Journal of Biomolecular NMR (2006) 35: 295–300
DOI 10.1007/s10858-006-9032-y

NMR structure note

The solution structure of the protein ydhA from *Escherichia coli*

Matthew Revington[a], Anthony Semesi[b], Adelinda Yee[b], Cheryl H. Arrowsmith[b] & Gary S. Shaw[a,*]

[a]*Department of Biochemistry, University of Western Ontario, London, N6A 5C1, Ontario, Canada;*
[b]*Ontario Centre for Structural Proteomics, University Health Network, Toronto, M5G 2C4, Ontario, Canada*

Received 17 February 2006; Accepted 15 May 2006

Key words: beta barrel, NMR, protein structure, structural proteomics

Biological context

The *Escherichia coli* gene ydhA (P28224) encodes a 109 amino acid protein of unknown structure and function. YdhA lies between genes for pyridoxamine 5′-phosphate oxidase (pdxH) and ydhH, a hypothetical protein of unknown function, in the *E. coli* genome. Insertions that disrupt the ydhA gene have no identifiable phenotype for *E. coli* under normal growth conditions and no effect on transcription of the neighboring pdxH-tyrS operon (Lam and Winkler, 1992). Under normal growth conditions the transcripts for ydhA are not detected suggesting that its expression may be regulated by specific factors or conditions that have not yet been identified (Lam and Winkler, 1992).

The 27 N-terminal residues of ydhA form a predicted periplasmic export signal sequence that should be cleaved after membrane translocation into the periplasmic space while the remaining 82 residues should constitute the final functional form. Sequence alignments identify significant similarity to several other putative periplasmic or outer membrane associated lipoproteins but none of known structure or defined function. Although most of the closely aligned sequences contained only an N-terminal signal sequence and a single ydhA-like domain there were two genes identified that contained the ydhA-like domain paired with another domain. The genes Q8XWH8 from *Ralstonia solanacearum* and Q8YFM6 from *Brucella melitensis* code for two-domain proteins with an N-terminal periplasmic signal sequence followed by a domain homologous to the *E. coli* protein HslJ and a C-terminal ydhA-like domain. HslJ has been identified as an outer membrane protein that confers novobiocin resistance (Lilic et al., 2003) and that may be upregulated by heat shock conditions (Chuang and Blattner, 1993). An *in vivo* interaction was detected between ydhA and Hsp15 a heat shock protein that has been associated mostly with binding of RNA, DNA and ribosomal associated proteins (Butland et al., 2005).

Sequence analysis therefore indicates that the ydhA protein could be part of a novel family of outer membrane associated proteins that are conserved in gram negative bacteria. This family of proteins would be expressed at very low levels under standard growth conditions and interact with proteins related to antibiotic resistance. An NMR structural study of ydhA was undertaken to gain further insight into its function and role in *E. coli.* Here we report the solution structure of ydhA as a slightly flattened eight stranded β-barrel that is stabilized by an intramolecular disulfide bond. This structure has been deposited in the RCSB Protein Data Bank with accession code 2F09. Structural similarity analysis shows it to be a member of either the calycin superfamily or of the β-barrel-sandwich

*To whom correspondence should be addressed. E-mail: gshaw1@uwo.ca hybrid superfamily. These proteins use the hydrophobic interior of the β-barrel to bind small hydrophobic substrates often as part of the outer membrane/periplasmic space transport processes.

Methods and results

The *E. coli* gene ydhA was cloned into a pET15b vector (Novagen, Madison, WI) for overexpression in the BL21(Gold DE3) strain of *E. coli* (Yee et al., 2000). The construct consisted of the codons for the 82 amino acid sequence of the mature protein without the N-terminal 27 residue periplasmic signal sequence (MTMKKLLIIILPV LLSGCSAF-NQLVER) but with a 20 residue N-terminal leader sequence containing a 6 × His tag (MGSSHH-HHHHSSGLVPRGSH). Residue num bering in this work refers to the 82 amino acid wild type ydhA with the N-terminal signal sequence cleaved and neglects the 20 residue leader sequence.

The BL21(Gold DE3) cells containing this plasmid were grown in M9 minimal media supplemented with ^{15}N labeled ammonium chloride and ^{13}C labeled glucose to produce uniformly ^{15}N and ^{13}C labeled protein. Expression was induced by the addition of 1.0 mM isothiopropylgalactoside. The cells were harvested, lysed and the protein was purified using Ni^{2+}-NTA affinity chromatography. Trial HSQC spectra were collected under several standard buffer and salt conditions. The optimal conditions for spectral quality and sample stability utilized a buffer containing 10 mM sodium phosphate, 400 mM sodium chloride, 10 mM dithiothreitol, 0.01% sodium azide, 1 mM benzamidine, pH 6.5, 90% H_2O, 10% D_2O for the final NMR sample. The final sample volume of ~300 μl at a concentration of 0.5 mM protein was placed in a Shigemi NMR Tube for data collection (Shigemi Inc., Allison Park, PA).

A single uniformly ^{15}N, ^{13}C labeled NMR sample was used for all experiments. A series of ^{15}N-edited HSQC spectra were collected at 15, 25, and 35 °C to identify the temperature at which the protein produced the spectrum with the most resolved signals and a number of peaks consistent with a monodisperse protein. At 25 °C, as shown in Figure 1, the HSQC spectrum showed uniform signal intensities for the ~100 well dispersed peaks, consistent with a 82 amino acid protein and sidechain amino groups. In this spectrum backbone amide correlations were observed for 77 of 82 residues. The exceptions were three prolines (Pro14, Pro22 and Pro80) and residues Met1 and Gln2 which likely exchanged rapidly with solvent. The resonances arising from the N-terminal His-Tag leader sequence were not assigned in this spectrum or in the three dimensional spectra (discussed below) although several weak peaks were observed that were likely produced by that sequence. All other NMR experiments were run at 25 °C.

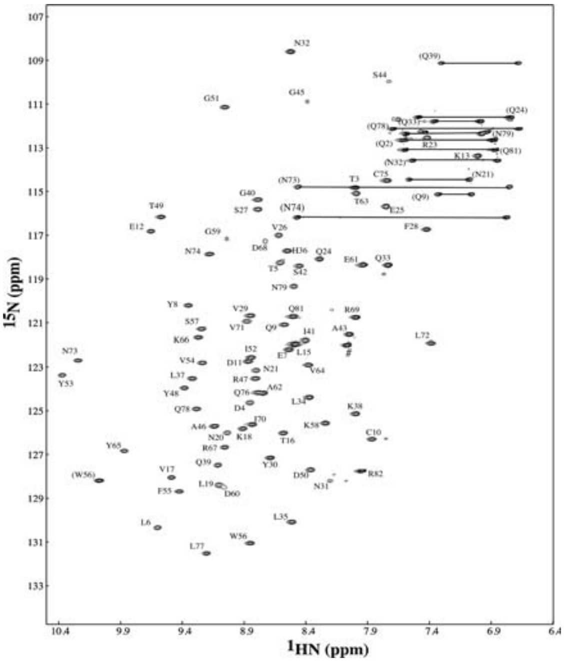


Figure 1. A 600 MHz ^{1}H–^{15}N HSQC spectrum of [U–^{15}N,^{13}C] ydhA at 298 K. The assignments are indicated on the figure, assignments in brackets indicate side chain tryptophan indole imide or pairs of asparagines and glutamine sidechain amino groups. Unassigned peaks that likely arise from the N-terminal His-tag are indicated with an # and were not assigned. Peaks assigned to ydhA but that were too weak to be seen at the contour level used are indicated by a *. Numbering refers to that of the wild type protein sequence (residues 1–82) and neglects the N-terminal leader sequence.

The aromatic sidechain assignment spectra HBCBCGCDCEHE, HBCBCGCDHD (Yamazaki et al., 1993) and the aliphatic sidechain assignment spectrum HCCH-COSY were collected on a 600 MHz Varian Inova spectrometer equipped with a ^{13}C enhanced HCN cold probe. All other data were collected using standard Varian Biopack sequences on a Varian INOVA 600 MHz spectrometer equipped with a *xyz* gradient probe. The backbone and partial sidechain assignments were derived from analysis of HNCO, HNCA, HNCACB (Grzesiek et al., 1992), CBCA(CO)NH (Grzesiek et al., 1993), HCC(CO)NH, and CC(CO)NH spectra. The ^{15}N NOESY-HSQC and ^{13}C NOESY-HSQC experiments were acquired using mixing times of 150 and 100 ms, respectively. Typical experimental parameters for the *xyz* gradient probe included the collection of 16 or 24 transients per increment and 32 complex planes in F2. For the cold probe only 8 transients per increment were acquired. Total data collection time was approximately 1 month. All data was processed using NMRPIPE software (Delagio et al., 1995) using standard in-house processing scripts. A $\pi/3$ shifted sine squared bell function was applied as apodization to the directly detected dimensions while $\pi/3$ shifted sine bell function was applied to the indirectly detected dimensions. Three-dimensional data sets were linearly predicted and zero filled in the F1 and F2 dimensions to 256 and 64 points respectively. NMRVIEW 5.2.2 (Johnson et al., 1994) was used for spectral analysis and assignment.

Analysis of the spectra resulted in the assignment of 836 of the 866 assignable protons in the sequence of wild type ydhA while no signals from the 20-residue leader sequence were assigned. About 22 of the 30 unassigned protons in ydhA were located in aromatic sidechains and could not be assigned due to degeneracy of those signals. Stereospecific assignments were derived from the automated NOE assignment algorithm in CYANA for the Hβ protons from Leu6, Tyr8, Gln9, Pro14, Lys18, His36, Ser57, Glu61, Leu72, Leu77, as well as the Hδ protons from Lys18 and the γ methyl protons of Val26 and Val71. The chemical shifts of the cysteine Cβ nuclei in the sequence were both downfield of 43 ppm, indicative of these residues being in the oxidized state and involved in a disulfide bond. Chemical shift assignments were deposited in the Biological Magnetic Resonance Data Bank under accession code 6955.

Although the NMR data was consistent with a monomer there was a possibility of dimer or higher order structures, possibly mediated by the disulfide bonds. To test the possibility of inter-subunit disulfide bonds the uniformly ^{13}C, ^{15}N labeled NMR sample was run on a Q-TOF2 Mass spectrometer using a Z-spray source in positive ion electrospray mode. The major species identified had a mass of 12367.7 Da, which is close to the calculated mass for the full length monomeric form having an intramolecular disulfide bond (12380.1 Da) having isotopic labeling levels $>99\%$.

We decided to proceed with calculations of a monomeric model for ydhA and then to compare estimates of the size of the protein derived from WATERSLED (Altieri et al., 1995) diffusion experiments with values calculated from the model. The monomeric form dictates that the disulfide bond must be intramolecular between the only cysteine residues in the sequence, Cys10 and Cys75. Initial structures were calculated using the program CYANA (Guntert et al., 1997) with NOE assignments being accomplished by iterative automated and manual approaches. A total of 1927 NOES were identified in the NOE experiments of which 1694 were assigned unambiguously by this approach. TALOS (Cornilescu et al., 1999) angular restraints were used in regions in which the Chemical Shift Index (CSI) (Wishart et al., 1992) predicted secondary structure and where 10 consistent database matches were identified. The error ranges for the angular restraints were set to 2 times that derived from TALOS or $\pm 20°$ whichever was greater. A total of 28 angular constraints from TALOS were defined by these criteria and for the remaining ϕ, ψ angles loose angular restraints for the allowed regions of the Ramachandran plot were used. Calculations that did not include any explicit disulfide constraints between Cys10 and Cys75 provided structures that placed these two residues in close proximity to each other in the tertiary structure. Therefore a disulfide bond constraint between Cys10 to Cys75 was included for further calculations. For the final set of structures from CYANA there were 7 cycles of refinement run with 100 structures calculated per cycle and the 20 structures with the lowest calculated target function were used to seed the next cycle. The 20 best structures from the seventh cycle were then further refined in the presence of explicit water using CNS (Nederveen et al., 2005).

The signal intensity, *S*(g), observed in WATERSLED experiments is a function of the gradient power level g, and the translational diffusion coefficient, *D*, and therefore can be interpreted in terms of molecular shape and size. The diffusion of ydhA was analyzed by fitting (PRISM, Graphpad Software, San Diego, CA) the *S*(g) observed in an array of one dimensional ^{1}H WATERSLED NMR spectra at increasing gradient strengths using Equation 1.

$$s(\mathrm{g}) = \mathrm{Ae}^{-\mathrm{dg}^2} \quad (1)$$

The value d obtained from the best fit is the single exponential decay rate of S(g) and is proportional to D. The d values of ydhA and dioxane, an internal standard with a known effective radius of hydration, R_{Hd}, of 2.12 Å, were determined from the same set of WATERSLED data and then used to calculate the effective radius of hydration of ydhA, R_{HydhA}, of 19.46 ± 2.03 Å using Equation 2.

$$R_{\mathrm{HydhA}} = R_{\mathrm{Hd}}\{d_{\mathrm{diox}}/d_{\mathrm{ydhA}}\} \quad (2)$$

The program HYDROPRO (Garcia De La Torre et al., 2000) was used to calculate a radius of gyration, R_{g}, of 13.4 Å from the solution structures of monomeric ydhA while an R_g of 19.5 Å was calculated for the same structure with the leader sequence present as a fully extended structure. The effective radius of hydration was calculated from the radius of gyration for a globular structure by Equation 3 (Wilkins et al., 1999),

$$R\mathrm{g} = (3/5)^{1/2} * \mathrm{R_H} \quad (3)$$

resulting in a calculated R_{HydhA} for ydhA of 17.3 Å and a 25.2 Å for the structure that included the extended leader sequence. The leader sequence is likely unstructured in solution and results in an observed R_{H} between the calculated values for the two structures. The size of the protein in solution derived from the WATERSLED experiments is therefore consistent with the values calculated from the monomeric solution structure.

The overlaid solution structures of the 20 best structures of ydhA are shown in Figure 2A. The stereo ribbon diagram shown in Figure 2B show that ydhA is an 8-stranded, anti-parallel, slightly flattened β-barrel that is stabilized by the disulfide

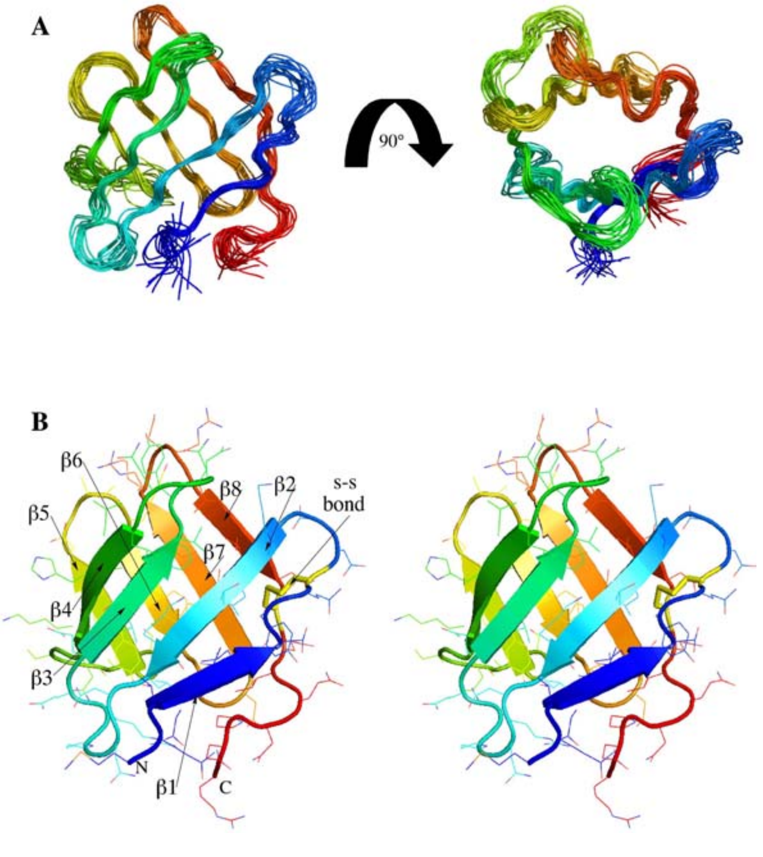
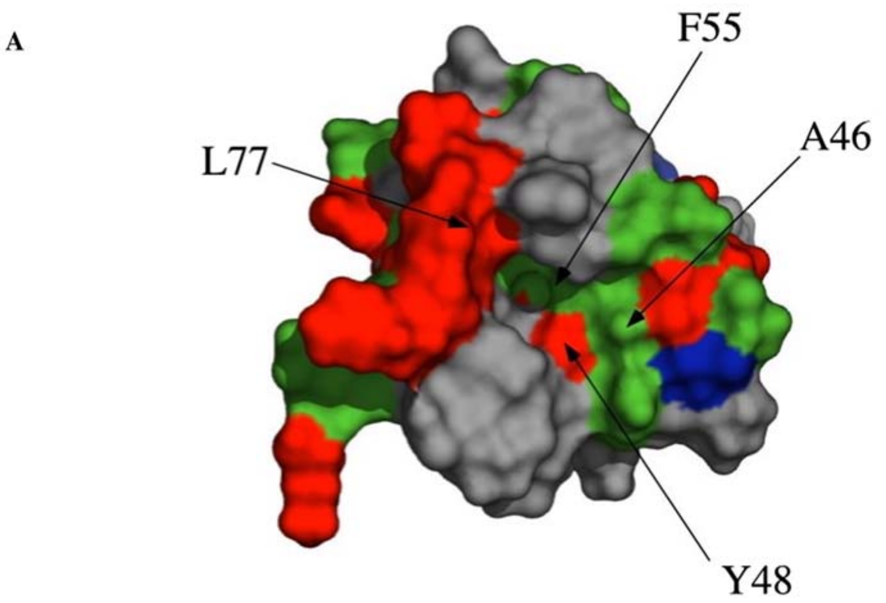


Figure 2. (A) An overlay of the 20 best solution structures of ydhA shown from the side of the β-barrel and from a 90° rotated view showing the slightly flattened barrel shape. (B) Stereo view of ribbon diagrams of ydhA with side chains shown. The 8β-strands in ydhA are labeled as β1 (3–7), β2 (13–19), β3 (25–29), β4 (35–39), β5 (47–50), β6 (54–57), β77 (61–66), and β8 (70–74). The position of the disulfide bond between Cys10 and Cys75 is indicated.

bond between Cys10 and Cys75, located at the ends of the first and eighth strands of the barrel respectively. The β-strands are made up of the residues 3–7 (β1), 13–19 (β2), 25–29 (β3), 35–39 (β4), 47–50 (β5), 54–57 (β6), 61–66 (β7), and 70–74 (β8). Examination of space filling model of the structure shows that there is access to the interior of the protein from the end of the barrel formed by loops connecting strands β2 and β3, β 4 and β5, β6 and β7 but the other end is blocked by aromatic side chains. All of the 16 side chains located on the interior of the barrel are hydrophobic (Leu, Val, Phe, Ile, Tyr, Ala) indicating that only hydrophobic substrates could be bound in the binding pocket. Mapping of the conserved residues from the BLAST alignment (Figure 3B) onto the structure (Figure 3A) revealed a conserved patch on the open end of the barrel formed by the ^{42}SASGAR^{48}Y and ^{55}FWSK^{59}G sequences and may indicate a conserved substrate binding or protein/protein interaction surface. The BLAST alignment also highlighted the conservation of Cys residues (10 and 75) that participate in the disulfide bond. Representative structures of ydhA were submitted to the structural homology programs PROFUNC (Laskowski et al., 2005) and DALI (Holm and Sander, 1993). The SSM subroutine in PROFUNC identified homology between elements



B

```
E. coli       -MQTDTLEYQC-DEKPLTVKL----NNPRQ---EVSFVYDNQLLHLKQGISASGARYTDGIYVFWSKGDEATVYKRDRIVLNNCQLQNPQR
S. sonnei     -MQTATLEYQC-DEKPLTVKL----NNPRQ---EVSFVYDNQLLHLKQGISASGARYTDGIYVFWSKGDEATVYKRDRIVLNNCQLQNPQR
S. flexneri   -MQTDTLEYQC-DEKPLTVKL----NNPRQ---EVSFVYDNQLLHLKQGLSASGARYSDGIYVFWSKGEEATVYKRDRIVLNNCQLQNPQR
S. typhim     -MQTDTLEYQC-DEKPLTVKV----NNPRE---EVSFVYDNKLLTLKQGISASGARYTDGIYVFWSQGESATVYKRDRIVLNNCQLQNPKR
S. typhi      -MQTDTLEYQC-DEKPLTVKA----NNPRE---EVSFVYDNKLLTLKQGISASGARYTDGIYVFWS-------------------------
Y. pestis     ----ETLHYQC-GTTKLTVTQ----NNKQD---NVSFIMDGKQLTLPQVRAASGAKYSDGHYTFWSKGDSAFIEREEKVIINDCVL-----
P. astros     ----ETLHYQC-GTMPLTVTL----QQGREAPSQVSFLLDGERLILPQVVSASGVRYSNDTYTFWSKGDRAFIQRGERVIVDDCVL-----
Y. enterl     --KNETLHYQC-GGTKLTVMQ----DNKQD---KVSLVLDGKQLTLPQVRAASGAKYSDNHYTFWSKGNTAFVERDEKVIINDC-------
R. solana     ------VRYACQDGKSLTVKYMTVGDTPL-----ATIVLDGKTLVFANVVSGSGARYASGRYVWWTKG-----------------------
Consensus      M:: T.:Y C     LTV      ::  :   :.:...: : L . :.   .SG.:Y :   Y .W::G::A . : ::...::C LQNP:R
```

Figure 3. (A) Surface representation of the ydhA structure colored with respect to residue conservation as depicted in the BLAST alignment in Figure 3B. The molecule is oriented to show the hydrophobic 'open' end of the barrel and the highly conserved patch of residues close to opening of the hydrophobic pocket in the centre of the β-barrel. Residue numbering refers to the wild type sequence and neglects the N-terminal leader sequence. (B) Alignment of sequences homologous to wild type ydhA identified using BLAST. The residues are colored red to indicate complete conservation, green to indicate a single conservative variation in the sequences shown, blue to indicate two conserved possibilities and black to indicate greater variation between the sequences. A consensus sequence is shown at the bottom where completely conserved residues are indicated by their single letter code, conserved polar positions are indicated by a colon and conserved non-polar positions are indicated by a period.

of secondary structure to several β-barrel avidin and streptavidin proteins however ydhA exhibits sequence similarities below 20% for members of the avidin family. The DALI program identified significant structural similarity with several members of the lipocalin family. Lipocalins are a family of 8-stranded anti-parallel β-barrel proteins stabilized by disulfide bonds that, in bacteria, are commonly found in the periplasmic space and bind a variety of substrates including lipids and antibiotics (Bishop, 2000). Lipocalin family members, however, usually have a small C-terminal helix and three structurally conserved regions that ydhA lacks. Both the avidin and lipocalin structural families are part of the calycin superfamily of β-barrel proteins that also includes some metalloproteinase proteins and fatty acid binding proteins. Based on a combination of sequence and structural factors ydhA does not clearly fit in any of the families of the calycin superfamily. Some similarity was also identified with the β-barrel-sandwich hybrid superfamily that contains a diverse set of protein including various periplasmic proteins (Table 1).

Table 1. Structural statistics for ydhA

Experimental restraints	Number
TALOS Dihedral	28
NOE Distance	
Intraresidue	785
Short($\| i–j\| = 1$)	433
Medium($2 < \| i–j \| < 5$)	142
Long($5 < \| i–j \|$)	334
Total	1694
Violations (20 structures)	
NOE violations > 0.3 Å	6
Dihedral constraint violations > 2.0°	0
Close contacts violations > 0.01 Å	6
Ramachandran statistics (%)	
Most favored/Additionally allowed	95.9
Generously allowed	4.1
Disallowed	0.0
RMSD to mean structure (Å)	
All residues (1–82)	
All atoms	1.834
Backbone	0.837
Secondary Structure (β1–β8 only)	0.360

Acknowledgement

The authors gratefully acknowledge the Ontario Research Development Challenge Fund, a CIHR Multiuser Maintenance Grant and the Canada Research Chairs Program for support. We would also like to thank Dr James (W.Y.) Choy of the University of Western Ontario for his help in implementing the WATERSLED translational diffusion experiments and Amanda Doherty-Kirby of the Dr Don Rix Protein Identification Facility for the mass spectrometry data.

References

Altieri, A.S., Hinton, D.P. and Byrd, R.A. (1995) *J. Am. Chem. Soc.*, **117**, 7566–7567.

Bishop, R.E. (2000) *Biochim. Biophys. Acta.*, **1482**, 73–83.

Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J. and Emili, A. (2005) *Nature*, **433**, 531–537.

Chuang, S.E. and Blattner, F.R. (1993) *J. Bacteriol.*, **175**, 5242–5252.

Cornilescu, G., Delagio, F. and Bax, A. (1999) *J. Biolmol. NMR*, **13**, 289–302.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR*, **6**, 277–293.

GarciaDeLa Torre, J., Huertas, M.L. and Carrasco, B. (2000) *Biophys. J.*, **78**, 719–730.

Grzesiek, S., Anglister, J. and Bax, A. (1993) *J. Magn. Reson. Ser. B.*, **101**, 114–119.

Grzesiek, S. and Bax, A. (1992) *J. Magn. Reson.*, **99**, 201–207.

Guntert, P., Mumenthaler, C. and Wuthrich, K. (1997) *J. Mol. Biol.*, **273**, 283–298.

Holm, L. and Sander, C. (1993) *J. Mol. Biol.*, **233**, 123–138.

Johnson, B.A. and Blevins, R.A. (1994) *NMR J. Biolmol.*, **4**, 603–614.

Lam, H.M. and Winkler, M.E. (1992) *J. Bact.*, **174**, 6033–6045.

Lilic, M., Jovanovic, M., Jovanovic, G. and Savic, D.J. (2003) *FEMS Microbiol. Lett.*, **224**, 239–246.

Laskowski, R.A., Watson, J.D. and Thornton, J.M. (2005) *Nucleic Acids Res.*, **33**, W89–W93.

Nederveen, A.J., Doreleijers, J.F., Vranken, W., Miller, Z., Spronk, C.A., Nabuurs, S.B., Guntert, P., Livny, M., Markley, J.L., Nilges, M., Ulrich, E.L., Kaptein, R. and Bonvin, A.M. (2005) *Proteins*, **59**, 662–672.

Wilkins, D.K., Grimshaw, S.B., Receveur, V., Dobson, C.M., Jones, J.A. and Smith, L.J. (1999) *Biochemistry*, **38**, 16424–16431.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) *Biochemistry*, **31**, 1647–1651.

Yamazaki, T., Forman-Kay, J.D. and Kay, L.E. (1993) *J. Am. Chem. Soc.*, **115**, 11054–11055.

Yee, A., Booth, V., Dharamsi, A., Engel, A., Edwards, A.M. and Arrowsmith, C.H. (2000) *Proc. Natl. Acad. Sci. USA*, **97**, 6311–6315.