WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

STRUCTURE NOTE

NMR structure of protein Cgl2762 from *Corynebacterium glutamicum* implicated in DNA transposition reveals a helix-turn-helix motif attached to a flexibly disordered leucine zipper

Kiran Kumar Singarapu,[1,2] Rong Xiao,[2,3] Dinesh K. Sukumaran,[1] Thomas Acton,[2,3] Gaetano T. Montelione,[2,3] and Thomas Szyperski[1,2]*

[1] Department of Chemistry, State University of New York at Buffalo, Buffalo, New York 14260–3000

[2] Northeast Structural Genomics Consortium

[3] Center of Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey 08854–5638

***Key words:** structural genomics; transposase; helix-turn-helix motif; leucine zipper; DNA binding; GFT NMR.*

INTRODUCTION

97-residue protein Cgl2762 (gil42602314, SwissProt/TrEMBL ID Q8NM20_CORGL, access number Q8NM20) encoded by gene Q8NM20_CORGL from *Corynebacterium glutamicum* has no significant sequence similarity with any protein with known three-dimensional structure. The protein was thus selected by the Protein Structure Initiative-2 of the United States National Institutes of Health and was assigned to the Northeast Structural Genomics consortium (NESG; http://www.nesg.org) (NESG target ID CgR3). Protein Cgl2762 belongs to the large Pfam[1] family PF01527, which currently contains 1985 proteins identified as "transposase proteins." This class of proteins has been shown to be essential for DNA transposition events.[2] All members of PF01527 contain an N-terminal helix-turn-helix (HTH) DNA binding domain, and many of the members (currently 1660 of 1985) also possess a C-terminal domain with a leucine zipper (LZ) motif. The HTH motif as well as the LZ motif, which mediates oligomerization of transposase components, is essential for the DNA binding in bacterial insertion sequences of IS911 transposase family.[3,4] PF01527 belongs to a "HTH clan" which comprises about 70 Pfam families. Here, we report the NMR solution structure of protein Cgl2762 which was solved using a protocol devised for high-throughput protein structure determination.[5]

MATERIALS AND METHODS

Protein Cgl2762 (Q8NM20_CORGL) was cloned, expressed, and purified following standard protocols developed by the NESG for production of uniformly *U*-^{13}C,^{15}N-labeled protein samples.[6] Briefly, the full length Cgl2762 gene from *Corynebacterium glutamicum* was cloned into a pET21 (Novagen) derivative, yielding the plasmid CgR3-21.2. The resulting construct contains eight nonnative residues at the C-terminus (LEHHHHHH) that facilitate protein purification. *Escherichia coli* BL21 (DE3) pMGK cells, a rare codon enhanced strain, were transformed with CgR3-21.2, and cultured in MJ9 minimal medium containing $(^{15}NH_4)_2SO_4$ and *U*-^{13}C-glucose as sole nitrogen and carbon sources. *U*-^{13}C, ^{15}N Cgl2762 was purified using an AKTAxpress (GE Healthcare) based two-step protocol consisting of IMAC (HisTrap HP) and gel filtration (HiLoad 26/60 Superdex 75) chromatography. The

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887–3585/suppmat/.

Grant sponsor: National Institutes of Health; Grant number: U54GM074958-01; Grant sponsor: National Science Foundation; Grant number: MCB 0416899. Support from the Publication Subvention Fund of the College of Arts and Sciences at SUNY Buffalo is acknowledged.

*Correspondence to: Thomas Szyperski, Department of Chemistry, State University at New York at Buffalo, Buffalo, NY 14260. E-mail: szypersk@chem.buffalo.edu

Received 13 September 2007; Accepted 19 September 2007

Published online 3 January 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21840

Table I
Statistics of NMR Structure of Protein Cgl2762

Conformationally restricting distance constraints	
Intraresidue [$i = j$]	330
Sequential [$(i - j) = 1$]	433
Medium range [$1 < (i - j) \leq 5$]	293
Long range [$(i - j) > 5$]	223
Total	1279
Dihedral angle constraints	
Φ	60
ψ	60
Number of constraints per residue	14.4
Number of long-range constraints per residue	2.3
Number of long-range constraints per residue (8–50)	5.3
Completeness of stereo-specific assignments[a] **(%)**	
$^{\beta}CH_2$	25 (13/52)
Val and Leu isopropyl groups	45 (5/11)
CYANA target function (Å^2)	0.35 ± 0.001
Average r.m.s.d. to the mean CYANA coordinates (Å)	
Regular secondary structure elements[b], backbone heavy	0.45 ± 0.09
Regular secondary structure elements, all heavy atoms	0.99 ± 0.12
Residues 7–49 backbone heavy atoms N, C^{α}, C′	0.52 ± 0.09
Residues 7–49 all heavy atoms	0.99 ± 0.10
Residues 62–72 backbone heavy atom	0.38 ± 0.09
Residues 76–82 backbone heavy atom	0.48 ± 0.13
Residues 86–91 backbone heavy atom	0.46 ± 0.17
Heavy atoms of molecular core (or best-defined SC)[c]	0.69 ± 0.08
PROCHECK rawscore[d] (φ and ψ/all dihedral angles)	0.40/0.14
PROCHECK *Z*-scores[d] (φ and ψ/all dihedral angles)	1.89/−0.83
MOLPROBITY *Z*-score/raw score[e]	−0.87/13.94
AutoQF R/P/DP scores (%)[f]	0.96/0.96/0.73
Ramachandran plot summary ordered residue ranges: 7–47, 59–72, 76–82, 87–91 (%)	
Most favored regions	97.5
Additionally allowed regions	2.5
Generously allowed regions	0.0
Disallowed regions	0
Average number of distance constraints violations per CYANA conformer (Å)	
0.2–0.5	0.0
>0.5	0
Average number of dihedral-angle constraint violations per CYANA conformer (°)	
>10	4.8

[a]Relative to pairs with nondegenerate chemical shifts.
[b]Residues 7–19, 26–34, 36–48 (α-helices).
[c]Includes 31 residues: 7, 11, 13–18, 20, 21, 23–27, 29–36, 38–45.
[d]Scores defined in Ref. 13.
[e]Scores defined in Ref. 14.
[f]Scores defined in Ref. 15.

final yield of purified *U*-^{13}C, ^{15}N Cgl2762 (>98% homogenous by SDS-PAGE; 11.5 kDa by MALDI-TOF mass spectrometry) was about 15 mg/L. The final sample of *U*-^{13}C, ^{15}N labeled Cgl2762 was prepared at a concentration of ~1.2 m*M* in 95% H_2O/5% D_2O solution containing 20 m*M* ammonium acetate (pH 4.5), 100 m*M* NaCl, 10 m*M* DTT, 5 m*M* $CaCl_2$, 0.02% NaN_3. An isotropic overall rotational correlation time of ~6 ns was inferred from ^{15}N spin relaxation times,[5] indicating that the protein is monomeric in solution. This conclusion was further confirmed by an analytic gel-filtration (Agilent Technologies) followed by a combination of static light scattering and refractive index (Wyatt Technology).[6]

All NMR spectra were recorded at 25°C on a Varian INOVA 750 spectrometer equipped with a cryogenic probe. Five through-bond correlated G-matrix Fourier transform[7] (GFT) NMR experiments[7–9] were collected for backbone and side chain resonance assignment (total measurement time: 51 h), and a 3D $^{15}N/^{13}C^{aliphatic}/^{13}C^{aromatic}$-resolved [^{1}H,^{1}H]-NOESY spectrum[9] (mixing time: 70 ms; measurement time: 24 h) was acquired to derive ^{1}H—^{1}H distance constraints. Spectra were processed and analyzed with the programs NMRpipe[10] and XEASY[11], respectively. Sequence specific backbone (H^N, H^{α}, N, C^{α}) and H^{β}/C^{β} resonance assignments were obtained with (4,3)D HNN$\underline{C}^{\alpha\beta}\underline{C}^{\alpha}$/$\underline{C}^{\alpha\beta}\underline{C}^{\alpha}$(CO)NHN[4] and (4,3)D $\underline{H}^{\alpha\beta}\underline{C}^{\alpha\beta}$(CO)NHN experiments using the program AUTOASSIGN.[12] Side-chain assignments were accomplished by using aliphatic and aromatic (4,3)D $\underline{HC}$CH. Assignments were obtained for 98% of the as-

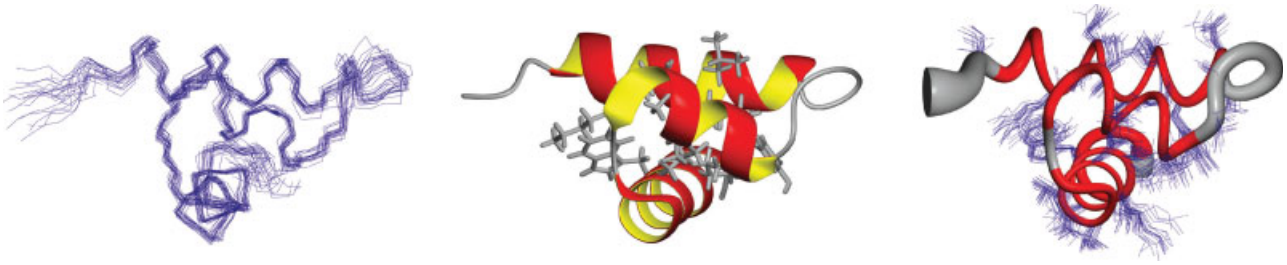

Figure 1

*NMR structure of the HTH domain (1–50) of protein Cgl2762. (**a**) The 20 CYANA conformers with the lowest residual CYANA target function representing the NMR solution structure are shown after superposition for minimal r.m.s.d. of the backbone heavy atoms N, C^{α}, and C′ atoms of the α-helices (Table I). (**b**) Ribbon drawing of the CYANA conformer with the lowest residual target function value (Table I). The α-helices I to III are shown in red and yellow, other polypeptide segments are in grey, and the N- and C-terminal boundaries of the polypeptide segment are indicated as "N" and "C." Side chains of the residues in the molecular core are shown in grey. Those are conserved within the HTH clan (Table SI; Fig. 3). (**c**) "Sausage" representation of 20 superimposed conformers in the orientation of (a). For the presentation of the backbone a spline function was draw through the C^{α} positions and the thickness of the cylindrical rod is proportional to the mean of the global displacements of the 20 CYANA conformers calculated after superposition as in described in (a). The α-helices are shown in red and other polypeptide segments are displayed in grey. The 31 best-defined side chains (Table I) are also displayed. The Figures were generated using the program MOLMOL.[21]*

signable backbone (excluding the N-terminal NH_3^+, the Pro ^{15}N, and the $^{13}C'$ shifts) and $^{13}C^{\beta}$, and for 94% of the side chain chemical shifts (excluding Lys NH_3^+, $ArgNH_2$, OH, side chain $^{13}C'$, and aromatic quaternary ^{13}C shifts; Table I). Stereo-specific assignments were obtained for 25% of the β-methylene groups exhibiting non-degenerate proton chemical shifts, and for 45% of the Val and Leu isopropyl moieties with non-degenerate chemical shifts (Table I). Chemical shifts were deposited in the BioMagResBank (accession code: 15086). Upper distance limit constraints for structure calculations were extracted from NOESY (Table I), and backbone dihedral angle constraints were derived from chemical shifts as described[16] for residues located in locally well defined α-helices I to VI (Table I) by using the program TALOS. The programs CYANA[17,18] and AUTOSTRUCTURE[19] were used in parallel to automatically assign long-range NOEs. The final structure calculations were performed using version 2.1 of CYANA[18] and CNS[20] with the DYANA constraints (see: http://www.las.jp/prod/cyana/eg).

RESULTS AND DISCUSSION

The statistics of the structure determination of protein Cgl2762 are summarized in Table I. Protein Cgl2762

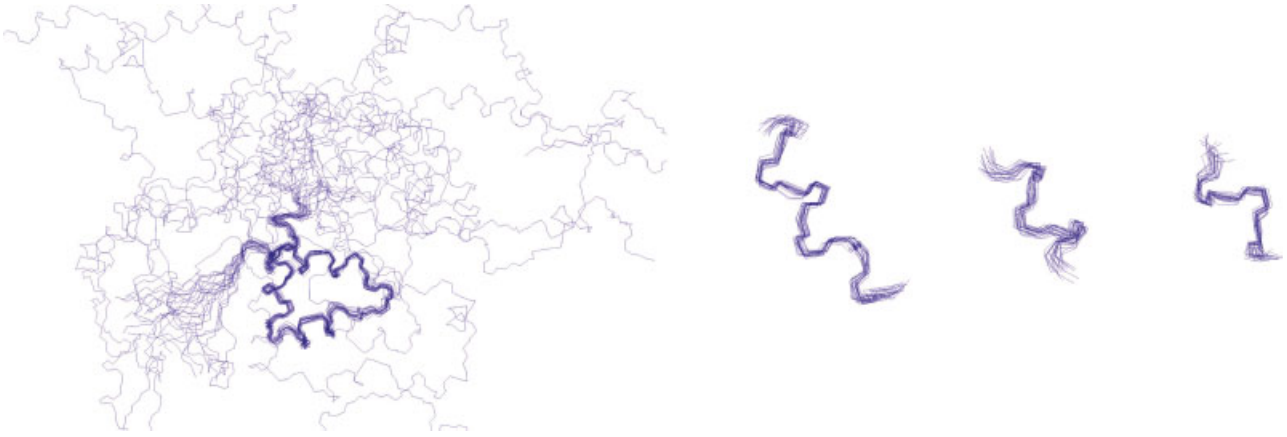

Figure 2

*(**a**) The 20 CYANA conformers representing the solution structure of protein Cgl2762 are shown after superposition of the α-helices (shown in bold) of the HTH domain (Fig. 1). The flexibly disordered nature of the C-terminal segment (52–97) is apparent. However, the three helices in the segment are locally well defined, as is demonstrated by superposition of polypeptide backbone heavy atoms of the individual helices only (Table I). (**b**) α-helix IV (62–72). (**c**) α-helix V (76–82). (**d**) α-helix VI.*

DOI 10.1002/prot

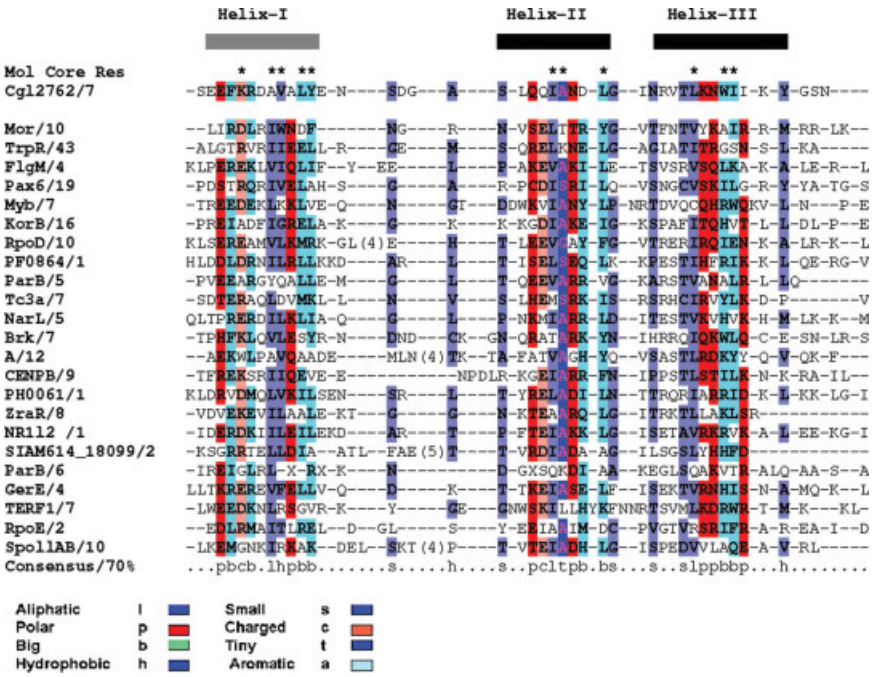

Figure 3

Structure based sequence alignment of the HTH domain (7–50) of protein Cgl2762 (shown at the top) with the 23 HTH domains identified using the program DALI (Table SI). Conserved residues are highlighted with colors, and the color coding scheme is indicated below the alignment. The gray bar indicates α-helix-I, and the black bars represent α-helices II and III which form in HTH motif. Residues (Lys 11, Ala 14, Val 15, Leu 17, Tyr 18, Ile 29, Ala 30, Leu 33, Leu 40, Trp 43, Ile 44) marked with asterisk (*) are involved in molecular core and are depicted in Figure 1(b). The figure was generated using the program CHROMA.[26]

comprises six α-helices (I to VI) with residues 7–19, 26–34, 36–48, 62–72, 76–82, and 86–91 (PDB ID: 2JN6). α-helices I and II are arranged in antiparallel fashion and α-helix III is arranged perpendicular to these α-helices, yielding a structurally well-defined N-terminal HTH motif (1–50), a common double-stranded DNA binding motif (Fig. 1). In contrast, α-helices IV to VI are only locally well defined (Fig. 2); no long range NOEs were identified between any of the α-helices IV, V, and VI, or between these helices and the HTH motif.

The intrinsically flexible nature of the C-terminal segment (52–97) was confirmed by the observation of quite intense peaks in 2D [^{15}N,^{1}H]-HSQC, which reflect long T_2-spin relaxation times arising from large scale and rapid motional modes superimposed onto the overall rotational tumbling. The polypeptide segment (62–81) exhibits the "heptead repeat" characteristic for LZ, which often function as DNA binding motifs.[22] The program MULTICOIL[23] predicts that the LZ of protein Cgl2762 does not serve to mediate homodimer formation, which is consistent with our experimental observations that the protein is monomeric in solution.

When a z-score > 4.0 is used as the cut-off criterion, a search with the HTH domain (1–51) of protein Cgl2762 for structurally similar proteins using the program DALI[24] yielded 23 HTH domain containing proteins from 20 different Pfam families (Table SI). The sequence identity between the HTH domain of protein Cgl2762 and these 23 HTH domains varies between only 9 and 26%. Nonetheless, the r.m.s.d. values calculated for the polypeptide backbone atoms of the HTH domains show that all domains are structurally quite similar, demonstrating that the HTH domain of protein Cgl2762 in well within the range of thus far documented structural variation of HTH domains. Furthermore, a structure based multiple sequence alignment for the HTH domains of Table SI and protein Cgl2762 using the program STRAP2[25] (Fig. 3) shows that highly conserved residues either play an integral role for the architecture of the HTH domains (i.e., in the amino acid residue numbering of protein Cgl2762: Glu 9, Lys 11, Ala 14, Val 15, Leu 17, Tyr 18, Gln 27, Gln 28, Ile 29, Ala 30, Asn 31, Leu 33, Thr 39, Leu 40, Trp 43, Ile 44, Tyr 47), or are (likely) involved in nonspecific DNA binding (Glu 9, Lys 11, Gln 27, Asn 36, Thr 39, Asn 42, Lys 46).

Taken together, the NMR structure of protein Cgl2762 contributes to developing an atomic resolution picture for the functioning of the members of Pfam family PF01527 in DNA transposition. It is the first representative structure for this large family, providing high novel model leverage.[27] The HTH motif structure mediates binding to double-stranded DNA, consistent with its

implication in DNA transposition. An important first aim of future research might focus on identifying the protein binding partners with which members of this family form the LZ.

REFERENCES

1. Kalinowski J, Bathe B, Bartels D, Bischoff N, Bott M, Burkovski A, Dusch N, Eggeling L, Eikmanns BJ, Gaigalat L, Goesmann A, Hartmann M, Huthmacher K, Kraemer R, Linke B, McHardy AC, Meyer F, Moeckel B, Pfefferle W, Puehler A, Rey D.A, Rueckert C, Rupp O, Sahm H, Wendisch VF, Wiegraebe I, Tauch A. The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. J Biotechnol 2003;104:5–25.
2. Haniford DB. Transpososome dynamics and regulation in Tn10 transposition. Crit Rev Biochem Mol Biol 2006;41:407–424.
3. Haren L, Polard P, Ton-Hoang B, Chandler M. Multiple oligomerisation domains in the IS911 transposase: a leucine zipper motif is essential for activity. J Mol Biol 1998;283:29–41.
4. Rousseau P, Gueguen E, Duval-Valentin G, Chandler M. The helix-turn-helix motif of bacterial insertion sequence IS911 transposase is required for DNA binding. Nucleic Acids Res 2004;32: 1335–1344.
5. Liu G, Shen Y, Atreya HS, Parish D, Shao Y, Sukumaran DK, Xiao R, Yee A, Lemak A, Bhattacharya A, Acton TA, Arrowsmith CH, Montelione GT, Szyperski T. NMR data collection and analysis protocol for high-throughput protein structure determination. Proc Natl Acad Sci USA 2005;102:10487–10492.
6. Acton TB, Gunsalus KS, Xiao R, Ma LC, Aramini JM, Baran MC, Chiang ,YW, Climent T, Cooper B, Denissova N, Douglas SM, Everett JK, Ho CK, Macapagal D, Paranji RK, Shastry R, Shih L-J, Swapna GVT, Wilson M, Wu M, Gerstein M, Inouye M, Hunt JF, Montelione GT. Robot cloning and protein production platform of the Northeast Structural Genomics Consortium. Methods Enzymol 2005;394:210–243.
7. Kim S, Szyperski T. GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J Am Chem Soc 2003;125:1385–1393.
8. Atreya HS, Szyperski T. G-matrix Fourier transform NMR spectroscopy for complete protein resonance assignment. Proc Natl Acad Sci USA 2004;101:9642–9647.
9. Shen Y, Atreya HS, Liu G, Szyperski T. G-matrix fourier transform NOESY based protocol for high-quality protein structure determination. J Am Chem Soc 2005;127:9085–9099.
10. Delaglio F, Grzesiek G, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 1995;6:277–293.
11. Bartels C, Xia T, Billeter M, Güntert P, Wüthrich K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 1995;6:1–10.
12. Zimmerman DE, Kulikowski CA, Feng W, Tashiro M, Chien C-Y, Ríos CB, Moy FJ, Powers R, Montelione GT. Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 1997;269:592–610.
13. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996;8:477–486.
14. Word JM, Bateman RC, Presley BK, Lovell SC, Richardson DC. Exploring steric constraints on protein mutations using MAGE/PROBE. Protein Sci 2000;9:2251–2259.
15. Huang YJ, Powers R, Montelione GT. Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J Am Chem Soc 2005;127:1665–1674.
16. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shifts and sequence homology. J Biomol NMR 1999;13:289–302.
17. Herrmann T, Güntert P, Wüthrich K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm CYANA. J Mol Biol 2002;319:209–227.
18. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program CYANA. J Mol Biol 1997;273:283–298.
19. Huang YJ, Moseley HN, Baran MC, Arrowsmith C, Powers R, Tejero R, Szyperski T, Montelione GT. An integrated platform for automated analysis of protein NMR structures. Methods Enzymol 2005;394:111–141.
20. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography and NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54:905–921.
21. Koradi R, Billeter M., Wüthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graphics 1996;14:51–55.
22. Lupas A. Coiled coils: new structures and new functions. Trends Biochem Sci 1996;21:375–382.
23. Wolf E, Kim SP, Berner B. MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci 1997;6:1179–1189.
24. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–602.
25. Gille C, Frömmel C. STRAP: editor for STRuctural Alignments of Proteins. Bioinformatics 2001;17:377–378.
26. Goodstadt L, Pontin CP. CHROMA: consensus-based colouring of multiple alignments for publication. Bioinformatics 2001;17:845–846.
27. Liu J, Montelione GT, Rost B. Novel leverage of structural genomics. Nature Biotechnol 2007;25:845–851.

 DOI 10.1002/prot