, Sbalzarini IF, Myers EW (2015) ClearVolume: open-source live 3D visualization for light-sheet microscopy. Nat Methods 12(6):480–481. https://doi.org/10.1038/nmeth.3372

Chapter 16

Observing the Assembly of Protein Complexes in Living Eukaryotic Cells in Super-Resolution Using refSOFI

Fabian Hertel, Gary C. H. Mo, Peter Dedecker, and Jin Zhang

Abstract

Few approaches are currently available that allow the detection of protein-protein interactions (PPIs) in super-resolution, and the observation of the assembly of protein complexes in living cells has been particularly challenging. We developed reconstituted fluorescence-based stochastic optical fluctuation imaging (refSOFI), which is based on bimolecular fluorescence complementation (BiFC) and SOFI, allowing us to detect protein complex assembly 30 min after the induction of complex formation. Here we describe how to use refSOFI to map the assembly of two proteins of interest into a complex within living cells at super-resolution.

Key words Super-resolution imaging, Stochastic optical fluctuation imaging (SOFI), Protein complexes, Protein-protein interactions, Bimolecular fluorescence complementation (BiFC)

1 Introduction

In the last decade, tremendous progress has been made in the field of super-resolution imaging, allowing us to characterize cellular structures below the diffraction limit, thus having a vast impact on research in several biological fields [1]. Several strategies have been developed to overcome the diffraction limit, which are either based on sophisticated illumination strategies, such as stimulated emission depletion (STED) [2] and saturated structured illumination microscopy (SSIM) [3], the localization of single molecules represented by photoactivated localization microscopy (PALM) [4], and stochastic optical reconstruction microscopy (STORM) [5], or higher-order correlation analyses of fluorescence intensity fluctuations in the case of stochastic optical fluctuation imaging (SOFI) [6, 7]. While some of these methods permitted a resolution of up to 20 nm in living cells under optimal conditions, the number of methods that could directly visualize the interaction of proteins beyond simple co-localization is rather limited. Recently, several methods that combine the bimolecular fluorescence complementa-

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_16,

tion (BiFC) of fluorescent proteins [8] with a super-resolution strategy have been introduced [9–12]. The approach we developed is using SOFI with BiFC-compatible fragments of Venus [13] or Dronpa MVF (DMVF), a variant of the photoswitchable fluorescent protein Dronpa [14] that we specifically optimized for this task. We termed this strategy reconstituted fluorescence-based SOFI (refSOFI) [12] and demonstrated that it is specifically advantageous to map protein interactions in living cells and to observe the assembly of protein complexes within 30 min when using Venus. We could show that in the case of subunits of the pore forming Ca^{2+} release-activated Ca^{2+} (CRAC) channel ORAI1 and stromal interaction molecule 1 (STIM1) at endoplasmic reticulum (ER)-plasma membrane (PM) junctions, super-resolution information is critical to assess the characteristics of clusters of STIM1/ORAI1 complexes. Here we describe how to apply refSOFI to detect protein complexes and to observe complex formation in living cells in a general way. In brief, first the proteins of interest are fused to BiFC-compatible fragments of Venus or DMVF and transiently expressed in the cells of interest. Subsequently, image series of the fluorescence intensity fluctuations are recorded, the cells can be stimulated if desired, and the data is analyzed with the software package Localizer [15], yielding super-resolution images of the complexes formed by the proteins of interest.

2 Materials

2.1 Plasmids and Molecular Cloning

1. Templates for molecular cloning: pcDNA3 [Lyn-FRB-DMVF-181-N], pcDNA3 [FKBP-DMVF-181C], pcDNA3 [Lyn-FRB-Venus-173-N], and pcDNA3 [FKBP-Venus-173-C] from our study [12].
2. Plasmids containing cDNA of the proteins of interest.
3. Restriction enzymes NheI, HindIII, BamHI, KpnI, and SpeI.
4. TAE buffer: 40 mM Tris, 20 mM acetate, and 1 mM EDTA.
5. Agarose.
6. Agarose gel DNA extraction kit.
7. Phusion High-Fidelity DNA Polymerase.
8. T4 DNA Ligase.
9. LB media: 1% (w/v) bacto-tryptone, 0,5% (w/v) yeast extract, and 1% (w/v) NaCl, autoclaved.
10. Ampicillin.
11. SOC media: 2% (w/v) bacto-tryptone, 0.5% (w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl, $MgCl_2$ 10 mM, and $MgSO_4$ 10 mM, autoclaved; 20 mM glucose has to be added after autoclaving using a sterile filtered stock solution.

12. Aliquots of chemically competent *E. coli* bacteria (DH5α strain).
13. LB agar plates with ampicillin (100 μg/mL): LB with 1.5% (w/v) agar, autoclaved, ampicillin added after cooling down to 60 °C from stock solution, poured in sterile plates.
14. Plasmid DNA miniprep kit.

2.2 Cell Cultivation

1. Growth media specific to the cell line of interest.
2. Lipofectamine 2000 (Thermo Fisher Scientific).
3. Opti-MEM® I Reduced Serum Medium (Thermo Fisher Scientific).
4. Hank's balanced salt solution (HBSS).
5. Imaging dishes (35 mm) with glass bottom.

2.3 SOFI Imaging

1. Download the software package Localizer [15], which also requires either the software Igor Pro (Wavemetrics) or MATLAB (MathWorks). Follow the included instructions to integrate the software.
2. Analysis computer with sufficient CPU power and memory to handle imaging files up to 1 GB.
3. Microscope with TIRF capability and appropriate filters to excite and detect Dronpa (excitation peak at 503 nm and emission peak at 518 nm) and Venus (excitation peak at 515 nm and emission peak at 528 nm), optical components that result in a camera image exhibiting a pixel size close to or below the diffraction limit, and a 488 nm laser (for DMVF or Venus) and/or a 514 nm laser (for Venus) with at least 25 mW output (*see* **Note 1**). Furthermore, the hardware and software should allow recording image sequences with an exposure time of 35 ms or less.
4. Imaging buffer: Hank's balanced salt solution (HBSS) with 2.0 g/L D-glucose, adjust pH to 7.4 with NaOH, and filter-sterilized using 0.22 μm filters.

3 Methods

3.1 Generating the Constructs

The protocol below describes a typical method to obtain expression plasmids containing c-terminal fusion constructs of the proteins of interest with the BiFC fragments of DMVF or Venus based on the constructs we generated. Accordingly, several alternative molecular cloning strategies can be used to combine the fragments with the proteins of interest, which are not covered here.

1. The original plasmids are derived from pcDNA3 and follow a certain cloning scheme; therefore the cDNA of the proteins of

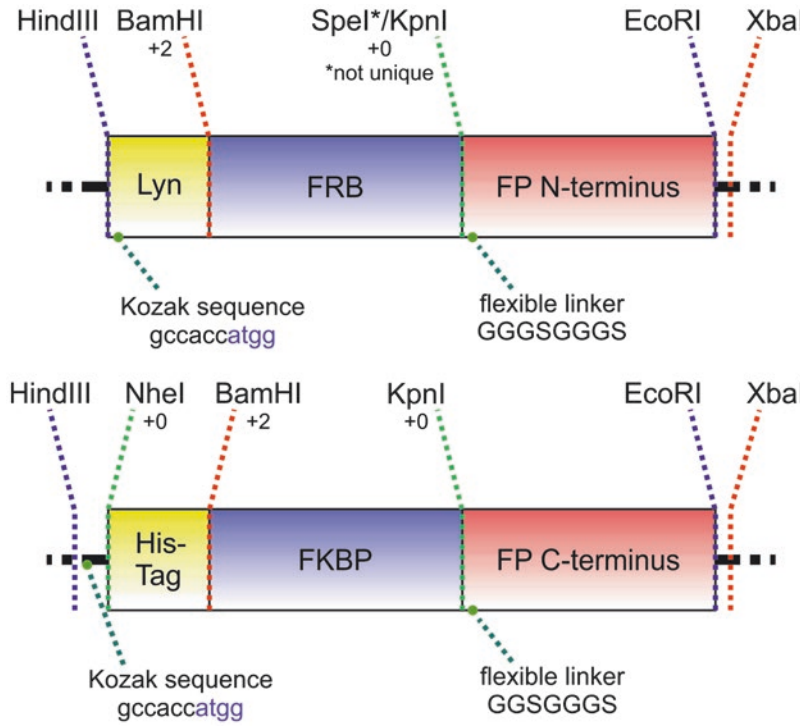


Fig. 1 Cloning scheme of the constructs in the template plasmids. All internal restriction sites except for BamHI are located in frame. The cDNA of FRB and FKBP and eventually the Lyn targeting sequence or the His-tag have to be replaced with the cDNAs of the proteins of interest

interest has to be furnished with the appropriate restriction sites in order to insert them. The plasmids contain the BiFC-suitable fragments of Venus or DMVF (*see* **Note 2**) but are constructed similarly otherwise. The plasmids possess restriction sites for NheI, HindIII, BamHI, SpeI, KpnI, EcoRI, and XbaI (5′–3′) and contain other designed features, illustrated in Fig. 1.

2. Design PCR primers for the cDNAs of your proteins of interest and add the suitable restriction sites so that it can be inserted into the original constructs (most likely HindIII and KpnI). Make sure that these restriction sites are not present in the original cDNA (*see* **Note 3**). Furthermore, add at least three extra bases to the 5′ end of the primer to ensure efficient restriction enzyme digestion and make sure that the cDNA is in frame. Also, add a Kozak sequence including a start codon when starting from the restriction site HindIII. The SpeI site is unique in the construct but not in the plasmid, so it can be utilized using alternative cloning strategies (*see* **Note 4**).
3. Prepare the samples for PCR reactions, for example, by mixing 33 μL water, 10 μL 5× Phusion HF Buffer, 1 μL 10 mM

dNTPs, 2.5 μL of 10 μM forward and reverse primer, and 0.5 μL template DNA and then adding 0.5 μL Phusion DNA Polymerase at the end. Run the PCR reactions in a thermocycler. The following is a sample protocol: preheat the block to 98 °C; 98 °C for 30 s; repeat that in bracket for 25 times: [98 °C for 10 s, 60 °C for 15 s, and 72 °C for 15 s per kb; 72 °C for 10 min]; hold temperature at 4 °C.

4. Add 6 μL 10× restriction enzyme buffer and 2 μL of the enzymes corresponding to the primers to each PCR product and incubate at 37 °C for 30 min.
5. Mix 1 μg of the original plasmids containing the fragments of the fluorescent proteins, 2 μL 10× restriction enzyme buffer, 1 μL of each enzyme corresponding to the PCRs, and water for a total volume of 20 μL. Incubate the reaction at 37 °C for 30 min.
6. Prepare a 1% agarose gel with TAE buffer, load the gel with the digested PCR products and original plasmids, and run an agarose gel electrophoresis at 110 V for 30 min.
7. Cut the DNA bands of the backbones of the original plasmids and the PCR products out of the gel and use an agarose gel DNA extraction kit to obtain the DNA.
8. Mix 2 μL of the backbone DNA solution, 13 μL of the PCR product, 4 μL of the 5× ligase buffer, and 1 μL T4 DNA ligase, and incubate at room temperature for 15 min.
9. Transform 5 μL of the ligation reactions into chemically competent DH5α bacteria using a common heat-shock protocol, spread the bacteria on LB agar plates with ampicillin, and incubate the plate over night at 37 °C.
10. Pick several single colonies from the plate to inoculate bacterial cultures (2 mL LB media with ampicillin). Incubate the bacterial cultures at 37 °C overnight.
11. Use a plasmid DNA miniprep kit to obtain the plasmids from the bacterial cultures. Verify that the plasmids contain the fusion construct of the cDNA of the proteins of interest and the fluorescent protein fragments, either by sequencing or by performing a test digest with the appropriate restriction enzymes.

3.2 Transfection of Cells

1. Plate cells of the cell line of interest in imaging dishes 1 day prior to the transfection. Select the number of cells to seed so that the confluency reaches 75–90% on the following day.
2. For each imaging dish, mix 50 μL Opti-MEM® with 0.5 μg of each construct and 50 μL Opti-MEM® with 1.5 μL Lipofectamine 2000. Add the DNA solution dropwise to the

Lipofectamine mix, and incubate for 15 min at room temperature (*see* **Note 5**). It can be useful to add an additional construct with a different colored FP serving as a marker for co-imaging and to facilitate imaging in case the interaction is constitutively weak.

3. Add the mix dropwise to the imaging dish, and keep it in an incubator under the standard conditions for the employed cell line for at least 24 h.

3.3 Imaging

1. Remove the growth media from the imaging dish, and wash the cells twice with HBSS before adding the imaging buffer (HBSS with glucose).
2. Perform the imaging of the dish on your microscopy setup at the appropriate temperature. Fix the dish to the microscope stage to prevent any movement.
3. Set the microscope up for imaging DMVF or Venus and any other appropriate markers (*see* **Note 6**).
4. Switch the imaging software to the "live" mode to directly observe the fluorescence intensity as it would be recorded in subsequent experiments. Focus and adjust the TIRF angle appropriately (*see* **Note 7**) to ensure a high intensity of the excitation light exclusively close to the surface of the imaging dish. Select an exposure time of 35 ms or less and adjust the laser power and EM gain so that the fluorescence intensity values make proper use of the camera dynamic range.
5. For reconstituted DMVF: Increase the laser power until you clearly observe visible fluorescence intensity fluctuations. Higher laser power improves the fluctuations behavior but increases photobleaching (permanent photodamage); therefore, the laser power has to be carefully chosen to allow imaging over 10–30 s without significant signal reduction due to photobleaching.
6. For reconstituted Venus: In our experience, the population of fluorescent Venus often has to be decreased (via photobleaching) in order to obtain suitable single-molecule fluorescence intensity fluctuations. If the cellular fluorescence is bright, saturated, and intensity fluctuations could not be observed, increase the 514 nm laser power to photobleach a fraction of the Venus molecules. This may only take several seconds, and single-molecule fluctuations should come into view. After the average fluorescence intensity decreased by approximately half, reduce the laser power and check for visible single-molecule fluorescence intensity fluctuations again. Repeat if necessary.
7. Acquire an image series. To obtain a second order refSOFI image (up to double resolution), record at least 200 images.

For a third order refSOFI image (up to triple resolution), record at least 800 images.

8. Use the imaging software to save your image sequence as a multipage TIFF file (a single TIFF file containing all of the images in the sequence).
9. If the interaction of the proteins of interest can be induced, record an image series before induction and multiple image series after induction. For Venus, the formation of clusters of protein complexes can be monitored 30 min after they occur, while for DMVF, the cells should be incubated for at least 3 h to allow reconstitution and maturation.
10. If an additional marker is used, also take images of it shortly before or after the image sequence for refSOFI is recorded. A marker that contains a photoswitchable FP such as rsTagRFP can also be used for SOFI, so that a super-resolution image of the protein complexes and the marker can be obtained from the same cell (e.g., in Fig. 2).

3.4 Image Data Analysis

Here we describe how we analyze the image data using Localizer and the graphical user interface (GUI) using the software Igor Pro. Localizer is also available as a subroutine for MATLAB and can be used analogously, though the GUI is specific to the Igor Pro version. *See* Fig. 3 for an illustration of the principles of refSOFI.

1. Open "Igor Pro" and select the menu "Localizer\Read CCD data\Read CCD data from disk…." Use the dialog to select a TIFF file containing an image sequence.

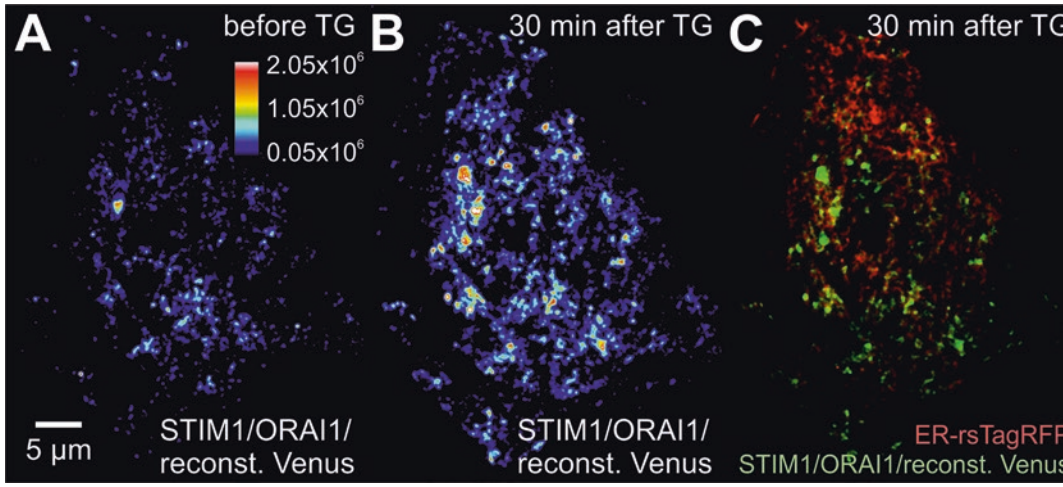


Fig. 2 Using refSOFI to investigate the assembly of STIM1 and ORAI1. (**a**) Representative refSOFI image of an unstimulated HeLa cell expressing STIM1-VC, VN-ORAI1, and ER-localized rsTagRFP (not shown). (**b**) Cell shown in (**a**) 30 min after treatment with 1 μM thapsigargin (TG). (**c**) Multicolor image showing the STIM1/ORAI1 complexes (green) and the ER-localized rsTagRFP (red) after the treatment

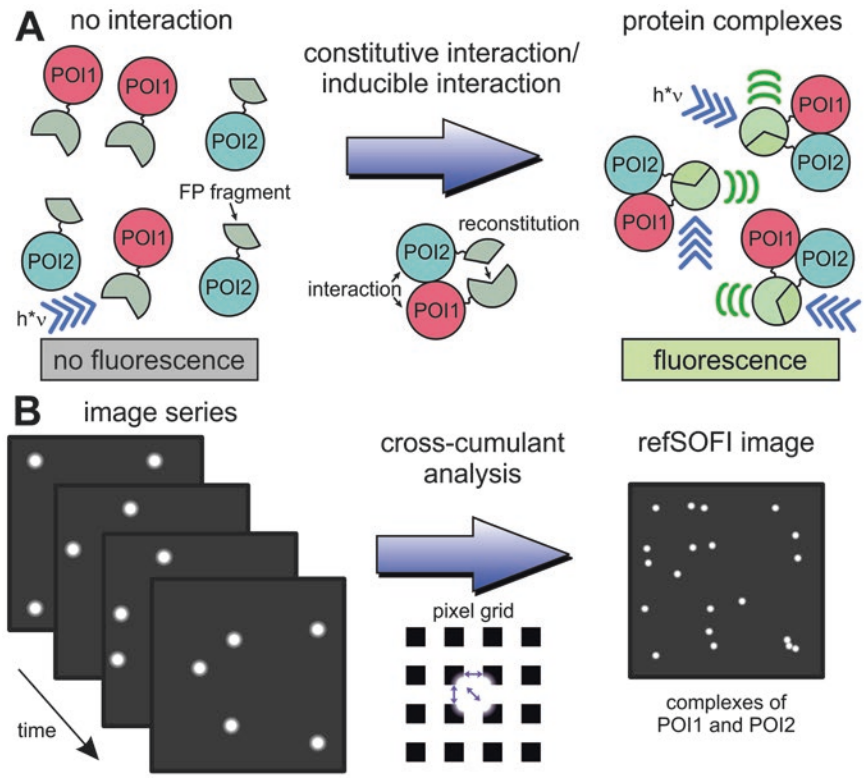


Fig. 3 Schematic representation of the principles important for refSOFI. (**a**) Two proteins of interest that are fused to fragments of DMVF or Venus are expressed in cells. Accordingly, there is no fluorescence when excited with 488 nm (DMVF) or 514 nm (Venus) when the proteins are not interacting. If interactions take place and the complementary fragments get into nanometer proximity, the fluorescent protein is reconstituted and provides fluorescence intensity fluctuations when excited. (**b**) An image sequence is acquired, and the software package Localizer is used to carry our cross-cumulant analysis, generating the refSOFI image with an increased resolution

2. Click on the button ">>" to open the localizer GUI. Select the tab "SOFI."
3. Choose the desired cumulant Order (second or third) and the number of pixel combinations. Subsequently calculate the combination weights by clicking on the button "Calc weights." The weights are then stored in an Igor Pro data structure (wave) and can be used for the same order and to analyze image sequences that were recorded under the same conditions. Select the wave and the combination weights.
4. Choose a range of frames within the image sequence that will be used to calculate a SOFI image. It is important to exclude frames at the beginning that exhibit strong photobleaching and frames at the end that barely show a signal. For that purpose, the average intensity trace of the image sequence can be calculated to aid with estimating an appropriate range. Select

the tab "Other," pick "Average Intensity Trace" from the drop-down list in the "Basis analysis" area, and click the button "Do it" in order to obtain the trace showing the average intensity with respect to the frame number. Furthermore, the image sequence has to be inspected for sections in which cells that are visibly moving by scrolling through it, and these sections have to be excluded as well (*see* **Note 8**).

5. Calculate the SOFI image by clicking the button "Do it."
6. The GUI also offers to apply the built-in "Richardson-Lucy Deconvolution" to the SOFI image. In order to do that, select the standard deviation of the point spread function (PSF) in pixel, choose the number of iterations, and click on the button "Do it" in the right bottom corner.
7. To save the results, the corresponding waves (Igor datasets) M_SOFI (SOFI image), M_DeconvolvedSOFI (deconvolved SOFI image), and M_SOFI_avg (average image of the image sequence) can be saved as waves using the menu "Data\Save Waves\Save Igor Binary…" or as a tiff file using "Data\Save Waves\Save Image…." A window opens, and the particular waves can be found in the folder "Packages\Localizer\LocalizerViewerX" where X represents the a counter that increments for every opened image stack. The command can then be executed directly ("Do It") or copied to the command line ("To Cmd Line"). Alternatively, one can make use of the ImageSave command (see appropriate documentation in Igor).

4 Notes

1. The required laser power is affected by a variety of microscope- and sample-dependent factors, such as camera sensitivity, optical efficiency, and protein expression. In order to examine the adequate excitation power required for a given system, cells expressing membrane targeted Dronpa or Venus (full fluorescent proteins) can be used. To avoid excessive photobleaching, follow the instruction in Subheading 3.3, **step 9**, and compare the decrease in signal between two consecutive SOFI images.
2. The fragments of the two fluorescent proteins DMVF and Venus that we utilized have different advantages. The fragments of DMVF have been optimized for BiFC assays with the reconstituted DMVF retaining the photophysical properties of Dronpa. It can be used to image complex assemblies that are formed by either constitutive or inducible but long-lasting interactions between proteins. The Venus fragments allow to detect complex formation within 30 min; therefore, it is most suitable to observe acute signal-induced protein assembly.

3. If the important restriction sites are also present in the cDNA of the protein of interest, restriction enzymes with compatible cohesive ends can be used to extend the PCR product. For example, BglII can be used to ligate to BamHI and XbaI to ligate to NheI.
4. If an alternative cloning strategy is used, ensure a flexible linker (e.g., GGGSGGGS) is used between the fragments and the proteins of interest.
5. The transfection protocol was optimized for the systems we have described. Different inserts may require optimization by testing several DNA/Lipofectamine ratios. Read the manufacturer's protocol for the most updated information.
6. The FKBP/FRB model system can be readily utilized to establish a refSOFI protocol on new equipment. The constructs we described can be transfected into HeLa cells; the interaction can be induced using 100 nM rapamycin.
7. While TIR condition is not necessary for refSOFI, the observation of single-molecule fluctuation benefits from substantial z-axis rejection, which lends directly to high signal-to-noise ratio in refSOFI and high image contrast. It is worthwhile to optimize this factor for better results.
8. For live-cell imaging, the effect of probe diffusion could be of concern. However, the accuracy of SOFI imaging under these conditions has been examined, and it has been suggested that diffusion has an overall positive effect on the imaging [16].

Acknowledgment

This work was supported by NIH (R35 CA197622, R01 DK073368, R01 GM111665, and R01 MH111516 to J.Z.).

References

1. Han R, Li Z, Fan Y, Jiang Y (2013) Recent advances in super-resolution fluorescence imaging and its applications in biology. J Genet Genomics 40:583–595. https://doi.org/10.1016/j.jgg.2013.11.003
2. Hell SW (2007) Far-field optical nanoscopy. Science 316:1153–1158. https://doi.org/10.1126/science.1137395
3. Gustafsson MGL (2005) Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution. Proc Natl Acad Sci U S A 102:13081–13086. https://doi.org/10.1073/pnas.0406877102
4. Betzig E, Patterson GH, Sougrat R et al (2006) Imaging intracellular fluorescent proteins at nanometer resolution. Science 313:1642–1645. https://doi.org/10.1126/science.1127344
5. Rust MJ, Bates M, Zhuang X (2006) Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Methods 3:793–795. https://doi.org/10.1038/nmeth929
6. Dertinger T, Colyer R, Vogel R et al (2010) Achieving increased resolution and more pixels with superresolution optical fluctuation

imaging (SOFI). Opt Express 18:18875–18885. https://doi.org/10.1364/OE.18.018875

7. Dedecker P, Mo GCH, Dertinger T, Zhang J (2012) Widely accessible method for superresolution fluorescence imaging of living systems. Proc Natl Acad Sci U S A 109:10909–10914. https://doi.org/10.1073/pnas.1204917109
8. Kodama Y, C-D H (2012) Bimolecular fluorescence complementation (BiFC): a 5-year update and future perspectives. BioTechniques 53:285–298. https://doi.org/10.2144/000113943
9. Liu Z, Xing D, QP S et al (2014) Super-resolution imaging and tracking of protein-protein interactions in sub-diffraction cellular space. Nat Commun 5:4443. https://doi.org/10.1038/ncomms5443
10. Nickerson A, Huang T, Lin L-J, Nan X (2014) Photoactivated localization microscopy with bimolecular fluorescence complementation (BiFC-PALM) for Nanoscale imaging of protein-protein interactions in cells. PLoS One 9:e100589. https://doi.org/10.1371/journal.pone.0100589
11. Xia P, Liu X, Wu B et al (2014) Superresolution imaging reveals structural features of EB1 in microtubule plus-end tracking. Mol Biol Cell 25:4166–4173. https://doi.org/10.1091/mbc.E14-06-1133
12. Hertel F, Mo GCH, Duwé S et al (2016) RefSOFI for mapping nanoscale organization of protein-protein interactions in living cells. Cell Rep 14:390–400. https://doi.org/10.1016/j.celrep.2015.12.036
13. Nagai T, Ibata K, Park ES et al (2002) A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. Nat Biotechnol 20:87–90. https://doi.org/10.1038/nbt0102-87
14. Ando R, Mizuno H, Miyawaki A (2004) Regulated fast nucleocytoplasmic shuttling observed by reversible protein highlighting. Science 306:1370–1373. https://doi.org/10.1126/science.1102506
15. Dedecker P, Duwé S, Neely R, Zhang J (2012) Localizer: fast, accurate, open-source, and modular software package for superresolution microscopy. J Biomed Opt 17:126008. https://doi.org/10.1117/1.JBO.17.12.126008
16. Vandenberg W, Dedecker P (2017) Effect of probe diffusion on the SOFI imaging accuracy. Sci Rep 7:44665. https://doi.org/10.1038/srep44665

Chapter 17

Detecting Purinosome Metabolon Formation with Fluorescence Microscopy

Anthony M. Pedley and Stephen J. Benkovic

Abstract

A long-standing hypothesis in the de novo purine biosynthetic pathway is that there must be highly coordinated processes to allow for enhanced metabolic flux when a cell demands purines. One mechanism by which the pathway meets its cellular demand is through the spatial organization of pathway enzymes into multienzyme complexes called purinosomes. Cellular conditions known to impact the activity of enzymes in the pathway or overall pathway flux have been reflected in a change in the number of purinosome-positive cells or the density of purinosomes in a given cell. The following general protocols outline the steps needed for purinosome detection through transient expression of fluorescent protein chimeras or through immunofluorescence in purine-depleted HeLa cells using confocal laser scanning microscopy. These protocols define a purinosome as a colocalization of FGAMS with one additional pathway enzyme, such as PPAT or GART, and provide insights into the proper identification of a purinosome from other reported cellular bodies.

Key words Purinosome, Metabolon, Purine metabolism, De novo purine biosynthesis, Fluorescence microscopy

1 Introduction

Our current understanding of enzymes can be credited to the tools and techniques of traditional in vitro enzymology. However, the removal of an enzyme from a cellular environment has largely downplayed those regulatory events that might contribute to the innate activity or behavior of an enzyme. These factors could include posttranslational modifications, ancillary protein-mediated allosteric modulation, and spatial organization. Therefore, the generation of intracellular reporters has provided a means to better understand how an enzyme functions within a cell and has brought to light the era of in-cell enzymology.

One way in-cell enzymology has reshaped our knowledge of enzymes is through the spatial organization of sequential metabolic pathway enzymes into supramolecular clusters called

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764,
https://doi.org/10.1007/978-1-4939-7759-8_17,

metabolons [1]. Since the initial observation of metabolon formation among enzymes in the tricarboxylic acid cycle [2], metabolons have been observed in glycolysis [3, 4], amino acid biosynthesis [5], and the de novo biosynthesis of purines and pyrimidines [6, 7]. Several of these metabolons were hypothesized for decades, but traditional in vitro techniques did not provide compelling evidence for their existence. Ultimately, the translation of commonly employed fluorescence microscopy techniques rapidly developed a tool belt in which one can effectively study these metabolons [8].

Here, we outline a method for visualizing a metabolon comprising all six enzymes within the de novo purine biosynthetic pathway by confocal laser scanning microscopy. The spatial organization of these enzymes in cells is referred to as a purinosome and has been the subject of recent reviews [9, 10]. Purinosome assembly has shown to be a reversible phenomenon whose phenotype is largely predominant when cellular conditions result in a high purine demand, such as in the G_1-phase of the cell cycle [6, 11]. Cellular conditions favoring purinosome formation were also shown to enhance the metabolic flux of the de novo purine biosynthetic pathway suggesting that the two observations are connected—a generalized hypothesis surrounding metabolon formation in cells [12]. Further characterization of purinosomes has unveiled a high degree of colocalization with cytoskeletal elements [13] and mitochondria [14] as well as interactions with molecular chaperones [15]. The interplay between all these different cellular elements has presented the purinosome as a highly regulated complex, whose composition and cellular localization have started to provide a fresh and more comprehensive perspective on the regulation of purine metabolism otherwise not readily recognized by more traditional means.

2 Materials

The original discovery and characterization of purinosomes were performed in the HeLa CCL-2 cervical carcinoma cell line under purine-depleted growth conditions [6]. Since then, purinosomes have been observed under similar growth conditions in human hepatocarcinoma liver cell line HepG2 [16] and its derivative C3A [15], sarcoma osteogenic cell line Saos-2 [16], human embryonic kidney cell line HEK293 [16], human skin cancer cell line A431 [15], human breast cancer cell line HTB-126 [6], primary human keratinocytes [16], and primary human dermal fibroblasts [11, 17]. The diversity in cell types bearing purinosomes, observed by transient expression of fluorescent chimeras of enzymes and/or immunofluorescence, illustrates that purinosome formation is a generalized phenomenon to likely denote elevated pathway usage and not limited to one cell type or genetic background.

The following list of materials has been validated for detecting purinosomes in purine-depleted HeLa CCL-2 cervical carcinoma cells. For these methods, a purinosome is defined as the colocalization of FGAMS (also referred to as PFAS) with one additional pathway enzyme (PPAT or GART) (*see* **Note 1**). Other combinations of plasmids and antibodies can be used to define a purinosome; however, at least two pathway enzymes should be imaged concurrently. If not, differentiating purinosomes from other cellular bodies, such as a recently discovered inhibitory FGAMS enzyme cluster, void of other pathway enzymes, might not be possible [18].

2.1 Materials for Cell Culture and Imaging

1. General mammalian cell culture disposables and instrumentation.
2. HeLa CCL-2 cervical carcinoma cell line (American Type Culture Collection).
3. 35 mm glass bottom tissue culture-treated culture dish.
4. 1× Dulbecco's phosphate buffered saline (without calcium and magnesium) solution (D-PBS).
5. 0.25% Trypsin with 2.21 mM ethylenediaminetetraacetic acid (EDTA).
6. Purine-depleted complete growth medium: Roswell Park Memorial Institute (RPMI) 1640 supplemented with 300 mg/L L-glutamine and 10% (v/v) dialyzed fetal bovine serum (FBS) (*see* **Note 2**).
7. Olympus Fluoview 1000 confocal laser scanning microscope equipped with appropriate lasers and filters for the selected fluorescent dyes and proteins.
8. ImageJ image analysis and visualization software [19].

2.2 Detection of Purinosomes in Living Cells Using Transient Expression of Fluorescently Labeled Protein Chimeras

1. Gibco™ Opti-MEM™ reduced serum medium or Eagle's Minimum Essential Medium (MEM) without fetal bovine serum.
2. Lipofectamine® 2000 transfection reagent.
3. Endotoxin-free plasmids encoding genes for FGAMS-EGFP and PPAT-mCherry (*see* **Note 3**).
4. Hank's Balanced Salt Solution (HBSS): 8.0 g/L sodium chloride, 400 mg/L potassium chloride, 140 mg/L calcium chloride, 1 g/L glucose, 60 mg/L potassium phosphate monobasic, 48 mg/L sodium phosphate dibasic anhydrous, 350 mg/L sodium bicarbonate, 98 mg/L magnesium sulfate anhydrous.
5. Optional: Hoechst 33342 (2′-[4-ethoxyphenyl]-5-[4-methyl-1-piperazinyl]-2,5′-bi-1H-benzimidazole trihydrochloride trihydrate) counterstain. Prepare a 1 μg/mL solution diluted in complete growth medium (*see* **Note 4**).

2.3 Immunofluorescence Detection of Endogenous Purinosomes in Fixed HeLa Cells

1. Fixative solution: 4% (v/v) electron microscopy grade paraformaldehyde in 1× D-PBS.
2. Permeabilization solution: 0.1% (v/v) Triton X-100 in 1× D-PBS.
3. Wash buffer (PBST): 0.1% (v/v) Tween-20 in 1× D-PBS.
4. Blocking buffer: 5% (v/v) normal donkey serum (serum of secondary antibody host) in PBST.
5. Primary antibody solution: 1:500 dilution of PFAS rabbit polyclonal antibody (Bethyl Laboratories) and 1:1000 dilution of GART mouse monoclonal antibody (Novus Biologicals) prepared in blocking buffer.
6. Secondary antibody solution: 1:1000 dilution of CF488A-conjugated donkey anti-rabbit immunoglobulin and 1:1000 dilution of CF568-conjugated donkey anti-mouse immunoglobulin prepared in blocking buffer.
7. Optional: DAPI (4′,6-diamidino-2-phenylindole) counterstain: 1:1000 dilution of a 300 μM DAPI solution prepared in 1× D-PBS into the secondary antibody solution.

3 Methods

The following methods serve as a starting point for detecting purinosome formation in purine-depleted HeLa CCL-2 cervical carcinoma cells. Optimization of transfection or immunostaining experimental conditions may be warranted for best results. Experimental notes have been added to assist in areas where optimization is often suggested. Further tips on optimizing transfection efficiency and/or cell viability post-transfection can be found on manufacturer's websites.

3.1 Detection of Purinosomes in Living Cells Using Transient Expression of Fluorescently Labeled Protein Chimeras

1. The day before transfection, seed purine-depleted HeLa cells at 0.8–1.0 × 10^5 cells per 35 mm glass bottom tissue culture-treated dish. Incubate the cells overnight at 37 °C under 5% CO_2 in purine-depleted growth medium.
2. The next day, check the cell confluency under an inverted microscope. Optimal results (transfection efficiency and cell viability) are obtained when cells are approximately 70–80% confluent the day of transfection.
3. Add 2.0 μg of endotoxin-free pFGAMS-EGFP and 2.0 μg of endotoxin-free pPPAT-mCherry (4.0 μg total) plasmids to 50 μL of Opti-MEM™ reduced serum medium in a microcentrifuge tube. Pipet up and down to mix. Let sit for 5 min at room temperature.

4. Add 4 μL of 1 mg/mL Lipofectamine® 2000 solution to 50 μL Opti-MEM™ reduced serum medium in a microcentrifuge tube. Pipet up and down to mix. Let sit for 5 min at room temperature.
5. After 5 min incubation of both the plasmid and Lipofectamine® 2000 solutions, add the Lipofectamine® 2000 solution to the plasmid solution. Pipet up and down to mix (*see* **Note 5**).
6. Carefully remove growth medium and wash the cells once with 1× D-PBS.
7. Add 1 mL of Opti-MEM™ medium to the 35 mm glass bottom dish. Swirl to cover the entire bottom of the dish.
8. Add the lipid:DNA mixture (100 μL) to the cells dropwise. Gently swirl to mix.
9. Incubate the cells for 4 h at 37 °C (5% CO_2).
10. After 4 h, carefully remove all Opti-MEM™ medium from the culturing dish. Replenish cells with enough purine-depleted growth medium to cover the bottom of the dish (1–2 mL) (*see* **Note 6**).
11. Incubate the cells at 37 °C (5% CO_2) for an additional 16–18 h.
12. Remove the cells from the incubator and look for adherence under an inverted microscope. No significant cell death (>25%) should be observed.
13. Optional: Carefully remove growth medium and replace with a 1 μg/mL Hoechst 33342 solution prepared in purine-depleted complete growth medium. Incubate for 20 min at 37 °C (5% CO_2) for an effective nuclear counterstain.
14. Carefully remove the purine-depleted growth medium from the cells, and wash the cells once with HBSS.
15. Carefully aspirate the HBSS from the cells and replace with enough HBSS to cover the bottom of the dish (1–2 mL) (*see* **Note 7**).
16. Immediately image cells using a confocal laser scanning microscopy using filters appropriate for detecting EGFP and mCherry fluorescent proteins. If time-lapse imaging is desired, use a live cell imaging medium such as FluoroBrite DMEM medium instead of HBSS. Representative images are shown in Fig. 1 (*see* **Note 8** for further definition of the purinosome based on image analyses).

3.2 Immunofluorescence Detection of Endogenous Purinosomes in Fixed HeLa Cells

1. The day before fixation, seed purine-depleted HeLa cells at 6.0–8.0 × 10^4 cells per 35 mm glass bottom tissue culture-treated dish. Incubate the cells overnight at 37 °C under 5% CO_2 in purine-depleted growth medium.
2. The next day, verify cell adherence under an inverted microscope.

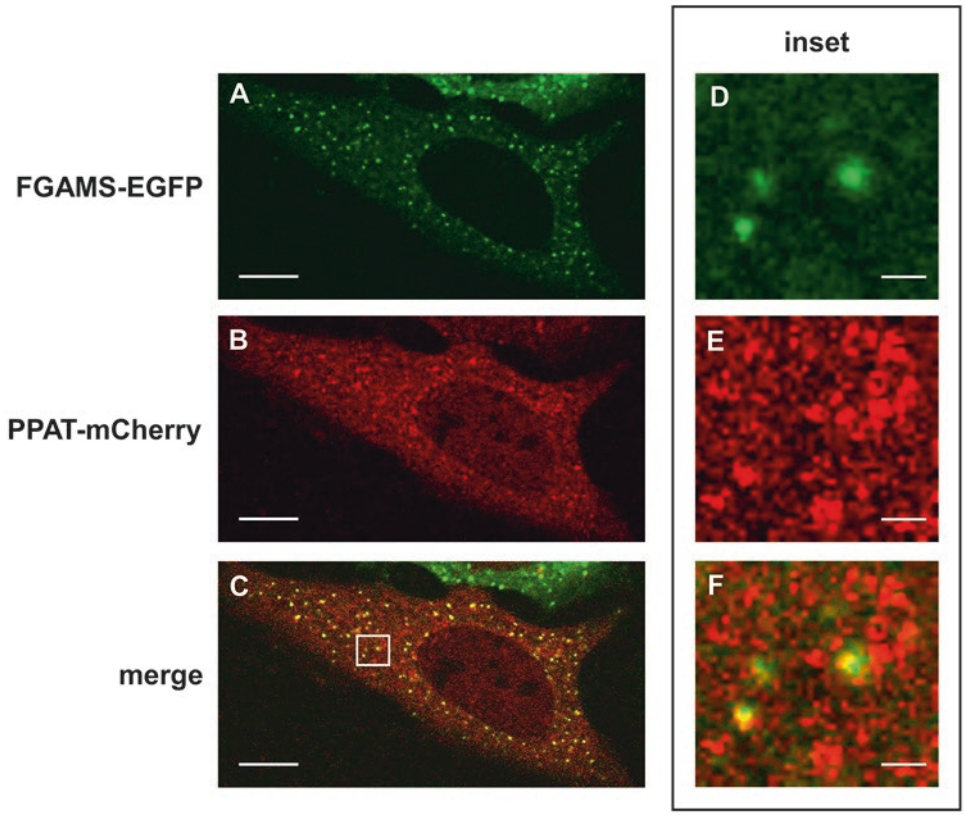


Fig. 1 Colocalization of transiently expressed FGAMS-EGFP and PPAT-mCherry to visualize purinosomes in purine-depleted HeLa cells. Purine-depleted HeLa cells were transiently transfected with plasmids encoding FGAMS-EGFP and PPAT-mCherry and allowed to express for 16 h prior to live cell imaging in HBSS using a 100× oil objective on an Olympus Fluoview 1000 confocal laser scanning microscope. Sequential imaging of EGFP and TRITC channels resulted in clustering of (**a**) FGAMS-EGFP (green) with (**b**) PPAT-mCherry (red), respectively. (**c**) Merging of the individual channels resulted in proper identification of purinosomes (yellow) as observed through the colocalization of FGAMS-EGFP with PPAT-mCherry. Inset shows an enlarged view of the individual EGFP (**d**), TRITC (**e**), and merged (**f**) channels. Scale bar: 10 μm (**a**–**c**) and 1 μm (**d**–**f**)

3. Carefully aspirate the medium away from the cells and wash the cells twice with enough 1× D-PBS to cover the bottom of the dish (1–2 mL).
4. Add 200 μL of fixative solution dropwise to the cells (*see* **Note 9**).
5. Incubate the samples in the fixative solution covered for 10 min at room temperature.
6. Carefully aspirate the fixative solution from the cells and wash the cells three times with enough 1× D-PBS to cover the bottom of the dish (1–2 mL). Each wash should last at least 5 min and be carried out on an orbital shaker at room temperature (*see* **Note 10**).
7. Add 200 μL of permeabilization solution dropwise to the cells.

8. Incubate the samples in the permeabilization solution for 10 min at room temperature on an orbital shaker.
9. Carefully aspirate the permeabilization solution from the cells, and wash the cells three times with enough 1× D-PBS to cover the bottom of the dish (1–2 mL). Each wash should last at least 5 min and be carried out on an orbital shaker at room temperature.
10. Block the cells with 200 μL of blocking buffer for 1 h at room temperature on an orbital shaker.
11. Carefully aspirate the blocking buffer, and add 200 μL primary antibody solution. For co-staining of FGAMS and GART, use a 1:500 dilution of PFAS rabbit polyclonal antibody and a 1:1000 dilution of GART mouse monoclonal antibody prepared in blocking buffer.
12. Incubate the samples in the primary antibody solution overnight at 4 °C on an orbital shaker. It is best practice to keep the samples covered to maintain dish humidity and minimize evaporation (*see* **Note 11**).
13. The next day, carefully aspirate the primary antibody solution from the cells and wash the cells four times with enough wash buffer (PBST) to cover the bottom of the dish (1–2 mL). Each wash should last at least 5 min and be carried out on an orbital shaker at room temperature.
14. Carefully aspirate the blocking buffer, and add 200 μL secondary antibody solution. For co-staining of FGAMS and GART, use a 1:1000 dilution of CF488A-conjugated donkey anti-rabbit IgG and a 1:1000 dilution of CF568-conjugated donkey anti-mouse IgG prepared in blocking buffer.
15. Optional: Add DAPI (300 nM final solution) to the same secondary antibody solution for an effective nuclear counterstain.
16. Incubate the samples in the secondary antibody solution for 2 h at room temperature on an orbital shaker. From this point on, all samples should be covered to prevent any photobleaching of the fluorescently labeled secondary antibodies.
17. Carefully aspirate the secondary antibody solution from the cells, and wash the cells four times with enough wash buffer (PBST) to cover the bottom of the dish (1–2 mL). Each wash should last at least 5 min and be carried out on an orbital shaker at room temperature.
18. Carefully aspirate the wash buffer from the cells and wash twice with enough 1× D-PBS to cover the bottom of the dish to remove any excess Tween-20. Each wash should last at least 5 min and be carried out on an orbital shaker at room temperature.

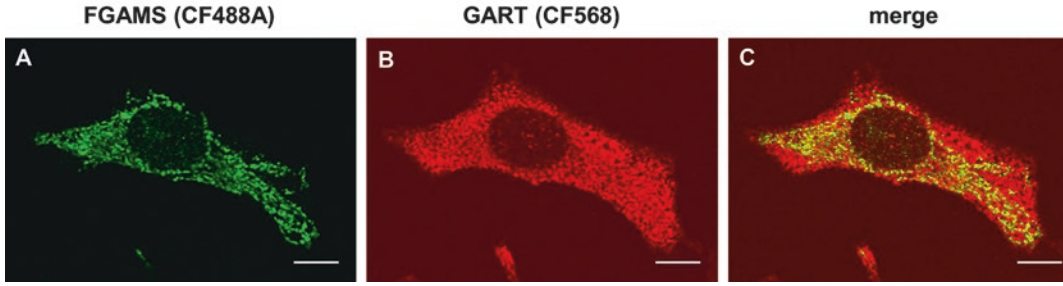


Fig. 2 Colocalization of endogenous FGAMS and GART for visualization of purinosomes by immunofluorescence. Purine-depleted HeLa cells were fixed and permeabilized prior to being probed for with FGAMS rabbit polyclonal antibody and GART mouse monoclonal antibody. Fluorescently labeled secondary antibodies CF488A-conjugated donkey anti-rabbit and CF568-conjugated donkey anti-mouse were used to visualize the expression and localization of FGAMS and GART, respectively. A representative image of an individual cell was captured using a 100× oil objective on an Olympus Fluoview 1000 confocal laser scanning microscope. Sequential imaging of CF488A and CF568 showed colocalization of (**a**) FGAMS with (**b**) GART as represented by the yellow puncta present in (**c**) the merged image. Scale bar: 10 μm

19. Add 1 mL of 1× D-PBS to the fixed cells.
20. Image the cells using a confocal laser scanning microscope equipped with filters appropriate for detecting CF488A and CF647 fluorescent dyes. Representative images are shown in Fig. 2.

4 Notes

1. Historically, FGAMS (also referred to as PFAS) has been used as the intracellular marker to denote purinosomes. While all the pathway enzymes have been shown to colocalize with FGAMS, the best combinations, based on reagents available, are between FGAMS and one of the other "core" purinosome proteins, PPAT, and GART [20].
2. HeLa cells are grown under purine-depleted growth conditions for at least two to three passages prior to purinosome detection for optimal results. Note that the doubling time of purine-depleted HeLa cells is approximately 28–32 h compared to HeLa cells cultured under normal growth conditions (approximately 20–24 h). Dialyzed FBS is prepared by extensively dialyzing FBS against 0.9% (w/v) sodium chloride prepared in water using a 10 kDa molecular weight cutoff dialysis membrane.
3. Other plasmids may be used for the detection of purinosomes; however, we strongly recommend FGAMS (PFAS) as one of the transient expressing proteins. All fluorescent protein chimeras of pathway enzymes are C-terminal fusions with the

exception of ATIC, where N-terminal fusions are required to prevent disruption of dimerization and enzyme activity. Molecular cloning details can be found in [6].

4. While both DAPI and Hoechst 33342 are popular counterstains for nuclei, Hoechst 33342 works best for live cells, whereas DAPI works best when cells have been fixed and permeabilized. Therefore, we recommend using Hoechst 33342 for live cell imaging (transient transfection-based methods) and DAPI for immunofluorescence-based detection of purinosomes.
5. Older manufacturer's (Invitrogen) protocols suggest a 20 min incubation with both solutions prior to adding the lipid:DNA mixture to the adherent cells. Newer protocols have eliminated the need for this incubation step. Both methods have been used with no detectable difference in transfection efficiency or number of purinosome-positive cells.
6. Opti-MEM™ is a modified form of Eagle's Minimum Essential Medium (MEM) that contains hypoxanthine. Long-term incubation of purine-depleted HeLa cells in this medium may result in loss of purinosome formation. Therefore, Opti-MEM™ must be swapped out with purine-depleted growth medium to achieve optimal results. If this is a concern or does not yield appropriate purinosome formation, try MEM without FBS instead of Opti-MEM™ medium.
7. At this point, the cells can be fixed to preserve purinosome complexation (*see* **steps 2–5** in Subheading 3.2) until imaging by confocal laser scanning microscopy can be performed. Be aware that extended period of time post-fixation might result in decreased fluorescence intensity of the fluorescent protein chimeras. It is best to perform the imaging immediately.
8. Caution must be taken when defining the purinosome metabolon in transient transfected models. Extraction of physical parameters from areas of high colocalization has provided a way to properly identify the purinosome from other well characterized non-membrane-bound cytoplasmic cellular bodies such as processing bodies (P-bodies), stress granules, and aggresomes [10]. These parameters include the overall purinosome diameter and density or number of purinosomes in a cell. Based on an analysis of over 200 purinosome containing HeLa cells, a purinosome has been defined as a cellular body showing colocalization between FGAMS and another pathway enzyme (such as PPAT or GART), having an FGAMS particle diameter between 0.2 and 0.9 μm and encompassing 50–1000 purinosomes per cell [11]. These features can be extracted from images collected and processed through an analysis and visual-

ization software like ImageJ as previously described [11]. Alternatively, co-transfection of other cellular markers (GFP-G3BP, GFP170*, GFP250) can be used to differentiate the purinosome from known stress granules and aggresomes in a similar fashion as outlined in Subheading 3.1 [15].

9. Paraformaldehyde is a neurotoxin and should be handled only in a biosafety cabinet or hood with appropriate personal protective equipment. Paraformaldehyde is also light sensitive and will degrade over time, so to prevent degradation during storage or use, cover all aliquots and samples in aluminum foil.
10. For weakly adherent cells, agitation during the washing steps could result in detachment of the cells. In those cases, carefully add the wash buffer to the side of the 35 mm glass bottom dish dropwise and do not perform washes on an orbital shaker.
11. Incubation of fixed cells with primary antibodies targeting FGAMS and GART can also be performed at room temperature for 3–4 h without detectable differences in immunostaining.

Acknowledgments

The authors wish to thank all current and prior members of the Benkovic Laboratory who have helped in generating and optimizing the methods outlined here. Financial support for this work was provided by the National Institutes of Health (NIH GM024129, S.J.B.).

References

1. Srere PA (1985) The metabolon. Trends Biochem Sci 10(3):109–110
2. Barnes SJ, Weitzman PD (1986) Organization of citric acid cycle enzymes into a multienzyme cluster. FEBS Lett 201(2):267–270
3. Ovadi J, Mohamed Osman IR, Batke J (1983) Interaction of the dissociable glycerol-3-phosphate dehydrogenase and fructose-1,6-bisphosphate aldolase. Quantitative analysis by an extrinsic fluorescence probe. Eur J Biochem 133(2):433–437
4. Puchulu-Campanella E, Chu H, Anstee DJ, Galan JA, Tao WA, Low PS (2013) Identification of the components of a glycolytic enzyme metabolon on the human red blood cell membrane. J Biol Chem 288(2):848–858. https://doi.org/10.1074/jbc.M112.428573
5. Islam MM, Nautiyal M, Wynn RM, Mobley JA, Chuang DT, Hutson SM (2010) Branched-chain amino acid metabolon: interaction of glutamate dehydrogenase with the mitochondrial branched-chain aminotransferase (BCATm). J Biol Chem 285(1):265–276. https://doi.org/10.1074/jbc.M109.048777
6. An S, Kumar R, Sheets ED, Benkovic SJ (2008) Reversible compartmentalization of de novo purine biosynthetic complexes in living cells. Science 320(5872):103–106. https://doi.org/10.1126/science.1152241
7. Evans DR, Guy HI (2004) Mammalian pyrimidine biosynthesis: fresh insights into an ancient pathway. J Biol Chem 279(32):33035–33038. https://doi.org/10.1074/jbc.R400007200
8. Kohnhorst CL, Schmitt DL, Sundaram A, An S (2016) Subcellular functions of proteins under

fluorescence single-cell microscopy. Biochim Biophys Acta 1864(1):77–84. https://doi.org/10.1016/j.bbapap.2015.05.014

9. Chitrakar I, Kim-Holzapfel DM, Zhou W, French JB (2017) Higher order structures in purine and pyrimidine metabolism. J Struct Biol 197(3):354–364. https://doi.org/10.1016/j.jsb.2017.01.003
10. Pedley AM, Benkovic SJ (2017) A new view into the regulation of purine metabolism: the purinosome. Trends Biochem Sci 42(2):141–154. https://doi.org/10.1016/j.tibs.2016.09.009
11. Chan CY, Zhao H, Pugh RJ, Pedley AM, French J, Jones SA, Zhuang X, Jinnah H, Huang TJ, Benkovic SJ (2015) Purinosome formation as a function of the cell cycle. Proc Natl Acad Sci U S A 112(5):1368–1373. https://doi.org/10.1073/pnas.1423009112
12. Zhao H, Chiaro CR, Zhang L, Smith PB, Chan CY, Pedley AM, Pugh RJ, French JB, Patterson AD, Benkovic SJ (2015) Quantitative analysis of purine nucleotides indicates that purinosomes increase de novo purine biosynthesis. J Biol Chem 290(11):6705–6713. https://doi.org/10.1074/jbc.M114.628701
13. An S, Deng Y, Tomsho JW, Kyoung M, Benkovic SJ (2010) Microtubule-assisted mechanism for functional metabolic macromolecular complex formation. Proc Natl Acad Sci U S A 107(29):12872–12876. https://doi.org/10.1073/pnas.1008451107
14. French JB, Jones SA, Deng H, Pedley AM, Kim D, Chan CY, Hu H, Pugh RJ, Zhao H, Zhang Y, Huang TJ, Fang Y, Zhuang X, Benkovic SJ (2016) Spatial colocalization and functional link of purinosomes with mitochondria. Science 351(6274):733–737. https://doi.org/10.1126/science.aac6054
15. French JB, Zhao H, An S, Niessen S, Deng Y, Cravatt BF, Benkovic SJ (2013) Hsp70/Hsp90 chaperone machinery is involved in the assembly of the purinosome. Proc Natl Acad Sci U S A 110(7):2528–2533. https://doi.org/10.1073/pnas.1300173110
16. Baresova V, Skopova V, Sikora J, Patterson D, Sovova J, Zikanova M, Kmoch S (2012) Mutations of ATIC and ADSL affect purinosome assembly in cultured skin fibroblasts from patients with AICA-ribosiduria and ADSL deficiency. Hum Mol Genet 21(7):1534–1543. https://doi.org/10.1093/hmg/ddr591
17. Fu R, Sutcliffe D, Zhao H, Huang X, Schretlen DJ, Benkovic S, Jinnah HA (2015) Clinical severity in Lesch-Nyhan disease: the role of residual enzyme and compensatory pathways. Mol Genet Metab 114(1):55–61. https://doi.org/10.1016/j.ymgme.2014.11.001
18. Schmitt DL, Cheng YJ, Park J, An S (2016) Sequestration-mediated downregulation of de novo purine biosynthesis by AMPK. ACS Chem Biol 11(7):1917–1924. https://doi.org/10.1021/acschembio.6b00039
19. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. Nat Methods 9(7):671–675
20. Deng Y, Gam J, French JB, Zhao H, An S, Benkovic SJ (2012) Mapping protein-protein proximity in the purinosome. J Biol Chem 287(43):36201–36207. https://doi.org/10.1074/jbc.M112.407056

Chapter 18

Analysis of Bacterial Pilus Assembly by Shearing and Immunofluorescence Microscopy

Areli Luna-Rico, Jenny-Lee Thomassin, and Olivera Francetic

Abstract

Bacterial surface appendages of the type 4 pilus superfamily play diverse roles in adherence, aggregation, motility, signaling, and macromolecular transport. Here we describe two analytical approaches to study assembly of type 4 pili and of pseudopili produced by type 2 protein secretion systems: the shearing assay and immunofluorescence microscopy. These complementary antibody-based methods allow for semiquantitative analysis of fiber assembly. The shearing assay can be scaled up to yield crude extracts of pili that can be further analyzed by electron and atomic force microscopy or by mass spectrometry.

Key words Type 4 pili, Type 2 secretion pseudopili, Pilus assembly, Shearing assay, Immunodetection, Tricine SDS-PAGE, Immunofluorescence microscopy

1 Introduction

Type 4 pili (T4P) are thin filaments exposed on the surface of many bacterial and archaeal species [1]. They promote motility, adherence, cell signaling, biofilm formation, DNA uptake, or protein secretion [2]. T4P mediate colonization of tissues by promoting the surface adhesion and attachment, as well as inter-bacterial contacts leading to aggregation and microcolony formation. These features make T4P key virulence factors of many human, animal, and plant pathogens [2]. Many T4P are dynamic and able to retract upon binding to a surface, promoting a form of motility called twitching. Active pilus retraction generates remarkably high forces that pull the bacterial body toward the attachment point [3]. In some bacteria, T4P have been implicated in electron transport and conductance [4].

T4P biogenesis and dynamic function rely on a set of conserved proteins that are part of the superfamily known as type 4 filament (Tff) nanomachines [1]. All Tff systems assemble helical fibers from protein subunits embedded in the plasma membrane. Found in many bacterial species, T4P have common evolutionary

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_18,

origins with archaeal pili and flagella [5, 6]. Another prominent member of this superfamily is the bacterial type 2 protein secretion system (T2SS) [7, 8]. T2SSs assemble periplasmic fibers called pseudopili, which can be visualized on the bacterial surface upon overproduction of the T2SS assembly machinery and/or of the major pseudopilin subunit [9, 10].

The close evolutionary relationship between T4P and T2SS is illustrated by the ability of the *Klebsiella oxytoca* pullulanase T2SS to assemble fibers from T4P subunits [10, 11]. The same system can be used to assemble fibers from major pseudopilins cloned from a variety of heterologous T2SSs [12]. Here, we describe the use of the Pul T2SS machinery reconstituted in *E. coli* and cloned on a moderate copy number plasmid to assemble filaments from its cognate major pseudopilin subunit PulG and from the *E. coli* T4P subunit PpdD. The same protocols can also be used to efficiently shear flagella or other types of pili (type 1 pili or pili from Gram-positive bacteria) from the bacterial surface, permitting downstream analyses in both the cell-bound and soluble fractions. Using antibodies specific for fiber subunits, this method allows for the global quantitative assessment of pilus assembly efficiency, expressed as a ratio between the assembled pilin subunits on the bacterial surface and the total amount of pilins, including both those assembled into periplasmic fibers and the pool of pilin subunits in the plasma membrane awaiting assembly. Immunofluorescence microscopy is a complementary assay that provides insight into the fiber length and number of pili per cell, giving an estimate for the number of active assembly machineries. In contrast to the global approach, immunofluorescence microscopy only allows one to visualize fibers assembled on the cell surface. The complementarity of these two approaches is illustrated in previous studies of T2SS models [13, 14].

2 Materials

Standard analytical grade chemicals are purchased from established commercial suppliers. Prepare all solutions using ultrapure water.

2.1 Bacterial Culture

1. *Escherichia coli* K-12 strain PAP7460 (MC4100 Δ*(lac-argF) U169 araD139 relA1 rpsL 150 ΔmalE444 malG501* [F′ *lacI*^{Q}Tn*10*]) (*see* **Note 1**).
2. Lysogeny broth (LB).
   Dissolve 10 g of bacto-tryptone, 5 g of yeast extract, and 10 g of NaCl in 1 L of distilled H_2O (dH_2O). If required, adjust the pH to 7.0 with 1 N NaOH. Autoclave at 121 °C for 20 min to sterilize.

3. Pilus-inducing solid media: in this case we used LB with 1.5% agar sterilized as above and supplemented, after cooling to 45–50 °C, with Ap (100 μg/mL), Cm (25 μg/mL), and 0.2% D-maltose.
4. An incubator, set to 30 °C.

2.2 Shearing of Pili from the Bacterial Surface

1. Sterile toothpicks or a platinum loop.
2. Vortex.
3. LB (around 3 mL per sample).
4. One 2 mL syringe with 26-gauge needle for each sample.
5. A spectrophotometer at 600 nm and disposable 1 mL cuvettes.
6. A benchtop microcentrifuge (4°C).
7. 100% trichloroacetic acid (TCA).
8. 100% acetone.
9. A ventilated fume hood.
10. SDS sample buffer (150 mM Tris-HCl (pH 6.8), 6% sodium dodecyl sulfate (SDS), 30% glycerol containing the tracking dye—typically bromophenol blue or phenol red at 0.05 mg/mL).

2.3 Tris-Tricine Denaturing Gel Electrophoresis

1. A vertical gel caster system and slab gel electrophoresis apparatus with a DC power supply.
2. 70% ethanol (v/v).
3. 40% acrylamide-bis solution (37.5:1), stored at 4 °C.
4. Tricine gel buffer (3×): 3 M Tris, 1 M HCl, 0.3% (w/v) SDS, pH 8.45.
5. 10% (w/v) ammonium persulfate (APS), stored in aliquots at −20 °C.
6. N,N,N,N′-Tetramethyl-ethylenediamine (TEMED), stored at room temperature.
7. Anode buffer 10× (for bottom reservoir): 1 M Tris, 0.225 M HCl, pH 8.9.
8. Cathode buffer 10× (for top reservoir): 1 M Tris, 1 M Tricine, 1% (w/v) SDS, pH 8.25 (if made correctly, pH does not need to be adjusted).

2.4 Immunoblotting

1. Nitrocellulose membranes (0.45 μm pore) optimized for ECL.
2. Whatman 3MM chromatography paper.
3. Flat tip tweezers.
4. A fast blotter (Pierce G2, Thermo Fisher Scientific) or any other semidry electro-transfer apparatus.

5. 1-Step Transfer Buffer (Thermo Fisher Scientific) or standard transfer buffer (25 mM Tris-HCl, 192 mM glycine, 10% ethanol, pH 8.3).
6. 0.2% Ponceau S (w/v) solution in 3% TCA. This solution is stored at room temperature and reused.
7. Tris-buffered saline with Tween-20 (TBST): 50 mM Tris, 150 mM NaCl, 0.05% Tween-20, pH 7.6.
8. Blocking solution: 5% nonfat dry milk in TBST or 1% BSA in TBST.
9. Primary antibody to detect pilin subunit, appropriately diluted in blocking solution (*see* **Note 2**).
10. Secondary antibody: horseradish peroxidase-coupled anti-rabbit antibody, diluted 1:40,000 in TBST.
11. Enhanced chemiluminescence Pierce ECL2 western blotting substrate.
12. A chemiluminescence detection apparatus.

2.5 Immunofluorescence Microscopy

1. A portable vacuum aspiration system that can be placed under the fume hood.
2. An ultrasonicator (with standard small probe).
3. Delicate task wipers.
4. Clear nail polish.
5. Dulbecco's phosphate-buffered saline (PBS) without $MgCl_2$ or $CaCl_2$.
6. Blocking solution: 1% BSA in PBS.
7. Poly-L-lysine hydrobromide stock solution (1 mg/mL in dH_2O) filter sterilized, store at −20 °C. Poly-L-lysine working solution (100 μg/mL). Prepare by diluting stock solution 1:10 with PBS.
8. Cleaned rectangular coverslips 22 mm × 22 mm. An ultrasonication protocol is used to clean the coverslips, as described in Subheading 3.4.
9. Microscope slides (75 mm × 26 mm).
10. Tissue culture test plates with six wells.
11. 37% paraformaldehyde (PFA) stock solution in dH_2O (pH 7.4), filter sterilize and store at −20 °C. Prepare working solution by diluting stock solution 1:10 in PBS, filter sterilize and store at −20 °C (*see* **Note 3**).
12. 1 M Tris-HCl pH 8.0.
13. Primary antibody of interest; for this protocol polyclonal rabbit antisera raised against pilus subunits diluted to 1:1000 in 1% BSA-PBS.

14. Secondary antibody; Alexa 488-coupled goat anti-rabbit IgG diluted to 1:200 in 1% BSA-PBS.
15. ProLong Gold Antifade Reagent with DAPI.
16. Microscopy tweezers with fine tips.
17. A fluorescence microscope equipped with a digital camera.
18. Immersion oil.

3 Methods

3.1 Shearing of Pili from the Cell Surface

Streak a dense bacterial lawn on pilus-inducing agar media, incubate at 30 °C for 48 h.

1. Scrape the bacteria off the plates with a sterile toothpick or sterile platinum loop and transfer into a microcentrifuge tube containing 1 mL of LB. Resuspend by pipetting until clumps are dissolved (*see* **Note 4**).
2. Measure the OD_{600nm} using 1 mL of LB as a blank. Dilute the sample, if necessary, to ensure the measurement is within the linear range of the spectrophotometer.
3. Normalize all samples to 1 mL, $OD_{600nm} = 1$.
4. Vortex suspensions continuously for 2 min at maximum speed.
5. Pass suspensions through a 26-gauge needle six times to shear pili from the cell surface (*see* **Note 5**).
6. Centrifuge the suspension 15 min at 16,000 × *g* in a microcentrifuge at 4 °C.

   Transfer 0.85 mL of supernatant (SF 1) from the topmost fraction to a clean microfuge tube while avoiding disturbing the pellet. Discard the remaining supernatant (0.15 mL) from the cell fraction (CF) and resuspend the pellet in 100 µL of SDS sample buffer. Reserve the CF sample on ice or at −20 °C for SDS-PAGE analysis.
7. Centrifuge the tube containing 0.85 mL of SF 1 for 10 min, at 16,000 × *g* at 4 °C. This step will remove remaining bacteria from the supernatant.
8. On ice, transfer 0.7 mL of supernatant off the topmost fraction into a clean microcentrifuge tube, now called sheared fraction 2 (SF 2) (Fig. 1).
9. Place the tube on ice. Add 1/10 volume (70 µL) of cold 100% TCA solution to obtain a final concentration of 10% TCA. Vortex immediately (s*ee* **Note 6**).
10. Incubate the solution on ice for 30 min.
11. Place the tubes in the same orientation in the microcentrifuge and pellet at 16,000 × *g* for 30 min to 1 h at 4 °C.

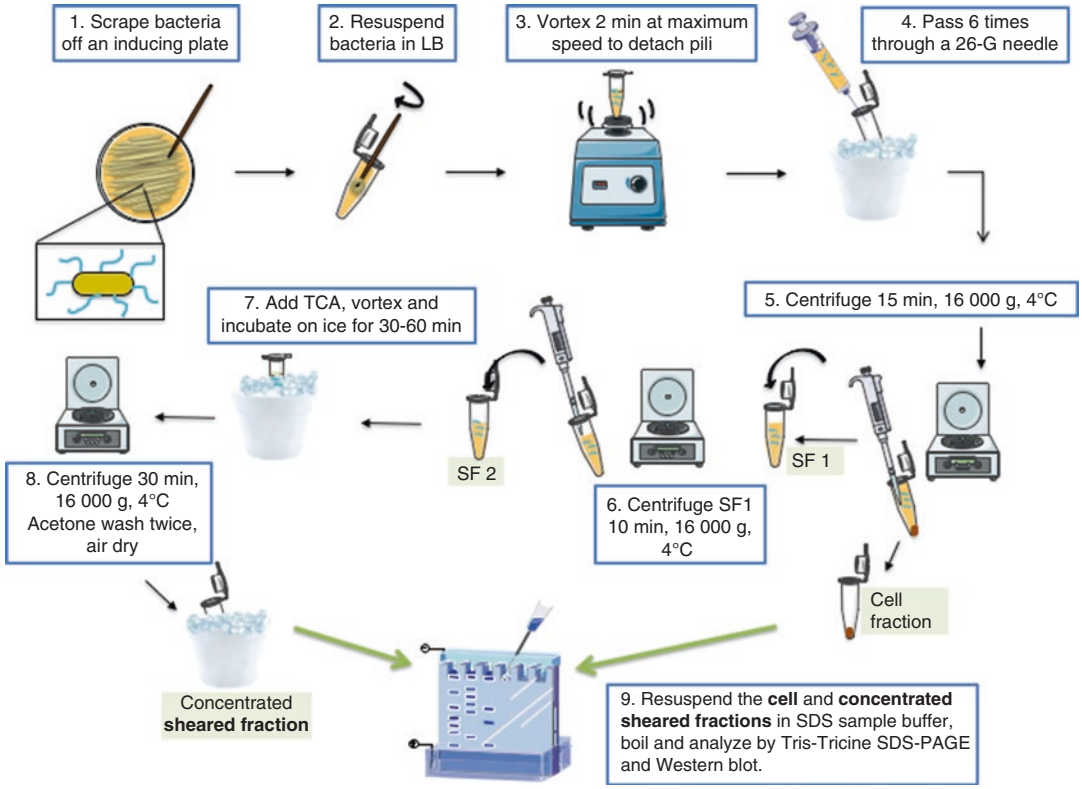


Fig. 1 Shearing assay workflow. Shown is a schematic of the workflow for a shearing assay, with individual steps numbered and indicated in boxes

12. Carefully aspirate the supernatant (*see* **Note 7**).
13. Add 1 mL of cold acetone (kept at −20 °C) to wash the pellet, without resuspending. Centrifuge at 4 °C for 2 min at 16,000 × *g*. Aspirate off the supernatant carefully.
14. Repeat acetone wash as above. Air-dry the pellet while keeping the Eppendorf tubes open on ice under the fume hood for 10–15 min. Resuspend in 70 μL of SDS sample buffer.
15. Pilins are relatively abundant proteins. Analyze equivalent amounts of sample (typically to 0.05 OD_{600nm} for CF and SF) by Tris-Tricine SDS-PAGE as described in Subheading 3.2 (Fig. 2). This electrophoresis system allows small proteins to be well resolved, rendering it highly suitable for pilins that have a size range from 10 to 20 kDa.

3.2 Tris-Tricine SDS-PAGE

Analysis of concentrated cell and sheared fractions on 10% Tris-Tricine SDS-PAGE [15] (*see* **Note 8**).

1. Wipe gel plates with 70% ethanol and cast the gel (any vertical slab gel apparatus is suitable).
2. Prepare 15 mL of the separating gel solution (enough for 4 BioRad Mini-Protean II gels or 2 Apelex Mini-Wide Vertigel

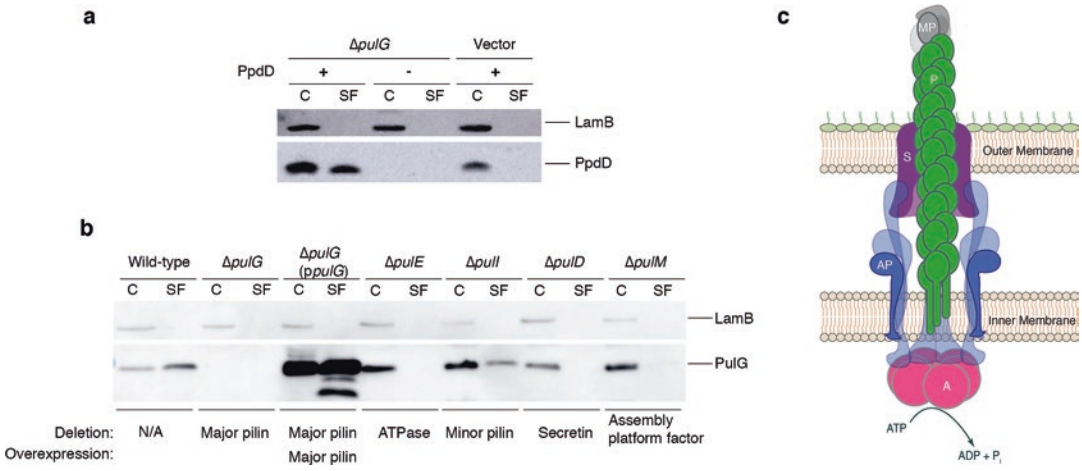


Fig. 2 Assembly of the PpdD and PulG pili by the Pul T2SS in *E. coli* K-12. (**a**) Immuno-detection of PpdD in cell fractions (C) and sheared fractions (SF) of *E. coli* K-12 strain PAP7460 co-expressing the Pul system and/or PpdD pilin indicated by (+), or co-expressed with the empty vector in each case indicated by (–) as negative controls. (**a**, **b**) The outer membrane protein LamB was used as an indicator for outer membrane contamination in the sheared fraction. Other abundant outer membrane markers can also be used. (**b**) Immuno-detection of PulG in cell fractions (C) or sheared fractions (SF) of *E. coli* K-12 strain PAP7460 co-expressing the Pul system (wild-type) or its derivatives harboring gene deletions (indicated) and a compatible empty vector pSU19 or its derivatives encoding major T2SS pseudopilin *pulG*. When applicable, the T2SS component missing due to gene deletion, or overexpression of a component gene is indicated underneath the immunoblot. (**c**) Schematic of an assembled T2SS/T4P machine. Missing components from panel b: major pilin (P), ATPase (A), minor pilin (MP), secretin (S), and assembly platform (AP) factor proteins are indicated

II gels) by mixing 6.25 mL dH_2O, 5 mL 3× gel buffer, and 3.75 mL acrylamide solution. Add 75 µL 10% APS and 30 µL TEMED, mix well and pour immediately between the glass plates leaving space for the stacking gel. Overlay gently with 1 mL dH_2O and allow to polymerize at room temperature. Once polymerized, pour out dH_2O and carefully wick out any remaining dH_2O using a piece of Whatman 3MM chromatography paper.

3. Clean the combs with 70% ethanol and prepare the stacking gel solution by combining 5.7 mL dH_2O, with 3.3 mL 3× gel buffer and 1 mL acrylamide. Add 90 µL 10% APS and 40 µL TEMED, mix well and pour on top of separating gel. Insert the comb immediately and leave at room temperature until polymerized.
4. Place the gel into the apparatus. Fill the top tank with 1× cathode buffer and the bottom with 1× anode buffer. Remove combs and wash the wells by pipetting gently.
5. Load a well with 2 µL pre-stained molecular weight marker and the equivalent of 0.05–0.1 OD_{600nm} of each CF and SF. Run the gel at a constant current of 30 mA per Biorad minigel (typically 2 h and 30 min), until the tracking dye reaches the bottom of the gel.

3.3 Western Blot and Immuno-detection

1. Remove the gel from the apparatus, separate the glass plates, and use a sharp tool to remove the stacking gel. Transfer the separating gel into a container containing 10 mL of transfer buffer and equilibrate 5 min with gentle agitation on a benchtop shaker.
2. Set up the electro-transfer: Prepare ten sheets of Whatman 3MM paper and cut the nitrocellulose membrane slightly larger than the size of the gel. Wet the anode (bottom) plate of the transfer apparatus with 3 mL transfer buffer. Soak the Whatman paper in transfer buffer and stack five sheets on the anode plate; use a plastic roller or tube to remove air bubbles between each layer. Wet the nitrocellulose membrane and place it on the stack, center the gel on the membrane, and cover with the remaining five sheets of presoaked Whatman paper. Wet the cathode (top plate), place it on top of the stack, press with even pressure to seal the plates together, and insert the closed cassette into the transfer apparatus.
3. Transfer proteins from the gel onto the membrane using the standard preset program for mixed molecular weight range (at 25 V and 1.3 A, for 7 min). Alternatively, standard transfer buffer and a semi-dry transfer apparatus can be used in this step; in that case transfer should proceed for 45 min at a constant current of 100 mA.
4. Remove the membrane from the transfer apparatus and stain with Ponceau S solution for 5 min. Destain with dH_2O using a squirt bottle, until protein bands appear, verify even transfer, and take a picture of the result.
5. Wash the membrane in 20 mL TBST with gentle agitation, until the red protein bands are no longer visible. Discard the destaining solution.
6. Cover the membrane in blocking solution and incubate for 1 h at room temperature with mild agitation on a benchtop shaker. Discard the blocking solution.
7. Add the primary antibody working solution and incubate for 1 h with mild agitation as above, at room temperature. Remove primary antibody. If desired, primary antibody solution can be stored at −20 °C and reused 5–10 times.
8. Wash the membrane four times for 10 min in TBST.
9. Add the secondary antibody working solution and incubate for 1 h with agitation. Discard the secondary antibody solution.
10. Wash the membrane four times for 10 min in TBST.
11. Blot excess liquid from the membrane using a sheet of Whatman filter paper. Prepare the ECL2 developing solution by mixing 2 mL of solution A with 50 μL of solution B in a dish made of glass or inert plastic (polyallomer). Soak the

membrane face down in this solution for 5 min to activate the chemiluminescent reaction.

12. Record the signal (emitted light) using the chemiluminescence imager.

Quantify the signal for the mature pilin in cell and sheared fractions using Image J or other software. Piliation efficiency is expressed as the percentage of total signal present in the sheared fraction, i.e., $E = d_{SF}/(d_{SF} + d_{CF}) \times 100$. E represents piliation efficiency (in %), and d represents the measured density of sheared fraction (SF) and cell fraction (CF) expressed in arbitrary units.

3.4 Ultrasonication of Coverslips for Immunofluorescence Microscopy

(*see* **Note 9**).

1. Place the coverslips on a rack inside a wide crystallization beaker and add dH_2O, so that the rack is fully immersed in liquid.
2. Place the ultrasonicator tip in the liquid close to the rack without touching it.
3. Ultrasonicate at 23–27% for 20 min using 30 s pulse and 30 s pause intervals.
4. Wash the rack twice with sterile dH_2O.
5. Wash the rack with 90% ethanol.
6. Wash three times with sterile dH_2O.
7. Aspirate the excess of water from in-between coverslips. Cover the beaker containing the coverslip rack with aluminum foil and dry at 42 °C overnight. The coverslips can be prepared days in advance and stored at room temperature protected from dust with parafilm and aluminum foil.

3.5 Immunofluorescence Microscopy

1. Place one clean coverslip per well in the 6-well tissue culture test plate. Coat the surface of each coverslip with 1 mL of the poly-L-lysine working solution, ensuring the entire surface is covered. Close the lid of the tissue culture test plate and incubate at 37 °C for 1 h.
2. Aspirate the liquid using a vacuum pump and dry at 37 °C (*see* **Note 10**).
3. Wash the coverslips three times with dH_2O and dry at 37 °C.
4. The bacterial samples for the immunofluorescence microscopy assay are grown using the same pilus-inducing conditions described for the shearing assay (Subheading 3.1). For immunofluorescence microscopy assay, bacteria are removed from the plate using a sterile inoculation loop and transferred into 1 mL PBS. Gently rotate the loop to resuspend the cells without shearing the pili (*see* **Note 11**).

5. Measure the OD_{600nm} using 1 mL PBS as a blank. Dilute the sample, if necessary, to ensure the measurement is within the linear range of the spectrophotometer.
6. Normalize all samples to 1 mL, 0.1 OD_{600nm} (*see* **Note 12**).
7. Cover the entire surface of the coverslip with the bacterial suspension (*see* **Note 13**).
8. Incubate the culture plate at room temperature, static for 1 h.
9. Wash three times with 1 mL PBS gently by adding and aspirating the liquid on the side. Avoid touching or disturbing the surface of the coverslip.
10. In a fume hood, add 1 mL of 3.7% PFA in PBS working solution and incubate the samples for 30 min at room temperature in the fume hood to fix the bacteria and pili to the coverslip.
11. In a fume hood, stop the reaction by adding 500 μL 1 M Tris-HCl (pH 8.0). Remove the liquid immediately and dispose in the appropriate toxic liquid waste container.
12. Add 1 mL 1 M Tris-HCl (pH 8.0) and incubate 5 min at room temperature.
13. Wash three times with 1 mL PBS.
14. Flood the surface of the coverslip with blocking solution (1% BSA in PBS). Incubate 1 h at room temperature without shaking. Use vacuum to aspirate blocking solution.
15. Cover the entire surface of the coverslip with primary antibody solution (500 μL), and incubate 1 h at room temperature without shaking. Aspirate off the primary antibody solution.
16. Wash the coverslip by flooding the coverslip with 1 mL PBS, gently adding PBS to the space next to the coverslip until the coverslip is fully covered with PBS. Incubate for 1 min without shaking and then aspirate all liquid. Repeat this step two more times.
17. Cover the entire surface of the coverslip with secondary antibody solution (500 μL of fluorophore-coupled anti-rabbit diluted in PBS), protect the samples from light, and incubate 1 h at room temperature without shaking (*see* **Note 14**). Aspirate off the liquid.
18. Wash three times with 1 mL PBS as in **step 16**.
19. Place the microscope slides on a piece of aluminum foil. Clean with tissue paper and 70% ethanol. Air-dry.
20. Put a 10 μL drop of the ProLong Gold Antifade Reagent with DAPI on the microscope slide to stain the bacterial nucleoid (*see* **Note 15**).
21. With microscopy tweezers (with the aid of a needle if necessary), carefully remove the coverslip from the tissue culture test

plate and use tissue paper to wick off excess liquid from the edge of the coverslip; be sure not to touch the fixed sample. Place the coverslip face down in the center of the DAPI-ProLong drop. Press down gently to remove air bubbles.

22. Apply a spot of nail polish to the corners of the coverslip. Let dry in the dark.
23. Apply nail polish around the edges of the coverslip to seal, let dry in the dark, and store the slides in the dark at room temperature.
24. Samples are examined using a fluorescence microscope equipped with 63× or 100× immersion oil objectives, using blue (DAPI) and green (Alexa 488) filters. Here, samples were visualized with an upright Axio Imager A2 microscope (Zeiss); images were captured using the AxioCam MRm digital camera connected to the microscope. Images were analyzed with Zen 2012 software (Fig. 3) (*see* **Note 16**).

4 Notes

1. For pilus assembly assays, PAP7460 was co-transformed with plasmid pCHAP8185 (Ap^R) encoding the complete set of *pul* genes from the *Klebsiella oxytoca* T2SS or its derivatives harboring gene deletions and a compatible empty vector pSU18/pSU19 (Cm^R) or its derivatives encoding major T4 pilin *ppdD* or major T2SS pseudopilin *pulG*. To induce the *pul* genes encoding the fiber assembly machinery, 0.2% D-maltose is added to LB Ap Cm plates. Depending on the expression system, different antibiotics and/or inducing media may be necessary.
2. Primary antibody is diluted to its working concentration in TBST-5% milk. Here, custom rabbit polyclonal antibodies raised against MalE-PpdD, PulGsp-His, and LamB were diluted to 1:1000, 1:2000, and 1:1000, respectively. To reduce the number of non-specific proteins visualized during immunodetection, the primary polyclonal antisera can be purified by adsorbing against a bacterial extract lacking the antigen of interest. Here, bacterial extracts were produced by growing the appropriate PAP7460 strain lacking the antigen of interest under *pul*-inducing conditions; cells were collected by centrifugation, resuspended in PBS, and broken open using a cell disrupter; and resulting cell extracts were then stored at −20 °C until used. To adsorb the antibodies, bacterial extracts (1 mL) were combined with undiluted primary serum (40 μL) in 1.5 mL Eppendorf tubes and incubated for 1 h at 37 °C. Antigen-antibody complexes and debris were removed

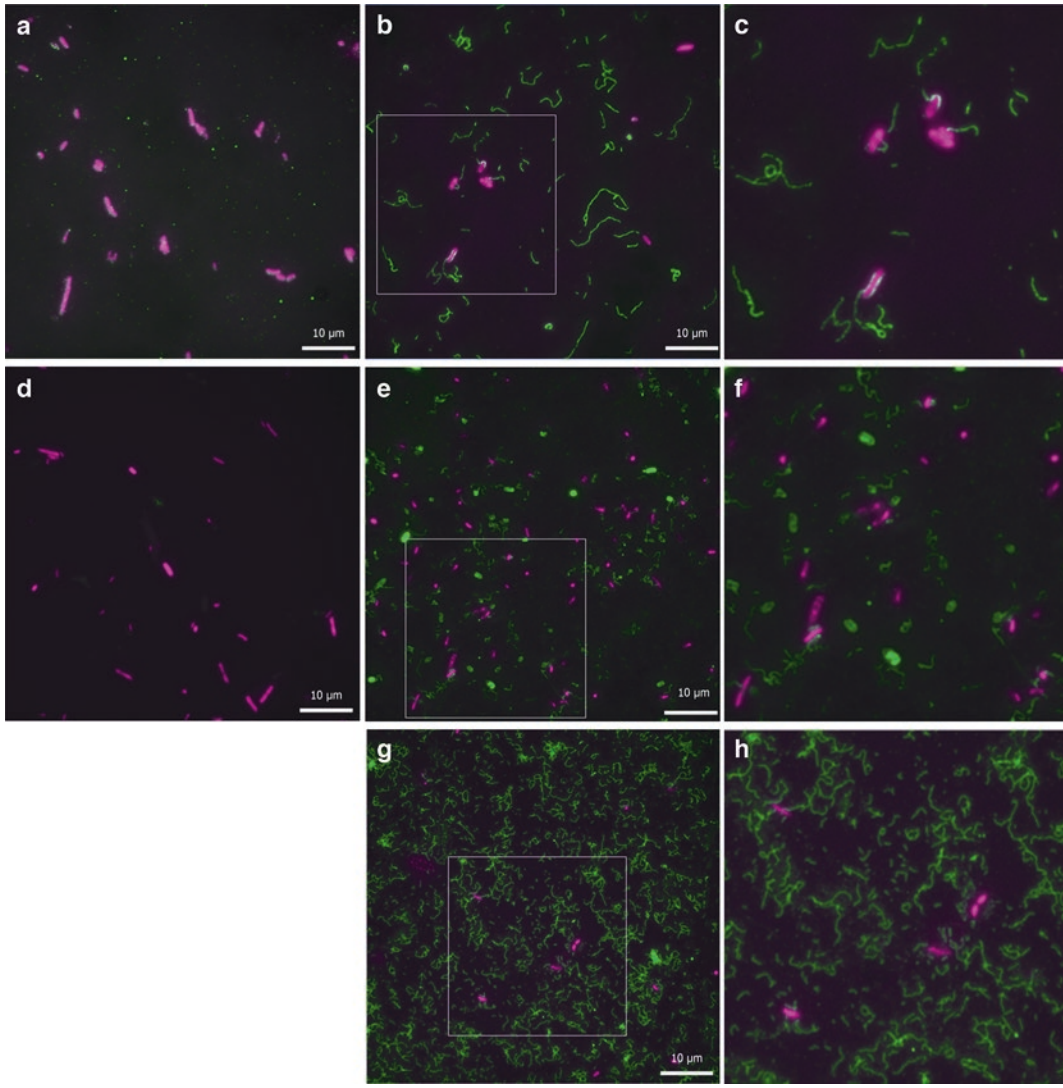


Fig. 3 Immunofluorescence microscopy analysis of PpdD (**a**–**c**) and PulG (**d**–**h**) fiber assembly. DAPI staining was used to detect DNA in fixed bacteria (magenta), and surface fibers were detected using PpdD (**a**–**c**) or PulG (**d**–**h**) antibodies (green). (**a**) and (**d**) Negative control cells expressing the Pul T2SS without the major pilin subunit from plasmid pCHAP8184 (*ΔpulG*) and empty vector; (**b**) cells co-expressing the Pul T2SS from pCHAP8184 and the *ppdD* gene encoding thr T4P subunit (from pCHAP8565) [13]. (**c**) Enlarged area of interest highlighted in panel **b**. (**e**) Bacteria expressing the complete set of *pul* genes from plasmid pCHAP8185 and empty vector. (**f**) Enlarged area of interest highlighted in panel e. (**g**) Bacteria co-expressing the *pul* genes from pCHAP8184 *(ΔpulG*) and the *pulG* gene encoding the major T2SS pseudopilus subunit from pCHAP8658 [14]

by centrifugation at 16,000 × *g* for 5 min at 4 °C. Clarified supernatants were transferred to a fresh tube and were diluted in TBST or PBST to the appropriate working concentration.

3. PFA is classified as a toxic, inflammable, harmful product; security measures must be taken for the preparation of the solution; weigh and dissolve the compound in a chemical hood; wear appropriate personal protective equipment such as gloves, goggles, and a mask equipped with a particle filter. PFA is soluble at alkaline pH. To prepare the stock solution, add 500 µL of 7.5 N NaOH to 100 mL of PFA in dH_2O and heat to 55 °C to dissolve. Do not overheat as it can denature the PFA. When fully dissolved, adjust the pH back to 7.4 with concentrated HCl. Filter and store frozen in aliquots.
4. After this point, maintain samples on ice or at 4 °C.
5. Passing samples through the 26 G needle is not required to shear PulG pseudopili; however, this step significantly improves the yield of PpdD T4P. Whether or not this step is required depends on the thickness and elasticity of the surface fibers and should be determined empirically for each system.
6. Addition of a carrier protein, such as BSA (final concentration 1 µg/mL), can improve TCA precipitation efficiency. Carrier proteins can also serve as useful loading and transfer controls after protein electro-transfer and can be tracked by Ponceau S staining.
7. At this step, the TCA-precipitated protein pellet may be invisible or may detach from the bottom of the tube; aspirate off the supernatant very carefully using a drawn-out Pasteur pipet or a micro-tip from the opposite side of where the pellet is expected to accumulate.
8. Acrylamide is a neurotoxin; wear appropriate personal protective equipment when handling, such as gloves and goggles. Discard contaminated waste in appropriate containers.
9. Ideally, this step should be performed in a closed chamber, devoid of personnel. Any person present in the room during ultrasonication must wear hearing protection (earmuffs).
10. At each drying step, it is important for the coverslip to be completely dry before proceeding to the next step.
11. It is critical to be very gentle at this step; too much movement will detach pili from the cell surface.
12. The amount of the sample analyzed will depend on the plasmid copy number; in this case we use 0.05 OD_{600nm} for medium- and 0.1 OD_{600nm} for low copy number plasmids.
13. Use wide-mouth pipette tips (cut the tip off with sterile scissors) when transferring bacterial suspensions to reduce shearing.

14. To prevent photo-bleaching, it is critical to protect samples from light during this step. Covering the tissue culture test plate with aluminum foil and placing samples in a dark room or a closed drawer will protect samples from light.
15. As in **Note 13**, use wide-mouth pipette tips to transfer the ProLong reagent to the microscopy slide. ProLong is an anti-fade reagent that attenuates photo-bleaching during fluorescence imaging; the ProLong reagent used here contains DAPI DNA stain. Two coverslips can be placed on one microscopy slide.
16. The numbers and lengths of pili can be quantified in a representative number of fields, and the data can be analyzed with appropriate methods to compare the piliation phenotypes of different strains grown under the same conditions. Examples of such semiquantitative analysis are provided in [13, 14].

Acknowledgments

The work in our group is funded by the Institut Pasteur, CNRS and ANR grant 14-CE09-0004. A.L.R. was funded by the Pasteur-Paris University PhD program. J.L.T. was funded by the ANR grant 14-CE09-0004 and by the NSERC postdoctoral fellowship. We thank Nadia Izadi-Pruneyre, Daniel Ladant and members of the NMR of Biomolecules and Biochemistry of Macromolecular Interactions Units for interest and support. We thank Servier Medical Art (http://www.servier.com/ Powerpoint-image-bank) as a source of drawings used in Fig. 1.

References

1. Berry JL, Pelicic V (2015) Exceptionally widespread nanomachines composed of type IV pilins: the prokaryotic Swiss Army knives. FEMS Microbiol Rev 39:134–154. https://doi.org/10.1093/femsre/fuu001
2. Strom MS, Lory S (1993) Structure-function and biogenesis of the type IV pili. Annu Rev Microbiol 47:565–596. https://doi.org/10.1146/annurev.micro.47.1.565
3. Mattick JS (2002) Type IV pili and twitching motility. Annu Rev Microbiol 56:289–314. https://doi.org/10.1146/annurev.micro.56.012302.160938
4. Reguera G, McCarthy KD, Mehta T et al (2005) Extracellular electron transfer via microbial nanowires. Nature 435:1098–1101. https://doi.org/10.1038/nature03661
5. Hobbs M, Mattick JS (1993) Common components in the assembly of type 4 fimbriae, DNA transfer systems, filamentous phage and protein-secretion apparatus: a general system for the formation of surface-associated protein complexes. Mol Microbiol 10:233–243. https://doi.org/10.1111/j.1365-2958.1993.tb01949.x
6. Peabody CR, Chung YJ, Yen MR et al (2003) Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. Microbiology 149:3051–3072. https://doi.org/10.1099/mic.0.26364-0
7. Thomassin J-L, Santos Moreno J, Guilvout I et al (2017) The trans-envelope architecture and function of the type 2 secretion system: new insights raising new questions. Mol Microbiol 105(2):211–226. https://doi.org/10.1111/mmi.13704

8. Korotkov KV, Sandkvist M, Hol WGJ (2012) The type II secretion system: biogenesis, molecular architecture and mechanism. Nat Rev Microbiol 10:336–351. https://doi.org/10.1038/nrmicro2762
9. d'Enfert C, Ryter A, Pugsley AP (1987) Cloning and expression in *Escherichia coli* of the *Klebsiella pneumoniae* genes for production, surface localization and secretion of the lipoprotein pullulanase. EMBO J 6:3531–3538
10. Sauvonnet N, Vignon G, Pugsley AP, Gounon P (2000) Pilus formation and protein secretion by the same machinery in *Escherichia coli*. EMBO J 19:2221–2228. https://doi.org/10.1093/emboj/19.10.2221
11. Sauvonnet N, Gounon P, Pugsley AP (2000) PpdD type IV pilin of *Escherichia coli* K-12 can be assembled into pili in *Pseudomonas aeruginosa*. J Bacteriol 182:848–854. https://doi.org/10.1128/JB.182.3.848-854.2000.Updated
12. Vignon G, Köhler R, Larquet E et al (2003) Type IV-like pili formed by the type II secreton: specificity, composition, bundling, polar localization, and surface presentation of peptides. J Bacteriol 185:3416–3428. https://doi.org/10.1128/JB.185.11.3416-3428.2003
13. Cisneros DA, Bond PJ, Pugsley AP et al (2011) Minor pseudopilin self-assembly primes type II secretion pseudopilus elongation. EMBO J 31:1041–1053. https://doi.org/10.1038/emboj.2011.454
14. Santos-Moreno J, East A, Guilvout I et al (2017) Polar N-terminal residues conserved in type 2 secretion pseudopilins determine subunit targeting and membrane extraction steps during fibre assembly. J Mol Biol 429(11):1746–1765. https://doi.org/10.1016/j.jmb.2017.04.005
15. Schägger H (2006) Tricine–SDS-PAGE. Nat Protoc 1:16–22. https://doi.org/10.1038/nprot.2006.4

Chapter 19

Expression, Purification, and Assembly of Archaellum Subcomplexes of *Sulfolobus acidocaldarius*

Paushali Chaudhury, Patrick Tripp, and Sonja-Verena Albers

Abstract

The archaellum assembly machinery and its filament consist of seven proteins in the crenarchaeon *Sulfolobus acidocaldarius*. We have so far expressed, purified, and biochemically characterized four of these archaellum subunits, namely, FlaX, FlaH, FlaI, and FlaF. FlaX, FlaH, and FlaI tightly interact and form the archaellum motor complex important for archaellum assembly and rotation. We have previously shown that FlaH forms an inner ring within a very stable FlaX ring, and therefore FlaX is believed to provide the scaffold for the assembly of the archaellum motor complex. Here we describe how to express and purify FlaX and FlaH and how the double ring structure both form can be obtained.

Key words Archaea, Archaellum, Archaellum motor complex, Type IV pili, Motility, Archaeal flagellum

1 Introduction

The archaellum is the archaeal motility structure and propels cells forward by the means of a rotary filament [1]. Its assembly machinery is evolutionary related to archaeal and bacterial type IV pili [2]. In contrast to the bacterial flagellum, where rotation of the filament relies on the proton motive force, the archaellar filament is rotated by the energy provided by ATP hydrolysis [3]. The archaellum is a relatively simple nanomachine as the minimal archaellum assembly machinery requires only seven proteins and is found in crenarchaea like *Sulfolobus acidocaldarius*, while euryarchaeal archaellum operons can contain up to 13 genes. The core components of the archaellum are the filament protein, the archaellins, which can, dependent on the species, be present in one up to five different copies. The proteins FlaF and probably also FlaG are extracellular components of the archaellum and anchor it in the cell envelope (Fig. 1a) [4]. The archaellum motor complex consists of FlaH, FlaI, and FlaJ where FlaJ is the only polytopic membrane protein. FlaI is an ATPase and involved in the assembly and

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_19,

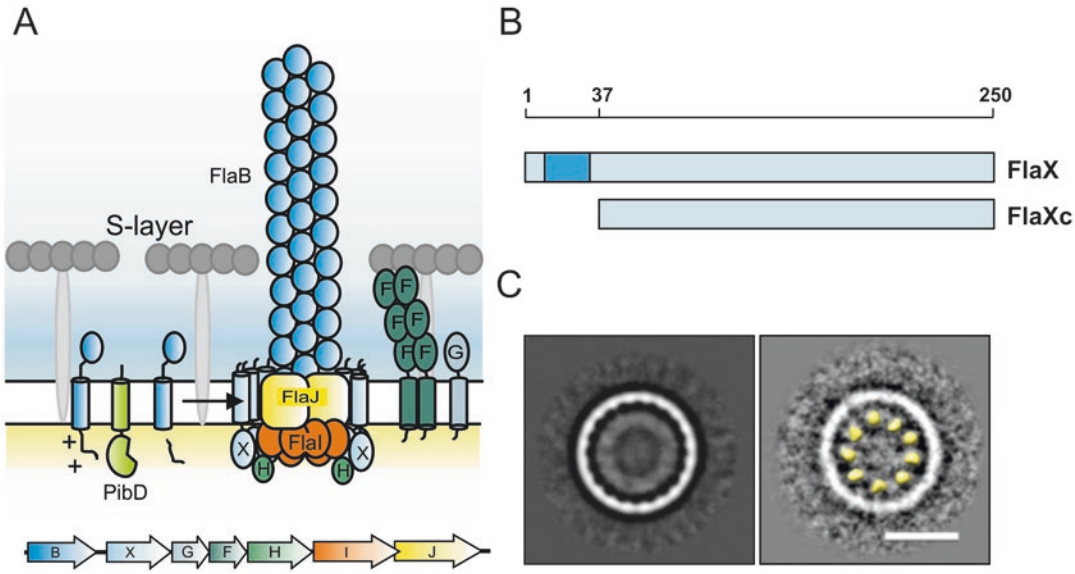


Fig. 1 (**a**) Working model of *S. acidocaldarius* archaellum and operon encoding the genes of the archaellum components. *Sulfolobus acidocaldarius* consist of single cell membrane enveloped by proteinaceous S-layer. FlaB, the filament-forming protein, is cleaved by a dedicated class III signal peptidase PibD, before its assembly into the filament. The motor component of the archaellum comprises of FlaH, FlaI, FlaX, and FlaJ, the polytopic membrane protein. The motor ATPase, FlaI, interacts with FlaH, a RecA family ATP-binding protein, and FlaX, a monotopic membrane protein. FlaX might act as a scaffold surrounding FlaH, FlaI, and FlaJ complex. FlaF and FlaG are two monotopic membrane proteins where soluble domain of FlaF can interact with the S-layer, and together they might form the stator complex of the archaellum. Genes are represented in the same color as of the respective proteins. (**b**) Schematic representation of the full-length and truncated FlaX where FlaXc is the truncation of first 37 amino acids. The predicted membrane domain is shown in dark blue. (**c**) FlaH forms a second ring inside FlaXc rings. The densities were superimposed with the FlaXc rings, and 20-fold symmetry was imposed to visualize the second ring which revealed nine to ten discrete particles of monomeric FlaH (yellow, filtered to 40 Å resolution). Scale bar = 20 nm. (**a**) and (**c**) were modified from ref. 6

subsequently the rotation of the archaellum [5]. FlaH interacts with FlaI in a nucleotide-dependent manner, and this interaction is essential for the assembly of the archaellum [6]. In crenarchaea the archaellum operon comprises one additional protein, FlaX. FlaX is a monotopic membrane protein, and its soluble part interacts with FlaI and FlaH in the cytoplasm [7]. Interestingly, the soluble part of FlaX forms large oligomeric rings with 15- to 23-fold symmetry and diameters ranging from 26 to 38 nm as measured by cryoelectron microscopy [8]. Therefore, FlaX is believed to be the scaffold that guides the assembly of the archaellum motor complex. In euryarchaea, FlaX is substituted by the proteins FlaC, D, and E. These are cytoplasmic proteins, and so far no biochemical data on these proteins are available, but it is proposed that these proteins act in a similar way as FlaX.

It was shown that FlaX, FlaH, and FlaI from *S. acidocaldarius* interact with nanomolar affinities with each other [7]; therefore

attempts were undertaken to isolate complexes of these proteins. As full-length FlaX heterologous expression was not successful so far, we used a truncated form of FlaX that was devoid of its N-terminal transmembrane domain, called FlaXc (Fig. 1b). We were able to obtain stable complexes of FlaXc and FlaH where FlaH forms an inner ring within the FlaXc ring (Fig. 1c) [6]. In the following, we will describe how we isolated FlaXc and FlaH and then reconstituted the stable FlaXc-FlaH complex.

2 Materials

For the preparations of all buffers, use ultrapure water [obtained by purifying deionized water (18 MΩ-cm at 25 °C)] and analytical grade reagents. All purification steps are performed at room temperature if not noted otherwise.

2.1 Protein Overproduction

1. *E. coli* BL21-DE3 RIL expression strain (Stratagene).
2. LB-amp-cam medium: dissolve 10 g tryptone, 5 g yeast extract, and 10 g NaCl in 1 L of demineralized water. Sterilize by autoclaving. Once the medium has cooled down to room temperature, add 1 mL of filter-sterilized 34 g/L chloramphenicol (cam) and 50 g/L ampicillin (amp), respectively.
3. LB-amp-cam plates: prepare LB and put 15 g/L agar prior to autoclaving. Afterward, cool down medium to approximately 40 °C and add antibiotics. Pour medium in standard plastic petridishes and allow the medium to solidify.
4. 0.5 M isopropyl-β-D-thiogalactopyranoside (IPTG), filter-sterilized in ultrapure water.

2.2 Protein Purification

1. Bandelin Sonopuls HD3100 with KE-76 probe.
2. DNaseI.

2.2.1 Cell Disruption

3. 50 mM Tris–HCl pH 8, 150 mM NaCl.
4. 50 mM Hepes pH 7.2, 150 mM NaCl.
5. Lysis buffer: 50 mM Tris–HCl pH 8, 150 mM NaCl, and 20 mM imidazole, 0.5% TritonX-100, EDTA-free protease inhibitor.

2.2.2 Affinity Chromatography and Sample Analysis

1. 50 mM MES pH 6, 150 mM NaCl.
2. 25 mM glycine-NaOH pH 10, 200 mM NaCl.
3. 2 M imidazole, pH 8.
4. Ni-NTA affinity chromatography beads (e.g., His-select Ni-affinity gel, Sigma).
5. Polypropylene gravity column for affinity chromatography (Bio-Rad): apply 2 mL 50% slurry of the Ni-NTA beads. Allow

the material to settle down. Wash and equilibrate the beads according to the respective purification protocol.

6. Ammonium sulfate.
7. SDS sample buffer (5×): 5 mL 1 M Tris–HCl pH 6.8, 2 g sodium dodecyl sulfate (SDS), 2 g 1,4-dithiothreitol (DTT), 10 mL glycerol, 10 mg bromophenol blue, add up to 20 mL with ultrapure water.
8. Dialysis tubing with 7000 Da molecular weight cutoff (e.g., Serva Membra-Cel, 22 mm diameter, 7000 Da MWCO). Activate membrane according to the manufacturer's instructions.
9. Kit for colorimetric protein concentration determination (e.g., Serva BCA Protein Assay Macro Kit).

3 Methods

3.1 Preparation of Expression Pre-culture

Day 1

1. Mix about 50–100 ng of *E. coli* expression plasmid for FlaXc (pSVA1911 having C-terminal His_6–tag Δ37 FlaX in pSA4 [8], *see* **Note 1**) and FlaH (pSVA2100 having N-terminal His_6–tag in pET Duet1 [6] and *see* **Note 2**) with 50 µL chemical competent *E. coli* BL21 DE3 Ril cells (preparation of *E. coli* chemical competent cells [9]) and transform using heat shock transformation protocol [10].
2. Use LB agar-amp-cmp to plate the transformed cells and incubate at 37 °C for 16 h.

Day 2

1. Examine the plate for single colony formation.
2. Pick a single colony for further protein expression experiment. Inoculate the single colony into pre-warmed 50 mL LB-amp-cmp.
3. Incubate the pre-culture at 37 °C shaker incubator under constant shaking (~150 rpm) for 16 h.

3.2 Protein Production

3.2.1 FlaXc Expression (Induction at 37 °C for 3 h)

Day 3

1. Inoculate the pre-culture in 2 L LB-amp-cmp to reach an initial OD_{600} of 0.05.
2. Grow the culture at 37 °C in a shaker incubator shaking at 150 rpm. Follow the OD_{600} until it has reached 0.5.
3. Induce protein expression using 0.3 mM isopropyl β-D-thiogalactopyranoside (IPTG). Continue cell cultivation for 3 h at 37 °C and 150 rpm shaker speed.

4. Harvest cells by centrifugation at 4000 × *g* for 20 min and proceed with FlaXc purification or otherwise freeze the cell pellet in liquid nitrogen and store in −80 °C until further use.

3.2.2 FlaH Expression (Induction at 18 °C for 16 h)

Day 3

1. Inoculate the pre-culture in 2 L LB-amp-cmp to reach an initial OD_{600} of 0.05.
2. Grow the culture at 37 °C in a shaker incubator shaking at 150 rpm. Follow the OD_{600} until it has reached 0.5.
3. Cool cells on ice for 30 min prior to induction at OD_{600} 0.5. Induce FlaH overproduction with 0.5 mM IPTG and continue growing the cells overnight at 18 °C in shaker incubator with a constant rotation of 150 rpm.

Day 4

1. Harvest the cells by centrifugation at 4000 × *g* for 20 min and either proceed with FlaH purification or freeze the FlaH overexpressed cell pellet in liquid nitrogen and store it in −80 °C until further use.

3.3 Protein Purification

3.3.1 FlaXc Purification

1. Resuspend obtained cells in 5 mL of lysis buffer per gram of pellet wet weight.
2. Disrupt cells containing FlaXc by sonication using Bandelin Sonopuls HD3100, KE-76 probe with 50% amplitude and 15 s interval between each pulse for 30 min.
3. Collect the supernatant with centrifugation at 10,000 × *g* for 30 min (*see* **Note 3**).
4. Optional heat step: incubate the supernatant at 70 °C for 20 min to precipitate *E. coli* proteins. Separate precipitated from soluble proteins by using centrifugation at 10,000 × *g* for 30 min. Proceed with the supernatant (*see* **Note 4**).
5. Pipette 1 mL Ni-NTA beads into gravity chromatography columns and equilibrate with 5 mL of deionized water and 5 mL of 50 mM Tris-HCl pH 8, 150 mM NaCl buffer.
6. Apply cell lysate (or heat-step supernatant) to the Ni-NTA column bed to enable FlaXc His-tag binding to Ni-NTA beads.
7. Wash the column with 10 mL of 25 mM glycine-NaOH pH 10, 200 mM NaCl buffer (*notice buffer change*).
8. Perform stepwise protein elution using 25 mM glycine-NaOH pH 10, 200 mM NaCl buffer containing increasing concentration of imidazole from 25 mM to 500 mM, in five consecutive steps (25, 100, 200, 300, 500 mM) and 2 mL volume each. Collect the elution in separate 2 mL Eppendorf reaction tubes and store at room temperature. Pure FlaXc should elute after 100 mM imidazole.

9. Check the protein purity on a Coomassie-stained SDS-PAGE [11].
10. Dialyze [12] pure FlaXc protein overnight against 25 mM glycine-NaOH pH 10, 200 mM NaCl buffer to remove imidazole.
11. Determine protein concentration using BCA assay (or comparable colorimetric assay) and store the pure protein at −80 °C after freezing it in liquid nitrogen.

3.3.2 FlaH Purification

1. Resuspend obtained cells in 5 mL 50 mM Hepes pH 7.2, 150 mM NaCl per gram of cell pellet wet weight. Additionally, add traces of DNase I and incubate the suspension on ice for 30 min for DNA digestion.
2. Lyse cells containing FlaH by sonication using Bandelin Sonopuls HD3100, KE-76 probe with 50% amplitude and 15 s interval between each pulse for 30 min.
3. Centrifuge the broken cells at 4600 × *g* for 20 min to separate the cell lysate and cell debris.
4. To collect the soluble lysate, centrifuge the supernatant at 20,000 × *g* for 20 min at 4 °C.
5. Pipette 1 mL Ni-NTA beads into a gravity chromatography column and equilibrate with 5 mL of deionized water and subsequently with 5 mL of 50 mM Hepes pH 7.2, 150 mM NaCl buffer. Pass the supernatant through pre-equilibrated 1 mL Ni-NTA beads.
6. Wash the column with 10 mL of 50 mM Hepes pH 7.2, 150 mM NaCl containing 20 mM imidazole.
7. Elute the protein using 50 mM MES pH 6, 150 mM NaCl (*notice buffer change*) containing 300 mM imidazole in ten 1 mL fractions.
8. *Heat step*: To obtain pure FlaH protein, perform a heat step by incubating the elution fractions at 50 °C in a water bath for 20 min followed by a short centrifugation on a tabletop centrifuge for 15 min at 10,000 × *g* (*see* **Note 5**).
9. Dialyze the heat stable FlaH protein overnight against 50 mM MES pH 6, 150 mM NaCl.
10. Estimate purity of the protein using Coomassie-stained SDS-PAGE [11].
11. Store the pure protein at −80 °C after freezing it in liquid nitrogen.

3.4 In Vitro Reconstitution of FlaXc-FlaH Complex

1. Precipitate purified FlaH using 80% ammonium sulfate (*see* **Note 6**).
2. Collect the protein precipitate using centrifugation on a tabletop centrifuge for 15 min at 10,000 × *g*.

3. Resuspend the white pellet in 500 μL of 25 mM glycine-NaOH pH 10, 200 mM NaCl buffer.
4. Dialyze the resuspended pellet overnight against 25 mM glycine-NaOH pH 10, 200 mM NaCl buffer at room temperature to remove residual ammonium sulfate.
5. Collect the sample from the dialysis tube and centrifuge at room temperature in a tabletop centrifuge for 15 min at 10,000 × *g*.
6. Check the concentration of regenerated FlaH samples by BCA assay and mix equimolar concentrations of FlaXc and FlaH to initiate FlaXc-FlaH complex formation.

4 Notes

1. The soluble domain of FlaX was used for the co-complex formation as full-length expression was not achieved.
2. FlaH gene we used in this study was codon optimized. The sequence can be obtained from the authors.
3. Usage of higher centrifugation speeds is not advisable for FlaXc purification. The protein is capable of forming large complexes and thus might be pelleted down at speeds greater than 10,000 × *g*.
4. The indicated heat step is not required for obtaining pure protein after the purification. However, it reduces the number of *E. coli* proteins in the lysate applied to the gravity column, therefore allowing for higher flow rates during chromatography.
5. In contrast to FlaXc purification, the heat step for the FlaH protocol is crucial for protein purity.
6. FlaH is most stable at pH 6, and accordingly, this condition is used for elution and storage. However, pH 10 is essential for FlaXc ring formation. Therefore, prior to mixing the proteins for complex formation, the buffer for FlaH is changed to pH 10 using ammonium sulfate precipitation and eventual resuspending in glycine-NaOH pH 10.

Acknowledgment

P.C. received funding from an ERC starting grant (ARCHAELLUM, 311523), and P.T. was supported by funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 686647 (MARA).

References

1. Shahapure R, Driessen RPC, Haurat MF, Albers SV, Dame RT (2014) The archaellum: a rotating type IV pilus. Mol Microbiol 91:716–723
2. Jarrell KF, Albers SV (2012) The archaellum: an old structure with a new name. Trends Microbiol 20:307–312
3. Streif S, Staudinger WF, Marwan W, Oesterhelt D (2008) Flagellar rotation in the archaeon *Halobacterium salinarum* depends on ATP. J Mol Biol 384:1–8
4. Banerjee A, Tsai CL, Chaudhury P, Tripp P, Arvai AS, Ishida JP, Tainer JA, Albers S-V (2014) FlaF is a β-sandwich protein that anchors the archaellum in the archaeal cell envelope by binding the S-layer protein. Structure 23:863–872
5. Reindl S, Ghosh A, Williams GJ, Lassak K, Neiner T, Henche A-L, Albers S-V, Tainer JA (2013) Insights into FlaI functions in archaeal motor assembly and motility from structures, conformations, and genetics. Mol Cell 49:1069–1082
6. Chaudhury P, Neiner T, D'Imprima E, Banerjee A, Reindl S, Ghosh A, Arvai AS, Mills DJ, van der Does C, Tainer JA, Vonck J, Albers S-V (2016) The nucleotide-dependent interaction of FlaH and FlaI is essential for assembly and function of the archaellum motor. Mol Microbiol 99:674–685
7. Banerjee A, Neiner T, Tripp P, Albers S-V (2013) Insights into subunit interactions in the *Sulfolobus acidocaldarius* archaellum cytoplasmic complex. FEBS J 280:6141–6149
8. Banerjee A, Ghosh A, Mills DJ, Kahnt J, Vonck J, Albers S-V (2012) FlaXc, a unique component of the crenarchaeal archaellum, forms oligomeric ring-shaped structures and interacts with the motor ATPase FlaI. J Biol Chem 287:43322–43330
9. Inoue H, Nojima H, Okayama H (1990) High efficiency transformation of *Escherichia coli* with plasmids. Gene 96:23–28
10. Froger A, Hall JE (2007) Transformation of plasmid DNA into *E. coli* using the heat shock method. J Vis Exp 6:e253
11. Shapiro AL, Viñuela E, V Maizel J (1967) Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels. Biochem Biophys Res Commun 28:815–820
12. Craig LC (1967) Techniques for the study of peptides and proteins by dialysis and diffusion. Methods Enzymol 11:870–905

Chapter 20

Unstable Protein Purification Through the Formation of Stable Complexes

Sylvia Eiler, Nicolas Levy, Benoit Maillot, Julien Batisse, Karine Pradeau Aubreton, Oyindamola Oladosu, and Marc Ruff

Abstract

Purification of proteins containing disordered regions and participating in transient complexes is often challenging because of the small amounts available after purification, their heterogeneity, instability, and/or poor solubility. To circumvent these difficulties, we set up a methodology that enables the production of stable complexes in large amounts for structural and functional studies. In this chapter, we describe the methodology used to establish the best cell culture conditions and buffer compositions to optimize soluble protein production and their stabilization through protein complex formation. Two examples of challenging protein families are described, namely, the human steroid nuclear receptors and the HIV-1 pre-integration complexes.

Key words Protein complex, Nuclear receptor, GR, TIF2, HIV, Integrase, Pre-integration complex

1 Introduction

Protein flexibility and disorder have been shown to be inherent properties of major protein families [1], particularly those involved in large transient complexes [2]. The intrinsic disorder is often an asset that explains the ability of such proteins to interact with multiple partners and to perform multiple functions [3]. Each function is related to a unique structure that results from interaction with small ligands, DNA, RNA, or partner proteins [4]. Purification of such proteins is often challenging because of the small amounts available, their heterogeneity, instability, or poor solubility. Despite standard procedures to optimize protein solubility and stability, some proteins remain very unstable and difficult to purify. For example, proteins containing intrinsically disordered domains tend to aggregate rather than form folded, stable proteins. Although the addition of chemicals in the solubilizing buffer may improve

Sylvvia Eiler and Nicolas Levy contributed equally in this work

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_20,

solubility, disordered domains fold only when they are in complex with their partner molecules. In this chapter, the aim is to describe how to produce and purify proteins that are largely unstable/insoluble when expressed alone, but that can be stabilized when present in a complex with other partners. To study such proteins, we developed a new methodology that enables the production of stable and functional complexes of proteins and/or protein domains in large amounts, allowing structural and functional studies. We describe the methods used for the optimization of cell culture, methods for deciphering the best solubilizing and stabilizing buffers, as well as methods for protein complex production through two examples: (1) the ligand binding domain of the human glucocorticoid nuclear receptor (hGR-LBD) in complex with a domain of the human transcriptional intermediary factor 2 (hTIF2) containing the 3 LXXLL motifs [5] and (2) the HIV-1 integrase (IN) in complex with its cellular partner LEDGF and a domain of the INI1 protein, part of the SWI/SNF complex [5–7]. In the first example, we had to develop a protocol different from the one previously described for the purification of steroid nuclear receptors [8–10]. Here we describe the production and purification of the hGR-LBD/TIF2 complex produced by co-expression of the two partners in the same cell. In the second example, we describe the production and purification of the IN/LEDGF and IN/LEDGF/INI1 complexes by solubilizing and purifying individual partner proteins in presence of solubilizing agents, which can be removed by dialysis for in vitro complex reconstitution.

2 Materials

Prepare all solutions using ultrapure water and analytical grade reagents.

Material and solutions must be sterile for cell culture.

The buffers used for purification are filtered through a 0.45 μm filter before use.

2.1 E. coli Cell Culture

1. LB medium.
2. LB medium agar plate with ampicillin (100 μg/mL) and kanamycin (50 μg/mL).
3. Small Petri plates: Gosselin Round Petri plate with vents H14, diameter 90 mm.
4. Big Petri plates: Greiner BioOne Petri Dish 145 × 20 mm with vents.
5. Sucrose 50%: Weigh 500 g sucrose in a 1 L graduated cylinder or a 1 L glass beaker and add water. When the sucrose is dis-

solved, make up to 1 L with water. Sterilize the solution by filtration through a 0.22 micron Stericup filter unit.

6. Ampicillin at 100 mg/mL: The stock solution is made by dissolving 1 g of ampicillin in 9 mL of deionized H_2O. After the antibiotic has dissolved, adjust volume of the solution to 10 mL with deionized H_2O and sterilize by filtration through a 0.22 micron filter.
7. Kanamycin at 50 mg/mL: Prepare the stock solution by dissolving 0.5 g of kanamycin in 9 mL of deionized H_2O. After the antibiotic has dissolved, adjust volume of the solution to 10 mL with deionized H_2O and sterilize by filtration through a 0.22 micron filter.
8. 0.8 M IPTG: The stock solution is made by dissolving 1 g of isopropyl-1-thio-beta-D-galactopyranoside in 5 mL of deionized H_2O. After IPTG has dissolved, sterilize by filtration through a 0.22 micron filter.
9. 10 mM dexamethasone: Prepare the stock solution by dissolving 39.25 mg of dexamethasone in 10 mL of ethanol.
10. Chemically competent cells: *E. coli* BL21 (DE3) host strain (Invitrogen).
11. pET expression plasmids [11] adapted to the Gateway cloning system [12]:
    (a) NusA-(His)6-thrombin-GR_LBD triple mutant C638A, W557T, and W712S (residue 524–777) plasmid (ampicillin resistance).
    (b) (His)6-thrombin-TIF2 (623–772) (kanamycin resistance).
    (c) IN expression vector: GST-P3C-IN (1–288) or (His)6-P3C IN (1–288) (ampicillin resistance).
    (d) LEDGF expression vector: (His)6-P3C-LEDGF (1–530) (ampicillin resistance).
    (e) INI1 expression vector: (His)6-Tev-INI1-IBD (174–289) described in [6] (ampicillin resistance).

2.2 Purification

1. Affinity column (GR/TIF2): 5 mL HiTrap Chelating column. The column is supplied free of metal ions and must be charged with the metal ion (0.1 M $ZnCl_2$ solution) before use (*see* **Note 1**).
2. 0.1 M $ZnCl_2$: weigh 13.6 g $ZnCl_2$ in a 1 L graduated cylinder and add 900 mL water. When the $ZnCl_2$ is dissolved, make up to 1 L with water. The pH of the solution should not exceed 5 (*see* **Note 2**).
3. Lysis and affinity binding buffer (GR/TIF2): 50 mM Na/K phosphate buffer pH 7.5, 250 mM NaCl, 10 mM dexamethasone, 10 mM beta-mercaptoethanol.

4. Affinity elution buffer (GR/TIF2): 50 mM Na/K phosphate buffer pH 7.5, 250 mM NaCl, 10 mM dexamethasone, 10 mM beta-mercaptoethanol, 500 mM imidazole.
5. Affinity column (IN, LEDGF, and INI1-IBD): 5 mL HisTrap FF Crude column. The column is supplied already charged with metal ions ($NiSO_4$).
6. Sonicator VibraCell 72412 with a sonication probe of 13 mm diameter and a temperature probe.
7. Lysis and affinity binding buffer (IN/LEDGF): 1 M NaCl, 7 mM CHAPS, 50 mM Hepes pH 7.5, 2 mM $MgCl_2$, 2 mM beta-mercaptoethanol.
8. Lysis and affinity binding buffer (INI1-IBD): 2 M NaCl, 20 mM CHAPS, 50 mM Hepes pH 7,5, 2 mM $MgCl_2$, 2 mM beta-mercaptoethanol, 5% glycerol.
9. Affinity elution buffer (IN, LEDGF, and INI1): lysis buffer + 500 mM imidazole.
10. Diisopropylfluorophosphonate (DIFP) at 100 mM.
11. Phenylmethanesulfonyl fluoride (PMSF) at 100 mM.
12. Centriprep concentration unit with a cutoff of 30 kDa for GR/TIF2 and 100 kDa for IN/LEDGF.
13. Dialysis bag: we routinely used dialysis bag (Standard RC tubing) with a 6–8 kDa molecular weight cutoff (MWCO) that allows removal of both salt and CHAPS.
14. Dialysis buffers (IN/LEDGF) with step baths: initial buffer (1 M NaCl, 7 mM CHAPS, 50 mM Hepes pH 7.5, 2 mM $MgCl_2$, 2 mM beta-mercaptoethanol); intermediate buffer (600 mM NaCl, 2 mM CHAPS, 50 mM Hepes pH 7.5, 2 mM $MgCl_2$, 2 mM beta-mercaptoethanol); and final dialysis buffer (500 mM NaCl, No CHAPS, 50 mM Hepes pH 7.5, 2 mM $MgCl_2$, 2 mM beta-mercaptoethanol).
15. Dialysis buffers (IN/LEDGF/INI1-IBD) with continuous dialysis: 1 L initial buffer (2 M NaCl, 20 mM CHAPS, 50 mM Hepes pH 7.5, 2 mM $MgCl_2$, 2 mM beta-mercaptoethanol, 5% glycerol) and 7 L final dialysis buffer (500 mM NaCl, No CHAPS, 50 mM Hepes pH 7.5, 2 mM $MgCl_2$, 2 mM beta-mercaptoethanol).
16. Peristaltic pump with dual variable flow.
17. Gel filtration column: HiLoad 16/60 Superdex 200 prep grade.
18. Gel filtration buffer (GR/TIF2): 50 mM Tris-HCl pH 8.0, 250 mM NaCl, 10 mM dexamethasone, 10 mM beta-mercaptoethanol.

19. Gel filtration buffer (IN/LEDGF/INI1): final dialysis buffer (*see* above).
20. Thrombin at 1 U/μL: thrombin from bovine plasma >1500 U/mg is dissolved in buffer 25 mM Tris pH 8.5, 50% glycerol at a concentration of 1 U/μL.
21. TEV protease at a final concentration of 1.5 mg/mL, His-Tag recombinant purified in house.
22. P3C protease at a final concentration of 1.5 mg/mL, GST-Tag recombinant purified in house.
23. Anion exchange column: 1 mL HiTrap Q.
24. IEX dilution buffer (GR/TIF2): 10 mM Tris pH 8.5, 10 mM dexamethasone, 5 mM beta-mercaptoethanol.
25. IEX binding buffer (GR/TIF2): 10 mM Tris pH 8.5, 10 mM NaCl, 10 mM dexamethasone, 5 mM beta-mercaptoethanol.
26. IEX elution buffer (GR/TIF2): 10 mM Tris pH 8.5, 1 M NaCl, 10 mM dexamethasone, and 5 mM beta-mercaptoethanol.
27. Mass spectrometry buffer (GR/TIF2): 50 mM ammonium acetate pH 8.5, 10 mM dexamethasone.
28. Zeba™ Spin Desalting Columns, 7K MWCO, 2 mL.
29. Gel filtration for the validation of complex formation (IN/LEDGF/INI1-IBD): Superose 6 10/300.
30. General basic buffer (GBB): 150 mM NaCl, 50 mM pH buffer, and additives depending on the protein.
31. SDS-PAGE equipment: gel cassette, tank, power generator, running gel buffer, stacking gel buffer, migration buffer, acrylamide, bis-acrylamide, SDS solution, Laemmli buffer, Coomassie Blue staining solution, destaining solution [13].

3 Methods

3.1 In Cellulo Nuclear Receptor Complex Production of GR/TIF2

3.1.1 GR/TIF2 Co-expression

1. Transformation: Put 20 ng GRtm LBD plasmid and 20 ng TIF2 plasmid in an Eppendorf tube and store on ice. Add 50 μL of *E. coli* BL21 (DE3) chemically competent chilled cells to the tube with the DNAs. Incubate 20 min on ice. Heat shock for 60 s at 42 °C in water bath. Leave for 2 min on ice. Add 0.5 mL LB without antibiotics. Incubate for 1 h at 37 °C. Plate 100 μL of transformed cells on a small LB medium agar plate with ampicillin (100 μg/mL) and kanamycin (50 μg/mL). Incubate overnight at 37 °C. After incubation check that there are isolated colonies on the plate.
2. Pre-culture: Put 1 mL LB medium with ampicillin (100 μg/mL) and kanamycin (50 μg/mL) in a 14 mL sterile falcon

tube. Use a sterile pipette tip or toothpick and select a single colony from the freshly streaked plate. Drop the tip or toothpick into the liquid LB with antibiotics and swirl. Incubate bacterial culture at 37 °C for 8–12 h in a shaking incubator. After incubation check for growth: the solution should be cloudy. Plate 1 mL of pre-culture on a big LB medium agar plate with ampicillin (100 μg/mL) and kanamycin (50 μg/mL). Incubate overnight at 37 °C. After incubation, there should be a uniform cell sheet. Add 10 mL LB medium to the plate and scrap off all colonies. This suspension will be used to seed the culture medium.

3. Culture: This protocol has been optimized using the procedures described in **Note 3**.

   Prepare 1 L LB medium in a 5 L flask and cover loosely with aluminum foil. Autoclave and allow to cool to room temperature. Add 250 mL sterile 50% sucrose. Add ampicillin (50 μg/mL final concentration), kanamycin (25 μg/mL final concentration), and dexamethasone (10 μg/mL final concentration) (*see* **Note 4**). Add the freshly prepared 10 mL pre-culture suspension. The cells are incubated at 37 °C up to an OD600nm of 0.6. The culture is then slowly cooled to 18 °C before adding 0.5 mM isopropyl-1-thio-beta-D-galactopyranoside and grown overnight. Cells are harvested by centrifugation at 4 °C at 4000 × *g*.

3.1.2 GR/TIF2 Complex Purification

Protein purity is analyzed by SDS-PAGE and protein concentrations measured by UV absorption at 280 nm.

1. Cell lysis: the cells are homogenized in the affinity binding buffer (the composition of the buffer has been optimized using the procedure described in **Note 5**) with a ratio of 10 mL buffer for 1 g cells. The cells are lysed by pulse sonication for 10 min (2 s on/2 s off) on ice with a temperature probe to keep the extract below 10 °C. Diisopropylfluorophosphonate (DIFP) (0.1 mM) and phenylmethanesulfonyl fluoride (PMSF) (0.1 mM) are added before, during, and after sonication. The extract is centrifuged at 100,000 × *g* for 1 h. The supernatant is separated from the pellet and further used for purification.
2. Purification: the purification is performed in a three-step procedure in the presence of dexamethasone—zinc affinity, gel filtration, and anion exchange chromatography. The crude extract is loaded onto a 5 mL zinc affinity column (HiTrap Chelating). Nonspecifically bound proteins are removed by 10 column volumes wash in affinity binding buffer. Elution is performed in a 15 column volumes imidazole gradient from 0 (affinity binding buffer) up to 0.5 M (affinity elution buffer). The sample is then concentrated on Centriprep (MWCO

30 kD) and analyzed on SDS-PAGE. The complex is further purified on a gel filtration Superdex 200 column equilibrated in the gel filtration buffer. Endoproteolytic cleavage of the fusion proteins is achieved using one unit of thrombin per milligram of fusion substrate and incubating at 4 °C overnight. The completeness of the proteolytic reaction is assessed by SDS-PAGE. Following the digestion step, the sample is diluted with IEX dilution buffer to decrease the salt concentration to 10 mM and further purified by anion exchange chromatography on HiTrap Q. The column is equilibrated in IEX binding buffer. A salt gradient from 10 mM IEX binding buffer to 1 M NaCl (IEX elution buffer) over 20 column volumes is used for elution. Peak fractions corresponding to the complex are pooled and concentrated by ultrafiltration. With this procedure, one can obtain 0.5 mg complex from a 3 L culture.

3.1.3 GR/TIF2 Complex Validation by Mass Spectrometry Analysis

All studies are performed using an electrospray time-of-flight mass spectrometry (ESI-TOF) mass spectrometer. The protein samples are submitted to buffer exchange in 50 mM ammonium acetate pH 8.5, 10 mM dexamethasone (*see* **Note 6**). The sample is continuously infused into the ion source at a flow rate of 4 mL/min using a Harvard Model 1 L syringe pump.

3.2 In Vitro Complex Reconstitution (IN/LEDGF, IN/LEDGF/INI1-IBD Complexes)

3.2.1 Production of Isolated Proteins

1. Transformation: Put 20 ng IN plasmid in an Eppendorf tube and store on ice. Add 50 μL of *E. coli* BL21 (DE3) chemically competent chilled cells to the tube with the DNA. Incubate 20 min on ice. Heat shock for 60 s at 42 °C in water bath. Leave for 2 min on ice. Add 0.5 mL LB without antibiotic. Incubate for 1 h at 37 °C. Plate 100 μL of transformed cells on a small LB medium agar plate with ampicillin (100 μg/mL). Incubate overnight at 37 °C. After incubation check that there are isolated colonies on the plate.
2. Pre-culture: Put 1 mL LB medium with ampicillin (100 μg/mL) in a 14 mL sterile falcon tube. Use a sterile pipette tip or toothpick and select a single colony from the freshly streaked plate. Drop the tip or toothpick into the liquid LB with antibiotics and swirl. Incubate bacterial culture at 37 °C for 8–12 h in a shaking incubator. After incubation check for growth: the solution should be cloudy. Plate 1 mL of pre-culture on a big LB medium agar plate with ampicillin (100 μg/mL). Incubate overnight at 37 °C. After incubation, there should be a uniform cell sheet. Add 10 mL LB medium to the plate and scrap off all colonies. This suspension will be used to seed the culture medium.
3. Culture: This protocol has been optimized using the procedures described in **Note 3**.

Prepare 1 L LB medium in a 5 L flask and cover loosely with aluminum foil. Autoclave and allow to cool to room temperature. Add 250 mL sterile 50% sucrose and ampicillin (50 μg/mL final concentration). Add the freshly prepared 10 mL preculture suspension. The cells are incubated at 37 °C up to an OD600nm of 0.6. The culture is then slowly cooled to 18 °C before adding 0.5 mM isopropyl-1-thio-beta-D-galactopyranoside and grown overnight. Cells are harvested by centrifugation at 4 °C at 4000 × *g*.

4. Repeat the **steps 1–3** for LEDGF and INI1-IBD.

3.2.2 Purification of Isolated Proteins

Single His-tagged proteins are expressed in *E. coli* and purified using immobilized metal ion affinity chromatography (IMAC) column, in 1 M NaCl/7 mM CHAPS (*see* Subheading 2 for more details) (Fig. 1a). This buffer is used during both lysis and purification, improving the protein solubilization and avoiding aggregation and/or precipitation: the high salt concentration increases the ionic strength, whereas the 7 mM CHAPS concentration is chosen to be above the CHAPS critical micelle concentration (CMC), allowing the CHAPS to form micelles.

3.2.3 In Vitro Assembly of Complexes

Partners predicted to form a complex are mixed together according to the expected stoichiometry and loaded in a dialysis bag with a molecular weight cutoff (MWCO) compatible with the molecular weight of the smallest subunit. For example, the IN/LEDGF complex is reconstituted mixing at a molar ratio 4:2 (4 IN for 2 LEDGF) according to previous studies [6, 7]. For this complex of 275 kDa, the MWCO should not be above 32 kDa, the size of the monomeric IN protein. To keep a slow exchange rate, we standardly use a cutoff of 6–8 kDa that is high enough to allow the release of both NaCl and CHAPS.

1. *Step dialysis (IN/LEDGF)* For the dialysis, the sample bag is immersed in a buffer volume equivalent to 50 sample volumes homogenized with a magnetic stirrer for at least 3 h at 4 °C. Several baths are done to gradually decrease the salt and the CHAPS concentration, allowing complex reconstitution: from an initial concentration of 1 M NaCl/7 mM CHAPS, we switch to a first bath of 600 mM NaCl/2 mM CHAPS. This CHAPS concentration is just below the CMC to softly switch from a high CHAPS concentration to no CHAPS at all. Omitting this step would most likely lead to the aggregation and/or precipitation of the complex members. Then, a final bath without CHAPS, at 500 mM NaCl, is used to keep the complex stable enough to allow functional and structural studies. Sample are then centrifuged (4000 × *g*; 10 min; 4 °C) to get rid of aggregated fraction that corresponds to single pro-

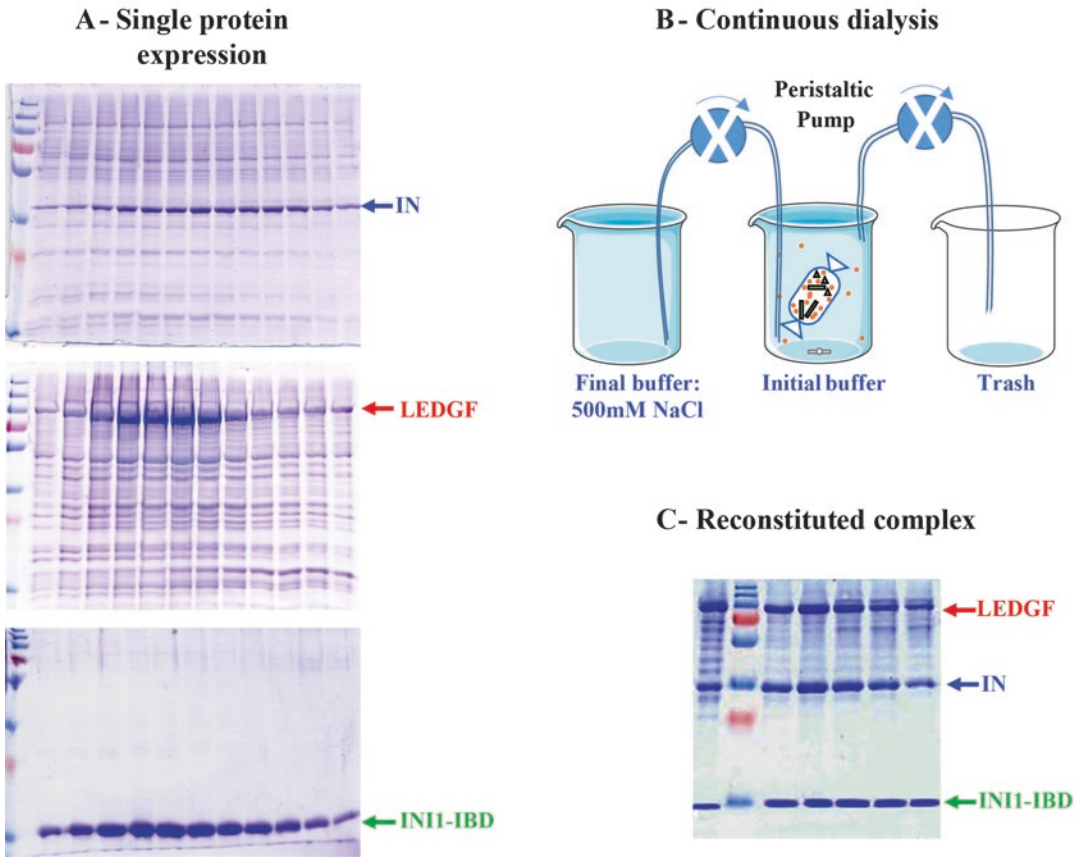


Fig. 1 IN/LEDGF/INI1-IBD purification and complex reconstitution. (**A**) Single purification of (His)6-IN; LEDGF; INI1-IBD: Coomassie blue-stained SDS-PAGE gels after affinity purification. (**B**) Scheme of the system used for continuous dialysis. (**C**) Reconstituted complex: Coomassie blue-stained SDS-PAGE gels after gel filtration

teins that have not associated into stable complexes and precipitate at such a final concentration of NaCl.

2. *Concentration and gel filtration* A final purification step is recommended to get a highly purified complex. The sample can be further purified either on an affinity column if a tag is still present on one complex member or on a gel filtration column to purify the expected complex according to its molecular weight. For this last purification strategy, the sample must be concentrated to a volume appropriate for the column used. Therefore, samples are concentrated on Amicon filters with a cutoff of 100 kDa according to the manufacturer's recommendation to concentrate high molecular weight complexes and remove any single protein in excess. A final round of purification on a gel filtration column is then performed in the final buffer to separate purified complex from aggregates and single proteins present in excess. Purity and dispersion of the complex into the collected fractions are assessed by SDS-PAGE and

Coomassie staining. Fractions of interest are either concentrated on Amicon filters (*see* above) or directly flash frozen in small aliquots in liquid nitrogen [14].

3. *Assembly of complexes via continuous flow dialysis (IN/LEDGF/INI1-IBD)* Some proteins require higher salt and detergent concentration to be solubilized. This is the case of the protein INI1, a member of the chromatin remodeling complex SWI/SNF. Such proteins are very difficult to purify, and even a small subdomain [INI1-IBD for integrase-binding domain—116 amino acids (174–289)] remains largely insoluble. To purify this domain, we use a concentration of 2 M NaCl and of 20 mM CHAPS. With this initial buffer, the reconstitution of the complex by batch dialysis would take too many baths to gradually decrease both NaCl and CHAPS concentrations from 2 M and 20 mM to 500 mM and no CHAPS, respectively. To integrate such proteins into complexes together with other partners, we use a specific continuous flow dialysis method. The dialysis bag is immerged into 1 L of initial buffer (2 M NaCl and 20 mM CHAPS). Using a peristaltic pump with dual flow, 5 L of final buffer (500 mM NaCl; no CHAPS) is gradually added directly into the initial bath with a 5 mL/min flow rate (Fig. 1b) to allow a slow exchange. An additional dialysis step in 2 L of final buffer is performed to get rid of any traces of CHAPS. As described above, the reconstituted complex is further purified on a gel filtration column equilibrated in the final buffer to recover a highly pure protein complex (Fig. 1c).

3.2.4 Validation of IN/LEDGF/INI1-IBD Complex Formation by Gel Filtration

To monitor the state of the complex between the different steps of the process (before dialysis, after dialysis, and after purification of the complex), a sample of each step is loaded on a size exclusion chromatography column for comparison.

If complex formation is successful, new high molecular mass species should appear after dialysis and should be further enriched after complex purification. An example is shown in Fig. 2.

A Superose 6 10/300 GL gel filtration column is equilibrated with 2 CV of the buffer corresponding to the analyzed step (pre-dialysis, post-dialysis, post-purification). A sample of the protein mix, typically 200 μg in a volume of 400 μl, is applied on the column at a flow rate of 0.5 mL/min.

Elution profiles are recorded and superimposed in the Unicorn software, or data were exported and plotted in Microsoft Excel. The comparison of the three different chromatograms shows the appearance and enrichment of high molecular mass (non-aggregates) species illustrating the formation of the complexes that can be confirmed by the SDS-PAGE analysis of the co-purified partners.

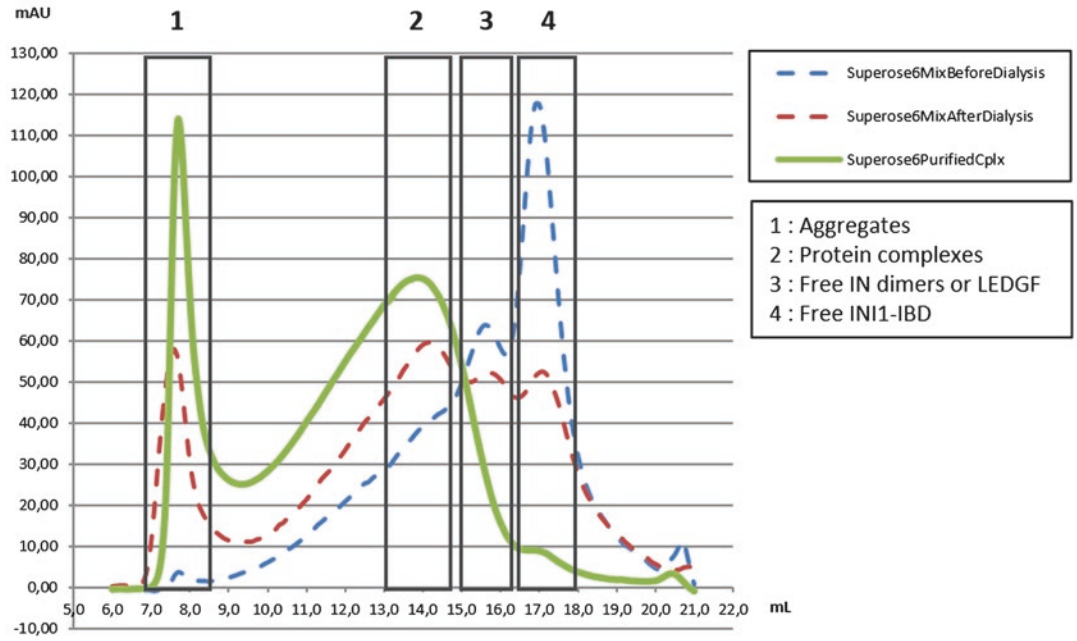


Fig. 2 Gel filtration validation of the reconstitution of IN/LEDGF/INI1-IBD complex. Before dialysis (blue dashed curve), the free proteins IN or LEDGF (same elution volume, peak 3) and INI1-IBD (peak 4) are eluted at the expected volume corresponding to the apparent molecular mass of the single partners. After dialysis (red dashed curve), the amount of free partners is diminished, and the appearance of higher molecular mass protein complex is monitored in peak 2. In the same time, proteins that are unable to associate in a stable complex and aggregate during dialysis are found in the excluded fraction (peak 1). After purification of the complexes (green curve), the high molecular mass species (peak 2) are enriched, and the free proteins (peaks 3 and 4) are eliminated

4 Notes

1. Use of $ZnCl_2$ rather than $NiSO_4$ for affinity column because of the high concentration of reducing agent in the buffer.
2. 0.1 M $ZnCl_2$ with pH < 5 otherwise formation of insoluble $Zn(OH)_2$.
3. *E. coli* cell culture optimization: the cell culture parameters are optimized for soluble protein expression. The parameters tested are the temperature for induction (37 °C, 30 °C, 25 °C, 20 °C, 18 °C), the duration of induction (1 h, 3 h, 6 h, overnight), cell density at the time of induction (determined by the OD of the culture at 600 nm), and the media (LB, TB, ZYM, auto-induction media), with or without 10% sucrose and/or any other specific additives (*see* **Note 7**). To reduce the number of combination, the most standard conditions are tested first (Table 1). For the first tests, protein production is induced by addition of IPTG once the culture has reached 0.8 OD. The LB media with sucrose is often used in combination with an

Table 1
Standard conditions for protein expression

Temperature of induction (°C)	37	25	25	18	18
Sucrose (g/100 mL)	0	0	10	0	10
Duration of induction (h)	3	6	6	Overnight	Overnight

induction temperature below 25 °C. Before IPTG induction, an aliquot is taken, centrifuged, and suspended in Laemmli buffer so as to adjust the OD to 10 [approximately 80 million of cells (*see* **Note 8**)]. After heating 10 min at 95 °C, the samples are centrifuged and kept at 4 °C or frozen before SDS-PAGE analysis. If the temperature of induction is different from the temperature of growth, the temperature of the incubator needs to be set at the right temperature (25 °C or 18 °C) before the culture reaches the density for induction (it takes 60 min for 1 L culture, *see* **Note 9**). The induction is done by adding IPTG at 500 μM final concentration. Before harvesting, an aliquot is again taken, centrifuged, and suspended in Laemmli buffer to adjust the OD to 10. After heating 10 min at 95 °C, the samples are centrifuged (total extract). The cell pellets are resuspended in the general basic buffer (GBB) with a ratio of 10% (w/v, 1 mL of buffer for 0.1 g of cells). Lysis is performed by pulsed sonication (2 s on/2 s off) on ice for 1 min per gram of cells. The total extract is centrifuged at 100,000 × *g* for 1 h. An aliquot of supernatant (soluble extract) and pellet (insoluble extract) are added to Laemmli buffer. After heating 10 min at 95 °C, the tubes are centrifuged. The samples of uninduced culture, total extract, soluble extract, and insoluble extract after induction are loaded on the same SDS-PAGE (run at 100 V). After Coomassie staining, the gel band intensities are analyzed with the software ImageJ [15].

4. For co-expression from different vectors with different antibiotic resistance in liquid medium, it is necessary to reduce the amount of antibiotics and sometimes to express proteins in absence of antibiotics.

5. Optimization of the buffer composition for protein solubility: to optimize protein solubility, several buffer compositions are tested for cell lysis. Key parameters such as ionic force (salt concentration: from low salt 50 mM to high salt 1 M), pH (type of buffer and pH: below and above the theoretical pI of the protein of interest), and detergents (CHAPS, NP40, DDM) are tested. A strong solubilizing detergent (Zwittergent 3–14) is used as a positive control.

   The pellet is thawed and homogenized with a ratio between volume of lysis buffer and weight of cells of 10% (10 mL of lysis buffer for 1 g of cells). Lysis is performed by pulsed

sonication (2 s on/ 2 s off) on ice in the lysis buffer. The size of the sonication probe (3 mm, 13 mm) and the related amplitude is chosen depending of sample volume (below 15 mL: 3 mm probe). The amount of cells determines the duration of the sonication, usually 1 min of effective sonication for 1 g of cells. The extract (total extract) is centrifuged at 100,000 × *g* for 1 h. The supernatant (soluble extract) is separated from the pellet (insoluble extract) and analyzed by SDS-PAGE. After the addition of Laemmli buffer and heat denaturation (10 min at 99 °C), 2 μL of the total extract, soluble extract, and the pellet are loaded on SDS-PAGE. The SDS-PAGE is run at 100 V (limiting parameter) to improve the separation of the proteins.

After staining, the SDS-PAGE gel is scanned and analyzed with the ImageJ software [15]. The comparison of the intensity of the band of the protein of interest between the total extract and the supernatant indicates the solubility efficiency of the lysis buffer. The ratio between the intensities of the band of interest in the soluble and total extract lanes represents the solubilizing efficiency of the lysis buffer tested. A possible follow-up is to load the supernatants on affinity beads (GST; 6XHIS; STREP, FLAG). After extensive washing and elution, all fractions are analyzed by SDS-PAGE. If no suitable buffer is found, other strategies are implemented based on co-expression or co-lysis to directly purify the protein complex.

6. The buffer exchange before mass spectrometry analysis is performed on Zeba™ Spin Desalting Columns. The column is equilibrated in 50 mM ammonium acetate pH 8.5, 10 mM dexamethasone. The sample is applied and centrifuged for 2 min at 1000 × *g*. The desalted sample is recovered in the flow-through.
7. Specific additive (depending on the protein), for example, zinc finger motif protein, requires addition of zinc cation in the media, and GR ligand (dexamethasone) is added to the cell culture medium.
8. To calculate the cell concentrations from the turbidity measurement, we use OD600 of 1.0 = 8×10^8 cells/mL.
9. Reduce the growth temperature slowly; otherwise there will be a large amount of chaperones co-expressed (from 37 °C to 20 °C in a minimum of 1 h).

Acknowledgments

This work was supported by grants from the CNRS, the INSERM, SIDACTION, and the French National Agency for Research against AIDS (ANRS), the support and the use of resources of the

French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05 and of Instruct, a Landmark ESFRI project. We wish to thank Robert Drillien (IGBMC) for his help and for useful suggestions about the manuscript. We would like to thank the members of the IGBMC Structural Biology and Genomics platform, the IGBMC cloning service headed by Paola Rossolillo, and the members of the IGBMC common services for their contribution.

References

1. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM (2014) Classification of intrinsically disordered regions and proteins. Chem Rev 114(13):6589–6631. https://doi.org/10.1021/cr400525m
2. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6(3):197–208. https://doi.org/10.1038/nrm1589
3. Dunker AK, Bondos SE, Huang F, Oldfield CJ (2015) Intrinsically disordered proteins and multicellular organisms. Semin Cell Dev Biol 37:44–55. https://doi.org/10.1016/j.semcdb.2014.09.025
4. Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, Gibson TJ, Davey NE (2014) Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev 114(13):6733–6778. https://doi.org/10.1021/cr400585q
5. Levy N, Eiler S, Pradeau-Aubreton K, Maillot B, Stricher F, Ruff M (2016) Production of unstable proteins through the formation of stable core complexes. Nat Commun 7:9. https://doi.org/10.1038/ncomms10932
6. Maillot B, Levy N, Eiler S, Crucifix C, Granger F, Richert L, Didier P, Godet J, Pradeau-Aubreton K, Emiliani S, Nazabal A, Lesbats P, Parissi V, Mely Y, Moras D, Schultz P, Ruff M (2013) Structural and functional role of INI1 and LEDGF in the HIV-1 preintegration complex. PLoS One 8(4):14. https://doi.org/10.1371/journal.pone.0060734
7. Michel F, Crucifix C, Granger F, Eiler S, Mouscadet JF, Korolev S, Agapkina J, Ziganshin R, Gottikh M, Nazabal A, Emiliani S, Benarous R, Moras D, Schultz P, Ruff M (2009) Structural basis for HIV-1 DNA integration in the human genome, role of the LEDGF/P75 cofactor. EMBO J 28(7):980–991. https://doi.org/10.1038/emboj.2009.41
8. Cura V, Gangloff M, Eiler S, Moras D, Ruff M (2008) Cleaved thioredoxin fusion protein enables the crystallization of poorly soluble ER alpha in complex with synthetic ligands. Acta Crystallogr Sect F Struct Biol Cryst Commun 64:54–57. https://doi.org/10.1107/s1744309107066444
9. Gangloff M, Ruff M, Eiler S, Duclaud S, Wurtz JM, Moras D (2001) Crystal structure of a mutant hER alpha ligand-binding domain reveals key structural features for the mechanism of partial agonism. J Biol Chem 276(18):15059–15065. https://doi.org/10.1074/jbc.M009870200
10. Eiler S, Gangloff M, Duclaud S, Moras D, Ruff M (2001) Overexpression, purification, and crystal structure of native ER alpha LBD. Protein Expr Purif 22(2):165–173. https://doi.org/10.1006/prep.2001.1409
11. Studier FW (2014) Stable expression clones and auto-induction for protein production in E. coli. Methods Mol Biol 1091:17–32. https://doi.org/10.1007/978-1-62703-691-7_2
12. Busso D, Delagoutte-Busso B, Moras D (2005) Construction of a set gateway-based destination vectors for high-throughput cloning and expression screening in Escherichia coli. Anal Biochem 343(2):313–321. https://doi.org/10.1016/j.ab.2005.05.015
13. J. Sambrook, E.F. Fritsch, T. Maniatis (1982) Molecular cloning: a laboratory manual, vol 1–3. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press
14. Deng J, Davies DR, Wisedchaisri G, Wu M, Hol WG, Mehlin C (2004) An improved protocol for rapid freezing of protein samples for long-term storage. Acta Crystallogr D Biol Crystallogr 60(Pt 1):203–204
15. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. Nat Methods 9(7):671–675

Chapter 21

Expressing Multi-subunit Complexes Using biGBac

Florian Weissmann and Jan-Michael Peters

Abstract

The reconstitution of recombinant protein complexes is facilitated by methods that allow coexpression of their subunits from a single vector. Here we describe a detailed step-by-step protocol for the biGBac cloning method which can be used to generate baculoviral transfer vectors coding for up to 25 subunits of a protein complex (Weissmann et al., Proc Natl Acad Sci U S A 113(19):E2564–E2569, 2016). biGBac is based on Gibson assembly reactions, optimized DNA linker sequences, and uses a hierarchical two-step assembly procedure. In the first assembly step, up to five expression cassettes are combined to generate a polygene cassette. In the second step, up to five polygene cassettes can then be combined to generate transfer vectors coding for up to 25 subunits.

Key words Protein complex, Baculovirus-insect cell expression, BEVS, Multigene expression, Gibson assembly

1 Introduction

Many cellular processes depend on multi-subunit protein complexes [1]. When individual subunits of a protein complex are coexpressed in heterologous cell systems, they often assemble into functional protein complexes that can be purified and then used for structural and functional studies. The baculovirus-insect cell system is an invaluable tool for the production of recombinant proteins and protein complexes that cannot be easily produced in *E. coli*, for example, because they might require chaperone and post-translational modification systems provided by eukaryotic host cells. Typically, the coding sequences for proteins to be expressed are cloned into expression cassettes on a baculoviral transfer vector before these expression cassettes are transferred onto a baculoviral genome using Tn7 transposition [2]. The yield and homogeneity of protein complex preparations can be improved if all subunits are expressed from a single baculoviral vector rather than from combinations of single subunit baculoviral vectors [3]. Prominent transfer vectors are the pFastBac vectors onto which one or two genes

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764,
https://doi.org/10.1007/978-1-4939-7759-8_21,

can be cloned. For the coexpression of several subunits, the MultiBac series of transfer vectors are particularly useful tools that utilize a range of cloning techniques including conventional restriction-ligation cloning, sequence- and ligation-independent cloning (SLIC) combined with Cre-LoxP recombination of "acceptor" and "donor" vectors (tandem recombineering), and uracil-specific excision reagent (USER) cloning [4–8].

Here we describe a detailed protocol for the biGBac cloning method [9] that uses Gibson assembly reactions [10] to generate baculoviral transfer vectors coding for up to 25 subunits of a protein complex. The biGBac assembly procedure (Fig. 1) uses three assembly levels corresponding to three types of biGBac cloning vectors (Fig. 2).

On the first level, the cDNAs of interest are individually cloned into gene expression cassettes (GECs) on the transfer vector pLIB (library vector). The GEC consists of the subunit cDNA flanked by a polyhedrin promoter for high expression levels in baculovirus-infected insect cells and a transcriptional terminator sequence.

On the second level, the GECs are amplified by PCR from pLIB templates using predefined primer sets (Table 1) that introduce DNA linker sequences for Gibson assembly at the fragment ends (Fig. 1a). Up to five PCR-amplified GECs are combined in a Gibson assembly reaction with a linearized pBIG1 vector to create a transfer vector coding for up to five subunits. This procedure introduces SwaI restriction sites between individual GECs and PmeI sites flanking the generated polygene cassette (PGC) for convenient analysis of clones (Fig. 1b). Digestion by PmeI additionally leads to the appearance of new linker sequences on the fragment ends of the PGC. This multi-GEC assembly step can be performed in five different versions of the pBIG1 vector—distinguished by the letters a, b, c, d, and e—that differ only in the linker sequences next to the PmeI sites (*see* Fig. 2b).

On the third level, up to five PGCs derived from different pBIG1 constructs can be combined after PmeI digestion with a compatible linearized pBIG2 vector (*see* Fig. 2c) in a Gibson assembly reaction to create a transfer vector that can encode up to 25 subunits. Products of this multi-PGC assembly step can be analyzed by SwaI digestion cleaving between individual GECs or by PacI digestion cleaving between PGCs (Fig. 1c).

All three types of biGBac vectors (pLIB, pBIG1, pBIG2) contain Tn7L and Tn7R sites as well as a gentamicin resistance marker, which can be used to generate recombinant baculoviral genomes using Tn7 transposition (Fig. 2). The linker sequences used in biGBac have been carefully selected and tested for high assembly efficiency and specificity in the Gibson assembly steps used in biGBac [9]. These linker sequences are encoded on the vector backbones and the predefined primer set.

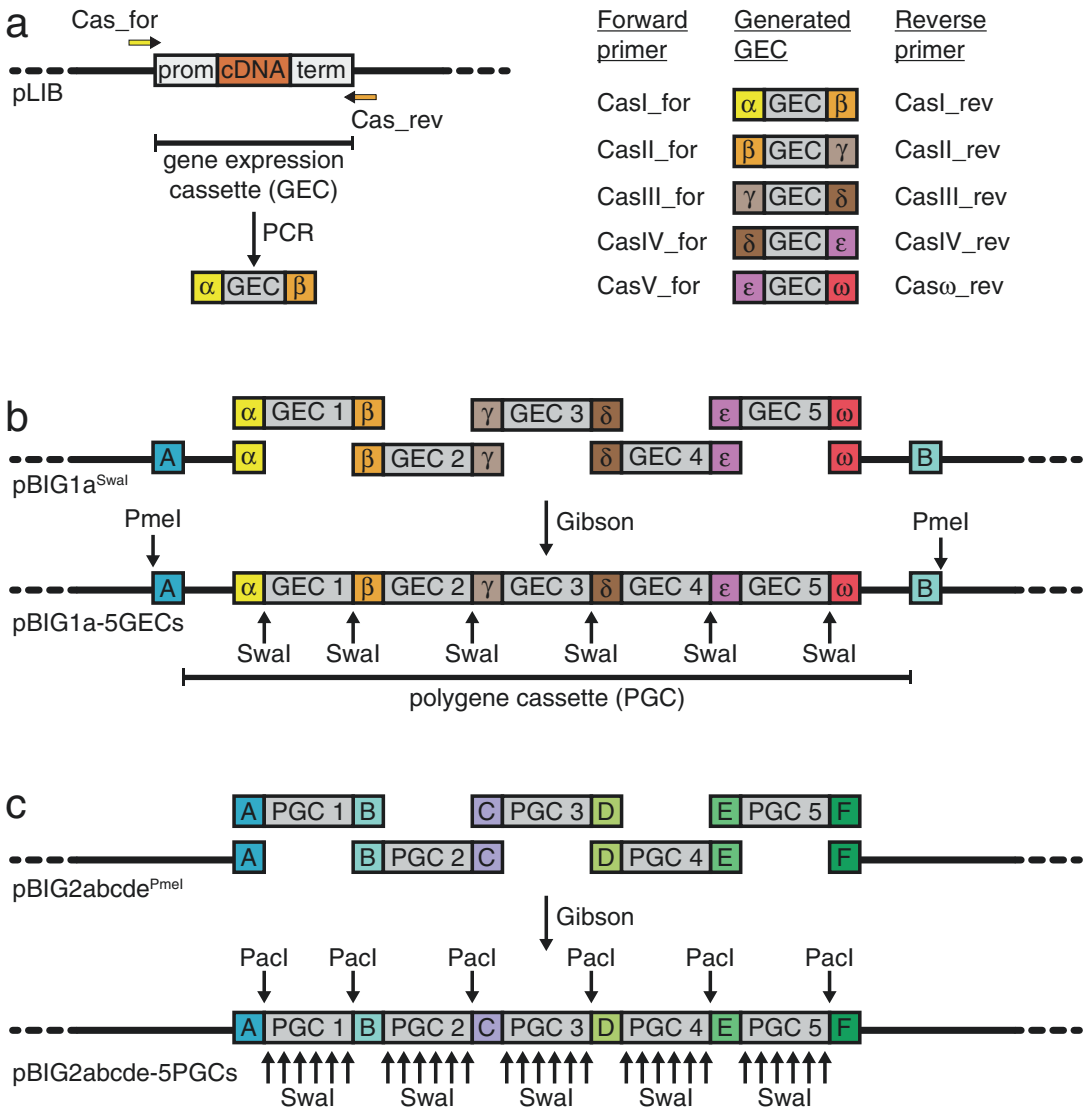


Fig. 1 biGBac assembly procedure. (**a**) Each cDNA is cloned into a gene expression cassette (GEC) in the pLIB vector. A GEC consists of a polyhedrin promoter (prom), the coding sequence of the protein to be expressed (cDNA), and a transcriptional terminator sequence (term). GECs are amplified by PCR from pLIB templates using predefined primer sets (Cas_for/Cas_rev primers; *see* Table 1) that introduce linker sequences for Gibson assembly (Greek letters). The "last" GEC of a pBIG1 assembly should carry the "omega" linker sequence (Casω_rev) to create an overlap with the pBIG1 vector. (**b**) In the first assembly step, up to five PCR-amplified GECs containing suitable linker sequences are combined in a Gibson assembly reaction with a linearized pBIG1 vector generating a polygene cassette (PGC). pBIG1 constructs are analyzed by SwaI digestion to release individual GECs or by PmeI digestion to release the PGC. After PmeI digestion the released PGC contains new linker sequences on the fragment ends (indicated as A, B, …). (**c**) In the second assembly step, up to five PGCs from different pBIG1 constructs are released by PmeI digestion and combined in a Gibson assembly reaction with a linearized pBIG2 vector. Generated pBIG2 constructs are analyzed by SwaI digestion to release individual GECs or by PacI digestion to release PGCs

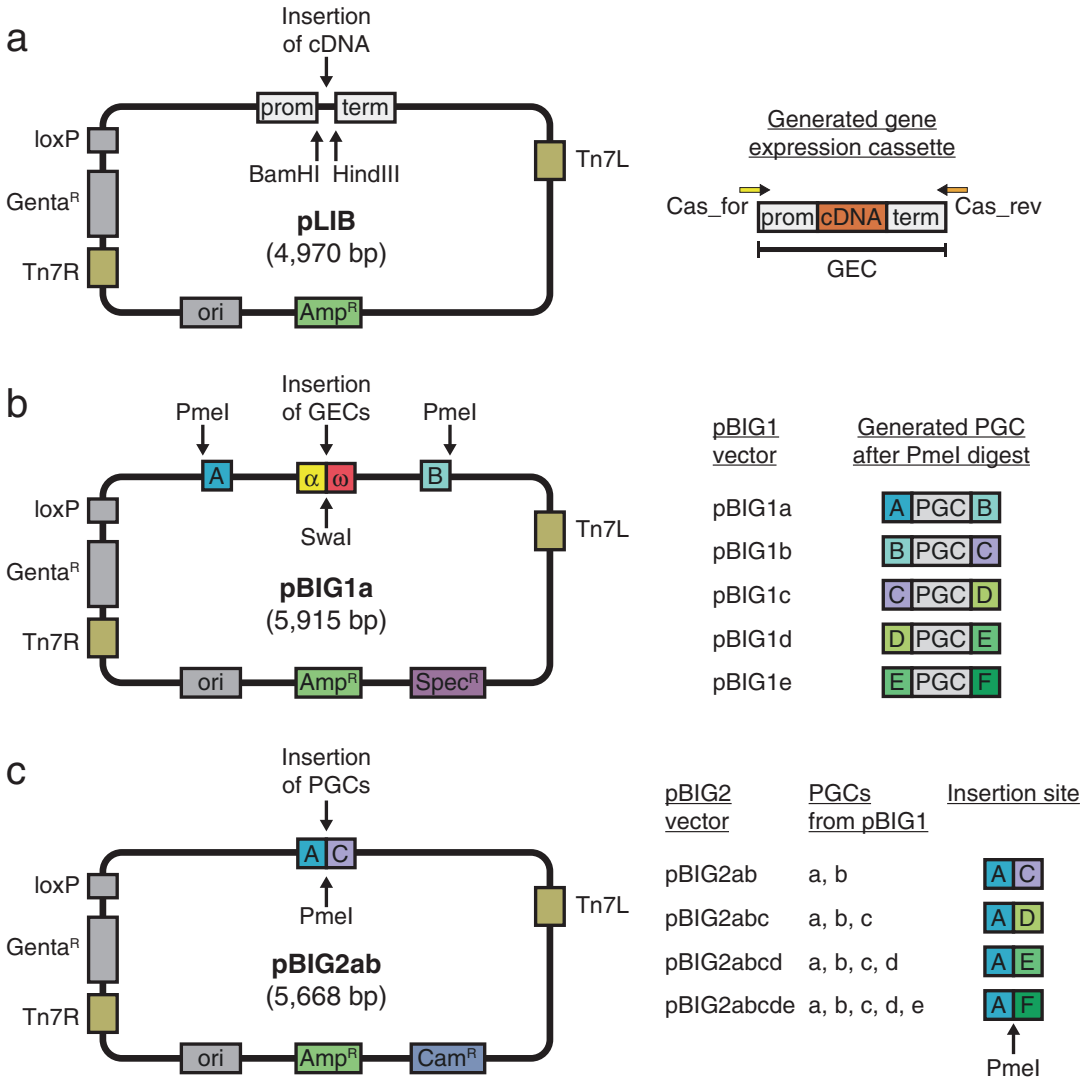


Fig. 2 Schematic representation of biGBac cloning vectors. (**a**) pLIB vector level: the coding sequence of a subunit (cDNA) is cloned into the BamHI-/HindIII-linearized pLIB vector, which generates a gene expression cassette (GEC) consisting of polyhedrin promoter (prom), cDNA, and transcriptional terminator sequence (term). Generated pLIB constructs can be used as templates for GEC amplification and multigene construct generation in pBIG1 or for transfer of its GEC to a baculoviral genome using Tn7 transposition (Tn7R, Tn7L sites and gentamicin resistance). pLIB is maintained with ampicillin (Amp^R). (**b**) pBIG1 vector level: amplified GECs from pLIB templates are inserted into SwaI-linearized pBIG1 vectors. All pBIG1 vectors contain the "alpha" and "omega" linker sequences for GEC insertion (α, ω) at the SwaI linearization site. PmeI digestion of pBIG1 constructs releases the generated polygene cassette (PGC), and new linker sequences (A, B, C, D, E, or F) appear on the fragment ends depending on which of the five pBIG1 vectors a, b, c, d, or e was used. Generated pBIG1 constructs can be used to generate larger multigene assemblies in pBIG2 vectors or to transfer its GECs onto a baculoviral genome using Tn7 transposition (Tn7R, Tn7L sites, and $Genta^R$). Spectinomycin resistance ($Spec^R$) is used as selection marker for pBIG1 construct generation. (**c**) PGCs from different pBIG1 constructs with compatible linker sequences are inserted into PmeI-linearized pBIG2 vectors. The name of the pBIG2 vector (ab, abc, …) indicates which pBIG1-derived PGCs can be combined. Generated pBIG2 constructs can be used to transfer its GECs onto a baculoviral genome using Tn7 transposition (Tn7R, Tn7L, and $Genta^R$). Chloramphenicol resistance (Cam^R) is used as selection marker for pBIG2 construct generation

Table 1
Predefined primer set for GEC amplification

Primer	Sequence
CasI_for	AACGCTCTATGGTCTAAAGATTTAAATCGACCTACTCCGGAATATTAATAGATC
CasI_rev	AAACGTGCAATAGTATCCAGTTTATTTAAATGGTTATGATAGTTATTGCTCAGCG
CasII_for	AAACTGGATACTATTGCACGTTTAAATCGACCTACTCCGGAATATTAATAGATC
CasII_rev	AAACATCAGGCATCATTAGGTTTATTTAAATGGTTATGATAGTTATTGCTCAGCG
CasIII_for	AAACCTAATGATGCCTGATGTTTAAATCGACCTACTCCGGAATATTAATAGATC
CasIII_rev	AAACTAAGCTATGTGAACCGTTTATTTAAATGGTTATGATAGTTATTGCTCAGCG
CasIV_for	AAACGGTTCACATAGCTTAGTTTAAATCGACCTACTCCGGAATATTAATAGATC
CasIV_rev	AAACCAAGTCAATGTCAGTGTTTATTTAAATGGTTATGATAGTTATTGCTCAGCG
CasV_for	AAACACTGACATTGACTTGGTTTAAATCGACCTACTCCGGAATATTAATAGATC
Casω_rev	AACCCCGATTGAGATATAGATTTATTTAAATGGTTATGATAGTTATTGCTCAGCG

The biGBac cloning method has been used successfully to reconstitute protein complexes such as the anaphase-promoting complex/cyclosome, mitotic checkpoint complex, cohesin, and kinetochore complexes [9, 11–13]. In principle, biGBac can be used for the coexpression of any proteins, not only for subunits of a protein complex, for example, to coexpress a chaperone to assist in protein folding or to coexpress a posttranslational modification enzyme with its substrate.

2 Materials

2.1 Generation of biGBac Multigene Transfer Vectors

1. biGBac vectors (Addgene kit #1000000088): pLIB, pBIG1a, pBIG1b, pBIG1c, pBIG1d, pBIG1e, pBIG2ab, pBIG2abc, pBIG2abcd, and pBIG2abcde.
2. 2× Gibson assembly master mix (e.g., NEB) (*see* **Note 1**).
3. Restriction endonucleases: BamHI, HindIII, PacI, PmeI, and SwaI.
4. Standard PCR reagents and equipment: thermocycler and high-fidelity DNA polymerase (e.g., Phusion polymerase, Thermo Fisher Scientific).
5. Predefined DNA oligonucleotide set for GEC amplification (e.g., Microsynth; PAGE-purified quality recommended): five Cas_for and five Cas_rev primers (Table 1).
6. Standard agarose gel electrophoresis reagents and equipment.
7. Miniprep and gel extraction kits.

8. PureLink PCR Purification Kit (Thermo Fisher Scientific).
9. Spectrophotometer.
10. Competent cells of a standard *E. coli* cloning strain (e.g., DH10B) (*see* **Note 2**).
11. Media and antibiotics for growing *E. coli* cultures: LB medium, LB-agar plates, ampicillin (use at 100 μg/ml), spectinomycin (use at 50 μg/ml), and chloramphenicol (use at 34 μg/ml).
12. Sequencing primers P1 (5′-TCAACAGGTTGAACTGCT GATC-3′) and P2 (5′-GGTGTAGCGTCGTAAGCTAAT AC-3′) and gene-specific sequencing primers.

2.2 Generation of Recombinant Baculoviruses from biGBac Transfer Vectors Using Tn7 Transposition

1. *E. coli* competent cells for site-specific transposition onto baculoviral genome with Tn7 (e.g., DH10EMBacY [14]).
2. Blue-white selection plates: LB-agar, 50 μg/ml X-gal, 0.1 mM IPTG, 50 μg/ml kanamycin, 10 μg/ml tetracycline, and 7 μg/ml gentamicin.
3. Isopropanol and ethanol.
4. Sf9 insect cells and cell culture media (Thermo Fisher Scientific).
5. Transfection reagent (e.g., Fugene 6, Promega).

3 Methods

We recommend, in particular for large complexes, to plan how to distribute subunits between biGBac vectors. It is recommended to first generate linearized stocks of the biGBac cloning vectors (*see* Subheading 3.1) and to clone each subunit into the pLIB vector to generate a "library" of pLIB constructs, each containing one subunit. Any additional DNA sequences required, for example, those encoding affinity tags or protease recognition sites, should be introduced at this stage (*see* Subheading 3.2). In the first assembly step, up to five subunits can be combined on one pBIG1 vector (*see* Subheading 3.3). It can be useful to collect subcomplexes on pBIG1 vectors that might be expressed directly from pBIG1-derived baculoviruses or to put subunits that will be mutagenized onto a separate pBIG1 vector. Up to five PGCs from different pBIG1 constructs can be combined on a pBIG2 vector (*see* Subheading 3.4). To determine which pBIG1 vectors are compatible with which pBIG2 vectors and to plan how to distribute subunits, refer to Fig. 2. Another consideration is that pBIG2 assemblies are typically more efficient than pBIG1 assemblies, in particular if pBIG1 assemblies combine several large subunits. To reduce the number of clones that need to be analyzed, it can be useful to distribute the largest subunits onto several pBIG1 constructs and to reduce the number of subunits per pBIG1 construct.

3.1 Preparation of Linearized Cloning Vectors

1. Linearize the pLIB vector by digestion with BamHI/HindIII: mix 5 µl 10× FastDigest buffer, 30 µl pLIB plasmid DNA (~300 ng/µl), 2 µl BamHI, 2 µl HindIII, and 11 µl water, and incubate at 37 °C for 3 h. Gel-purify using a 0.7% agarose gel and a gel extraction kit.
2. Linearize the five pBIG1 vectors by digesting each pBIG1 vector with SwaI: mix 5 µl 10× NEBuffer 3.1, 30 µl pBIG1 plasmid DNA (~300 ng/µl), 1 µl SwaI, and 14 µl water, and incubate at 25 °C overnight. Add another 2 µl SwaI for 2 h. Heat-inactivate SwaI at 65 °C for 20 min and purify the linearized plasmid using a PCR purification kit (*see* **Note 3**).
3. Linearize the four pBIG2 vectors by digesting each pBIG2 vector with PmeI: mix 5 µl 10× CutSmart buffer, 30 µl pBIG2 plasmid DNA (~300 ng/µl), 1 µl PmeI, and 14 µl water, and incubate at 25 °C overnight. Add another 2 µl PmeI for 2 h. Purify the linearized plasmid using a PCR purification kit.

3.2 Cloning of pLIB Library Constructs

1. PCR-amplify the cDNA of a protein complex subunit using oligonucleotide primers containing the 5′ extensions 5′-CCACCATCGGGCGCGGATCCA... (followed by start-ATG and gene-specific sequences) for the forward primer and 5′-TCCTCTAGTACTTCTCGACAAGCTT... (followed by reverse complement of stop codon and gene-specific sequences) for the reverse primer. In case the cDNA sequence contains a PmeI recognition site (GTTTAAAC), amplify the cDNA as two overlapping PCR products choosing a Gibson overhang that introduces a silent mutation to the PmeI site (*see* **Note 4**). At this stage, tags can be introduced to subunits (*see* **Note 5**).
2. Purify the PCR products by gel extraction from a 0.7% (w/v) agarose gel using a gel extraction kit, elute DNA in 30 µl EB buffer and determine DNA concentration using a spectrophotometer.
3. Prepare the Gibson assembly reaction by mixing 100 ng BamHI-/HindIII-digested pLIB vector DNA (*see* Subheading 3.1, **step 1**) and a 5× molar excess of PCR-amplified cDNA in a volume of 5 µl (*see* **Note 6**). Add water to 10 µl total volume. Preheat a thermocycler block to 50 °C with heated lid.
4. Perform the Gibson assembly reaction by adding 10 µl 2× Gibson assembly master mix on ice and mix by pipetting up and down. Immediately (*see* **Note 7**) transfer the tube to the preheated 50 °C thermocycler and incubate for 1 h (*see* **Note 8**).
5. Mix the finished Gibson assembly reaction by flicking the tube and transform a standard *E. coli* cloning strain such as DH10B (*see* **Note 2**). Recover the transformed bacteria in LB medium at 37 °C for 60 min, spread on LB-ampicillin agar plates and incubate at 37 °C overnight.

6. Pick 2–6 colonies per pLIB construct and grow 5 ml cultures at 37 °C in LB-ampicillin medium overnight and isolate plasmid DNA using a miniprep kit.
7. Digest 1.5 µl of plasmid DNA (~80–500 ng/µl) with BamHI/HindIII: mix 1 µl 10× FastDigest buffer, 0.2 µl FastDigest BamHI, 0.2 µl FastDigest HindIII, 1.5 µl DNA, and 7.1 µl water, and incubate at 37 °C for 1–3 h, and analyze by electrophoresis on a 0.8% (w/v) agarose gel.
8. Sequence clones that show the expected restriction pattern by Sanger sequencing using sequencing primers P1 and P2 and gene-specific primers.

3.3 Cloning of pBIG1 Constructs Coding for Up to Five Subunits

1. Choose up to five pLIB constructs containing the cDNAs of subunits to be combined on a pBIG1 vector, and use them as templates in PCR reactions to amplify linker sequence-flanked GECs: mix 20 µl 5× HF buffer, 2 µl dNTPs (10 mM each), 0.5 µl CasX_for primer (100 µM) (*see* Table 1), 0.5 µl CasX_rev primer (100 µM) (*see* Table 1), 0.5 µl pLIB template, 1 µl Phusion HF polymerase, and 75.5 µl water. Use the Casω_rev primer for the "last" cassette to generate a Gibson overlap to the vector (*see* Fig. 1b). Set up the PCR reactions on ice and run the following PCR program: 95 °C for 3 min, 42× (98 °C for 15 s, 65 °C for 20 s, 72 °C for 30 s/kb), 72 °C for 5 min; store at 4 °C (*see* **Note 9**).
2. Purify the PCR-amplified GECs using the PureLink PCR Purification Kit. Mix 400 µl high-cutoff binding buffer B3 with 100 µl PCR product. Perform the purification according to the manufacturer's instructions. Elute in 30 µl E1 buffer. Determine DNA concentration using a spectrophotometer (*see* **Note 10**). Confirm successful amplification and purity of the GECs by running 3 µl aliquots on a 0.8% (w/v) agarose gel.
3. Prepare a Gibson assembly reaction by mixing 100 ng SwaI-digested pBIG1 vector (*see* Subheading 3.1, **step 2**), a 5× molar excess of each PCR-amplified GEC, and water to a volume of 10 µl (*see* **Note 11**). Preheat a thermocycler block to 50 °C with heated lid.
4. Perform a Gibson assembly reaction by adding 10 µl 2× Gibson assembly master mix on ice, and mix by pipetting up and down. Immediately (*see* **Note 7**) transfer the tube to the preheated 50 °C thermocycler, and incubate for 1 h (*see* **Note 8**).
5. Mix the finished reaction thoroughly by flicking the tube, and transform a standard *E. coli* cloning strain like DH10B (*see* **Note 2**). Recover the transformed bacteria in LB medium at 37 °C for 60 min, spread on LB-spectinomycin agar plates, and incubate at 37 °C overnight.

6. Pick six colonies per pBIG1 construct, grow 5 ml cultures in LB-spectinomycin medium at 37 °C overnight, and isolate plasmid DNA using a miniprep kit.
7. Analyze clones by two separate restriction digests: digest each clone with SwaI to release individual expression cassettes from the vector backbone by mixing 1 μl 10× NEB 3.1 buffer, 1.2 μl DNA (typically ~100–500 ng/μl), 0.5 μl SwaI (10 U/μl), and 7.3 μl water. Incubate at 25 °C for 2 h. Separately, digest each clone with PmeI to release the generated polygene cassette from the vector backbone using 1 μl 10× NEB CutSmart buffer, 1.2 μl DNA, 0.2 μl PmeI (10 U/μl), and 7.6 μl water. Incubate at 25 °C for 2 h. Analyze both restriction digests on a 0.8% agarose gel (*see* Fig. 3a).
8. Sequence clones that show the expected restriction patterns (*see* **Note 12**) by Sanger sequencing using gene-specific primers.

3.4 Cloning of pBIG2 Constructs Coding for Up to 25 Subunits

1. Choose up to five compatible pBIG1 constructs containing the GECs to be combined and a suitable pBIG2 vector (e.g., pBIG1a, pBIG1b, pBIG1c can be combined on pBIG2abc; *see* Fig. 2), and set up a PmeI digest to release the polygene cassettes (*see* **Note 13**): mix 33 ng PmeI-digested pBIG2 vector (*see* Subheading 3.1, **step 3**) with a 5× molar excess of each pBIG1 plasmid to be combined in a volume of 8 μl (*see* **Note 14**). Add 1 μl NEB CutSmart buffer and 1 μl PmeI and mix by pipetting. Incubate the digestion at 37 °C for 90 min.
2. Preheat a thermocycler block to 50 °C. Perform the Gibson assembly reaction by adding 10 μl 2× Gibson assembly master mix on ice and mix by pipetting up and down. Immediately transfer the tube to the preheated 50 °C thermocycler block and incubate for 1 h (*see* **Notes 7** and **8**).
3. Mix the finished assembly reaction thoroughly by flicking the tube, and transform a standard *E. coli* cloning strain such as DH10B (*see* **Note 2**). Recover transformed bacteria in LB medium at 37 °C for 60 min, spread on LB-chloramphenicol agar plates, and incubate at 37 °C overnight.
4. Pick 2–6 colonies per pBIG2 construct, grow 5 ml cultures in LB-chloramphenicol medium at 37 °C overnight, and isolate plasmid DNA using a miniprep kit.
5. Analyze clones by two separate restriction digests: digest each clone with SwaI to release individual expression cassettes by mixing 1 μl 10× NEB 3.1 buffer, 2.5 μl DNA (typically ~100–500 ng/μl), 1 μl SwaI (10 U/μl), and 5.5 μl water. Incubate at 25 °C for 2 h. Separately, digest each clone with PacI to release pBIG1-derived polygene cassettes by mixing 1 μl 10× NEB CutSmart buffer, 0.8 μl DNA, 0.5 μl PacI (10 U/μl), and

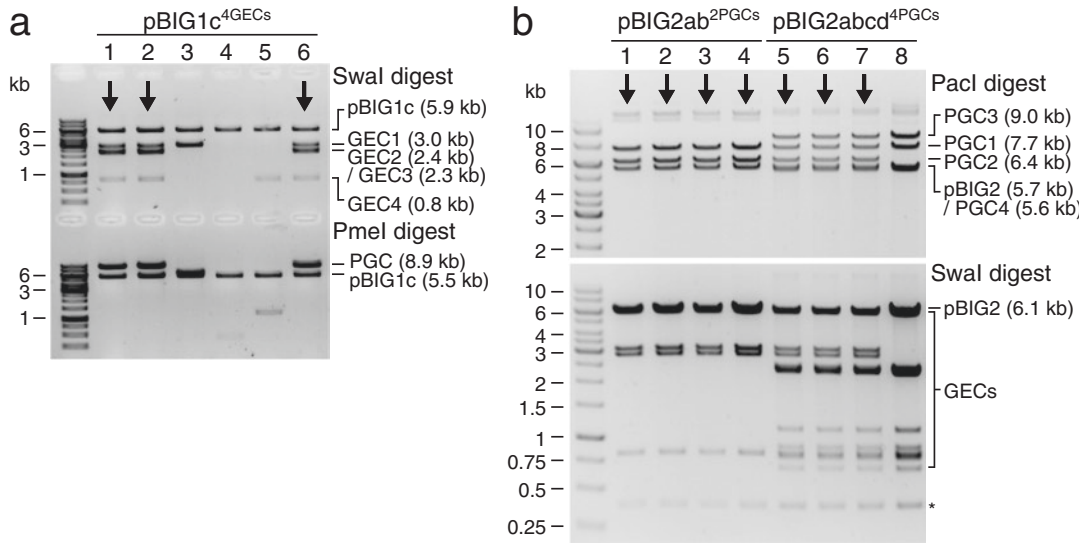


Fig. 3 Restriction analysis of pBIG1 and pBIG2 constructs. (**a**) Analytical digests of six clones of a 4 GEC–pBIG1 assembly (pBIG1c:APC3/APC6/APC7/APC12) were analyzed on a 0.8% agarose gel. Clones 1, 2, and 6 show the presence of all 4 GECs in the SwaI digest as well as a PGC of the expected size in the PmeI digest (correct clones indicated by arrows). (**b**) Analytical digests of four clones of a 2 PGC (4 GECs)–pBIG2ab assembly (lanes 1–4) and four clones of a 4 PGC (12 GECs)–pBIG2abcd assembly (lanes 5–8) coding for APC/C subcomplexes. Clones 1–4 show the presence of two PGCs derived from pBIG1a and pBIG1b constructs in the PacI digest, as well as the expected restriction pattern in the SwaI digest. Clones 5–7 show bands corresponding to four PGCs derived from pBIG1a, pBIG1b, pBIG1c, and pBIG1d constructs at the expected sizes in the PacI digest, as well as the expected restriction pattern in the SwaI digest (correct clones indicated by arrows). Clone 8 is incorrect, because it does not show a band for PGC2 in the PacI digest and is missing the corresponding bands in the SwaI digest. Note that SwaI digests of pBIG2 constructs show an additional band at 370 bp (indicated by an asterisk) that does not represent a GEC but corresponds to sequences between PGCs

7.7 μl water. Incubate at 37 °C for 2 h. Analyze the SwaI digest on a 1.0% agarose gel and the PacI digest on a 0.6% agarose gel (*see* Fig. 3b). Select clones that show the expected restriction patterns. Note that SwaI digestion of pBIG2 constructs gives an additional band at 370 bp, corresponding to sequences between the polygene cassettes.

3.5 Generation of Recombinant Baculoviruses from biGBac Constructs

Any biGBac construct (pLIB, pBIG1, pBIG2) can be used to generate recombinant baculoviruses using Tn7 transposition. Here, a protocol is given for Tn7 transposition of cassettes from biGBac vectors onto the EmBacY baculoviral genome, which contains YFP as a fluorescent reporter [14].

1. Transform competent DH10-EmBacY cells with a generated biGBac construct (pLIB-, pBIG1-, or pBIG2-derived). Electroporation is recommended for constructs with a plasmid size >15–20 kb. Recover transformed bacteria in LB medium at 37 °C for 6 h. Plate the recovered bacteria on blue-white selection LB-agar plates containing X-gal, IPTG, kanamycin,

tetracycline, and gentamicin. Incubate at 37 °C for 24–48 h until blue and white colonies can be clearly distinguished. Grow a white colony in LB medium containing kanamycin, tetracycline, and gentamicin at 37 °C for 16 h.

2. Isolate bacmid DNA by alkaline lysis using buffers from a Miniprep kit and isopropanol precipitation: resuspend the bacterial pellet in 250 μl P1, lyse with 250 μl P2, and neutralize with 350 μl N3. Clear the lysate by centrifugation at 20,000 × *g* for 10 min. Mix 750 μl supernatant with 750 μl isopropanol, and invert and incubate on ice for 5 min. Pellet the DNA precipitate by centrifugation at 20,000 × *g*, 20 min, 4 °C. Wash the DNA pellet with 70% ethanol, and centrifuge at 20,000 × *g*, 5 min, 4 °C.
3. In a laminar flow hood, remove the supernatant, air-dry the DNA pellet, and resuspend in 40 μl sterile water. Transfect Sf9 insect cells, amplify the baculovirus, and express protein using standard procedures.

4 Notes

1. The Gibson assembly reactions can also be performed using a homemade 1.33× Gibson assembly master mix or by mixing individual components as described [9, 10].
2. The Gibson assembly reaction products from the biGBac assembly steps can be transformed into any standard *E. coli* cloning strain. For assembly products with a size >15 kb, electroporation is recommended, because chemical transformation can show a bias for smaller incorrectly assembled by-products. For chemical transformation, use, e.g., 5 μl reaction product to transform 50 μl chemically competent cells. For electroporation, use, e.g., 0.4 μl reaction product for 40 μl electrocompetent cells.
3. Complete linearization of biGBac cloning vector stocks is crucial to avoid empty vector colonies.
4. PmeI sites (GTTTAAAC) are rare, but must be removed if present in a cDNA sequence, because they are incompatible with the second assembly step. To remove a PmeI site from a cDNA sequence, a silent mutation can be introduced by cloning the cDNA into pLIB using two PCR fragments that overlap with a Gibson overhang (~20 bp; Tm > 50 °C), which contains the silent mutation. While PmeI is the only site that needs to be removed for the biGBac assembly procedure, it can be useful to check cDNA sequences at this point also for the presence of SwaI (ATTTAAAT) and PacI (TTAATTAA) sites, because they will produce additional fragments on analytical agarose

gels when analyzing multigene constructs later in the assembly procedure. There is no need to remove SwaI or PacI sites.

5. Small tags, e.g., His-, StrepII-, or Flag-tags, can typically be encoded on a primer. To introduce larger tags or fusion proteins, generate an additional PCR product that encodes the tag or fusion protein and overlaps with the cDNA sequence with a Gibson overhang (~20 bp; Tm > 50 °C). Separate PCR products should have a minimum size of ~100 bp.
6. When using gel-extracted DNA in a Gibson assembly reaction, we recommend to limit the amount of gel-extracted DNA to 5 μl in a 20 μl Gibson reaction to avoid recovering a low number of colonies. If necessary, scale down the amount of DNA accordingly without changing the vector/insert ratio.
7. The Gibson assembly reaction mixture should be set up on ice and transferred directly from ice to 50 °C without incubation on ice or room temperature. One of the enzymatic activities in Gibson assembly reactions, T5 exonuclease, creates single-stranded DNA ends that might form secondary structures at lower temperatures, which could reduce assembly efficiency.
8. Alternatively, instead of a 60-min incubation at 50 °C, a two-step incubation of 25 min at 50 °C and 10 min at 64 °C can be used. Taq ligase, one of the enzymatic activities in Gibson assembly reactions, is highly active at 64 °C.
9. The Cas_for and Cas_rev primers work best with an annealing temperature of 65 °C. Depending on the choice of thermocycler, it might be necessary to use a higher annealing temperature if unspecific PCR by-products are observed. It can be useful to choose a more precise temperature control mode in the settings of the PCR program to avoid short drops below the specified annealing temperature (e.g., the setting "Simulate Mastercycler gradient" for Eppendorf thermocyclers).
10. The high-cutoff binding buffer is used to avoid binding of potentially present primer dimers and short unspecific PCR products <300 bp to the column material. Binding to the column material also depends on the pH, which can vary depending on the used PCR buffer. Addition of 10 μl 3 M sodium acetate pH 5.0 to the PCR product before purification can improve recovery in combination with certain PCR buffers. Typical yields are 50–200 ng/μl in 30 μl elution volume. Elution efficiency might also be increased by preheating the elution buffer to 70 °C before applying it to the column.
11. Note that a GEC is 577 bp bigger than the inserted cDNA sequence. High DNA concentrations of the purified GECs are important for successful pBIG1 assemblies. If necessary, scale down the amount of DNA without changing the vector/

GECs ratio. For the pBIG1 assembly with four GECs shown in Fig. 3a, the following components were used.

Component	Generated by	Size (bp)	c (ng/μl)	$m_{5\times\ excess}$ (ng)	V (μl)
pBIG1c^{SwaI}	SwaI digest	5915	44	100	2.27
GEC1	CasI_for/CasI_rev	3052	226	258	1.14
GEC2	CasII_for/CasII_rev	2440	135	206	1.53
GEC3	CasIII_for/CasIII_rev	2275	132	192	1.46
GEC4	CasIV_for/Casω_rev	835	81	71	0.87
Water					2.73

12. Typically analyzing six clones of a pBIG1 assembly is sufficient to find correct clones. In some cases, it can be necessary to analyze more clones. If unspecific smaller PCR by-products seem to get incorporated into the assembly, try increasing the annealing temperature in the PCR and check the purity of the affected pLIB template. In certain cases, especially when several GECs are large and a correct clone is difficult to obtain, it can be useful to distribute the large GECs over different pBIG1 constructs, which will be combined on the pBIG2 level.
13. "Empty" pBIG1 vectors can also be used as placeholders in pBIG2 assemblies. For example, if expression cassettes present on pBIG1a and pBIG1c are to be combined, an "empty" pBIG1b plasmid can be used and the three plasmids combined on pBIG2abc.
14. Use a 5× molar excess of each pBIG1 construct over the PmeI-linearized pBIG2 vector. For example, for the pBIG2 assembly with 4 PGC (12 GECs) shown in Fig. 3b lanes 5–8, the following components were used.

Component	Plasmid size (bp)	c (ng/μl)	$m_{5\times\ excess}$ (ng)	V (μl)
pBIG2abcd^{PmeI}	5670	36	33	0.92
pBIG1a-PGC1	13150	164	383	2.33
pBIG1b-PGC2	11900	223	346	1.55
pBIG1c-PGC3	14500	367	422	1.15
pBIG1d-PGC4	11100	241	323	1.34
Water				0.71

Acknowledgments

We would like to thank Georg Petzold and Brenda Schulman and her laboratory members for their invaluable contributions during development and validation of the biGBac technique. Research in the laboratory of J.-M.P. is supported by Boehringer Ingelheim, the Austrian Science Fund (SFB-F34 and Wittgenstein award Z196-B20), the Austrian Research Promotion Agency (headquarter grants FFG-834223 and FFG-852936, Laura Bassi Centre for Optimized Structural Studies grant FFG-840283), and the European Union (Seventh Framework Programme Grant 227764 MitoSys).

References

1. Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92(3):291–294
2. Luckow VA, Lee SC, Barry GF, Olins PO (1993) Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in Escherichia coli. J Virol 67(8):4566–4579
3. Bieniossek C, Imasaki T, Takagi Y, Berger I (2012) MultiBac: expanding the research toolbox for multiprotein complexes. Trends Biochem Sci 37(2):49–57. https://doi.org/10.1016/j.tibs.2011.10.005
4. Berger I, Fitzgerald DJ, Richmond TJ (2004) Baculovirus expression system for heterologous multiprotein complexes. Nat Biotechnol 22(12):1583–1587. https://doi.org/10.1038/nbt1036
5. Fitzgerald DJ, Berger P, Schaffitzel C, Yamada K, Richmond TJ, Berger I (2006) Protein complex expression by using multigene baculoviral vectors. Nat Methods 3(12):1021–1032. https://doi.org/10.1038/nmeth983
6. Vijayachandran LS, Viola C, Garzoni F, Trowitzsch S, Bieniossek C, Chaillet M, Schaffitzel C, Busso D, Romier C, Poterszman A, Richmond TJ, Berger I (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. J Struct Biol 175(2):198–208. https://doi.org/10.1016/j.jsb.2011.03.007
7. Zhang Z, Yang J, Barford D (2016) Recombinant expression and reconstitution of multiprotein complexes by the USER cloning method in the insect cell-baculovirus expression system. Methods 95:13–25. https://doi.org/10.1016/j.ymeth.2015.10.003
8. Fitzgerald DJ, Schaffitzel C, Berger P, Wellinger R, Bieniossek C, Richmond TJ, Berger I (2007) Multiprotein expression strategy for structural biology of eukaryotic complexes. Structure 15(3):275–279. https://doi.org/10.1016/j.str.2007.01.016
9. Weissmann F, Petzold G, VanderLinden R, Huis In 't Veld PJ, Brown NG, Lampert F, Westermann S, Stark H, Schulman BA, Peters JM (2016) biGBac enables rapid gene assembly for the expression of large multisubunit protein complexes. Proc Natl Acad Sci U S A 113(19):E2564–E2569. https://doi.org/10.1073/pnas.1604935113
10. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA III, Smith HO (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods 6(5):343–345. https://doi.org/10.1038/nmeth.1318
11. Qiao R, Weissmann F, Yamaguchi M, Brown NG, VanderLinden R, Imre R, Jarvis MA, Brunner MR, Davidson IF, Litos G, Haselbach D, Mechtler K, Stark H, Schulman BA, Peters JM (2016) Mechanism of APC/CCDC20 activation by mitotic phosphorylation. Proc Natl Acad Sci U S A 113(19):E2570–E2578. https://doi.org/10.1073/pnas.1604929113
12. Yamaguchi M, VanderLinden R, Weissmann F, Qiao R, Dube P, Brown NG, Haselbach D, Zhang W, Sidhu SS, Peters JM, Stark H, Schulman BA (2016) Cryo-EM of mitotic checkpoint complex-bound APC/C reveals reciprocal and conformational regulation of ubiquitin ligation. Mol Cell 63(4):593–607. https://doi.org/10.1016/j.molcel.2016.07.003
13. Brown NG, VanderLinden R, Watson ER, Weissmann F, Ordureau A, Wu KP, Zhang W,

Yu S, Mercredi PY, Harrison JS, Davidson IF, Qiao R, Lu Y, Dube P, Brunner MR, Grace CRR, Miller DJ, Haselbach D, Jarvis MA, Yamaguchi M, Yanishevski D, Petzold G, Sidhu SS, Kuhlman B, Kirschner MW, Harper JW, Peters JM, Stark H, Schulman BA (2016) Dual RING E3 architectures regulate multiubiquitination and ubiquitin chain elongation by APC/C. Cell 165(6):1440–1453. https://doi.org/10.1016/j.cell.2016.05.037

14. Trowitzsch S, Bieniossek C, Nie Y, Garzoni F, Berger I (2010) New baculovirus expression tools for recombinant protein complex production. J Struct Biol 172(1):45–54. https://doi.org/10.1016/j.jsb.2010.02.010

Part II

Computational Methods

Chapter 22

Computational Modelling of Protein Complex Structure and Assembly

Jonathan N. Wells, L. Therese Bergendahl, and Joseph A. Marsh

Abstract

Sequence and structure space are nowadays sufficiently large that we can use computational methods to model the structure of proteins based on sequence similarity alone. Not only useful as a standalone tool, homology modelling has also had a transformative effect on the ease with which we can solve crystal structures and electron density maps. Another technique—molecular dynamics—aims to model protein structures from first principles and, thanks to increases in computational power, is slowly becoming a viable tool for studying protein complexes. Finally, the prediction of protein assembly pathways from three-dimensional structures of complexes is also now becoming possible.

Key words Protein interactions, Template-based modelling, Docking, Molecular dynamics, Assembly

1 Introduction

Frequently, experimental methods for determining protein complex structure are not possible or, tellingly, are simply no longer the best use of a researcher's time and money. Prediction of protein structure from sequence, first prophesied by Anfinsen in 1973 [1], has been a long-standing challenge in biology that has until recently been impossible in practice, for want of both sequence data and computational power. However, the genomic era has seen exponential increases in both of these areas, along with a similar expansion in the number of experimentally determined protein structures. Accordingly, there has been a concomitant improvement in our ability to predict structures computationally. Two important and long-running competitions have been instrumental in driving progress in the field. Both competitions—CASP (Critical Assessment of Protein Structure Prediction) [2] and CAPRI (Critical Assessment of Prediction of Interactions [3])—are now an integral part of the community, providing much-needed benchmarks and progress reports that enable both users and developers to keep track of the rapidly changing state of the art.

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764,
https://doi.org/10.1007/978-1-4939-7759-8_22,

Broadly speaking, the field of protein structure modelling can be divided into two overlapping subgroups, albeit separated mostly by minor philosophical differences. Older, and currently more practical, are top-down approaches based on the use of templates for the structures being modelled. These templates are selected based on sequence similarity with the target protein or complex. In contrast, there is equal interest in prediction of protein structure and behavior from first principles—an approach exemplified by molecular mechanics (MM) and molecular dynamics (MD) simulations. Template-based modelling (TBM) and MD represent opposite sides of the protein modelling community, but in practice, there is a great degree of overlap between the two, and most of the methods described below borrow elements from both.

2 Top-Down Modelling of Protein Complex Structure

The extent of the improvement in predictive power is such that, for individual sequences with close sequence similarity to known structures, it is usually possible to produce structures that are within a few angströms of the experimentally determined version, as measured by root-mean-square deviation of residue distances and other metrics [4, 5]. This is the process known as TBM, and nowadays it is routinely used to facilitate experimental structure determination of single protein chains.

For protein complexes, regardless of whether the structures of individual subunits are already known, it is often possible to reach a realistic approximation of the correct complex structure by a combination of homology modelling and molecular docking. Though most applicable for investigating protein-ligand interactions, molecular docking combined with homology modelling is becoming increasingly viable for protein-protein interactions, as evidenced by numerous recent studies and the results from the CASP and CAPRI competitions [6–9].

2.1 Template-Based Modelling

TBM is based on the principle that the degree of sequence divergence in homologous proteins is closely related to their structural similarity [10]. Once a suitable template is found—realistically with at least 40% sequence identity with the target protein—the sequences are aligned, and conserved regions are used to map fragments of the target onto the template structure. This is followed by replacement of the loop regions and final model refinement.

There are numerous methods based on extensions of this basic protocol that enable modelling of complete protein complexes, in addition to individual subunits [11–13]. One important and widely used strategy for template identification is threading, or dimeric threading, in the case of modelling complexes [14, 15]. Threading differs slightly from approaches based

solely on sequence homology, in that it relies more on fold recognition than sequence similarity; this is assessed by a scoring function, with the template that is eventually selected being the one that minimizes this function. As a general rule, threading is used when the target sequence has particularly low sequence similarity with other known proteins, but in practice, most modern software takes these decisions out of the hands of the user. For a broad overview of the currently available software and experimental strategies for TBM, readers should see the recent review on the topic by Szilagyi and Zhang [16].

2.2 Prediction of Protein-Protein Interfaces

Template-based methods work by mapping the sequence of the target proteins of interest onto a template of the protein complex, without explicitly modelling the interface until later refinement steps. In contrast, "molecular docking" begins with subunits in their monomeric form and models the interface directly by attempting to minimize the potential energy landscape of the bound proteins. This is done by sampling the conformational space of the two proteins with respect to each other and scoring the different interfaces that can be formed between them [17].

These two steps (sampling and scoring) can be handled in a number of ways. Conformational sampling is computationally very intensive, since two chains moving in three dimensions produce six degrees of freedom from the get-go, and allowing sidechain and backbone movements increases this number drastically. Thus, at least initially, almost all docking methods assume the proteins of interest to be rigid bodies and fix the orientation of one protein with respect to the other. This reduces the complexity of the sampling problem greatly. Fast Fourier transforms [18] were the first method to make molecular docking possible, but other approaches have subsequently been developed too. These include Monte Carlo search [19] and normal mode analysis [20], the former of which is of note because of its use in RosettaDock [21, 22]—one of the most popular and successful docking programs. A second popular program is HADDOCK [23, 24], which uses a gradient-based search method.

Scoring of interfaces is another nontrivial problem. As in the wider field of structure modelling (which encompasses docking), solutions to this problem tend to take the form of either physical or empirical, knowledge-based methods. In the former camp, force field scoring functions [25] are used to model the energy of the system in a given conformation and typically involve a large number of parameters relating to attributes such as van der Waals interactions and intramolecular strain energies. The latter consists of conceptually simpler techniques including counting of intermolecular contacts [26] and scoring based on prior knowledge of statistically likely interactions [27] gleaned from sources such as the PDB.

3 De Novo Structure Prediction

The docking methods described above are contingent on having structures available for the proteins whose interactions you are trying to model. However, it is often the case that there is no solved structure or suitable template for homology modelling available. In the past, this would have meant that the best one could do would be to try and predict secondary structure regions and likely interfaces or infer the presence of binding domains from homologous sequences using tools such as JPred4 [28] or databases such as PFAM and UniProt [29, 30]. With this in mind, researchers have begun to make inroads into de novo structure prediction, as well as looking at ways in which the difficulties of true de novo prediction can be circumvented using sequence information alone.

3.1 Using Protein Coevolution to Infer Intermolecular Contacts

Using the evolutionary sequence record to inform structure prediction is an idea that has been around for a long time [31], but has been held back by the fact that it is very challenging to distinguish coevolving sites that indicate direct amino acid contacts from transitive ones. For example, if residues A and B are in contact, and residues B and C are in contact, then A and C may also show a strong coevolutionary signal, despite interacting via an intermediary residue. Over the entire protein sequence, this blurring of coevolutionary signal is sufficient to prevent meaningful structure prediction. The major breakthrough in tackling this problem was achieved with the development of an algorithm named direct coupling analysis [32, 33], which extends Shannon's concept of mutual information [34] and enables direct and indirect residue contacts to be distinguished from each other. This has now been implemented in a more generally applicable and user-friendly format by Deborah Marks and co-workers, enabling the method's widespread use in the structure-prediction community [35–37].

Though originally used for single protein structures, this method is also applicable to protein complexes, as intermolecular contacts are subject to many of the same coevolutionary pressures as intramolecular ones; as such, EVcouplings (from the Marks group) has recently been applied to protein complexes [37]. Out of a set of 82 protein complexes with unsolved structures, 32 had a sufficiently good sequence record as to be able to predict the entire complex de novo, whereas others were sufficient to predict intermolecular contacts, but not the entire structure. Unfortunately, a limitation of this technique is identification of homomeric contacts, since without additional information these cannot be distinguished from intramolecular interactions. A related issue is that many nominally heteromeric interactions arise from homomeric interactions between genes which have undergone duplication and subsequent genetic drift [38–40]. In such cases, it may not be

possible to acquire a sufficient number of sequences for structure/interaction prediction, particularly if the proteins in question have diverged recently.

3.2 Molecular Dynamics of Protein Complexes

At present, we are still a long way from saturation of structure and sequence space, as is clear from recent studies sampling viral and prokaryotic genomes [9, 41, 42]. Consequently, a substantial proportion of the protein universe is beyond the reach of either TBM or EVcouplings. One intriguing possibility would be to use the protein sequence to model the structures of protein complexes and simulate their behavior completely from first principles, in a true de novo approach. Applied mathematics has given us quantum mechanical expressions that describe the behavior of chemical molecules down to the level of the electronic and nuclear structure. These methods are not yet practical for anything other than very small chemical systems, but the rate at which both hardware and software has improved over recent years seems to indicate a bright future. As a former editor of Nature suggested nearly 30 years ago [43], we might very well reach a stage where we can present a mathematical description of a section of DNA and use "a little algebra" to describe the function of polymerase during transcription. As exciting, or possibly far-fetched, as that thought is, the truth is that a quantum mechanical calculation of a protein complex carries with it a lot of information that is not needed.

A better approach for simulating protein complexes is to use a method known as molecular mechanics (MM) or force field methods. Although MM methods exist that explicitly include also some electronic configurations of all atoms as parameters in the model, these are not practical for systems beyond a few atoms in size. For biological macromolecules on the scale of protein complexes, empirical force fields are commonly used. These force fields depend on the concept of using cumulative physical forces to describe an approximation of the potential energy of the system. In MM a function for the potential energy with the general form $U(R)$, where $R = \{r_1, \ldots, r_n\}$ describes the coordinates of the atoms in the system, is constructed based on experimental molecular data (e.g., from crystal structures or spectroscopy studies). It is this same potential energy function that is minimized in molecular docking simulations. In slightly more detail, $U(R)$ can be described as follows:

$$
\begin{aligned}
U &= E_{\text{bonding}} + E_{\text{non bonding}} \\
E_{\text{bonding}} &= E_{\text{bonds}} + E_{\text{angles}} + E_{\text{torsions}} \\
E_{\text{non bonding}} &= E_{\text{electrostatic}} + E_{\text{Van der Waals}}
\end{aligned}
$$

The static description of a protein complex does not reveal all of its function and activity. In order to get a thorough understanding of the various states that are thermally accessible for the system—the thermodynamic ensemble—as well as the likelihood of it populating

any of these states, statistical thermodynamics is needed. From a computational point of view, this often means using molecular dynamics (MD) simulations to estimate equilibrium and average behavior of a protein complex. MD is a simulation method that models the behavior of all the atoms or molecules in a dynamical system. This is typically achieved by numerically solving Newton's equations of motion for the entire system of particles under the influence of the chosen force field. Essentially, following the dynamics of a protein complex in space as well as in time then leads to a large amount of information on the states that are available for the protein, tying together the protein's sequence to its dynamical behavior and function. Being rooted in basic physical principles, and unlike most of the other techniques we have discussed, MD is enormously versatile and is used widely in other fields outside of biology. From our perspective, much of its power lies in its ability to span multiple cellular scales—it has been used to study processes ranging from the molecular details of ligand binding to the assembly of virus capsids [44, 45].

At present, computational power is only just beginning to reach the level at which the assembly of protein complexes can be modelled from scratch. The two major obstacles barring further progress are the size of proteins and the timescales on which assembly takes place. Currently, only small proteins or protomers can be modelled and even then only across very short timescales. This is unfortunate, since folding and protein complex assembly take place over timescales of microseconds to minutes.

One way of decreasing the simulation time needed to describe certain dynamical events is to use accelerated MD approaches. These all utilize the notion that most stable states that exist are minima on the potential energy landscape [46]. By biasing the potential, these energy minima become shallower, and the system is more likely to escape and transition to a distinct state within more manageable simulation times. In combination with the massively parallelizable capabilities of a GPU processor, these methods promise an interesting future for the simulation of protein-protein interaction dynamics [47].

A recent exciting study by Plattner and colleagues [48] has also had some success using hidden Markov models in combination with thousands of MD simulations starting from different points, selected so as to maximize the efficiency of conformational space exploration. This enabled them to model the bacterial ribonuclease barnase in complex with its inhibitor barstar with impressive accuracy. Associated, dissociated, and transition states were observed across 2 ms of modelling time, and predicted thermodynamic parameters were in strong agreement with those obtained from independent experimental results.

4 Modelling Protein Complex Assembly Pathways

In addition to computationally modelling protein complex structures, it is also possible to use computational methods to model or predict the assembly pathways of protein complexes. First, Levy et al. showed that the subcomplexes formed during disassembly of homomeric complexes observed in vitro using electrospray mass spectrometry could almost always be predicted from crystal structures on the basis of interface size, with smaller interfaces being weaker and thus disassociating first during disassembly [49]. Subsequently, this strategy was extended to heteromeric complexes, with a model in which disassembly is predicted to proceed via a pathway that exposes the least amount of buried interface area at each step [50, 51] or, conversely, where assembly is predicted to proceed via the formation of the largest possible interface at each step [52]. Moreover, further support for the biological relevance and accuracy of computational predictions has come from the observations that the order of protein subunit-encoding genes in prokaryotic operons appears to be highly optimized for the predicted order of protein complex assembly [52] and that patterns of protein complex subunit degradation are closely related to the predicted order of assembly [53]. Finally, the importance of thinking about complexes in terms of assembly pathways has been underscored by the demonstration that the patterns of quaternary structure topologies observed in protein complexes of known structure can largely be explained through a model that invokes simple combinations of basic assembly steps [54].

5 Conclusions

Computational modelling of protein complexes continues to go from strength to strength. Though not a panacea for the difficult challenges we still face in structural biology, as the diversity of sequences and structures increases, so too will our ability to leverage computational power to fill in the gaps through homology modelling. MD too is finally approaching a point at which we can use it to study real-time processes of complex assembly, and with tantalizing hints of progress in the world of quantum computing, the future looks bright for the study of protein complexes.

Acknowledgment

J.M. is supported by a Medical Research Council Career Development Award (MR/M02122X/1).

References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230. https://doi.org/10.1126/science.181.4096.223
2. Moult J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. Proteins Struct Funct Genet 23:ii–iv. https://doi.org/10.1002/prot.340230303
3. Janin J, Henrick K, Moult J et al (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins Struct Funct Genet 52:2–9. https://doi.org/10.1002/prot.10381
4. Haas J, Roth S, Arnold K et al (2013) The protein model portal–a comprehensive resource for protein structure and model information. Database 2013:bat031–bat031. https://doi.org/10.1093/database/bat031
5. Moult J, Fidelis K, Kryshtafovych A et al (2016) Critical assessment of methods of protein structure prediction: progress and new directions in round XI. Proteins 84:4–14. https://doi.org/10.1002/prot.25064
6. Jiang Z-Y, Chu H-X, Xi M-Y et al (2013) Insight into the intermolecular recognition mechanism between Keap1 and IKKβ combining homology modelling, protein-protein docking, molecular dynamics simulations and virtual alanine mutation. PLoS One 8:e75076. https://doi.org/10.1371/journal.pone.0075076
7. Rajapaksha H, Petrovsky N (2014) In silico structural homology modelling and docking for assessment of pandemic potential of a novel H7N9 influenza virus and its ability to be neutralized by existing anti-hemagglutinin antibodies. PLoS One 9:e102618. https://doi.org/10.1371/journal.pone.0102618
8. Agostino M, Mancera RL, Ramsland PA, Fernández-Recio J (2016) Optimization of protein-protein docking for predicting Fc-protein interactions. J Mol Recognit 29:555–568. https://doi.org/10.1002/jmr.2555
9. Lensink MF, Velankar S, Kryshtafovych A et al (2016) Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. Proteins 84:323–348. https://doi.org/10.1002/prot.25007
10. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826
11. Chen H, Skolnick J (2008) M-TASSER: an algorithm for protein quaternary structure prediction. Biophys J 94:918–928. https://doi.org/10.1529/biophysj.107.114280
12. Tuncbag N, Gursoy A, Nussinov R, Keskin O (2011) Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat Protoc 6:1341–1354. https://doi.org/10.1038/nprot.2011.367
13. Guerler A, Govindarajoo B, Zhang Y (2013) Mapping monomeric threading to protein-protein structure prediction. J Chem Inf Model 53:717–725. https://doi.org/10.1021/ci300579r
14. Bowie J, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–170. https://doi.org/10.1126/science.1853201
15. Lu L, Lu H, Skolnick J (2002) Multiprospector: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins Struct Funct Genet 49:350–364. https://doi.org/10.1002/prot.10222
16. Szilagyi A, Zhang Y (2014) Template-based structure modeling of protein-protein interactions. Curr Opin Struct Biol 24:10–23. https://doi.org/10.1016/j.sbi.2013.11.005
17. Huang S-Y (2014) Search strategies and evaluation in protein–protein docking: principles, advances and challenges. Drug Discov Today 19:1081–1096. https://doi.org/10.1016/j.drudis.2014.02.005
18. Katchalski-Katzir E, Shariv I, Eisenstein M et al (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A 89:2195–2199. https://doi.org/10.1073/pnas.89.6.2195
19. Hart TN, Read RJ (1992) A multiple-start Monte Carlo docking method. Proteins Struct Funct Genet 13:206–222. https://doi.org/10.1002/prot.340130304
20. Zacharias M (2005) ATTRACT: protein-protein docking in CAPRI using a reduced protein model. Proteins 60:252–256. https://doi.org/10.1002/prot.20566
21. Lyskov S, Gray JJ (2008) The RosettaDock server for local protein-protein docking. Nucleic Acids Res 36:W233–W238. https://doi.org/10.1093/nar/gkn216
22. Zhang Z, Schindler CEM, Lange OF, Zacharias M (2015) Application of enhanced sampling Monte Carlo methods for high-resolution protein-protein docking in Rosetta. PLoS One 10:e0125941. https://doi.org/10.1371/journal.pone.0125941
23. Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: a protein-protein docking

approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737. https://doi.org/10.1021/ja026939x

24. van Zundert GCP, Rodrigues JPGLM, Trellet M et al (2016) The HADDOCK2.2 Web server: user-friendly integrative modeling of biomolecular complexes. J Mol Biol 428:720–725. https://doi.org/10.1016/j.jmb.2015.09.014
25. Kynast P, Derreumaux P, Strodel B (2016) Evaluation of the coarse-grained OPEP force field for protein-protein docking. BMC Biophys 9:4. https://doi.org/10.1186/s13628-016-0029-y
26. Böhm H-J (1998) Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. J Comput Aided Mol Des 12:309–309. https://doi.org/10.1023/A:1007999920146
27. Sasse A, de Vries SJ, Schindler CEM et al (2017) Rapid design of knowledge-based scoring potentials for enrichment of near-native geometries in protein-protein docking. PLoS One 12:e0170625. https://doi.org/10.1371/journal.pone.0170625
28. Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. Nucleic Acids Res 43:W389–W394. https://doi.org/10.1093/nar/gkv332
29. Finn RD, Coggill P, Eberhardt RY et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279–D285. https://doi.org/10.1093/nar/gkv1344
30. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169. https://doi.org/10.1093/nar/gkw1099
31. Altschuh D, Lesk AM, Bloomer AC, Klug A (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol 193:693–707. https://doi.org/10.1016/0022-2836(87)90352-4
32. Weigt M, White RA, Szurmant H et al (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A 106:67–72. https://doi.org/10.1073/pnas.0805923106
33. Lunt B, Szurmant H, Procaccini A et al (2010) Inference of direct residue contacts in two-component signaling. Methods Enzymol 471:17–41. https://doi.org/10.1016/S0076-6879(10)71002-8
34. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
35. Marks DS, Colwell LJ, Sheridan R et al (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS One 6:e28766. https://doi.org/10.1371/journal.pone.0028766
36. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30:1072–1080. https://doi.org/10.1038/nbt.2419
37. Hopf TA, Schärfe CPI, Rodrigues JPGLM et al (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife. https://doi.org/10.7554/eLife.03430
38. Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18:1283–1292. https://doi.org/10.1093/oxfordjournals.molbev.a003913
39. Wagner A (2003) How the global structure of protein interaction networks evolves. Proc R Soc B Biol Sci 270:457–466. https://doi.org/10.1098/rspb.2002.2269
40. Fokkens L, Hogeweg P, Snel B (2012) Gene duplications contribute to the overrepresentation of interactions between proteins of a similar age. BMC Evol Biol 12:99. https://doi.org/10.1186/1471-2148-12-99
41. Brum JR, Ignacio-Espinoza JC, Kim E-H et al (2016) Illuminating structural proteins in viral "dark matter" with metaproteomics. Proc Natl Acad Sci U S A 113:2436–2441. https://doi.org/10.1073/pnas.1525139113
42. Mukherjee S, Seshadri R, Varghese NJ et al (2017) 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nat Biotechnol. https://doi.org/10.1038/nbt.3886
43. Maddox J (1989) Towards the calculation of DNA. Nature 339:577. https://doi.org/10.1038/339577a0
44. Buch I, Giorgino T, De Fabritiis G (2011) Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proc Natl Acad Sci U S A 108:10184–10189. https://doi.org/10.1073/pnas.1103547108
45. Zhao G, Perilla JR, Yufenyuy EL et al (2013) Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. Nature 497:643–646. https://doi.org/10.1038/nature12162
46. Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys 120:11919–11929. https://doi.org/10.1063/1.1755656
47. Friedrichs MS, Eastman P, Vaidyanathan V et al (2009) Accelerating molecular dynamic

simulation on graphics processing units. J Comput Chem 30:864–872. https://doi.org/10.1002/jcc.21209

48. Plattner N, Doerr S, De Fabritiis G, Noé F (2017) Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. Nat Chem 9:1005–1011. https://doi.org/10.1038/nchem.2785
49. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA (2008) Assembly reflects evolution of protein complexes. Nature 453:1262–1265. https://doi.org/10.1038/nature06942
50. Marsh JA, Hernández H, Hall Z et al (2013) Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell 153:461–470
51. Hall Z, Hernández H, Marsh JA et al (2013) The role of salt bridges, charge density, and subunit flexibility in determining disassembly routes of protein complexes. Structure 21:1325–1337. https://doi.org/10.1016/j.str.2013.06.004
52. Wells JN, Bergendahl LT, Marsh JA (2016) Operon gene order is optimized for ordered protein complex assembly. Cell Rep 14:679–685. https://doi.org/10.1016/j.celrep.2015.12.085
53. McShane E, Sin C, Zauber H et al (2016) Kinetic analysis of protein stability reveals age-dependent degradation. Cell 167:803–815.e21. https://doi.org/10.1016/j.cell.2016.09.015
54. Ahnert SE, Marsh JA, Hernández H et al (2015) Principles of assembly reveal a periodic table of protein complexes. Science 350:aaa2245. https://doi.org/10.1126/science.aaa2245

Chapter 23

Inferring and Using Protein Quaternary Structure Information from Crystallographic Data

Sucharita Dey and Emmanuel D. Levy

Abstract

A precise knowledge of the quaternary structure of proteins is essential to illuminate both their function and their evolution. The major part of our knowledge on quaternary structure is inferred from X-ray crystallography data, but this inference process is hard and error-prone. The difficulty lies in discriminating fortuitous protein contacts, which make up the lattice of protein crystals, from biological protein contacts that exist in the native cellular environment. Here, we review methods devised to discriminate between both types of contacts and describe resources for downloading protein quaternary structure information and identifying high-confidence quaternary structures. The use of high-confidence datasets of quaternary structures will be critical for the analysis of structural, functional, and evolutionary properties of proteins.

Key words Protein quaternary structure, Protein interactions, Crystallography, Protein Data Bank, Promiscuous interactions, Homomers , Homo-oligomers, Biological assembly, Crystal contact

1 Introduction

Protein structure is classically described using a hierarchy with four levels. The polypeptide sequence is the primary structure, helices and beta-stands form secondary structures, the 3D-fold is the tertiary structure, and multiple polypeptide chains can assemble into a quaternary structure (QS). In the cell, most proteins adopt a QS. For example, the lamina network in the nucleus, the nuclear pore complex or ribosomes are large structures consisting of multiple protein chains (Fig. 1a). In these complexes, different genes contribute to the assembly, and so we refer to them as hetero-oligomers. Alternatively, a QS composed of a single gene product is referred to as a homo-oligomer or "homomer," and 30–50% of proteins self-assemble into such structures [2, 3]. For example, in the glycolytic pathway, eight out of ten enzymes form homomers (Fig. 1b).

A precise knowledge of QS is required to best understand the function [4, 5] and evolution [2, 6] of proteins and their networks of interactions. At the level of protein networks, for example, the

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_23,

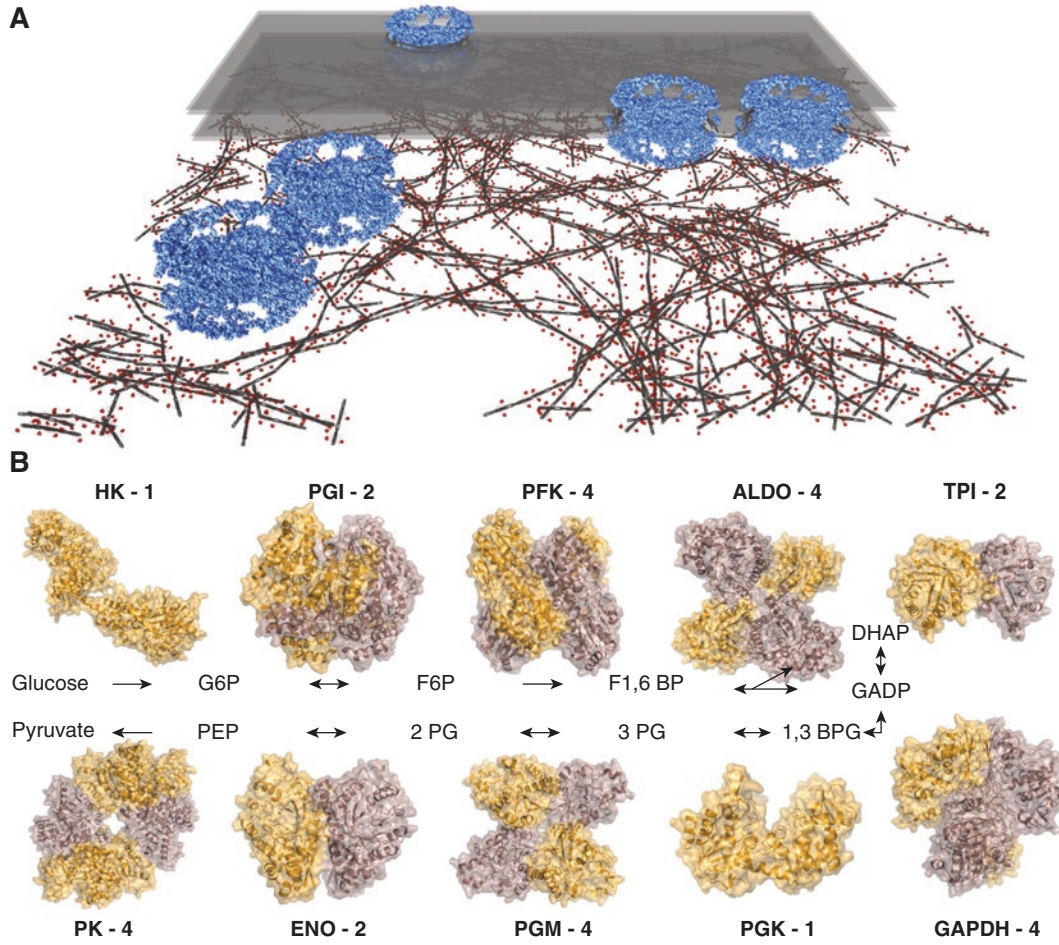


Fig. 1 QS at the nuclear envelope and in the glycolytic pathway. (**a**) The molecular architecture of the lamina meshwork is revealed in situ, by cryo-electron tomography. A- and B-type lamina monomers consist of a long helical segment and a globular domain. Four monomers from both types can assemble through their helical segment into 3.5-nm-thick filaments (gray) decorated by their globular domains (red). The tomogram also highlights nuclear pores (blue), which are large multi-protein complexes with pseudo-D8 symmetry. Image courtesy of Ohad Medalia [1]. (**b**) The glycolytic pathway includes ten enzymes, of which eight form homo-oligomers. The number of subunits composing each enzyme is given next to its abbreviation. *HK* hexokinase (PDB code 1DGK), *PGI* phosphoglucose isomerase (PDB code 1IAT), *PFK* phosphofructokinase (PDB code 4PFK), *ALDO* fructose 1,6 bisphosphate aldolase (PDB code 4ALD), *TPI* triose phosphate isomerase (PDB code 1TIM), *GAPDH* glyceraldehyde-3-phosphate dehydrogenase (PDB code 3GPD), *PGK* phosphoglycerate kinase (PDB code 2XE7), *PGM* phosphoglycerate mutase (PDB code 4PGM), *ENO* enolase (PDB code 4ENL), *PK* pyruvate kinase (PDB code 1A3W). Note that the structures shown come from different organisms, but reflect the most common QS state for each enzyme. For a more detailed structural view of glycolysis, we refer the reader to an excellent description by D. Goodsell (doi:10.2210/rcsb_pdb/mom_2004_2)

duplication of a gene coding for a homomer will result in two interacting duplicates [7]. Early analyses of protein networks revealed that homologous proteins often interact [7–10], which is largely caused by this effect, where duplicated gene products maintain an ancestral homomeric interaction. Numerous protein com-

plexes originated via this route. They include, among many others, hemoglobin, nucleosomes, or the 20S core of the proteasome. Geneticists often assume that gene duplication introduces functional redundancy and robustness, whereby each gene can compensate for a loss of function of its duplicate. However, among complexes that originated in homo-oligomers, gene duplication can create the opposite effect [11]. The 20S proteasome from *S. cerevisiae* is one such example. It contains 14 homologous proteins forming two stacked rings of 7 subunits each. All the subunits are homologous, but they cannot compensate for each other. Indeed, most of these genes are essential for survival [12, 13].

At the level of individual proteins, oligomerization is also important for function and regulation [4, 5, 14–17]. In particular, allosteric regulation is a prominent feature associated with homo-oligomerization, whereby changes in one subunit can propagate to other subunits to induce coordinated, switch-like responses [18]. In the glycolytic pathway, for example, pyruvate kinase (PK) undergoes a dramatic conformational change and gets activated upon binding of fructose 1,6 bisphosphate, a precursor in the pathway [19]. Evidence linking homo-oligomerization to function or catalysis exists for at least half of the homomers in the pathway (Table 1). Mutations that impair oligomerization can indeed disrupt allosteric regulation. More generally, impairing oligomerization results in a loss of contacts between subunits, which can destabilize protein structure and disrupt function.

Besides the glycolytic pathway, oligomerization also plays important regulatory roles [4, 24–26]. Homomers are found in virtually all cellular process, including signal transduction [27, 28] and transcriptional regulation [29]. For example, caspases often trigger apoptosis following a change in their oligomeric state [30]. And the well-known tumor suppressor protein p53 is a homo-tetramer (dimer of dimers), with mutations in the tetramerization domain impairing its DNA-binding activity [31].

Table 1
Relating oligomerization and function in glycolytic enzymes

Name	No. Sub.	Description	Reference
PFK	4	Mutations of residues at the tetramer interface were found to affect tetramer formation, enzyme catalysis, and regulation	Webb et al. [20]
TPI	2	Mutation at the TPI dimerization interface affects dimerization, causes TPI deficiency, and affects enzyme activity	Ralser et al. [21]
PGM	4	Mutations at the interface induce dissociation into monomers and dimers, which possess ~35% of the tetramer activity	White et al. [22]
PK	4	The allosteric regulation is accomplished through the oligomeric organization of the enzyme, which is usually a tetramer	Mattevi et al. [23]

At the same time, the function associated with QS can be undetermined, and it is often not clear why a particular oligomeric state would be required for a particular function. For example, Hsp27 can form high-order oligomers containing up to 24 subunits, and mutations can disrupt these oligomers while maintaining the chaperone activity [32]. In another example, the photoreceptor for ultraviolet-B (UV-B) light is a dimer. While the mutation of particular salt-bridging residues prevents dimer formation, the receptor can still perceive UV-B and initiate signaling in its monomeric state [33]. These examples reflect that proteins are likely to explore different oligomeric states during evolution. In some cases, those changes can be linked to function [34, 35] or stability [36, 37], but in other cases, the gain and loss of self-interactions may be random [38]. This notion may seem surprising but can be explained by the ease with which homo-oligomers can evolve [39–42], through gain/loss of loops [43, 44] or even single point mutations [34, 45].

The impact of QS on the function and evolution of proteins means that it is critical to consider it. The richest source of information on protein QS is the Protein Data Bank (PDB), the central repository for structural information obtained by X-ray crystallography, NMR spectroscopy, and electron microscopy [46, 47]. Currently, most of the structural information in the PDB comes from X-ray crystallography, with over 118,000 structures solved. However, QS information is not readily available from these structures, because X-ray crystallography provides atomic coordinates of the asymmetric unit (ASU) only. At the molecular level, a crystal is formed by an infinite lattice of ASUs (Fig. 2), and a QS may be made from one or more ASUs or from parts of several ASUs [4, 40]. A critical challenge underlying analysis and interpretation of protein structures is, therefore, to discriminate fortuitous crystal-packing contacts from functional protein-protein contacts that make up the biologically relevant QS (Fig. 2).

In this chapter, we review concepts and methods developed to characterize protein QS. We focus on computational methods

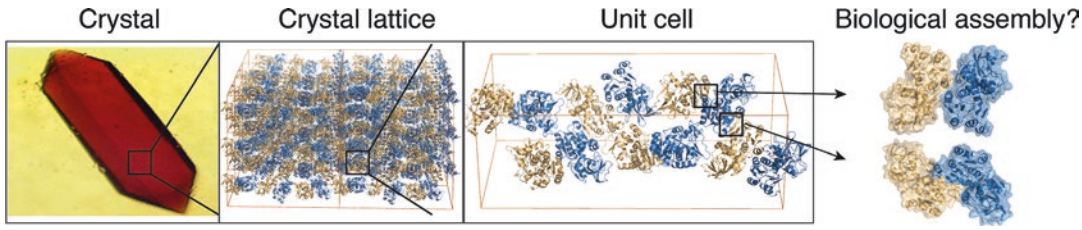


Fig. 2 From a crystal to a QS. The determination of a protein's structure by X-ray crystallography requires the formation of a protein crystal. Protein crystals are formed by a repetitive pattern of interacting protein copies forming a lattice. The repetitive pattern from which the lattice is constructed by symmetry operations is the asymmetric unit (ASU). Numerous crystal contacts are found both within the ASU and across different ASUs. These contacts are hard to distinguish from biological contacts that make up biological assemblies. Image of the crystal courtesy of Zhang et al. [48]

applicable to X-ray crystallographic data and how they can help identifying biological contacts and biological assemblies.

2 Discriminating Biological Interfaces from Crystal Contacts

The crystallization of proteins involves the formation of crystal contacts—that is, non-biological interfaces that make up the crystal lattice. To infer the QS of a protein from its crystallographic structure, one must, therefore, discriminate crystal contacts from biological protein-protein interfaces.

Interface properties used to discriminate between biological and crystal contacts can be broadly categorized into two types: (1) physicochemical and shape properties deduced directly from the structure and (2) comparative properties that make use of external data not present in the structure itself.

2.1 Structural Descriptors

A widely employed feature to discriminate biological and crystal contacts is buried surface area (BSA)—the solvent accessible area buried upon complexation. This feature was first used by Janin who observed an average BSA of 285 Å^2 per subunit in crystal contacts with no point group symmetry. Only 1% of these contacts buried over 800 Å^2 per subunit, which is an average value for biological interfaces [49]. Nowadays, BSA is still used as one of the main criteria for discriminating biological and non-biological contacts by different methods and servers. PQS was the first to implement BSA along with other features [50] to predict entire assemblies.

The area of an interface can also be subdivided into different regions. Chakrabarti and Janin divided interfaces into core and rim [51], with the former including residues with at least one fully buried atom at the interface. Levy revisited this subdivision to facilitate the study of interface evolution [52], but for classifying biological and crystal contacts, the most informative subdivision comes from Schärer et al. [53]. According to this definition, "core" residues are those burying a large fraction (>95%) of their surface upon complexation [54].

Interface area has been most often measured based on the differences of solvent accessibility in the free versus bound states of the proteins [55, 56]. However, alternative approaches exist. A notable alternative is Voronoi tessellation, the formalism of which was proposed by Georgy Voronoi in 1907 [57]. A historical account of the use and application of Voronoi tessellation to protein structure is given in [58]. Generally, Voronoi tessellation presents the advantage of a mathematically defined description of the interface [59] and thereby gives access to specific geometric descriptors such as local curvature. A method called DiMoVo distinguishes biological from crystal contacts using interface features derived from Voronoi tessellation [60].

Chemically, crystal contacts bury more polar residues on average when compared to biological interfaces [61, 62]. This difference of chemical composition was exploited by Bahadur et al. to generate a score ("Rp") based on residue propensities to discriminate biological from crystal contacts [63]. It was also observed that crystal-packing contacts have a comparatively low density of hydrogen bonds, a low fraction of fully buried atoms and a higher level of hydration [64]. Most of these features were previously implemented in the PQS prediction method, with some modifications.

Geometrically, biological interfaces are less planar [65] and show tighter atomic packing than crystal interfaces [64, 66]. Other features preferentially associated with biological interfaces include long secondary structure segments [67], low predicted free energy of association [50, 68], or low entropy of amino acids [69], which minimizes entropy loss upon binding.

The features discussed in this section have been implemented in numerous web servers and methods that tackle the problem of interface classification and are discussed in a review by Janin et al. [64]. Different methods are listed in Table 2 in chronological order, along with a brief description.

2.2 Comparative Approaches

The evolutionary conservation of traits across organisms is a powerful mean to assess their function. For example, it has been applied to detect functional regulatory elements in genomes, [79] functional post-translational modifications in proteins, [80] and can also serve to predict the functional relevance of crystallographic interfaces.

Interface conservation can be assessed in two ways. A first approach consists of measuring the conservation of amino acids forming the interface. Amino acids mediating contacts between subunits indeed tend to be more constrained in their identity than amino acids exposed to the solvent [81–83]. As a result, at the sequence level, conserved surface amino acids are predictive of interaction interfaces. This concept has been explored early on [84–86] and, recently, has been implemented in EPPIC to provide global predictions of biological interfaces across the PDB [87].

Another approach employed for assessing interface conservation is structural superposition. In this case, the residues mediating the interface need not be conserved, as long as the subunits interact with a similar geometry. Early surveys of protein structures have shown that interfaces indeed tend to be conserved during evolution [88–90]. Interface geometry conservation across different crystal forms of the same protein was employed to distinguish biological and crystal contacts [91], whereby interfaces consistently observed across crystal forms being likely biologically relevant. This notion was later extended to homologous proteins and led to the protein common interface database (ProtCID) [76]. Another resource that provides clusters of homologous interfaces is InterEvol [92].

Table 2
Methods developed to discriminate biological and non-biological protein-protein interfaces

Method	Property used for classification	Dataset used for testing	Prediction type	Location/URL
Janin [49]	Buried surface area (BSA)	Own dataset (1320 crystal contacts from 152 monomers)	Interface	–
PQS [50]	Empirical weighted score combining BSA, number of core residues, desolvation energy, salt bridges, disulfide bonds	–	Assembly	Offline, succeeded by PISA
PITA [70]	BSA and pairwise statistical potential based on types of atom pairs	Own dataset (218 proteins)	Assembly	–
NOXclass [66]	Combines six interface properties using SVM. Features include interface area, area ratio of interface to surface, interface residue composition, gap volume index, interface sequence conservation	Own dataset of obligate, non-obligate, and crystal packing interactions (243 interactions)	Interface	http://noxclass.bioinf.mpi-inf.mpg.de
PISA [68]	Models the free energy change involved in interface formation using a sophisticated model that includes enthalpy and entropy of proteins and the solvent	From PITA (218 proteins) [70]	Assembly	http://www.ebi.ac.uk/pdbe/pisa/
PiQSi [71]	Curation of the literature and manual inspection of the structures	Bahadur et al. (2003, 2004) [63, 72], Ponstingl et al. (2000, 2003) [70, 73]	Assembly	www.piqsi.org
COMP [65]	Combination of interface shape, complementarity, hydrophobicity, and electrostatic potential	Own dataset of 282 homodimers and 111 crystal contacts	Interface	–
DiMoVo [60]	SVM that includes multiple features: interface area, number of core residues and their associated Voronoi volume, frequency of residue types and pair types, as well as distances between their geometric center	Ponstingl et al. (2003) [70], Zhu et al. (2006) [66]	Interface	http://fifi.ibbmc.u-psud.fr/DiMoVo

(continued)

Table 2
(continued)

Method	Property used for classification	Dataset used for testing	Prediction type	Location/URL
IPAC [74]	Bayesian classifier that takes into account multiple features that include interface area, packing, chemical complementarity, desolvation energy, and interface composition	Ponstingl et al. (2000, 2003) [70, 73], Benchmark3.0 (Hwang et al. 2008) [75]	Assembly	pallab.serc.iisc.ernet.in/IPAC
ProtCID [76]	Interface geometry conservation across crystal forms and homologous structures	Ponstingl et al. (2000) [73], Bahadur et al. (2004) [63]	Interface	http://dunbrack2.fccc.edu/ProtCiD/Default.aspx
EPPIC [53, 54, 77]	Interface sequence conservation together with structural features including BSA and ratio of number of "core" and "rim" residues	Own dataset (78 monomers, 74 oligomers)	Interface and assembly	http://www.eppic-web.org/ewui/
IChemPIC [78]	Random forest classifier with 45 descriptors for interface size, chemical complementarity, and buriedness	Own dataset (300 interfaces)	Interface	bioinfo-pharma.u--strasbg.fr/IChemPIC

Recently, we extended the concept further, by structurally superposing full QS, which may contain more than two polypeptide chains. This strategy is called QSalign, and we describe it in the next section.

Comparative approaches relying on evolutionary information will likely improve over time, as our coverage of the sequence and structure space increases [54]. In contrast, methods that depend on physicochemical properties of the structure itself are inherently static. It will also be interesting to compare the impact of structure quality on the prediction performance of both types of approaches.

3 Resources on Protein Quaternary Structure Information

Faced with the task of analyzing protein structure, scientists must retrieve biologically relevant quaternary states. The PDB is the central repository for information on protein structures solved to date [46]. Most of the structures deposited in PDB have been solved by X-ray crystallography, and we saw that it is not trivial to annotate the QS from the crystal structure. As a result, one has to rely on methods that predict QS based on features discussed in Subheading 2. While numerous methods have been developed to distinguish biological and crystal contacts, comparatively few provide information at the level of biological assemblies. To date, the method PISA is the state-of-the-art for predicting biological assemblies [68], and PDB relies on its predictions to annotate biological assemblies when no information is available from the authors. We also note that an update to the EPPIC server is now providing assembly information and was released concomitantly with this book chapter [77].

3.1 Biological Assemblies in the Protein Data Bank

The PDB has a long and fascinating history [93]. It was created in 1971 at the Brookhaven National Laboratory and initially contained seven entries only. Later, in 1998, the management changed to a consortium called the Research Collaboratory for Structural Bioinformatics (RCSB PDB). In 2003, the PDB network expanded with the Macromolecular Structure Database at the European Bioinformatics Institute (MSD-EBI) [94] and PDB Japan (PDBj) at the Institute for Protein Research at Osaka University, Japan [95]. The rapid development of technologies used in structure determination and the fundamental role of structure in biological research resulted in exponential growth. The archive that started with only seven structures now contains more than a hundred thousand. The information made available by the PDB also became wider, with the integration of external resources such as GenBank, UniProt, domain databases, model organism databases, and many others.

Apart from the atomic coordinates of the asymmetric unit, the PDB provides information on the probable biological assembly.

This information is derived using three sources: (1) information provided by the authors, if any, (2) the prediction from PISA, and (3) manual curation by the PDB teams. The retrieval of all assemblies coordinates can be achieved with rsync, using the following command as detailed on the wwpdb.org website (https://www.wwpdb.org/download/downloads):

```
rsync -rlpt -v -z --delete rsync.ebi.ac.uk::pub/databases/rcsb/pdb/data/biounit/coordinates/divided/
```

A caveat of using precomputed information of biological assemblies is that chains duplicated by symmetry operations bear the same name. An alternative source of coordinate files for biological assemblies with unique chain names can be found on the 3D Complex website (http://www.3dcomplex.org, latest version accessible through www.piqsi.org). Another helpful feature of these structures is the renumbering of residues according to the SEQRES sequence, but a drawback is the lack of regular updates. Alternatively, each biological assembly can be generated based on the coordinate of the asymmetric unit, using symmetry operations detailed in the "REMARK 350" section or from the mmCIF format files using the attributes "pdbx_struct_assembly," "pdbx_struct_assembly_gen," and "pdbx_struct_oper_list." The first two attributes show how to generate each biological assembly for the structure and present details about it, while the third one gives the transformations required for generating the biological assembly (https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies#Anchor-Biol).

Importantly, biological assemblies available from the PDB are limited by experimental knowledge on QS and, in the absence of such information, are limited by predictions of PISA. As a result, a non-negligible fraction of biological assemblies does not reflect a biological QS. The large-scale manual curation involved with the PiQSi database suggests that non-biological QS may represent ~15% of assemblies found in PDB. To assist end users to filter out such cases, we make available predictions derived from QSbio (discussed below).

3.2 PISA

PISA stands for "Protein Interfaces, Surfaces, and Assemblies" and was established in 2005 by Krissinel and Henrick [68]. PISA supplanted the Probable Quaternary Structure (PQS) server [50], which was the first of a kind to determine oligomeric states of protein assemblies. Beyond the classification into biological or crystal contacts, the identification of QS from crystallographic data is a complex task because a potentially infinite number of assemblies need to be enumerated. To accomplish this task, PISA represents the crystal as a periodic graph and employs an efficient backtracking algorithm. Each assembly that is enumerated is assessed by an energy scoring function that models the enthalpy and entropy of both the proteins and the solvent.

PISA is available as part of the CCP4 suite (http://www.ccp4.ac.uk/index.php) [96] and can be executed locally on a protein

structure of choice. The method can also be queried through a web interface, and coordinates of assemblies can be retrieved automatically with the following command given in the website http://www.ebi.ac.uk/pdbe/pisa/pi_download.html:

```
wget http://www.ebi.ac.uk/pdbe/pisa/cgi-bin/multimer.pdb?pdbcode:n,m
```

where "pdbcode" is a four-letter code and using a value equal to 1 for *n* and *m* to get the most likely assembly.

3.3 PiQSi

PiQSi, which stands for "Protein Quaternary Structure Investigation" [71], is a manually annotated database containing QS annotations of ~15,000 proteins, of which ~2600 are nonredundant. PiQSi relies on the representation of protein complexes as graphs, as introduced in 3D Complex [97]. The graph representation provides a quick overview of the number of subunits, their type (identical, homologous, structurally unrelated subunits), and the contact sizes between them. Given a particular entry, the PiQSi server displays homologues along with their graph representation to facilitate the visual comparison of QS within a protein family.

3.4 QSalign and Anti-QSalign Compare Quaternary Structures Across Homologues

QSalign relies on the same premise as other comparative approaches, i.e., that evolutionary conservation is indicative of function (Fig. 3). A unique feature of QSalign, however, is that it considers entire QS as units and does not attempt to decompose them into pairwise interfaces. Thus, full assemblies are being superposed, and structural similarity, if found, is used as evidence of their functional relevance. The multichain superposition employed in QSalign relied on the Kpax algorithm [99] together with a heuristic for chain-chain mapping. The high degree of redundancy found in PDB, whereby the same structure often exists for multiple species, enabled annotating ~70% of homo-oligomers on the basis of their evolutionary conservation. A remarkable aspect of QSalign was the accuracy of its annotations, which showed an error rate below ~5%, about threefold lower than that observed with other methods on the same datasets [98].

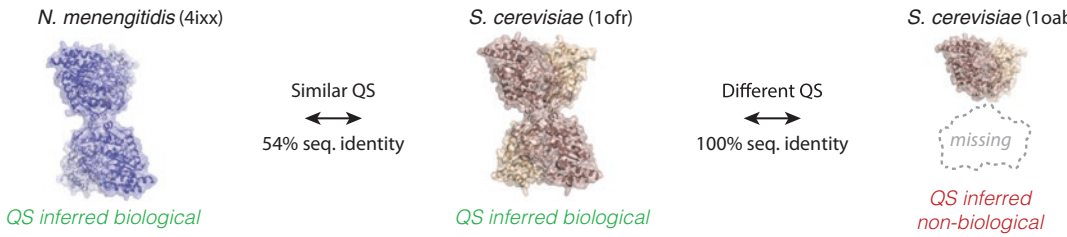


Fig. 3 Conservation of QS geometry. The enzyme 3-deoxy-d-arabino-heptulosonate 7-phosphate synthase (DAH7PS) is a tetramer in *Neisseria meningitidis* (PDB code 4ixx). A similar tetramer is found in *Saccharomyces cerevisiae* (PDB code 1ofr), which shares 54% identity in sequence. The geometric conservation of the tetramers suggests they are biologically relevant. This information can be used subsequently to correct entries showing identical sequence but different QS (PDB code 1oab). This example illustrates how QSalign operates to annotate QS [98]

The concept of QS conservation can also be applied to monomers. In that case, however, a reverse approach must be employed, where the absence of QS in homologues can be used as predictive information of a monomeric state. This reverse approach, anti-QSalign, provides annotations for ~80% of monomers, with accuracies comparable to that of other approaches [98].

3.5 QSbio, a Resource to Focus on Biologically Relevant Assemblies in the PDB

Meta-predictors integrating multiple methods generally perform better than any of the individual methods they rely on. This motivated the creation of QSbio, which combines predictions of PISA, EPPIC, and, when available, QSalign/anti-QSalign. The predictions of all three methods were mapped onto PDB assemblies and, depending on their agreement and disagreement, enabled assigning an error probability to each structure. At one extreme, very high confidence can be placed in PDB assemblies supported by all methods. And at the other extreme, one should avoid using PDB assemblies supported by no single method. Overall, QSbio assigns one of the five "confidence tags" to PDB assemblies: "very high," "high," "medium," "low," and "very low." While 63% of annotated structures were assigned a "high" or "very high" confidence, a significant 10% had a confidence predicted to be "very low" [98].

Importantly, the PDB often makes available multiple potential assemblies for any given protein structure. Such multiplicity can occur when multiple assemblies are observed in the asymmetric unit but can also occur when uncertainty exists and multiple options are left open. We thus analyzed how the fractions of structures in the different confidence classes change after we keep a single assembly (the one of highest confidence—best biological unit "BU") per PDB entry (Table 3). The statistics show that 74% of proteins in the PDB can be used with high confidence, 14% of structures should be avoided as they are assigned a low confidence, and finally 12% of structures show an intermediate level of confidence.

Table 3
QSbio confidence categories across all structures and when only the best biological assembly per structure is kept

	Very high	High	Medium	Low	Very low
Total number	51,050	18,217	14,995	14,335	11,499
Fraction	0.46	0.17	0.14	0.13	0.10
Num. of best BU	38,816	13,207	8720	6809	3671
Fraction of best BU	0.55	0.19	0.12	0.09	0.05

The confidence annotations of QSbio are available for download at www.QSbio.org and can be used in conjunction with biological assemblies retrieved from the PDB. Information about individual structures is also available in the "structural analysis" page of each structure, accessible on the PDBe website (PDB in Europe, www.pdbe.org) [47] (Fig. 4).

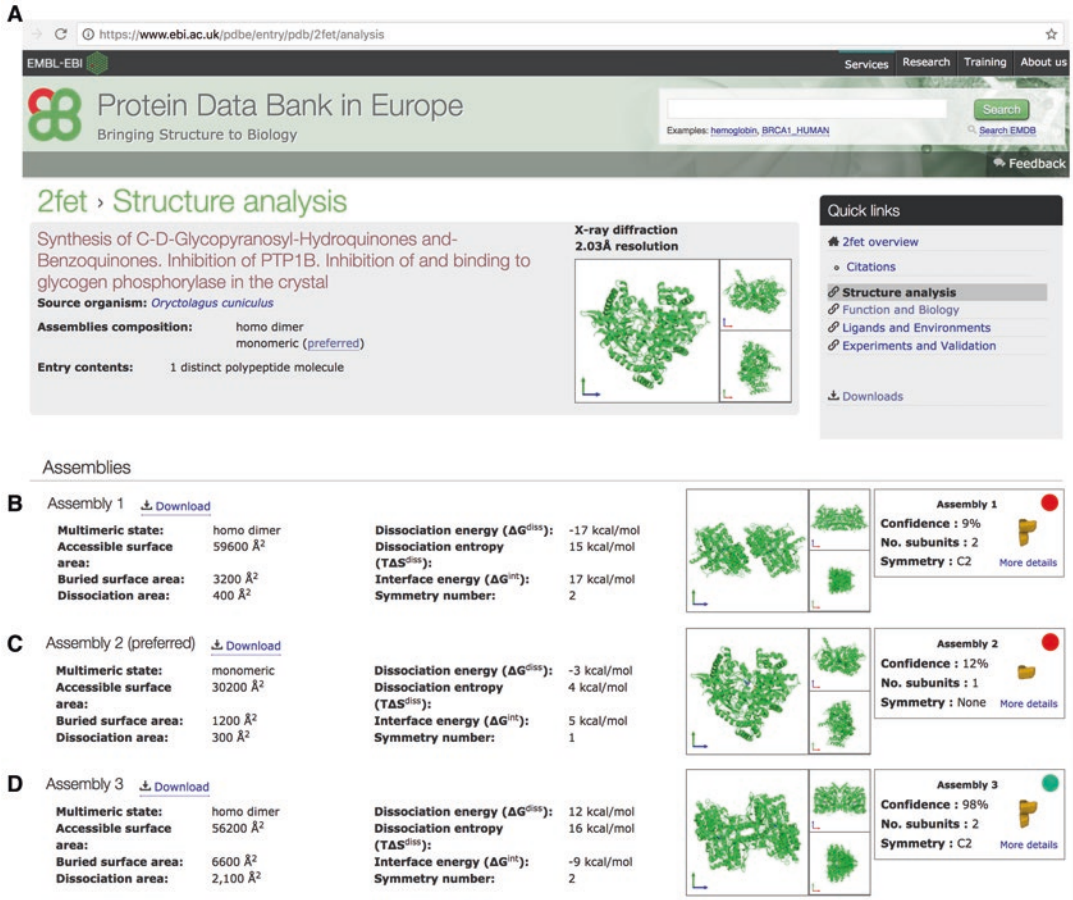


Fig. 4 Accessing QSbio confidence estimate on the PDBe website. (**a**) The "structure analysis" page of a particular protein displays assemblies available in the PDB. Here, we show the structure analysis page for a glycogen phosphorylase from rabbit (PDB code: 2FET). Whenever an annotation for a particular assembly is available from QSbio, the widget depicted on the right-hand side is shown. The widget provides information from the 3D Complex database on the number of subunits and symmetry of each biological assembly. It also provides an estimate of the confidence one can place in a particular biological assembly, as estimated in QSbio (www.QSbio.org, [98]). (**b**) Assembly 1 is a dimer. QSbio places a low confidence in this structure, suggesting this assembly is likely non-biological. (**c**) Assembly 2 is a monomer, and QSbio also places a low confidence in it. (**d**) Assembly 3 is a dimer in which the interface is different from the dimer seen in (**b**). QSbio places high confidence in this assembly because a homologous dimer exists in *E. coli*

4 Conclusion

In recent years, the exponential growth of the PDB has been making "omics data" amenable to structural interpretation. Indeed, the coverage of proteomes by protein structures is reaching significant figures [100–102]. In this respect, the use of high-confidence QS will be crucial to maximizing the potential of structural data. To this aim, we described methods and repositories that provide information on protein QS and that will facilitate its use in future analyses of protein structure.

Acknowledgments

We thank Ohad Medalia for providing the lamina meshwork image in Fig. 1 and William Cramer for providing the crystal photography in Fig. 2. We thank the PDBe team and in particular Sameer Velankar for the integration of QSbio into PDBe assembly pages. This work was supported by a VATAT fellowship to S. Dey, by the Israel Science Foundation and the I-CORE Program of the Planning and Budgeting Committee (grant nos. 1775/12 and 2179/14), by the Marie Curie CIG Program (project no. 711715), by the HFSP Career Development Award to E. D. Levy (award no. CDA00077/2015), and by a research grant from AM. Boucher. E.D. Levy is incumbent of the Recanati Career Development Chair of Cancer Research.

References

1. Turgay Y, Eibauer M, Goldman AE, Shimi T, Khayat M, Ben-Harush K, Dubrovsky-Gaupp A, Sapra KT, Goldman RD, Medalia O (2017) The molecular architecture of lamins in somatic cells. Nature 543(7644):261–264. https://doi.org/10.1038/nature21382
2. Levy ED, Teichmann S (2013) Structural, evolutionary, and assembly principles of protein oligomerization. Prog Mol Biol Transl Sci 117:25–51. https://doi.org/10.1016/B978-0-12-386931-9.00002-7
3. Marsh JA, Teichmann SA (2015) Structure, dynamics, assembly, and evolution of protein complexes. Annu Rev Biochem 84:551–575. https://doi.org/10.1146/annurev-biochem-060614-034142
4. Goodsell DS, Olson AJ (2000) Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 29:105–153. https://doi.org/10.1146/annurev.biophys.29.1.105
5. Ali MH, Imperiali B (2005) Protein oligomerization: how and why. Bioorg Med Chem 13(17):5013–5020. https://doi.org/10.1016/j.bmc.2005.05.037
6. D'Alessio G (1999) Evolution of oligomeric proteins. The unusual case of a dimeric ribonuclease. Eur J Biochem 266(3):699–708
7. Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA (2007) Evolution of protein complexes by duplication of homomeric interactions. Genome Biol 8(4):R51
8. Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) Binding properties and evolution of homodimers in protein-protein interaction networks. Nucleic Acids Res 33(11):3629–3635
9. Orlowski J, Kaczanowski S, Zielenkiewicz P (2007) Overrepresentation of interactions between homologous proteins in interactomes. FEBS Lett 581(1):52–56. https://doi.org/10.1016/j.febslet.2006.11.076
10. Levy ED, Pereira-Leal JB (2008) Evolution and dynamics of protein interactions and networks. Curr Opin Struct Biol 18(3):349–357. https://doi.org/10.1016/j.sbi.2008.03.003
11. Diss G, Gagnon-Arsenault I, Dion-Cote AM, Vignaud H, Ascencio DI, Berger CM, Landry CR (2017) Gene duplication can impart fragility, not robustness, in the yeast protein inter-

action network. Science 355(6325):630–634. https://doi.org/10.1126/science.aai7685
12. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294(5550):2364–2368. https://doi.org/10.1126/science.1065810
13. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 40(Database issue):D700–D705. https://doi.org/10.1093/nar/gkr1029
14. Nooren IM, Thornton JM (2003) Diversity of protein-protein interactions. EMBO J 22(14):3486–3492
15. Griffin MD, Gerrard JA (2012) The relationship between oligomeric state and protein function. Adv Exp Med Biol 747:74–90. https://doi.org/10.1007/978-1-4614-3229-6_5
16. Matthews JM, Sunde M (2012) Dimers, oligomers, everywhere. Adv Exp Med Biol 747:1–18. https://doi.org/10.1007/978-1-4614-3229-6_1
17. Perica T, Marsh JA, Sousa FL, Natan E, Colwell LJ, Ahnert SE, Teichmann SA (2012) The emergence of protein complexes: quaternary structure, dynamics and allostery. Colworth Medal Lecture. Biochem Soc Trans 40(3):475–491. https://doi.org/10.1042/BST20120056
18. Monod J, Wyman J, Changeux JP (1965) On the nature of allosteric transitions: a plausible model. J Mol Biol 12:88–118
19. Mattevi A, Valentini G, Rizzi M, Speranza ML, Bolognesi M, Coda A (1995) Crystal structure of Escherichia coli pyruvate kinase type I: molecular basis of the allosteric transition. Structure 3(7):729–741
20. Webb BA, Forouhar F, Szu FE, Seetharaman J, Tong L, Barber DL (2015) Structures of human phosphofructokinase-1 and atomic basis of cancer-associated mutations. Nature 523(7558):111–114. https://doi.org/10.1038/nature14405
21. Ralser M, Heeren G, Breitenbach M, Lehrach H, Krobitsch S (2006) Triose phosphate isomerase deficiency is caused by altered dimerization—not catalytic inactivity—of the mutant enzymes. PLoS One 1:e30. https://doi.org/10.1371/journal.pone.0000030
22. White MF, Fothergill-Gilmore LA, Kelly SM, Price NC (1993) Dissociation of the tetrameric phosphoglycerate mutase from yeast by a mutation in the subunit contact region. Biochem J 295(Pt 3):743–748
23. Mattevi A, Bolognesi M, Valentini G (1996) The allosteric regulation of pyruvate kinase. FEBS Lett 389(1):15–19
24. Marianayagam NJ, Sunde M, Matthews JM (2004) The power of two: protein dimerization in biology. Trends Biochem Sci 29(11):618–625. https://doi.org/10.1016/j.tibs.2004.09.006
25. Hashimoto K, Madej T, Bryant SH, Panchenko AR (2010) Functional states of homooligomers: insights from the evolution of glycosyltransferases. J Mol Biol 399(1):196–206. https://doi.org/10.1016/j.jmb.2010.03.059. S0022-2836(10)00334-7 [pii]
26. Bergendahl LT, Marsh JA (2017) Functional determinants of protein assembly into homomeric complexes. Sci Rep 7(1):4932. https://doi.org/10.1038/s41598-017-05084-8
27. Lemmon MA, Schlessinger J (1994) Regulation of signal transduction and signal diversity by receptor oligomerization. Trends Biochem Sci 19(11):459–463
28. Changeux JP, Edelstein SJ (2005) Allosteric mechanisms of signal transduction. Science 308(5727):1424–1428. https://doi.org/10.1126/science.1108595
29. Funnell AP, Crossley M (2012) Homo- and heterodimerization in transcriptional regulation. Adv Exp Med Biol 747:105–121. https://doi.org/10.1007/978-1-4614-3229-6_7
30. Renatus M, Stennicke HR, Scott FL, Liddington RC, Salvesen GS (2001) Dimer formation drives the activation of the cell death protease caspase 9. Proc Natl Acad Sci U S A 98(25):14250–14255. https://doi.org/10.1073/pnas.231465798
31. Chene P (2001) The role of tetramerization in p53 function. Oncogene 20(21):2611–2617. https://doi.org/10.1038/sj.onc.1204373
32. Chavez Zobel AT, Lambert H, Theriault JR, Landry J (2005) Structural instability caused by a mutation at a conserved arginine in the alpha-crystallin domain of Chinese hamster heat shock protein 27. Cell Stress Chaperones 10(2):157–166
33. Heilmann M, Velanis CN, Cloix C, Smith BO, Christie JM, Jenkins GI (2016) Dimer/monomer status and in vivo function of salt-bridge mutants of the plant UV-B photoreceptor UVR8. Plant J 88(1):71–81. https://doi.org/10.1111/tpj.13260
34. Perica T, Chothia C, Teichmann SA (2012) Evolution of oligomeric state through geometric coupling of protein interfaces. Proc Natl Acad Sci U S A 109(21):8127–8132. https://doi.org/10.1073/pnas.1120028109
35. Perica T, Kondo Y, Tiwari SP, McLaughlin SH, Kemplen KR, Zhang X, Steward A, Reuter N, Clarke J, Teichmann SA (2014)

Evolution of oligomeric state through allosteric pathways that mimic ligand binding. Science 346(6216):1254346. https://doi.org/10.1126/science.1254346
36. Cohen-Khait R, Dym O, Hamer-Rogotner S, Schreiber G (2017) Promiscuous protein binding as a function of protein stability. Structure 25(12):1867–1874.e3. https://doi.org/10.1016/j.str.2017.11.002
37. Bershtein S, Mu W, Shakhnovich EI (2012) Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. Proc Natl Acad Sci U S A 109(13):4857–4862. https://doi.org/10.1073/pnas.1118157109
38. Lynch M (2013) Evolutionary diversification of the multimeric states of proteins. Proc Natl Acad Sci U S A 110(30):E2821–E2828. https://doi.org/10.1073/pnas.1310980110
39. Lukatsky DB, Zeldovich KB, Shakhnovich EI (2006) Statistically enhanced self-attraction of random patterns. Phys Rev Lett 97(17):178101. https://doi.org/10.1103/PhysRevLett.97.178101
40. Lukatsky DB, Shakhnovich BE, Mintseris J, Shakhnovich EI (2007) Structural similarity enhances interaction propensity of proteins. J Mol Biol 365(5):1596–1606
41. Andre I, Strauss CE, Kaplan DB, Bradley P, Baker D (2008) Emergence of symmetry in homooligomeric biological assemblies. Proc Natl Acad Sci U S A 105(42):16148–16152. https://doi.org/10.1073/pnas.0807576105
42. Schulz GE (2010) The dominance of symmetry in the evolution of homo-oligomeric proteins. J Mol Biol 395(4):834–843. https://doi.org/10.1016/j.jmb.2009.10.044
43. Akiva E, Itzhaki Z, Margalit H (2008) Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. Proc Natl Acad Sci U S A 105(36):13292–13297. https://doi.org/10.1073/pnas.0801207105
44. Hashimoto K, Panchenko AR (2010) Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. Proc Natl Acad Sci U S A 107(47):20352–20357. https://doi.org/10.1073/pnas.1012999107
45. Garcia-Seisdedos H, Empereur-Mot C, Elad N, Levy ED (2017) Proteins evolve on the edge of supramolecular self-assembly. Nature 548(7666):244–247. https://doi.org/10.1038/nature23320
46. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28(1):235–242
47. Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P, Hendrickx PM, Lagerstedt I, Mir S, Fernandez Montecelo MA, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Sanz-Garcia E, Sen S, Slowley RA, Wainwright ME, Deshpande MS, Iudin A, Sahni G, Salavert Torres J, Hirshberg M, Mak L, Nadzirin N, Armstrong DR, Clark AR, Smart OS, Korir PK, Kleywegt GJ (2016) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. Nucleic Acids Res 44(D1):D385–D395. https://doi.org/10.1093/nar/gkv1047
48. Zhang H, Kurisu G, Smith JL, Cramer WA (2003) A defined protein-detergent-lipid complex for crystallization of integral membrane proteins: the cytochrome b6f complex of oxygenic photosynthesis. Proc Natl Acad Sci U S A 100(9):5160–5163. https://doi.org/10.1073/pnas.0931431100
49. Janin J (1997) Specific versus non-specific contacts in protein crystals. Nat Struct Biol 4(12):973–974
50. Henrick K, Thornton JM (1998) PQS: a protein quaternary structure file server. Trends Biochem Sci 23(9):358–361
51. Chakrabarti P, Janin J (2002) Dissecting protein-protein recognition sites. Proteins 47(3):334–343
52. Levy ED (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution. J Mol Biol 403(4):660–670. S0022-2836(10)01016-8[pii]. https://doi.org/10.1016/j.jmb.2010.09.028
53. Scharer MA, Grutter MG, Capitani G (2010) CRK: an evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts. Proteins 78(12):2707–2713. https://doi.org/10.1002/prot.22787
54. Duarte JM, Srebniak A, Scharer MA, Capitani G (2012) Protein interface classification by evolutionary analysis. BMC Bioinformatics 13:334. https://doi.org/10.1186/1471-2105-13-334
55. Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55(3):379–400
56. Chothia C, Janin J (1975) Principles of protein-protein recognition. Nature 256(5520):705–708
57. Georgy F (1907) Voronoi. Nouvelles applications des parametres continusa la théorie des formes quadratiques premier mémoire: sûr quelques propriétés des formes quadratiques positives parfaits. Journal für die reine und angewandte Mathematik 133:97–178
58. Poupon A (2004) Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. Curr Opin Struct Biol 14(2):233–241. https://doi.org/10.1016/j.sbi.2004.03.010
59. Cazals F, Proust F, Bahadur RP, Janin J (2006) Revisiting the Voronoi descrip-

tion of protein-protein interfaces. Protein Sci 15(9):2082–2092. https://doi.org/10.1110/ps.062245906
60. Bernauer J, Bahadur RP, Rodier F, Janin J, Poupon A (2008) DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. Bioinformatics (Oxford, England) 24(5):652–658. https://doi.org/10.1093/bioinformatics/btn022
61. Miller S, Lesk AM, Janin J, Chothia C (1987) The accessible surface area and stability of oligomeric proteins. Nature 328(6133):834–836. https://doi.org/10.1038/328834a0
62. Jones S, Thornton JM (1995) Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 63(1):31–65
63. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2004) A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 336(4):943–955
64. Janin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41(2):133–180. S0033583508004708 [pii]. https://doi.org/10.1017/S0033583508004708
65. Tsuchiya Y, Nakamura H, Kinoshita K (2008) Discrimination between biological interfaces and crystal-packing contacts. Adv Appl Bioinform Chem 1:99–113
66. Zhu H, Domingues FS, Sommer I, Lengauer T (2006) NOXclass: prediction of protein-protein interaction types. BMC Bioinformatics 7:27
67. Pal A, Chakrabarti P, Bahadur R, Rodier F, Janin J (2007) Peptide segments in protein-protein interfaces. J Biosci 32(1):101–111
68. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372(3):774–797
69. Liu Q, Li Z, Li J (2014) Use B-factor related features for accurate classification between protein binding interfaces and crystal packing contacts. BMC Bioinformatics 15(Suppl 16):S3. https://doi.org/10.1186/1471-2105-15-S16-S3
70. Ponstingl H, Kabir T, Thornton JM (2003) Automatic inference of protein quaternary structure from crystals. J Appl Cryst 36(5):1116–1122
71. Levy ED (2007) PiQSi: protein quaternary structure investigation. Structure 15(11):4
72. Bahadur RP, Chakrabarti P, Rodier F, Janin J (2003) Dissecting subunit interfaces in homodimeric proteins. Proteins 53(3):708–719
73. Ponstingl H, Henrick K, Thornton JM (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 41(1):47–57
74. Mitra P, Pal D (2011) Combining Bayes classification and point group symmetry under Boolean framework for enhanced protein quaternary structure inference. Structure 19(3):304–312. https://doi.org/10.1016/j.str.2011.01.009
75. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z (2008) Protein-protein docking benchmark version 3.0. Proteins 73(3):705–709. https://doi.org/10.1002/prot.22106
76. Xu Q, Dunbrack RL Jr (2011) The protein common interface database (ProtCID)—a comprehensive database of interactions of homologous proteins in multiple crystal forms. Nucleic Acids Res 39(Database issue):D761–D770. https://doi.org/10.1093/nar/gkq1059
77. Bliven S, Lafita A, Parker A, Capitani G, Duarte JM (2017) Automated evaluation of quaternary structures from protein crystals. Acta Cryst Sec A A73:a117. https://doi.org/10.1101/224717
78. Da Silva F, Desaphy J, Bret G, Rognan D (2015) IChemPIC: a random forest classifier of biological and crystallographic protein-protein interfaces. J Chem Inf Model 55(9):2005–2014. https://doi.org/10.1021/acs.jcim.5b00190
79. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C,

Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146):799–816. https://doi.org/10.1038/nature05874

80. Beltrao P, Albanese V, Kenner LR, Swaney DL, Burlingame A, Villen J, Lim WA, Fraser JS, Frydman J, Krogan NJ (2012) Systematic functional prioritization of protein posttranslational modifications. Cell 150(2):413–425. https://doi.org/10.1016/j.cell.2012.05.036
81. Valdar WS, Thornton JM (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins 42(1):108–124
82. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13(1):190–202. https://doi.org/10.1110/ps.03323604
83. Franzosa EA, Xia Y (2009) Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol 26(10):2387–2395. https://doi.org/10.1093/molbev/msp146
84. Elcock AH, McCammon JA (2001) Identification of protein oligomerization states by analysis of interface conservation. Proc Natl Acad Sci U S A 98(6):2990–2994. https://doi.org/10.1073/pnas.061411798
85. Guharoy M, Chakrabarti P (2005) Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Acad Sci U S A 102(43):15447–15452. https://doi.org/10.1073/pnas.0505425102
86. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38(Web Server issue):W529–W533. https://doi.org/10.1093/nar/gkq399
87. Baskaran K, Duarte JM, Biyani N, Bliven S, Capitani G (2014) A PDB-wide, evolution-based assessment of protein-protein interfaces. BMC Struct Biol 14:22. https://doi.org/10.1186/s12900-014-0022-0
88. Aloy P, Ceulemans H, Stark A, Russell RB (2003) The relationship between sequence and interaction divergence in proteins. J Mol Biol 332(5):989–998
89. Winter C, Henschel A, Kim WK, Schroeder M (2006) SCOPPI: a structural classification of protein-protein interfaces. Nucleic Acids Res 34(Database issue):D310–D314. https://doi.org/10.1093/nar/gkj099
90. Stein A, Ceol A, Aloy P (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res 39(Database issue):D718–D723. https://doi.org/10.1093/nar/gkq962
91. Xu Q, Canutescu AA, Wang G, Shapovalov M, Obradovic Z, Dunbrack RL Jr (2008)

Statistical analysis of interface similarity in crystals of homologous proteins. J Mol Biol 381(2):487–507. https://doi.org/10.1016/j.jmb.2008.06.002
92. Faure G, Andreani J, Guerois R (2012) InterEvol database: exploring the structure and evolution of protein complex interfaces. Nucleic Acids Res 40(Database issue):D847–D856. https://doi.org/10.1093/nar/gkr845
93. Berman HM (2008) The Protein Data Bank: a historical perspective. Acta Crystallogr A 64(Pt 1):88–95. https://doi.org/10.1107/S0108767307035623
94. Velankar S, McNeil P, Mittard-Runte V, Suarez A, Barrell D, Apweiler R, Henrick K (2005) E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res 33(Database issue):D262–D265. https://doi.org/10.1093/nar/gki058
95. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10(12):980. https://doi.org/10.1038/nsb1203-980
96. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr 67(Pt 4):235–242. https://doi.org/10.1107/S0907444910045749
97. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA (2006) 3D complex: a structural classification of protein complexes. PLoS Comput Biol 2(11):e155
98. Dey S, Ritchie DW, Levy ED (2018) PDB-wide identification of biological assemblies from conserved quaternary structure geometry. Nat Methods 15(1):67–72. https://doi.org/10.1038/nmeth.4510
99. Ritchie DW, Ghoorah AW, Mavridis L, Venkatraman V (2012) Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. Bioinformatics (Oxford, England) 28(24):3274–3281. https://doi.org/10.1093/bioinformatics/bts618
100. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B, Orengo C (2009) PSI-2: structural genomics to cover protein domain family space. Structure 17(6):869–881. https://doi.org/10.1016/j.str.2009.03.015
101. Mosca R, Ceol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. Nat Methods 10(1):47–53. https://doi.org/10.1038/nmeth.2289
102. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci U S A 111(10):3733–3738. https://doi.org/10.1073/pnas.1321614111

Chapter 24

Searching and Extracting Data from the EMBL-EBI Complex Portal

Birgit H. M. Meldal and Sandra Orchard

Abstract

The Complex Portal (www.ebi.ac.uk/complexportal) is an encyclopedia of macromolecular complexes. Complexes are assigned unique, stable IDs, are species specific, and list all participating members with links to an appropriate reference database (UniProtKB, ChEBI, RNAcentral). Each complex is annotated extensively with its functions, properties, structure, stoichiometry, tissue expression profile, and subcellular location. Links to domain-specific databases allow the user to access additional information and enable data searching and filtering. Complexes can be saved and downloaded in PSI-MI XML, MI-JSON, and tab-delimited formats.

Key words Complex portal, Protein-protein interactions, Database, Bioinformatics, Protein function, Protein structure, Molecular pathways

1 Introduction

Biological processes are driven by the action of proteins and fine-tuned by their interactions with other proteins, small molecules, and nucleic acids. Many proteins have no function as monomeric chains but are constitutively found as obligate members of defined protein complexes. Many proteins can function in several distinct complexes in different molecular pathways. One major question posed by experimental biologists can be: Is my target protein a member of a complex and what does this complex do? A Complex Portal entry [1] summarizes our knowledge about a given complex. This includes the complex participants, identified uniquely by their accession numbers (UniProtKB [2] for proteins, ChEBI [3] for small molecules, or RNAcentral [4] for ncRNAs) and the relative stoichiometry of each participant, if known. Complex function is described both as free-text and structured Gene Ontology terms [5, 6], and each entry contains extensive cross-referencing to external databases that deal with specific aspects of a given complex, such as ChEMBL [7] for drug-binding information,

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_24,

Electron Microscopy Data Bank (EMDB) [8] and PDB [9] for visualization of their 3D structure, IntEnz [10] for enzyme classifications, MatrixDB [11] for extracellular matrix components, and Reactome [12] for molecular reactions and pathways. Each complex is manually annotated to the Gene Ontology, and these annotations are exported from the Complex Portal to the Gene Ontology database.

All complexes are annotated with an indication of the level of experimental evidence which confirms their existence. This evidence may be physical interaction data derived from an IMEx Consortium partner [13], a consortium of protein interaction databases, or structural evidence from EMDB [8] or the PDB [9]. In some cases, experimental evidence is derived from experiments using proteins from a mix of related species. When a complex has been fully experimentally verified in one species, it may also be described as an orthologue in closely related species; for example, a human complex may also be curated in mouse and rat. Similarly, within a single species, paralogous complexes containing members from the same gene family may also be described. Complexes that lack experimental protein-protein interaction evidence are curated if their existence is accepted by experts in the field. Such complexes often include proteins with transmembrane domains that are difficult to isolate. Terms have been added to Evidence and Conclusion Ontology (ECO) [14] specifically to describe the level of confidence a user may have in the existence of a protein complex (Table 1), and each entry in the Complex Portal is annotated with the appropriate ECO term. As stated above, when physical interaction evidence is available, cross-references to the experimental data in an IMEx database are provided.

Complexes can be saved and downloaded in PSI-MI XML, MI-JSON, and tab-delimited ComplexTAB formats. Users can save and download bespoke lists of complexes or download all complexes (as folders per species) from the Complex Portal home page or ftp site.

The Complex Portal is a free-to-use database and is always open to offers of expert help from new curators. We also encourage the community to request new complexes for curation. Please contact us through our home page.

2 Materials

All you need is a web browser. We recommend Google Chrome, Mozilla Firefox, or Safari.

Table 1
ECO annotation codes

Accession number	Term name	When used
ECO:00000353	Physical interaction evidence used in manual annotation	If the whole complex has been purified from one species in one experiment
ECO:00005543	Biological systems reconstruction evidence by experimental evidence from mixed species used in manual assertion	If a complex has been purified in a single experiment but complex proteins come from more than one species. Substituted proteins must be homologous between species
ECO:00005610	Biological systems reconstruction evidence based on homology evidence used in manual assertion	If the complex has been inferred from an experimentally proven complex in a closely related species or from a complex with a protein derived from the same protein family with the same proven function
ECO:00005544 (child of ECO:00005610)	Biological systems reconstruction evidence based on orthology evidence used in manual assertion	If the complex has been inferred from an experimentally proven complex in a closely related species
ECO:00005546 (child of ECO:00005610)	Biological systems reconstruction evidence based on paralogy evidence used in manual assertion	If the complex has been inferred from an experimentally proven complex with proteins derived from the same protein family with the same proven function
ECO:00005547	Biological systems reconstruction evidence by experimental evidence based on inference from background scientific knowledge used in manual assertion	If the complex is generally regarded as existing but no physical interaction evidence is available for this or a related species

3 Methods

3.1 Searching for Complexes Using a Single Keyword or Identifier (See Note 1)

1. Go to www.ebi.ac.uk/complexportal.
2. Enter your search term, such as a UniProtKB AC [2], a gene symbol, or the name of a complex, into the search box and hit <Enter>, or click on the magnifier symbol (Fig. 1). A maximum of ten complexes are displayed per results page (Fig. 2) (*see* **Note 2**).
3. Scroll through the list of resulting complexes and select a complex.
4. The details page contains information pertaining to the complex:

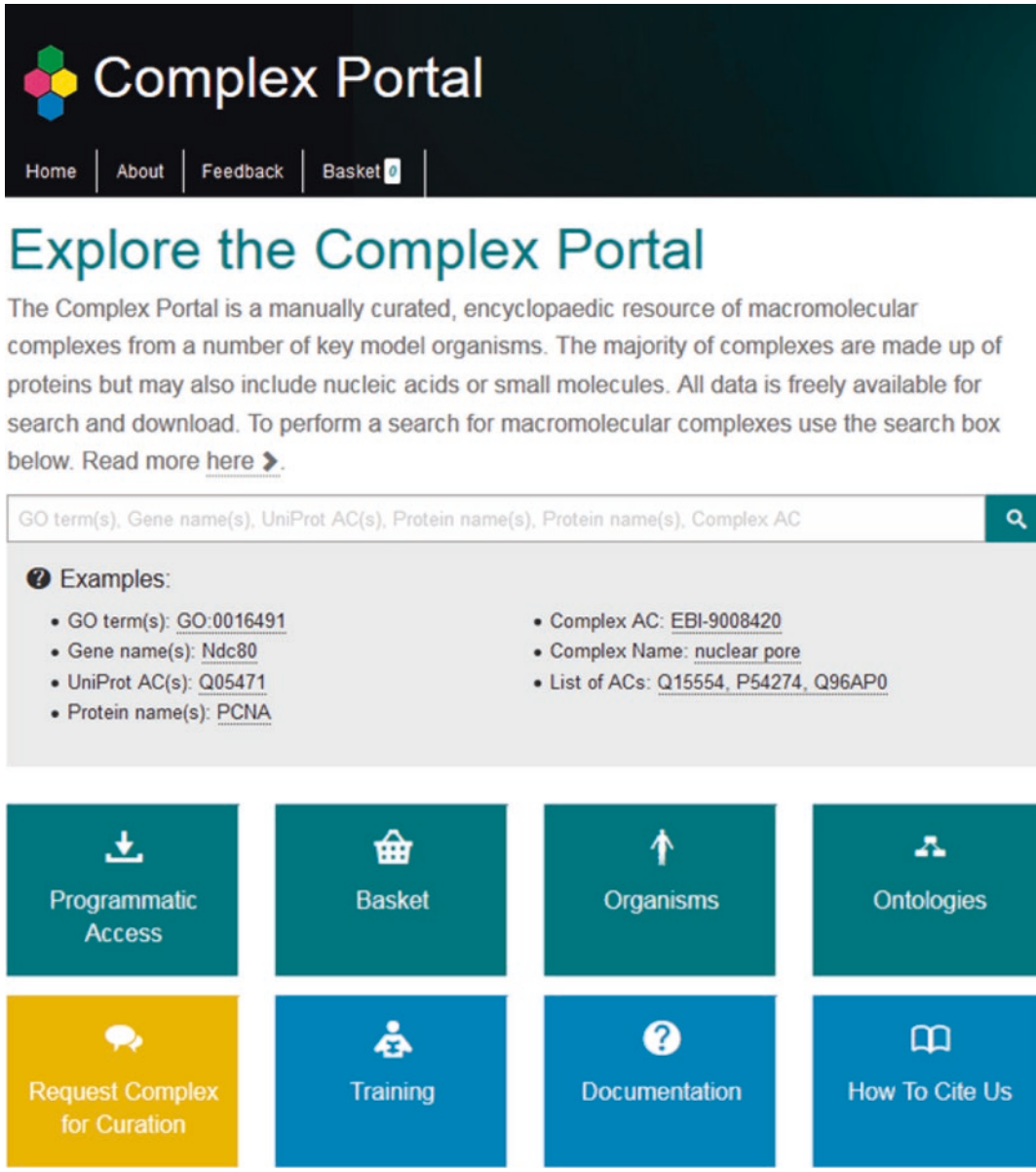


Fig. 1 Home page with search examples

- *Unique identification:* The recommended name, species, and evidence confidence level, with a link to the experimental interaction evidence when available, are displayed in the header section (Fig. 3).
- *Complex participants:* Details about the participants of each complex are displayed in both tabular and graphic style (Fig. 4). The ComplexViewer [15] creates a diagram on the fly with the most up-to-date information from the database, including information on stoichiometry and known binary

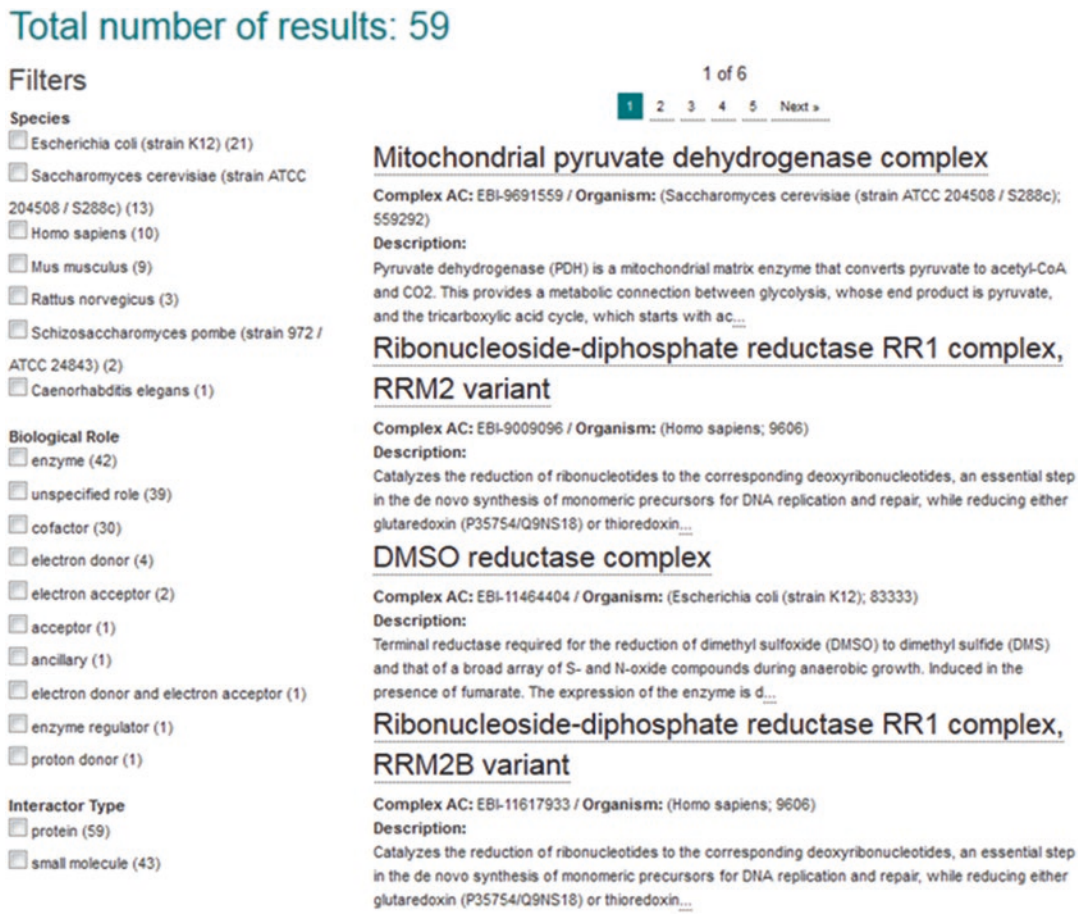


Fig. 2 Search results for GO:0016491 oxidoreductase activity. Note, there are (at time of submission) 59 results from 7 species, split over 6 results pages

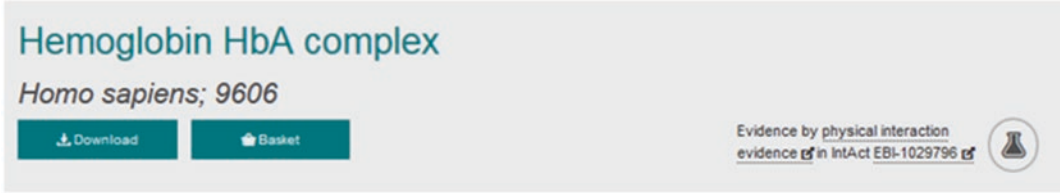


Fig. 3 Entry header for the human hemoglobin HbA complex, the main adult hemoglobin complex

interactions, to give a picture of the complex topology (*see* **Note 3**). Clicking on a protein symbol (open circle) opens up as a bar providing more detail about the binding region (Fig. 4b). Use the button below the diagram to expand all protein sequences simultaneously. Hold shift + click to zoom into the amino acid sequence. Hovering the mouse over the binding region in the protein sequence, or the binding edge,

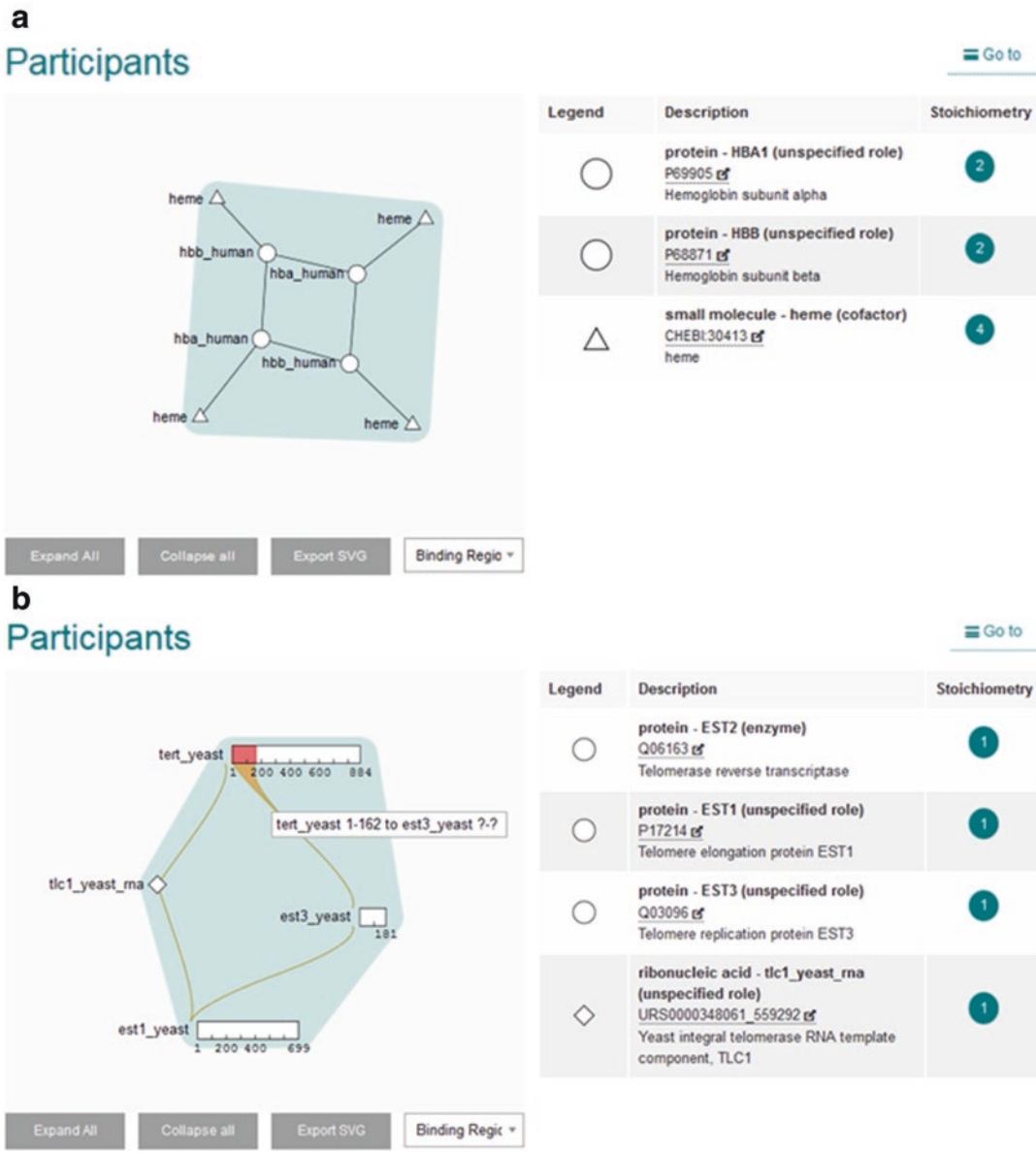


Fig. 4 ComplexViewer diagram and participant table. The diagram is interactive and depicts known topology, stoichiometry, and binary interacting regions. The table contains the legend for the symbols and links to further information on the molecules and participant stoichiometry. (**a**) Human hemoglobin contains a small molecule, the heme, which also acts as a cofactor. The topology is well defined but not the individual binding regions. (**b**) Yeast telomerase holoenzyme. This complex includes the telomerase RNA TLC1. The binding region on the enzyme TERT is defined (see pop-up); all other regions are undefined (see "?-?" region in pop-up)

displays the start and end residues of the binding regions as a pop-up. By default the viewer displays manually curated binary interacting regions; however, changing the feature type using the dropdown menu above the viewer enables the display of features imported from UniProtKB [2] records, structural

domains taken from the SUPERFAMILY database, or allows the user to visualize the stoichiometry of a complex by randomly assigning a color to each unique participant (to a maximum of 20). This enables the user to compare binary binding regions with known structural domains and other features of the protein. The ComplexViewer diagram can be reset or downloaded as svg file using the buttons below the diagram. The participant table displays the name, accession number, stoichiometry, and role of the participant in the complex. Click on the hyperlinked accession numbers to access additional data on the participant molecules in the reference database for this participant (UniProtKB [2], ChEBI [3], or RNAcentral [4]).

- *Complex function:* The function of the complex, rather than that of the individual components, is summarized in a free-text field and formalized using Gene Ontology [5, 6] annotations (Fig. 5). If applicable, Enzyme Commission numbers, ligands, and natural and commonly occurring external agonists and antagonists (e.g., ethanol, nicotine) are also listed. If a human complex has been curated into the Reactome database [12], each instance of this is displayed as an embedded Reactome pathway. As Reactome assigns a new accession number to a complex when it occurs in a different cellular location or has varying modifications (oxidations states, phosphorylations), there are often several instances of the same complex in Reactome, and they may be split across several pathways. Users can search through the complex and pathway instances using the table, and the diagram will update upon selection. Click on the Reactome logo in the diagram space to open the pathway diagram directly in the Reactome website.
- *Complex properties:* A free-text field describes several properties of the complex, such as size, molecular weight, assembly details, or specific binding characteristics (Fig. 6). Cross-referenced structures in EMBD [8] and PDB [9] are displayed using the LiteMol Viewer. Toggle through the different structures by selecting PDB accession numbers. Clicking on the PDBe logo opens the entry for this structure in PDBe (www.ebi.ac.uk/pdbe).
- *Gene expression and cellular location:* The Gene Expression Atlas (GXA) viewer has been embedded in order to display the tissue expression patterns of proteins in a given complex, as indicated by transcript level (Fig. 7) (*see* **Note 4**). The subcellular location is indicated using Gene Ontology [5, 6] Cellular Compartment terms.
- *Diseases:* If defects in a human complex cause a disease, details are provided in a free-text annotation and linked to reference resources such as the Experimental Factor Ontology (EFO) [16], Orphanet [17], or Human Phenotype Ontology (HP) [18].

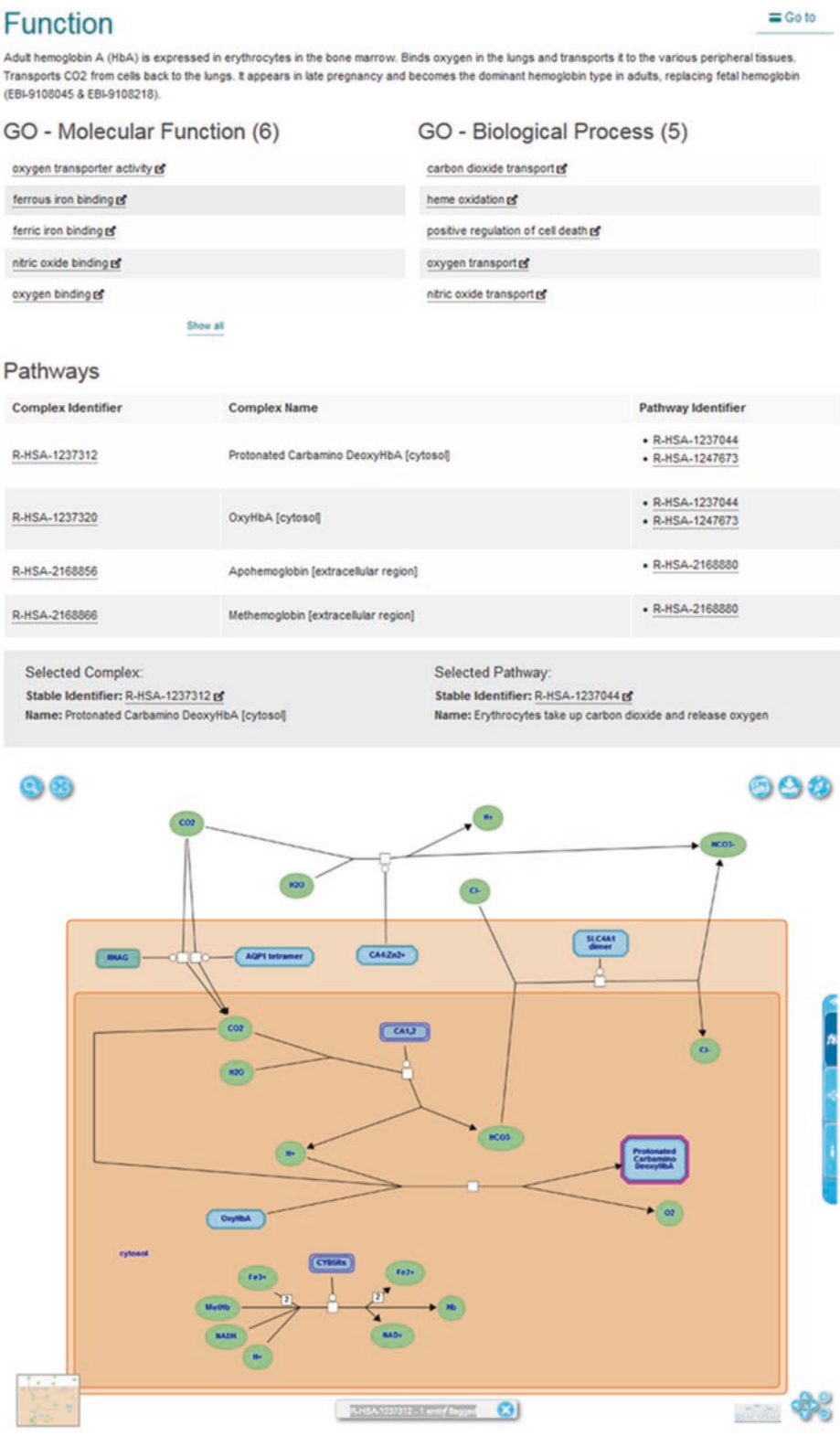


Fig. 5 Complex function. Note that any table with more than five rows is collapsed by default to avoid very long pages. Select a Reactome complex identifier to navigate through different complexes and select the pathway identifier to find complexes in different pathways

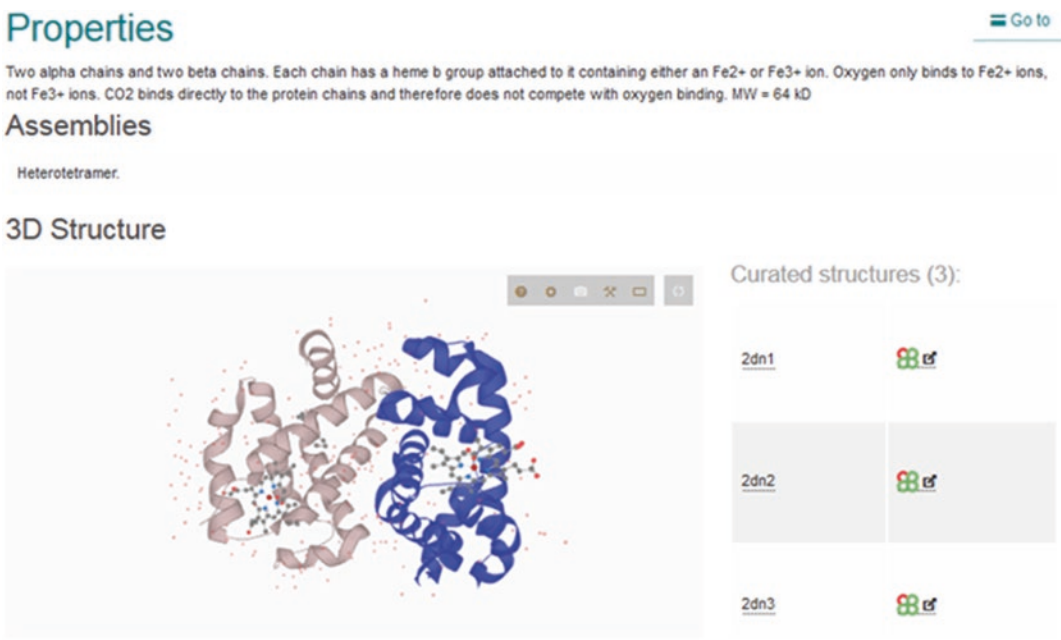


Fig. 6 Complex properties. Toggle through the available structures for the human hemoglobin one by one by selecting the appropriate accession number

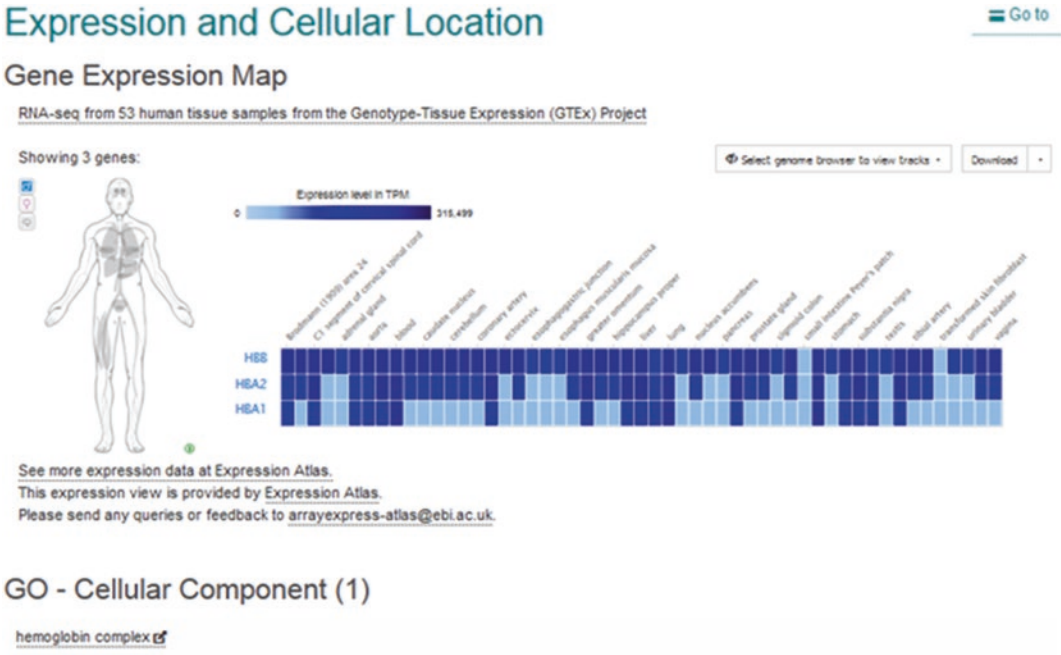


Fig. 7 Gene expression data and subcellular location. Click on the GXA experiment title at the top of the GXA viewer to access the full entry in the Expression Atlas (http://www.ebi.ac.uk/gxa). Cellular locations are annotated to GO terms

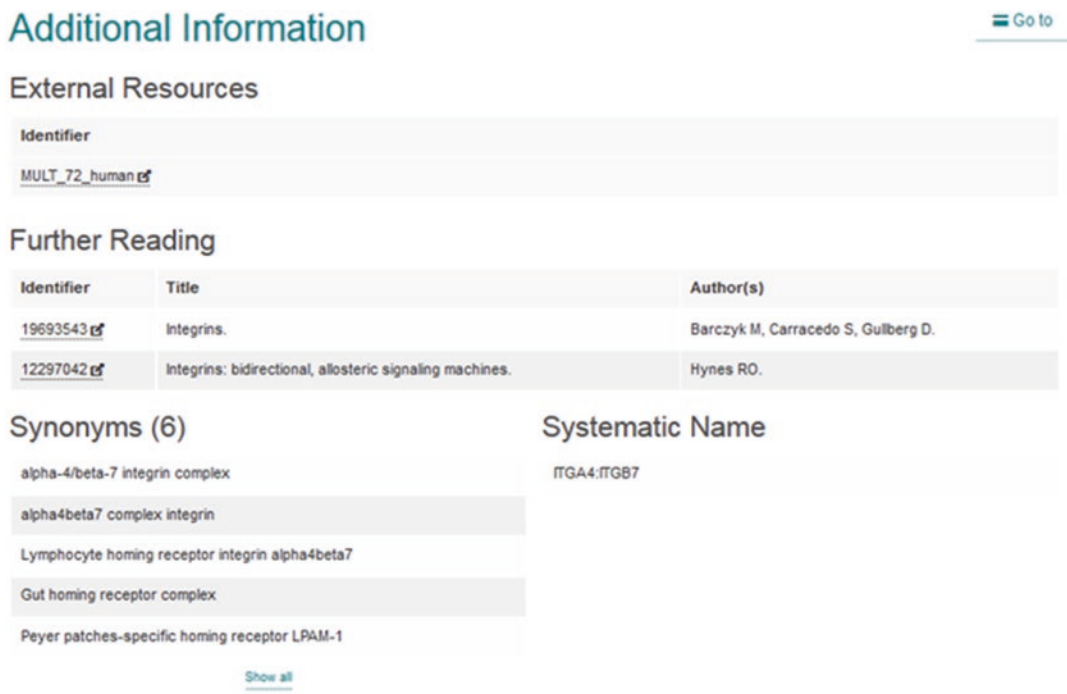


Fig. 8 External identifiers, publications, and other names

Cross-references to ChEMBL [7] provide additional information on complexes with drug-binding evidence.

- *External resources:* This section contains identifiers for this complex in any other external databases, e.g., MatrixDB (http://matrixdb.univ-lyon1.fr/).
- *Additional reading:* Lists publications linking to experimental and functional evidence and additional information such as reviews with links to Europe PMC [19] (Fig. 8).
- *Synonyms and systematic names:* All alternative names a complex may be known by are listed at the end of the entry. A systematic name is also added. The systematic name is essentially a concatenation of gene names in alphanumeric order and, if known, details of their stoichiometry (Fig. 8).

3.2 Searching for Complexes Using a List of Identifiers

Searching with a list of identifiers by default performs an "OR" search, meaning it returns all complexes that contain at least one participant that matches any one of the search terms. Identifiers should either be comma, paragraph, tab, or space delimited. To retrieve only complexes that contain all of the identifiers in a list, connect the search terms with the Boolean AND (e.g., Q16602 O60894* vs Q16602 AND O60894*) (*see* **Note 5**).

3.3 Searching for Complexes by Species

- *Searching by organism:* From the home page, click on the organism tile (Fig. 1), and find your target species on this list. Click on the species pictogram, which will perform a search for all complexes from this species. Alternatively, the data may be accessed directly via the programmatic access tile, selecting the required download format, followed by the appropriate species file.
- *Text search:* Using either the appropriate scientific name, common name, or taxonomy identifier returns all complexes for this species. Common names should, however, be used with care, as these may often be used as synonyms in the names of other species.
- *Targeted field search:* Prefixing the search term with species: followed by the UniProt short label or NCBI taxID limits the search to the organism field, e.g., species:human or species:9606.

3.4 Searching for Complexes by Other Identifiers

Perform a search using identifiers, including their synonyms, from the following databases: ChEMBL [7], ChEBI [3], EMDB [8], Gene Ontology [5, 6], IntEnz [10], MatrixDB [11], PDB [9], Reactome [12], RNAcentral [4], and UniProtKB [2] (including specific isoforms and PRO chain IDs). For example, a search for GO:0016491 returns all complexes that are annotated to oxidoreductase activity (Fig. 2).

3.5 Performing an Advanced Search

- *Text string:* You can search with any given string of text. The search engine regards these all as individual search terms and returns complexes containing any one of these terms in its entry. Use the Boolean AND operator (*see* **Note 5**) to search for complexes that contain all of the search terms in a string or search for an exact match by adding double quotes around the search string (e.g., alpha AND polymerase returns many complexes but "alpha polymerase" returns none).
- *Complex query language (CQL):* To perform more complex queries, search fields can be specified using a query syntax based on Lucene. For example, species:9606 AND alias:cdk1 returns only human complexes containing CDK1 as participant. For further details of this syntax, go to http://www.ebi.ac.uk/complexportal/documentation/query_syntax.

3.6 Saving and Downloading Complexes and Using the Web Service

- *Saving and downloading complexes from the details page:* Click on the basket sign in the header section of the details page to save the complexes in the basket (Fig. 3) (*see* **Note 6**). All saved complexes can be accessed by selecting the basket link in the page header (Fig. 1) (*see* **Note 7**). Complexes can be downloaded in a range of formats from the link next to the basket icon.
- *Download access from the home page:* Select the programmatic access tile on the home page (Fig. 1) and then select the required format (*see* **Note 8**).

- *By ftp:* Access all complexes via ftp in XML3.0 format (ftp://ftp.ebi.ac.uk/pub/databases/intact/complex/current/psi30/) or tab-delimited format (ftp://ftp.ebi.ac.uk/pub/databases/intact/complex/current/complextab/) (*see* **Note 8**).
- *By web service:* Our web service can be accessed at http://www.ebi.ac.uk/intact/complex-ws/ and currently provides three different methods: <search>, <details> and <export>. Each method returns an MI-JSON file.

4 Notes

1. The most specific results are retrieved when using accession numbers (ACs). We recommend UniProtKB [2], ChEBI [3], or RNAcentral [4] ACs. To retrieve all forms of a protein, including isoforms and splice variants, UniProt ACs must be appended with a wildcard (*), e.g., O60894*. Keyword searches may return spurious results from mentions of the term in descriptions or synonyms. Gene symbols may also be ambiguous, e.g., CDC2 is a gene symbol synonym for mammalian CDK1 [P06493] and yeast POL3 [P15436].
2. Results are ranked using a custom search engine algorithm, with the best matched complexes ranked top. When searching with a text term, the ranking may not be obvious as the term may appear in an unexpected field such as a gene or protein synonym. The use of the filters on the left-hand side column will refine the search by selecting a target species, the molecule type present in the complex (proteins, small molecules, different types of nucleic acids) or the biological role of an interactor (e.g., enzyme or electron donor).
3. In the future there will be links directly from the edges in the ComplexViewer to the experimental evidence. At the time of publication of this article, this feature is still under development.
4. Currently, the Gene Expression Atlas (GXA) viewer only displays expression data for a single experiment per species. The experiment is chosen by the GXA team of curators. In a future instance of the viewer, the user may be able to choose the experiment from a range of GXA experiments, including protein expression data.
5. You can use the Boolean operators AND, OR, and NOT in any search. If you search on a list and do not identify any operators, the search engine performs a default search using OR between each term in the list.
6. Complexes are only saved in the current instance of your browser; updating the browser version or clearing the cache will empty the basket.

7. There will also be an option to save selected complexes from the search results page, but that function was not yet implemented at the time of print.
8. We recommend using the PSI-MI XML3.0 format rather than PSI-MI XML2.5 format as it uses a much more intuitive data structure for the representation of complexes and also contains more information. The MI-JSON format contains the same data as the XML file but was developed for access by visualization tools and web services. The ComplexTAB allows a tab-delimited representation of the protein complexes and is recommended for use by large-scale data producers looking for clusters of molecules in, for example, transcriptomic or proteomic data that correspond to protein complexes. Further details on this format are given at ftp://ftp.ebi.ac.uk/pub/databases/intact/complex/current/complextab/README.htm. Complexes are organized on the Complex Portal ftp site in one species per folder.

Acknowledgment

This work was supported by European Molecular Biology Laboratories core funding, BBSRC Midas Grant (BB/L024179/1), and the Wellcome Trust (WT101477MA) (Complex Viewer, as part of PRIDE Atlas). We thank Mila Rodrigues and Luana Perfetto for their curation efforts, Maximilian Koch for developing the website, Xavier Watkins and Sangya Pundir for their user experience design expertise, and Colin Combe for writing ComplexViewer for this project.

References

1. Meldal BHM, Forner-Martinez O, Costanzo MC, Dana J, Demeter J, Dumousseau M, Dwight SS, Gaulton A, Licata L, Melidoni AN, Ricard-Blum S, Roechert B, Skyzypek MS, Tiwari M, Velankar S, Wong ED, Hermjakob H, Orchard S (2015) The complex portal–an encyclopaedia of macromolecular complexes. Nucleic Acids Res 43(Database issue):D479–D484. https://doi.org/10.1093/nar/gku975
2. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45(D1):D158–D169. https://doi.org/10.1093/nar/gkw1099
3. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Res 44(D1):D1214–D1219. https://doi.org/10.1093/nar/gkv1031
4. The RNAcentral Consortium (2017) RNAcentral: a comprehensive database of non-coding RNA sequences. Nucleic Acids Res 45(D1):D128–D134. https://doi.org/10.1093/nar/gkw1008
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29

6. Gene Ontology Consortium (2015) Gene ontology consortium: going forward. Nucleic Acids Res 43(Database issue):D1049–D1056. https://doi.org/10.1093/nar/gku1179
7. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res 42(Database issue):D1083–D1090. https://doi.org/10.1093/nar/gkt1031
8. Lawson CL, Patwardhan A, Baker ML, Hryc C, Garcia ES, Hudson BP, Lagerstedt I, Ludtke SJ, Pintilie G, Sala R, Westbrook JD, Berman HM, Kleywegt GJ, Chiu W (2016) EMDataBank unified data resource for 3DEM. Nucleic Acids Res 44(D1):D396–D403. https://doi.org/10.1093/nar/gkv1126
9. Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S (2017) Protein data bank (PDB): the single global macromolecular structure archive. Methods Mol Biol 2017(1607):627–641. https://doi.org/10.1007/978-1-4939-7000-1_26
10. Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, Bairoch A, Schomburg D, Tipton KF, Apweiler R (2004) IntEnz, the integrated relational enzyme database. Nucleic Acids Res 32(Database issue):D434–D437
11. Launay G, Salza R, Multedo D, Thierry-Mieg N, Ricard-Blum S (2015) MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. Nucleic Acids Res 43(Database issue):D321–D327. https://doi.org/10.1093/nar/gku1091
12. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P (2016) The reactome pathway knowledgebase. Nucleic Acids Res 44(D1):D481–D487. https://doi.org/10.1093/nar/gkv1351
13. Orchard S, Kerrien S, Abbani S, Aranda B, Bhate J, Bidwell S, Bridge A, Briganti L, Brinkman FS, Cesareni G, Chatr-aryamontri A, Chautard E, Chen C, Dumousseau M, Goll J, Hancock RE, Hannick LI, Jurisica I, Khadake J, Lynn DJ, Mahadevan U, Perfetto L, Raghunath A, Ricard-Blum S, Roechert B, Salwinski L, Stümpflen V, Tyers M, Uetz P, Xenarios I, Hermjakob H (2012) Protein interaction data curation: the international molecular exchange (IMEx) consortium. Nat Methods 9(4):345–350. https://doi.org/10.1038/nmeth.1931
14. Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M (2014) Standardized description of scientific evidence using the evidence ontology (ECO). Database (Oxford) 2014:bau075. https://doi.org/10.1093/database/bau075
15. Combe CW, Sivade MD, Hermjakob H, Heimbach J, Meldal BHM, Micklem G, Orchard S, Rappsilber J (2017) ComplexViewer: visualization of curated macromolecular complexes. Bioinformatics. 33(22):3673–3675. PMID:29036573. https://doi.org/10.1093/bioinformatics/btx497
16. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H (2010) Modeling sample variables with an experimental factor ontology. Bioinformatics 26(8):1112–1118. https://doi. org/10.1093/bioinformatics/btq099
17. www.orpha.net/
18. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJ, DeMare LE, Devereau AD, de Vries BB, Firth HV, Freson K, Greene D, Hamosh A, Helbig I, Hum C, Jähn JA, James R, Krause R, F Laulederkind SJ, Lochmüller H, Lyon GJ, Ogishima S, Olry A, Ouwehand WH, Pontikos N, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI, Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E, Veltman MW, Vulliamy T, Yu J, von Ziegenweidt J, Zankl A, Züchner S, Zemojtel T, Jacobsen JO, Groza T, Smedley D, Mungall CJ, Haendel M, Robinson PN (2017) The human phenotype ontology in 2017. Nucleic Acids Res 45(D1):D865–D876. https://doi.org/10.1093/nar/gkw1039
19. Europe PMC Consortium (2015) Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Res 43(Database issue):D1042–D1048. https://doi.org/10.1093/nar/gku1061

Chapter 25

Automated Computational Inference of Multi-protein Assemblies from Biochemical Co-purification Data

Florian Goebels, Lucas Hu, Gary Bader, and Andrew Emili

Abstract

Biology has amassed a wealth of information about the function of a multitude of protein-coding genes across species. The challenge now is to understand how all these proteins work together to form a living organism, and a crucial step for gaining this knowledge is a complete description of the molecular "wiring circuits" that underlie cellular processes. In this chapter, we describe a general computational framework for predicting multi-protein assemblies from biochemical co-fractionation data.

Key words Protein-protein interaction, Bioinformatics, Machine learning, Systems biology, Protein interaction prediction, Protein complex prediction, Python, Docker, Cytoscape

1 Introduction

Previously, in Chapter 12, we discussed in detail how to plan and execute the co-fractionation (e.g., non-denaturing chromatography) part of the biochemical purification/mass spectrometry (BP/MS) experimental pipeline, while in this chapter we provide an in-depth description of the computational part required for proteomics data processing, analysis, and interpretation. Specifically, we describe EPIC (**E**lution **P**rofile-based **I**nference of Protein **C**omplex Membership), a software toolkit which can automatically generate confidence binary protein interactions and predict the memberships of corresponding stable multi-protein assemblies from raw co-elution proteomics data [1]. The EPIC is accessible to the public via a GitHub (https://github.com/BaderLab/EPIC) or Docker Hub (https://hub.docker.com/r/baderlab/bio-epic/) repository.

Since it does not rely of achieving purity, co-fractionation is a practical but imperfect experimental approach to characterize multi-protein complexes. Our computational workflows have been optimized to minimize the number of spurious protein pairs that are predicted to interact because they simply happen to co-elute at the same time (due to similar biophysical behavior during

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_25,

chromatography) but which are actually functionally unrelated (we refer to such events as the "chance co-elution" problem). Toward this end goal, we apply basic statistical criteria to measure protein similarity based on their respective biochemical fractionation profiles, followed by more sophisticated machine learning to exploit publicly available supporting functional association evidence to guide the selective filtering of biologically irrelevant correlations.

We demonstrated the practical utility and real-world performance of this co-fractionation data analysis pipeline, which was first used to predict 13,993 high-confidence physical interactions among 3006 stable protein complexes in human [2] and in a follow-up experiments that identified 981 conserved metazoan complexes [3]. Below, we outline implementation of the stand-alone EPIC software designed to facilitate such analyses by biologists lacking computational expertise.

2 Materials

As EPIC is a computational pipeline, the only physical equipment required is suitable computer infrastructure (e.g., Linux- or Mac OSX-enabled machine). However, we provide suggestions for implementation as well as minimal and recommended specs. Moreover, we list both required and optional software for running EPIC.

2.1 Equipment

1. Working computer (Mac OSX/Linux-based) (*see* **Note 1**).
   - Minimal: one core, 8 GB RAM.
   - Recommended: four cores, 8 GB RAM.
2. Internet connection (optional).
   - Required for automatic generation of reference data set and automatic download of STRING and GeneMANIA.
   - Alternatively the user can supply own reference clusters and functional annotation scores as flat file (*see* below for file formats).

2.2 Supplementary Software

1. Docker (mandatory).
2. Cytoscape [4] (optional but highly recommended) (*see* **Note 2**).
3. We highly recommend basic understanding for navigating a Jupyter script.

2.3 File Formats

There are three main types of input files used in EPIC: elution profile data, reference protein complexes, and functional annotation data. Example files for Worm (taxid 6239) can be found in the test_data directory inside the EPIC GitHub repository (https://github.com/BaderLab/EPIC/tree/master/test_data).

1. Elution Profile Data

This is a tab delimited file or data matrix containing the elution profiles for all the proteins detected by mass spectrometry in one distinct co-fractionation experiment. For example data, see https://github.com/BaderLab/EPIC/tree/master/test_data/elution_profiles. Multiple experiments will result in multiple co-elution profiles (i.e., one file for each experiment). The header is located on the first line and contains the names for each fraction, while each subsequent row contains the various protein IDs (accessions/descriptions) and the corresponding detection values (e.g., spectral counts) recorded in each fraction.

2. Reference Protein Complexes (Optional)

   The user may supply a custom set of reference protein complexes (e.g., CORUM [5], IntAct [6], GO [7]) for use in training the EPIC scoring algorithm (*see* **Note 3**). In this file, each complex is summarized in one line by concatenating all member protein IDs with tab-delimited characters. Example reference complexes for Worm can be found here https://github.com/BaderLab/EPIC/blob/master/test_data/Worm_reference_complexes.txt.

3. Functional Annotation Data (Optional)

   EPIC uses functional associations as additional features to minimize chance co-elution, and in this step the user can provide a predefined set of functional associations (*see* **Note 4**). The data in this file should be on protein interaction level and will be added as additional features to each candidate PPI without further modifying the added features. In this file each column represents a functional association score, and each row consists of protein pair followed by available functional association scores (columns are tab separated). This file has a header row, which contains each column respective functional annotation score name. **Note 4** contains some examples for species-specific functional annotation resources, and Subheading 3.3.5 lists the default sources used in EPIC (e.g., https://raw.githubusercontent.com/BaderLab/EPIC/master/test_data/Wormnet_funanno.txt).

3 Methods

The EPIC software mostly runs automatically, and thus the most labor-consuming part for establishing the computational scoring pipeline is setting up docker and starting EPIC. However, this step can be easily completed within an hour. EPIC runs automatically and has on average a runtime of 40 min per co-elution score per experiment, divided by the number of available computer cores. The most computationally heavy part is generating the co-elution scores for all pair-wise protein combinations.

3.1 Installing Required Software

To run EPIC, it is mandatory to install docker, which is a lightweight virtual machine, which will enable operation of the entire EPIC pipeline.

- Docker. For Macintosh instructions see https://docs.docker.com/docker-for-mac/. For Linux see https://docs.docker.com/engine/installation/.
- Cytoscape. Cytoscape is available from http://www.cytoscape.org/.

Once docker is installed, one needs to change the assigned memory to at least 6 GB. This is achieved by selecting docker, followed by preference, and then selecting advanced options.

3.2 Installing EPIC

Once docker is installed, EPIC can be installed. This step can take time, depending on the available Internet speed, since the EPIC image is roughly 8 GB in size.

1. Open a terminal.
2. Enter the following command:

```
$ docker pull baderlab/bio-epic
```

3. Create a folder on your machine named EPIC.
4. Within this folder, create another subfolder for data (e.g., My_EPIC_PROJ).
5. Move all project-relevant co-fractionation data files into this folder (e.g., copy chromatographic elution files into the MY_EPIC_PROJ folder).

3.3 Running EPIC

1. Open/select a terminal window.
2. Navigate to the previously generated EPIC folder (*see* **Note 5**).
3. Download the EPIC start script (https://github.com/BaderLab/EPIC/blob/master/src/start-EPIC), and put it in the EPIC folder, and double click the file.
4. Open a browser and enter http://localhost:8888/tree.
5. Once the web page is finished loading, click on the EPIC.ipynb symbol.
6. When running EPIC for the first time, it is recommended to go through the EPIC script in a step-by-step wise manner by repeatedly pressing the play button.
7. Press play until an input directory selection appears (*see* Fig. 1a), and select a folder from the list (e.g., MY_EPIC_PROJ). From now on, we no longer indicate if a user has to press play to reach the next input mask, rather we describe what to do at each input mask.

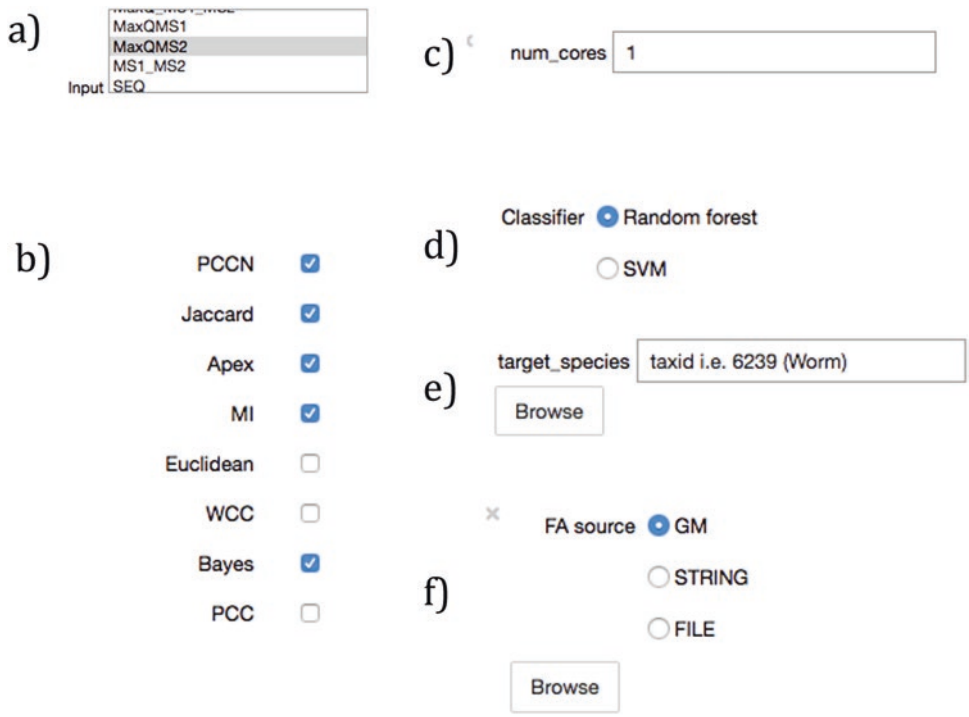


Fig. 1 Overview of different widgets for configuring the EPIC Jupyter notebook, they show options for selecting: (**a**) input data directory, (**b**) co-elution scores, (**c**) number of cores, (**d**) machine learning classifier, (**e**) reference data, and (**f**) functional annotation

3.3.1 Selecting Input Features

There are eight protein similarity score features available in total (Fig. 1b): Pearson with Poisson noise (PCCN), Jaccard, Apex, Mutual Information (MI), Euclidean, weighted cross correlation (WCC), Bayes correlation, and Pearson correlation coefficient. A short description for each feature follows below (citations provided as needed):

PCCN: To reduce the spurious correlations caused by fractions with low peptide counts, PCCN correlation is calculated by averaging multiple Pearson correlation values that are computed by taking the raw counts and adding a round of small value Poisson noise to them [2].

Jaccard: Determines co-elution based on the number of overlapping fractions two proteins are detected together in.
Apex: This score is one when the largest or peak signal (highest spectral count) of each of two proteins occurs in the same fraction together or else zero if otherwise [2].

Mutual Information: The mutual dependence between two variables (e.g., protein spectral counts) is used to identify statically significant protein pairs.

Euclidean: The Pythagorean theorem is used to calculate the Euclidean distance between two proteins by considering each fraction as an independent dimension.

WCC: A weighted correlation-based metric takes into account small possible shifts in the patterns of two proteins that co-fractionate together [8].

Bayes: Bayes correlation identifies statistical significant protein correlations [9].

PCC: Pearson correlation coefficient of two proteins calculated based on their respective co-elution profile patterns.

3.3.2 Number of Cores

Increasing the number of computer cores (if available) will greatly reduce the runtime of EPIC (Fig. 1c).

3.3.3 Machine Learning Classifier

Currently supported options are support vector machine [10] and random forest classifiers [11]; we recommend initially using the random forest classifier (Fig. 1d).

3.3.4 Reference Data

The user can either have reference complexes automatically generated from CORUM, GO, and IntAct by supplying a valid taxonomic (species) ID (taxid) (Fig. 1e) or supply a custom set of reference protein assemblies (*see* Subheading 2.3).

3.3.5 Functional Annotation Data

Analogous to the reference data, the user can either have it automatically obtained using EPIC (Fig. 1f) or by supplying custom data (*see* Subheading 2.3). For automatic generation the user can select either to use STRING (https://string-db.org/) [12] or GeneMANIA (http://genemania.org/) [13] as source. When using STRING we exclude "experimental," "database," and "combined_score" scores from the database to avoid circular reasoning in the training phase. We recommend using GeneMANIA if the target species is available in both databases, since we observed better performance for predicting Worm protein complexes when using GeneMANIA.

Once this step is completed, the user can either run the script cell by cell (pressing run cell and select next, i.e., play button) or run the entire EPIC script by selecting run cell and below. When running for the first time, we recommend to run cell by cell, so the user can check the output for each cell, and for repeated reruns the user can select "run cell and below" to run all cells automatically without human supervision.

3.4 EPIC Output

Once the Jupyter script is completed, it will generate an initial graphical overview of the generated protein clusters using Cytoscape.js in its second to last step. At the end, EPIC will generate an output folder with various result files in a specified input directory named My_EPIC_PROJ_out, including the following files:

1. Out.scores.txt: Raw co-elution scores for all candidate PPIs.
2. Out.roc.png – precision-recall curve for predicted PPIs [14].

3. Out.pr.png – receiver operating characteristic (ROC) curve for predicted PPIs [15].
4. Out.rf.cutoff.png – shows precision and recall values across all confidence cut-off values.
5. Out.pred.txt – predicted protein interactions with classifier confidence values.
6. Out.clust.txt – predicted multi-protein clusters.

The Out.scores.txt contains the features used for predicting the protein associations, while the Out.roc.png, Out.pr.png, and Out.rf.cutoff.png files give an overview of the classifiers performance (*see* **Note 6**). The Out.pred.txt and Out.clust.txt contain the main predicted outputs (PPI and clusters) generated by EPIC.

3.4.1 EPIC with Cytoscape

The last cell of the Jupyter script is used to visualize the generated protein clusters using Cytoscape. This step is optional but recommended.

1. Start the locally installed Cytoscape on your machine. This is done outside of the Jupyter script.
2. In Cytoscape, select Apps, and then select app manager.
3. In the search mask, enter clusterMaker2, and select clusterMaker2 from the selection, and press install. This step needs only to be performed once.
4. Switch back to the Jupyter script, and run the last cell, followed by switching back to Cytoscape.
5. In Cytoscape, select Layout, followed by yFiles Layout, and finally select organic. Now there should be one group of nodes (proteins) per cluster showing all associated interactions.
6. Use the mouse to select a cluster, and then select Apps and clusterMaker visualizations, and finally select JTree HeatMapView.
7. In the "Node attributes for cluster" field, select all the fractions that are displayed. Check the "use only selected nodes/edges for clusters" box, and press the OK button.

4 Notes

1. The central component for improving the runtime of EPIC is assigning it more cores if available. It is **important to assign the number of cores to the docker engine** so that EPIC can use those cores. For most normal use cases where you have 4–5 experiments (around 1000 fractions), EPIC can completely run between a night and an afternoon.
2. The main advantage of using Cytoscape with EPIC is visualizing both the network of protein complexes and PPIs that are generated, as well as the supporting co-fractionation data for

each putative protein member in heat-map format to confirm profile similarity. Each edge in the Cytoscape network provides the EPIC derived confidence score, and the user can adjust edge thickness (cutoff values) to define data consistency within a cluster. No prior knowledge of Cytoscape is required; however, it is encouraged to become familiar with network style and layout formats (see http://wiki.cytoscape.org/Cytoscape_User_Manual#Visual_Styles and http://wiki.cytoscape.org/Cytoscape_User_Manual/Navigation_Layout).

3. When supplying a custom set of reference complexes, there are certain aspects the user needs to be aware of. First, the automatically generated reference set is based on experimentally inferred complexes retrieved from the CORUM, IntAct, and GO curation databases, so if the user wants to use a custom set, it is recommended to use different sources. The most important thing to be aware of is to refrain using complexes derived from functional genomics, since using this resource will result in circular reasoning because EPIC uses functional-based features for boosting PPI scores. In case the user wants to use complexes derived from functional annotation, then it is recommended to run EPIC using only experimental evidences. Also, we liked to note when generating the reference set, the user should not use complexes derived using non-biochemically based experimental methods (e.g., yeast two hybrid assays) because these tend to overlap poorly with biochemical data (e.g., co-fractionation). In brief, we highly recommend users to generate their reference complexes using complexes that are manually curated and were verified by low-throughput experimental methods.
4. When deciding which functional associations to use for enhancing learning/scoring, we typically observed best performance using species-specific and tissue-specific data (when available). For example, when predicting complex membership by co-fractionation analysis of *H. sapiens*, *C. elegans*, or *M. musculus* protein extracts, we observed optimal performance using supporting functional associations from HumanNet (http://www.functionalnet.org/humannet/about.html) [16], WormNet (http://www.functionalnet.org/wormnet/) [17], and MouseNet (http://www.functionalnet.org/mousenet/) [18], respectively. If wanting to combine multiple resources to boost prediction confidence, the user needs to combine these data into one single functional annotation file; public data integration tools like GeneMANIA (http://genemania.org/) [13] facilitate this.
5. A user can select any folder as an input folder; however, it is highly recommended to create individual project folders within EPIC.

6. Precision-recall curves provide an overview of classifier performance indicating how many protein associations can be classified with a certain precision. Receiver operating characteristic (ROC) curves indicate how well the classifier can distinguish false-positive from false-negative interactions. The precision-recall plot for various classifier confidence values is intuitive, since it shows the precision (i.e., relative fraction of predicted interactions that are correctly classified) and recall (i.e., relative fraction of positive interactions that are correctly classified) across all possible classifier confidence cutoff values.

References

1. Lucas Hu Ming FG, Cuihong Wan, Gary Bader, Andrew Emili (2018) EPIC: elution profile-based inference of protein complex membership. Under revision.
2. Havugimana PC et al (2012) A census of human soluble protein complexes. Cell 150(5):1068–1081
3. Wan C et al (2015) Panorama of ancient metazoan macromolecular complexes. Nature 525(7569):339–344
4. Shannon P et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
5. Ruepp A et al (2010) CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res 38(suppl 1):D497–D501
6. Kerrien S et al (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40(D1):D841–D846
7. Gene Ontology C (2015) Gene ontology consortium: going forward. Nucleic Acids Res 43(Database issue):D1049–D1056
8. Wehrens, R. and M.R. Wehrens, Package 'wccsom'. 2015
9. Sánchez-Taltavull D et al (2016) Bayesian correlation analysis for sequence count data. PLoS One 11(10):e0163595
10. Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300
11. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
12. Szklarczyk D et al (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res 45(D1):D362–D368
13. Warde-Farley D et al (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 38(suppl_2):W214–W220
14. Davis J and Goadrich M 2006. The relationship between precision-recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning. ACM
15. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36
16. Lee I et al (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res 21(7):1109–1121
17. Lee I et al (2010) Predicting genetic modifier loci using functional gene networks. Genome Res 20(8):1143–1153
18. Kim WK, Krumpelman C, Marcotte EM (2008) Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. Genome Biol 9(1):S5

Chapter 26

A Multiscale Computational Model for Simulating the Kinetics of Protein Complex Assembly

Jiawen Chen and Yinghao Wu

Abstract

Proteins fulfill versatile biological functions by interacting with each other and forming high-order complexes. Although the order in which protein subunits assemble is important for the biological function of their final complex, this kinetic information has received comparatively little attention in recent years. Here we describe a multiscale framework that can be used to simulate the kinetics of protein complex assembly. There are two levels of models in the framework. The structural details of a protein complex are reflected by the residue-based model, while a lower-resolution model uses a rigid-body (RB) representation to simulate the process of complex assembly. These two levels of models are integrated together, so that we are able to provide the kinetic information about complex assembly with both structural details and computational efficiency.

Key words Protein complex assembly, Multiscale modeling, Coarse-grained simulation, Protein association rate, Kinetic Monte Carlo, Diffusion-reaction algorithm

1 Introduction

Proteins form high-order complexes to carry out their diverse functions in cells [1, 2]. In order to maintain the proper functions, natural evolution has developed specific assembling pathways for these complexes [3, 4]. Any mistake along the pathways of complex assembly can lead to severe biological consequences [5]. Moreover, in a crowded cellular environment, the assembly of protein complexes is often under kinetic, rather than thermodynamic, control [6, 7]. Therefore, to study the kinetics of protein complex assembly is of paramount importance. Unfortunately, relative to the intensive studies made for the structural determination of protein complexes, the dynamic aspects of their assembling pathways have just started to be understood. In addition to the recently developed experimental techniques such as super-resolution microscopy [8], electron microscopy [9], and native mass spectrometry [10], a large variety of computational models have also been developed to simulate the

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_26,

association of protein complexes. However, among these models, high-resolution methods based on molecular dynamic simulations can hardly approach the full time scale of assembly processes for large protein complexes [11–27]. In contrast, low-resolution models fail to provide a quantitative description of the structure and energetics of protein complexes [28–35].

In this chapter, we outline a computational framework to simulate the kinetics of protein complex assembly. The framework consists of models on two different scales [36]. The higher-resolution simulation uses residue-based coarse-grained (CG) models of protein structure to evaluate the binding rates between each pair of subunits in a complex, whereas the lower-resolution model uses a rigid-body (RB) representation to simulate the process of complex assembly. By feeding the binding rates calculated from the residue-based simulations into the lower-resolution simulations, two levels of models are integrated together so that assembly of specific protein complexes can be studied with both structural detail and computational efficiency.

2 Materials

2.1 Information Needed as Input Parameters

The following information is needed as input parameters for simulations:

1. The structure (atomic coordinates) of the entire protein complex in PDB format.
2. The translational and rotational diffusion constants of each subunit in the complex. These constants can be obtained by curve fitting to the data that were calculated by a precise boundary element method [37, 38].
3. The dissociation constants (K_d) which quantify the binding stability for all pairs of individual subunits in the complex. For instance, a heterotrimer that contains two types of subunits (A and B) includes two types of binding interfaces (Fig. 1a). One is between subunit A and B, while the other is between two subunits A. The dissociation constants through both AB and AA binding are needed.
4. The on rates (k_{on}) of binding which quantify the kinetics of association for all pairs of individual protein subunits in the complex.

2.2 Residue-Based Model for Simulation Protein Association

2.2.1 Model Representation

The atomic structure of proteins was reduced to a simplified model in which each residue is represented by two sites [39]. One is the position of its Cα atom, while the other is the representative center of a side chain selected based on the specific properties of a given amino acid.

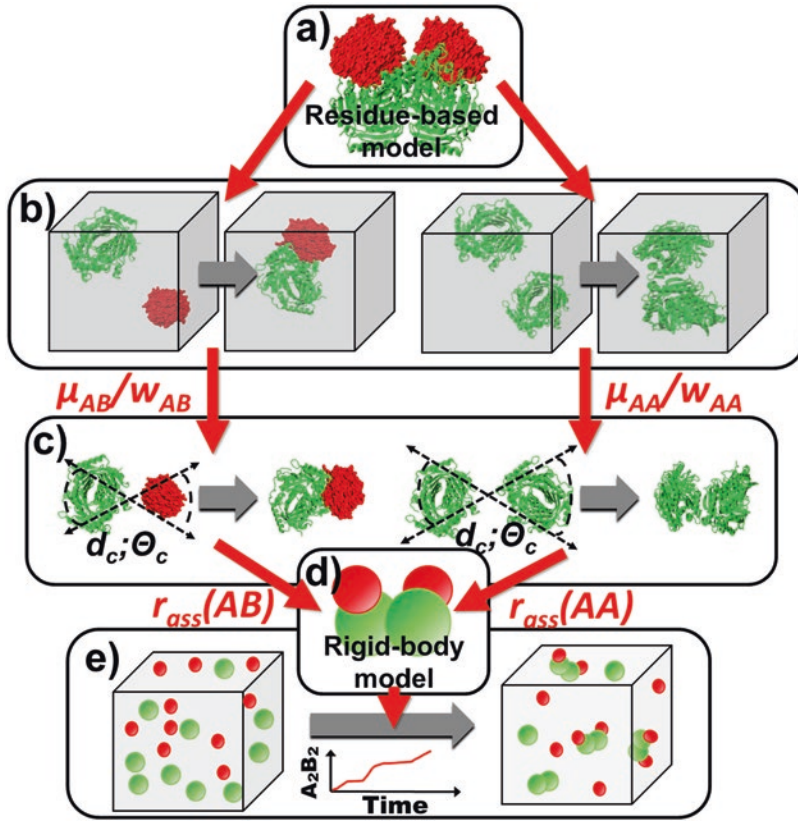


Fig. 1 There are two levels of models in the schematic framework of our multiscale simulation method. The structural details of each protein subunits in a complex can be reflected by the residue-based model (**a**). We first adjust the parameters in the energy functions of residue-based simulation to reproduce the experimentally measured values of k_{on} for each pair of subunits (**b**). Given the same energy parameter, the rate of association r_{ass} for each pair of subunits is then estimated by the same residue-based model but with a different boundary condition (**c**). Finally, the derived values of r_{ass}, together with the diffusion constants and binding affinities of interacting subunits, are used to guide the simulation with a rigid-body-based representation (**d**). The rigid-body simulations which contains a large number of protein subunits in the system are able to provide the kinetic information about complex assembly, such as how many final complexes or kinetic intermediate are formed along the simulation time (**e**)

2.2.2 Simulation Algorithm

The kinetic Monte Carlo (KMC) simulation starts from an initial conformation in which a pair of proteins was randomly placed. The translational and rotational diffusions are then carried out within each simulation step for both proteins in the system. Specific boundary conditions are applied after the diffusions. The new binding conformation is evaluated by either GO-like potential [40] or coarse-grained physical-based energy functions [39]. The probability of acceptance for this new conformation after diffusion depends on the calculated binding energy. At the end of each simulation step, the distances between all intermolecular interfacial pairs were calculated to determine how many native contacts were recovered. When at least three native contacts were recovered, we assumed that the

two proteins formed an encounter complex and the current simulation trajectory was terminated. Otherwise, the simulation ended when it reached the predefined maximal duration (*see* **Note 1**).

2.2.3 Boundary Condition

Two different boundary conditions are used in our study. The first is the periodic boundary condition. In the periodic boundary condition, two proteins are initially placed in a three-dimensional periodic box at random positions (Fig. 1b). During simulation, if one protein exists from one side of the box, it will immediately enter the opposite side. In the second boundary condition, a different initial conformation is constructed. Specifically, the binding interfaces of two proteins are placed randomly, but within the given distance cutoff d_c, and the range of their packing angles is within the cutoff Θ_c (Fig. 1c). Consequently, in the following simulation, two molecules either formed an encounter complex or separated far away from each other by the end of each simulation trajectory.

2.2.4 Energy Functions Between Proteins

The binding between proteins was originally evaluated by a GO-like potential [40] which gives scores for all pairs of native contact. Any pair of Cα atoms between residues i in one protein and j in the other is defined as a native contact if its corresponding distance in the native structure is smaller than 7.5 Å. An adjustable parameter μ defines the energy depth of the GO potential. It can be used to control the rate of binding.

In our more recent study, the total energy of binding between two proteins is described by a simple physics-based potential function consisting of three terms [39]. The first component is the electrostatic interaction which was previously used in the Kim-Hummer model [41, 42]. The second component is the hydrophobic interaction, which is calculated by summing the hydrophobic scores of all contact residue pairs. The hydrophobic scores of a contact residue pair are taken from a previous study by Kyte and Doolittle [43]. The excluded volume effect during protein binding is taken into account as the third component. Finally, a weight parameter w which determines the relative contributions between the hydrophobic and electrostatic interactions can be used to control the rate of binding.

2.3 Rigid-Body Model for Simulation Protein Complex Assembly

In the rigid-body-based model, proteins are simplified as spherical rigid bodies with various radii (*see* **Note 2**). Multiple binding sites are assigned on the surface of each rigid body [44]. The spatial assignment of each binding site depends on the quaternary arrangement of the protein complex under study (Fig. 1d).

2.3.1 Model Representation

2.3.2 Simulation Algorithm

A diffusion-reaction algorithm is developed to simulate the assembly kinetics [44]. As the initial configuration, a large number of proteins with all species of subunits in the complex are randomly distributed in a 3D simulation box (Fig. 1e). The number of subunits and the size of simulation box are determined by the concentrations and the stoichiometry of the complex. Followed by the initial conformation, the system is evolved by an iteration of diffusion-reaction process. Molecules are first chosen to undergo random diffusions with the periodic boundary condition. The amplitude of diffusions for all molecules is determined by their corresponding diffusion coefficients. If a complex is formed during the process of assembly, all its subunits will move together, with a relatively smaller diffusion coefficient. After diffusions, any pair of subunits that fulfill the binding criteria has the probability to associate together, by the corresponding on rate. In contrast, any associated pair of subunits has the probability to break into separate monomers, by the corresponding on rate and binding affinity.

3 Methods

3.1 Calibrate the Energy Function in Residue-Based Simulations

For each different pair of interacting protein subunits in a complex, the following steps of operation will be carried out sequentially (Fig. 1b):

1. 10^4 residue-based simulation trajectories are generated with periodic boundary condition, using either GO-like or physics-based potential functions. The default value of μ in the GO-like potential or w in the physics-based energy function is used as initial condition.
2. The on rate k_{on} is derived by counting how many complexes are associated from these simulation trajectories.
3. The calculated k_{on} from the simulation is compared with the experimentally measured value. The value of μ or w is adjusted accordingly, if the calculated k_{on} is either weaker or stronger than the corresponding experimental value.
4. The first step is repeated using the adjusted value of μ or w, so that the new k_{on} is calculated.
5. The procedure from the first to the fourth step is iterated until the calculated k_{on} fits reasonably well with the experimental value. Consequently, the calibrated parameter μ or w is used to derive the association rate r_{ass} in Subheading 3.2.

If no experimental k_{on} is available for a specific pair of protein subunits, Subheading 3.1 will be skipped. The association rate r_{ass} for this pair of subunits is directly calculated in Subheading 3.2 by using the physics-based energy function with the default value of weight constant w (*see* **Note 3**).

3.2 Derive the Association Rate r_{ass} for All Pairs of Subunits in the Complex

The r_{ass} in the rigid-body-based model is the rate of association between two interacting proteins under the given binding criteria. It is a specific parameter resulting from the coarse-grained nature of rigid-body-based model and depends on the choice of different binding criteria, such as the distance and orientation between two proteins. For each different pair of interacting protein subunits in a complex, we can derive r_{ass} using the specific calibrated energy functions described above. In detail, the following steps of operation will be carried out sequentially for each pair of subunits in a complex (Fig. 1c):

1. 10^4 residue simulation trajectories are generated with the second type of boundary condition, in which the initial conformation is constructed by placing two proteins within the given distance cutoff d_c and the range of their packing angles within the cutoff Θ_c (*see* **Note 4**). Either GO-like or physics-based potential functions can be used with the calibrated parameter μ or w from Subheading 3.1. The maximal length of each trajectory equals Δt_{RB}, which is the simulation time step of the rigid-body-based model (*see* **Note 5**).
2. The dimerization probability between protein subunits ρ is calculated by counting how many complexes are associated from these simulation trajectories.
3. The association rate r_{ass} can be calculated as $r_{ass} = \rho/\Delta t_{RB}$. The values of r_{ass} for all pairs of interacting protein subunits in a given complex are derived for the simulation of complex assembly which will be introduced in Subheading 3.3.

3.3 Simulate the Complex Assembly by Rigid-Body-Based Model

Based on calculated r_{ass} for all pairs of subunits in the complex, the following steps of rigid-body simulation will be carried out to study the kinetics of complex assembly (Fig. 1e):

1. The diffusion constants for all protein subunits in the complex are calculated by a precise boundary element method.
2. The off rate k_{off} which characterizes the kinetics of dimer dissociation is calculated for all pairs of protein subunits in the complex using the equation $k_{off} = k_{on} \times K_d$, in which k_{on} is the on rate and K_d is the dissociation constant for a corresponding pair of protein subunits.
3. The radii of rigid bodies for all subunits are determined by the given three-dimensional structure of the complex.
4. The number of binding sites for each subunit and their relative positions are assigned on the surface of its corresponding rigid body based on the quaternary organization of the protein complex.
5. After determining the size of simulation box, the initial conformation of the rigid-body simulation is constructed by randomly placing rigid bodies for all types of subunits in the box.

The number of rigid bodies for each type of subunit is determined by the concentrations and the stoichiometry of the complex.

6. The simulations are carried out by giving the desired number of trajectories and the length of simulation time for each trajectory.
7. Collect information from the simulation trajectories, and analyze the simulation result, such as the number of protein complexes and different intermediate states formed along the simulation time.

Using the above framework of multiscale simulation procedure, we can study how mutations affect the kinetics of protein complex assembly (*see* **Note 6**) and evaluate how protein complex assembly can be regulated by solvation effect (*see* **Note 7**).

4 Notes

1. The conformational changes are not considered during our study of complex assembly. Previous studies have illustrated that conformational changes are important in protein complex assembly. Although the effect of conformational flexibility cannot be reflected by the rigid-body model, it can be estimated by our residue-based simulation. For instance, the elastic network model (ENM) [47] has been integrated into the current model of our residue-based simulation so that protein conformations can be changed during association.
2. In the current stage of the study, each protein subunit in the lower-resolution simulation is simplified by a spherical rigid body. Therefore, our method will not be able to be applied to protein complexes containing subunits of non-globular shapes. In the future, our method can be improved by using non-spherical rigid bodies. Furthermore, by applying a domain-based representation in which each globular domain is represented by a rigid body, our method can be extended to protein complexes that contain multi-domain protein subunits.
3. As we mentioned in Subheading 3.1, if no experimental k_{on} is available for a specific pair of protein subunits, the association rate for this pair of subunits is directly calculated by using the physics-based energy function with the default value of weight constant w. On the other hand, if no experimental dissociation constants are available for a specific pair of protein subunits, computational methods can be used to predict either the absolute values of wild-type binding affinity, such as *PPEPred* [45]. Other computational methods such as *BindProfX* [46] can predict the relative changes of binding affinity due to mutations at the binding interfaces.

4. In the second boundary condition of the residue-based simulation, the binding interfaces of two proteins are initially placed within the given values of the distance cutoff d_c and the range of packing angles Θ_c. The same values of distance cutoff and range of packing angles should be used in the rigid-body simulations as criteria for binding in order to pass the calculated value of r_{ass} from the higher-resolution model to the lower-resolution model.
5. To derive the association rate r_{ass} for all pairs of subunits in the complex (Subheading 3.2), the maximal length of each simulation trajectory should be equal the time step of the rigid-body-based simulation. By the definition of r_{ass} and Δt_{RB}, if two molecules that meet the binding criteria, association will occur at the probability of $r_{ass} \times \Delta t_{RB}$ within each time step of rigid-body simulation. To estimate the value of r_{ass}, residue-based simulations should be carried out with the same time scale. Consequently, each trajectory of residue-based simulation consists of n steps so that the total length of simulation time for each trajectory satisfies $\Delta t_{RB} = n \times \Delta t$, in which Δt is the time step of residue-based simulation.
6. Our previous study demonstrated that our residue-based simulation method can capture the effects of single- and double-point mutations on the association rates [39]. Therefore, the framework of our multiscale model can be used to study how mutations affect the kinetics of protein complex assembly by applying the same procedure to both wild-type protein complex and to its mutant systems.
7. The concentration of ions around two interacting proteins is an important factor controlling the rate of their association. The ionic strength is an adjustable parameter in our residue-based simulation. Our tests showed that the residue-based model can reproduce the effect of the ionic strength on associations [39]. Therefore, by changing the value of ionic strength, our multiscale method will also be able to evaluate how protein complex assembly can be regulated by solvation effect.

Acknowledgments

This work was supported in part by the National Institutes of Health (Grant No. R01GM120238) and a start-up grant from the Albert Einstein College of Medicine.

References

1. Ali MH, Imperiali B (2005) Protein oligomerization: how and why. Bioorg Med Chem 13(17):5013–5020. https://doi.org/10.1016/j.bmc.2005.05.037
2. Levy ED, Teichmann S (2013) Structural, evolutionary, and assembly principles of protein oligomerization. Prog Mol Biol Transl Sci 117:25–51. https://doi.org/10.1016/b978-0-12-386931-9.00002-7
3. Marsh JA, Hernandez H, Hall Z, Ahnert SE, Perica T, Robinson CV, Teichmann SA (2013) Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell 153(2):461–470. https://doi.org/10.1016/j.cell.2013.02.044
4. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA (2008) Assembly reflects evolution of protein complexes. Nature 453(7199):1262–1265. https://doi.org/10.1038/nature06942
5. Ellis RJ (2007) Protein misassembly: macromolecular crowding and molecular chaperones. Adv Exp Med Biol 594:1–13. https://doi.org/10.1007/978-0-387-39975-1_1
6. Gabdoulline RR, Wade RC (2002) Biomolecular diffusional association. Curr Opin Struct Biol 12(2):204–213
7. Zhou HX (2010) Rate theories for biologists. Q Rev Biophys 43(2):219–293. https://doi.org/10.1017/S0033583510000120
8. Picco A, Irastorza-Azcarate I, Specht T, Boke D, Pazos I, Rivier-Cordey AS, Devos DP, Kaksonen M, Gallego O (2017) The in vivo architecture of the exocyst provides structural basis for exocytosis. Cell 168(3):400–412. e418. https://doi.org/10.1016/j.cell.2017.01.004
9. Gilmore BL, Winton CE, Demmert AC, Tanner JR, Bowman S, Karageorge V, Patel K, Sheng Z, Kelly DF (2015) A molecular toolkit to visualize native protein assemblies in the context of human disease. Sci Rep 5:14440. https://doi.org/10.1038/srep14440
10. Heck AJ (2008) Native mass spectrometry: a bridge between interactomics and structural biology. Nat Methods 5(11):927–933. https://doi.org/10.1038/nmeth.1265
11. Wieczorek G, Zielenkiewicz P (2008) Influence of macromolecular crowding on protein-protein association rates--a Brownian dynamics study. Biophys J 95(11):5030–5036. https://doi.org/10.1529/biophysj.108.136291
12. Ermakova E (2005) Lysozyme dimerization: Brownian dynamics simulation. J Mol Model 12(1):34–41. https://doi.org/10.1007/s00894-005-0001-2
13. Forlemu NY, Njabon EN, Carlson KL, Schmidt ES, Waingeh VF, Thomasson KA (2011) Ionic strength dependence of F-actin and glycolytic enzyme associations: a Brownian dynamics simulations approach. Proteins 79(10):2813–2827. https://doi.org/10.1002/prot.23107
14. Long H, Chang CH, King PW, Ghirardi ML, Kim K (2008) Brownian dynamics and molecular dynamics study of the association between hydrogenase and ferredoxin from Chlamydomonas Reinhardtii. Biophys J 95(8):3753–3766. https://doi.org/10.1529/biophysj.107.127548
15. Frembgen-Kesner T, Elcock AH (2010) Absolute protein-protein association rate constants from flexible, coarse-grained Brownian dynamics simulations: the role of intermolecular hydrodynamic interactions in barnase-barstar association. Biophys J 99(9):L75–L77. https://doi.org/10.1016/j.bpj.2010.09.006
16. Zimmer MJ, Geyer T (2012) Do we have to explicitly model the ions in brownian dynamics simulations of proteins? J Chem Phys 136(12):125102. https://doi.org/10.1063/1.3698593
17. Dlugosz M, Huber GA, McCammon JA, Trylska J (2011) Brownian dynamics study of the association between the 70S ribosome and elongation factor G. Biopolymers 95(9):616–627. https://doi.org/10.1002/bip.21619
18. Huber GA, Kim S (1996) Weighted-ensemble Brownian dynamics simulations for protein association reactions. Biophys J 70(1):97–110. https://doi.org/10.1016/S0006-3495(96)79552-8
19. Rojnuckarin A, Livesay DR, Subramaniam S (2000) Bimolecular reaction simulation using weighted ensemble Brownian dynamics and the University of Houston Brownian Dynamics program. Biophys J 79(2):686–693. https://doi.org/10.1016/S0006-3495(00)76327-2
20. Zou G, Skeel RD (2003) Robust biased Brownian dynamics for rate constant calculation. Biophys J 85(4):2147–2157. https://doi.org/10.1016/S0006-3495(03)74641-4
21. Zhou HX (1993) Brownian dynamics study of the influences of electrostatic interaction and diffusion on protein-protein association kinetics. Biophys J 64(6):1711–1726. https://doi.org/10.1016/S0006-3495(93)81543-1
22. Northrup SH, Erickson HP (1992) Kinetics of protein-protein association explained by

Brownian dynamics computer simulation. Proc Natl Acad Sci U S A 89(8):3338–3342

23. Merlitz H, Rippe K, Klenin KV, Langowski J (1998) Looping dynamics of linear DNA molecules and the effect of DNA curvature: a study by Brownian dynamics simulation. Biophys J 74(2 Pt 1):773–779. https://doi.org/10.1016/S0006-3495(98)74002-0
24. Mereghetti P, Gabdoulline RR, Wade RC (2010) Brownian dynamics simulation of protein solutions: structural and dynamical properties. Biophys J 99(11):3782–3791. https://doi.org/10.1016/j.bpj.2010.10.035
25. Lin J, Beratan DN (2005) Simulation of electron transfer between cytochrome C2 and the bacterial photosynthetic reaction center: Brownian dynamics analysis of the native proteins and double mutants. J Phys Chem B 109(15):7529–7534. https://doi.org/10.1021/jp045417w
26. De Rienzo F, Gabdoulline RR, Menziani MC, De Benedetti PG, Wade RC (2001) Electrostatic analysis and Brownian dynamics simulation of the association of plastocyanin and cytochrome f. Biophys J 81(6):3090–3104. https://doi.org/10.1016/S0006-3495(01)75947-4
27. Haddadian EJ, Gross EL (2006) A Brownian dynamics study of the interactions of the luminal domains of the cytochrome b6f complex with plastocyanin and cytochrome c6: the effects of the Rieske FeS protein on the interactions. Biophys J 91(7):2589–2600. https://doi.org/10.1529/biophysj.106.085936
28. Hattne J, Fange D, Elf J (2005) Stochastic reaction-diffusion simulation with MesoRD. Bioinformatics 21(12):2923–2924. https://doi.org/10.1093/bioinformatics/bti431
29. Ander M, Beltrao P, Di Ventura B, Ferkinghoff-Borg J, Foglierini M, Kaplan A, Lemerle C, Tomas-Oliveira I, Serrano L (2004) SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks. Syst Biol (Stevenage) 1(1):129–138
30. Rodriguez JV, Kaandorp JA, Dobrzynski M, Blom JG (2006) Spatial stochastic modelling of the phosphoenolpyruvate-dependent phosphotransferase (PTS) pathway in Escherichia Coli. Bioinformatics 22(15):1895–1901. https://doi.org/10.1093/bioinformatics/btl271
31. Stiles J, Bartol TM (2001) Monte Carlo methods for simulating realistic synaptic microphysiology using MCell. Computational Neuroscience:87–127
32. Andrews SS, Bray D (2004) Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. Phys Biol 1(3–4):137–151. https://doi.org/10.1088/1478-3967/1/3/001
33. Ridgway D, Broderick G, Lopez-Campistrous A, Ru'aini M, Winter P, Hamilton M, Boulanger P, Kovalenko A, Ellison MJ (2008) Coarse-grained molecular simulation of diffusion and reaction kinetics in a crowded virtual cytoplasm. Biophys J 94(10):3748–3759. https://doi.org/10.1529/biophysj.107.116053
34. Frazier Z, Alber F (2012) A computational approach to increase time scales in Brownian dynamics-based reaction-diffusion modeling. J Comput Biol 19(6):606–618. https://doi.org/10.1089/cmb.2012.0027
35. Lee B, LeDuc PR, Schwartz R (2008) Stochastic off-lattice modeling of molecular self-assembly in crowded environments by Green's function reaction dynamics. Phys Rev E 78(3). https://doi.org/10.1103/PhysRevE.78.031911
36. Xie ZR, Chen J, Wu Y (2016) Multiscale model for the assembly kinetics of protein complexes. J Phys Chem B 120(4):621–632. https://doi.org/10.1021/acs.jpcb.5b08962
37. Aragon S (2004) A precise boundary element method for macromolecular transport properties. J Comput Chem 25(9):1191–1205. https://doi.org/10.1002/jcc.20045
38. Aragon S, Hahn DK (2006) Precise boundary element computation of protein transport properties: diffusion tensors, specific volume, and hydration. Biophys J 91(5):1591–1603. https://doi.org/10.1529/biophysj.105.078188
39. Xie ZR, Chen J, Wu Y (2017) Predicting protein-protein association rates using coarse-grained simulation and machine learning. Sci Rep 7:46622. https://doi.org/10.1038/srep46622
40. Hills RD Jr, Brooks CL 3rd (2009) Insights from coarse-grained go models for protein folding and dynamics. Int J Mol Sci 10(3):889–905. https://doi.org/10.3390/ijms10030889
41. Kim YC, Hummer G (2008) Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. J Mol Biol 375(5):1416–1433. https://doi.org/10.1016/j.jmb.2007.11.063
42. Ravikumar KM, Huang W, Yang S (2012) Coarse-grained simulations of protein-protein association: an energy landscape perspective.

Biophys J 103(4):837–845. https://doi.org/10.1016/j.bpj.2012.07.013

43. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132
44. Xie Z-R, Chen J, Wu Y (2014) A coarse-grained model for the simulations of biomolecular interactions in cellular environments. J Chem Phys 140:054112
45. Su Y, Zhou A, Xia X, Li W, Sun Z (2009) Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. Protein Sci 18(12):2550–2558. https://doi.org/10.1002/pro.257
46. Xiong P, Zhang C, Zheng W, Zhang Y (2017) BindProfX: assessing mutation-induced binding affinity change by protein Interface profiles with pseudo-counts. J Mol Biol 429(3):426–434. https://doi.org/10.1016/j.jmb.2016.11.022
47. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80(1):505–515

Chapter 27

Flexible Protein-Protein Docking with SwarmDock

Iain H. Moal, Raphael A. G. Chaleil, and Paul A. Bates

Abstract

The atomic structures of protein complexes can provide useful information for drug design, protein engineering, systems biology, and understanding pathology. Obtaining this information experimentally can be challenging. However, if the structures of the subunits are known, then it is often possible to model the complex computationally. This chapter provide practical guidelines for docking proteins using the SwarmDock flexible protein-protein docking method, providing an overview of the factors that need to be considered when deciding whether docking is likely to be successful, the preparation of structural input, generation of docked poses, analysis and ranking of docked poses, and the validation of models using external data.

Key words Molecular modelling, Docking, Protein-protein interaction, Computational chemistry

1 Introduction

The functions of protein complexes are products of the specific geometrical arrangement of the subunits from which it is formed. The position, orientation, and conformation of each subunit is established by the formation of specific intermolecular contacts which anchor the subunit in place and lower the free energy of the bound state. Information about the structure can aid in tasks such as the identification of energetic hotspots and potential sites for the design of molecules which can mimic or prevent a natural interaction. It can also help in elucidating the mechanisms through which pathological mutations alter functions, aid library design for engineering high binding affinity, and allow the identification of overlapping binding sites. While in many cases the structure of an interaction can be resolved using NMR, X-ray crystallography, or high-resolution electron microscopy, these experiments are not guaranteed success and can be expensive and time-consuming. However, if the structures of the unbound constituents of the interaction have been resolved, or a high-quality model can be generated by homology modelling, then it may be possible to generate a structure of the interaction computation-

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_27,

ally using protein-protein docking. Multiple servers and standalone programs for docking are currently available [1–14], and while the focus of this chapter is on using SwarmDock [15, 16], many of the principles also apply to other approaches. Most of the tasks required to perform the docking are handled automatically in the SwarmDock server [17, 18], which can be accessed at https://bmm.crick.ac.uk/~svc-bmm-swarmdock/. Below is a brief overview of the algorithm, with the following sections outlining the steps involved in docking, finishing with a case study.

2 Overview of SwarmDock

The purpose of protein-protein docking is to produce structural models of interacting proteins, ranked such that the top-ranked models are most likely to be close to the native structure. This typically proceeds as a sequence of steps: an initial conformational search, filtering of models using an efficient scoring function and/or clustering, refinement of the resultant structures, and finally ranking of the refined structures. In practice, not all docking pipelines employ all these steps. While it is possible to mix and match different search, filtering, refinement, and scoring protocols, the success of the later steps is usually greatest when applied to structures generated in the same way in which the method was developed [19–21].

SwarmDock is a flexible docking method that optimizes the conformation and the relative position and orientations of the subunits (Fig. 1a). The set of accessible states is specified by the orientation and position of the smaller of the two proteins with respect to the larger and the conformation of both binding partners. The conformations are modelled by a linear combination of normal coordinates, which in many cases encapsulates the conformational changes observed when proteins bind to one another [22]. In this framework, each potential docked pose is typically characterized by the three Cartesian coordinates for relative position of the binding partners, four quaternion terms for their relative orientation, and five coefficients for each binding partner that specify their conformation. Any given set of values for these 17 parameters corresponds to a structure. In SwarmDock, these parameters are optimized to find the set of values that minimize the interaction energy of the two binding partners (Fig. 1b, i–iv). The interaction energy is calculated using the DComplex potential function [23]. The optimization is performed by a population-based memetic algorithm, in which a swarm of 350 parameter combinations are initially sampled. This is done by combining a modified particle swarm optimization global search [24] with a local search [25]. On each iteration of the algorithm, the energies and past histories of the swarm members are used to determine

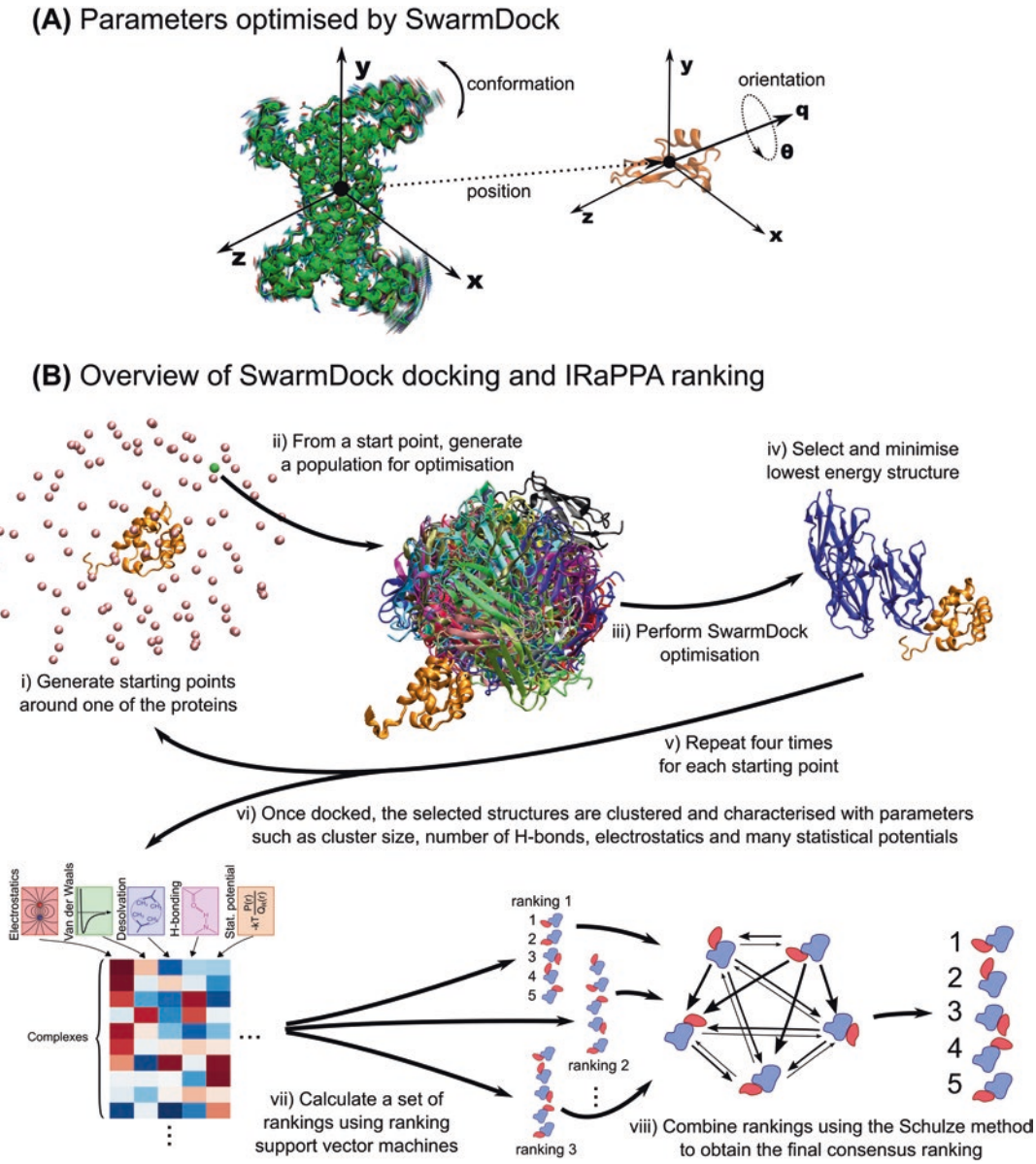


Fig. 1 Overview of the SwarmDock algorithm. (**a**) The parameters optimized during the docking process. (**b**) A summary of the steps taken for docking with SwarmDock (i–v) and ranking with IRaPPA (vi–viii)

the subsequent positions to be sampled, using a population-based memetic algorithm that combines a global search for identifying the broad low-energy regions of parameter space within which the correctly bound structure is more likely to be found, with a local search for refining the structures into the minima of the energy landscape. The optimization is performed around 240 times, with the initial sampling of each run focusing on different overlapping regions of parameter space corresponding to evenly spaced areas

surrounding the surface of the larger of the two binding partners. Once the search is complete, the resultant structures are subsequently clustered and ranked, either with a pairwise potential function [26], using the local energy structure of the binding region [27], or using the IRaPPA method (Fig. 1b, vi–viii) [28].

3 Factors Influencing the Success Rate of Protein-Protein Docking

The success rate for docking depends on various factors that can be taken into consideration when choosing whether or not docking is a viable option and when interpreting the results of a docking calculation (Fig. 2).

1. *Structural resolution and homology models.* During docking, prior modelling errors can accumulate with docking errors, so using the highest-quality available starting structures is advised, ideally high-resolution (<2 Å) crystal structures for both binding partners. At coarser resolution, some surface loops and the side chains of long amino acids are inferred during structural determination. When the resolution exceeds 3 Å, these defects can become severe, and docking is significantly less likely to succeed. If a structure is not available, then a homology model can be used. In this case, models built from sequences with over 70% sequence identity are of a very high quality and are suitable for docking. Below this docking is less likely to succeed, particularly below 50% sequence identity. Models built with below 30% sequence identity are not suitable for docking.
2. *Protein flexibility.* Proteins which bind as rigid molecules (<1 Å Cα RMSD conformational change upon binding) have much higher success rates for docking than more flexible proteins; beyond 2 Å RMSD success rates drop precipitously. Disordered proteins cannot currently be modelled using SwarmDock. While it is not possible to know the exact extent of conformational change upon binding without knowing the structure of the bound complex, protein flexibility can be estimated from the

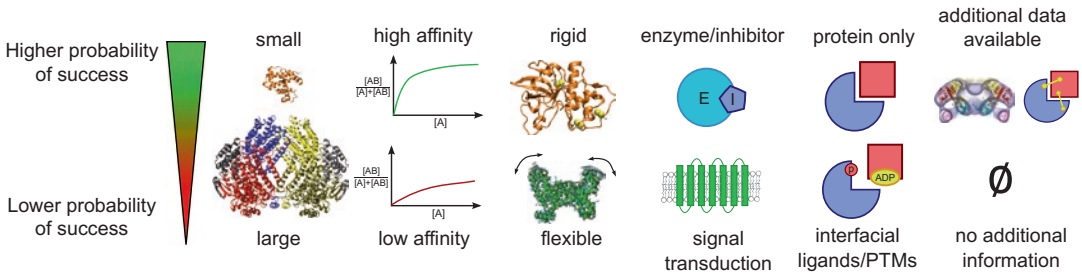


Fig. 2 Factors influencing the feasibility of docking

geometry and dynamics using simple calculations [29–32]. Further, some classes of proteins such as cytokines and venoms and other extracellular proteins are typically rigid, which can be inferred from multiple disulfide bridges and a tightly packed core.

3. *Binding affinity*. As high-affinity complexes have more chemically complementary interactions across the binding interface, higher affinity is associated with better docking success rates. Indeed, a recent study showed that the chance of finding a near-native pose in the top 10-ranked solution using SwarmDock increased from 45% for low-affinity interactions (<9 kcal/mol) to 65% for high-affinity interactions (>13 kcal/mol) [28].
4. *Biological role*. Some categories of interaction have higher success rates than others. For instance, enzyme-inhibitor complexes are typically high-affinity rigid-body interactions for which top 10 success rates can be over 65% [28]. On the other hand, interactions involved in signal transduction, such as receptor-ligand interactions, tend to be weak and transient and are consequently harder to dock.
5. *Protein size*. Larger proteins have a greater surface area and thus present more opportunities for the production of false-positive docked poses.
6. *Nonprotein components and nonstandard amino acids*. Interactions that are mediated by direct contact with nonprotein components such as ions, small molecules, oligonucleotides, porphyrin groups, or posttranslational modifications may not dock correctly if these components make an important contribution to the binding energy, as parameters for these moieties are not available in the DComplex energy function. Nevertheless, sometimes these can be accommodated indirectly (*see* next section), and even when they cannot, docking is often still successful if the contribution from the standard amino acids is substantial.
7. *Additional information*. Greater success rates can be achieved if experimental or evolutionary information is used to constrain the regions of search space explored or to filter away poses that are incompatible with that data. Details of the types of data and how they can be incorporated are covered later in this chapter.

4 Structure Selection and Preparation

The SwarmDock server accepts atomic coordinates as PDB files, which in some cases can be taken directly from the protein data bank and used as input. The server will only dock contiguous chains composed of the standard 20 amino acids, each separated with a "TER" statement, and for which each atomic coordinate is

specified only once. However, the server does contain automated procedures for dealing with nonstandard amino acids, alternative atom locations, and missing side chains and loops, outlined below. The user is encouraged to check the PDB files bearing in mind the following factors, to ensure that the files are correctly formatted and that docking proceeds as intended:

1. *Structure selection*. Generally, crystallographic structures are recommended over NMR or homology models. If multiple structures are available, then the higher-resolution structure is advised, although another structure of similar resolution may be preferable if the wwPDB validation report indicates that it is of higher quality or that a region suspect of being relevant for binding gives a better fit to the density map. If a crystal structure is of particularly poor resolution, then it is advisable to use a high-quality homology model or NMR structure if available. NMR structures can also be assessed by their Worldwide Protein Data Bank (wwPDB; https://www.wwpdb.org) validation report. Usually the first structure in an NMR ensemble satisfies the greatest number of restraints. However, it is worth checking the entire ensemble for loops or terminal portions of the chain that appear disordered. If such disorder exists, then that region may be truncated to avoid unwanted tip effects arising during the normal mode calculations. If different conformational states are available, it may be worth performing the docking multiple times with multiple structures.
2. *Stoichiometry and symmetry*. As a rule of thumb, the stoichiometry and symmetry of a subunit in the unbound state are conserved in the bound state, although there are many exceptions. Sometimes a protein may appear multiple times within the same unit cell despite acting as a monomer in solution, and at other times the protein may interact with a symmetrical copy of itself in an adjacent cell. Consequently, it is advisable to use the biological assembly to perform the docking.
3. *Alternative atom locations*. The electron density maps for many PDB files contain regions that are ambiguous over which rotamer a side chain adopts, and these are flagged with alternative location indicators. By default, the SwarmDock server will select the first conformation, typically indicated with "A." However, you may wish to delete this conformation in the PDB file if an alternative conformation has a higher occupancy factor.
4. *Missing side chains*. If a side chain is missing or has missing atoms, the SwarmDock will attempt to model it using SCWRL [33].
5. *Disordered loops*. Frequently loops are not resolved crystallographically and are omitted from a PDB file. If a region of non-consecutive amino acid numbering is found within the

file, then the SwarmDock server will attempt to model it using the Loopy program [34], replacing the missing residues with poly-alanine. If you do not wish for this to happen, the protein can be split into two chains separated by a "TER" entry.

6. *Nonstandard components, posttranslational modifications, and synthetic peptides.* By default, the server will strip all HETATM records, with some exceptions for nonstandard amino acids. For these, the list of which can be found on the SwarmDock website, the residue will be reverted to the nearest available amino acid, typically its precursor amino acid. In some cases, this will have no effect, such as for D-peptides, as the intramolecular energy is not a component of the SwarmDock scoring function. For others, such as the reversion of selenomethionine to methionine, the change will be negligible. However other changes may be functionally significant, and it may be possible to modify the starting structure to mimic the modified protein. For instance, when modelling a phosphorylation-dependent interaction, instead of allowing phosphoserine to be reverted to serine by the server, the residue can be changed to aspartic acid, a phosphomimetic of phosphoserine for which pS to D mutations typically result in constitutively active proteins.

5 Docking

Once structures are uploaded and repaired by the server if needed, the docking process proceeds automatically. While this does not require input from the user, a technical summary of these steps is outlined below. Those only interested in applying the method can skip to the next section.

5.1 Minimization of Starting Structures

In order to eliminate any severe clashes, the structure is minimized using CHARMM [35] in a vacuum with the CHARMM19 force field: 50 steps of steepest descent, 100 steps of conjugate gradient, and 200 steps of adopted basis Newton-Raphson.

5.2 Normal Mode Calculation

All-atom normal modes are calculated using the ElNeMo program [36], in which all non-hydrogen atoms are unit masses and a force constant of 1 is used for all atom pairs within 10 Å distance of one another. The Hessian is constructed using *pdbmat* and diagonalized using *diagrtb*. The number of normal modes used in the receptor and ligand can be selected when submitting a job to the server and by default is set to the recommended level of five each.

5.3 Generation of Starting Positions

Approximately 120 points are generated surrounding the larger of the two binding partners, each around 15 Å away from the surface of the protein [16]. This is done by first approximating the protein

as an ellipse, spacing points around the ellipse, and then projecting them to 15 Å beyond the $E = 0$ potential energy isosurface of the protein. These points act as starting points for each of the SwarmDock runs.

5.4 Conformational Search

Each SwarmDock run begins by initializing a population of 350 candidate solutions (particles) with random orientations (quaternion) and random position drawn from a Gaussian distribution centered on the starting position for each Cartesian coordinate (σ = 10 Å) and with normal mode coefficients drawn from a Gaussian distribution centered on zero ($\sigma = 3$). The vector of these values form the initial position in search space ($\boldsymbol{Xi}$), and the binding energies of each particle are calculated. Each particle is also assigned a direction of travel in the form of a velocity through search space ($\boldsymbol{vi}$), which initially has zero magnitude. On each iteration of the algorithm (t), the following steps are taken:

1. The velocity of each position ($\boldsymbol{vi}$) is calculated as $\boldsymbol{vi}(t + 1) = w\,\boldsymbol{vi}(t) + c1r1,i(pi(t) - \boldsymbol{Xi}(t)) + c2r2,i(pn,i(t) - \boldsymbol{Xi}(t)) + r3,i(p\mathrm{r}$ and $(t) - \boldsymbol{Xi}(t))$. Here, $c1 = c2 = 2.05$, $w = 0.8$, r^* are random numbers drawn uniformly from the [0,1] range, $p$rand is the position of a randomly selected particle, pi is the lowest energy position found by particle i, and pn,i is the lowest energy position found by a particle in the neighborhood of particle i. A ring topology neighborhood of size $k = 114$ is used [37]. A limit is placed on the velocity of 5 Å in the Cartesian directions, 0.2 rad in the angular part of the quaternion, and 0.5 Å in the spatial part.
2. Then, the position ($\boldsymbol{Xi}$) of each member is updated by the formula $\boldsymbol{Xi}(t + 1) = \boldsymbol{Xi}(t) + \boldsymbol{vi}(t + 1)$.
3. The energy of each member of the population is calculated using DComplex [23].
4. The lowest energy particle is subjected to a local search using the Solis and Wets method [25], with initial step sizes of 0.5 Å, 5° and 0.25 Å for the Cartesian and angular and spatial parts of the quaternion. The step size doubles or halves after five consecutive successful or unsuccessful moves, and the search terminates after five consecutive halvings of the step size.

After 600 iterations, the structure of the lowest energy position found during the run is then returned. This process is repeated four times from each of the starting positions, generating a total of around 480 docked poses. Finally, each of the poses are minimized in CHARMM in the same way as the input structures.

6 Post-processing

The server offers three options for post-processing and ranking of the generated poses.

1. *Original SwarmDock server protocol:* Selecting this option first characterizes the poses using the side-chain potential function reported by Tobi [26]. Then a clustering approach is used in which decoys add either added to a cluster or form a new cluster in ascending order of energy. The lowest energy structure constitutes the first member of the first cluster. If the second lowest energy structure is within 3 Å of the first, then it becomes the second member of that cluster; otherwise it forms a new cluster of its own, which is added to the end of a list of clusters. Each subsequent decoy is compared, in list order, to the first member of each existing cluster, and if it is within 3 Å of the first member of that cluster, it is added to it. If it is not added to any cluster, it forms a new cluster. Once all poses have been clustered, the ranked list of clusters is returned to the user.
2. *Searching for funnel-like energy structures:* The second option available consists of finding poses at the tip of funnel-like energy structures [27]. In this approach the decoys are used as states in a Markov chain, in which transition probabilities between structurally similar decoys depend on the energy difference between the decoys. Ranks are determined by the equilibrium population of each decoy. The performance is similar to the original method.
3. *Integrative Ranking of Protein-Protein Assemblies (IRaPPA):* The IRaPPA method, illustrated in Fig. 1b (vi–viii), exploits methods originally developed for informational retrieval and electoral voting tasks [28]. It first characterizes the decoys with a large number of descriptors from the CCharPPI web server [38], such as molecular mechanics energy terms and statistical potentials. It then combines them with an ensemble of ranking support vector machines to produce an ensemble of ranks, which it then combines into a consensus ranking. This approach produces higher-quality rankings than the other two in terms of the rank of near-native structures and the quality of high-ranking near-natives, but at high computational cost. It is thus the recommended option if only one or two complexes need to be docked.

An alternative method in which clusters of docked poses are enriched by sampling conformational space, and the clusters are ranked based on the distributions of their properties [39], is not currently available on the server.

7 Incorporating Experimental Information

There are two ways in which additional information can be incorporated into the docking process. The first is to use the information during the docking to either favor structures consistent with that data or to restrict the conformational search. The second is to use it to validate docking predictions or filter out final docked poses that are inconsistent with the data. The SwarmDock server offers a limited capability of restricting the search to the region surrounding a residue or residues implicated in binding. It does this by excluding starting points that are on the other side of the molecule. Other methods, such as HADDOCK [40], are specifically designed to incorporate experimental data by having an energy term reflecting how well a structure satisfies experimental restraints. The favored way of using experiment data with SwarmDock, however, is as an external method of validating docked structures; if a high-ranking decoy is consistent with data that is completely separate from the docking itself, then it gives much higher confidence in the accuracy of that structure. The following gives some of the types and sources of experimental data and method to evaluate their congruence with the docked poses.

7.1 Low-Resolution Data

Low-resolution structural data can give us information about the overall geometry of the complex such as whether it is more oblate or prolate in character, but generally not enough to deduce the atomic coordinates or positioning of secondary structure elements.

1. Small-angle neutron scattering (SANS) and small-angle X-ray scattering (SAXS) curves can tell us the distribution of the angles at which neutrons or x-rays are scattered from a sample. To compare this with docked poses, tools such as CRYSON [41], for SANS, and CRYSOL [42] for SAXS can be used to calculate synthetic curves from docked structures and the correlation between the experimental and synthetic curves used to assess congruence with the data.
2. Collision cross-section data from mass spectrometry can tell us the effective cross-sectional area of a sample. The program MOBCAL can be used to calculate the rotationally averaged collision cross-section from atomic coordinates [43].
3. Cryo-electron microscopy is a powerful method of creating 3D reconstructions of a structure by combining many electron microscopy images. The Multifit module of the Integrative Modelling Platform (IMP) software package can be used to fit docked poses to the density data as well as provide an overlap score of how good the fit is [44, 45].

4. Nuclear magnetic resonance (NMR) residual dipolar coupling and pseudo-contact shifts can give information about the relative orientation of the binding partners.

7.2 Information About the Participation of Individual Residues

Other types of data give information about individual residues that are likely to participate in the interaction through direct contacts.

1. Site-directed mutation coupled with affinity or *Ki* measurements of the wild-type and mutant structures can be used to calculate ΔΔG, the change in binding affinity upon mutation. Sites that are solvent exposed in the unbound and, upon mutation, significantly reduce binding are predominantly located at binding sites [46]. Of particular interest are mutations to alanine, which sever the side chain. Similar information can be obtained from deep mutational scanning data [47].
2. Binding interfaces evolve more slowly than the non-binding surface, and residues critical for binding are commonly conserved. Conserved residues that are solvent exposed in the unbound structure, in particular amino acids that are enriched at binding sites such as the hydrophobics, may be involved in the interaction. Residue conservation maps can be generated for both protein receptor and ligand (see, e.g., the online conservation mapping protocol at (https://bmm.crick.ac.uk/~chalei01/mapconservation)). Surface residues identified to have high levels of conservation, particularly if a number of them form a prominent conserved patch, can be selected in the SwarmDock server to restrict the conformational search space. Other patterns to look out for are conserved apolar patches and protrusions that may form anchor residues upon complexation [48]. Another approach to assessing how well the structure fits the evolutionary data is to estimate the binding score of interologs, a topic that is covered in detail in the following chapter by Nadaradjane et al.
3. Residues with labile protons, such as alcohols, can exchange those protons if solvent exposed. If these moieties become buried upon binding, proton exchange is prevented. This can be detected by methods such as H/D exchange and NMR cross-saturation.
4. Often a complex can be too large for structural elucidation using classical NMR techniques. However, if the assignment of atoms to chemical shifts is available for one of the binding partners, it can be titrated and the chemical shift perturbations used to identify regions in which the chemical environment is altered by either direct contact with the binding partner or conformational changes.

7.3 Information About Specific Interaction Pairs: The Methods Below Can Give Information About the Spatial Separation of Pairs of Residues, One on Each Binding Partner

1. While the thermodynamic consequences of single-point mutations (ΔΔG) can identify residues which contribute to binding, double-mutant cycles can be employed to find nonadditive interactions from spatially proximate residue pairs. To do this, affinity is measured for the wild type, for single-point alanine mutations of two residues, one on each binding partner, and for the double alanine mutant. If the residue contributions are additive, then the ΔΔG of the double mutant will be equal to the sum of the ΔΔG values for the two single mutants [49]. Spatially distinct mutations tend to be additive, as each mutation removes the contribution of different atomic interactions. For interacting residues, however, the ΔΔG of the double mutant is less destabilizing than the sum of the two single-point mutations, as the contribution from atomic contacts between the two residues is removed if either of the residue is mutated; thus the effect of the second mutation is lower.
2. Spatially proximate residues can also be identified by chemical cross-linking. In these experiments the complex is treated by compounds that react with two protein sites, thereby creating a link between them which can be detected using mass spectrometry. The detection of an intermolecular cross-link gives information about how close the two residues are and also informs us that the residues are on the periphery of the binding site, where there is enough space to accommodate the cross-linking agent.

8 An Example Docking: The LCP2/FLNA Interaction

An interaction between lymphocyte cytosolic protein 2 (LCP2) and filamin A (FLNA) was first predicted computationally from the observations that the two proteins are co-expressed at the same time and place, have mutual binding partners, and have co-occurrence of post-translational modifications [50]. This prediction was confirmed more recently during a proteome-wide chemical cross-linking experiment as an interaction found with 99% confidence [51], using a disuccinimidyl sulfoxide (DSSO) cross-link between the sterile alpha motif (SAM) domain at the N-terminus of LCP2 (residue K81) and the third Ig-like domain of FNLA (residue K498). This information was published during the developing of IRaPPA, and we investigated whether this interaction would make a good test case for the new ranking scheme [28]. This section outlines the steps taken in structure selection, docking, and the incorporation of the cross-linking data.

A high-resolution (1.72 Å) structure of the third, fourth, and fifth Ig-like domains of FLNA had been solved [52] (pdb: 4M9P). As this structured fitted the electron density well, and had no disordered loops, posttranslational modifications, or bound ions or ligands, it was an ideal candidate for protein-protein docking. For

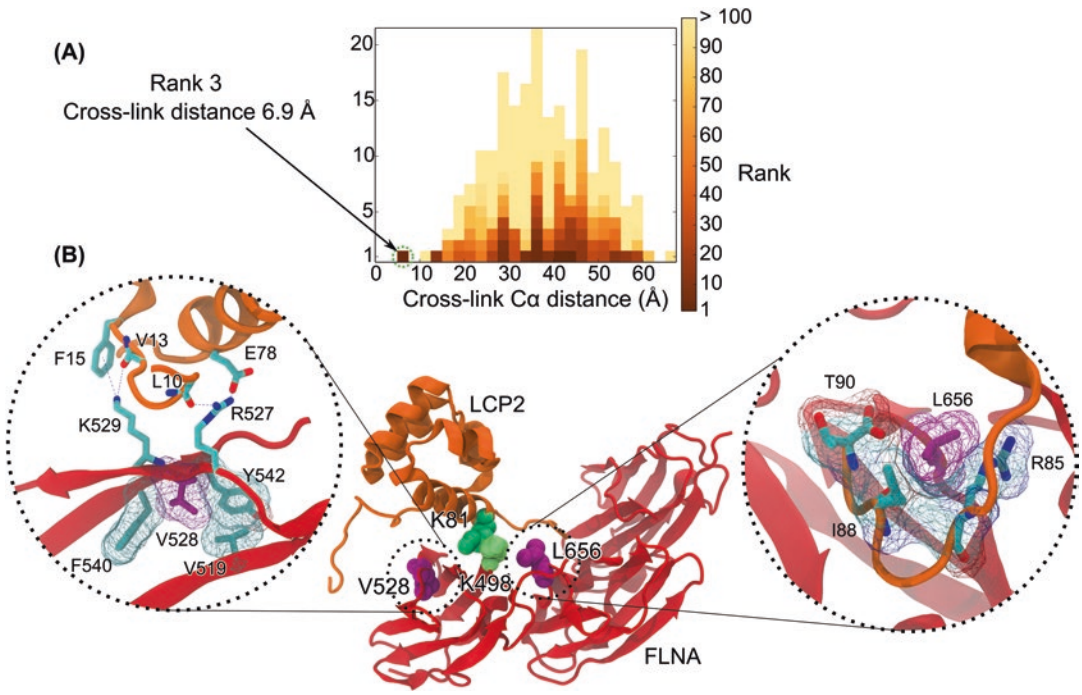


Fig. 3 Docking results for the LCP2/FLNA interaction. (**a**) The distribution of cross-linker distances for docked poses, colored by rank. (**b**) The structure of the low-energy pose consistent with cross-linking and mutation data. Image adapted from [28]

the SAM domain of LCP2, an ensemble of 20 NMR structures had been determined by the high-throughput RIKEN structural genomics/proteomics initiative as part of the Protein 3000 project (pdb 2EAP). This structure included many off-rotamer side chains, as well as unresolved regions at either side of the domain. No homologs had been solved; indeed the reason this protein was targeted by the structural genomics initiative was because this region of sequence space had not been structurally characterized, and so homology modelling was not an option. Nevertheless, the NMR structures had a well-defined core region and no nonstandard components, so the structure with the least restraint violations was selected for docking.

The structures were docked using the SwarmDock server with the IRaPPA ranking method. For each of the docked poses, the Cα distance between the DSSO interlinked lysine residues was calculated (Fig. 3a). Residues coupled in this way are always solvent exposed, such that the linker can be located without clashing with protein atoms, and typically have Cα distances in the 6–15 Å range [53]. One docked pose, which was low in energy (ranked third according to IRaPPA), had an LCP2_K81-FLNA_K498 Cα distance of 6.9 Å and has both lysine residues located at the periphery of the binding site such that a DSSO cross-link could be accommodated. Of further interest was the involvement of two residues in

FLNA for which mutations had been implicated in periventricular heterotopia, a condition in which gray matter is incorrectly located in the brain (V528 and L656). The two mutations are on different adjacent domains of FLNA and are 18 Å apart. Nevertheless, both are involved in the predicted interaction, either as direct interactions between L656 with T90 and I88 on LCP2 or as indirect support for a binding loop in the case of V528, in particular the residues at either side which form hydrogen bonds and a cation-π interaction with LCP2 (Fig. 3). Thus, despite some concerns regarding the suitability of LCP2 structure for docking, this example demonstrates how the SwarmDock server can be used to predict the structure of a previously uncharacterized and unstudied interaction. Although docking predictions can rarely be taken at face value, this example also shows how experimental information, in this case a combination of mass spectrometry data and the reconciliation of the common pathology of two disparate mutations, can be used to select and validate a prediction. We hope it has also shown how docking can help in answering questions of biological importance and in the generation of hypothesis regarding the role of an interaction, in this case by postulating a role for the LCP2/FLNA interaction in the development of the brain.

Acknowledgments

This work was supported by the European Molecular Biology Laboratory [IHM], the Biotechnology and Biological Sciences Research Council [Future Leader Fellowship BB/N011600/1 to IHM], and the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001003), the UK Medical Research Council (FC001003), and the Wellcome Trust (FC001003) [R.A.G.C., P.A.B.].

References

1. Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 34:310–314
2. Lyskov S, Gray JJ (2008) The RosettaDock server for local protein-protein docking. Nucleic Acids Res 36:W233–W238
3. Garzon JI, Lopéz-Blanco JR, Pons C et al (2009) FRODOCK: a new approach for fast rotational protein-protein docking. Bioinformatics 25(19):2544–2551
4. Macindoe G, Mavridis L, Venkatraman V et al (2010) HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Res 38:W445–W449
5. Mashiach E, Schneidman-Duhovny D, Peri A et al (2010) An integrated suite of fast docking algorithms. Proteins 78(15):3197–3204
6. Huang S-Y, Zou X (2010) MDockPP: a hierarchical approach for protein-protein docking and its application to CAPRI rounds 15-19. Proteins 78(15):3096–3103
7. Pierce BG, Hourai Y, Weng Z (2011) Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS One 6(9):e24657
8. Jiménez-García B, Pons C, Fernández-Recio J (2013) pyDockWEB: a web server for rigid-body protein-protein docking using electro-

statics and desolvation scoring. Bioinformatics 29(13):1698–1699
9. van Zundert GCP, Bonvin AMJJ (2014) Modeling protein-protein complexes using the HADDOCK webserver. Methods Mol Biol 1137:163–179
10. Viswanath S, Ravikant DVS, Elber R (2014) DOCK/PIERR: web server for structure prediction of protein-protein complexes. Methods Mol Biol 1137:199–207
11. Esquivel-Rodriguez J, Filos-Gonzalez V, Li B, Kihara D (2014) Pairwise and multimeric protein-protein docking using the LZerD program suite. Methods Mol Biol 1137:209–234
12. de Vries SJ, Schindler CEM, Chauvot de Beauchêne I, Zacharias M (2015) A web interface for easy flexible protein-protein docking with ATTRACT. Biophys J 108(3):462–465
13. Kozakov D, Hall D, Xia B et al (2017) The ClusPro web server for protein-protein docking. Nat Protoc 12(2):255–278
14. Lee H, Seok C (2017) Template-based prediction of protein-peptide interactions by using GalaxyPepDock. Methods Mol Biol 1561:37–47
15. Moal IH, Bates PA (2010) SwarmDock and the use of normal modes in protein-protein docking. Int J Mol Sci 11(10):3623–3648
16. Li X, Moal IH, Bates PA (2010) Detection and refinement of encounter complexes for protein–protein docking: taking account of macromolecular crowding. Proteins 78(15):3189–3196
17. Torchala M, Moal IH, Chaleil RA et al (2013) SwarmDock: a server for flexible protein-protein docking. Bioinformatics 29(6):807–809
18. Torchala M, Bates PA (2014) Predicting the structure of protein-protein complexes using the SwarmDock web server. Methods Mol Biol 1137:181–197
19. Vajda S, Hall DR, Kozakov D (2013) Sampling and scoring: a marriage made in heaven. Proteins 81(11):1874–1884
20. Moal IH, Torchala M, Bates PA, Fernandez-Recio J (2013) The scoring of poses in protein-protein docking: current capabilities and future directions. BMC Bioinformatics 14:286
21. Barradas-Bautista D, Moal IH, Fernández-Recio J (2017) A systematic analysis of scoring functions in rigid-body protein docking: the delicate balance between the predictive rate improvement and the risk of overtraining. Proteins. https://doi.org/10.1002/prot.25289
22. Hayes TW, Moal IH (2017) Modeling protein conformational transition pathways using collective motions and the LASSO method. J Chem Theory Comput 13(3):1401–1410
23. Liu S, Zhang C, Zhou H, Zhou Y (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins 56(1):93–101
24. Kennedy J, Eberhart R (1995) Particle swarm optimization. IEEE International Conference on Neural Networks, Perth
25. Solis FJ, Wets RJ-B (1981) Minimization by random search techniques. Math Oper Res 6(1):19–30
26. Tobi D (2010) Designing coarse grained-and atom based-potentials for protein-protein docking. BMC Struct Biol 10:40
27. Torchala M, Moal IH, Chaleil RA et al (2013) A Markov-chain model description of binding funnels to enhance the ranking of docked solutions. Proteins 81(12):2143–2149
28. Moal IH, Barradas-Bautista D, Jiménez-García B et al (2017) IRaPPA: information retrieval based integration of biophysical models for protein assembly selection. Bioinformatics 33(12):1806–1813. https://doi.org/10.1093/bioinformatics/btx068
29. Dobbins SE, Lesk VI, Sternberg MJE (2008) Insights into protein flexibility: the relationship between normal modes and conformational change upon protein-protein docking. Proc Natl Acad Sci U S A 105(30):10390–10395
30. Karaca E (1993) Bonvin AMJJ (2011) a multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes. Structure 19(4):555–565
31. Marsh JA, Teichmann SA (2011) Relative solvent accessible surface area predicts protein conformational changes upon binding. Struct 19(6):859–867
32. Chen H, Sun Y, Shen Y (2017) Predicting protein conformational changes for unbound and homology docking: learning from intrinsic and induced flexibility. Proteins 85(3):544–556
33. Wang Q, Canutescu AA, Dunbrack RL (2008) SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. Nat Protoc 3(12):1832–1847
34. Soto CS, Fasnacht M, Zhu J et al (2008) Loop modeling: sampling, filtering, and scoring. Proteins 70(3):834–843
35. Brooks BR, Brooks CL, Mackerell AD et al (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30(10):1545–1614
36. Suhre K, Sanejouand Y-H (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates

for molecular replacement. Nucleic Acids Res 32:W610–W614
37. Engelbrecht AP (2005) Fundamentals of computational swarm intelligence. Wiley, Hoboken, NJ
38. Moal IH, Jimenez-Garcia B, Fernandez-Recio J (2015) CCharPPI web server: computational characterization of protein-protein interactions from structure. Bioinformatics 31(1):123–125
39. Pfeiffenberger E, Chaleil RAG, Moal IH, Bates PA (2017) A machine learning approach for ranking clusters of docked protein-protein complexes by pairwise cluster comparison. Proteins 85(3):528–543
40. van Zundert GCP, Rodrigues JPGLM, Trellet M et al (2016) The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. J Mol Biol 428(4):720–725
41. Svergun DI, Richard S, Koch MH et al (1998) Protein hydration in solution: experimental observation by x-ray and neutron scattering. Proc Natl Acad Sci U S A 95(5):2267–2272
42. Svergun D, Barberato C, Koch MHJ (1995) CRYSOL – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 28(6):768–773
43. Shvartsburg AA, Jarrold MF (1996) An exact hard-spheres scattering model for the mobilities of polyatomic ions. Chem Phys Lett 261(1–2):86–91
44. Lasker K, Sali A, Wolfson HJ (2010) Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. Proteins 78(15):3205–3211
45. Russel D, Lasker K, Webb B et al (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol 10(1):e1001244
46. Moal IH, Fernández-Recio J (2012) SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinformatics 28(20):2600–2607
47. Fowler DM, Fields S (2014) Deep mutational scanning: a new style of protein science. Nat Methods 11(8):801–807
48. Andreani J, Faure G, Guerois R (2012) Versatility and invariance in the evolution of homologous heteromeric interfaces. PLoS Comput Biol 8(8):e1002677
49. Reichmann D, Rahat O, Albeck S et al (2005) The modular architecture of protein-protein binding interfaces. Proc Natl Acad Sci U S A 102(1):57–62
50. McDowall MD, Scott MS, Barton GJ (2009) PIPs: human protein-protein interaction prediction database. Nucleic Acids Res 37:D651–D656
51. Liu F, Rijkers DTS, Post H, Heck AJR (2015) Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat Methods 12(12):1179–1184
52. Sethi R, Seppälä J, Tossavainen H et al (2014) A novel structural unit in the N-terminal region of filamins. J Biol Chem 289(12):8588–8598
53. Kao A, Chiu C-l, Vellucci D et al (2011) Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. Mol Cell Proteomics. https://doi.org/10.1074/mcp.M110.002212

Chapter 28

Protein-Protein Docking Using Evolutionary Information

Aravindan Arun Nadaradjane, Raphael Guerois, and Jessica Andreani

Abstract

The structural modeling of protein complexes by docking simulations has been attracting increasing interest with the rise of proteomics and of the number of experimentally identified binary interactions. Structures of unbound partners, either modeled or experimentally determined, can be used as input to sample as extensively as possible all putative binding modes and single out the most plausible ones. At the scoring step, evolutionary information contained in the joint multiple sequence alignments of both partners can provide key insights to recognize correct interfaces. Here, we describe a computational protocol based on the InterEvDock web server to exploit coevolution constraints in protein-protein docking methods. We provide methodology guidelines to prepare the input protein structures and generate improved alignments. We also explain how to extract and use the information returned by the server through the analysis of two representative examples.

Key words Protein docking, Protein interactions, Protein structure, Protein scoring, Evolutionary information, Coevolution, Bioinformatics, InterEvDock, InterEvolAlign, Complex interface

1 Introduction

Protein-protein interactions orchestrate many of the cellular pathways through complex and dynamic networks of cross talks taking place at the level of protein surfaces. Understanding the molecular logic associated with these networks represents a major challenge, which can be best tackled nowadays through a combination of low- to high-throughput proteomics, structural biology, and functional assays. In this framework, bioinformatics methods are particularly interesting for integrating and predicting the properties of the interaction network at different scales [1]. Here, we will present a computational method that can be used to predict the structure of complexes between proteins through a server-based interface. The server belongs to the broad class of protein docking methods [2], in which the structural knowledge of the binding proteins is used to predict their most likely assembly modes. Such information can, for instance, help decipher how two proteins

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764,
https://doi.org/10.1007/978-1-4939-7759-8_28,

compete or synergize with each other at the surface of a third partner. It is also very useful to design mutants that specifically abrogate one interaction of the network while keeping the rest unaffected [3, 4]. Reaching sufficiently precise models of interfaces can even allow the design of compensatory mutants, in which one mutation in a partner disrupts the interaction while a second mutation, in the other partner, can compensate for the deleterious effect of the first one [5].

The field of protein docking has greatly benefited from the community-wide assessment of prediction methods promoted by the CAPRI experiment, in which different docking strategies have been compared with each other in a blind manner [6–8]. To predict the structure of a complex, most currently successful methods rely on a two-step procedure. First, protein structures of the unbound partners are considered as rigid bodies, and a thorough sampling of the possible assemblies between them can be efficiently performed using methods such as FFT-based algorithms [9]. This step typically generates between 10^4 and 10^5 solutions—called decoys—which have to be clustered and ranked so as to extract a limited set of likely solutions. The second step of docking focuses on a limited set of selected decoys and integrates flexibility and more refined scoring metrics. Because the second step requires considerably more computer resources than the first one, it is crucial that efficient pruning of unlikely solutions is performed at the initial rigid-body step.

The InterEvDock server [10], which will be presented in this chapter, is useful for the first step of docking and proposes an original combination of methods to increase the performance of initial selection of rigid-body docking solutions. One of the specificities of the method is to exploit the evolutionary information contained in the multiple sequence alignments of the binding partners and give more credit to the decoys that are compatible with all the sequences contained in the alignments and not only the queried ones. This constraint is powerful because structures of complexes tend to be conserved in evolution provided that the sequences of the binding partners did not diverge to less than 30% sequence identity [11–13]. Below this threshold, assembly modes may still be conserved, but care should be taken since there are a number of examples showing different binding orientations. Evolutionary information can be integrated in the scoring of complexes in different manners. First, the conservation of patches at the surface of proteins can provide valuable hints about the location of binding surfaces [14, 15]. However, these proxies lack specificity since we do not know which residues are facing each other in the interface. To circumvent this limitation, other approaches were recently developed with the aim of extracting co-varying pairs of residues [16–18]. While such methods work in a number of specific cases, their application is still currently limited to cases in which a large

number of sequences can be collected in the alignment (typically several hundred non-redundant sequences). The InterEvDock server proposes a third track for recognizing coevolved interfaces. The general principle of the method was inspired by a large-scale pairwise comparison of interfaces whose subunits in contact were two-by-two homologues [19]. We could quantify at the residue level the type of structural plasticity, which occurred at the interface of protein complexes. From this analysis, we derived a scoring scheme named InterEvScore [20], which assesses through residue-based, two- and three-body statistical potentials whether contacts between pairs or triplets of residues of a possible interface (docking decoy) are often observed in real interfaces, compared to protein surfaces. This scoring is not only performed at the level of a single species but also computed for all sequences contained in the multiple sequence alignments of the binding partners (co-alignments). Unlike the covariation approaches described above, a performance gain, thanks to accounting for coevolution of interfaces in InterEvScore, is observed with alignments containing as few as 10 (and up to 100) sequences.

The principle of InterEvScore is to analyze interface likelihood at the residue level. To complement this coarse-grained representation, two additional scoring metrics were combined with InterEvScore in the InterEvDock server to best account for the quality of the atomic interactions. First, the FRODOCK method [21], which is used to sample the decoys scored by InterEvDock, implements an FFT-based algorithm with a hybrid physics- and statistical-based potential which is integrated in the InterEvDock scoring framework. Second, we also took advantage of the SOAP-PP statistical atomic potential [22], which was found to provide highly complementary scores to those calculated by InterEvScore [23]. Altogether, the three scoring metrics InterEvScore, FRODOCK, and SOAP-PP are calculated for each docking decoy, and a consensus score is used to finally rank the most likely solution. The way this consensus is calculated was optimized based on a benchmark of 85 reference protein complexes which showed that we could reach a success rate of up to 49% on docking cases belonging to the rigid-body category [10]. To reach the best performance, users have to be careful with the generation of the co-alignments. These co-alignments can be generated automatically by the InterEvDock server for the simple cases. Sometimes, it is not trivial to recover in an automatic manner the correct pair of sequences corresponding to the correct orthologous pair of interacting partners. For those cases, we have developed the InterEvolAlign server [13] which circumvents some of the difficulties by implementing a reciprocal blast procedure which increases the confidence that sequences added in the alignment are the proper orthologs of the queried proteins. These more difficult cases will be discussed further in the chapter.

Our purpose here is to present step-by-step usage of the InterEvDock docking server and how to analyze docking results, as well as step-by-step usage of the InterEvolAlign web server. We will use two case studies described in Table 1 to illustrate this protocol. Case #1 is an interface between two regulatory subunits of the proteasome from yeast *Saccharomyces cerevisiae*, Nas6, and Rpt3 (C-terminal AAA+ ATPase domain) [24]. This case is rather standard from the point of view of protein docking and interesting from the point of view of evolution, since Rpt3 is highly conserved in eukaryotes from yeast to human, while Nas6 exists in yeast and human but is not conserved in some species, for instance, insects. Case #2 comes from the protein docking benchmark version 4 [25] and is an interface between adenylyl cyclase and its stimulatory heterotrimeric G-protein alpha subunit (Gsalpha) [26]. This case is technically interesting because adenylyl cyclase is actually a homodimeric protein. In both cases, the structures of the unbound protein partners are available and used as input for docking in order not to bias the search toward the correct interface by using bound structures presenting perfect shape complementarity. In both cases, rigid-body protein docking can generate models of acceptable or better quality according to the CAPRI criteria [27], and we can use the information about how the two protein partners coevolved in our docking protocol. The structure of the protein complex has been experimentally determined so that we can compare it to the models identified by InterEvDock.

Three major procedures will be described in the methods below: how to perform a docking run with the InterEvDock web server (Subheading 3.1), how to analyze results from this docking run (Subheading 3.2), and how to create fine-tuned multiple sequence alignments for two binding partners using InterEvolAlign (Subheading 3.3). These procedures in their standard version will be illustrated using case #1, referring the reader to the appropriate notes for a variant of case #1 and for case #2. The result analysis provided in Subheading 4 gives details about why the variants are interesting.

2 Materials

2.1 Input Data

1. The only input files that are mandatory to run InterEvDock are the atomic 3D coordinate files of two proteins that are known to interact directly (*see* **Note 1**) and for which you want to obtain structural models of the interface. These coordinates can be obtained either from experimental structural data deposited in the Protein Data Bank (PDB) or from structural models. Protein sequences can also be used as input (*see* **Note 7**).

   If the atomic coordinates of one (or both) protein(s) of interest constitute a whole single chain of a PDB structure,

Table 1
Details about the two docking cases used to illustrate our docking protocol

	Partner A	Partner A PBD id	Partner A Chain id	Partner A residue delimitations	Partner B	Partner B PBD id	Partner B chain id	Partner B residue delimitations	Protein complex PDB id	Protein complex chains
#1	Nas6	1WG0	A	–	Rpt3	4CR4	K	153–417	2DZN	A:B
#2	Adenylyl cyclase	1AB8	AB	–	Gsalpha	1AZT	A	–	1AZS	AB:C

then the PDB identifier and chain can be directly used as input in the InterEvDock web server interface. Otherwise, the atomic coordinates of the experimental structure or model (in PDB format) can be pasted or uploaded to the web server. Sometimes, structural coordinates that will be used as input need to be prepared specifically (*see* Subheading 3.1 and **Note 2**).

Structural models can be built for many proteins of unknown structure [28]. We suggest using the HHpred web server [29] to look for available structures or structural templates for the two proteins of interest. The MPI Bioinformatics Toolkit [30] can also be used to subsequently generate structural models using MODELLER [31]. In case the sequence identity between the protein of interest and the identified template is low (typically below 35%), our HHalign-Kbest web server [32] can be used to generate suboptimal alignments and then generate the corresponding structural models using MODELLER and evaluate them.

2. Although the InterEvDock web server can automatically generate multiple sequence alignments for the two proteins of interest, in some cases the user may wish to build the multiple sequence alignments prior to running InterEvDock. In this case, multiple sequence alignments can be given as input to the InterEvDock web server together with the atomic coordinates of the two proteins of interest. *See* **Note 3** on requirements about the multiple sequence alignments provided to InterEvDock. See case study details and **Notes 4** and **5** for examples of why it can be interesting to use custom multiple sequence alignments as input to InterEvDock.
3. To run the InterEvolAlign co-alignment building web server, only the sequences of two proteins (in FASTA format) are required.

2.2 Software

Many programs can be used for structure and multiple sequence alignment preparation and visualization. This protocol reports instructions using the following two widely used programs.

1. The PyMOL molecular visualization system (the PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC) is used in this protocol for input data manipulations and for output analysis. PyMOL can be obtained from http://pymol.org/.
2. Jalview [33] is used in this protocol to inspect and manipulate multiple sequence alignments. Jalview is a free program that can be downloaded from http://jalview.org/.

   Both programs support a variety of operating systems and hardware.

3 Methods

3.1 Running the InterEvDock Docking Protocol

The following steps for running the InterEvDock protocol on the web server are illustrated in Fig. 1.

1. Go to the InterEvDock job submission page:

   http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::InterEvDock2

   Relevant documentation can be found at:

   http://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock2/

2. Input the two protein structures to be used for docking into the appropriate fields. In the simplest situation, no specific structure preparation is required: either the atomic coordinates of the user's own structure or structural model can be uploaded to the web server or a PDB identifier can be provided as input. This is the case, for instance, for Nas6 in case #1 which is a single-domain protein: simply choose the PDB as the input database and select "1WG0A" (four-letter PDB identifier and one-letter PDB chain).

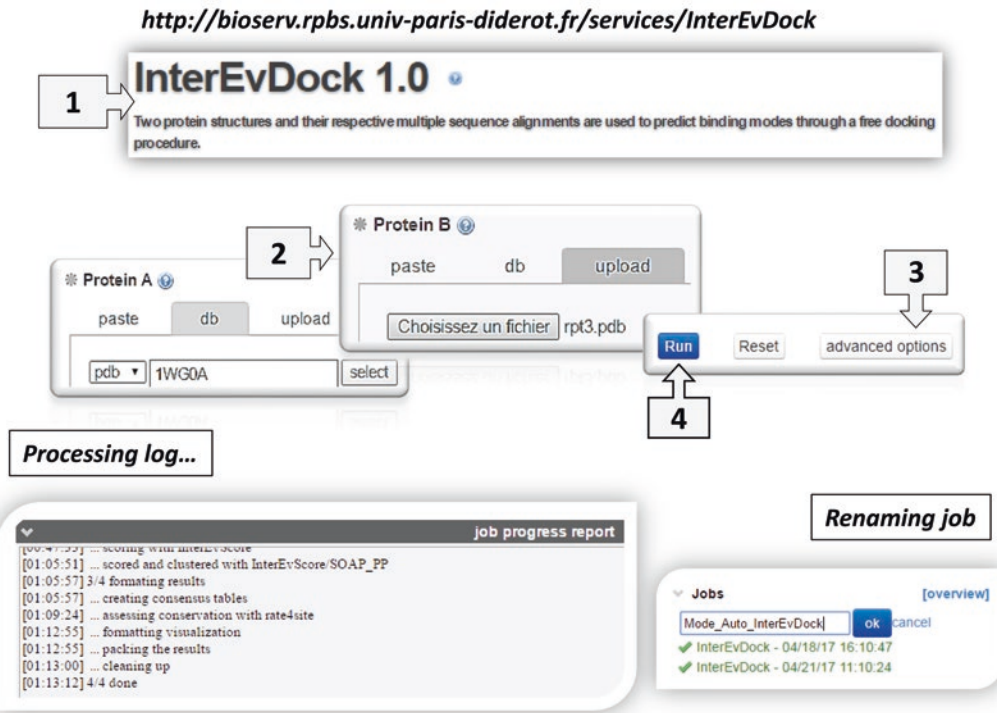


Fig. 1 Running the InterEvDock web server for protein-protein docking using evolutionary information

Simple structure manipulations to prepare input structures can be done with PyMOL. For Rpt3 in case #1, only one domain of the PDB structure will be used as input, so we need to provide InterEvDock with coordinates for only the AAA+ ATPase domain from Rpt3 (residues 153–417). This can be done by selecting those residues from PDB identifier 4CR4 and chain K using PyMOL and saving them into a coordinate file:

```
PyMOL> fetch 4CR4
PyMOL> extract rpt3, chain K and resi 153-417 in 4CR4
PyMOL> save rpt3.pdb, rpt3
```

This coordinate file rpt3.pdb can then be uploaded to the appropriate InterEvDock field.

In case #2, the two chains of the alpha subunit need to be merged. *See* **Notes 2** and **7** for more details about input format requirements and how to prepare input structures for InterEvDock when one partner is actually homomultimeric, i.e., contains multiple copies of the same protein.

3. InterEvDock can either compute automatically the co-alignments (default case) or take as input co-alignments generated independently by the user. In most cases, the automatic mode is sufficient. Therefore, in the standard situation, this step can be skipped. For case #1, a standard run with automatic computation of the co-alignments is sufficient to get final results of reasonable quality.

   Cases in which it might be interesting to build co-alignments independently are illustrated by our two examples. If users want to input their own co-alignments, special care is needed for formatting (*see* **Note 3**).

   In case #1, because Nas6 is not conserved in some species, generating co-alignments that do not include undesired homologues (mainly paralogs of Nas6 in species where Nas6 does not have an ortholog) gives more relevant coevolutionary information about the interface and therefore improves docking results (*see* Subheading 4.1). Building a multiple sequence alignment that limits the risks of including undesired homologues can be a tedious task. For this purpose, we developed the InterEvolAlign web server [13]; *see* **Note 4** on when to use InterEvolAlign to create input co-alignments for InterEvDock and protocol Subheading 3.3 on how to use InterEvolAlign.

   In case #2, because the alpha subunit is a homodimer, the automated co-alignment generation method in InterEvDock will not work. Therefore, custom input co-alignments need to be used for this case; *see* **Note 5** on how to create co-alignments for cases with multiple chains. *See* **Note 7** for new features in the co-alignment of oligomeric inputs.

4. Press "Run." The following steps will be undertaken automatically by the InterEvDock web server:
   - Rigid-body docking with FRODOCK [21], followed by clustering and generation of 10,000 decoys.
   - Scoring and clustering using the FRODOCK scoring function [21] and the SOAP-PP scoring function [22].
   - If co-alignments were not provided as input, multiple sequence alignment generation for both partners (*see* **Note 4** for details).
   - Scoring and clustering with InterEvScore [20]
   - Computation of the top ten (consensus) models, the Rate4Site [34] conservation scores, the consensus interface residues, and the structure preparation for visualization.

The InterEvDock web interface offers user-friendly options to reuse server inputs or server outputs in multiple runs (*see* **Note 6**).

3.2 Analyzing InterEvDock Results

The following steps for online analysis of InterEvDock results are illustrated in Fig. 2.

3.2.1 Online Analysis

1. In the visualization panel, switch between models using the "Complex" dropdown menu. Scroll down to the "Top 10 best InterEvDock consensus models" textbox for a list of the ten best InterEvDock models, starting with the most likely.
2. In the visualization panel, switch to "Color by Residue Interface Consensus." Use the slider to tune the contrast. Scroll down to the "5 most likely residues involved in the complex interface" to find out the five residues on each chain that are most often found in the interface of the best InterEvDock models. This information can be used to guide mutagenesis of the interface. The likelihood that at least one of these residues is in the interface is very high: up to 90% according to a recent benchmark [10].
3. In the visualization panel, switch to "Color by Conservation (Rate4Site)." Use the slider to tune the contrast. This maps the conservation score calculated by Rate4Site [34] onto the structure of the models. Color code is a gradient from red (more conserved) to yellow (mild conservation) to white (more diverse). This analysis is useful to identify the most conserved regions on the surface of both partners, since interface regions are more conserved than non-interface surface [19, 35, 36].
4. In the visualization panel, when hovering over a residue in the structure, this residue gets highlighted in yellow, and residue information appears in the bottom right of the panel. For instance, the "5 most likely residues involved in the complex interface" can be identified on the structure.

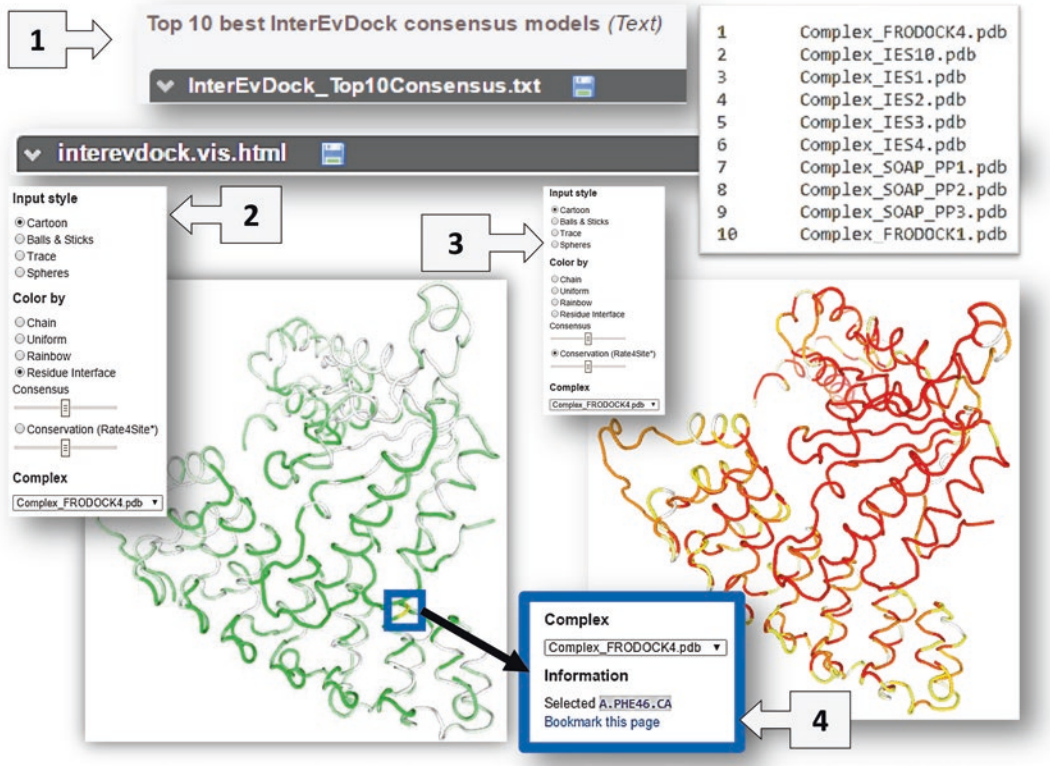


Fig. 2 Online analysis of InterEvDock results

3.2.2 *Offline Analysis*

The following steps for offline analysis of InterEvDock results are illustrated in Fig. 3.

1. Download the results of the InterEvDock run from the server by clicking the "save" icon next to "results.zip". Extract the archive to a convenient location.
2. Double-click on start_analysis.pml. This will load the top ten models from each score into PyMOL and post-process them.
3. Look at the loaded models. "Conservation index" is mapped on chainA_conservation (white-yellow-red scale) and chainB_conservation (white-cyan-blue scale) and "consensus probability of residues to be in interface" on the chainA_res_consensus and chainB_res_consensus (white-green scale) objects.
4. Analysis of a specific interface can be done using the instruction:

```
PyMOL> analyze_complex Complex_IES10
```

This will show the interface residues in model Complex_IES10 as sticks and display hydrogen bonds at the interface as dashed lines.

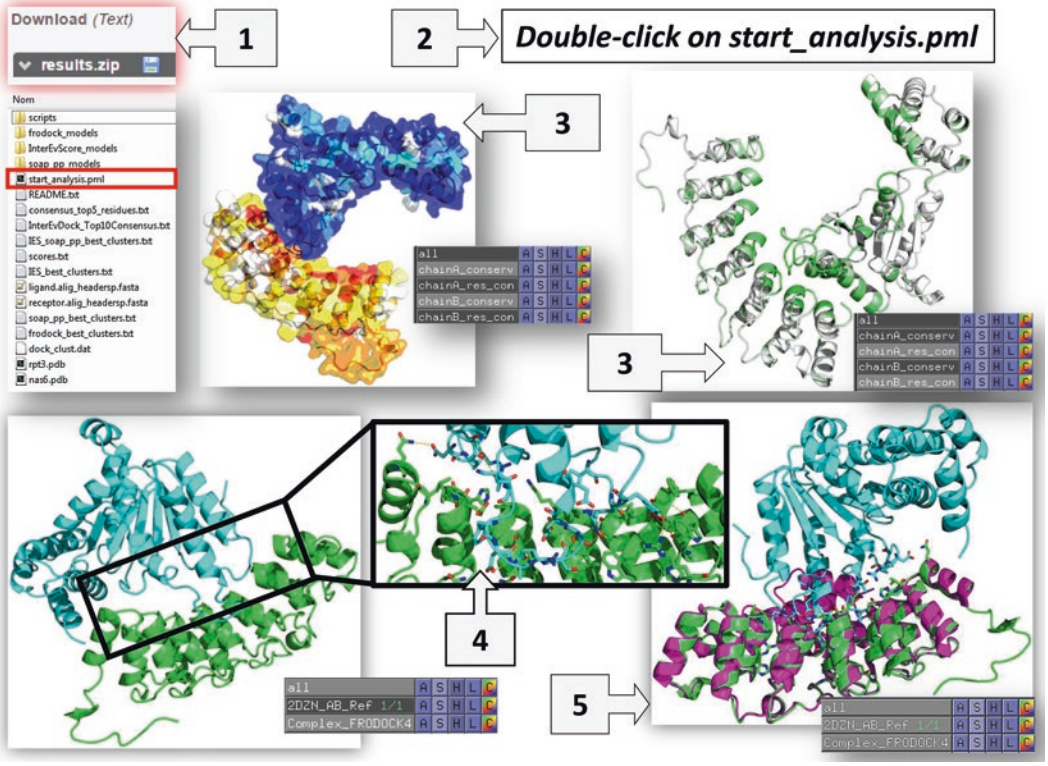


Fig. 3 Offline analysis of InterEvDock results

Visualization of interface conservation and interface residue consensus mapping on a specific model can be done using the following instructions:

```
PyMOL> align chainB_conservation, /Complex_IES10//B
PyMOL> align chainB_res_consensus, /Complex_IES10//B
```

(Chain A objects do not need to be aligned.)

5. Models can be compared to a reference complex, for instance, in benchmark cases or when the structure of a homologous complex is known. In this case, load the reference complex into the PyMOL session and align it to the model of interest. For case #1, the reference complex is PDB structure 2DZN, chains A and B. To display this reference complex, use the following commands in PyMOL:

```
PyMOL> fetch 2DZN
PyMOL> extract 2DZN_AB, 2DZN and chain A or 2DZN and chain B
```

```
PyMOL> delete 2DZN
PyMOL> as cartoon, 2DZN_AB
```

To compare the reference structure with the docking models, align the first chain of the reference structure to the chain A of all models:

```
PyMOL> align 2DZN_AB and chain A, chainA_
conservation
```

3.3 Using InterEvolAlign to Generate Co-alignments

The following steps for generating co-alignments using the InterEvolAlign web server are illustrated in Fig. 4.

1. Go to the InterEvolAlign web server at:

   http://biodev.cea.fr/interevol/interevalign.aspx.

2. Input the two protein sequences. InterEvolAlign needs as input the two sequences of the query proteins in FASTA format. To get those for protein structure inputs, PyMOL can be used. For case #1:

```
PyMOL> fetch 1WG0
PyMOL> save nas6.fasta, 1WG0 and chain A
```

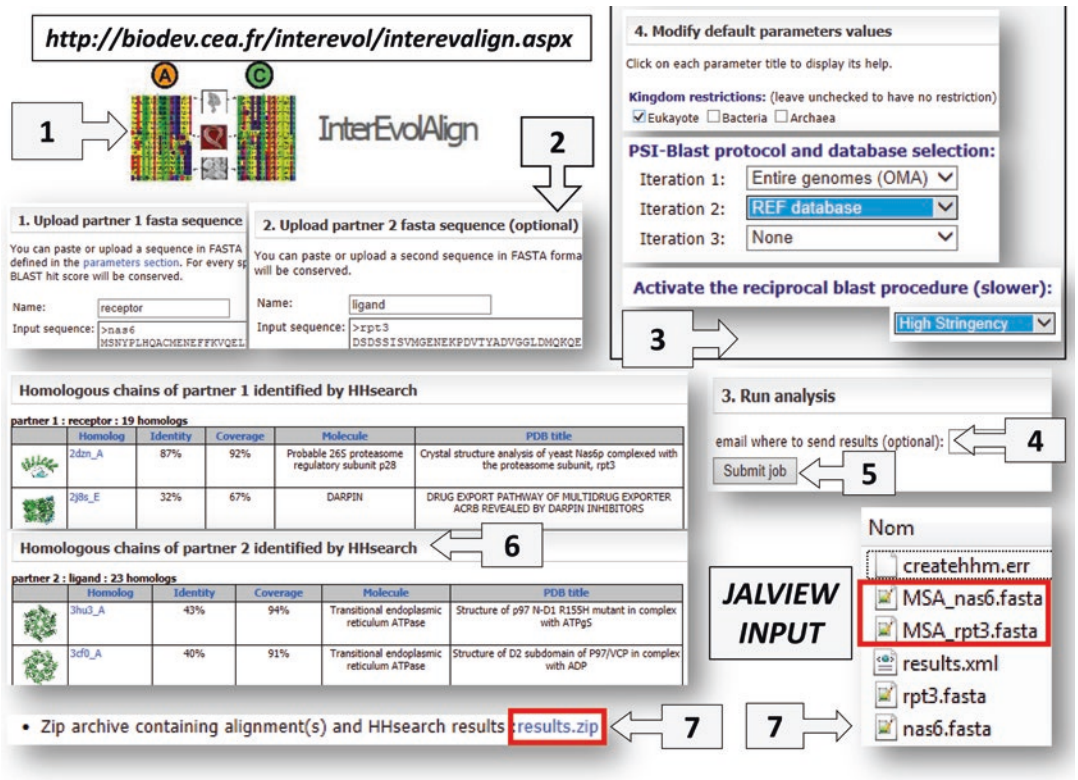


Fig. 4 Running the InterEvolAlign web server to generate multiple sequence alignments for two protein partners

```
PyMOL> fetch 4CR4
PyMOL> extract rpt3, chain K and resi 153-
417 in 4CR4
PyMOL> save rpt3.fasta, rpt3
```

The two protein sequences (in case #1: nas6.fasta and rpt3.fasta) can be pasted or uploaded into the relevant fields of the InterEvolAlign web server.

3. InterEvolAlign options can be tuned as needed. In general, it is recommended to use the following options for a first trial run:
   - Restrict the search to the relevant superkingdom. In the case of Nas6/Rpt3, choose "Eukaryote."
   - Use two iterations: iteration 1 on the entire genomes (OMA) database (to recover all possible orthologs in fully sequenced genomes) and then iteration 2 on the REF database (to retrieve more sequences).
   - Activate the reciprocal blast procedure, with either medium or high stringency depending on the characteristics of the case at hand. For Nas6/Rpt3, choose high stringency because we know that ankyrin repeats such as Nas6 have many paralogs.
4. Optionally, input your email address to receive an email when results are ready (typically around 30 min when no other requests are running).
5. Click "Submit job" to start the InterEvolAlign procedure. InterEvolAlign will perform an iterative search on the selected databases.
6. Check the results page to see how many sequences were retrieved by the search. The results page also provides a list of homologous structures for each partner identified by HHsearch [37]. This can be useful if models need to be built for the two proteins of interest.
7. Download and extract the results.zip archive. Jalview can be used to view and analyze alignments. This is useful to decide, for instance, whether to rerun InterEvolAlign with different options. Alignments can also be downloaded separately or viewed interactively with Jalview from the results web page.

4 Case Studies

4.1 Case Study #1: Nas6/Rpt3

InterEvDock implements a cross-consensus selection to single out models from the ten best models identified by each of the three scoring functions (FRODOCK, InterEvScore, and SOAP-PP) that are similar to models identified by other methods in their top 50 selection. The top ten best models of the three scores are considered

in the final ranking. At first, the best three models of every method are selected (four models for InterEvScore) to create a consensus of ten models. This selection is then modified if other models were identified by at least two different methods.

In the Nas6-Rpt3 example, a model such as Complex_IES10 (acceptable according to the CAPRI classification) was not part of the initial consensus selection. However, since a similar model was also identified by SOAP-PP and FRODOCK among their top 50 models (Complex_SOAP_PP12 and Complex_FRODOCK47), Complex_IES10 was re-ranked favorably among the top ten consensus selections of InterEvDock as top two.

When comparing the reference complex (2DZN, chains A and B) with the Nas6/Rpt3 models, the top five consensus residues most likely to be involved in the interface can be analyzed. Three out of five residues predicted for chain A (Nas6) actually belong to the interface (S45, K117, W73), and the remaining two are close by (F46, W119). On the contrary, none of the five residues predicted as most likely to be involved in the interface in chain B (Rpt3) is actually present in the 2DZN crystal structure. Those five residues predicted on chain B form a patch on the opposite side of Rpt3 compared to the binding site of Nas6. This highlights the difficulty of selecting a correct docking model when most models involve the wrong binding region on Rpt3.

When looking at the residue conservation mapped on the surface of chains A and B, it is clear that the most conserved region on the surface of Nas6 is involved in binding Rpt3, while the Rpt3 surface is overall quite conserved, and the most conserved region is actually not the region binding Nas6, but rather the region containing the top five residues (wrongly) predicted as most likely to be involved in the interface.

It is also interesting to consider the results of an additional InterEvDock run using customized co-alignments (those generated by protocol Subheading 3.3). By using InterEvolAlign, the quality of the alignments was optimized. It has a direct impact on the selection of the most likely model: using these optimized co-alignments, the model ranked first in the top ten consensus is Complex_IES2.pdb which is actually of medium quality according to the CAPRI classification. Analyzing both sets of co-alignments (for instance, using Jalview) reveals that the limiting factor in the co-alignments automatically generated by InterEvDock was Rpt3, for which too little sequence diversity was retrieved.

4.2 Case Study #2: Adenylyl Cyclase and Its Activator Gsalpha

When comparing the best models selected by InterEvDock with the reference structure (PDB identifier 1AZS, chains A/B against C), the best InterEvScore model Complex_IES1.pdb (ranked second in the top ten consensus classifications) is of acceptable quality according to the CAPRI criteria. No model of better CAPRI quality is found among the top 150 models (top 50 models ranked by each of the three scoring functions).

Three out of five residues ranked in the top five interface consensus on adenylyl cyclase (F898, F991, and N992), and five out of five on the GSalpha chain (Q236, N239, I235, W281, R232) actually belong to the interface.

5 Notes

1. Note of caution about protein-protein interactions.

   InterEvDock assumes the interaction between the two submitted proteins has been experimentally validated. It is not designed to predict either the likelihood or the strength of the interaction.

2. How to prepare input structures for InterEvDock.

   Each protein structure input should contain a single-chain identifier. The size of each submitted protein should be between 10 and 3000 amino acids. The web server is currently not able to perform docking with nucleic acids.

   For case #2, the alpha subunit is homodimeric, but residues from both monomers need to be saved into a single chain (with no intermediate TER records) that will be submitted to the InterEvDock web server. (*see* **Note 7** for new features allowing to handle oligomers as inputs). This can be done using PyMOL:

   ```
   PyMOL> fetch 1AB8
   PyMOL> extract 1AB8_AB, 1AB8 and chain A or 1AB8 and chain B
   PyMOL> alter 1AB8_AB, chain="A"
   PyMOL> set pdb_use_ter_records, 0
   PyMOL> save 1AB8_AB.pdb, 1AB8_AB
   ```

3. Requirements for co-alignments given as direct input to InterEvDock.

   In order for the coevolutionary information to be correctly taken into account in the docking process, the user has to ensure that the multiple sequence alignments contain sequences in exactly the same species for the two proteins and that these species appear in the exact same order in the two alignments. This is always the case for alignments generated automatically by InterEvDock or by our dedicated InterEvolAlign web server.

4. How InterEvDock generates co-alignments and when to prepare co-alignments using InterEvolAlign.

   The InterEvDock server generates multiple sequence alignments for partners A and B with a single blast request [38] against the UniRef database (the Uniprot consortium, 2014) and keeps only sequences present in the same species. The first hit found for each species is selected as the putative orthologous

sequence. Sequence redundancy is removed for sequence identities >90%. Only the 100 sequences closest to the input sequences are kept. Sequences with less than 30% sequence identity are discarded. Sequence coverage with the input sequence is required to be larger than 75%. This procedure is generally sufficient to produce coevolutionary information about the interaction that can be used for scoring by InterEvScore.

InterEvolAlign [13] is useful when we need to tune more specifically the properties of the co-alignments. It has three major features:

- A systematic reciprocal blast procedure on the sequences of the alignment can be activated to prune sequences unlikely to correspond to true orthologs. The runtime is longer with this option, but typically in the case of Nas6, this is exactly what we need to remove homologues from insect species. Two levels of stringency are available; in the case of Nas6, the high-stringency threshold yields the best result.
- InterEvolAlign can run on the OMA database [39] which is restricted to organisms which are fully sequenced. It limits the risk that partially sequenced organisms can introduce significant noise if the expected ortholog is not present in the sequence database.
- The parameters of the search can be tuned and adapted to all types of sequences with either many or few homologous sequences.

Generally, InterEvolAlign generates alignments with fewer sequences than the automatic InterEvDock procedure and takes longer, but the content of the alignment can be more specifically tuned. Therefore, InterEvolAlign should be used preferentially in specific cases, such as for Nas6 which we know does not have orthologs in insect species.

5. How to prepare multiple sequence alignments in cases where one partner contains multiple copies of the same chain. (*see* **Note 7** for new features released in InterEvDock2).

   The first step is to use a blank run of InterEvDock to generate co-alignments for a single copy of each chain. For instance, in the case of NAD (P) transhydrogenase subunit alpha, a blank run of InterEvDock using PDB identifiers 1AB8A and 1AZTA should be run. Then the multiple sequence alignment for partner A (one chain of the alpha subunit) should be used to build an alignment for the A-A homodimer:

   - Display the alignment for partner A (receptor.alig_headersp.fasta in the results.zip archive) with Jalview.
   - Through the "File" menu, output to text box in CLUSTAL format.

- Duplicate the alignment in the text box (by copy-pasting all lines).
- Create a "New Jalview Window" from the duplicated alignment.

After saving the new alignment in FASTA format, advanced options in InterEvDock can be used to submit it as "Protein A co-alignment," where protein A is the structure of the dimeric alpha subunit (*see* **Note 2**). The alignment for partner B retrieved from the blank InterEvDock run (ligand.alig_headersp.fasta in the results.zip archive) should be reused for partner B in this new run.

6. Tips about using the InterEvDock web interface.

   In the InterEvDock web interface, jobs can be renamed in the left margin for easier identification. A nice feature of the Mobyle environment in which InterEvDock is running is that all files (inputs or outputs) are kept as data bookmarks which can be retrieved later by the user for subsequent usage (top left). Their names can be changed as in the case of the job name.

7. New features in InterEvDock2 with respect to InterEvDock.

   With the release of the new InterEvDock2 server several features became available. The sequence of a partner can be submitted instead of its 3D structure. In that case, a homology model of the subunit is generated prior to docking. It is possible to submit the structure of an oligomeric assembly as input. In that case, the server is now able to automatize the operations described in **Notes 2** and **5**.

References

1. Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 7(3):188–197. https://doi.org/10.1038/nrm1859
2. Vajda S, Kozakov D (2009) Convergence and combination of methods in protein-protein docking. Curr Opin Struct Biol 19(2):164–170. https://doi.org/10.1016/j.sbi.2009.02.008
3. Dreze M, Charloteaux B, Milstein S, Vidalain PO, Yildirim MA, Zhong Q, Svrzikapa N, Romero V, Laloux G, Brasseur R, Vandenhaute J, Boxem M, Cusick ME, Hill DE, Vidal M (2009) 'Edgetic' perturbation of a C. Elegans BCL2 ortholog. Nat Methods 6(11):843–849. https://doi.org/10.1038/nmeth.1394
4. Wang Y, Sahni N, Vidal M (2015) Global edgetic rewiring in cancer networks. Cell Syst 1(4):251–253. https://doi.org/10.1016/j.cels.2015.10.006
5. Kadota Y, Amigues B, Ducassou L, Madaoui H, Ochsenbein F, Guerois R, Shirasu K (2008) Structural and functional analysis of SGT1-HSP90 core complex required for innate immunity in plants. EMBO Rep 9(12):1209–1215. https://doi.org/10.1038/embor.2008.185
6. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ, Critical Assessment of PI (2003) CAPRI: a critical assessment of predicted interactions. Proteins 52(1):2–9. https://doi.org/10.1002/prot.10381
7. Wodak SJ, Janin J (2017) Modeling protein assemblies: critical assessment of predicted interactions (CAPRI) 15 years hence.: 6TH CAPRI evaluation meeting April 17-19 Tel-Aviv, Israel. Proteins 85(3):357–358. https://doi.org/10.1002/prot.25233
8. Lensink MF, Velankar S, Wodak SJ (2017) Modeling protein-protein and protein-peptide

complexes: CAPRI 6th edition. Proteins 85(3):359–377. https://doi.org/10.1002/prot.25215
9. Huang SY (2014) Search strategies and evaluation in protein-protein docking: principles, advances and challenges. Drug Discov Today 19(8):1081–1096. https://doi.org/10.1016/j.drudis.2014.02.005
10. Yu J, Vavrusa M, Andreani J, Rey J, Tuffery P, Guerois R (2016) InterEvDock: a docking server to predict the structure of protein-protein interactions using evolutionary information. Nucleic Acids Res 44(W1):W542–W549. https://doi.org/10.1093/nar/gkw340
11. Aloy P, Ceulemans H, Stark A, Russell RB (2003) The relationship between sequence and interaction divergence in proteins. J Mol Biol 332(5):989–998
12. Levy ED, Boeri Erba E, Robinson CV, Teichmann SA (2008) Assembly reflects evolution of protein complexes. Nature 453(7199): 1262–1265. https://doi.org/10.1038/nature06942
13. Faure G, Andreani J, Guerois R (2012) InterEvol database: exploring the structure and evolution of protein complex interfaces. Nucleic Acids Res 40(Database issue):D847–D856. https://doi.org/10.1093/nar/gkr845
14. Ofran Y, Rost B (2007) ISIS: interaction sites identified from sequence. Bioinformatics 23(2):e13–e16. https://doi.org/10.1093/bioinformatics/btl303
15. Res I, Mihalek I, Lichtarge O (2005) An evolution based classifier for prediction of protein interfaces without using protein structures. Bioinformatics 21(10):2496–2501. https://doi.org/10.1093/bioinformatics/bti340
16. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A 106(1):67–72. https://doi.org/10.1073/pnas.0805923106
17. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30(11):1072–1080. https://doi.org/10.1038/nbt.2419
18. Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H, Grishin NV, Baker D (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. elife 4:e09248. https://doi.org/10.7554/eLife.09248
19. Andreani J, Faure G, Guerois R (2012) Versatility and invariance in the evolution of homologous heteromeric interfaces. PLoS Comput Biol 8(8):e1002677. https://doi.org/10.1371/journal.pcbi.1002677
20. Andreani J, Faure G, Guerois R (2013) InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. Bioinformatics 29(14):1742–1749. https://doi.org/10.1093/bioinformatics/btt260
21. Garzon JI, Lopez-Blanco JR, Pons C, Kovacs J, Abagyan R, Fernandez-Recio J, Chacon P (2009) FRODOCK: a new approach for fast rotational protein-protein docking. Bioinformatics 25(19):2544–2551. https://doi.org/10.1093/bioinformatics/btp447
22. Dong GQ, Fan H, Schneidman-Duhovny D, Webb B, Sali A (2013) Optimized atomic statistical potentials: assessment of protein interfaces and loops. Bioinformatics 29(24): 3158–3166. https://doi.org/10.1093/bioinformatics/btt560
23. Yu J, Andreani J, Ochsenbein F, Guerois R (2017) Lessons from (co-)evolution in the docking of proteins and peptides for CAPRI rounds 28-35. Proteins 85(3):378–390. https://doi.org/10.1002/prot.25180
24. Nakamura Y, Umehara T, Tanaka A, Horikoshi M, Padmanabhan B, Yokoyama S (2007) Structural basis for the recognition between the regulatory particles Nas6 and Rpt3 of the yeast 26S proteasome. Biochem Biophys Res Commun 359(3):503–509. https://doi.org/10.1016/j.bbrc.2007.05.138
25. Hwang H, Vreven T, Janin J, Weng Z (2010) Protein-protein docking benchmark version 4.0. Proteins 78(15):3111–3114. https://doi.org/10.1002/prot.22830
26. Tesmer JJ, Sunahara RK, Gilman AG, Sprang SR (1997) Crystal structure of the catalytic domains of adenylyl cyclase in a complex with Gsalpha.GTPgammaS. Science 278(5345): 1907–1916
27. Lensink MF, Mendez R, Wodak SJ (2007) Docking and scoring protein complexes: CAPRI 3rd edition. Proteins 69(4):704–718. https://doi.org/10.1002/prot.21804
28. Mosca R, Ceol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. Nat Methods 10(1):47–53. https://doi.org/10.1038/nmeth.2289
29. Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33(Web Server):W244–W248. https://doi.org/10.1093/nar/gki408
30. Alva V, Nam SZ, Soding J, Lupas AN (2016) The MPI bioinformatics toolkit as an integrative platform for advanced protein sequence and structure analysis. Nucleic Acids Res 44(W1):W410–W415. https://doi.org/10.1093/nar/gkw348

31. Webb B, Sali A (2016) Comparative protein structure modeling using MODELLER. Curr Protoc Bioinformatics 54:5 6 1–5 6 37. https://doi.org/10.1002/cpbi.3
32. Yu J, Picord G, Tuffery P, Guerois R (2015) HHalign-Kbest: exploring sub-optimal alignments for remote homology comparative modeling. Bioinformatics 31(23):3850–3852. https://doi.org/10.1093/bioinformatics/btv441
33. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9): 1189–1191. https://doi.org/10.1093/bioinformatics/btp033
34. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18(Suppl 1):S71–S77
35. Teichmann SA (2002) The constraints protein-protein interactions place on sequence divergence. J Mol Biol 324(3):399–407
36. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13(1):190–202. https://doi.org/10.1110/ps.03323604
37. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7):951–960. https://doi.org/10.1093/bioinformatics/bti125
38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
39. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C (2011) OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res 39(Database):D289–D294. https://doi.org/10.1093/nar/gkq1238

Chapter 29

Modeling Structure and Dynamics of Protein Complexes with SAXS Profiles

Dina Schneidman-Duhovny and Michal Hammel

Abstract

Small-angle X-ray scattering (SAXS) is an increasingly common and useful technique for structural characterization of molecules in solution. A SAXS experiment determines the scattering intensity of a molecule as a function of spatial frequency, termed SAXS profile. SAXS profiles can be utilized in a variety of molecular modeling applications, such as comparing solution and crystal structures, structural characterization of flexible proteins, assembly of multi-protein complexes, and modeling of missing regions in the high-resolution structure. Here, we describe protocols for modeling atomic structures based on SAXS profiles. The first protocol is for comparing solution and crystal structures including modeling of missing regions and determination of the oligomeric state. The second protocol performs multi-state modeling by finding a set of conformations and their weights that fit the SAXS profile starting from a single-input structure. The third protocol is for protein-protein docking based on the SAXS profile of the complex. We describe the underlying software, followed by demonstrating their application on interleukin 33 (IL33) with its primary receptor ST2 and DNA ligase IV-XRCC4 complex.

Key words Small-angle X-ray scattering (SAXS), Protein-protein docking, Conformational heterogeneity, Multi-state models, Conformational ensembles

1 Introduction

SAXS has become a widely used technique for structural characterization of molecules in solution [1]. A key strength of the technique is that it provides information about conformational and compositional states of the system in solution. Moreover, SAXS profiles can be rapidly collected for a variety of experimental conditions, such as ligand-bound and unbound protein samples, ligand titration series, different temperatures, or pH values [2]. The experiment is performed with ~15 μl of the sample at the concentration of ~1.0 mg/ml. It usually takes only a few minutes on a well-equipped synchrotron beam line [1, 3]. The SAXS profile of a macromolecule, I(q), is computed by subtracting the SAXS profile of the buffer from the SAXS profile of the macromolecule in the

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764,
https://doi.org/10.1007/978-1-4939-7759-8_29,

buffer. The profile can be converted into an approximate distribution of pairwise atomic distances of the macromolecule (i.e., the pair distribution function) via a Fourier transform. The challenge lies in data interpretation since the profiles provide rotationally, conformationally, and compositionally averaged information about protein solution conformation(s).

Computational approaches for modeling a macromolecular structure based on its SAXS profile can be classified based on the system representation into ab initio and atomic resolution modeling methods [4, 5]. On the one hand, the ab initio methods search for coarse three-dimensional shapes represented by dummy atoms (beads) that fit the experimental profile [6–8]. On the other hand, atomic resolution modeling approaches generally rely on an all atom representation to search for models that fit the computed SAXS profile to the experimental one [9]. Therefore, atomic resolution modeling can be used only if an approximate structure or a comparative model of the studied molecule or its components is available. With the increasing number of structures in the Protein Data Bank (PDB) [10] that can serve as templates for comparative modeling of a large number of sequences [11], we have focused our own efforts on atomic resolution modeling with SAXS profiles [12–17].

SAXS-based atomic modeling can be used in a wide range of applications, such as comparing solution and crystal structures, modeling of a perturbed conformation (e.g., modeling active conformation starting from non-active conformation), structural characterization of flexible proteins, assembly of multi-domain proteins starting from single-domain structures, assembly of multi-protein complexes, fold recognition and comparative modeling, modeling of missing regions in the high-resolution structure, and determination of biologically relevant states from the crystal [18–20]. Several software packages and web servers are available for some of these tasks, including ATSAS [21] and pyDockSAXS [22, 23]. Here, we describe how our tools can be used to facilitate addressing several of these questions (Fig. 1). Specifically, we describe three protocols for modeling atomic structures based on SAXS profiles. The first protocol is for comparing solution and crystal structures including comparative modeling and modeling of missing regions and determination of the oligomeric state. The second protocol is for multi-state modeling (finding a set of conformations and their corresponding weights that fit the data) based on the SAXS profile and single-input structure. The third protocol is for protein-protein docking based on the SAXS profile of the complex. We describe the underlying software, followed by demonstrating their application on interleukin 33 (IL33) with its primary receptor ST2 (BIOISIS ST2ILP) [24] and DNA ligase IV-XRCC4 complex.

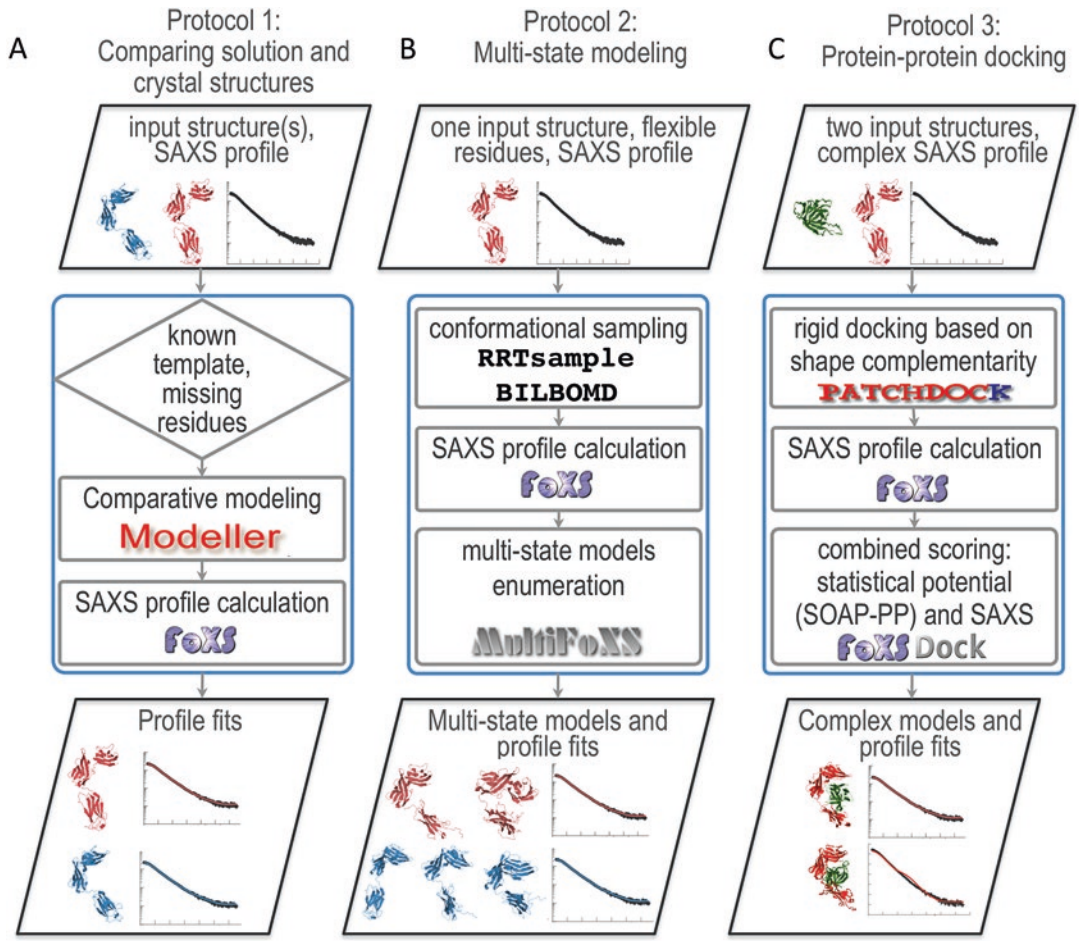


Fig. 1 Overview of the input and output of the three protocols: (**a**) comparing solution and crystal structures, (**b**) multi-state modeling, and (**c**) protein-protein docking

2 Materials

2.1 Software

The following software packages are used in the protocols described below:

1. Integrative Modeling Package (IMP)—a software package that includes SAXS module can be downloaded from http://salilab.org/imp/download.html and is available in binary form for most common machine types and operating systems; alternatively, it can be rebuilt from the source code; either the stable 2.7.0 release of IMP or a recent development version should be used. The code related to the protocols described here is mainly in the saxs, foxs, kinematics, multi_state, and integrative_docking IMP modules.

2. BILBOMD—a web server for multi-state modeling accessible from http://sibyls.als.lbl.gov/bilbomd.
3. PatchDock—a software for protein-protein docking can be downloaded from http://bioinfo3d.cs.tau.ac.il/PatchDock/.
4. MODELLER—a software for comparative modeling of protein structures can be downloaded from https://salilab.org/modeller/download_installation.html.
5. Gnuplot (http://www.gnuplot.info/) is used for plotting by the scripts provided with the examples used here.

The example files and scripts can be downloaded from https://modbase.compbio.ucsf.edu/foxs/mmb_files.zip.

3 Methods

3.1 Comparing Solution and Crystal Structures

Rapid and accurate computation of the SAXS profile of a given atomic structure and its comparison with the experimental profile is a basic component in any SAXS-based atomic modeling. FoXS is a program that is based on the IMP SAXS module that performs this task [13, 16]. The profiles are calculated using the Debye formula [25].

$$I(q) = \sum_{i=1}^{N}\sum_{j=1}^{N} f_i(q) f_j(q) \frac{\sin(q d_{ij})}{q d_{ij}} \quad (1)$$

where the intensity, $I(q)$, is a function of the momentum transfer $q = (4\pi \sin\theta)/\lambda$ and where 2θ is the scattering angle and λ is the wavelength of the incident X-ray beam; $f_i(q)$ is the atomic form factor, d_{ij} is the distance between atoms i and j, and N is the number of atoms in the molecule. In our model, the form factor $f_i(q)$ takes into account the displaced solvent as well as the hydration layer:

$$f_i(q) = f_v(q) - c_1 f_s(q) + c_2 s_i f_w(q) \quad (2)$$

where $f_v(q)$ is the atomic form factor in vacuo [26], $f_s(q)$ is the form factor of the dummy atom that represents the displaced solvent [27], s_i is the fraction of solvent accessible surface of the atom i [28], and $f_w(q)$ is the water form factor. The computed profile is fitted to the experimental data with adjustment of the excluded volume (c_1) and hydration layer density (c_2) parameters. The fit score is computed by minimizing the χ function with respect to c, c_1, and c_2:

$$\chi = \sqrt{\frac{1}{S}\sum_{i=1}^{s}\left(\frac{I_{\exp}(q_i) - cI(q_i, c_1, c_2)}{\sigma(q_i)}\right)^2} \quad (3)$$

where $I_{exp}(q)$ and $I(q)$ are the experimental and computed profiles, respectively, $\sigma(q)$ is the experimental error of the measured profile, S is the number of points in the profile, and c is the scale factor.

3.1.1 Inputs

The input to FoXS is one or more structure files in the PDB format and an experimental SAXS profile. The profile is specified in a text file with three columns: q in Å^{-1} units, intensity $I(q)$, and error $\sigma(q)$

```
# q            intensity     error
0.185480E-01  0.192175E+03  0.639769E+01
0.191560E-01  0.197885E+03  0.575226E+01
0.197640E-01  0.196492E+03  0.472259E+01
```

In addition, FoXS has several optional input parameters. Maximal q value determines the range for calculating the profile (default 0.5 Å^{-1}) and is controlled by –q option. The sampling resolution of the profile is controlled by the –s option that sets the number of points in the profile (default 500). The profile will be sampled at the resolution equal to the maximal q value divided by the number of profile points. For example, if the q_{max} value is 0.5 Å^{-1} and the user asks for 1000 profile points, the resulting profile will be uniformly sampled at the interval of 0.0005 Å^{-1}. The range of fit parameters (c_1 and c_2) is controlled by --min_c1 (default 0.99), --max_c1 (default 1.05), and --min_c2 (default −2.00), --max_c2 (default 4.00) options. By default hydrogen atoms are considered implicitly, unless -h option is specified. FoXS supports residue-level coarse graining by specifying -r option. This option is recommended only for very big structures where atomic resolution calculation is not feasible. It is also possible to adjust the background of the experimental profile (-b option, disabled by default) and use a constant in profile fitting (-o option, disabled by default). It is possible to write the profile to file before it is summed up using c_1 and c_2 parameters (-p option, disabled by default). This profile file will have six columns with different contributions to the intensity, and it is used in multi-state modeling by MultiFoXS (below). Another useful option is -m, which specifies how to read PDB files with multiple models. By default FoXS reads the first model only (-m 1). Alternatively, each model can be read into a separate structure (-m 2) or all models into a single structure (-m 3). If -g is specified, FoXS will print a script file for display of the fit file in Gnuplot.

3.1.2 Running FoXS

Here, we compare the SAXS profile of the ST2-IL33 complex [24] to the crystal structure (PDB 4kc3) using default program options:

```
> foxs 4kc3.pdb complex.dat
```

3.1.3 FoXS Output

The output is the values of χ, c_1, and c_2 for the resulting fit:

```
4kc3.pdb complex.dat Chi = 3.26 c1 = 1.04 c2 = 4.0
```

The program also outputs two files: the computed SAXS profile file (4kc3.pdb.dat) and the fit file between the computed profile

and the experimental one (4kc3_complex.dat). The format of the computed profile file is identical to the format of the input experimental file: three columns *(q, I(q), σ(q))*. The fit file contains four columns: *q*, experimental intensity, computed intensity, and the error of the experimental intensity:

```
#  q        exp_intensity    model_intensity error
0.01855     192.17500305     183.99343682     6.39769
0.01916     197.88499451     182.93727554     5.75226
0.01976     196.49200439     181.86241725     4.72259
```

The computed profile does not fit the experimental data within the noise (χ = 3.26). We hypothesized that the several loops and the C-terminal histidine tag that are unresolved in the crystal structure explain the difference (*see* **Note 1**).

3.1.4 Running MODELLER to Complete the Structure

We used MODELLER v9.8 [11] to add the missing fragments as follows. Two template structures were used: the ST2-IL33 complex (PDB 4kc3) and the IL33 structure (PDB 2kll). The additional IL33 structure was used since it does not have missing fragments. The corresponding MODELLER alignment file (fill.ali) and the script file (model_mult.py) are provided in the download zipfile. MODELLER v9.8 was run as follows:

```
> mod9.8 model_mult.py
```

After the models are generated, each candidate can be fitted to the experimental SAXS profile using FoXS (we repeat Part 2):

```
> foxs st2_il33.B999900*.pdb complex.dat
st2_il33.B99990001.pdb complex.dat Chi=1.88 c1=1.02 c2=4.0
st2_il33.B99990002.pdb complex.dat Chi=1.80 c1=1.03 c2=4.0
st2_il33.B99990003.pdb complex.dat Chi=1.61 c1=1.03 c2=4.0
st2_il33.B99990004.pdb complex.dat Chi=1.64 c1=1.03 c2=4.0
st2_il33.B99990005.pdb complex.dat Chi=1.53 c1=1.03 c2=4.0
st2_il33.B99990006.pdb complex.dat Chi=1.71 c1=1.03 c2=4.0
st2_il33.B99990007.pdb complex.dat Chi=1.85 c1=1.04 c2=3.4
st2_il33.B99990008.pdb complex.dat Chi=1.74 c1=1.03 c2=4.0
st2_il33.B99990009.pdb complex.dat Chi=1.83 c1=1.02 c2=4.0
st2_il33.B99990010.pdb complex.dat Chi=1.57 c1=1.03 c2=4.0
```

The resulting models have a significantly better fit than the crystal structure (1.5 < χ < 1.9), with the best χ value of 1.5 (Fig. 2a), which is within the experimental noise [29]. The fit plot along with the difference weighted by the error (Fig. 2a, **Note 2**) was generated from the fit files using plotFit.pl script (available in the zipfile scripts folder) that relies on Gnuplot:

```
> plotFit.pl 4kc3_complex.dat 2 x-ray
st2_il33.B99990005_complex.dat 3 model
```

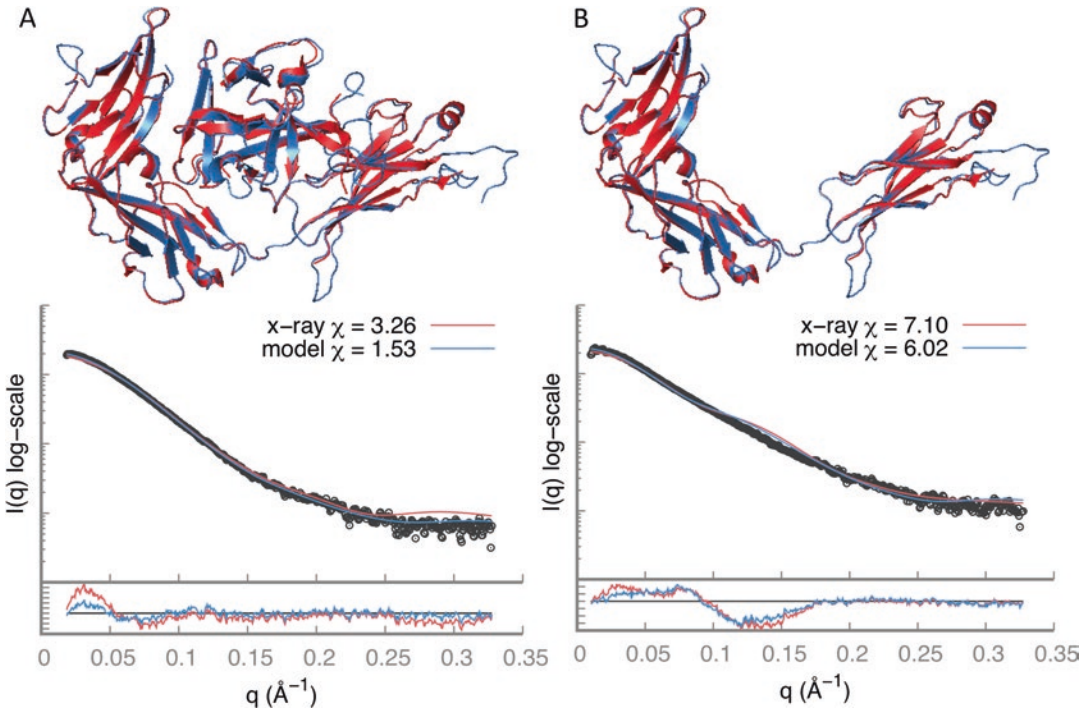


Fig. 2 Comparing solution and crystal structures: (**a**) ST2-IL33 complex and (**b**) ST2. The crystal structure is in red and the models with missing fragments added are in blue

The ST2 chain extracted from the crystal structure of the complex (PDB 4kc3, chain B) does not fit the experimental profile either (Fig. 2b, χ = 7.1). In this case, addition of missing atoms improved the fit only slightly (Fig. 2b, χ = 6.0), in contrast to the ST2-IL33 complex:

```
> foxs 4kc3B.pdb st2.pdb st2.dat
4kc3B.pdb st2.dat Chi = 7.1 c1 = 1.05 c2 = 4.0
st2.pdb st2.dat Chi = 6.0 c1 = 1.05 c2 = 4.0
```

We concluded that the ST2 solution structure is different from its crystal structure in complex with IL33. Therefore, we used multi-state modeling to sample the conformational space of ST2 and fit one or more conformations to the SAXS profile.

3.2 Multi-state Modeling

Multi-state modeling protocol addresses conformational heterogeneity in solution by relying on a SAXS profile. The input is a single atomic structure (or a comparative model), a list of flexible residues, and a SAXS profile for the protein. The protocol proceeds in three stages (Fig. 1b).

In the first stage, the conformations of the input structure are generated. We provide two methods for conformational sampling: RRTsample and BILBOMD. RRTsample explores the space of the

φ and ψ main-chain dihedral angles of the user-defined flexible residues with a rapidly exploring random trees (RRTs) algorithm [30–33]. Since the sampling uses internal coordinates, the sampled structure cannot contain cycles and can work with linear or tree-like arrangements of rigid bodies. The RRT algorithm samples the conformational space by leveraging an iteratively constructed nearest-neighbor linked tree. This iterative strategy expands the tree toward unexplored regions of the conformational space and significantly improves the efficiency compared to random sampling. In contrast, BILBOMD works in Cartesian coordinates and is not limited to tree-like topologies of the input structure. In BILBOMD molecular dynamics (MD) simulation is used to explore the conformational space. A common strategy is to perform the MD simulation on the linkers between the domains at very high temperature, where the additional kinetic energy prevents the molecule from becoming trapped in a local minimum. The MD simulation or RRT-based sampling provides a pool of atomistic models for SAXS profile calculation and fitting to the experimental profile in the subsequent steps.

In the second stage, a SAXS profile is pre-calculated for each sampled conformation using FoXS. To avoid data overfitting, the method sets a single pair of free parameters (c_1 and c_2) for each multi-state model, rather than using a different pair for each conformation. Therefore, at this stage the different parts contributing to the profile intensity are pre-calculated without summing up using c_1 and c_2 parameters.

In the third stage, best-scoring multi-state models are enumerated using the multi-state scoring function and branch and bound combinatorial optimization. Given N input conformations and their computed SAXS profiles, we look for multi-state models (subsets of conformations and their weights) of size n ($n << N$), such that the corresponding sum of weighted SAXS profiles fits the experimental SAXS profile. The score of a multi-state model is:

$$\chi = \sqrt{\frac{1}{S}\sum_{i=1}^{s}\left(\frac{I_{\exp}(q_i) - c\sum_{n} w_n I_n(q_i, c_1, c_2)}{\sigma(q_i)}\right)^2} \qquad (4)$$

where $I_n(q, c_1, c_2)$ and w_n are the computed profile and the corresponding weight, respectively, for each of the N states in the model; this equation minimizes data overfitting by using a single set of c_1 and c_2 values for all N states. In each "branch" step, we extend K ($K = 10{,}000$) best-scoring models of size n to KN models of size $n + 1$ by addition of each of the N input conformations. In the "bound" step, we select K best-scoring models out of the total KN models for the next iteration. Therefore, generation of K multi-state models of size $n + 1$ from K multi-state models of size n requires KN SAXS score calculations. This greedy approach avoids

the exponential growth in scale of enumeration while still hopefully producing the best-scoring multi-state models.

This protocol is modular and can work with a different method for generating conformations, such as normal mode analysis [34, 35] and KGSrna for RNA molecules [36, 37]. We provide two examples for multi-state modeling protocol: ST2 models are generated with RRT-based sampling using IMP software, and human DNA ligase IV-XRCC4 complex models are generated by BILBOMD web server.

3.2.1 ST2 Multi-state Modeling with IMP

Inputs

The input to the multi-state modeling protocol is a structure file in the PDB format, a text file with the list of flexible residues, and an experimental SAXS profile. Flexible residues list specifies to the RRT-based sampling program which φ and ψ angle to sample. Those residues divide the input protein into rigid bodies and linkers. This list should contain linkers or hinge regions between the rigid protein domains. HingeProt [38] can be used to identify hinges automatically. Flexible loops should not be specified, as the program cannot handle cycles, the current implementation is limited to linear or tree-like topologies. The flexible residues file contains one residue per line, specified as residue index in the PDB file and chain identifier.

ST2 consists of three immunoglobulin-like domains (D1–D3). Based on the previous studies, we defined the linker between the D2 and D3 domains as flexible, as we did the C-terminal histidine tag (residues 203–208 and 318–327, respectively). The flexible residues (Fig. 3a) are defined using hinges.dat file:

```
203 B
204 B
...
208 B
318 B
...
327 B
```

The whole protocol can be ran using runMultiFoXS.pl script as follows:

```
> runMultiFoXS.pl st2.pdb hinges.dat st2.dat -
```

Below we provide a step-by-step instruction with the goal to explain to the advanced user the various program options of each step.

Running RRT-Based Sampling

We run the conformational sampling program with the input PDB file and flexible residues file as follows:

```
> rrt_sample st2.pdb st2.dat –i 100000 –n 10000
```

The program continues to run until it performs the specified number of iterations (-i, default 100) or until it generates the specified number of conformations (-n, default 100). Here, we ask to

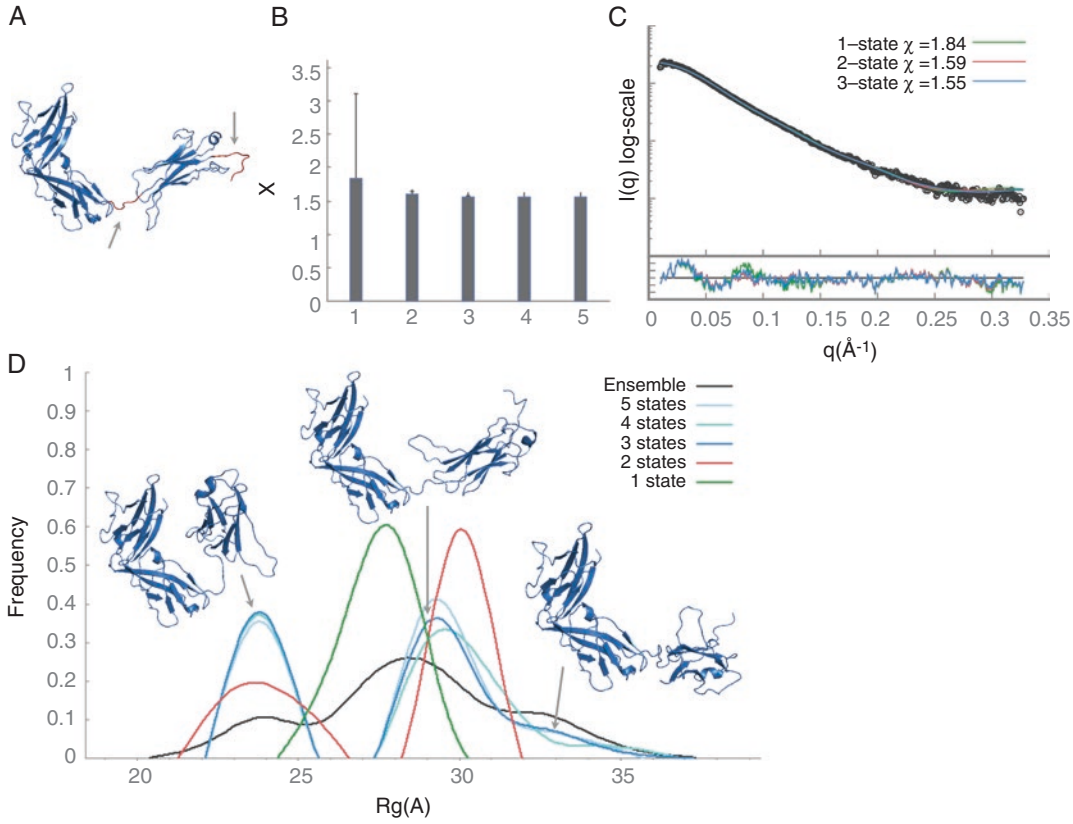


Fig. 3 ST2 multi-state modeling with IMP. (**a**) Flexible residues that were sampled by RRTsample are colored red. (**b**) The lowest χ value for N-state models ($N = 1...5$). (**c**) Fits between the experimental profile (black) and the best-scoring one-, two-, and three-state models (green, red, and blue, respectively). (**d**) R_g distribution of the best-scoring multi-state models

generate 10,000 conformations (*see* **Note 3**). The program has several optional parameters. When a new node is added to the tree, a collision-free path is generated between the closest tree node and the new node by a linear interpolation between the sampled angles of the two nodes. The conformations of the path are very close to each other; as a result the program saves every tenth conformation by default. This number can be controlled with -p option. When the number of the sampled rotatable angles (degrees of freedom) is high (>30), it might be hard to find moves that allow changing all the degrees of freedom at once. Therefore, the program supports random selection of a smaller number of degrees of freedom to sample in each iteration (-a, default 0 -all degrees of freedom are sampled). When there are more than 15 flexible residues, it is recommended to set this number to 10. The radii scaling parameter is controlled by -s option ($0.5 < s < 1.0$). The sampling can start only from collision-free conformation. If decreasing the scaling parameter does not help, the structure has to be minimized to

remove steric clashes (*see* **Note 4**). When sampling multi-chain structure, we often want to maintain the relative position of specific domains from different peptide chains as they are in the input structure by connecting them into a single rigid body. For example, this option is useful when a protein is a dimer where each monomer consists of two domains connected by a flexible linker and the first domain is the one involved in the dimerization, such as ATG7 (PDB 3vh1). In the ATG7 case, we want to maintain the dimerization interface intact and move only the *N*-terminal domains of each monomer. This is supported by -c option that receives a text file with a pair of residues one from each dimerization domain:

```
326 A 513 B
```

This will link two rigid bodies into a single one: the rigid body that residue 326 (chain A) belongs to and the rigid body that residue 513 (chain B) belongs to. Definition of two flexible linkers (between the N-terminal and C-terminal domains of each monomer) and one bridging region (for the dimerization domains) will result in three rigid bodies connected by two linkers (Fig. 4a). The same option also enables to maintain ligands position with respect to a protein by specifying an atom number from a ligand and from a protein. For example, to sample the structure of the calmodulin protein (PDB 1cll) with the calcium atoms, we will connect the four calcium atoms as follows:

1135 166
1136 420
1137 724
1138 1019

where 1135–1138 are the indexes of calcium atoms in the PDB file and 166, 420, 724, and 1019 are the indexes of the oxygen OD1 atoms in the aspartate residues that are closest to the calcium atom (Fig. 4b).

The rrt_sample program writes the conformations into PDB files named nodesX.pdb. By default 100 conformations are written to each PDB file using MODEL/ENDMDL to separate between the conformations. This number can be modified by -m option.

Running SAXS Profile Calculation

Here, we run FoXS to pre-calculate the profiles as explained above (-p option). Since the sampled conformations are models in the PDB format files, we use -m 2 option to read each model into a separate structure:

```
> foxs -m 2 -p nodes[1-9].pdb nodes10.pdb
```

This will pre-calculate SAXS profiles for the first 1000 conformations and write them into nodesX_mY.pdb.dat files, where X is the number of the original PDB file and Y is the model number in this file (*see* **Note 5**).

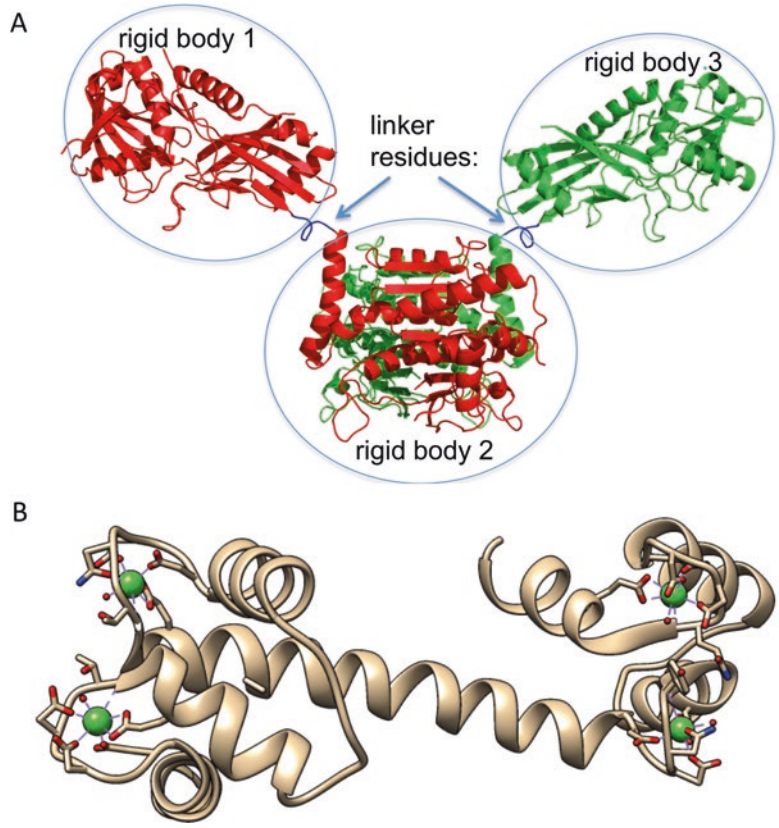


Fig. 4 Defining rigid bodies for RRTsample. (**a**) Connecting two domains from two chains into a single rigid body (PDB 3vh1). After the lower domains are connected (rigid body 2), we obtain a linear topology with three rigid bodies connected by two linkers (blue). (**b**) The calcium atoms (green) in the calmodulin (ODB 1cll) are linked to the protein by creating a connection with one of the oxygen atom of the aspartate

Running Multi-state Models Enumeration

We prepare the experimental SAXS profile file (st2.dat) and a file with the names of pre-computed profiles from the previous step:

```
> ls nodes*.pdb.dat > filenames
```

and run multi-state enumeration as follows:

```
> multi_foxs st2.dat filenames --max_c2 4.0
```

There are several optional parameters accepted by the program. The maximal number of states is set with -s option (default 10). The number of good-scoring multi-state models retained in each "bound" step is set with -k option (default 1000). It is recommended to increase this number to 10,000 for a large number of input profiles (>10,000). The minimal weight for a conformation to be included in the multi-state model is set with -w option (default 5%). Prior to the enumeration of multi-state models, the program clusters the profiles based on similarity as measured by the χ score. The clustering threshold is set with -t option (default 0.3), and it is

defined as the percentage of the χ score of the best-scoring conformation. For example, if the best-scoring conformation has a χ score of 2.0 when compared to the experimental profile, the default clustering threshold will be 0.6. We use the error bars from the experimental profile to ensure that the χ values are comparable. The maximal q value to consider in the experimental profile is set by -q option. The range of fit parameters (c_1 and c_2) is controlled by the same options as in FoXS, --min_c1 (default 0.99), --max_c1 (default 1.05), --min_c2 (default −0.5), and --max_c2 (default 2.00) with smaller range for c_2 parameter to avoid data overfitting. In the ST2 example, we increased the default c_2 range because we obtained c_2 value of 4.0 in the fit of the complex structure (Subheading 3.1.4).

Output Analysis

The generated ensembles of multi-state models are written into ensembles_size_X.txt files (X stands for the number of states) that are formatted as follows:

```
==> ensembles_size_1.txt <==
1 |  1.84 | x1 1.84 (1.05, 4.00)
    0   | 1.00 (1.00, 1.00) | nodes80_m93.pdb.dat (0.001)
2 |  2.08 | x1 2.08 (1.05, 4.00)
    1   | 1.00 (1.00, 1.00) | nodes63_m14.pdb.dat (0.001)
3 |  2.19 | x1 2.19 (1.05, 4.00)
    2   | 1.00 (1.00, 1.00) | nodes72_m54.pdb.dat (0.001)
==> ensembles_size_2.txt <==
1 |  1.59 | x1 1.59 (1.05, 4.00)
  212   | 0.73 (0.69, 0.08) | nodes81_m36.pdb.dat (0.053)
 1229   | 0.27 (0.29, 0.06) | nodes8_m76.pdb.dat (0.016)
2 |  1.59 | x1 1.59 (1.05, 4.00)
  212   | 0.71 (0.69, 0.08) | nodes81_m36.pdb.dat (0.053)
 1112   | 0.29 (0.34, 0.05) | nodes9_m78.pdb.dat (0.015)
==> ensembles_size_3.txt <==
1 |  1.55 | x1 1.55 (1.05, 4.00)
  637   | 0.49 (0.47, 0.08) | nodes43_m93.pdb.dat (0.417)
 1270   | 0.36 (0.36, 0.04) | nodes98_m82.pdb.dat (0.399)
 1541   | 0.15 (0.16, 0.04) | nodes40_m34.pdb.dat (0.016)
...
```

The first line is a summary of scores and fit parameters for a multi-state model: the first column is a number/rank of the multi-state model (sorted by score), a χ value for the fit to SAXS profile, and a pair of c_1 and c_2 values (in brackets) that optimize the fit to data are in the third column. In the ST2 example above, the χ values of the best-scoring one-, two-, and three-state models are 1.84, 1.59, and 1.55, respectively. After the model summary line, the file contains information about the states (one line per state). For example, the best-scoring two-state model consists of confor-

mation numbers 212 and 1229, with the weights of 0.73 and 0.27, respectively. The first conformation is model 36 in the nodes81.pdb file, and the second conformation is model 76 in the nodes8.pdb file. The numbers in brackets after the conformation weight are an average and a standard deviation of the weight calculated for this conformation across all good-scoring multi-state models of this size. The number in brackets after the filename is the fraction of good-scoring multi-state models that contain this conformation.

The program also outputs fit files (multi_state_model_X_Y_1.dat, where X is the number of states and Y is the number/rank of the multi-state model) between the weighted sum of profiles of the multi-state models and the experimental SAXS profile for the ten best-scoring models. The fit file is the same as in FoXS and contains four columns: q, experimental intensity, computed intensity, and the error of the experimental intensity.

Selecting the Representative Multi-state Models

Usually, the best explanation of the data is obtained by minimizing the number of conformations that resulted in the data (Occam's razor principle). Therefore, we are looking at the minimal number of states that enables fitting the data within the noise. However, the program usually produces a large ensemble of multi-state models with the same number of states and equally good scores. The conformations belonging to these multi-state models are generally neither accurate nor precise, but they provide a general shape for representative states. We describe these conformations using more robust structural features, such as radius of gyration (R_g) and maximal distance (D_{max}). Next, we analyze the distribution of R_g values for the ensemble of good-scoring multi-state models to estimate the number of possible states [39]. The number of peaks in the R_g distribution is a lower-bound estimate on the number of states, and the width of the peak is indicative of the state precision [40].

We generate the R_g distribution as follows. First, we calculate the radius of gyration for all the sampled conformations:

```
> runRg.pl nodes?.pdb nodes??.pdb nodes???.pdb
```

Second, we generate the distribution for best-scoring N-state models ($N = 1..0.5$) using plotHistograms.pl script:

```
> plotHistograms.pl 5 100 1.75
```

where 5 is the maximal number of states to consider, 100 is an ensemble size (the number of top-scoring multi-state models to analyze for each number of states), and 1.75 Å is a bin size in the R_g distribution. The output of the script is the R_g distribution plot (hist.png, Fig. 3d) and the χ values plot (chis.png, Fig. 3b). The χ values plot displays the χ values for the best-scoring N-state model ($N = 1..0.5$), where the error bar indicates the range of χ values for the top 100 multi-state models. We can use plotFit.pl script to generate the fit plot for

the top-scoring one-, two-, and three-state model as follows (the output is written to multi_state_model_1_1_1.eps file):

```
> plotFit.pl multi_state_model_1_1_1.dat 1 1-state
multi_state_model_2_1_1.dat 2 2-state multi_state_mod-
el_3_1_1.dat 3 3-state
```

For ST2 the χ value improved significantly even for a single-state model (χ = 1.8, Fig. 3b, c) compared to the crystal structure (χ = 6.0, Fig. 2b). The fit is even better with two- or three-state models (χ = 1.6), as expected. To estimate the number of states in solution, we examined the R_g distribution (Fig. 3d). The R_g distribution in the initial pool of 10,000 conformations is almost uniform (black line). The top-scoring one-state models (green line) have R_g in the range of 25–30 Å. The R_g distribution of the two-state models (red line) has two peaks: one corresponding to closed conformations at 21–26.5 Å and the other corresponding to open conformations at 28–32 Å. For three-state models (blue line), the R_g distribution has three peaks: the first peak at 22–25.5 Å overlaps with the closed conformations peak of two-state models, the second peak at 27–31 Å corresponds to open conformations that are similar to the IL33 binding conformation in the crystal structure, and the third low-frequency peak at 32–36 Å represents structures that are more open than the crystal structure. For models with four or five states, we observe three peaks overlapping with the peaks of the three-state models. Therefore, based on multi-state modeling results, we can conclude that ST2 exists in multiple states in solution, corresponding to a wide range of open and closed conformations (*see* **Note 6**). Upon IL33 binding, there is a population shift to the IL33 binding conformation.

3.2.2 DNA Ligase IV-XRCC4 Multi-state Modeling with BILBOMD

BILBOMD is a stand-alone web server that performs all the multi-state modeling stages: conformational sampling, SAXS profile calculation, and multi-state models enumeration (Fig. 5). The conformational sampling is based on the minimal molecular dynamics (MD) simulation using CHARMM version 40b [41]. SAXS profile calculation and enumeration of multi-state models use FoXS [16] and MultiFoXS [17] programs, respectively. The entire protocol is fully automated and does not require user interaction.

Inputs

BILBOMD web server accepts the following inputs (Fig. 5a):

1. Protein initial model in the PDB format where each peptide chain is uploaded as a separate segment/file.
2. A text file const.inp that defines rigid bodies and the connections of segments to maintain complex architecture.
3. Experimental SAXS profile file in a three-column format as in FoXS.

4. R_g min and R_g max values in Å for restraining the movement extent in the conformational sampling. A maximum of ten parallel simulations will be started in the range defined by these values.
5. Extent of conformational sampling that determines the number of conformers that will be generated for each R_g value.
6. An email address the results will be sent to.

We used BILBOMD web server to model solution state of the multi-protein complex human DNA ligase IV-XRCC4 (LigIV) [42]. LigIV is composed of N-terminal DNA-binding and catalytic domains connected by a long linker (50 residues) to the C-terminal tandem BRCT domain that interacts directly with the coiled-coil stalk domain of XRCC4 [43]. XRCC4 also contains long unfolded C-terminal regions (~120 residues) (Fig. 5c). The initial atomistic model was built by comparative modeling of the LigIV N-terminal consisting of DNA-binding domain (DBD), the nucleotidyltransferase domain (NTD), and OB-fold domain (OBD). Structures of the homologous human DNA ligases I [44] (PDB 1x9n) and III [45] (PDB 3l2p) were used as templates. The N-terminal domains model was connected to the crystal structure of XRCC4-BRCT [43] (PDB 3ii6) via addition of the flexible linker regions in between the catalytic core domains of LigIV and the tandem BRCT domain by MODELLER v9.8 [11]. MODELLER was also used to add the partially unfolded C-terminal regions of XRCC4.

BILBOMD input PDB files are the individual peptide chains of the complex corresponding to three segments. In our case segments 1 and 2 are the two chains of the XRCC4 dimer. Segment 3 is the DNA ligase IV chain containing DBD, NTD, OBD, and BRCT domains. Definition of rigid bodies is required for conformational sampling and can be done by clicking on "Create const. inp File" button on the BILBOMD input page (Fig. 5a). Rigid bodies are defined by selecting a relevant residue range in each segment (Fig. 5b). Residues that do not belong to rigid bodies are defined as the flexible regions for the MD simulations. In the XRCC4-BRCT complex, we would like to maintain the XRCC4 dimer interface and the XRCC4-BRCT interface (the XRCC4-BRCT crystal structure [43], PDB 3ii6) as one rigid body during MD simulation. Therefore, we group residues 1–210 from segments 1 and 2 (the XRCC4 dimer) and residues 611–833 from segment 3 (BRCT domain) into a single rigid body (Fig. 5b, domain 1 box). We also define rigid bodies for DBD, NTD, and OBD domains (Fig. 5b, domains 2–4) and a few shorter fragments in the XRCC4 dimer (Fig. 5b, domains 5–8). The server automatically creates a visualization of the rigid bodies and flexible regions. The rigid bodies are displayed as circles with the circle size proportional to the number of residues, while the flexible regions are

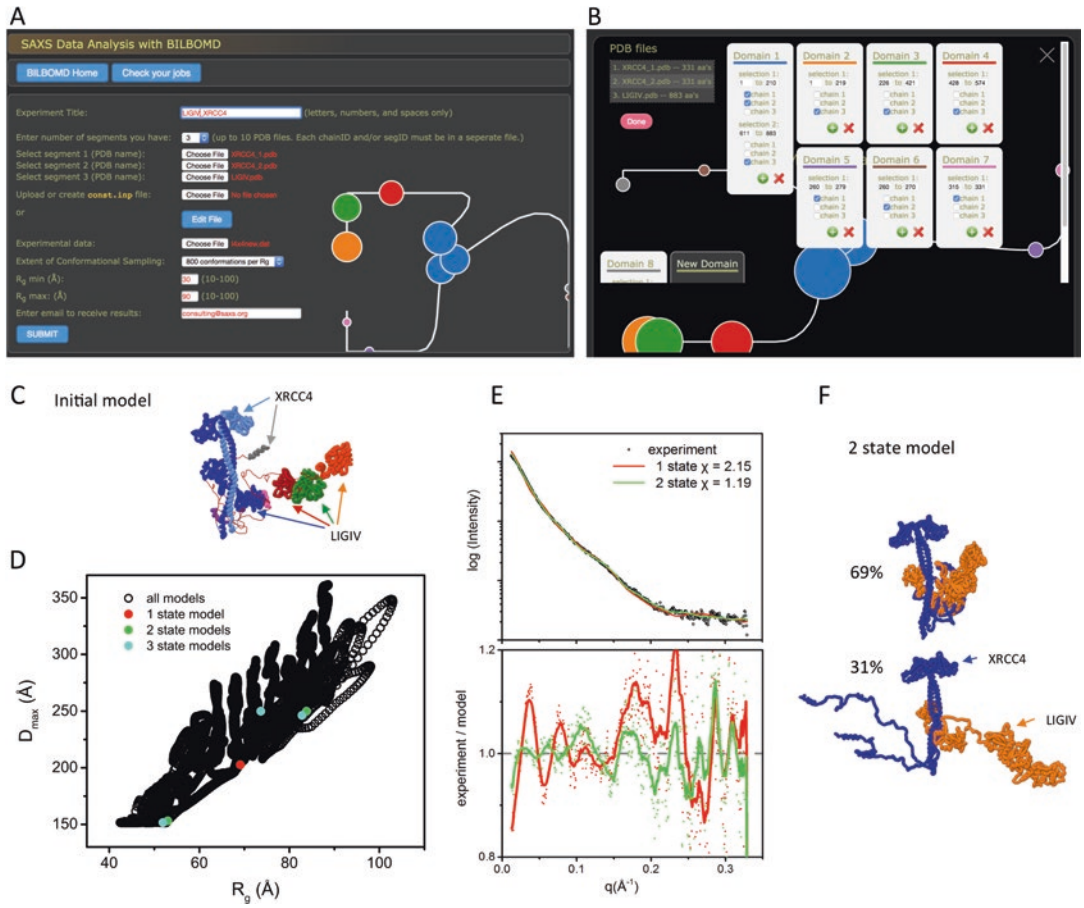


Fig. 5 DNA ligase IV-XRCC4 multi-state modeling with BILBOMD. (**a**) Web server input page. (**b**) Screenshot of the server interface for rigid bodies definitions ("Create const.inp File" option) for DNA ligase IV-XRCC4 complex, including visualization with circles and lines. (**c**) The initial model colored as the domain selections in the panel B. Flexible linkers are colored red. (**d**) R_g vs. D_{max} plot derived from foxs_rg.out output file with values for top-scoring one-state, two-state, and three-state models. (**e**) Fits between the experimental profile (black) and the best-scoring one- and two-state models (red and green, respectively). (**f**) Conformations of the top-scoring two-state model and their corresponding weights

shown as lines that connect to the circles (Fig. 5a, b). The definitions are written to the const.inp file:

```
define fixed1 sele (resid 1:210 .and. segid 1) end
define fixed2 sele (resid 1:210 .and. segid 2) end
define fixed3 sele (resid 611:883 .and. segid 3) end
cons fix sele fixed1 .or. fixed2 .or. fixed3 end
shape desc dock1 rigid sele (resid 1:219 .and. segid 3) end
shape desc dock2 rigid sele (resid 226:421 .and. segid 3) end
shape desc dock3 rigid sele (resid 428:574 .and. segid 3) end
shape desc dock4 rigid sele (resid 260:279 .and. segid 1) end
```

```
shape desc dock5 rigid sele (resid 260:279 .and. segid 2) end
shape desc dock6 rigid sele (resid 315:331 .and. segid 1) end
shape desc dock7 rigid sele (resid 315:331 .and. segid 2) end
return
```

where the fixed1, fixed2, and fixed3 define rigid domains of XRCC4 dimer (segid 1 and 2) and BRCT domain in LigIV (segid 3). These domains are connected into a single rigid body by cons fix command for maintaining their position during MD simulation. dock1-3 define the three rigid bodies of DBD, NTD, and OBD in the DNA ligase IV (segid 3). dock4-7 define small rigid regions in XRCC4 C-terminus (segid 1 and 2) (Fig. 5b, c). The user can also upload a revised const.inp file from a previous BILBOMD run.

Once the rigid bodies are defined, we submit additional inputs required by the server: the experimental SAXS profile, R_g min = 30 Å and R_g max = 90 Å values, the extent of the conformational sampling (800 conformations per R_g), and an email address (Fig. 5a).

BILBOMD Run

In the first phase, BILBOMD performs minimization of the entire model to eliminate steric clashes and optimize bond length and angles. This minimization enables to upload input structures with imperfect chain connectivity that may be present due to manual modeling of loops or linkers. In the second phase, the linkers connecting the defined rigid bodies are heated up to 1500 K. In the production phase, a maximum of ten parallel MD simulations are initiated at the various R_g increments within the R_g min and R_g max range (see movie in the ligase folder). One conformer for each simulation (corresponding to a specific R_g) is recorded every 0.5 ps. The length of the simulations is determined by the "Extend of conformational sampling" input parameter with the option to record up to 800 conformers (400 ps) per simulation. The final trajectory files are split into PDB format files with one conformer, out_X_Y_Z.pdb, where X corresponds to the R_g value, Y is a simulation step, and Z is the time in simulation. Next, BILBOMD precomputes SAXS profiles for all the conformers using FoXS and enumerates multi-state models using MultiFoXS (see Subheading 3.2.1 on ST2 multi-state modeling, Parts 3–5).

BILBOMD Outputs

Top-scoring multi-state models for 1–5 states are delivered by email. Additional output includes foxs_rg.out file with the list of R_g and maximal dimension (D_{max}) values for all generated models. Plotting R_g vs. D_{max} values and its comparison to the selected models enables additional visualization and validation of the conformational space of the multi-state models (Fig. 5d).

In the LigIV example, the χ values of the best-scoring one-, two-, and three-state models are 2.15, 1.19, and 1.05, respectively. The best-scoring two-state model consists of conformer 30_1_24500 and 90_4_398500, with the weights of 0.61 and 0.39, respectively. This model has a significantly better fit to the

experimental data in comparison to the best-scoring one-state model (Fig. 5e). These two states are derived from the simulation restrained by R_g values of 30 Å and 90 Å and were recorded at the time step 24,500 and 398,500, respectively (Fig. 5f).

3.3 Protein-Protein Docking

While many structures of single-protein components are increasingly available, structural characterization of their complexes remains challenging. Methods for modeling assembly structures from individual components frequently suffer from large errors, due to protein flexibility and inaccurate scoring functions. SAXS profile of the complex can significantly improve the success rate of protein-protein docking [14, 15]. The input to protein-protein docking protocol is the structures of the docked proteins in the PDB format and a SAXS profile of their complex. The protocol proceeds in three stages (Fig. 1c).

In the first stage, the proteins are docked using PatchDock, which is an efficient rigid-docking method that maximizes geometric shape complementarity [46, 47]. Protein flexibility is accounted for by a geometric shape complementarity scoring function, which allows for a small amount of steric clashes at the interface. The configurational sampling precision can be controlled by the resolution of the surface representation (i.e., the minimal distance between surface points used to generate docking models) and clustering parameters (*see* **Note 7**).

In the second stage, a SAXS profile is calculated for each docking model and is compared to the experimental SAXS profile using FoXS. It is possible that the complex sample in the SAXS experiment contained a mixture of monomers and complexes. Therefore, the SAXS scoring can optionally rely on a multi-state weighted scoring function (Eq. 4). This option is extremely useful for docking of transient complexes.

In the third stage, combined SAXS and statistical potential (SOAP-PP) [48] scores are calculated. To calculate the combined score, SAXS χ scores and statistical potential scores are normalized with respect to all the docking models. The combined score is the sum of the normalized Z-scores. The normalization of the scores allows us to avoid the use of weights for the terms of the combined score.

3.3.1 Inputs

The input to the protocol is two structure files in the PDB format and an experimental SAXS profile of the complex. Here, we will use the ST2 model with missing residues added (PDB 4kc3), the IL33 NMR structure (PDB 2kll), and the SAXS profile of the ST2-IL33 complex.

3.3.2 Docking

We can run all the steps of the protocol using idock IMP script as follows:

```
> idock st2.pdb 2kll.pdb --saxs complex.dat --patch_
dock patch_dock_path
```

The script accepts several optional parameters. The sampling precision of PatchDock (*see* **Note 7**) can be controlled by --precision option (1, normal; 2, medium; and 3, high precision). The usage of multi-state scoring function that accounts for monomer contributions is controlled by the --weighted_saxs_score option (default = False). There is a special parameter set for docking antibody-antigen and enzyme-inhibitor complexes (--complex_type AA or EI; *see* **Note 8**).

3.3.3 Results

The output file results_saxs_soap.txt is a list of complex models computed via rigid docking sorted by a combined SAXS and statistical potential scores:

```
# | Score |fil| ZScore| saxs | Zscore| soap     |
Zscore | Transformation
1 | -5.72 | + | -3.93 | 1.68 | -1.71 | -2765.39 |
-4.006 | 0.69 0.53 -2.03 45.73 -41.86 16.25
2 | -5.61 | + | -3.85 | 1.43 | -1.81 | -2616.54 |
-3.794 | 0.69 0.09 -2.00 43.54 -41.20 16.74
3 | -4.67 | + | -3.21 | 1.55 | -1.76 | -1997.67 |
-2.910 | -0.98 -0.76 3.10 55.82 -33.84 -4.72
4 | -4.66 | + | -3.20 | 1.39 | -1.83 | -1942.27 |
-2.831 | 0.33 0.31 -1.77 47.94 -42.51 17.87
5 | -4.64 | + | -3.19 | 1.94 | -1.62 | -2076.15 |
-3.022 | 0.52 0.47 -2.09 48.75 -41.56 15.09
```

Each line corresponds to one docking model; the models are ranked by the total score, best first (second column). The individual SAXS and SOAP-PP score/z-score pairs are also shown (columns 5–6 and 7–8, for SAXS and SOAP-PP, respectively). The last column is a transformation (three rotation angles and a translation vector) that transforms the second protein relative to the first (the first molecule is kept fixed).

To generate the output PDB files, we use PathDock transOutput.pl script that takes as an input the output file and a range of docking models ranks, applies the transformation on the second molecule, and produces complex files. To generate top ten models, we run the script as follows:

```
> patch_dock_path/transOutput.pl results_saxs_soap.txt
1 10
```

The ST2-IL33 example illustrates the benefit of docking restrained by a SAXS profile: the model with the best combined SAXS and energy score has a relatively low interface-RMSD from the crystal structure of 3.5 Å (Fig. 6a), while the model ranked as top scoring by the SAXS score alone has a much larger interface-

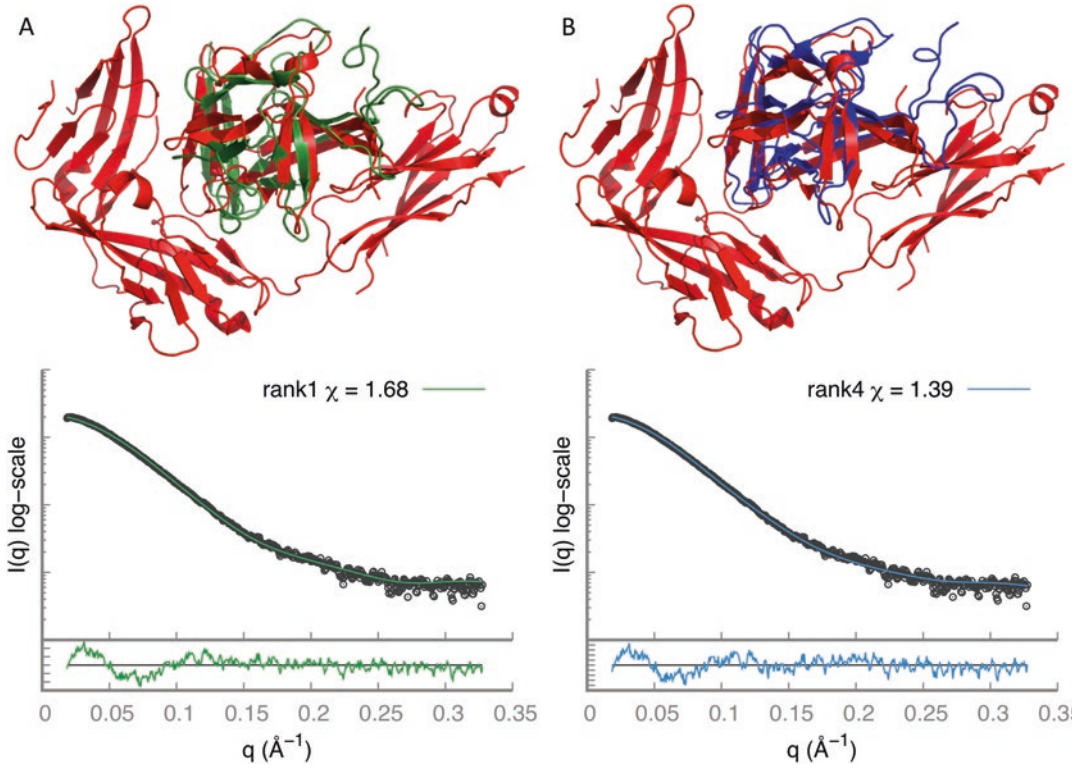


Fig. 6 Protein-protein docking. (**a**) Superposition (according to ST2) between the top-scoring model (green) and the crystal structure (red). (**b**) Superposition (according to ST2) between the top fourth scoring model (green) and the crystal structure (red). The fit between the model and the SAXS profile of the complex is below

RMSD of 18.2 Å. The model ranked as fourth has even higher accuracy with interface-RMSD of 1.7 Å (Fig. 6b).

4 Conclusion

The three protocols presented here facilitate the use of SAXS data in a variety of molecular modeling applications, such as comparing solution and crystal structures, structural characterization of flexible proteins, assembly of multi-protein complexes, and modeling of missing regions in the high-resolution structure. Atomic resolution representation of the modeled system provides strong constraints on possible solutions consistent with SAXS data, thus making SAXS-based modeling helpful for characterizing biomolecular systems. To maximize the accuracy of the predictions, the protocols rely on: (1) scoring functions for fitting multi-state mod-

els with single set of fitting parameters to reduce data overfitting, (2) efficient deterministic approach for enumeration of multiple states, and (3) advanced methods for exhaustive sampling of conformations and complexes. The SAXS-based modeling of the ST2-IL33 and DNA ligase IV-XRCC4 complexes provided an illustration of the protocols usage. The protocols are also available as web services [17] from http://salilab.org/foxs, http://salilab.org/multifoxs, http://salilab.org/foxsdock, and http://sibyls.als.lbl.gov/bilbomd.

5 Notes

1. Structures determined by X-ray crystallography are often missing coordinates for some of the residues, sugars, or His-tags. Since a SAXS profile is experimentally determined for the entire protein, the crystal structure will not perfectly fit the SAXS profile. To improve the SAXS fit, it is highly recommended to add all the missing fragments. Here, we did it using MODELLER.
2. SAXS intensity decreases rapidly and by orders of magnitude over the measured q-range, and depending upon how the data are presented, regions of significant misfit of the scattering profile may not be apparent. A straightforward and intuitive approach to visualizing the quality of a model fit over the entire measured q-range of a SAXS profile that takes into account relative errors is an error-weighted difference plot of $\frac{I_{\exp}(q) - cI_{\mathrm{mod}}(q)}{\sigma(q)}$ vs. q as shown in Fig. 2.
3. It is recommended to generate at least 10,000 conformations for an adequate coverage of the conformational space. Moreover, to test the sampling convergence, it is important to run the sampling protocol at least twice and validate that the results are similar.
4. The rrt_sample program needs to start from collision-free conformation. If the input conformation includes steric clashes, they can be resolved by MODELLER. Simply, run MODELLER using your input structure as a template.
5. The profile calculation for the sampled conformations can be trivially parallelized by running FoXS in parallel on different nodesX.pdb files.
6. The conformations generated by the multi-state modeling are generally neither accurate nor precise, but they provide representative states to the peaks in the R_g distribution. The wider the peak, the lower is the precision of the representative conformations from that peak.

7. SAXS profiles for two docking models with similar interface can differ significantly. This is because a small rotation relative to the interface can lead to significant change in the overall shape. Therefore, it is recommended to sample with higher sampling precision.
8. PatchDock has special protocols for enzyme-inhibitor and antibody-antigen complexes. For enzyme-inhibitor complexes, the docking search space is limited to the enzyme cavities. For antibody-antigen complexes, the docking search space is limited to the antibody complementarity-determining regions (CDRs) that are detected automatically.

Acknowledgments

We thank Drs. Andrej Sali, John Tainer, Ben Webb, David Agard, Friedrich Foerster, Seung Jong Kim, Hiro Tsuruta, Tsutomu Matsui, Lester Carter, Greg Hura, Riccardo Pellarin, Barak Raveh, Patrick Weinkam, and many others who contributed to our SAXS-based modeling efforts over the years. SAXS at the Advanced Light Source SIBYLS beamline in supported by National Institutes of Health (NIH) grants CA92584, DOE BER Integrated Diffraction Analysis Technologies (IDAT) program and NIGMS grant P30 GM124169-01, ALS-ENABLE.

References

1. Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL 2nd, Tsutakawa SE, Jenney FE Jr, Classen S, Frankel KA, Hopkins RC, Yang SJ, Scott JW, Dillard BD, Adams MW, Tainer JA (2009) Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat Methods 6(8):606–612. https://doi.org/10.1038/nmeth.1353. nmeth.1353 [pii]
2. Hura GL, Budworth H, Dyer KN, Rambo RP, Hammel M, McMurray CT, Tainer JA (2013) Comprehensive macromolecular conformations mapped by quantitative SAXS analyses. Nat Methods 10(6):453–454. https://doi.org/10.1038/nmeth.2453
3. Dyer KN, Hammel M, Rambo RP, Tsutakawa SE, Rodic I, Classen S, Tainer JA, Hura GL (2014) High-throughput SAXS for the characterization of biomolecules in solution: a practical approach. Methods Mol Biol 1091:245–258. https://doi.org/10.1007/978-1-62703-691-7_18
4. Putnam CD, Hammel M, Hura GL, Tainer JA (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys 40(3):191–285. https://doi.org/10.1017/S0033583507004635. S0033583507004635 [pii]
5. Rambo RP, Tainer JA (2013) Super-resolution in solution x-ray scattering and its applications to structural systems biology. Annu Rev Biophys 42:415–441. https://doi.org/10.1146/annurev-biophys-083012-130301
6. Chacon P, Moran F, Diaz JF, Pantos E, Andreu JM (1998) Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm. Biophys J 74(6):2760–2775. https://doi.org/10.1016/S0006-3495(98)77984-6. S0006-3495(98)77984-6 [pii]
7. Svergun DI (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys J 76(6):2879–2886. https://doi.org/10.1016/S0006-3495(99)77443-6. S0006-3495(99)77443-6 [pii]
8. Svergun DI, Petoukhov MV, Koch MH (2001) Determination of domain structure of proteins from X-ray solution scattering. Biophys J 80(6):2946–2953. https://doi.org/10.1016/S0006-3495(01)76260-1. S0006-3495(01)76260-1 [pii]

9. Petoukhov MV, Svergun DI (2005) Global rigid body modeling of macromolecular complexes against small-angle scattering data. Biophys J 89(2):1237–1250. https://doi.org/10.1529/biophysj.105.064154. S0006-3495(05)72771-5 [pii]
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
11. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815. https://doi.org/10.1006/jmbi.1993.1626. S0022-2836(83)71626-8 [pii]
12. Förster F, Webb B, Krukenberg KA, Tsuruta H, Agard DA, Sali A (2008) Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. J Mol Biol 382(4):1089–1106. https://doi.org/10.1016/j.jmb.2008.07.074. S0022-2836(08)00943-1 [pii]
13. Schneidman-Duhovny D, Hammel M, Sali A (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res 38(Suppl):W540–W544. https://doi.org/10.1093/nar/gkq461. gkq461 [pii]
14. Schneidman-Duhovny D, Hammel M, Sali A (2011) Macromolecular docking restrained by a small angle X-ray scattering profile. J Struct Biol 173(3):461–471. https://doi.org/10.1016/j.jsb.2010.09.023. S1047-8477(10)00292-3 [pii]
15. Schneidman-Duhovny D, Rossi A, Avila-Sakar A, Kim SJ, Velazquez-Muriel J, Strop P, Liang H, Krukenberg KA, Liao M, Kim HM, Sobhanifar S, Dotsch V, Rajpal A, Pons J, Agard DA, Cheng Y, Sali A (2012) A method for integrative structure determination of protein-protein complexes. Bioinformatics 28(24):3282–3289. https://doi.org/10.1093/bioinformatics/bts628
16. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2013) Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophys J 105(4):962–974. https://doi.org/10.1016/j.bpj.2013.07.020
17. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A (2016) FoXS, FoXSDock and MultiFoXS: single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. Nucleic Acids Res 44(W1):W424–W429. https://doi.org/10.1093/nar/gkw389
18. Hammel M (2012) Validation of macromolecular flexibility in solution by small-angle X-ray scattering (SAXS). Eur Biophys J 41(10):789–799. https://doi.org/10.1007/s00249-012-0820-x
19. Schneidman-Duhovny D, Kim SJ, Sali A (2012) Integrative structural modeling with small angle X-ray scattering profiles. BMC Struct Biol 12(1):17. https://doi.org/10.1186/1472-6807-12-17
20. Rambo RP, Tainer JA (2011) Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. Biopolymers 95(8):559–571. https://doi.org/10.1002/bip.21638
21. Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, Gajda M, Gorba C, Mertens HDT, Konarev PV, Svergun DI (2012) New developments in the ATSAS program package for small-angle scattering data analysis. J Appl Crystallogr 45(2):342–350. https://doi.org/10.1107/S0021889812007662
22. Pons C, D'Abramo M, Svergun DI, Orozco M, Bernado P, Fernandez-Recio J (2010) Structural characterization of protein-protein complexes by integrating computational docking with small-angle scattering data. J Mol Biol 403(2):217–230. https://doi.org/10.1016/j.jmb.2010.08.029. S0022-2836(10)00891-0 [pii]
23. Jimenez-Garcia B, Pons C, Svergun DI, Bernado P, Fernandez-Recio J (2015) pyDockSAXS: protein-protein complex structure by SAXS and computational docking. Nucleic Acids Res 43(W1):W356–W361. https://doi.org/10.1093/nar/gkv368
24. Liu X, Hammel M, He Y, Tainer JA, Jeng US, Zhang L, Wang S, Wang X (2013) Structural insights into the interaction of IL-33 with its receptors. Proc Natl Acad Sci U S A 110(37):14918–14923. https://doi.org/10.1073/pnas.1308651110
25. Debye P (1915) Zerstreuung von Röntgenstrahlen. Ann Phys 351(6):809–823
26. Svergun D, Barberato C, Koch MHJ (1995) CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 28(6):768–773
27. Fraser RDB, MacRae TP, Suzuki E (1978) An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules. J Appl Crystallogr 11(6):693–694
28. Connolly ML (1983) Solvent-accessible surfaces of proteins and nucleic acids. Science 221(4612):709–713
29. Rambo RP, Tainer JA (2013) Accurate assessment of mass, models and resolution by small-angle scattering. Nature 496(7446):477–481. https://doi.org/10.1038/nature12070
30. LaValle SM, Kuffner JJ (2001) Rapidly-exploring random trees: progress and pros-

pects. In: Algorithmic and computational robotics: New Directions, pp. 293–308
31. Amato NM, Song G (2002) Using motion planning to study protein folding pathways. J Comput Biol 9(2):149–168
32. Cortes J, Simeon T, Ruiz de Angulo V, Guieysse D, Remaud-Simeon M, Tran V (2005) A path planning approach for computing large-amplitude motions of flexible molecules. Bioinformatics 21(Suppl 1):i116–i125. https://doi.org/10.1093/bioinformatics/bti1017
33. Raveh B, London N, Schueler-Furman O (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins 78(9):2029–2040. https://doi.org/10.1002/prot.22716
34. Suhre K, Sanejouand YH (2004) On the potential of normal-mode analysis for solving difficult molecular-replacement problems. Acta Crystallogr D Biol Crystallogr 60:796
35. Ma JP (2005) Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure 13:373
36. Fonseca R, Pachov DV, Bernauer J, van den Bedem H (2014) Characterizing RNA ensembles from NMR data with kinematic models. Nucleic Acids Res 42(15):9562–9572. https://doi.org/10.1093/nar/gku707
37. Fonseca R, van den Bedem H, Bernauer J (2015) KGSrna: efficient 3D kinematics-based sampling for nucleic acids. In: Przytycka TM (ed) Research in computational molecular biology: 19th annual international conference, RECOMB 2015, Warsaw, Poland, April 12–15, 2015, Proceedings. Springer International Publishing, Cham, pp. 80–95. doi:https://doi.org/10.1007/978-3-319-16706-0_11
38. Emekli U, Schneidman-Duhovny D, Wolfson HJ, Nussinov R, Haliloglu T (2008) HingeProt: automated prediction of hinges in protein structures. Proteins 70(4):1219–1227. https://doi.org/10.1002/prot.21613
39. Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. J Am Chem Soc 129(17):5656–5664. https://doi.org/10.1021/ja069124n
40. Carter L, Kim SJ, Schneidman-Duhovny D, Stohr J, Poncet-Montange G, Weiss TM, Tsuruta H, Prusiner SB, Sali A (2015) Prion protein-antibody complexes characterized by chromatography-coupled small-angle X-ray scattering. Biophys J 109(4):793–805. https://doi.org/10.1016/j.bpj.2015.06.065
41. Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30(10):1545–1614. https://doi.org/10.1002/jcc.21287
42. Williams GJ, Hammel M, Radhakrishnan SK, Ramsden D, Lees-Miller SP, Tainer JA (2014) Structural insights into NHEJ: building up an integrated picture of the dynamic DSB repair super complex, one component and interaction at a time. DNA Repair (Amst) 17:110–120. https://doi.org/10.1016/j.dnarep.2014.02.009
43. Wu PY, Frit P, Meesala S, Dauvillier S, Modesti M, Andres SN, Huang Y, Sekiguchi J, Calsou P, Salles B, Junop MS (2009) Structural and functional interaction between the human DNA repair proteins DNA ligase IV and XRCC4. Mol Cell Biol 29(11):3163–3172. https://doi.org/10.1128/MCB.01895-08
44. Pascal JM, O'Brien PJ, Tomkinson AE, Ellenberger T (2004) Human DNA ligase I completely encircles and partially unwinds nicked DNA. Nature 432(7016):473–478. https://doi.org/10.1038/nature03082
45. Cotner-Gohara E, Kim IK, Hammel M, Tainer JA, Tomkinson AE, Ellenberger T (2010) Human DNA ligase III recognizes DNA ends by dynamic switching between two DNA-bound states. Biochemistry 49(29):6165–6176. https://doi.org/10.1021/bi100503w
46. Duhovny D, Nussinov R, Wolfson HJ (2002) Efficient unbound docking of rigid molecules. In: Guigó R, Gusfield D (eds) Second International Workshop, WABI 2002, Rome, Italy. Lecture notes in computer science. Springer Berlin, Heidelberg, pp. 185–200. doi:https://doi.org/10.1007/3-540-45784-4
47. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 33(Web Server issue):W363–W367. https://doi.org/10.1093/nar/gki481. 33/suppl_2/W363 [pii]
48. Dong GQ, Fan H, Schneidman-Duhovny D, Webb B, Sali A (2013) Optimized atomic statistical potentials: assessment of protein interfaces and loops. Bioinformatics 29(24):3158–3166. https://doi.org/10.1093/bioinformatics/btt560

Chapter 30

Modeling the Structure of Helical Assemblies with Experimental Constraints in Rosetta

Ingemar André

Abstract

Determining high-resolution structures of proteins with helical symmetry can be challenging due to limitations in experimental data. In such instances, structure-based protein simulations driven by experimental data can provide a valuable approach for building models of helical assemblies. This chapter describes how the Rosetta macromolecular package can be used to model homomeric protein assemblies with helical symmetry in a range of modeling scenarios including energy refinement, symmetrical docking, comparative modeling, and de novo structure prediction. Data-guided structure modeling of helical assemblies with experimental information from electron density, X-ray fiber diffraction, solid-state NMR, and chemical cross-linking mass spectrometry is also described.

Key words Structure prediction, Structure determination, Helical symmetry, Helical assemblies, Fibrils, Fibers, Rosetta

1 Introduction

Polymers with helical symmetry are utilized in a wide variety of biological functions, such as information storage, structural support, and muscle movement [1]. Helical assembly formation is also associated with disease states of proteins, such as amyloid formation and hemoglobin polymerization in sickle-cell anemia. Detailed structural insight into the organization of helical assemblies at the atomic level is central to understanding their role in biological functions and disease. Unfortunately, helical symmetry is rarely compatible with formation of 3D crystals, which means that structure determination typically requires the use of methods that provide information at lower resolution. A further complication is that helical assemblies are often more structural heterogeneous than globular proteins. Due to these experimental challenges, computational structural modeling can be a critical tool in developing detailed structural models of helical assemblies. Such modeling is best carried out with constraints derived from experimental data

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764,
https://doi.org/10.1007/978-1-4939-7759-8_30,

from methods as cryo-electron microscopy, X-ray fiber diffraction, nuclear magnetic resonance, and cross-linking mass spectrometry. In this chapter, we present how homomeric assemblies with helical symmetry can be modeled at the atomic level using the Rosetta macromolecular modeling package [2] and how experimental data can be incorporated into the modeling approach. Rosetta provides a flexible software environment where a large variety of modeling tasks can be carried out including energy refinement, homology modeling, protein-protein docking, and de novo structure prediction [2] and applied to the modeling of arbitrarily complex symmetrical systems [3]. Our goal here is to demonstrate how such modeling task can be tailored toward the study of helical protein assemblies.

2 Materials

Instructions for downloading Rosetta can be found at www.rosettacommons.org. Compiling Rosetta requires the presence of a few different software packages including a C++ compiler and Python, which are typically found on a unix-type system. Perl and Python are also required to run some of the scripts for setting up the input files to Rosetta. Some of the modeling tasks described in this chapter can be run on a desktop computer, while many of the more computationally complex tasks require access to a computer cluster. Depending on the modeling scenario, additional sequence data, starting structural models, and experimental data may be required. Example files and instructions for simulating helical structures in Rosetta with X-ray fiber diffraction and solid-state NMR data can be found at http://www.cmps.lu.se/biostruct/people/ingemar-andre/fiber-diffraction/. These examples can readily be tailored toward modeling of other types of experimental data discussed in this chapter.

3 Methods

3.1 Defining Symmetry of the System

The symmetry operations that are available for helical symmetry [4] are (I) translation parallel to the helical axis (z axis), (II) n-fold rotation around the z axis, (III) screw displacement combining translation parallel to z together with a rotation around z, and (IV) a twofold rotation about a line passing through the z axis and perpendicular to it. On top of this, each structural unit repeating in the helix can have internal symmetry.

There are a number of possible descriptions of helical symmetry (Fig. 1). One such description is rise (**z**), helical rotation angle (**phi**), and radius (**r**). In X-ray fiber diffraction, the helical geometry is described by the following parameters: number of units (**u**) per

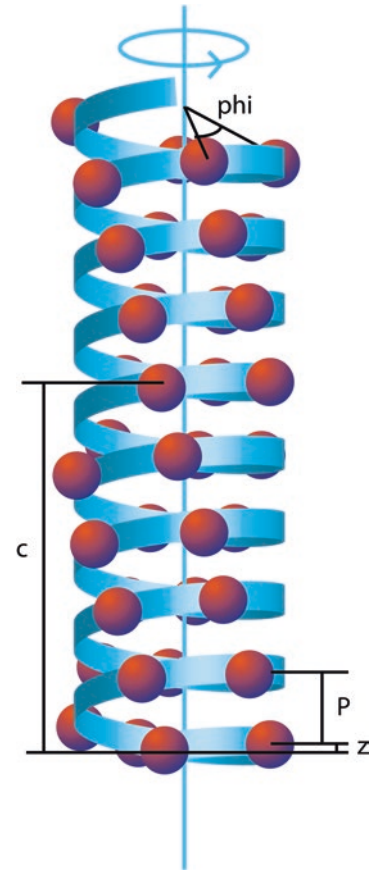


Fig. 1 Geometry of a system with helical symmetry. Red subunits are separated by the rise (**z**) and helical rotation angle (**phi**). Each turn the helix is translated with the value of the pitch (**P**). After a helical repeat distance (**c**), a subunit appears at the same location (identical x and y coordinates) in the helix

number of turns (**t**) over the repeat distance (**c**) giving rise to a pitch (**P**). For example, an alpha-helix has 18 residues per 5 turns, with **N** = 18/5 = 3.6 residues per turn. **phi** is 100°, the rise **z** 1.5 Å resulting in a pitch of **P** = 1.5×3.6 = 5.4 Å. Tobacco mosaic virus has 49 subunits over 3 turns with a helical repeat distance **c** of 68.83 Å and a measured pitch **P** of 22.92 Å [5]. Typically, the pitch and helical repeat distance are estimated from experimental data, and the number of subunits per turn is assigned subsequently. The following relations can be used to convert between the parameters:

$$\mathbf{N} = \mathbf{u}/\mathbf{t}$$

$$\mathbf{c} = \mathbf{u}\times\mathbf{z} = \mathbf{t}\times\mathbf{P}$$

Rosetta has the capability of modeling arbitrarily complex symmetries. A Rosetta protocol must be told which symmetry should be applied to the given system. This is done by providing a symmetry definition file. The details of the symmetry definition file

format are found in DiMaio et al. [3]. There are symmetries that have not been implemented in the auxiliary scripts described in the next section. The steps involved in creating custom-made symmetry definition files will not be described here but can be found in DiMaio et al. [3] and the Rosetta manual.

3.2 Defining Symmetry of the System for an Existing Helical System

When an existing helical structure or model is used in Rosetta, the first step is to analyze the symmetry of the system (*see* **Note 1**) with a script distributed with Rosetta (*Rosetta/main/source/src/apps/public/make_symmdef_file.pl*, *see* **Note 2**). The script is used to analyze the symmetry of a helical protein system and to generate a symmetry definition file. The input to the script is a PDB file of a helically symmetric segment, which two chains should be used to calculate the symmetry of the system and the number of subunits to explicitly model in Rosetta. For complex helical assemblies, each subunit can have many neighbors and a large number of subunits that may have to be modeled to represent the energetics of the complete helical system. The output of the script is a symmetry definition file, a PDB file with a single subunit that is used together with the symmetry definition file to generate the complete assembly inside Rosetta, and a reference PDB model representing the subunits modeled inside Rosetta.

3.3 Defining Symmetry of the System for De Novo Modeling

There are scenarios in which a preexisting helical model is not available and the helical parameters need to be determined during the simulation. Cases include global protein-protein docking and de novo folding simulations. A script is distributed with Rosetta (*Rosetta/tools/fiber_diffraction/make_helix_denovo.py* and *make_helix_denovo_cnsm.py*) for setting up symmetry definition files (*see* **Note 3**) for these modeling scenarios. The first can be used to model standard helical symmetry, while the second is employed for systems with both rotational and screw axis symmetries. In the scripts, helical symmetry is specified using the fiber diffraction convention, where the user is required to provide a helical rise (**z**) and number of subunits (**u**) and turns (**t**) over a helical repeat. The symmetry definition file can be manually altered to specify the starting position of subunits and whether rotational randomization should be done prior to simulation.

3.4 Sampling the Rigid Body Degrees of Freedom

A standard system with helical symmetry can be described by 6 rigid body degrees of freedoms (dofs): two translational dofs, the radius and the helical rise; one rotational degree of freedom controlling phi; and three rotational dofs controlling the rotation of each subunit around their center of mass. Which of these degrees of freedom that are allowed to vary during a simulation can be controlled by editing the symmetry definition file (*see* **Note 4**). It is also possible to control the translational and rotational step sizes during the simulation in the script and the ranges of values used for

parameter initialization at the beginning of the simulation. A standard heterodimeric protein-protein docking simulation involves the sampling of 6 degrees of freedom, indicating that global symmetrical docking with helical symmetry should also be feasible. Indeed, we have confirmed this in a benchmarking study [6]. Nonetheless, it is considerably more complex to efficiently explore the 6 dofs in helical symmetry than in standard heterodimeric protein-protein docking. It may be necessary to control which order the dofs are sampled (*see* **Note 5** and DiMaio et al. [3] for how to do this) and to tweak step sizes used during translational and rotational sampling. If experimental data is available defining the pitch and the number of subunits per turn, it is advisable to lock these values during the simulation (*see* **Note 5**).

3.5 Energy Refinement of a Helical Model

Once a helical model has been analyzed to extract symmetry information, the structure can be modeled through a host of Rosetta application. One central task is to refine the energy of the structure by allowing rigid body, sidechain, and backbone degrees of freedom to vary. This can be done with the stand-alone Rosetta relax application [7] or through calling this function through the Rosetta scripts framework [8] (*see* **Note 6**).

Energy refinement with the relax application in Rosetta is called by

```
relax.linuxgccrelease @relax_flags
```

where relax_flags contains options for the relax application. Alternatively, the same protocol can be called through the Rosetta script framework:

```
rosetta_scripts.linuxgccrelease @flags_relax -
parser:protocol relax.xml
```

where relax.xml is an xml script calling the relax protocol in Rosetta. This xml script has to be custom made by the user, but many examples exist in the Rosetta tutorials for how to generate these.

3.6 Homology Modeling of a Helical Fiber in the Context of Helical Symmetry

There are two alternative approaches to carry out comparative modeling of a helical protein structure. The first approach is to extract a monomer from a helical structure and use it as a template for comparative modeling. The symmetry definition file generated from the template helical structure can then be used to produce the helical homology model. This involves replacing the template master subunit with the monomeric model produced by comparative modeling and then applying the symmetry from the template symmetry definition file. The modeled subunit may have to be aligned with template monomer prior to using it in Rosetta for this to work. Alternatively, comparative modeling can be carried out directly in the context of the symmetrical helical model. The

comparative modeling framework, RosettaCM [9], is symmetry aware and can be instructed straightforwardly to model with helical symmetry (*see* **Note 6**).

Comparative modeling is run in Rosetta through the rosetta_scripts application:

```
rosetta_scripts.linuxgccrelease @flag_cm -
parser:protocol rosetta_cm.xml
```

where flag_cm contains additional options for the Rosetta run and rosetta_cm.xml is a xml script generated by a Python script distributed with Rosetta, *Rosetta/tools/protein_tools/setup_RosettaCM.py*, which prepares all input files and the xml file.

3.7 Docking of Subunits in Helical Symmetry

A protein subunit can be docked in helical symmetry using symmetrical protein docking in Rosetta [6]. There are several different scenarios. For instance, a protein model may be available, and a refinement of the rigid body position is desired. In this scenario, the symmetry of the helical assembly model is analyzed and used as a starting point for the docking refinement. Refinement may involve a coarse-grained phase ("docking perturbation") or just the all-atom refinement ("docking local refine") phase. Alternatively, if no starting helical model is available, global sampling of rigid body conformational space can be attempted ("global docking"). As mentioned in Subheading 3.4 and **Note 4**, which rigid body degrees of freedom that are sampled during a simulation can be controlled in the symmetry definition file.

Protein-protein docking can be run with the stand-alone SymDock application as

```
SymDock.linuxgccrelease @docking_flags
```

where docking_flags contains options for the docking run or through Rosetta scripts:

```
rosetta_scripts.linuxgccrelease @flags_dock -
parser:protocols symdock.xml
```

where symdock.xml is a xml script calling the symmetrical docking protocol in Rosetta. This xml script has to be custom made by the user, but many examples exist in the Rosetta tutorials for how to generate these.

3.8 De Novo Modeling

For systems with small subunit sizes (typically less than 80 residues) and where additional experimental data is available, complete de novo structure prediction of a helical assembly can be feasible. This involves simultaneous sampling of backbone, sidechain, and rigid body degrees of freedom with the fold-and-dock protocol [10, 11]. Following Subheading 3.4 and **Note 4**, the sampled rigid body degrees of freedom can be controlled in the symmetry definition file.

Fold and dock can be run through the stand-alone minirosetta application as

```
minirosetta.linuxgccrelease @fad_flags
```

where fad_flags contain the options for the run.

3.9 Modeling Helical Assemblies with Experimental Constraints

There are three main experimental methods used to gain insight into the atomic structures of helical assemblies: cryo-electron microscopy [12], solid-state nuclear magnetic resonance [13], and X-ray fiber diffraction [14]. Each of these methods can in favorable cases result in high-resolution data down to 3 Å. However, heterogeneous sample preparations and sparsity of experimental data often reduce the resolution of the data below the level required to identify atomic features. In these scenarios, the data alone is not sufficient to determine an atomic structure, and more emphasis has to be put on the structural modeling and the force field. A powerful approach to define the structure of these types of assemblies is to record a number of structural data of the system with different methods and combine these to build an atomic model with higher precision. Rosetta has the capabilities to simultaneously model structures using a multitude of experimental data. In this chapter, the use of data from electron density, solid-state NMR, X-ray fiber diffraction, and cross-linking mass spectrometry in Rosetta is described. Typically, experimental data is introduced into Rosetta as a pseudo-energy function where perfect agreement with data results in a zero energy contribution and any deviation is penalized. Alternatively, experimental data can be converted into distance or angular constraints, where perfect agreement results in zero constraint penalty score.

3.10 Modeling Helical Assemblies with Electron Density

Cryo-electron microscopy (cryoEM) has rapidly developed over the last decade due to improvements in electron detectors and reconstruction algorithms. A range of cryoEM structures with 2–5 Å resolution has been presented [15]. A resolution below 4.5 Å is needed to observe side chains, while medium resolution (4.5–6 Å) is sufficient to observe individual helices [15]. Rosetta has a large palette of methods to build and refine protein structures into medium- to high-resolution electron density maps from cryoEM, described in DiMaio et al. [15] (*see* **Note 7**). Lower-resolution maps can still be useful in modeling helical assemblies but do not provide sufficient data to define the atomic structure. Medium-resolution maps can be combined with symmetrical protein docking with preexisting structures or molecular models to generate a helical assembly model. Alternatively, external tools for fitting of atomic models into electron density maps can be used to consecutively place subunits into the helical density, followed by symmetrization with *make_symmdef_file.pl* to generate the symmetry input for Rosetta modeling and refinement.

3.11 Modeling Helical Assemblies with X-Ray Fiber Diffraction Data

For experimental samples of helical fibrils with large degrees of alignment, X-ray fiber diffraction can provide data down to 3 Å resolution and enable determination of atomic structures of helical assemblies [14]. DNA, helical viruses, and collagen have traditionally been structurally characterized using X-ray fiber diffraction. Compared to X-ray crystallography, fiber diffraction data has considerably lower information content and does not reach as high resolution [14]. Hence, structure modeling plays a more central role in the structure determination process [16]. Even if only lower-resolution data can be determined, it can provide some useful constraints in modeling, such as the helical parameters. Fiber diffraction is modeled as a pseudo-energy in Rosetta. There is a low-resolution and high-resolution fiber diffraction score in Rosetta, which means that any modeling protocol can be driven by X-ray fiber diffraction data [16] (*see* **Note 8**). The fiber diffraction data has to be preprocessed to extract intensity profiles across layer line, which is a standard procedure for analysis of well-ordered noncrystalline data. The energy function is a weight times the R-factor or χ^2 between experimental data and model fiber diffraction spectra (*see* **Note 8**). To reduce overfitting, a work and free set is created using Shannon sampling theory [16]. The work and free set is generated by a Rosetta application, *FiberDiffractionFreeSet*. Note that the R-factor in fiber diffraction is considerable lower than those observed in X-ray crystallography due to the cylindrical symmetry of the system. R-factors of 8–15% are often observed for solved structures. However, R-factors can be artificially low because of overfitting, since a cross-validation has not been consistently used in solving fiber diffraction structures [16].

3.12 Modeling Helical Assemblies with NMR Data

Various solid-state magic angle spinning (MAS) NMR measurements can be used to characterize the structure of helical assemblies. Chemical shift anisotropy (CSA) and dipolar coupling (DC) both provide information about the bond orientation of individual bond vectors in the protein. Nonetheless, the two data sources give distinctly different information so the combination of CSA and DC is a powerful constraint to determine the local structure of protein subunits in a helical assembly. In combination with X-ray fiber diffraction data, which provides long-range constraints, the helical assembly can be defined more precisely [17]. Dipolar couplings and CSA are implemented as pseudo-energy functions in Rosetta. 2D ^{13}C-^{13}C correlation spectroscopy can also generate atomic distance constraints, which can be encoded as NOE-type constraints in Rosetta (*see* **Note 9**). For example, Morag and colleagues used ^{13}C-^{13}C constraints from MAS NMR in combination with fold-and-dock simulations in Rosetta to determine the structure of M13 phage structure [18]. Selective labeling can be used to distinguish inter- vs intramolecular distances, which can be a challenging problem in NMR structure determination of symmetrical systems.

3.13 Modeling Helical Assemblies with Cross-Linking Mass Spectrometry

Cross-linking mass spectrometry can provide low-resolution distance constraints in structure modeling. A typical experiment involves the use of a cross-linker that connects lysine residues. The upper bound for the Cβ-Cβ distance of cross-linked lysine residues with BS3 cross-linker is 21.4 Å, so a single distance constraint provides little information, but such data can be useful in distinguishing between competing models if a significant number of cross-links can be determined. A difficulty in analyzing data from homomeric systems is distinguishing intra- vs. intermolecular cross-links. However, the use of proteins that are partially labeled with isotopes can alleviate this problem [19] in a similar fashion to solid-state NMR. There are a number of possible mathematical forms for the constraints to model the cross-linking data within Rosetta (*see* **Note 10**).

4 Notes

1. Michael Palmer at the University of Waterloo has developed a script, MakeMultimer.py, that generates the structure of an assembly from the BIOMT records in the PDB file. The script can be downloaded or can be run through a server at http://watcut.uwaterloo.ca/tools/makemultimer. For example, for Hibiscus latent Singapore virus (HLSV) with PDB id 3PDM

   ```
   python MakeMultimer.py 3PDM.pdb
   ```

   the script uses the same chain id for all identical subunits in the file. In preparation for analysis with symmetry definition script, chain ids should be added for two consecutive subunits in the assembly (chains S + T in our example below).

2. To generate the symmetry definition file for the assembly generated in **Note 1,** the following command can be used:

   ```
   make_symmdef_file.pl -m HELIX -a S -b T -t
   17 -p 3pdm_mm1_chain.pdb > 3pdm.symm
   ```

   where command line switches **a** and **b** specify the subunits to expand the symmetry around (selected to be at the center of the subunits in the input PDB; *see* **Note 1**). Switch **t** specifies how many subunits to expand in each direction. There are around 16.3 subunits per turn, so to get subunits above and below subunit S, we add 17 on each side to make it 35 subunits in total (Fig. 2). The master subunit with chain id S will be relabeled to chain A by the script and placed in a file with the ending _INPUT.pdb. This file and the symmetry definition (3pdm.symm in the above example) are used to model symmetry of the system.

3. To generate a symmetry definition file for de novo modeling with standard symmetry, the following command can be used:

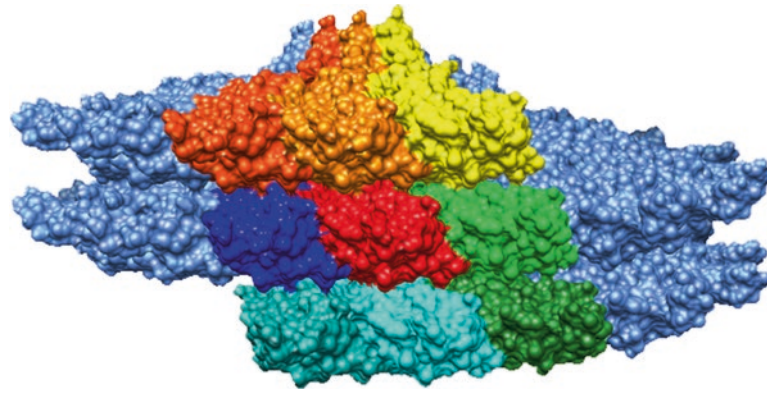


Fig. 2 Modeled symmetrical system of Hibiscus latent Singapore virus (PDB ID: 3PDM) [21]. The central subunit (red, the master subunit) is surrounded by all its nearest neighbors in the structure when a system of 35 subunits is used for modeling

```
make_helix_denovo.py -p 2.9 -n 40 -v 5 -u
27> helix_denovo.symm
```

where command line switch **p** specifies the helical rise (=**z**), **v** the number of turns (=**t**), and u the number of subunits over a helical repeat. **phi** is calculated from these values as $360 \times \mathbf{t}/\mathbf{u} = 66.67°$. **n** is the number of modeled subunits.

The default symmetry definition file generated by the script does not involve translation of subunits along the radial direction and does not specify random reorientation of subunits at initialization of the symmetry of the system in Rosetta. The relevant line of the symmetry definition file that specify the sampled rigid body degrees of freedom is

set_dof JUMP_0_0_0_to_subunit x angle_x angle_y angle_z

Changing this line to

set_dof JUMP_0_0_0_to_subunit x(50) angle_x(0:360) angle_y(0:360) angle_z(0:360)

specifies that the chains should be translated 50 Å perpendicular to the helical axis (along the x axis) and that the subunit should be randomly rotated around its center of mass. This is a good starting point for global docking and de novo folding. To apply these transformations, an additional Rosetta flag is needed (-symmetry:initialize_rigid_body_dofs).

4. When an existing helical assembly is analyzed with *make_symmdef_file.pl*, all degrees of freedom are defined by default. For example, analysis of 3PDM results in a section of the symmetry definition file with the following definitions:

set_dof JUMP_0_0_0 z(1.438897195189) angle_z

set_dof JUMP_0_0_0_to_com x(59.1390148517849)

set_dof JUMP_0_0_0_to_subunit angle_x angle_y angle_z

Here z controls the helical rise, angle_z controls the helical rotation angle, x controls the radius, and angle_x, angle_y, and angle_z control the rotation of the subunit around its center of mass. If all three lines are kept in the symmetry definition file, all 6 degrees of freedom can vary during the simulation.

The default de novo script does not allow sampling of the helical rise and helical rotation angle. Such sampling can be added with an additional line

set_dof JUMP_0_0_0 z(1.2:1.8) angle_z(-5:5).

Here we sample the helical rise (z) and helical rotation angle (angle_z) and initialize them randomly in a range. The angle values are a displacement, not an absolute number. If the symmetry definition generated by *make_helix_denovo.py* specifies that the angle_z is 45° (by setting the number of subunits and the number of turns), the above setting will randomize the value uniformly in the range 40–50° (not −5 to 5°). The value for the helical rise is an absolute value, meaning that z is initialized uniformly in the range 1.2–1.8 Å. This is generally true for rotations and translations.

5. The default mode of the relax application is not to move the rigid body degrees of freedom. To add this feature, an additional Rosetta flag is introduced (-relax:jump_move true). The degrees of freedoms that can vary during the simulation are specified in the symmetry definition file. In the symmetry definition file referred to in **Note 3**, there are four symmetrical degrees of freedom (dof): one translational dof controlling the radius of the helix and three rotational dofs controlling the orientation of each subunit. During docking steps, translational moves are often so-called slide moves, where subunits are translated until they come into contact with another chain. When there are several translational dofs in a system, this involves a multidimensional slide move. The order which this is done can be controlled. There are several options for this. For example, if

   ```
   slide_type ORDERED_SEQUENTIAL
   slide_order JUMP1 JUMP2
   ```

   is added to the symmetry definition file, sliding will happen along JUMP1 first and then JUMP2.

6. Comparative modeling requires the use of a fragment library. A fragment library can be generated locally using a script within Rosetta. However, it is often more convenient to use

the Robetta server (http://www.robetta.org) to produce fragments. RosettaCM is a Python script that generates a Rosetta script xml file that is run through the rosetta_scripts application in Rosetta (*Rosetta/tools/protein_tools/setup_RosettaCM.py*).

7. There are a number of demos and tutorials distributed with Rosetta describing how to run with electron density data; see, for example, Rosetta/demos/public/cryo_em_tutorial/ and Rosetta/public/electron_density_structure_refinement.
8. Detailed installation information and tutorials and running examples are found at http://www.cmps.lu.se/biostruct/people/ingemar-andre/fiber-diffraction/. For all-atom modeling, the computational cost of computing an X-ray fiber diffraction increases with the square of the number of atoms in a subunit. To reduce computation cost, Rosetta can be run with GPU acceleration (requires modified compilation of Rosetta; *see* http://www.cmps.lu.se/biostruct/people/ingemar-andre/fiber-diffraction for details). The low-resolution score function is called *fiberdiffdens* in the Rosetta score function, while the all-atom score is called *fiberdiffraction*. The default scoring function uses χ^2 during refinement. To change to R-factor, an additional flag is added (-fiber_diffraction:rfactor_refinement).

   Fiber diffraction data has the following format in Rosetta:

   62.523 0.101 0

   The first column is intensity, the second column is the reciprocal distance (R) in $Å^{-1}$, and the third column is the layer line.
9. Distance constraints from solid-state NMR can be introduced as NMR-styled restraints in Rosetta with accounting for the r^{-6} distance dependence of through-space NOEs. For example,

   AtomPair CG2 10 C 15 BOUNDED 1.500 7.000 0.300 NOE

   specifies a NOE distance constraint between CG2 of residue and C of residue 15. The constraint has a bounded energy with a lower bound of 1.5 Å and an upper bound of 7 Å, with a standard deviation of 0.3 Å. The Rosetta manual provides a more detailed description of the functional form of the constraint. To enable the constraints, a Rosetta score function with atom pair constraints turned on must be employed (referred to as *atom_pair_constraint* in the score function).

   To turn on CSA and dipolar coupling, *csa* and *dc* must be turned on in the energy function. The file format for dipolar coupling is

   7 N 7 H 2.700 0.1

where the columns are residue number for the first atom [7] in the bond for the measured dipolar coupling, second is the atom type (N), third is the residue number for the second atom in bond for the measured dipolar coupling (H), the fourth is the value of the dipolar coupling in Hz (2.7), and the fifth is the measurement error (0.1). This exemplifies data for an N-H dipolar coupling.

The file format for chemical shift anisotropy is

7 N 56.3 79.0 224.0 179.4 9.4

where column 1 is residue number [7], column 2 atom number (N), column 3 σ_{11} in ppm's (56.3), column 4 σ_{22} (79.0), column 5 σ_{33} (224.0), column 6 the experimental chemical shift anisotropy in ppm's (179.4), and column 7 the experimental error (9.4). This exemplifies data for backbone amide nitrogen CSA.

10. Protein flexibility can result in cross-links at larger distances than the expected 21.4 Å, as measured from a crystal structure. It is therefore useful to allow a bit of fuzziness in the distance cutoff. For example, we have used a SIGMOID constraint function in Rosetta to model lysine-lysine cross-links [19] with the functional form

$$\text{score} = \frac{1}{\left(1 + e^{25-d}\right)}$$

where d is the distance between Cβ atoms in cross-linked lysine residues.

5 Discussion

The Rosetta molecular modeling framework [2] provides a toolbox of methods to model homomeric assemblies with helical symmetry [3]. The starting point can be an existing helical structure, a homologous helical system, a single subunit, or the amino acid sequence. Modeling can involve full exploration of all degrees of freedom in the system—backbone, sidechain, and rigid body—and the extent of rigid body sampling can be controlled in detail (Fig. 3). The ability of Rosetta to employ a range of experimental data as constraints during modeling makes it a powerful system for data-guided structure determination and computational engine for exploring structural hypothesis. The machinery is also well suited to design proteins with helical symmetry, as we have recently demonstrated by designing a protein fibril de novo [20].

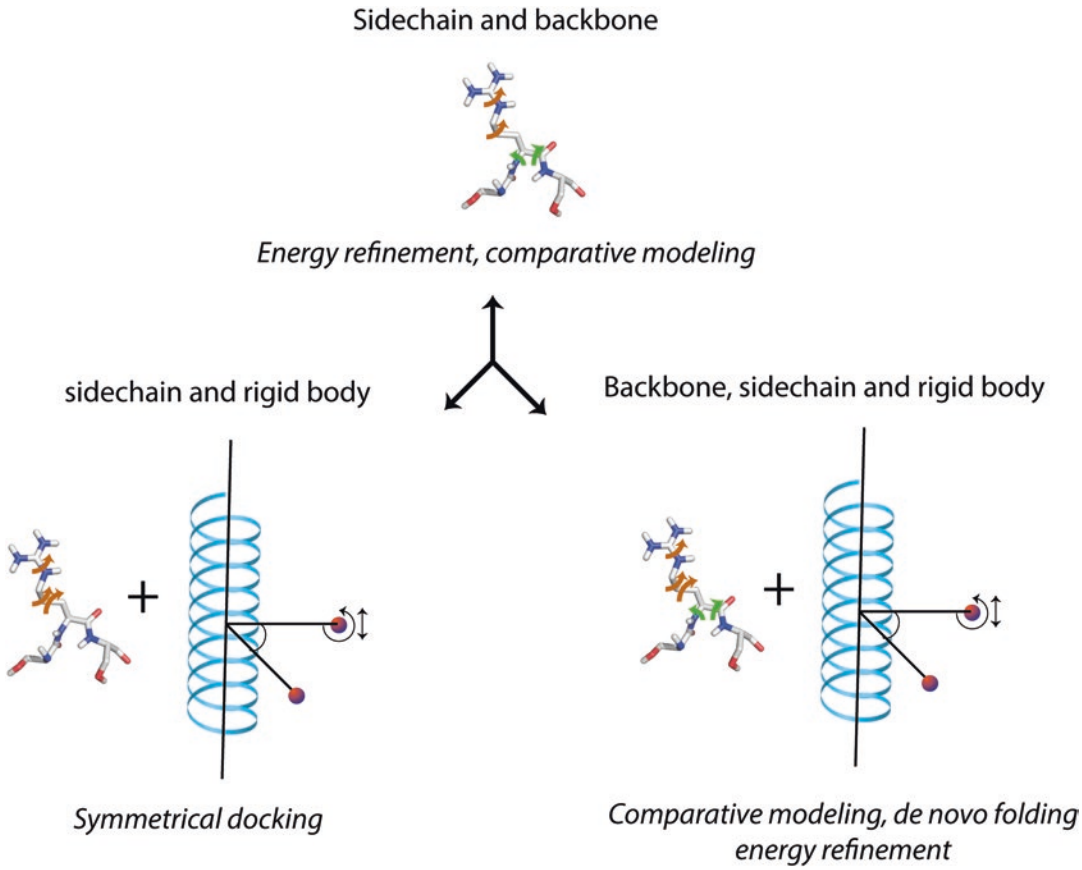


Fig. 3 Helical modeling scenarios. A helical system can be modeled with varying number of degrees of freedoms (dofs) explored during the simulation: (1) Backbone and sidechain conformational dofs can be explored while keeping the parameters for the helical symmetry fixed (*top*). (2) Exploring the sidechain dofs and the parameters for helical symmetry (*bottom left*). (3) Exploring sidechain and backbone dofs together with the parameters for helical symmetry (*bottom right*)

References

1. Goodsell DS, Olson AJ (2000) Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 29:105–153
2. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YE, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popovic Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 487:545–574. https://doi.org/10.1016/B978-0-12-381270-4.00019-6
3. DiMaio F, Leaver-Fay A, Bradley P, Baker D, Andre I (2011) Modeling symmetric macromolecular structures in Rosetta3. PLoS One 6(6):e20450. https://doi.org/10.1371/journal.pone.0020450
4. Klug A, Crick FHC, Wyckoff HW (1958) Diffraction by helical structures. Acta Crystallogr 11(3):199–213. https://doi.org/10.1107/S0365110x58000517
5. Kendall A, McDonald M, Stubbs G (2007) Precise determination of the helical repeat of tobacco mosaic virus. Virology 369(1):226–227. https://doi.org/10.1016/j.virol.2007.08.013
6. Andre I, Bradley P, Wang C, Baker D (2007) Prediction of the structure of symmetrical protein assemblies. Proc Natl Acad Sci U S A 104(45):17656–17661
7. Tyka MD, Keedy DA, Andre I, Dimaio F, Song Y, Richardson DC, Richardson JS, Baker D

(2011) Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol 405(2):607–618. https://doi.org/10.1016/j.jmb.2010.11.008

8. Fleishman SJ, Leaver-Fay A, Corn JE, Strauch EM, Khare SD, Koga N, Ashworth J, Murphy P, Richter F, Lemmon G, Meiler J, Baker D (2011) RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. PLoS One 6(6). https://doi.org/10.1371/journal.pone.0020161. doi: ARTN e20161
9. Song Y, DiMaio F, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D (2013) High-resolution comparative modeling with RosettaCM. Structure 21(10):1735–1742. https://doi.org/10.1016/j.str.2013.08.005
10. Das R, Andre I, Shen Y, Wu Y, Lemak A, Bansal S, Arrowsmith CH, Szyperski T, Baker D (2009) Simultaneous prediction of protein folding and docking at high resolution. Proc Natl Acad Sci U S A 106(45):18978–18983. https://doi.org/10.1073/pnas.0904407106
11. Ramisch S, Lizatovic R, Andre I (2015) Exploring alternate states and oligomerization preferences of coiled-coils by de novo structure modeling. Proteins 83(2):235–247. https://doi.org/10.1002/prot.24729
12. Sachse c (2015) Single-particle based helical reconstruction—how to make the most of real and Fourier space. AIMS Biophys 2(2):219–244
13. Yan S, Suiter CL, Hou G, Zhang H, Polenova T (2013) Probing structure and dynamics of protein assemblies by magic angle spinning NMR spectroscopy. Acc Chem Res 46(9):2047–2058. https://doi.org/10.1021/ar300309s
14. Chandrasekaran R, Stubbs G (2012) In: Arnold E, Himmel D, Rossmann MG (eds) International tables for crystallography Vol. F: crystallography of biological macromolecules, 2nd edn. Wiley, Chichester, p 583–592
15. DiMaio F, Chiu W (2016) Tools for model building and optimization into near-atomic resolution electron Cryo-microscopy density maps. In: Resolution revolution: recent advances in cryoem, vol 579. Elsevier, Amsterdam, pp 255–276. https://doi.org/10.1016/bs.mie.2016.06.003
16. Potrzebowski W, Andre I (2015) Automated determination of fibrillar structures by simultaneous model building and fiber diffraction refinement. Nat Methods 12(7):679–684. https://doi.org/10.1038/nmeth.3399
17. Lipsitz RS, Tjandra N (2003) N-15 chemical shift anisotropy in protein structure refinement and comparison with NH residual dipolar couplings. J Magn Reson 164(1):171–176. https://doi.org/10.1016/S1090-7807(03)00176-9
18. Morag O, Sgourakis NG, Baker D, Goldbourt A (2015) The NMR-Rosetta capsid model of M13 bacteriophage reveals a quadrupled hydrophobic packing epitope. Proc Natl Acad Sci U S A 112(4):971–976. https://doi.org/10.1073/pnas.1415393112
19. Boelt SG, Norn C, Rasmussen MI, Andre I, Ciplys E, Slibinskas R, Houen G, Hojrup P (2016) Mapping the Ca2+ induced structural change in calreticulin. J Proteome 142:138–148. https://doi.org/10.1016/j.jprot.2016.05.015
20. Kaltofen S, Li C, Huang PS, Serpell LC, Barth A, Andre I (2015) Computational de novo design of a self-assembling peptide with predefined structure. J Mol Biol 427(2):550–562. https://doi.org/10.1016/j.jmb.2014.12.002
21. Tewary SK, Oda T, Kendall A, Bian W, Stubbs G, Wong SM, Swaminathan K (2011) Structure of hibiscus latent Singapore virus by fiber diffraction: a nonconserved his122 contributes to coat protein stability. J Mol Biol 406(3):516–526. https://doi.org/10.1016/j.jmb.2010.12.032

Chapter 31

Selecting Conformational Ensembles Using Residual Electron and Anomalous Density (READ)

Loïc Salmon, Logan S. Ahlstrom, James C. A. Bardwell, and Scott Horowitz

Abstract

Heterogeneous and dynamic biomolecular complexes play a central role in many cellular processes but are poorly understood due to experimental challenges in characterizing their structural ensembles. To address these difficulties, we developed a hybrid methodology that combines X-ray crystallography with ensemble selections typically used in NMR studies to determine structural ensembles of heterogeneous biomolecular complexes. The method, termed READ, for residual electron and anomalous density, enables the visualization of heterogeneous conformational ensembles of complexes within crystals. Here we present a detailed protocol for performing the ensemble selections to construct READ ensembles. From a diverse pool of binding poses, a selection scheme is used to determine a subset of conformations that maximizes agreement with the X-ray data. Overall, READ is a general approach for obtaining a high-resolution view of dynamic protein-protein complexes.

Key words Crystallography, Ensemble, Conformational dynamics, Protein structure, Structural biology

1 Introduction

The majority of proteins perform their cellular functions through interactions with other proteins [1]. Thus, obtaining high-resolution structural models of protein-protein interactions yields indispensible insight into biological function and can facilitate the development of therapeutic strategies. However, many protein-protein interactions involve highly dynamic protein-protein association (e.g., intrinsically disordered protein-partner interactions in cell signaling and regulation networks [2]) and thus are particularly challenging for traditional X-ray crystallography and NMR methods. Although recent advances in X-ray crystallography [3] and NMR spectroscopy [4–6] allow one to analyze biomolecules in terms of ensembles as opposed to single structures, the information from these techniques is often limited by system size and

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_31,

complexity. X-ray crystallography can provide valuable information on small-scale conformational changes, but observing heterogeneous conformational states falls beyond the reach of current crystallographic techniques. Such studies can be done with NMR, but the systems must exhibit particular properties such as a certain conformational exchange rate, favorable interactions with external alignment media, or sufficient chemical shift dispersion.

To further our ability to obtain high-resolution structural ensembles of dynamic protein-protein interactions, we need to develop new approaches that combine the benefits of experiments and computational strategies in a transferable manner. The elucidation of ensembles of RNA, DNA-protein complexes, and highly flexible proteins by NMR has provided valuable insights into many complex biological problems [4–7]. NMR-based approaches can derive highly heterogeneous ensembles to describe complex and flexible interactions, a feat that current crystallographic methods are unable to replicate. Here, we combine crystallographic measurements with an NMR-like ensemble approach [8, 9], yielding a general method for observing structural and dynamic properties of proteins and protein complexes [10]. The approach, termed READ, for residual electron and anomalous density, determines the conformational ensembles of heterogeneous and dynamic proteins by X-ray crystallography. We will first describe the overall approach in general terms and then provide methodological details.

2 Materials

2.1 Residual Electron and Anomalous Densities

1. The foremost requirement for READ is sturdy and highly reproducible crystals of the protein or complex of interest that diffract to ~3 Å or higher resolution. Given the ambiguous nature of the electron density, we recommend validating the presence of any components of interest via crystal washing followed by gel or mass spectrometry analysis.
2. Initial attempts at modeling the structure should be made using standard crystallographic model building and refinement procedures, taking the refinement to as close to completion as possible. The resulting structure may later be used as part of a molecular dynamics simulation of the crystal, and the refined electron density will be compressed for use in the selection procedure.
3. Methods for incorporating anomalous scatterers into crystallization components can vary depending on the system. We have successfully used peptide synthesis with unnatural amino acids like iodophenylalanine [10]. Other potential methodologies include amber stop codon suppression [11] to incorporate iodine-containing amino acids; chemical labeling of natural

amino acids, such as tyrosine-iodine modifications [12]; or binding of heavy metals to cysteine residues. How many anomalous substitutions are required will likely vary depending on the system. As a general guideline, we recommend beginning by substituting non-conserved residues or using chemically similar substitutions.

2.2 Conformational Pool

1. The second prerequisite to using the READ protocol is to obtain a large conformational pool of the dynamic component of the crystal. Depending on the size and level of mobility of the substrate, different computational approaches could be used. For cases in which the dynamic fragment is especially large, undergoes high-amplitude conformational changes, or translates considerably within the crystal, we have found a coarse-grained approach to be effective [10, 13, 14]. Monte Carlo-based approaches such as flexible-meccano [15] could also be useful, or for more computationally amenable systems, all-atom simulations [16] could be applicable.
2. The pool of conformations in the crystalline environment should sample all possible conformational space. The crystalline environment can be simulated either through multiple pseudocrystal environments [10] or by altering the simulation's boundary conditions to match the crystal's dimensions and space group. It will often be more appropriate to favor extensive sampling of space over increased accuracy of local structural details, in which case a coarse-grained approach is especially appropriate.

2.3 Computational Resources

1. The compute time for the selection procedure will depend on various parameters including the amount of experimental data, the size of the conformational pool, and the exact parameters used. For simple cases, a selection should run on a modern desktop computer in a few hours or less. Therefore, standard selection procedures can be easily setup using a single desktop. Extensive testing, multiple cross-validations, or bootstrapping simulations may, however, require a small calculation cluster in order to run in a time-efficient manner (*see* **Note 1**). Additionally, the refinement of large conformational ensembles through classical crystallographic approaches will greatly benefit from using a small computer cluster with a crystallographic refinement package such as Phenix [17] or CCP4 [18] installed.
2. To run the READ selection, the computer or cluster must have MATLAB 2015b or later installed. The READ selection program is implemented as a MATLAB toolbox called READ.mltbx. To install the toolbox, simply double-click READ.mltbx, which will install the code, and add it to your path. The

MATLAB toolbox can be downloaded from https://bitbucket.org/umarsdev/mcdb-bardwell-ga. Example data for the selection can be downloaded from https://bitbucket.org/umarsdev/mcdb-bardwell-ga_datafiles.

3 Selection Algorithm

1. The selection of the conformational ensemble can be carried out using various algorithms. Selection procedures can be performed using a Monte Carlo Metropolis approach [9] or using a genetic algorithm [8]. Both approaches can potentially work, but in our experience, a genetic algorithm is computationally more efficient when adequately optimized and was successfully used in our recently published example [10].
2. The overall READ procedure uses the residual electron and anomalous data to select sub-ensembles from a large structural pool (*see* **Note 2**). Starting from a pool of conformations of typically 10,000 members, the selection is initiated by randomly generating X ensembles from this pool, where each ensemble contains N conformers. In a typical selection, X is set to 100, and N typically ranges from 1 to 20 [10]. Then, K evolution steps are iteratively performed. K is typically 50,000. Tweaking these parameters will affect the outcome of the selection, and the parameters need to be optimized as part of this protocol. Each evolution step has two phases. The first phase creates new ensembles to diversify the set of possible solutions, and the second phase performs a selection to determine which of the available ensembles best fit the data.
3. New ensembles are generated in the first phase. $4X$ new ensembles are produced by four methods: X by reproduction, X by external mutations, X by internal mutations, and X randomly. Reproduction generates a new ensemble by randomly selecting conformers from two parent ensembles. Mutations are performed by randomly changing one conformer in the ensemble with a new conformer either using conformations already present (internal mutation) or not present (external mutation) in one of the selected ensembles. At the end of this phase, the original X ensembles are combined with the $4X$ new ensembles, providing a total of $5X$ ensembles. This procedure is implemented so that a given conformation cannot appear more than once in any ensemble and any ensemble cannot be present more than once at each step. This ensures maximal diversity in the sampling of the conformational space.
4. In the second phase, the ensembles are evaluated. The $5X$ ensembles are clustered randomly into T tournaments. Within each tournament, the ensembles are compared to determine

which best fits the data. The best-fitting ensembles from each tournament are retained, forming a new set of X ensembles. During the course of the selection procedure, the number of tournaments (T) successively decreases, typically from X to 1, increasing the selection pressure. Low selection pressure is important at the beginning of the selection to ensure sufficient diversity, whereas increasing the selection pressure later improves the convergence of the result [19]. In the final rounds of the selection, only a single tournament is used to produce the single best-fitting ensemble.

5. To select the best-fitting ensembles, a χ^2 function is used in the tournaments. Best-fitted ensembles are those that minimize this target function by reducing the difference between the experimental data and that back-calculated from the ensembles. This χ^2 function is defined as

$$\chi^2 = \chi^2_{elec} + \lambda\chi^2_{anom}$$

where χ^2_{elec} and χ^2_{anom} are the contributions arising from the electron density and the anomalous signal positions, respectively, and λ is a scaling factor that allows one to adjust the relative weight of the two data types.

6. For the anomalous data, it is necessary to predict the position of the anomalous atoms from the coordinates. However, the initial conformational sampling may be obtained for the sequence of the non-modified peptide chain and thus might not explicitly include the modifications containing the anomalous atoms. Instead, the anomalous atom positions are inferred from the geometry of the conformers. For example, using iodophenylalanine, the fixed aromatic ring dictates a distance between the Cα and the iodine scatterer of 6.5 Å. Thus, the selection searches for iodines as close as possible to 6.5 Å from Cα of the mutated residue [10]. This particular approach does not take into account any angular information from the conformer; however, angular information could also be encoded if the experimental information and computational procedure allow for it. Although this approach avoids assumptions regarding backbone angles, it does introduce the problem that anomalous atoms can possibly be placed in positions that cause steric clashes or induce backbone distortions.
7. The anomalous data target function is encoded into the selection using

$$\chi^2_{iodo} = c^2 \sum_i \left(1 - \exp\left[-\left(\frac{D_i^{calc} - D_i^{\exp}}{c\delta^{anom}}\right)\right]^2\right)$$

where i represents the different anomalous atom positions, D_i^{calc} is the closest distance between the anomalous atom position and the corresponding Cα of any of the conformers in the ensemble, $D_i^{\exp}$ is the corresponding expected distance, δ^{anom} is a weight estimating the uncertainty in the anomalous atoms' position, and c is fixed at a canonical value of 2.9846 (*see* **Note 3**). The value of $D_i^{\exp}$ depends on the nature of the amino acid bearing the anomalous scatter and represents the distance between the anomalous scatter atom and the Cα of the same residue (e.g., 6.5 Å in the case of iodophenylalanine). The value of δ^{anom} in this protocol was the mean squared displacement of the scatterer, derived from the B-factors of the atoms after crystallographic refinement of the anomalous scatterers with a fixed occupancy. In this protocol, occupancy was fixed at 0.5, and crystallographic refinement was performed in Phenix [17].

8. In general, electron density maps contain an enormous amount of data. The number of data points in a typical map is computationally too expensive for practical direct use with this selection procedure. As a result, procedures to compress this information are required. The first compression approach is to only describe the electron density in the space covered by the conformational pool. Importantly, any density outside the initial conformational sampling might be important for understanding the system, but it cannot be recovered by any selection procedure if it is not initially sampled. Removal of data not sampled by the conformational pool dramatically decreases the number of points needed to describe the data by shifting from the entire experimental three-dimensional space of the asymmetric unit to the ensemble of coordinates that account for all the conformers in the pool. Second, the electron density is binned in three-dimensional space to create a three-dimensional electron density histogram. Third, the electron density is filtered for signal intensity and sufficiently averaged to further compress the data. Finally, computing the electron density contribution of each conformer before the selection, and then using these precomputed values as an input in the selection greatly speed the procedure. Thus, in the selection procedure, only simple operations such as averaging over a limited number of points are required to determine the fit of the ensembles to the electron density.

9. For the electron density evaluation, the following target function is used:

$$\chi^2_{elec} = \sum_i \left(\frac{\rho_i^{calc} - \rho_i^{\exp}}{\delta^{elec}} \right)^2$$

where i spans the different residual electron density data points (discussed in Subheading 4.1), ρ_i^{exp} is the experimental value, ρ_i^{calc} is the predicted value from the considered ensemble, and δ^{elec} represents a constant weight estimating the noise in the residual electron density map (*see* **Note 4**).

4 Methods

4.1 Preparing the Electron Densities

1. Prepare individual density maps. Before starting the READ procedure, each conformer in the initial large conformational pool must be crystallographically refined, one at a time, together with any fixed atoms in the structure. The refinement procedure should include at least B-factor refinement, but could include other parameters as well. The output Cα atoms of the conformers from this refinement then serve as the input structures for the selection. Depending on the conformational pool, geometric restraints may be necessary during refinement to ensure that the conformers maintain energetically reasonable conformations. This crystallographic refinement can be accomplished with any modern crystallographic software package such as Refmac [20] or Phenix [17], but a computing cluster will likely be required to efficiently refine the entire conformational pool (*see* **Note 5**).
2. Average the electron density maps. Taking the refined electron density for each conformer, an average map over the simulation must be produced for subsequent back-prediction of the conformer electron density. This averaged map will allow consistent sampling of the electron density within the crystal space. The average 2mFo-DFc map from all of the refined conformational pool snapshots is calculated using Phenix:

   Run command phenix.average_map_coeffs file_list=file_list.dat labin_mtz="FP=2FOFCWT PHIB=PH2FOFCWT" [output_file=mapfile.mtz]

   where file_list.dat has spaces or line breaks separating the different mtz file names in the folder containing the refined files. This averaged electron density map then serves as the basis for the electron density used in the selection procedure.
3. Extract map values from the averaged electron density map. After creating the averaged density maps, the contribution of each conformer to the averaged electron density can be estimated. In this procedure, the electron density was estimated for each conformer using the 2mFo-DFc map value of the refined atom using the eight-point interpolation function in cctbx [21]. The script eight_point.py can be used to loop over the conformational pool and output these map values. The

name of the model PDB in line 47 eight_point.py must be changed to that of the most up-to-date refined structure of the protein, and the folder name in line 68 must be changed to that containing the individually refined conformational pool. The name of the mtz file containing the electron density data is specified in line 29 of eight_point.py.

4. Bin and filter the electron density. Even after this reduction, the map values will still require typically ~100,000 data points, which remain computationally too expensive for this procedure to be practical. To further compress the information, the map values are binned in three-dimensional space. The binning procedure averages the map values extracted in **step 3** to create a three-dimensional histogram of the electron density. The bin size can be adjusted to fine-tune the balance between data compression level and accuracy. The resolution of the binning is an important parameter in the procedure. Testing different bin sizes and selecting the most appropriate size can be aided using the cross-validation test described below. To ensure correct addition of the bins, the same binning procedure must be performed for all the conformers using the exact same grid.

   Three additional filters were also included to further reduce the data (*see* **Note 6**). The first and second filters pertain to the electron density level. High map values corresponding to fully rigid parts of the system that should be handled with standard crystallographic approaches are removed from the selection procedure using the first threshold. Conversely, regions of especially low density are likely noise and can be removed using the second filter. Even removing a region with no density will modify the outcome of the selection, as a region with a zero-density level is treated differently than not having the region in the selection. The third filter ensures sufficient sampling. Selecting for a region that is extremely poorly sampled (e.g., only a few conformers within the ~10,000 pool) will not account for the possible structural heterogeneity in the region. If those regions are for some reason judged as essential in the description of the system, the initial sampling should be improved to pass this filter by changing the approach to generating the initial large conformational pool (*see* **Note 7**).

   To bin the electron density, run density_histogram.m. This program gives as outputs the compressed electron density for the target as well as all the conformers. The resulting compressed electron density map will be used as an input in the selection procedure, which will only need to average these values over the conformers selected in a given ensemble. To set the noise level of the electron density maps, set errorX1k. The different filters are inputted in the electron density binning procedure using the parameters densminX1k, densmaxX1k, and popmin.

4.2 Preparing the Anomalous Data

1. The anomalous atoms are limited in number and therefore can be directly used in the selection procedure. Depending on the space group, the positioning of the anomalous atoms, and the method for generating the conformational pool, you may also need to generate crystallographically identical anomalous atoms in neighboring asymmetric units. This symmetry generation can be easily accomplished in PyMOL. After manually merging all the anomalous atoms to the same PDB file using a text editor, open the file in PyMOL. Then, click on the PDB object, choose generate > symmetry mates, and choose the size required to produce enough duplicate anomalous atoms to cover the required space.
2. Convert the PDB file into .sym format using the script iodine_sym.sh, changing the file name in the first line of the script to input PDB file name.
3. After the positions of the anomalous scatterers are defined, it is necessary to precompute the distance between this position and the corresponding Cα of that residue within each conformer. This can be achieved using the script iodo_dist.m, using a text file containing a list of the .sym anomalous signal positions in which each line contains the complete path to the a .sym file, as well as a text file listing the file names for the input PDB files from the conformational pool.

4.3 Running the Selection

1. After the experimental data and conformers have been prepared, one can run the selection. Starting with the initial pool, the selection searches for the ensemble that best reproduces the experimental data. A few general parameters will be key in the process and need to be optimized as described below.
2. To run the selection with default parameters, set chi2scale=1, nbGen=50,000, and nTrials=1, and run the following commands from the MATLAB command window:

```
outputFile = 'output';
iofile = '/path/to/data/lists/list_anomalous_data';
densfile = '/path/to/data/lists/list_density'
params = processgalist(outputFile, iofile, densfile, 'chi2scale',
1, 'nbGen', 50000, 'nTrials', 1);
```

3. Set the selection ensemble size. The first parameter to consider is the number of conformers in the ensemble (*see* **Note 8**). In general, increasing the number of conformers will better reproduce the data but can lead to overfitting unless controlled for using cross-validation tests as described in Subheading 4.4. Usually, as the number of conformers in the ensemble increases, the data reproduction will reach a plateau value, indicating that adding one or a few extra conformers will not significantly

affect the ability to fit the data. Degraded data reproduction may sometimes be observed with an increased number of conformers if the initial pool does not contain enough usable conformers. As an example, if the pool contains only 20 conformations that are close to the correct solution, asking it to select 30 conformations will force it to keep inadequate conformations to fill out the requested ensemble size. The selection should be performed at a variety of ensemble sizes, ranging from one to tens or hundreds of conformers, depending on the system. Validation tests are then used to select the appropriate ensemble size for analysis (*see* below).

To choose the number of conformations in the ensemble, set ensSize to the desired value. Typically, ensSize is varied between 1 and 20, and the results are then compared.

4. Set the data weighting. The second key parameter within the selection determines the relative weighting of the residual and anomalous data. In the ideal case, the weighting should allow for the reproduction of each type of experimental data within experimental error and with contributions reflecting their information content. It is worth noting that both the number of data points and the estimated errors affect the weight of each data type. If both datasets have similar information content, obtaining a similar final weight in the data reproduction can be reasonable (*see* **Note 9**).

   The relative weighting of the anomalous positions compared to the electron density data can be set using chi2scale.

5. Set the number of selection iterations. The easiest parameter to define is the number of iterations in the selection. This parameter should be large enough to allow for convergence of the algorithm; however, lowering the number will improve computational efficiency. One can estimate this parameter by running a few long initial selections and observing when convergence is reached. Convergence can be judged by monitoring the output files reporting on the χ^2 evolution during the selection. This number can be influenced by the amount of experimental data, the size of the conformational pool, and the other selection parameters.

To set the number of iterations, set the input parameter nbGen.

4.4 Testing the Validity and Accuracy of the Selected Ensemble

1. Test the ensemble by replication. After performing the selection procedure, several tests must be used to estimate the validity and accuracy of the selected ensemble. First, the selection should be repeated several times using the exact same conditions, as the selection process is based on stochastic approaches and thus may not always converge on the same results (*see* **Note 10**). This step can be used to determine if there is a well-defined single solution or a set of relatively equivalent solu-

tions that are able to reproduce the experimental data to a roughly similar extent. If several solutions exist, they should also present similar conformational properties, such as location in space, secondary structure content, or tertiary contacts. If the solutions do not present similar properties, it suggests that the amount of experimental information in the selection is not sufficient to properly define the selected ensemble.

To increase the number of replicated selections, increase the parameter nTrials or simply run the program again.

2. Test the approach using simulated data. It is helpful to test that the selection approach can recapitulate known structures pulled from the conformational pool (*see* **Note 11**). This can be accomplished by generating a few realistic ensembles of specific characteristics that deviate from the average of the conformational pool using the procedures described below. These ensembles are used to generate noise-corrupted synthetic datasets that are then used as the data for the selection. The advantage of this approach is that the selected ensemble can then be directly compared to the known target ensemble. The ability to correctly capture the features of the test ensemble can characterize the validity of the procedure (*see* **Note 12**).

   Simulated electron density maps can be easily generated using the phenix.fmodel tool (https://www.phenix-online.org/documentation/reference/fmodel.html). Condensed map generation is then performed as described above, with the inclusion of noise corruption using the density_histogram_target.m script. Similarly, noise-corrupted pseudo-experimenal anomalous data can be generated using the scriptiodine_dist_target.m.

3. Test the ensemble via cross-validation and choose the ensemble size. In a tenfold holdout cross-validation procedure, only 90% of the data are used, and 10% are randomly removed from the selection. At the end of the selection, the 10% unused data are back-predicted from the final ensemble to test the ensemble's ability to predict the unused experimental data. This process is repeated, typically ten times, removing a different subset of the data each time (*see* **Note 13**). The exact procedure will depend on the system under consideration, but a safe approach is to use the ensemble with the lowest number of conformers that successfully passes this cross-validation test.

   Holdout cross-validation tests can be seen as similar in essence to the R_{Free} crystallographic test. In fact, assessing the quality of an ensemble on data that are not actively used in the fitting procedure makes cross-validation tests more stringent than a direct fit and greatly helps in identifying potential overfitting or dataset inconsistencies. Thus, cross-validations can be useful in selecting ensemble size, checking data quality and self-consistency, and assessing the predictive power of an

ensemble. For example, if the data are overfit or inconsistent, it will not pass this test, and the predictive power of the selected ensemble will be worse than the initial pool. Additional external evaluations of the ensemble can also be used as validations, such as using the ensemble in a full crystallographic refinement procedure and computing its R_{free}.

To run the cross-validation, run the cross-validation form of the program using the "processgalist_crossval" function. In this version of the program, the additional parameter "indices" sets the fold number of the cross-validation.

4. Test the approach using bootstrapping. Bootstrapping can be used to estimate the precision of the selection. To perform bootstrapping, a random subset of data, typically on the order of 1/e (~37%) of the data [22], are removed from the fit and replaced by an equal number of repeated data points. This bootstrapped dataset is then used for the selection. The procedure is repeated a few hundred times and the precision of the selection evaluated using the obtained ensembles. Variations in the selected bootstrapped ensembles can be used to estimate the error in properties of the selected ensemble (*see* **Note 14**).

   To perform the bootstrapping procedure, generate, for example, 200 new datasets matching the bootstrapping strategy, and run "processgalist_bootstrapping" function. In this version, the parameter nTrials is used to set the number of bootstrapped datasets.

5 Notes

1. Genetic algorithms can easily be encoded to allow for parallel computing, which can allow for the treatment of more complex systems. The MATLAB program provided here can run multiple selections in parallel.
2. *See* Fig. 2 in Horowitz et al. [10] for a diagram of an example procedure.
3. Using this target function provides a classical χ^2 for data near the target value but reaches a plateau for data points that are very poorly reproduced and thus better handles outliers. This function is especially useful to avoid the situation in which poor reproduction of a single anomalous atom dominates the target function.
4. Residual electronic density remains between 0 and 1, making the use of the more complex χ^2 unnecessary.
5. The most efficient refinement protocol will likely be case-dependent.

6. It is important to note that these filters must be handled with care, as they explicitly remove experimental information about specific regions.
7. These filters may or may not be required for different systems, and their precise adjustment will also be system-specific.
8. Searching for a single conformer corresponds to a static description, and searching for more conformers corresponds to a dynamic description.
9. Obtaining a good estimate of experimental uncertainties helps in determining the scaling factor between different data types.
10. Convergence properties may be improved by changing the parameters used in the selection.
11. For an example of the results of a simulated data test, refer to Supplementary Fig. 5 from Horowitz et al. [10].
12. Using simulated data can also be a way to estimate the amount of experimental data required to accurately characterize a given system.
13. Holding out 10% of the data is a typical value. This percentage may be adjusted depending on the system. The value must be sufficiently high to provide a stringent test but low enough to not completely destabilize the selection procedure.
14. For an example of the results of a bootstrapping analysis, refer to Supplementary Fig. 7 from Horowitz et al. [10].

Acknowledgment

This work was supported by the National Institutes of Health (R01-GM102829 to J.C.A.B. and K99/R00-GM120388 to S.H.). J.C.A.B. is a Howard Hughes Medical Investigator. The authors would like to thank S. Rocchio for comments on the manuscript and M. Mourao for useful discussions. The authors would also like to thank C. Stockbridge and the LSA-IT development team for the assistance in coding.

References

1. Keskin O, Gursoy A, Ma B et al (2008) Principles of protein-protein interactions: what are the preferred ways for proteins to interact? Chem Rev 108(4):1225–1244. https://doi.org/10.1021/cr040409x
2. Wright PE, Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol 16(1):18–29. https://doi.org/10.1038/nrm3920
3. Fraser JS, van den Bedem H, Samelson AJ et al (2011) Accessing protein conformational ensembles using room-temperature X-ray crystallography. Proc Natl Acad Sci U S A 108(39):16247–16252. https://doi.org/10.1073/pnas.1111325108
4. Salmon L, Blackledge M (2015) Investigating protein conformational energy landscapes and atomic resolution dynamics from nmr dipolar couplings: a review. Rep Prog Phys 78(12):126601. https://doi.org/10.1088/0034-4885/78/12/126601

5. Salmon L, Yang S, Al-Hashimi HM (2014) Advances in the determination of nucleic acid conformational ensembles. Annu Rev Phys Chem 65:293–316. https://doi.org/10.1146/annurev-physchem-040412-110059
6. Venditti V, Egner TK, Clore GM (2016) Hybrid approaches to structural characterization of conformational ensembles of complex macromolecular systems combining nmr residual dipolar couplings and solution x-ray scattering. Chem Rev 116(11):6305–6322. https://doi.org/10.1021/acs.chemrev.5b00592
7. Jensen MR, Zweckstetter M, Huang JR et al (2014) Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using nmr spectroscopy. Chem Rev 114(13):6632–6660. https://doi.org/10.1021/cr400688u
8. Nodet G, Salmon L, Ozenne V et al (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from nmr residual dipolar couplings. J Am Chem Soc 131(49):17908–17918. https://doi.org/10.1021/ja9069024
9. Salmon L, Bascom G, Andricioaei I et al (2013) A general method for constructing atomic-resolution rna ensembles using nmr residual dipolar couplings: the basis for interhelical motions revealed. J Am Chem Soc 135(14):5457–5466. https://doi.org/10.1021/ja400920w
10. Horowitz S, Salmon L, Koldewey P et al (2016) Visualizing chaperone-assisted protein folding. Nat Struct Mol Biol 23(7):691–697. https://doi.org/10.1038/nsmb.3237
11. Wals K, Ovaa H (2014) Unnatural amino acid incorporation in E. coli: current and future applications in the design of therapeutic proteins. Front Chem 2:15. https://doi.org/10.3389/fchem.2014.00015
12. Santrucek J, Strohalm M, Kadlcik V et al (2004) Tyrosine residues modification studied by maldi-tof mass spectrometry. Biochem Biophys Res Commun 323(4):1151–1156. https://doi.org/10.1016/j.bbrc.2004.08.214
13. Kmiecik S, Gront D, Kolinski M et al (2016) Coarse-grained protein models and their applications. Chem Rev 116(14):7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163
14. Salmon L, Ahlstrom LS, Horowitz S et al (2016) Capturing a dynamic chaperone-substrate interaction using nmr-informed molecular modeling. J Am Chem Soc 138(31):9826–9839. https://doi.org/10.1021/jacs.6b02382
15. Ozenne V, Bauer F, Salmon L et al (2012) Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics 28(11):1463–1470. https://doi.org/10.1093/bioinformatics/bts172
16. Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106(5):1589–1615. https://doi.org/10.1021/cr040426m
17. Miller BL, Goldberg DE (1995) Genetic algorithms, tournament selection, and the effects of noise. Complex Syst 9(3):193–212
18. Afonine PV, Grosse-Kunstleve RW, Echols N et al (2012) Towards automated crystallographic structure refinement with phenix. Refine. Acta Crystallogr D D68:352–367. https://doi.org/10.1107/S0907444912001308
19. Winn MD, Ballard CC, Cowtan KD et al (2011) Overview of the ccp4 suite and current developments. Acta Crystallogr D Biol Crystallogr 67(Pt 4):235–242. https://doi.org/10.1107/S0907444910045749
20. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53(Pt 3):240–255. https://doi.org/10.1107/S0907444996012255
21. Grosse-Kunstleve RW, Sauter NK, Moriarty NW et al (2002) The computational crystallography toolbox: crystallographic algorithms in a reusable software framework. J Appl Crystallogr 35:126–136. https://doi.org/10.1107/S0021889801017824
22. Press WH (2007) Numerical recipes: the art of scientific computing, 3rd edn. Cambridge University Press, Cambridge

Erratum to: Characterizing Intact Macromolecular Complexes Using Native Mass Spectrometry

Elisabetta Boeri Erba, Luca Signor, Mizar F. Oliva, Fabienne Hans, and Carlo Petosa

Erratum to:
Chapter 9 in: Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_9

The chapter author provided the below additional text to be added in the acknowledgement section. This has now been updated in the revised version of the book.

"We thank Paul Sauer, Jennifer Timm and Daniel Panne for providing the yeast CAF1 complex."

The updated online version of this chapter can be found at https://doi.org/10.1007/978-1-4939-7759-8_9

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8_32,

INDEX

Joseph A. Marsh (ed.), *Protein Complex Assembly: Methods and Protocols*, Methods in Molecular Biology, vol. 1764, https://doi.org/10.1007/978-1-4939-7759-8, © Springer Science+Business Media, LLC, part of Springer Nature 2018

L

M

N

O

P

Q

R

S

T

X

Y