

Short Communication

Solution structure of a soluble fragment derived from a membrane protein by shotgun proteolysis**Mark D. Allen¹, Mary Christie^{2,3}, Peter Jones¹, Benjamin T. Porebski⁴, Brendan Roome^{2,3}, Stefan M.V. Freund¹, Ashley M. Buckle⁴, Mark Bycroft¹, and Daniel Christ^{2,3,*}**

¹MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK, ²Department of Immunology, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, Sydney, NSW 2010, Australia, ³Faculty of Medicine, St Vincent's Clinical School, The University of New South Wales, Darlinghurst, Sydney, NSW 2010, Australia, and ⁴Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3800, Australia

*To whom correspondence should be addressed. E-mail: d.christ@garvan.org.au

Edited by Miroslaw Cygler

Received 20 October 2014; Revised 16 February 2015; Accepted 9 March 2015

Abstract

We have previously reported a phage display method for the identification of protein domains on a genome-wide scale (shotgun proteolysis). Here we present the solution structure of a fragment of the *Escherichia coli* membrane protein yrfF, as identified by shotgun proteolysis, and determined by NMR spectroscopy. Despite the absence of computational predictions, the fragment formed a well-defined beta-barrel structure, distantly falling within the OB-fold classification. Our results highlight the potential of high-throughput experimental approaches for the identification of protein domains for structural studies.

Key words: membrane proteins, NMR spectroscopy, phage display, protein domains, proteolysis

Introduction

The majority of proteins in nature are multi-domain proteins consisting of several independently folding units of structure (Liu and Rost, 2004). Domain boundaries can sometimes be predicted from sequence information alone, based on motifs or through homology with domains of known three-dimensional (3D) structure (Kelley *et al.*, 2000; Gough *et al.*, 2001; Shi *et al.*, 2001; Letunic *et al.*, 2012; Punta *et al.*, 2012). Computational approaches have thereby allowed the assignment of structural information to approximately half of all protein-coding sequences (Chothia *et al.*, 2003; Madera *et al.*, 2004). However, the domain structure of the remainder of the proteome largely remains unknown.

The limitations of computational approaches have inspired the development of experimental approaches for the identification of protein domains (Hart and Waldo, 2013). Strategies include genetic fusion with reporter proteins such as green fluorescent protein, dihydrofolate reductase or beta-lactamase (Cabantous and Waldo, 2006; Dyson

et al., 2008; D'Angelo *et al.*, 2011; Pedelacq *et al.*, 2011), and the use of proteolysis for high-throughput domain selection (shotgun proteolysis) (Christ and Winter, 2006). Shotgun proteolysis is based on random DNA fragmentation, followed by display of the encoded polypeptides on phage, and selection for protease resistance (Christ and Winter, 2006). Based on the classic limited proteolysis method (Porter, 1959), the phage method is capable of identifying segments of structure that strongly correlate with protein domains as defined by bioinformatics predictions (Gough *et al.*, 2001; Christ and Winter, 2006). However, in addition to proteins of known structure, our previous analysis of the *Escherichia coli* proteome by shotgun proteolysis had also identified a large number of fragments of unknown structure (Christ and Winter, 2006). Many of these fragments had been derived from poorly characterized membrane or membrane-associated proteins, including a segment of the *E.coli* protein yrfF (Christ and Winter, 2006). Here we report the solution structure of this fragment as determined by NMR spectroscopy.

Materials and methods

Shotgun proteolysis selection

Segments of the *E.coli* W3110 genome (Bachmann, 1972) were selected by shotgun proteolysis as previously described (Christ and Winter, 2006). In brief, genomic DNA was fragmented to 300–1200 bp using a Nebulizer device (Invitrogen), repaired with Pfu

polymerase and DNA adapters were ligated to the ends of the fragmentation products. Fragments were then cloned into phagemid vector pW656 and electroporated into *E.coli* TG1 (Gibson, 1984) to yield a library of 3×10^7 clones. After rescue with KM13 helper phage (Kristensen and Winter, 1998), phages were incubated with 20 µg/ml of trypsin protease at 10°C for 10 min in TBS buffer supplemented

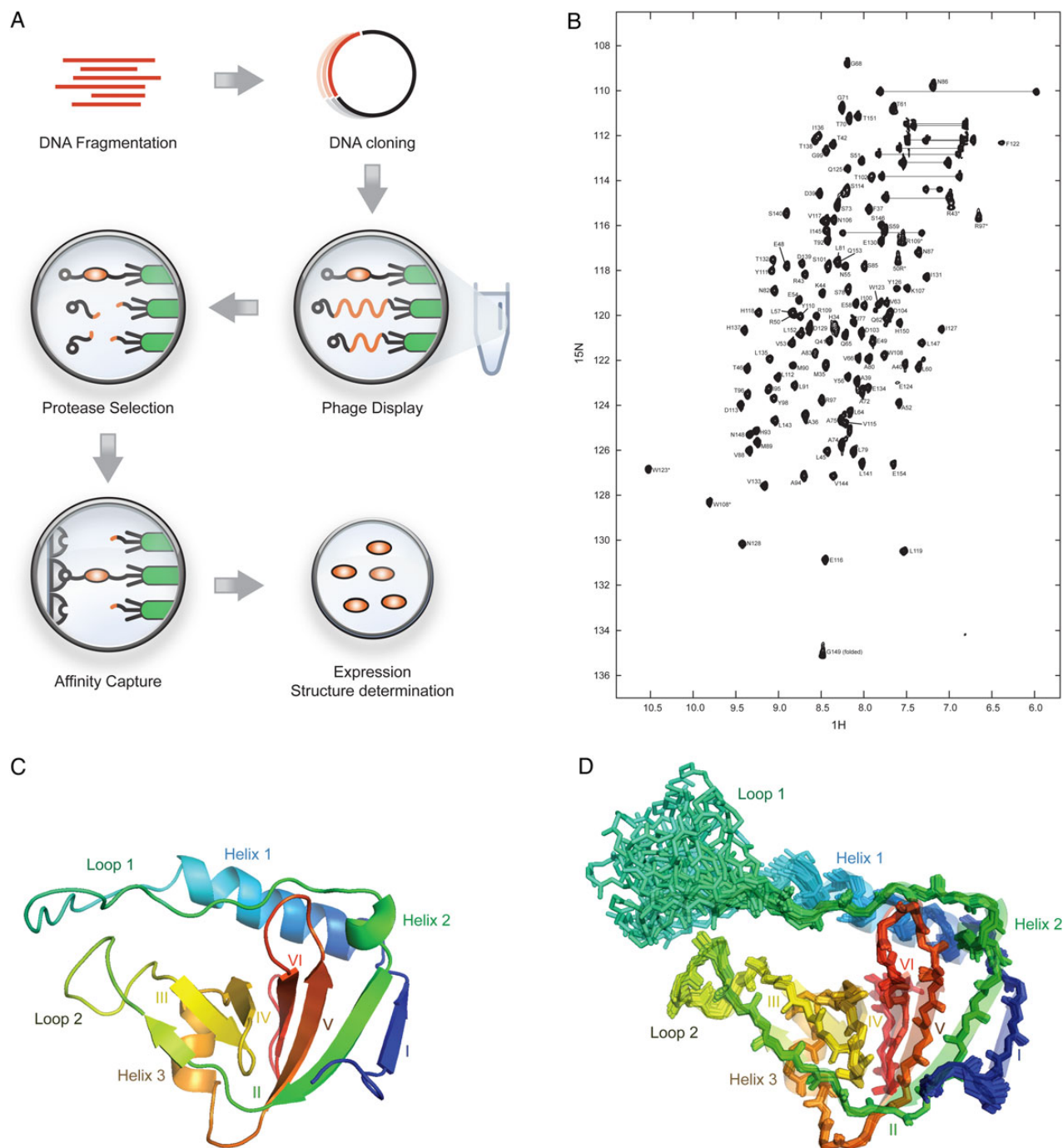


Fig. 1 (A) Selection of protease resistant polypeptides by shotgun proteolysis. A library of DNA fragments were cloned into a phagemid vector, displayed on phage and protease resistant polypeptides captured using an N-terminal affinity tag. (B) 2D $[^{15}\text{N}-^1\text{H}]$ HSQC spectrum of *E.coli* yrfF (residues 36–154) recorded at pH 7.0 and 293 K. The spectrum was recorded on a Bruker Avance 600 MHz spectrometer with 1024 and 512 complex points along the t_1 and t_2 dimensions, respectively, at a protein concentration of 1.5 mM in 95% H_2O and 5% D_2O . Peaks are labeled with single-letter amino-acid code followed by their sequence number. (C) NMR structure of *E.coli* yrfF (residues 36–154; PDB ID 4UZM) shown in cartoon representation. (D) Ensemble of backbone atoms from 20 NMR-derived structures in stick representation. Images were generated using PyMOL. Secondary structure elements are highlighted.

with calcium (25 mM Tris–HCl, 137 mM NaCl, 1 mM CaCl₂, pH 7.4). Protease resistant clones were captured utilizing an N-terminal barnase affinity tag, used to infect *E. coli* TG1, and analyzed by DNA sequencing.

Expression and purification of a shotgun proteolysis fragment

For further characterization, one of the selected fragments (residues 37–154 of the *E. coli* open reading frame gil16131275|refl NP_417857.1| encoding the putative membrane protein yrfF, clone 14, see supplementary information of Christ and Winter (2006)) was cloned into a modified pRSETa vector (Dodd *et al.*, 2004) containing a TEV cleavage site to allow for removal of the poly-histidine tag. An additional alanine residue was incorporated at the N-terminus to facilitate TEV cleavage, resulting in a final construct incorporating residues 36–154 of yrfF.

Protein was expressed in C41 cells grown either in 2xYT medium or in K-MOPS minimal media (Neidhardt *et al.*, 1974) containing ¹⁵NH₄Cl and/or [¹³C]-glucose for the production of isotopically labeled protein. Proteins were purified using Ni-NTA affinity chromatography. Following TEV cleavage and depletion by Ni-NTA affinity chromatography, a final purification step was carried out using a Superdex 75 gel-filtration column, yielding ~30 mg of soluble protein per liter of shaking flask culture.

NMR spectroscopy

Protein samples were prepared for NMR spectroscopy experiments at a concentration of 1.5 mM in 90% H₂O, 10% D₂O, containing 20 mM potassium phosphate, pH 6.5, 100 mM NaCl and 5 mM β-mercaptoethanol. Spectra were acquired using Bruker DRX800 or DRX600 spectrometers equipped with pulsed field gradient triple resonance at 25°C, and referenced relative to external sodium 2,2-dimethyl-2-silapentane-5-sulfonate for proton and carbon signals, or liquid ammonia for that of nitrogen. Assignments were obtained using standard NMR methods using ¹³C/¹⁵N-labeled, ¹⁵N-labeled, 10% ¹³C-labeled and unlabeled protein samples (England and Wand, 1987; Bax *et al.*, 1991). Backbone assignments were obtained using the following standard set of two-dimensional (2D) and 3D heteronuclear spectra: ¹H–¹⁵N HSQC (Fig. 1), HNCACB, CBCA(CO)NH, HNCACO, HNCO, HBHA(CO)NH and ¹H–¹³C HSQC. Additional assignments were made using 2D TOCSY and DQF-COSY spectra. A set of distance constraints were derived from 2D NOESY spectra recorded from a 1.5 mM samples with a mixing time of 100 ms. Hydrogen bond constraints were included for a number of backbone amide protons whose signals were still detected after 10 min in a 2D ¹H–¹⁵N HSQC spectrum recorded in D₂O (pH 6.5). Candidates for the acceptors were identified using the program HBPLUS for the hydrogen bond donors that were identified by the H–D exchange experiments. When two or more candidates of acceptors were found for the same donor in different structures, the most frequently occurring candidate was selected. For hydrogen bond partners, two distance constraints were used where the distance ^(D)H–O^(A) corresponded to 1.5–2.5 Å and ^(D)N–O^(A) to 2.5–3.5 Å. Torsional angle constraints were obtained from an analysis of C', N, C_α, H_α and C_β chemical shifts using the program TALOS (Cornilescu *et al.*, 1999). The stereospecific assignments of H_β resonances determined from DQF-COSY and HNHB spectra were confirmed by analyzing the initial ensemble of structures. Stereospecific assignments of H_γ and H_δ resonances of Val and Leu residues, respectively, were assigned using a fractionally ¹³C-labeled protein sample (Neri *et al.*, 1989).

The 3D structures of the yrfF domain were calculated using the standard torsion angle dynamics-simulated annealing protocol in the program CNS 1.2 (Brunger, 2007). Structures were accepted where no distance violation was >0.25 Å and no dihedral angle violations >5° (20 accepted structures) (Table I). The backbone dynamics of the yrfF domain were investigated using steady-state {¹H–¹⁵N} nuclear Overhauser enhancement (NOE) experiments (Kay *et al.*, 1989; Barbato *et al.*, 1992). Final coordinates have been deposited in the Protein Data Bank (PDB accession no. 4UZM).

Sequence analysis and modeling

Iterative Hidden Markov Model (HMM) searches were performed using the JackHMMER (Finn *et al.*, 2011) web server against the UniProtKB (Magrane and Consortium, 2011) sequence database until convergence was reached. Redundancy reduction was performed with CD-HIT (Huang *et al.*, 2010). Phylogenetic trees were generated using PhyML (Guindon *et al.*, 2010). Structure predictions were performed using I-TASSER (Yang *et al.*, 2014) and EVfold (Marks *et al.*, 2011). The alignment for EVfold was built with a 99.9% redundancy-reduced alignment from the HMM search results.

Results

Structural features of the yrfF shotgun proteolysis fragment

Initial analyses revealed that the yrfF fragment identified by shotgun proteolysis (Fig. 1A) was soluble, expressed at high levels in bacteria and displayed a well-dispersed NMR ¹H–¹⁵H HSQC spectrum (Fig. 1B). Further analyses revealed that a compact domain composed of a six-stranded beta-barrel (Fig. 1C). The barrel is assembled from beta-strands formed by residues 41–44 (strand I), 87–98 (strand II), 109–112 (strand III), 115–118 (strand IV), 131–137 (strand V) and

Table I. Summary of conformational constraints and statistics

| | |
|--|-------------------|
| Structural constraints | |
| Intra-residue | 992 |
| Sequential | 678 |
| Medium-range ($2 \leq i-j \leq 4$) | 411 |
| Long-range ($ i-j > 4$) | 940 |
| Dihedral angle constraints | 36 |
| TALOS constraints | 188 |
| Distance constraints for 44 hydrogen bonds | 88 |
| Total | 3333 |
| Statistics for accepted structures | |
| Statistical parameters (±SD) | |
| RMS deviation for distance constraints | 0.0093 ± 0.0004 Å |
| RMS deviation for dihedral constraints | 0.182 ± 0.027° |
| Mean CNS energy term (kcal mol ⁻¹ ± SD) | |
| E (overall) | 200.91 ± 7.47 |
| E (van der Waals) | 60.95 ± 3.60 |
| E (distance constraints) | 20.21 ± 1.70 |
| E (dihedral and TALOS constraints) | 1.38 ± 0.40 |
| RMS deviations from the ideal geometry (±SD) | |
| Bond lengths | 0.0020 ± 0.0001 Å |
| Bond angles | 0.422 ± 0.0070° |
| Improper angles | 0.358 ± 0.013° |
| Average atomic RMSD from the mean structure (±SD) | |
| Residues 36–154 (N, C _α , C atoms) | 0.330 ± 0.078 Å |
| Residues 36–154 (all heavy atoms) | 0.705 ± 0.054 Å |

N- and C-termini of the domain are both highly defined, while residues within the loop region connecting helices 1 and 2 (residues 63–82) are not well defined and display an absence of long-range NOE restraints (Fig. 1D). Analysis of $\{^{15}\text{N}\text{--}^1\text{H}\}$ -nuclear Overhauser enhancement values revealed that this region is indeed dynamic (Fig. 2A).

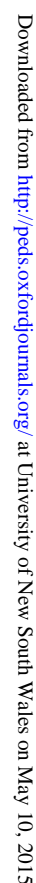


Fig. 2 (A) Plot of the fractional ^1H - ^{15}N heteronuclear backbone enhancement of *E. coli* yrfF (residues 36–154). Structural models of the yrfF fragment (residues 36–154) predicted from **(B)** I-TASSER and **(C)** EVfold shown in gray. For comparison, the experimental NMR structure is shown. **(D)** Secondary structure of the yrfF fragment (residues 36–154) as predicted by PsiPred and determined from the NMR ensemble. C = coil, S = beta strand, H = alpha helix.

Computational approaches fail to accurately predict the yrff fragment structure

Although our earlier analysis of the *E. coli* genome had not revealed any structural assignments (Christ and Winter, 2006), we decided to investigate whether more recent computational methods would be capable of predicting the yrff fragment structure. For this purpose, we utilized I-TASSER a homology approach based on HMMs (Yang *et al.*, 2014) and the evolutionary sequence variation method EVfold (Marks *et al.*, 2011). However, the two computational methods resulted in the generation of structural models, which not only differed considerably between each other, but also showed little resemblance to the structure of the yrff fragment as determined by NMR spectroscopy (Fig. 2B and C). Indeed, both methods failed to predict any sizeable part of the structure with large deviations observed (root-mean-square deviations of 10.5 Å (I-TASSER) and 14.6 Å (EVfold) over 119 C-alpha atoms). In contrast, secondary structure prediction of the yrff fragment was broadly in line with experimental observations, although PsiPred (Jones, 1999) failed to assign several shorter elements (Fig. 2D).

Structural analyses indicate that the yrff fragment is a distant member of the OB-fold family

The absence of accurate predictions for the yrff fragment suggested that it might form an 'orphan fold', with the structure reported here representing the sole representative in the PDB. However, it could also be envisaged that structural similarities might be simply be 'hidden' due to low sequence similarity and evolutionary divergence. To investigate this question, the searches were carried out using the Dali server and the solution structure of the yrff fragment, to identify structural similarities with other proteins (Dietmann *et al.*, 2001). These analyses revealed distant, but detectable similarities to a range of OB-fold containing proteins, including the aspartyl-tRNA synthetase from *Sulfolobus tokodaii* (Fig. 3A and B).

The OB-fold represents a common structural class and is observed in proteins derived from all kingdoms of life; it consists of a

five-stranded beta-barrel capped by an α -helical element (Arcus, 2002). This canonical structure can be observed in strands II–VI of the yrff domain, which form an OB-fold with helix 3 serving as the capping helix. OB-folds can bind to a wide variety of biological molecules including proteins, nucleic acids and carbohydrates; this functional divergence can hinder their recognition based on sequence conservation alone (Arcus, 2002). Despite binding to a diverse range of molecules, ligands are generally bound at a common face of the OB-fold, which in the case of the yrff domain corresponds to β -strands III and IV, and the loops connecting β -strands II–III, III–IV and V–VI.

Despite broadly falling with the OB-fold class, several features of the yrff fragment display considerable variation from previously reported structures. This structural divergence is reflected by overall low Dali Z-scores (≤ 4.1). In particular, extended N-terminal structural features formed by strand I, helix 1 and the loop 1 region represent non-canonical variations of the OB-fold (Fig. 3A). This terminal extension (residues 36–86) embraces one side of the beta-barrel burying ~ 1400 Å² of the domain surface. The interaction shields hydrophobic residues on strands II, V and VI from solvent, and covers the end of the beta-barrel on the opposite face of the capping helix.

Sequence analyses and structural assignments

In order to identify sequence homologs of the yrff fragment, we performed iterative HMM searches against the UniprotKB database (Magrane and Consortium, 2011). This revealed a set of 5401 homologous sequences. After removal of highly conserved sequences (90% identity cutoff), a total of 145 non-redundant protein sequences were identified. These protein sequences are largely classified as members of the IgaA family and predominantly derived from gram-negative bacteria (Supplementary Fig. S1). They are devoid of structural annotations within the yrff homology region, as determined by HMM searches of the PDB (Bernstein *et al.*, 1978), UniProt KB (Magrane and Consortium, 2011) and Pfam (PF07095) (Finn *et al.*, 2014) (excluding the structure reported here). In contrast, experimental determination of the yrff fragment structure allowed for the assignment of structural information to this protein family. This is exemplified by the construction of a structural model (Fig. 3C) for the most distant identified homolog R4R201_PSEPH (Supplementary Fig. S1), an otherwise uncharacterized protein from *Pseudomonas protegens*.

Discussion

Our results demonstrate that protein fragments identified by shotgun proteolysis are suitable for structural studies, as exemplified by the yrff fragment solution structure reported here. The yrff protein, from which the fragment had been derived from, is an inner membrane protein essential for the viability of *E. coli* (Yong *et al.*, 2013). The protein is a multi-span membrane protein and the identified shotgun proteolysis fragment corresponds to a putative cytoplasmic segment located between transmembrane helices 1 and 2. The yrff homolog IgaA has been shown to control the transcription of a range of genes involved in the maintenance of cell wall integrity, cell division and motility (Garcia-Calderon *et al.*, 2009). IgaA plays an important role in the virulence of *Salmonella enterica* through attenuation of the Rcs system, although molecular details remain unknown (Garcia-Calderon, *et al.*, 2009). In *E. coli*, it has also recently been shown that RscF directly interacts with the yrff periplasmic

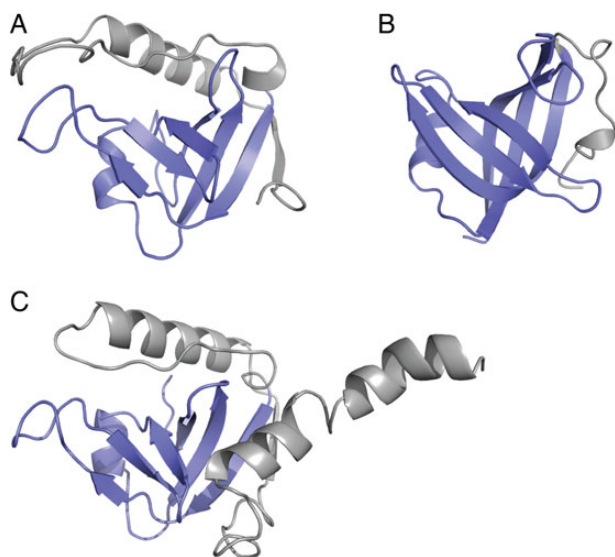


Fig. 3 Cartoon representation of the (A) yrff fragment and (B) *S. tokodaii* aspartyl-tRNA synthetase. OB-fold domains highlighted in blue. N-terminal additions to the OB-fold domain in gray. (C) Structural model of the distant yrff homolog R4R201_PSEPH generated by I-TASSER using the NMR structure reported here as a template.

domain to trigger the Rcs phosphorelay in response to envelope stress (Cho et al., 2014).

Although some aspects of the yrfF system have thereby been described in the literature, detailed mechanistic insights have so far remained elusive, possibly due to the absence of structural information for this protein family. The solution structure of the shotgun proteolysis fragment reported here not only provides molecular insights into the *E.coli* yrfF protein, but also allows the assignment of structural information to distant homologs observed in other species. Our findings highlight the potential of experimental approaches as an alternative and/or supplement to computational means for the identification of protein structure.

Supplementary data

Supplementary data are available at PEDS online.

References

- Arcus, V. (2002) *Curr. Opin. Struct. Biol.*, **12**, 794–801.
- Bachmann, B.J. (1972) *Bacteriol. Rev.*, **36**, 525–557.
- Barbato, G., Ikura, M., Kay, L.E., Pastor, R.W. and Bax, A. (1992) *Biochemistry*, **31**, 5269–5278.
- Bax, A., Ikura, M., Kay, L.E., Barbato, G. and Spera, S. (1991) *Ciba Found. Symp.*, **161**, 108–119. discussion 119–135.
- Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1978) *Arch. Biochem. Biophys.*, **185**, 584–591.
- Brunger, A.T. (2007) *Nat. Protoc.*, **2**, 2728–2733.
- Cabantous, S. and Waldo, G.S. (2006) *Nat. Methods*, **3**, 845–854.
- Cho, S.-H., Szewczyk, J., Pesavento, C., et al. (2014) *Cell*, **159**, 1652–1664.
- Chothia, C., Gough, J., Vogel, C. and Teichmann, S.A. (2003) *Science*, **300**, 1701–1703.
- Christ, D. and Winter, G. (2006) *J. Mol. Biol.*, **358**, 364–371.
- Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.
- D'Angelo, S., Velappan, N., Mignone, F., Santoro, C., Sblattero, D., Kiss, C. and Bradbury, A.R. (2011) *BMC Genom.*, **12** Suppl 1, S5.
- Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M. and Holm, L. (2001) *Nucleic Acids Res.*, **29**, 55–57.
- Dodd, R.B., Allen, M.D., Brown, S.E., Sanderson, C.M., Duncan, L.M., Lehner, P. J., Bycroft, M. and Read, R.J. (2004) *J. Biol. Chem.*, **279**, 53840–53847.
- Dyson, M.R., Perera, R.L., Shadbolt, S.P., Biderman, L., Bromek, K., Murzina, N. V. and McCafferty, J. (2008) *Nucleic Acids Res.*, **36**, e51.
- Englander, S.W. and Wand, A.J. (1987) *Biochemistry*, **26**, 5953–5958.
- Finn, R.D., Clements, J. and Eddy, S.R. (2011) *Nucleic Acids Res.*, **39**, W29–W37.
- Finn, R.D., Bateman, A., Clements, J., et al. (2014) *Nucleic Acids Res.*, **42**, D222–D230.
- Garcia-Calderon, C.B., Casadesus, J. and Ramos-Morales, F. (2009) *J. Bacteriol.*, **191**, 2743–2752.
- Gibson, T.J. (1984). University of Cambridge, Cambridge.
- Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001) *J. Mol. Biol.*, **313**, 903–919.
- Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. (2010) *Syst. Biol.*, **59**, 307–321.
- Hart, D.J. and Waldo, G.S. (2013) *Curr. Opin. Struct. Biol.*, **23**, 403–408.
- Huang, Y., Niu, B., Gao, Y., Fu, L. and Li, W. (2010) *Bioinformatics*, **26**, 680–682.
- Jones, D.T. (1999) *J. Mol. Biol.*, **292**, 195–202.
- Kay, L.E., Torchia, D.A. and Bax, A. (1989) *Biochemistry*, **28**, 8972–8979.
- Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2000) *J. Mol. Biol.*, **299**, 499–520.
- Kristensen, P. and Winter, G. (1998) *Fold Des.*, **3**, 321–328.
- Letunic, I., Doerks, T. and Bork, P. (2012) *Nucleic Acids Res.*, **40**, D302–D305.
- Liu, J. and Rost, B. (2004) *Nucleic Acids Res.*, **32**, 3522–3530.
- Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C. and Gough, J. (2004) *Nucleic Acids Res.*, **32**, D235–D239.
- Magrane, M. and Consortium, U. (2011) *Database (Oxford)*, **2011**, bar009.
- Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R. and Sander, C. (2011) *PLoS One*, **6**, e28766.
- Neidhardt, F.C., Bloch, P.L. and Smith, D.F. (1974) *J. Bacteriol.*, **119**, 736–747.
- Neri, D., Szyperski, T., Otting, G., Senn, H. and Wuthrich, K. (1989) *Biochemistry*, **28**, 7510–7516.
- Pedelacq, J.D., Nguyen, H.B., Cabantous, S., et al. (2011) *Nucleic Acids Res.*, **39**, e125.
- Porter, R.R. (1959) *Biochem. J.*, **73**, 119–126.
- Punta, M., Coghill, P.C., Eberhardt, R.Y., et al. (2012) *Nucleic Acids Res.*, **40**, D290–D301.
- Shi, J., Blundell, T.L. and Mizuguchi, K. (2001) *J. Mol. Biol.*, **310**, 243–257.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. and Zhang, Y. (2014) *Nat. Methods*, **12**, 7–8.
- Yong, H.T., Yamamoto, N., Takeuchi, R., Hsieh, Y.J., Conrad, T.M., Datsenko, K.A., Nakayashiki, T., Wanner, B.L. and Mori, H. (2013) *Genes Genet. Syst.*, **88**, 233–240.