J Struct Funct Genomics (2007) 8:37–44
DOI 10.1007/s10969-007-9018-3

ORIGINAL PAPER

Functional insights from structural genomics

Farhad Forouhar · Alexandre Kuzin · Jayaraman Seetharaman · Insun Lee · Weihong Zhou · Mariam Abashidze · Yang Chen · Wei Yong · Haleema Janjua · Yingyi Fang · Dongyan Wang · Kellie Cunningham · Rong Xiao · Thomas B. Acton · Eran Pichersky · Daniel F. Klessig · Carl W. Porter · Gaetano T. Montelione · Liang Tong

Received: 8 March 2007 / Accepted: 28 May 2007 / Published online: 23 June 2007

Abstract Structural genomics efforts have produced structural information, either directly or by modeling, for thousands of proteins over the past few years. While many of these proteins have known functions, a large percentage of them have not been characterized at the functional level. The structural information has provided valuable functional insights on some of these proteins, through careful structural analyses, serendipity, and structure-guided functional screening. Some of the success stories based on structures solved at the Northeast Structural Genomics Consortium (NESG) are reported here. These include a novel methyl salicylate esterase with important role in plant innate immunity, a novel RNA methyltransferase (*H. influenzae* yggJ (HI0303)), a novel spermidine/spermine *N*-acetyltransferase (*B. subtilis* PaiA), a novel methyltransferase or AdoMet binding protein (*A. fulgidus* AF_0241), an ATP:cob(I)alamin adenosyltransferase (*B. subtilis* YvqK), a novel carboxysome pore (*E. coli* EutN), a proline racemase homolog with a disrupted active site (*B. melitensis* BME11586), an FMN-dependent enzyme (*S. pneumoniae* SP_1951), and a 12-stranded *β*-barrel with a novel fold (*V. parahaemolyticus* VPA1032).

Keywords Methyltransferase · Spermine · Acetyltransferase

Introduction

Structural genomics efforts sponsored by the Protein Structure Initiative (PSI) at the National Institutes of Health (NIH) have produced experimental structures of more than 2,000 proteins over the past few years [1, 2]. The structures provide templates useful for modeling tens of thousands of protein sequence [3]. While many of these are for proteins with known biological/biochemical functions, a large percentage of them are on proteins for which the functions were not known. In fact, many of these proteins are annotated as 'hypothetical' in the sequence database. As structure is more conserved than sequence over the evolutionary time scale, this structural information can provide evidence for previously unrecognized evolutionary relationships, and hypotheses about biochemical functions. However, a systematic approach for function assignment from structural information is not yet feasible. In most cases, functional insights are derived on a protein-by-protein basis, through careful structural and sequence analyses, serendipity, and structure-guided activity screening.

F. Forouhar · A. Kuzin · J. Seetharaman · I. Lee ·
W. Zhou · M. Abashidze · Y. Chen · W. Yong ·
L. Tong (✉)
Department of Biological Sciences, Northeast Structural
Genomics Consortium, Columbia University, New York, NY
10027, USA
e-mail: ltong@columbia.edu

H. Janjua · Y. Fang · D. Wang · K. Cunningham ·
R. Xiao · T. B. Acton · G. T. Montelione
Center for Advanced Biotechnology and Medicine, Northeast
Structural Genomics Consortium, Rutgers University,
Piscataway, NJ 08854, USA

E. Pichersky
Department of Molecular, Cellular and Developmental Biology,
University of Michigan, Ann Arbor, MI 48109, USA

D. F. Klessig
Boyce Thompson Institute for Plant Research, Tower Road,
Ithaca, NY 14853, USA

C. W. Porter
Department of Pharmacology and Therapeutics, Roswell Park
Cancer Institute, Buffalo, NY 14263, USA

Before crystallization experiments are carried out on a target protein at the Northeast Structural Genomics Consortium (NESG), the protein is routinely searched against the sequence and literature databases to identify its possible function, and more importantly, possible substrates/ligands that can bind to it. Such substrates/ligands are then included in the crystallization trials. It is well accepted that binding of such substrates/ligands can often stabilize the protein, thereby improving the likelihood of crystallization and/or the quality of the crystals. Moreover, the resulting structures of the complexes will provide valuable insights into the functions of the target protein.

In this paper, we will describe several proteins for which the structural information determined by the NESG using X-ray crystallography has provided significant, often unexpected, insights into their biochemical functions. In our published work, we have reported several examples of novel biochemical function discovery using 3D structural information, including identification of a novel aspartate dehydrogenase in bacteria [4], a novel RNA methyltransferase [5], an NAD-dependent oxidoreductase with a novel fold [6], a novel ubiquitin-like modifier [7], an enzyme in phenazine biosynthesis [8], a methyl salicylate esterase with important functions in plant immunity [9], a spermidine/spermine acetyltransferase (SSAT) [10], and a tryptophan dioxygenase [11]. We will briefly summarize some of these discoveries here, and then describe the structures and possible functions of several new proteins. The functional insights derived from the structural information indicate new biochemical/biological experiments that can be pursued on these targets.

Materials and methods

Protein target selection

Protein targets for NESG are selected primarily from protein domain families that constitute the multi-cellular eukaryotic proteomes, as described elsewhere [12]. Many of the domain families include both prokaryotic and eukaryotic members. In most cases, the structures determined by the NESG are prokaryotic members of these very broadly conserved domain families.

Protein expression and purification

The production of recombinant proteins was carried out as part of the high-throughput protein production process of the NESG [13]. The genes of interest were cloned into bacterial expression vectors and over-expressed at 17°C in *Escherichia coli* BL21(DE3) pMGK cells, a rare codon enhanced strain. MJ9 minimal media [14] supplemented with selenomethionine, lysine, phenylalanine, threonine, isoleucine, leucine and valine were used for the production of selenomethionine-labeled proteins [15]. Initial growth was carried out at 37°C until the OD_{600} of the culture reached 0.6–0.8 unit. Protein expression was induced by the addition of isopropyl-β-D-thiogalactopyranoside (IPTG) at a final concentration of 1 mM, the cells were harvested after overnight incubation.

The recombinant proteins were purified by standard methods, using nickel-affinity column (HisTrap^{TM} HP) followed by size-exclusion column (HiLoad^{TM} 26/60 Superdex 75 pg). The purified proteins were concentrated, flash frozen in aliquots, and stored at –80°C. Sample purity and molecular weight were verified by SDS-PAGE and MALDI-TOF mass spectrometry, respectively.

Protein crystallization

Initial conditions for crystallization were determined using data generated for 1,536 crystallization conditions under oil, in microtiter plates, using technologies developed by the High Throughput (HTP) Crystallization Facility at the Hauptman-Woodward Institute [16]. These data were used as starting points for crystallization optimization at room temperature using the hanging-drop vapor diffusion method. Crystals were cryo-protected by soaking in the reservoir solution supplemented with 20–25% (v/v) glycerol or ethylene glycol.

For AF_0241 (NESG ID GR27), the reservoir solution contained 100 mM Na-acetate (pH 4.6) and 18% (w/v) PEG4000. For YvqK (BSU33150, NESG ID SR128), 2 µl of protein solution containing YvqK (10 mg/ml), 5 mM Tris (pH 7), 100 mM NaCl, and 5 mM DTT were mixed with 2 µl of the reservoir solution consisting of 50 mM MES (pH 6.2) and 12% PEG 20K. For EutN (NESG ID ER316), 2 µl of protein solution containing EutN (10 mg/ml), 5 mM Tris (pH 7), 100 mM NaCl, and 5 mM DTT were mixed with 2 µl of the reservoir solution consisting of 100 mM HEPES (pH 7.5) and 23% PEG 20 k. For BME11586 (NESG ID LR31), 2 µl of protein solution containing BME11586 (10 mg/ml), 5 mM Tris (pH 7), 100 mM NaCl, and 5 mM DTT were mixed with 2 µl of the reservoir solution consisting of 20% PEG 3350 and 200 mM potassium thiocyanate.

For SP_1951 (NESG ID SpR27), 2 µl of protein solution containing SP_1951 (10 mg/ml), 5 mM Tris (pH 7.5), 100 mM NaCl, and 5 mM DTT were mixed with 2 µl of the reservoir solution consisting of 10% PEG 3350, 100 mM tartrate, and 200 mM sodium choloride. For VPA1032 (NESG ID VpR44), 2 µl of protein solution containing VPA1032 (10 mg/ml), 5 mM Tris (pH 7.5), 100 mM NaCl, and 5 mM DTT were mixed with 2 µl of the reservoir solution consisting of 100 mM sodium cacodylate (pH 6.5), 20% PEG 8K, and 200 mM magnesium acetate.

Data collection and processing

X-ray diffraction data were collected at the X4A beamline of National Synchrotron Light Source. Both single- and multiple-wavelength anomalous diffraction data were used to solve the structures. The diffraction images were processed and scaled with the HKL package [17]. The data processing statistics can be found in the PDB entries of the structures described here.

Structure determination and refinement

The selenium sites were located with the program SnB [18] and/or Shelx [19]. SOLVE/RESOLVE [20] was used for phasing the reflections and automated model building. The program COMO was used for molecular replacement calculations [21]. Manual model building was carried out with the programs O [22], XtalView [23], and/or Coot [24]. The program CNS was used for structure refinement [25]. The refinement statistics can be found in the PDB entries of the structures described here.

Results and discussion

A novel methyl salicylate esterase with important role in plant innate immunity

Salicylic acid binding protein 2 (SABP2) was identified based on its high affinity for salicylic acid (SA), which is an important signal for plant defense responses [26, 27]. Sequence analysis suggested that SABP2 is a member of the α/β hydrolase superfamily, but its substrate was not known. We determined the structure of SABP2 from tobacco in complex with SA (NESG ID AR2241, PDB entries 1Y7H, 1Y7I, 1XKL) [9]. The structure unexpectedly showed that SA is bound in the active site, indicating that it could be a product after hydrolysis. This structural insight was confirmed by biochemical studies, which demonstrated strong esterase activity for SABP2 towards methyl salicylate and related compounds. The structural and biochemical data suggest a model where SA is transported from the site of infection to other parts of the plant in the form of its methyl ester and SABP2 is crucial for the regeneration of SA and the activation of systemic immune responses.

A novel RNA methyltransferase

The protein yggJ (HI0303) from *Haemophilus influenzae* was annotated as a hypothetical protein of unknown function in the sequence database. It is a member of a widely conserved protein family, present in a large number of prokaryotic genomes as well as in *Arabidopsis thaliana*. The structure of this protein (NESG ID IR73, PDB entry 1NXZ) unexpectedly showed remarkable homology to those of known RNA methyltransferases, including RrmA, MT1, and YibK [28–30], even though they share less than 15% amino acid sequence identity [5]. Detailed structural analyses identified possible binding sites for *S*-adenosylmethionine (AdoMet or SAM) and the RNA substrate in yggJ (HI0303), and we proposed that this protein is a novel RNA methyltransferase [5]. Recent studies confirmed our prediction, showing that yggJ (HI0303) is responsible for the methylation of U1498 in the 16S ribosomal RNA [31].

A novel SSAT

The *pai* operon in *Bacillus subtilis* contains two genes, *paiA* and *paiB*, and is involved in negative control of sporulation as well as other cellular processes [32]. It was suggested that PaiA may be a DNA binding protein, and a putative helix-turn-helix motif was recognized in its sequence [32]. However, this motif actually also corresponds to one of the motifs of *N*-acetyltransferases. Therefore, acetyl-CoA was included in the crystallization solution, which was crucial for obtaining good quality diffraction [10].

The structure of PaiA confirms that it has the *N*-acetyltransferase fold (NESG ID SR64, PDB entry 1TIQ) [10]. Serendipitously, the structure also revealed the binding of an oxidized CoA dimer in the active site. While one CoA molecule is fully ordered and therefore mimics the CoA in the reaction, only the linear portion of the other CoA molecule is ordered, suggesting it may mimic the substrate of the acetylation reaction. Careful structural analyses led to the hypothesis that PaiA may prefer linear, positively charged molecules as substrates, and polyamines are good candidates for such molecules [10]. Biochemical studies showed that PaiA has strong activity towards spermine, spermidine and aminopropylcadaverine, confirming that PaiA is a novel SSAT [10].

Spermidine/spermine acetyltransferase catalyzes the first reaction in both the degradation and the export pathways for polyamines, which have important roles in many biological processes, including DNA binding and stability, chromatin condensation, RNA binding and conformation, mRNA translation, protein binding. Therefore, PaiA may function in regulating intracellular polyamine concentrations and/or binding capabilities.

A novel AdoMet binding protein/methyltransferase

The open reading frame AF_0241 of *Archaeoglobus fulgidus* belongs to a large family of proteins (COG1720, PFAM ID UPF0066) that includes homologs predominantly from prokaryotes, but also from plants, *Drosophila*, and mammals (human, rat, mouse and others). The function of these proteins is not known. The structure of AF_0241 contains a six-stranded anti-parallel β-barrel (NESG ID GR27, PDB entry 2NV4, Fig. 1a). The backbone fold of this β-barrel has been observed in the structures of riboflavin synthase and other proteins [33, 34], with a *Z* score of 5.5, rms distance of 2.0 Å, and sequence identity of 11% based on the program Dali [35]. However, AF_0241 contains many extended loops between the β-strands of this barrel, in contrast to the compact loops in riboflavin synthase. AF_0241 is a dimer, both in solution and in the crystal, through contacts at the sides of the two β-barrels (Fig. 1a).

Unexpectedly, we observed clear electron density for an *S*-adenosylmethionine (AdoMet) molecule in the structure (Fig. 1b). AdoMet was not included in our crystallization solution, suggesting that it may be a natural ligand for AF_0241. The molecule is bound at the top of the β-barrel, where many of the extended loop segments are located (Fig. 1a). In contrast, the binding site in riboflavin synthase is at the side of the β-barrel [34]. The adenine base is sandwiched between the side chains of Met57 and Leu113 (Fig. 1b). The carboxylate group of methionine is recognized by the side chains of Arg82 and Lys122 (from the other monomer), and the ammonium ion is recognized by Gln22.

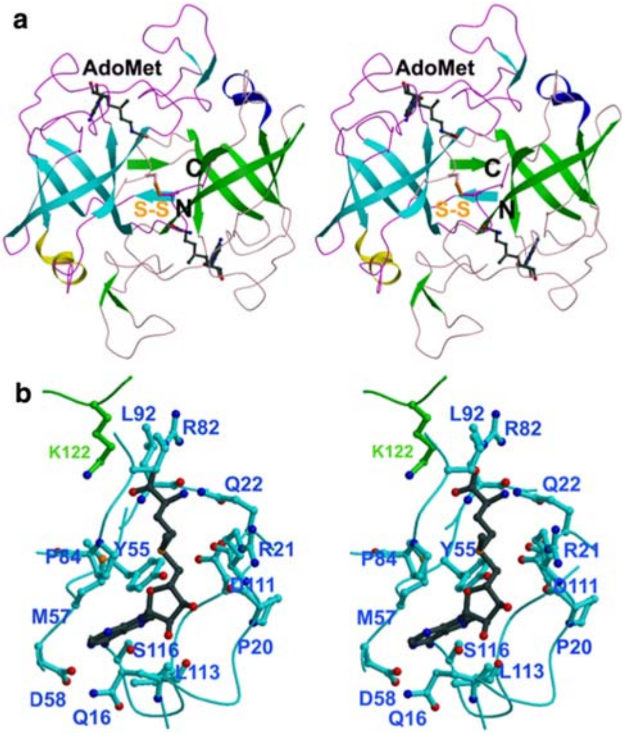

Fig. 1 Structure of AF_0241 from *Archaeoglobus fulgidus* (NESG ID GR27). (**a**) Schematic representation of the AF_0241 dimer. The β-strands, α-helices and loops are colored in cyan, yellow and magenta for one monomer. A fortuitous (non-native) disulfide bond, between the side chains of Cys130 in the two monomers, is indicated. (**b**) Binding mode of AdoMet to AF_0241. All the figures are created with Molscript [50], and rendered with Raster3D [51]

This mode of AdoMet recognition has not been observed before, suggesting that the AF_0241 family of proteins (COG1702, PFAM ID UPF0066) may be novel AdoMet binding proteins. Further studies are needed to demonstrate whether these proteins also possess methyltransferase activity, and if so, what the substrates are for this activity. In many of the homologs, this protein is actually a domain of a larger protein, suggesting that AF_0241 may bind AdoMet and provide this cofactor for methyltransferase activity that is carried out by the other domain(s).

An ATP:cob(I)alamin adenosyltransferase

Bacillus subtilis YvqK is annotated as a hypothetical protein in the sequence database. The crystal structure of YvqK contains six α-helices, five of which are arranged in a helical bundle reminiscent of an ion-channel fold (NESG ID SR128, PDB entry 1RTY, Fig. 2a). The structure reveals a tightly associated trimer, and a phosphate ion is bound in the center of the trimer (Fig. 2a).

The closest structural homolog of YvqK is the TA1434 protein of *Thermoplasma acidophilum* (PDB entry 1NOG) [36], with a Z score of 23.9 and rms distance of 1.2 Å for 147 equivalent Cα atoms based on the program Dali [35]. The two proteins share 42% amino acid sequence identity. TA1434 is an ATP:cob(I)alamin adenosyltransferase (ATR) [36], which transfers the 5′-deoxyadenosyl moiety of ATP to Co(I) of cobalamin (vitamin B12). Crystal structures of human ATR (PDB entry 2IDX) [37] and PduO-type ATP:Co(I)rrinoid adenosyltrasnferase from *Lactobacillus reuteri* (PDB entry 2NT8) [38], both in complex with ATP and Mg^{2+}, have recently been reported. The N-terminal residues of these proteins become ordered in the ATP complex, and residues Arg190, Arg191, and Glu193 in human ATR interact with the tri-phosphate of ATP. As YvqK shares conserved residues with these other enzymes, it is highly likely that YvqK is an ATP:cob(I)alamin adenosyltransferase in *B. subtilis*.

Although residues that interact with ATP have been identified, the mode of cobalamin binding in these enzymes is still unclear. Interestingly, the N-terminal five residues, Met1-Lys2-Leu3-Tyr4-Thr5, in one of the molecules of YvqK in our crystal are well ordered. Moreover, preceding the initiating methionine, a well-ordered electron density that can accommodate a chemical entity of 150–300 Da is also observed, although its exact identity remains to be

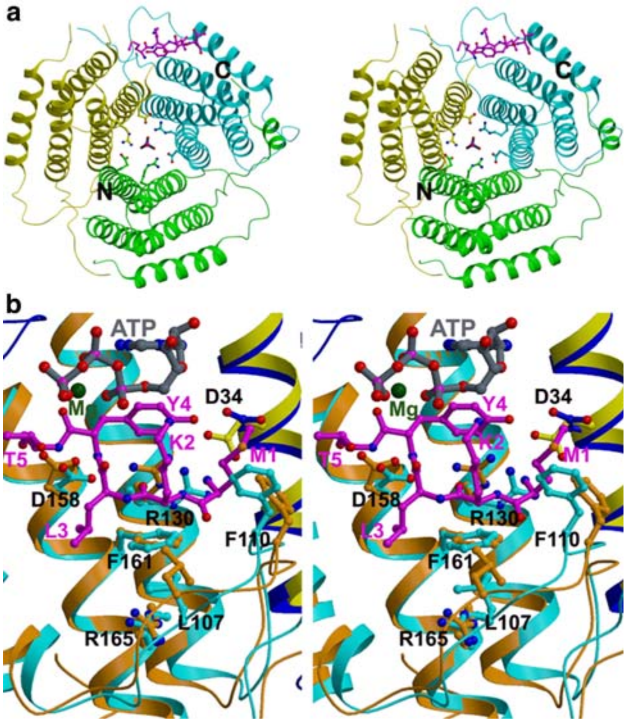

Fig. 2 Structure of YvqK from *Bacillus subtilis* (NESG ID SR128) (**a**) Schematic representation of the YvqK trimer. α-helices and loops are colored in yellow for molecule A, cyan for molecule B, and green for molecule C. The N-terminal 5 residues are shown as ball-and-stick models in magenta. (**b**) Structural overlay of the active sites of YvqK (molecules B and A, in cyan and yellow) and human ATR (molecules A and C, darkorange and blue) [37]. The N-terminal five residues of molecule B of YvqK are shown as ball-and-stick models in magenta. ATP and Mg^{2+} are shown as ball-and-stick in gray and darkgreen, respectively

determined (Fig. 2b). These N-terminal residues and the attached chemical group reside in a cavity next to ATP, as shown in the structural overlay of YvqK with human ATR (Fig. 2b). It is likely that this is the binding site for the cobalamin. Several highly conserved residues (Asp34, Leu107, Arg126, Asp158, Arg165, Phe170, and Phe221) may be important for this binding (Fig. 2b).

A new carboxysome pore

Escherichia coli ethanolamine utilization protein EutN has been annotated as an EutN/carboxysome structural protein CcmL. The crystal structure of EutN contains a central five-stranded β-barrel, with an α-helix at the open end of this barrel (NESG ID ER316, PDB entry 2HD3, Fig. 3a). The structure also contains three additional β-strands, which helps the formation of a tight hexamer, with a hole in the center (Fig. 3b). The structural information therefore suggests that EutN forms a pore, with an opening of 26 Å in diameter on one face (Fig. 3b) and 14 Å on the other face (Fig. 3c). A similar structure has recently been reported on another carboxysome shell protein, CcmK2, which also forms a hexameric pore [39]. Unlike CcmK2, whose pore is surrounded by basic residues, the large opening of EutN is predominantly covered by acidic residues (Fig. 3b). Therefore, it is possible that EutN/CcmL is a pore that is involved in the transport of positively-charged molecules for the carboxysome or another compartment.

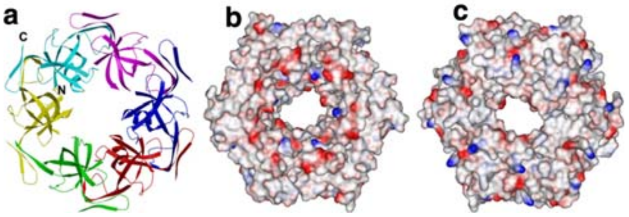

Fig. 3 Structure of EutN from *E. coli* (NESG ID ER316). (**a**) Schematic representation of the EutN homo-hexamer. The six monomers are colored in yellow, cyan, magenta, blue, red, and green. (**b**) Surface charge representation of the front face of the EutN pore. (**c**) Surface charge representation of the back face of the EutN pore

In *E. coli* and *Salmonella typhimurium*, *EutN* is one of 17 genes in the *eut* operon, which is involved in the cobalamin-dependent degradation of ethanolamine [40]. The operon also encodes an ATP:cob(I)alamin adenosyltransferase, although it shares no sequence homology with the YvqK protein described in the previous section.

A different biochemical activity for a proline racemase homolog

BME11586 from *Brucella melitensis* has been annotated as a proline racemase in the sequence database. The proline racemase from *Trypanosoma cruzi*, the causative agent of Chagas disease, is crucial for its evasion of host immune responses and therefore the enzyme is an attractive target for drug discovery against this pathogen [41]. The crystal structure of BME11586 contains five α-helices and 20 β-strands (NESG ID LR31, PDB entry 1TM0, Fig. 4a). The overall structure of the protein has the double hot-dog fold, each of which contains one helix wrapped around by nine β-strands. An additional β-strand, at the N-terminus, helps the tetramerization of this protein (Fig. 4a).

The closest structural homolog of BME11586 is the proline racemase from *Trypanosoma cruzi* (TcPRACA) [42], with rms distance of 2.2 Å for 306 residues and 30% sequence identity. However, TcPRACA is only a dimer. Most importantly, the two catalytic cysteine residues in TcPRACA, Cys130 and Cys300, are replaced by a serine and a threonine residue, Ser90 and Thr255, in BME11586 (Fig. 4b). Therefore, it is probably unlikely that BME11586 can possess proline racemase activity. The

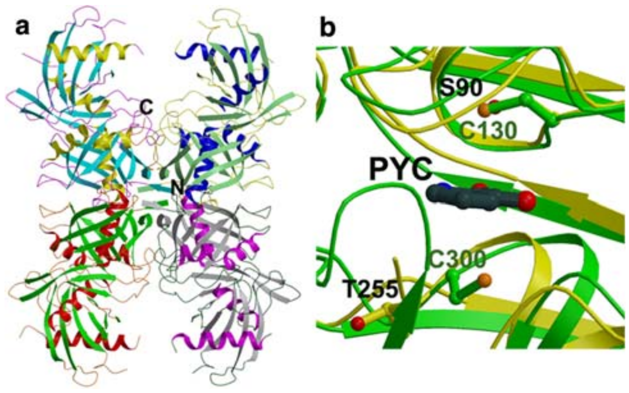

Fig. 4 Structure of BME11586 from *Brucella melitensis* (NESG ID LR31). (**a**) Schematic representation of the BME11586 tetramer. α-helices, β-strands, loops are colored in yellow, cyan, and magenta for molecule A, blue, palegreen, and yellow for molecule B, magenta, lightgray, and darkgreen for molecule C, red, green, darkorange for molecule D. (**b**) Structural overlay of the active sites of BME11586 (yellow) and TcPRACA (green) [42]. The inhibitor pyrrole-2-carboxylate (PYC) is shown as a ball-and-stick model in darkgray. The side chains of S90 and T255 (both in yellow) of BME11586, and C130 and C300 (both in green) of TcPRACA, are shown as ball-and-stick models

catalytic activity and the substrate of BME11586 remain to be determined.

An FMN-dependent enzyme

SP_1951 from *Streptococcus pneumoniae* belongs to the tryptophan repressor binding protein (WrbA) superfamily. WrbA was originally co-purified and co-immunoprecipitated with the tryptophan repressor protein TrpR [43], although later studies revealed that WrbA had no specific effect on the DNA binding ability of TrpR [44]. The crystal structure of SP_1951 contains five β-strands and nine α-helices that are arranged in two domains, an N-terminal Rossmann-fold domain comprised of a parallel five-stranded β-sheet surrounded by six helices and a C-terminal α-helical domain consisting of three helices (NESG ID SpR27, PDB entries 1SQS and 2OYS, Fig. 5a). SP_1951 is a tightly associated dimer (Fig. 5a), supported by our light scattering studies in solution. FMN is bound at the dimer interface, between the Rossmann-fold of one monomer and the C-terminal α-helical domain of the other monomer (Fig. 5a). There are essentially no conformational changes in the enzyme upon FMN binding.

Closest structural homologs of SP_1951 include two WrbA proteins from other bacteria [45], as well as several other flavoproteins. The WrbA proteins from *D. radiodurans* and *P. aeruginosa* are tetrameric, but it maintains a similar dimer interface as that seen for SP_1951. However, the C-terminal helical domain observed in SP_1951 is much smaller in WrbA.

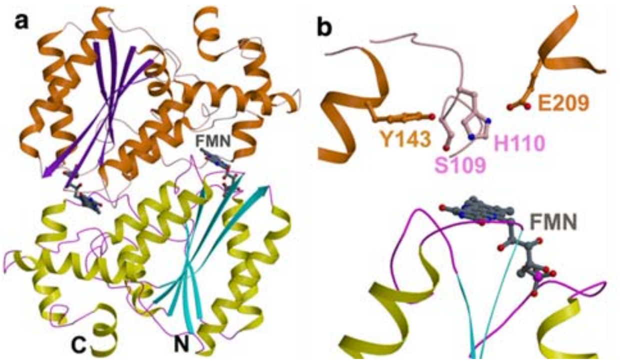

Fig. 5 Structure of SP_1951 from *Streptococcus pneumoniae* (NESG ID SpR27). (**a**) Schematic representation of the SP_1951 dimer. α-helices, β-stands, and loops are colored in yellow, cyan and magenta for one monomer. (**b**) The putative active site of SP_1951. The side chains of residues that possibly interact with the unknown substrate are shown as ball-and-stick models

Inspection of the residues near FMN, especially those from the C-terminal α-helical domain, reveals that the cavity near the flavin moiety of FMN is predominantly hydrophobic, suggesting that the natural substrate of SP_1951 is likely hydrophobic in nature. The side chains of Ser109, His110, Tyr143 and Glu209 are located near the flavin moiety of FMN (Fig. 5b), and they may be important for the catalysis by this protein.

A novel β-barrel fold

The crystal structure of VPA1032 from *Vibrio parahaemolyticus* contains 14 β-strands and nine α-helices that are arranged in two domains (NESG ID VpR44, PDB entry 1ZBP, Fig. 6), an N-terminal tetratricopeptide repeat (TPR) domain comprising five helices and a C-terminal domain consisting of the remaining four α-helices and all

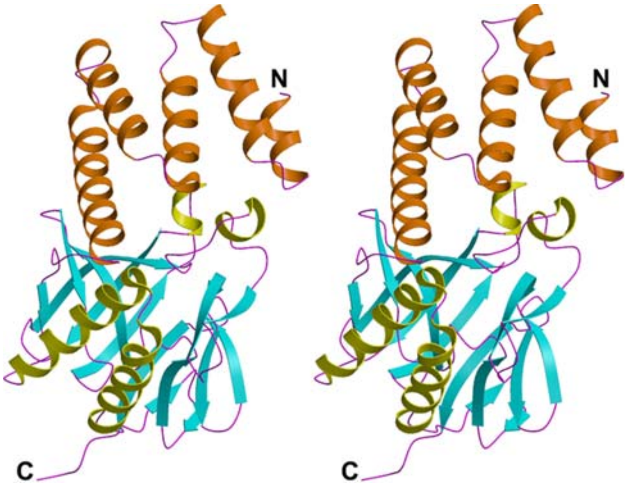

Fig. 6 Crystal structure of VPA1032 from *Vibrio parahaemolyticus* (NESG ID VpR44)

the β-strands. The 12-stranded β-barrel in the C-terminal domain does not appear to have a close structural homolog, based on the program Dali [35], suggesting that it may represent a new fold. VPA1032 is predominantly monomeric in solution, although a trimeric species can also be observed (data not shown). A loosely associated trimer is present in the crystal, but the functional relevance of this trimerization remains to be determined.

VPA1032 is a member of the impaired-in-nitrogen-fixation (ImpE) protein family, sharing 30% sequence identity with ImpE from *Rhizobium leguminosarum*, which is part of an operon with 14 different genes [46]. A close homolog has recently been identified as one of the virulence genes of the pathogenic bacterium *Vibrio vulnificus* [47]. The structural information on VPA1032 may help the identification of the function of this family of proteins.

In summary, the cases described here demonstrate that crystal structures can be a powerful source for functional insights. At the same time, successful functional assignment based on structural information requires careful structural analysis and activity screening, and sometimes serendipity, especially considering the fact that proteins with similar structures can have different functions (or substrates). With the large number of structures that have been determined, it is becoming increasingly common that a protein of interest belongs to a family for which structural information is already available. However, it is oftentimes much more difficult to define the exact substrate/function of the individual proteins of a family. Attempts at predicting the substrate/function of a protein by systematically analyzing its structure are having success in some cases [48, 49], but this remains an extremely challenging task in general. Careful structural analysis coupled with direct experimental studies probably represent the best approach for successful functional annotations at the present time.

Acknowledgements We thank Randy Abramowitz and John Schwanof at the NSLS for setting up the beamlines; G. DeTitta of Hauptman Woodward Research Institute for crystallization screening. This research was supported by grants from the Protein Structure Initiative of the National Institutes of Health (P50 GM62413 and U54 GM074958).

References

1. Chandonia J-M, Brenner SE (2006) Science 311:347
2. Terwilliger TC (2004) Nature Struct Mol Biol 11:296
3. Liu J, Montelione GT, Rost B (2007) Nat Biotech in press
4. Yang Z, Savchenko A, Yakunin A, Zhang R, Edwards A, Arrowsmith C, Tong L (2003) J Biol Chem 278:8804
5. Forouhar F, Shen J, Xiao R, Acton TB, Montelione GT, Tong L (2003) Proteins 53:329
6. Forouhar F, Lee I, Benach J, Kulkarni K, Xiao R, Acton TB, Montelione GT, Tong L (2004) J Biol Chem 279:13148
7. Cort JR, Chiang Y, Zheng D, Montelione GT, Kennedy MA (2002) Proteins 48:733
8. Blankenfeldt W, Kuzin AP, Skarina T, Korniyenko Y, Tong L, Beyer P, Janning P, Thomashow LS, Mavrodi DV (2004) Proc Natl Acad Sci USA 101:16431
9. Forouhar F, Yang Y, Kumar D, Yang C, Fridman Y, Park SW, Chiang Y, Acton TB, Montelione GT, Pichersky E, Klessig DF, Tong L (2005) Proc Natl Acad Sci USA 102:1773
10. Forouhar F, Lee I, Vujcic S, Vujcic J, Shen J, Vorobiev SM, Xiao R, Acton TB, Montelione GT, Porter CW, Tong L (2005) J Biol Chem 280:40328
11. Forouhar F, Anderson JLR, Mowat CG, Vorobiev SM, Hussain A, Abashidze M, Bruckmann C, Thackray SJ, Seetharaman J, Tucker T, Xiao R, Ma L-C, Zhao L, Acton TB, Montelione GT, Chapman SK, Tong L (2007) Proc Natl Acad Sci USA 104:473
12. Wunderlich Z, Acton TB, Liu J, Kornhaber G, Everett J, Carter P, Lan N, Echols N, Gerstein M, Rost B, Montelione GT (2004) Proteins 56:181
13. Acton TB, Gunsalus K, Xiao R, Ma L, Aramini J, Baron MC, Chiang Y, Clement T, Cooper B, Denissova N, Douglas S, Everett JK, Palacios D, Paranji RH, Shastry R, Wu M, Ho C-H, Shih L, Swapna GVT, Wilson M, Gerstein M, Inouye M, Hunt JF, Montelione GT (2005) Methods Enzymol 394:210
14. Jansson M, Li Y-C, Jendeberg L, Anderson S, Montelione GT, Nilsson B (1996) J Biomol NMR 7:131
15. Doublie S, Kapp U, Aberg A, Brown K, Strub K, Cusack S (1996) FEBS Lett 384:219
16. Luft JR, Collins RJ, Fehrman NA, Lauricella AM, Veatch CK, DeTitta GT (2003) J Struct Biol 142:170
17. Otwinowski Z, Minor W (1997) Method Enzymol 276:307
18. Weeks CM, Miller R (1999) J Appl Cryst 32:120
19. Sheldrick GM (1990) Acta Cryst A46:467
20. Terwilliger TC (2003) Meth Enzymol 374:22
21. Jogl G, Tao X, Xu Y, Tong L (2001) Acta Cryst D57:1127
22. Jones TA, Zou JY, Cowan SW, Kjeldgaard M (1991) Acta Cryst A47:110
23. McRee DE (1999) J Struct Biol 125:156
24. Emsley P, Cowtan KD (2004) Acta Cryst D60:2126
25. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang J-S, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Acta Cryst D54:905
26. Du H, Klessig DF (1997) Plant Phyisol 113:1319
27. Kumar D, Klessig DF (2003) Proc Natl Acad Sci USA 100:16101
28. Nureki O, Shirouzu M, Hashimoto K, Ishitani R, Terada T, Tamakoshi M, Oshima T, Chijimatsu M, Takio K, Vassylyev DG, Shibata T, Inoue Y, Kuramitsu S, Yokoyama S (2002) Acta Cryst D58:1129
29. Zarembinski TI, Kim Y, Peterson K, Christendat D, Dharamsi A, Arrowsmith CH, Edwards A, Joachimiak A (2003) Proteins 50:177
30. Lim K, Zhang H, Tempczyk A, Krajewski W, Bonander N, Toedt J, Howard A, Eisenstein E, Herzberg O (2003) Proteins 51:56
31. Basturea GN, Rudd KE, Deutscher MP (2006) RNA 12:426
32. Honjo M, Nakayama A, Fukazawa K, Kawamura K, Ando K, Hori M, Furutani Y (1990) J Bacteriol 172:1783
33. Liao D-I, Wawrzak Z, Calabrese JC, Viitanen PV, Jordan DB (2001) Structure 9:399
34. Gerhardt S, Schott AK, Kairies N, Cushman M, Illarionov B, Eisenreich W, Bacher A, Huber R, Steinbacher S, Fischer M (2002) Structure 10:1371
35. Holm L, Sander C (1993) J Mol Biol 233:123
36. Saridakis V, Yakunin A, Xu X, Anandakumar P, Pennycooke M, Gu J, Cheung F, Lew JM, Sanishvili R, Joachimiak A, Arrowsmith CH, Christendat D, Edwards AM (2004) J Biol Chem 279:23646
37. Schubert HL, Hill CP (2006) Biochem 45:15188
38. St Maurice M, Mera PE, Taranto MP, Sesma F, Escalante-Semerena JC, Rayment I (2007) J Biol Chem 282:2596

39. Kerfeld CA, Sawaya MR, Tanaka S, Nguyen CV, Phillips M, Beeby M, Yeates TO (2005) Science 309:936
40. Kofoid E, Rappleye C, Stojiljkovic I, Roth J (1999) J Bacteriol 181:5317
41. Reina-San-Martin B, Degrave W, Rougeot C, Cosson A, Chamond N, Cordeiro-da-Silva A, Arala-Chaves M, Coutinho A, Minoprio P (2000) Nat Med 6:890
42. Buschiazzo A, Goytia M, Schaeffer F, Degrave W, Shepard W, Gregoire C, Chamond N, Cosson A, Berneman A, Coatnoan N, Alzari PM, Minoprio P (2006) Proc Natl Acad Sci USA 103:1705
43. Yang W, Ni L, Somerville RL (1993) Proc Natl Acad Sci USA 90:5796
44. Grandori R, Khalifah P, Boice JA, Fairman R, Giovanielli K, Carey J (1998) J Biol Chem 273:20960
45. Gorman J, Shapiro L (2005) Prot Sci 14:3004
46. Bladergroen MR, Badelt K, Spaink HP (2003) Mol Plant Microbe Interact 16:53
47. Kim YR, Lee SE, Kim CM, Kim SY, Shin EK, Shin DH, Chung SS, Choy HE, Progulske-Fox A, Hillman JD, Handfield M, Rhee JH (2003) Infect Immun 71:5461
48. Binkowski TA, Joachimiak A, Liang J (2005) Prot Sci 14:2972
49. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM (2006) Proteins 62:479
50. Kraulis PJ (1991) J Appl Cryst 24:946
51. Merritt EA, Bacon DJ (1997) Meth Enzymol 277:505