PROTEINS: Structure, Function, and Bioinformatics 65:480–489 (2006)

Structure Determination of a New Protein From Backbone-Centered NMR Data and NMR-Assisted Structure Prediction

K. L. Mayer,[1,2] Y. Qu,[1,2] S. Bansal,[1,2] P. D. LeBlond,[1,2] F. E. Jenney Jr.,[1,2] P. S. Brereton,[1,2] M. W. W. Adams,[1,2] Y. Xu,[1,2] and J. H. Prestegard[1,2]*

[1]*Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia, 30602*
[2]*Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia, 30602*

ABSTRACT **Targeting of proteins for structure determination in structural genomic programs often includes the use of threading and fold recognition methods to exclude proteins belonging to well-populated fold families, but such methods can still fail to recognize preexisting folds. The authors illustrate here a method in which limited amounts of structural data are used to improve an initial homology search and the data are subsequently used to produce a structure by data-constrained refinement of an identified structural template. The data used are primarily NMR-based residual dipolar couplings, but they also include additional chemical shift and backbone-nuclear Overhauser effect data. Using this methodology, a backbone structure was efficiently produced for a 10 kDa protein (PF1455) from *Pyrococcus furiosus*. Its relationship to existing structures and its probable function are discussed. Proteins 2006;65:480–489.**

Key words: protein structure prediction; structural genomics; residual dipolar couplings; *Pyrococcus furiosus*; simulated annealing

INTRODUCTION

The structural genomics initiative set as one of its objectives the production of protein structures with novel folds.[1,2] The hope was that computational methods could use these structures to produce structural models for additional proteins with as little as 30% sequence identity to these experimentally determined structures. Finding targets with novel folds for structure determination has been challenging. Selection is frequently based on choosing proteins from sequence-clustered families that contain no members with experimental structures. In the Pfam classification,[3] for example, there are an estimated 7868 families (Pfam 17.0), only a third of which have one or more members with a structure in the protein data bank (PDB).[4] Logic dictates choosing targets for structure determination from families with no PDB representative. Despite the logic of this procedure, most structures of selected proteins still fall into one of the already well populated fold families.[1] This suggests the need for consideration of alternate procedures in both target selection and structure determination. For example, might a limited amount of easily obtained structural information allow classification to a previously populated fold family in the absence of a high level of sequence identity, and if such a classification were made, might a structure determination be facilitated by knowing its fold classification? Here we report the results of a new approach that answers these questions in the affirmative. It has two parts; it uses a homology search guided by a limited amount of NMR data (residual dipolar couplings, RDCs) to improve threading identifications of structural templates; it then produces a structure for the protein by refinement of a model threaded to templates, using additional NMR data from RDCs and backbone centered nuclear Overhauser effects (NOEs). The result is an expanded identification of a structural representatives for some previously unpopulated sequence-based Pfam families.

NMR is not generally regarded as an efficient approach to structure determination. Traditional methods are based primarily on NOEs.[5,6] These provide short range constraints that are most effective when significant numbers of NOEs between protons on sidechains packed at the protein core can be identified. This means that resonances from sidechains, in addition to the more easily identified backbone resonances, must be assigned. The assignment task extends the time required for structure determination to a period that can span many weeks.[6]

Abbreviations: DSS, dimethyl-silapentane-sufonate; DTT, dithiothreitol; HSQC, heteronuclear single quantum coherence; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; RDC, residual dipolar coupling; TOCSY, total correlation spectroscopy.

Grant sponsor: The National Institutes of Health's Protein Structure Initiative; Grant number: GM062407; Grant sponsor: DOE; Grant number: FG05-95ER20175; Grant sponsor: The Geogia Research Alliance and the National Science Foundation; Grant number(s): NSF/DBI-0354771, NSF/ITR-IIS-0407204.

*Correspondence to: J. H. Prestegard; CCRC, 315 Riverbend Rd, Athens, GA 30602. E-mail: jpresteg@ccrc.uga.edu

Received 22 November 2005; Revised 19 April 2006; Accepted 24 May 2006

Published online 22 August 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21119

There are other types of structurally useful NMR data that do not require such extensive resonance assignment and analysis time. We have previously reported the use of RDC data in both target screening and structure determination under conditions where no prior resonance assignment is required.[7–9] Other authors have suggested similar procedures, although most do require prior resonance assignment.[10–15] The key to all of these procedures is the fact that constraints from RDC data (at least those from directly bonded pairs of atoms such as $^{15}N-^{1}H$ in amide bonds and $^{13}C-^{1}H$ in Cα—Hα bonds) are orientational in character and can utilize information from backbone positions without the close approach of the nuclear pairs being observed. We rely heavily on RDCs for the target screening and structure refinement procedure presented here, but we proceed with significantly less RDC data than required for most previous methods.

The test case is the 10 kDa (88 residues) protein from hyperthermophilic archaeon *Pyrococcus furiosus* corresponding to the coding sequence PF1455. This protein is not currently assigned to a PfamA family, but it has homologs in closely related archaeal species (*P. horikoshii*, *P. abyssi*, and *Thermococcus kodakaraensis*); none of which currently have experimentally determined structures. PF1455 also has a 32-residue C-terminal segment that is included in Pfam01978, a family that contains a sugar-specific transcriptional regulator (TrmB, in *P. furiosus* this is encoded by PF1743). This family also has no experimental structural representative.

To further minimize the probability of producing a structure with a known fold, the additional step of threading the sequence of PF1455 against a structural database was taken. Threading programs are normally used to identify templates for structural modeling. When no suitable template can be found, however, the probability of a sequence producing a novel structure should rise. In our case, both Genthreader[16,17] and SP3[18] threading programs were initially used to identify templates. These programs score templates based on favorable residue–residue contacts among other factors. In neither case was a template with a score higher than a "guess" (E value less than 5.5 or Z-score greater than 5.2) identified. Subsequently, we used the 3D-Jury protocol to identify templates.[19] This protocol compares results from a suite of threading programs and scores based on similarity of identified templates. The highest Jscore was 34, well below the level of a confident prediction. Although the highest scoring templates shared a ferredoxin-like fold, available evidence indicated that PF1455 could adopt a novel fold that would be valuable in modeling many new sequence-based homologs. The results from this study will show that the fold is, in fact, not novel, but can be classified as a common ferredoxin-like fold. However, we also show that the probable fold can be identified, and a refined structure produced, with a small amount of easily obtained NMR data and the use of new computationally assisted structure determination techniques.

MATERIALS AND METHODS

Gene Cloning, Protein Expression, and Purification

Cloning of PF1455

PCR primers were designed based on the *P. furiosus* (DSM 3638) genome sequence obtained from the NCBI GenBank file (RefSeq NC_003413). The PCR product was cloned using standard techniques[20] into the expression vector pET-24d (Novagen, Madison, WI), which had been modified to include a His-tag (MAHHHHHHGS-) at the N-terminus to facilitate immobilized metal affinity chromatography purification. *Escherichia coli* strain TOP10 (Invitrogen, Carlsbad, CA) was used for cloning and *E. coli* strain BL21(DE3)Star (Invitrogen), supplemented with the pRIL plasmid encoding tRNAs for rare codons (Stratagene, La Jolla, CA) was used for expression and short-term glycerol stock storage.

Growth medium

M9 minimal medium[20] was utilized for cell growth and expression, with glucose (0.3% w/v) as the carbon source and 0.1% (w/v) ammonium-^{15}N chloride (Isotec, Miamisberg, OH) as the nitrogen source. The medium was further supplemented with thiamine hydrochloride (0.0174% w/v), a vitamin mixture,[21] and a metal mixture[22] modified to contain the recommended concentration of iron and zinc and one tenth the recommended concentration of the other trace metals. Kanamycin and chloramphenicol were added to final concentrations of 100 and 25 μg/mL, respectively. For ^{13}C-labeling of the recombinant protein, either D-glucose-2-^{13}C and D-glucose-1-^{13}C (Cambridge Isotope Laboratories, Andover, MA) at 0.1 and 0.2% (w/v), respectively, or D-glucose-$^{13}C_6$ (Isotec) at 0.3% (w/v) were substituted for the unlabeled glucose. The 2-^{13}C, 1-^{13}C labeling procedure produces a protein with a near-random distribution of ^{13}C at a level of approximately 16% in addition to uniform (U) labeling with ^{15}N at approximately 98%.[8]

Large-scale expression

Glycerol stocks were used to inoculate 30-mL starter cultures of the M9 minimal medium, which were shaken at 210 rpm at 37°C for 18 h. The following day, 15 mL of the culture was used to inoculate 1 liter of M9 minimal medium in a 2.5 L Fernbach flask. The culture was incubated at 37°C (150 rpm) for 7.5 h, induced with 1 m*M* isopropyl-β-D-1-thiogalactopyranoside and then incubated at 37°C (150 rpm) for a further 7 h. The cells were harvested by centrifugation (5487*g*, 30 min), resuspended in 20 mL of buffer A (20 m*M* sodium phosphate, 500 m*M* NaCl, 10 m*M* imidazole, pH 8.0), and then stored at −20°C.

Purification of recombinant protein

The resuspended cell pellet was thawed, lysed by sonication using a Sonifier 450 (Branson Ultrasonics, Danbury, CT), the cell extract was clarified by centrifugation (29000*g*, 45 min) and loaded onto 3 mL of NiNTA immo-

bilized metal affinity chromatography resin (Novagen) in a 1.5 × 10 cm Econo-column (Bio-Rad, Hercules, CA), preequilibrated with at least 10 bed volumes of buffer A. The column was washed with 5 bed volumes of buffer A and the protein block-eluted with 3.3 bed volumes of 300 m*M* imidazole in buffer A. The eluted protein was concentrated to <1 mL, diluted to 15 mL with buffer B (20 m*M* Tris, 2 m*M* dithiothreitol, pH 8.0), and loaded at 3 mL/min onto a 5 mL HiTrap Sepharose Q column (GE Healthcare, Piscataway, NJ), preequilibrated with buffer B. The column was washed with 10 bed volumes of buffer B and the protein was eluted with a linear gradient (20 bed volumes) of 0 to 1*M* NaCl in buffer B in 3 mL fractions. Fractions that absorbed at 280 nm were visualized on a 4–20% Criterion SDS-PAGE gel (Bio-Rad) and fractions with a major protein with an apparent molecular weight of 10 kDa were pooled. The combined fractions were concentrated to <2 mL and loaded at 1 mL/min onto a Superdex 30 16/60 gel filtration column (GE Healthcare), preequilibrated with 20 m*M* Tris, 300 m*M* NaCl, 2 m*M* DTT, pH 8.0. Fractions (3 mL each) were collected and the PF1455-containing fractions were pooled, concentrated to ~500 μL, diluted to 6 mL with 20 m*M* Tris, 50 m*M* NaCl, 2 m*M* DTT, pH 8.0 (to give a final concentration of ~70 m*M* NaCl) and concentrated to ~500 μL. Approximately 40 μg of protein was run on an SDS-PAGE gel to assess purity.

NMR Data Collection and Analysis

NMR sample preparation, including alignment

For measurements under isotropic conditions, a sample of 16% ^{13}C/U-^{15}N labeled PF1455 was prepared at a concentration of 2 m*M* in 20 m*M* Tris and 70 m*M* NaCl at pH 7. All samples also contained 2 m*M* DTT, 0.02% azide, 1 m*M* DSS and 10% D_2O. An anisotropic sample is required for the measurement of RDCs. After isotropic data collection, the 16% ^{13}C/U-^{15}N PF1455 sample was used to prepare two partially aligned samples to satisfy this requirement. A sample with pf1 phage as the alignment medium[23] was prepared which contained 1.14 m*M* PF1455 and 10 mg/mL phage in 11 m*M* Tris and 40 m*M* NaCl. After equilibration at room temperature overnight at 22°C, the sample showed a deuterium splitting of 8.8 Hz when placed in the magnet. A second aligned sample was prepared using C12E5 bicelles as the alignment medium.[24] This sample contained 1.2 m*M* in PF1455 in 3% (w/v) C12E5/hexanol at a molar ratio of 0.98 in 20 m*M* Tris and 70 m*M* NaCl at pH 7. After equilibration at room temperature overnight at 22°C, the sample showed a deuterium splitting of 13 Hz when placed in the magnet. An additional isotropic sample containing ^{15}N PF1455 instead of ^{13}C/^{15}N PF1455 was also prepared and used for the ^{15}N edited NOE, TOCSY, and HNHA experiments to be described below.

NMR data collection

NMR data were collected on a Varian Unity Inova 600 MHz spectrometer at 298 K using a conventional z-gradient triple resonance probe or a z-gradient triple resonance cryogenic probe (Varian, Palo Alto, CA). Two experiments were run using the conventional probe for measurement of RDCs: a soft HNCA-E-COSY[7,25] and a ^{15}N IPAP-HSQC.[26] Data were acquired for the isotropic and the pf1 phage sample using both experiments to provide a complete set of ^{15}N—^{1}HN, ^{13}CA—^{1}HA, and ^{1}HA—^{1}HN RDCs. Only the ^{15}N IPAP-HSQC data were collected on the C12E5 sample to provide a partial data set in a second alignment medium. Data collection for the soft HNCA-E-COSY included 72 t1, 16 t2, and 2048 t3 points collected over 72 h. Data collection for the ^{15}N IPAP-HSQC included 256 t1 points, and 2048 t2 points collected over 12 h. RDCs were calculated as the difference of the coupling measured in the aligned and isotropic conditions.

For assignment purposes, a standard HNCAHA experiment was collected on the isotropic sample. This aided in making sequential residue connections. In addition, ^{15}N-edited NOESY, ^{15}N-edited TOCSY, and HNHA data sets were collected on the isotropic ^{15}N sample using the standard probe and pulse sequences as implemented in the spectrometers employed. The 3-D ^{15}N experiments were collected with 64 t1, 16 t2, and 2048 t3 points over 16 h. Chemical shift assignments were determined and used to assign RDCs to specific residues as well as to assign NOE cross-peaks. NOEs involving HN—Hα and HN—HN connections were identified for use in energy minimization.

NMR data processing and analysis

All data were processed using NMRPipe and visualized using NMRDraw.[27] Peaks were picked using the automatic picking procedure in NMRDraw. Arbitrary assignments were automatically transferred in from the HSQC and the splittings (J or J + D) calculated using a series of tcl scripts modified from NMRDraw. Intraresidue and interresidue designations were automatically assigned for the HNCA-E-COSY based on the isotropic $^{3}J_{HNH\alpha}$ value (zero for interresidue). Text files containing chemical shifts and splittings were inserted into a mySQL database. A table of RDCs was generated from the difference between splittings in aligned and isotropic datasets.

Peaks in NOE and TOCSY data sets were automatically picked in NMRDraw and manually analyzed to provide information on secondary structure and backbone to backbone distance restraints. TOCSY spectra were manually analyzed to determine amino acid types and these were combined with Cα to Cα connectivities from the HNCA-E-COSY experiment to make sequential assignments. The NOEs were used along with Cα chemical shifts and $^{3}J_{HNHA}$ scalar coupling values from the isotropic soft HNCA-E-COSY experiment to identify secondary structure types.[28] To aid in structure determination, a set of short range and long range interresidue NOEs was identified and treated more quantitatively. Peak intensities were compared separately with average intensities

of HA—HN and HN—HN peaks from regions known to be in well defined α-helices and classed based on intensities being the same or stronger than the HN—HN standard (strong), the same as the HN—HA standard (medium) or weaker than either (weak). Initially, chemical shift tables were used to identify interresidue NOE cross-peaks that could be unambiguously assigned based on chemical shifts. In making these assignments errors of 0.1 ppm were used for both HN shifts and HA shifts. After an initial refinement, additional interresidue NOEs previously classified as ambiguous were resolved based on distances between proton pairs that were incompatible with observation of NOEs (>5 Å). These were used in a second stage of refinement.

Modeling of the Structure of PF1455

Template identification using RDC-PROSPECT

To obtain starting structural models of PF1455, the protein threading program RDC-PROSPECT[15] was used to find structural homologs in the SCOP[29] database that best match the experimental RDC data. No sequence homology and secondary structure information was used in RDC-PROSPECT in this prediction, so this is best used in conjunction with conventional threading programs. Using only the $^{15}N-^{1}H$ RDC data from the phage oriented set, RDC-PROSPECT identified the PDB structures 1cc8 and 1dd3 as top hits. Using the $^{15}N-^{1}H$ RDC data from C12E5 bicelle oriented set, the PDB structures 1cc8 and 1fvq were identified. 1cc8 and 1fvq both occur in lists from conventional threading (though not at the top of the lists), while 1dd3 does not. Moreover, 1dd3, an all α protein, does not agree with NOE and chemical shift patterns indicative of secondary structure. Hence, only the coordinates of 1cc8 and 1fvq were supplied to the homology modeling program MODELLER[30] to produce starting points for PF1455 structure determination.

Refinement using XPLOR-NIH

Refinement of the initial structure was carried out by implementing a simulated annealing protocol in the XPLOR-NIH package.[31] The initial stage, done in vacuum, used RDCs from both media (except for a set of 15% randomly excluded values to be used in validation), the unambiguous set of short range and long-range NOE data, and scalar HN—HA three bond couplings from the HNCA experiment. The RDCs were treated as harmonic constraints in the SANI module of XPLOR. Weightings of various RDC data sets were chosen to give convergence of back-calculated RDCs to approximately the estimated error in measurement (4 Hz for HN RDCs). Principal alignment parameters and rhombicity parameters were initially estimated from the distribution of H—N couplings in each medium and then refined by using order parameters back-calculated from the structure produced after the first cycle of minimization. These alignment parameters were scaled to correct for the expected difference in sizes of C—H and N—H couplings so that couplings could be entered as actual values in Hz. A set of axis coordinates was added for each medium and these were allowed to rotate freely with respect to protein coordinates to adjust to the preferred orientation of the alignment tensor. NOE constraints were employed as flat-bottomed parabolas with upper limits at 2.7, 3.0, and 5.0 Å for strong, medium and weak constraints. Torsional constraints based on secondary structure identification from TALOS[32] and dihedral angles from idealized structures were used along with molecular shape constraints (radius of gyration) to keep the protein structure compact.

For simulated annealing, the system was first equilibrated at 400 K for 1000 steps, followed by 10000 steps of cooling from 400 to 1 K. At the 1 K temperature step the NOE force constant was increased from 2 to 50, the dihedral constant was kept constant at 700, and dipolar coupling force constant was increased from 0.005 to 2. The annealing sequence was then repeated 20 times. This cycle was started independently 100 times from the predicted structure and the final set of structures overlaid to give RMSD values for the backbone atoms. A number of parameters (maximum temperature, number of annealing cycles, weighting factors) were varied to optimize the extent of convergence and exploration of conformational space.

A second stage of refinement was performed by refining the protein structure in explicit water. The inclusion of water can help in removing artifacts (if any) due to unrealistic representation of molecular forces in the vacuum refinement. It can also facilitate the rearrangement of certain bonds (hydrogen or electrostatic) that may have higher costs of disruption in the vacuum simulation. We used the water refinement protocol from Linge et al.[33] This consisted of three stages, a heating stage from 0 to 400 K in steps of 100 K with 150 steps of molecular dynamics at each temperature, a short refinement stage with 1000 steps at 400 K, and a cooling stage from 400 to 0 K in steps of 20 K with 100 steps of molecular dynamics at each temperature. This cycle was repeated four times, independently starting with each predicted structure from the refinement step in vacuum. The final types and numbers of constraints used are summarized in Table I along with error estimates and weighting factors for various data types.

Assessment of structure quality

To assess the progress of refinement and validate final structures, RDCs were back-calculated from structures in each case using the program REDCAT,[34] and these were compared with experimental values. For validation, coordinates corresponding to the atom pairs giving rise to the randomly excluded RDCs were entered into the program and the best set of order parameters from a fit to data used in refining the structure were used to back-calculate RDCs. Calculated RDCs were then plotted versus experimental RDCs and a quality factor[35] was calculated to evaluate structure reliability.

TABLE I. Parameters Used for the RDC Restraints in XPLOR-NIH

Type of data	Total # of data	Sequential	Long range	Force constant
HN-Phage[a]	63	—	—	4
HACA-Phage[b]	61	—	—	4
HN-C12E5[c]	54	—	—	8
Unambiguous NOE[d]	54	45	9	50
Secondary NOE[d]	37	18	19	50
Dihedral	116	—	—	700
Radius of gyration	—	—	—	50

[a]The values of Rhombicity (R) and Anisotropy (Da) are −4.66 and 0.5288, respectively.
[b]The values of Rhombicity (R) and Anisotropy (Da) are 8.06 and 0.5288, respectively.
[c]The values of Rhombicity (R) and Anisotropy (Da) are 2.6 and 0.443, respectively.
[d]There were six NOE violations in the refinement step.

RESULTS

Given sequence homology and traditional threading results indicating a probable unique structure, the initial intention was to proceed to a structure determination using a de novo structure determination method such as our previously described backbone RDC approach.[36] However, the limited numbers of RDC data obtained, under sample conditions required for this protein, prevented convergence to a unique structure using an approach based entirely on unassigned RDCs. In the course of collecting additional data, we obtained sufficient numbers of assigned RDCs to consider a second round of screening for known structure motifs using assigned RDC data as opposed to just sequence data. There are now a number of programs that allow threading of new sequences through structural databases using limited sets of experimental data.[10,15] It makes a great deal of sense to use such programs to screen for structure homology as new data are acquired. In this way, structural homologs that may have been missed with sequence-based screening can be found.

Data Collected

With additional assignments, some NOE data, and some scalar coupling data, we were also able to determine secondary structure and use agreement with this structure as validation for new predictions. Experiments for the actual resonance assignment were chosen to be compatible with the partial carbon labeling achieved in media containing a mixture of C1 and C2 labeled glucose. Partial labeling with ^{13}C excludes experiments relying on $^{13}C—^{13}C$ magnetization transfers, but for small proteins this is not a severe limitation, and such labeling simplifies measurements of certain RDCs. As a first step, collections of resonances belonging to specific residues were assigned to sequential positions in fragments by matching the interresidue and intraresidue Cα shifts from the HNCA-E-COSY experiment and the Hα shifts from the HNCAHA experiment. The HNCA-E-COSY experiment also gives RDC data that are described below. In this manner, three fragments were generated that corresponded to three regions of the amino acid sequence separated by proline residues. Assignment of the fragments to the sequence using easily observed TOCSY patterns for residues such as alanine, threonine, and valine was relatively straightforward, and resulted in sequence-specific backbone assignments of 85% of the protein residues. No resonances were observed in any spectra for residues G76 through K87 except for the C-terminal residue A88, which is most likely the result of increased dynamics in this region. This segment accounts for nearly all missing assignments. Assignments for all observed resonances have been deposited with the BMRB (accession number 7073).[37]

RDCs were measured for PF1455 under phage and C12E5 alignment conditions. After elimination of sets contaminated by noise or spectral overlap, a total of 124 rdcs were available from the phage sample, 63 NH RDCs from an IPAP-HSQC experiment and 61 HACA RDCs from the HNCA-E-COSY experiment. A total of 54 NH RDCs were measured for the C12E5 aligned sample from an additional IPAP-HSQC experiment.

Identification of a Structural Homolog

We chose the program RDC-PROSPECT to carry out a homology search. It can use a variety of assigned RDC data in a structural search[15]; however, we employed it using just the set of HN RDCs from the phage medium and the E12C5 medium. The PDB structures 1cc8 and 1dd3 were identified as the two top hits using RDC-PROSPECT and the phage data. The PDB structures 1cc8 and 1fvq were identified as the top hits using the E12C5 data. Analysis of secondary structure predictions from additional NMR data allowed selection of preferred structures for assignment. Analysis of the NOE patterns, $^{3}J_{HNHA}$, and $^{13}C\alpha$ and $^{1}H\alpha$ chemical shift indices resulted in the determination of the secondary structure pattern of PF1455 [Fig. 1(a)]. This analysis suggests that the protein contains four antiparallel β-strands and two α-helices arranged in a β–α–β–β–α–β topology. This topology matches 1cc8 and 1fvq very well, as depicted in Figure 1. In particular, the sequence segments with red chemical shift deviation bars and small coupling constants, indicative of α-helices, overlay well with α-helical segments in 1cc8 and 1fvq. Furthermore, the top templates predicted by the sequence-based threading programs, although having low confidence scores, belong to the same protein structure family as 1cc8 and 1fvq. Based on the common fold, the RDC-dependent prediction results, and the sequence-dependent prediction results, it is very likely that PF1455 is structurally homologous to 1cc8 and 1fvq, despite its low sequence identities (11 and 4%). Therefore, we used the coordinates of 1cc8 and 1fvq along with the homology modeling program MODELLER[30] to produce initial structures

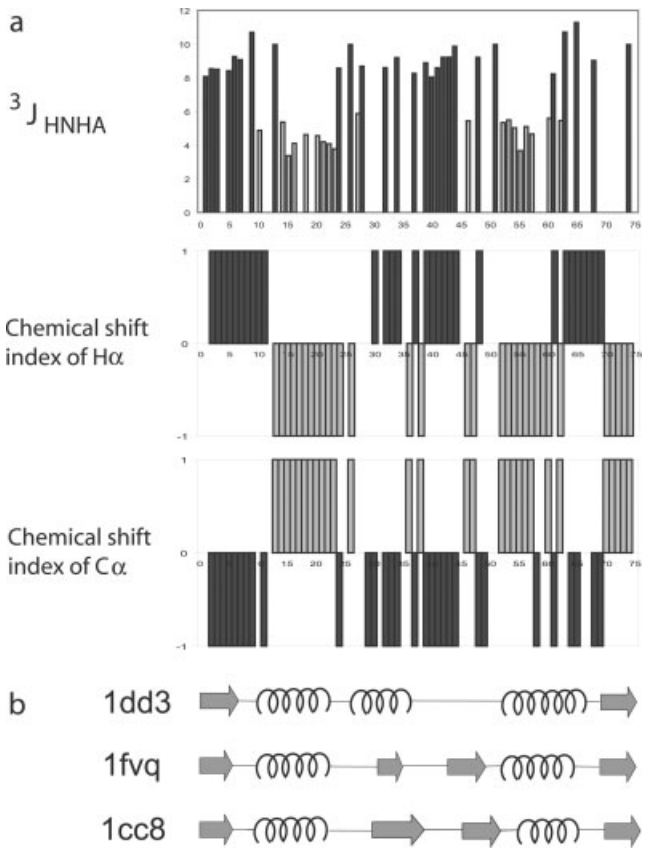

Fig. 1. (**a**) Secondary structure patterns for PF1455 were determined using $^3J_{HNHA}$ scalar couplings and chemical shift index of Hα and Cα as shown in Figure 1(a). The secondary structure patterns of the modeled proteins are depicted in Figure 1(**b**). The model for 1cc8 exactly matches the secondary structure pattern determined experimentally.

of PF1455. The initial structure from the 1cc8 template is depicted in Figure 2(a).

Refinement of a Structural Model

The initial models of PF1455 (minus the C-terminal 12 residues for which no data were obtained) were refined using XPLOR-NIH,[31] first with the entire RDC data set and the unambiguous NOEs, and then with all RDCs and 91 NOEs as restraints. The small number of distance constraints and the multiple minima inherent in RDC-based constraints prohibit using a normal high temperature simulated annealing protocol for this refinement.[38,39] Since there is good reason to believe the starting structures should be close to the final structures, a low temperature simulated annealing protocol was used, followed by a water refinement step (see methods section). This protocol was optimized to explore the maximum of unconstrained conformational space while still allowing convergence of back-calculated RDCs to experimental values (within experimental error). A trial starting from the 1cc8 model and not using RDC and NOE constraints produced a structure with a backbone RMSD from the starting structure of 2.6 Å, indicating a reasonable extent of structure exploration.

Application of the refinement procedure to the 1cc8 model using 85% of the RDC constraints and all of the NOE constraints moved structures from RDC quality factors of 0.75 and 0.96 for the initial structure in two media to quality factors of 0.27 and 0.34 for the final structure. These smaller values represent a significant improvement in structural agreement with RDC data.[35] In addition, only 2 of the 91 NOE constraints remained in violation for the best structures. A plot of back-calculated versus experimental RDCs for the set of data used in the above minimizations is given in Figure 3(a) and a plot for the 15% randomly excluded data is given in Figure 3(b). The lowest energy refined structure was used for back-calculations in each case. The scatter in the plots approximately reflects the expected experimental accuracy (4 Hz). The quality factor for the excluded data is 0.62.

The refined structure of PF1455 produced from the 1cc8 model is presented in Figure 2(b). The structure has preserved the two helices of the starting structure, three of the four strands, and the backbone in other regions lies close to the position of the original strands. However, the first helix is shorter and the second helix displays a somewhat different angle relative to the strands. Overall, the backbone atoms show a 2.4 Å RMSD deviation from the initial model. The structure shown is a single minimum energy representation. However, this is just one member of a set produced by starting the annealing protocol from the same initial structure 100 times. A set of 10 overlayed, fully converged structures, is presented in Figure 2(c). This set shows a backbone RMSD over the set that averages 1.5 Å. According to literature comparisons,[32] 1.8 and 2.5 Å X-ray structure would yield a quality factor of 0.25. The quality factors for RDCs given above suggest a structural precision of approximately 3 Å.

As an additional test of precision, we can compare a structure refined starting from the 1fvq model. The initial model from 1fvq deviates from the initial model from 1cc8 by an RMSD over backbone atoms of 3.3 Å and from the final structure starting from the 1cc8 model by an RMSD of 4.2 Å. After a refinement procedure similar to that used for the 1cc8 model, the best structure starting from 1fvq deviates from its starting point by an RMSD of 3.7 Å and converges to within an RMSD of 3.4 Å of the final 1cc8 model. However, this structure has four as opposed to two NOE violations. Three of these are in the 12–28 segment that encompasses the first short helix, suggesting that the annealing procedure is unable to properly position this helix. If the N-terminus, including this helix is excluded from comparison, the RMSD over backbone atoms is 2.6 Å. Ramachandran analysis also suggests that the 1fvq structure is of somewhat lower quality. For the final structures starting from 1cc8 and 1fvq 62% and 58% of the residues are in the most favored region, 25 and 27% are in the additionally allowed region, 9 and 12% are in the generously allowed region and 4 and 3% residues are

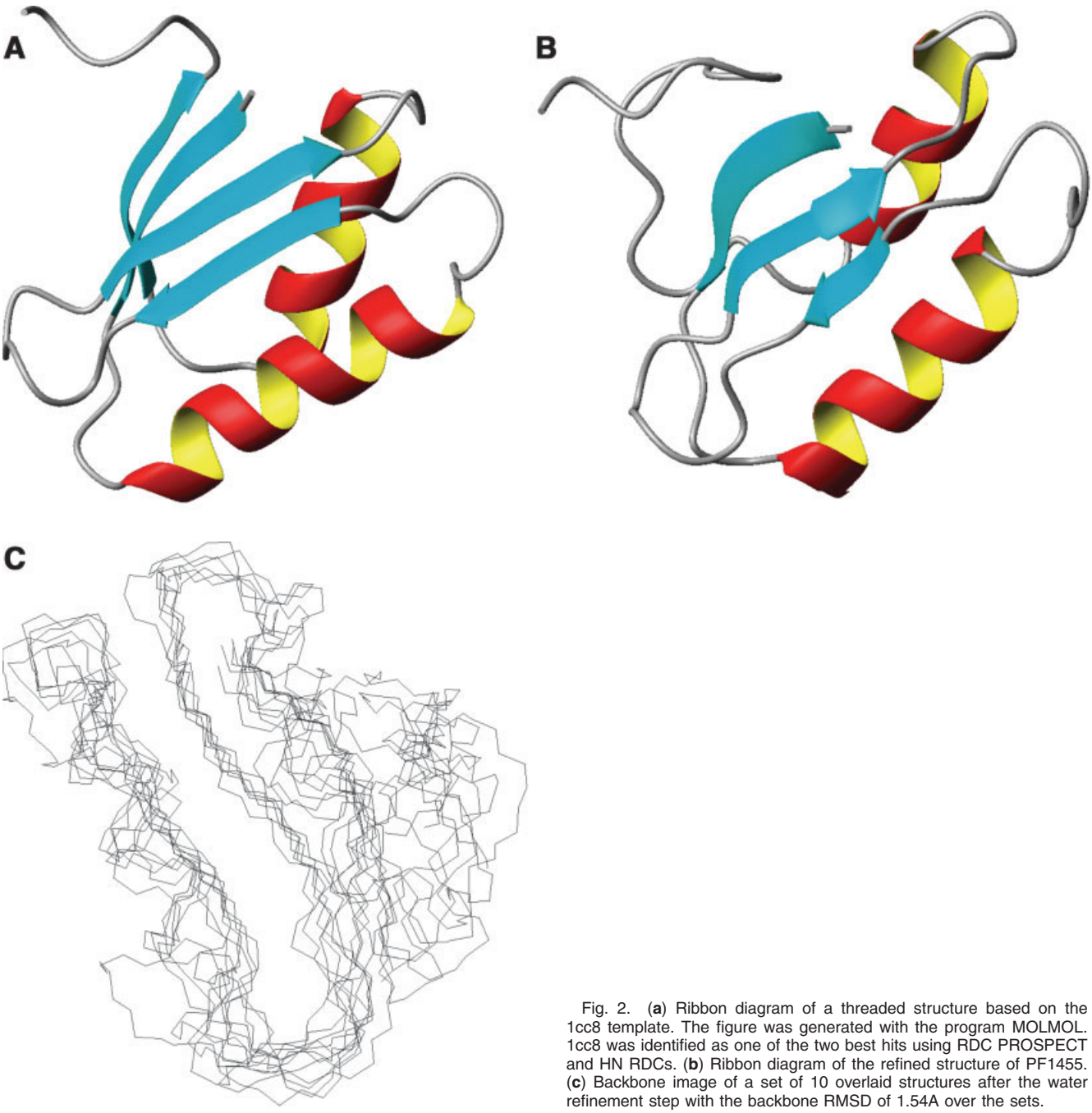

Fig. 2. (**a**) Ribbon diagram of a threaded structure based on the 1cc8 template. The figure was generated with the program MOLMOL. 1cc8 was identified as one of the two best hits using RDC PROSPECT and HN RDCs. (**b**) Ribbon diagram of the refined structure of PF1455. (**c**) Backbone image of a set of 10 overlaid structures after the water refinement step with the backbone RMSD of 1.54A over the sets.

in the disallowed region. The set of structures shown in Figure 2(c) for the 1cc8 model has been deposited with the PDB (2F40).

DISCUSSION

The model of PF1455 presented in Figure 2(b), like 1cc8 and 1fvq, can be described as an α and β structure with a ferredoxin-like fold [(β–α–β) × 2], a very well populated fold family. Nevertheless, it is now a structural representative in a small, but previously (structurally) uncharacterized sequence-clustered family. The structure produced is, of course, just a backbone structure, but it is of sufficient quality to not only identify the protein fold, but also properly place residues on the fold topology. The assignment of backbone resonances and the backbone structure can also provide an excellent starting point for more complete structural exploration when this effort is justified. The backbone NMR assignments will facilitate collection of additional sidechain assignments and the backbone structure can be used to reduce ambiguity in NOEs by setting limits on closest

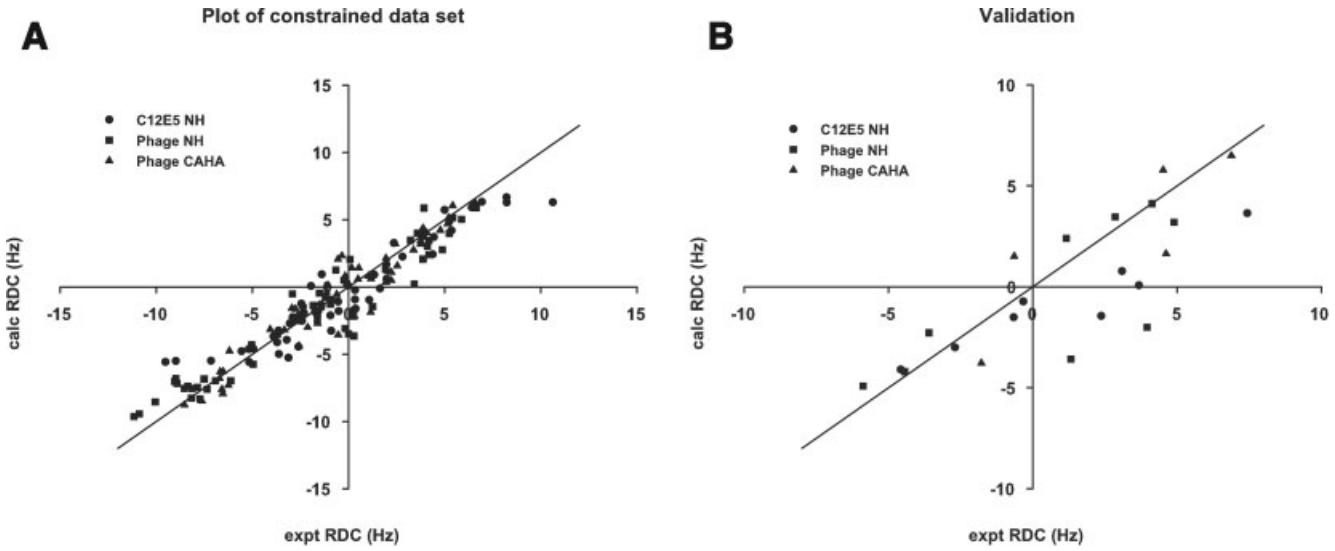

Fig. 3. Plot of calculated RDC vs experimental RDC for NH and CAHA couplings from Phage and NH couplings from C12E5(PEG) media for a constrained set (**a**) and a randomly 15% excluded data set (**b**). From the above data, three outliers were removed as they arise from regions lacking recognizable secondary structure that are not likely to be well constrained by data on adjacent residues.

approach of protons on particular residue pairs. There is also considerable research ongoing into methods for building sidechains into experimental backbone structures.[40–42] This is certainly successful with very high precision backbone structures. The estimated precision of backbone atom positions in our structures is close to 3 Å. It is unclear if this precision is adequate to use the backbone as a basis for accurately building in side chain positions.

Aside from purely structural interest, the backbone structure produced provides an excellent starting point for the identification of possible functional characteristics. PF1455 is annotated as a hypothetical protein in the TIGR database but can be classified as a heavy metal binding/transport/detoxification protein stemming from similarities between residues 12 and 60 of PF1455 and members of the InterPro[43] IPR006121 family, a large group represented in all three domains. However, for PF1455 the similarity is characterized by a modest *e* value (0.000755), and PF1455 lacks the two conserved cysteine residues for metal-binding (it has no Cys residues). It does have three ExxE motifs, is negatively charged (pI 4.9), and has an interesting D–E–D cluster (residues 21, 46, 53) between the two α-helices that could serve as a metal binding site. FprintScan[44] shows extremely weak similarity (e value of 2.4) for two regions (residues 14–28 and 40–48) to the von Willebrand factor, type A (IPR002035). This is a three element motif found in proteins including the integrin I domain. These often form surface metal binding sites that may be important in protein–protein interactions.

Having a structure for PF1455 allows a direct search for other structural homologs. Using the DALI program,[45] the highest similarity score (7.2), was with 1mwy, an intracellular Zn transporting protein.[46] The next two highest scores (6.1 and 5.8) belong to a Cu binding metallochaperone (1qup)[47] and a Cu transporting ATPase (1aw0).[48] All of these proteins have similar RMSD deviations of their backbones from the refined PF1455 structure (2.2–2.5 Å) but have less than 10% sequence identity.

The general notion of a protein involved in protein–protein interactions and metal binding is supported to some extent by biochemical data. PF1455 is predicted to be part of a two gene operon that also contains PF1454, which encodes a much larger Cys-rich protein (68.7 kDa, 11 Cys). Both of these genes are expressed in *P. furiosus* under the usual laboratory growth conditions, as shown by DNA microarray data.[49] Their expression is also coregulated, as both are up-regulated in response to both cold shock[49] and iron limitation (Menon A., unpublished data). PF1454 is a homolog of MoaA, which contains a 4Fe–4S cluster and is involved in the pathway that converts GTP ultimately to tungstopterin, which is used in the synthesis of iron-tungstoenzymes. Although tungsten is assimilated as the oxyanion rather than as a cation, it would not be unreasonable for PF1455 to serve as an iron-dependent chaperone or regulator, perhaps binding ferrous iron via the carboxylates of the DED cluster. Attempts to coexpress PF1454 and PF1455 and examine their metal-binding properties are underway. Structural work on PF1455 may thus have laid the groundwork for future studies of the nature of protein–protein and metal–protein interactions in pterin biosynthesis in *P. furiosus* and related organisms.

CONCLUSIONS

Much of the current structural genomics initiative is directed at the search for novel protein folds so as to fill out fold space, or at least to provide structural representatives for families clustered on the basis of sequence.

The structure presented here does not contribute from this perspective. However, by using a homology search that incorporated a limited amount of experimental data and a structure refinement procedure based on backbone RDCs and NOEs, structural similarity to proteins with a common fold was quickly identified and a backbone structure was efficiently produced. This in turn led to suggestions as to possible functional roles. With conventional screening, a traditional NMR structure determination requiring three to four times more data acquisition time would have been undertaken, and it is unlikely that this would have led to additional conclusions.

There are likely to be a significant number of proteins for which a structural homolog can be identified or a structure refined by the procedure described. A recent analysis, which included the *Pyrococcus furiosus* genome, suggested that homologs for about 60% of the proteins could be found by a combination of conventional sequence and threading based searches.[50] The refinement procedures described should be applicable to a significant fraction of the 60% judged to be small enough for study by NMR (35% of the genes in the *P. furiosus* genone (767 of 2198) are predicted to encode proteins of less than 20 kDa).[51] Excluding membrane proteins (20%) and those that are too large for NMR, about 17% of the total genome should have conventionally identified homologs that can be refined using the methods described ($60\% \times 0.80 \times 0.35 = 17\%$). While the single application we have presented does not allow an estimate of the success rate in finding homologs for the remaining 40% of the genome, the large number of new protein structures that fall into existing fold families would suggest it to be high. Thus, it is clear that routine application of experimentally assisted homology searches, such as that presented here, should be considered as a part of future targeting of proteins for structure determination in structural genomics programs.

REFERENCES

1. Todd AE, Marsden RL, Thornton JM, Orengo CA. Progress of structural genomics initiatives: an analysis of solved target structures. J Mol Biol 2005;348:1235–1260.
2. Norvell JC, Machalek AZ. Structural genomics programs at the US National Institute of General Medical Sciences—Foreword. Nat Struct Biol 2000;7:931.
3. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32:D138–D141.
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
5. Wuthrich K. NMR studies of structure and function of biological macromolecules (Nobel lecture). Angew Chem Int Ed Engl 2003; 42:3340–3363.
6. Montelione GT, Zheng DY, Huang YPJ, Gunsalus KC, Szyperski T. Protein NMR spectroscopy in structural genomics. Nat Struct Biol 2000;7:982–985.
7. Tian F, Valafar H, Prestegard JH. A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. J Am Chem Soc 2001;123: 11791–11796.
8. Valafar H, Mayer KL, Bougault CM, LeBlond PD, Jenney FE, Brereton PS, Adams MWW, Prestegard JH. Backbone solution structures of proteins using residual dipolar couplings: application to a novel structural genomics target. J Struct Funct Genomics 2004;5:241–254.
9. Prestegard JH, Mayer KL, Valafar H, Benison GC. Determination of protein backbone structures from residual dipolar couplings. Nucl Magn Reson Biol Macromol Part C. Methods Enzymol 2005;394:175–209.
10. Rohl CA, Baker D. De novo determination of protein backbone structure from residual dipolar couplings using rosetta. J Am Chem Soc 2002;124:2723–2729.
11. Meiler J, Baker D. Rapid protein fold determination using unassigned NMR data. Proc Natl Acad Sci USA 2003;100:15404–15409.
12. Haliloglu T, Kolinski A, Skolnick J. Use of residual dipolar couplings as restraints in ab initio protein structure prediction. Biopolymers 2003;70:548–562.
13. Andrec M, Harano Y, Jacobson MP, Friesner RA, Levy RM. Complete protein structure determination using backbone residual dipolar couplings and sidechain rotomer prediction. J Struct Funct Genomics 2002;2:103–111.
14. Delaglio F, Kontaxis G, Bax A. Protein structure determination using molecular fragment replacement and nmr dipolar couplings. J Am Chem Soc 2000;122:2142,2143.
15. Qu YX, Guo JT, Olman V, Xu Y. Protein structure prediction using sparse dipolar coupling data. Nucleic Acids Res 2004;32: 551–561.
16. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
17. McGuffin LJ, Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 2003;19: 874–881.
18. Xu Y, Xu D. Protein threading using PROSPECT: design and evaluation. Protein: Struct Funct Genet 2000;40:343–354.
19. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
20. Sambrook JDR. Molecular cloning, a laboratory manual. New York: Cold Spring Harbor Laboratory; 2001. 999 pp.
21. Venters RA, Calderone TL, Spicer LD, Fierke CA. Uniform C-13 isotope labeling of proteins with sodium-acetate for nmr-studies—application to human carbonic anhydrase-Ii. Biochemistry 1991;30:4491–4494.
22. Studier FW. Protein production by auto-induction in high-density shaking cultures. Protein Expr Purif 2005;41:207–234.
23. Hansen MR, Hanson P, Pardi A. Filamentous bacteriophage for aligning RNA, DNA, and proteins for measurement of nuclear magnetic resonance dipolar coupling interactions. Methods Enzymol 2000;317:220–240.
24. Ruckert M, Otting G. Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. J Am Chem Soc 2000;122:7793–7797.
25. Weisemann R, Ruterjans H, Schwalbe H, Schleucher J, Bermel W, Griesinger C. Determination of H(N), H-α and H(N), C$'$ coupling-constants in C-13, N-15-labeled proteins. J Biomol NMR 1994;4:231–240.
26. Ottiger M, Delaglio F, Bax A. Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. J Magn Reson 1998;131:373–378.
27. Delaglio F, Grzesiek S, 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.
28. Wishart DS, Sykes BD. The C-13 chemical-shift index—a simple method for the identification of protein secondary structure using c-13 chemical-shift data. J Biomol NMR 1994;4:171–180.
29. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP—a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
30. Sali A, Blundell TL. Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
31. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. J Magn Reson 2003;160:65–73.
32. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 1999;13:289–302.

33. Linge JP, Williams MA, Spronk C, Bonvin A, Nilges M. Refinement of protein structures in explicit solvent. Proteins: Struct Funct Genet 2003;50:496–506.
34. Valafar H, Prestegard JH. REDCAT: a residual dipolar coupling analysis tool. J Magn Reson 2004;167:228–241.
35. Bax A. Weak alignment offers new NMR opportunities to study protein structure and dynamics. Protein Sci 2003;12:1–16.
36. Prestegard JH, Mayer KL, Valafar H, Benison GC. Determination of backbone structures from residual dipolar couplings. Methods Enzymol 2005;394:175–209.
37. Doreleijers JF, Mading S, Maziuk D, Sojourner K, Yin L, Zhu J, Markley JL, Ulrich EL. BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. J Biomol NMR 2003;26:139–146.
38. Seidel RD, Amor JC, Kahn RA, Prestegard JH. Conformational changes in human Arf1 on nucleotide exchange and deletion of membrane-binding elements. J Biol Chem 2004;279:48307–48318.
39. Chou JJ, Li SP, Bax A. Study of conformational rearrangement and refinement of structural homology models by the use of heteronuclear dipolar couplings. J Biomol NMR 2000;18:217–227.
40. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. Progress on modeling of protein structures and interactions. Science 2005;310:638–642.
41. Goldsmith-Fischman S, Honig B. Structural genomics: computational methods for structure analysis. Protein Sci 2003;12:1813–1821.
42. Canutescu AA, Shelenkov AA, Dunbrack RL. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003; 12:2001–2014.
43. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Pointing CP, Quevillon E, Selengut J, Sigrist CJA, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. InterPro, progress and status in 2005. Nucleic Acids Res 2005;33:D201–D205.
44. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 2003;31:400–402.
45. Holm L, Sander C. Dali—a network tool for protein-structure comparison. Trends Biochem Sci 1995;20:478–480.
46. Banci L, Bertini L, Ciofi-Baffoni S, Finney LA, Outten CE, O'Halloran TV. A new zinc-protein coordination site in intracellular metal trafficking: solution structure of the Apo and Zn(II) forms of ZntA(46–118). J Mol Biol 2002;323:883–897.
47. Lamb AL, Wernimont AK, Pufahl RA, Culotta VC, O'Halloran TV, Rosenzweig AC. Crystal structure of the copper chaperone for superoxide dismutase. Nat Struct Biol 1999;6:724–729.
48. Gitschier J, Moffat B, Reilly D, Wood WI, Fairbrother WJ. Solution structure of the fourth metal-binding domain from the Menkes copper-transporting ATpase. Nat Struct Biol 1998;5:47–54.
49. Weinberg MV, Schut GJ, Brehm S, Datta S, Adams MWW. Cold shock of a hyperthermophilic archaeon: *Pyrococcus furiosus* exhibits multiple responses to a suboptimal growth temperature with a key role for membrane-bound glycoproteins. J Bacteriol 2005;187:336–348.
50. Shah M, Passovets S, Kim DS, Ellrott K, Wang L, Vokler I, LoCascio P, Xu D, Xu Y. A computational pipeline for protein structure prediction and analysis at genome scale. Bioinformatics 2003;19:1985–1996.
51. Poole FL, Gerwe BA, Hopkins RC, Schut GJ, Weinberg MV, Jenney FE, Adams MWW. Defining genes in the genome of the hyerthermophilic archaeon *Pyrococcus furiosus*: implications for all microbial genomes. J Bacteriol 2005;187:7325–7332.