NMR structure determination of the protein NP_344798.1 as the first representative of Pfam PF06042

Biswaranjan Mohanty,
Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA, and Joint Center for Structural Genomics, (http://www.jcsg.org), La Jolla, CA 92037, USA. Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

Pedro Serrano,
Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA, and Joint Center for Structural Genomics, (http://www.jcsg.org), La Jolla, CA 92037, USA

Michael Geralt, and
Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA, and Joint Center for Structural Genomics, (http://www.jcsg.org), La Jolla, CA 92037, USA

Kurt Wüthrich[*]
Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA, and Joint Center for Structural Genomics, (http://www.jcsg.org), La Jolla, CA 92037, USA. Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

Abstract

We present the NMR structure determination of the protein NP_344798.1, which forms a CCA-adding enzyme head-domain architecture and is the first structural representative of the Pfam protein family PF06042. Its structure can now serve as a template for homology modeling of the other 785 members of this protein family. With 191 residues, NP_344798.1 is the largest single-domain protein structure determined so far with the J-UNIO protocol for automated NMR structure determination. The present work thus also shows that J-UNIO based exclusively on automated projection spectroscopy (APSY) and 3D heteronuclear-resolved [^{1}H,^{1}H]-NOESY experiments, can successfully be used to obtain high-quality NMR structures of protein domains with up to 200 residues.

Keywords

APSY; CCA-adding enzymes; J-UNIO; Nucleotidyltransferase superfamily; Solution NMR; Protein structure

[*]wuthrich@scripps.edu.
Present Address: B. Mohanty, Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences, Monash University, 381 Royal Parade, Parkville, Victoria 3052, Australia

Biological context

The protein NP_344798.1 from *Streptococcus pneumoniae* (TIGR4) belongs to the Pfam protein family PF06042, which currently contains 786 sequences from 739 species (http://pfam.xfam.org/family/PF06042). Based on sequence analyses, NP_344798.1 has been identified as a member of the nucleotidyltransfererase (NTase) -fold superfamily and annotated as a protein domain of unknown function, DUF925 (Kuchta et al. 2009). The bioinformatics data further suggest that all PF06042 members contain a single domain and are active NTases, as they contain the characteristic functional catalytic residues (Kuchta et al. 2009). Nonetheless, due to scarcity of experimental data, no specific biological role of these enzymes could as yet be ascertained. The NMR core of the Joint Center for Structural Genomics (JCSG) targeted NP_344798.1 to obtain a first structure for the PF06042 family, and the J-UNIO protocol for automated NMR structure determination (Serrano et al. 2012) was applied to this 191-residue protein. The structure now presents a foundation for obtaining new insights into the function of proteins in this family by structure comparison with other related structures in the PDB, experimental studies of substrate binding and specificity, and homology modeling of other proteins in PF06042.

Methods

Protein preparation

The protein NP_344798.1 was expressed in *E. coli*, using the pSpeedET-NP_344798.1 plasmid produced by the JCSG crystallomics core, and was purified following our standard protocol (Serrano et al. 2012). Micro-scale exploratory experiments were pursued with the [u-^{15}N]-protein (Pedrini et al. 2013; Serrano et al. 2012). For the NMR structure determination, a 1.2 mM solution of the [u-^{13}C,^{15}N]-labeled protein was prepared, with 20 mM sodium phosphate at pH 6.0, 50 mM sodium chloride and 4.5 mM NaN_3 in 5% (v/v) D_2O/95% H_2O. To remove oxygen prior to data collection, the sample was treated with Argon gas in the NMR tube.

NMR spectroscopy

2D [^{15}N,^{1}H]-HSQC spectra and APSY-NMR datasets (Hiller et al. 2005; 2008) were recorded on a Bruker Avance 600 MHz spectrometer equipped with a CPTCI HCN z-gradient cryoprobe. For the 5D APSY-HACACONH and 5D APSY-CBCACONH experiments, 24 2D projections were acquired, and a 4D APSY-HACANH experiment was acquired with 31 projections. 32 transients were accumulated for each APSY projection. The total acquisition time for the three experiments was 84 h. The digital resolution of the 2D projections was 2048 × 128 complex points for 4D APSY-HACANH, 2048 × 96 points for 5D APSY-HACACONH, and 2048 × 90 points for 5D APSY-CBCACONH. Prior to Fourier transformation, the time domains data were multiplied in both dimensions with a 45°-shifted sine bell (DeMarco and Wüthrich 1976) and zero-filled to 256 complex points in the indirect dimension.

3D [^{1}H,^{1}H]-NOESY-^{15}N-HSQC, 3D [^{1}H,^{1}H]-NOESY-^{13}C$_{ali}$-HSQC and 3D [^{1}H,^{1}H]-NOESY-^{13}C$_{aro}$-HSQC spectra were recorded on a Bruker Avance 800 MHz spectrometer

equipped with a TXI HCN z-gradient room temperature probe. The mixing time was 60 ms and the relaxation delay was 1 s. The 3D [^{1}H,^{1}H]-NOESY-^{15}N-HSQC spectrum was acquired with $2048 \times 90 \times 320$ complex points, a spectral width of 30 ppm and with the carrier at 118 ppm. For the 3D [^{1}H,^{1}H]-NOESY-^{13}C$_{ali}$-HSQC spectrum, $2048 \times 100 \times 330$ complex points were acquired, with a spectral width of 32 ppm and the carrier at 33 ppm. The 3D [^{1}H,^{1}H]-NOESY-^{13}C$_{aro}$-HSQC spectrum was acquired with $2048 \times 80 \times 330$ complex points, a spectral width of 31 ppm and the carrier at 122 ppm. The NOESY data sets were zero-filled to $2048 \times 256 \times 512$ complex points, multiplied with a squared cosine window in both proton dimensions and with a 45°-shifted sine bell (DeMarco and Wüthrich 1976) in the ^{15}N or ^{13}C dimension, and processed using Topspin 2.1. The three NOESY experiments were acquired in 9 days. Chemical shifts were referenced to DSS.

NMR structure determination with J-UNIO

Automated chemical shift assignment with UNIO-MATCH 2.0.1 (Volk et al. 2008) and UNIO-ATNOS/ASCAN 2.0.1 (Fiorito et al. 2008) was interactively validated and extended based on the NOESY data, using CARA (Keller 2004). Structure calculation and validation was performed following the J-UNIO protocol (Serrano et al. 2012), using the software UNIO-ATNOS-CANDID (Herrmann et al. 2002a, b) and CYANA (Güntert et al. 1997). The 40 conformers with the smallest residual target function values from cycle 7 of the structure calculation were energy-minimized with OPALp (Luginbühl et al. 1996; Koradi et al. 2000). 20 energy-minimized conformers were selected on the basis of the J-UNIO validation criteria (Serrano et al. 2012) to represent the protein structure, which was analyzed with MOLMOL (Koradi et al. 1996).

Results

In exploratory micro-scale experiments, NP_344798.1 was assessed for protein structure determination from its NMR-Profile (Pedrini et al. 2013). 177 out of 181 backbone amide cross peaks expected from the amino acid sequence were observed, which included a few peaks with low intensity. Nonetheless, most of the peaks were well-resolved and exhibited quite uniform intensities, indicating that the protein was a promising target for structure determination by solution NMR, in spite of its rather large size. This was further supported by the fact that 167 of the 177 observed ^{15}N–^{1}H cross peaks had intensities above the T$_{HSQC}$ threshold (Pedrini et al. 2013), so that good quality APSY-NMR data sets could be expected. For confirmation, a new NMR-Profile was recorded with the protein solution used for the structure determination, which confirmed the conclusions from the microscale experiments (Fig. 1, a and b).

NMR structure determination

Initial automated polypeptide backbone assignment was based on the three experiments 4D APSY-HACANH, 5D APSY-HACACONH and 5D APSY-CBCACONH (Hiller et al. 2008; Serrano et al. 2012). Fig. 1c shows that about 80% of the ^{1}H^N, ^{15}N, ^{13}C$^\alpha$, ^{1}H$^\alpha$ and ^{13}C$^\beta$ signals were correctly assigned by the automated backbone assignment routine UNIO-MATCH (Volk et al. 2008). Missing assignments and three erroneous assignments were crowded in the polypeptide segments D20–M23, N43–N49, N97–H107, P156–H159 and

R173–Q180. With regard to the structure characterization, it is of interest that the signal intensities of part of the backbone amide cross peaks were very weak or broadened beyond detection in the segments 97–107 and 173–180 (see the Discussion). The automated UNIO-MATCH backbone assignments were interactively validated against the 3D heteronuclear-resolved [^{1}H,^{1}H]-NOESY data sets which were recorded for the subsequent collection of conformational constraints (Serrano et al. 2012). In addition to the identification and correction of the three erroneous assignments, this resulted in an extension of the backbone assignments to about 95%. The remaining missing assignments include that only the chemical shifts of $^{13}C^\alpha$, $^1H^\alpha$ and $^{13}C^\beta$ were obtained for residues M2 and H59, and that H102, P156, P158, R173 and L174 remained unassigned (Fig. 1c).

The near-complete polypeptide backbone assignments (Fig 1c) provided the foundation for the structure determination with the J-UNIO protocol. The automated UNIO-ANTOS/ASCAN routine, which uses as additional input the aforementioned three 3D heteronuclear-resolved [^{1}H,^{1}H]-NOESY spectra (Fiorito et al. 2008), yielded about 75% of the expected assignments. Interactive validation of this result lead to assignments in the extent of 87%. The missing assignments, in addition to the unassigned backbone chemical shifts (Fig. 1c), include exclusively peripheral side chain atoms of Met, Arg, Lys, His and Trp. Subsequent identification of ^{1}H–^{1}H NOEs with the program UNIO-ANTOS/CANDID (Herrmann et al. 2002a, 2002b) in combination with the torsion angle dynamics program CYANA for the structure calculations (Güntert et al. 1997) yielded a final input of 4,090 distance constraints, including 1138 long-range constraints. With a backbone RMSD of 0.66 Å and an all-heavy-atom RMSD of 1.05 Å (see Table 1 for the complete statistics of the structure calculation), high precision of the structure determination was obtained, which compares favorably with the results of interactive structure determinations.

NMR structure of the protein NP_344798.1

Two different presentations of the structure are shown in Fig. 2, i.e., a bundle of 20 NMR conformers and a ribbon diagram of the conformer closest to the mean coordinates of the bundle. NP_344798.1 exhibits an α/β-topology with seven β-strands, seven α-helices, and two 3_{10}-helices. The arrangement of the regular secondary structures along the amino acid sequence is shown in Fig. 1c. Strands β1–β5 form a strongly twisted antiparallel β-sheet, and the two short strands β6 and β7 form a parallel β-sheet near the C-terminus of the protein. The larger β-sheet is well shielded from the solvent by the spatial arrangement of the helices α1 to α5, whereas the two-stranded β-sheet near the C-terminus is partially solvent-accessible in spite of its association with the helices α6 to α9. A homology search using DALI (Holm and Rosenstrom 2010) revealed that the fold is similar to the architecture of the catalytic head domain of class II CCA-adding enzymes (DALI Z-score > 9).

Discussion

The structure determination of NP_344798.1 expands the structural coverage of the genomic protein universe to a Pfam family with presently 786 members. Close similarities of the molecular architecture with the catalytic head domain of class II CCA-adding enzymes provides a lead for functional studies, which has been followed up by NMR studies of

substrate binding (to be published). Based on comparison with CCA-adding enzymes, there are indications that the two less precisely defined polypeptide segments of residues 97–107 and 173–180 are at or near the substrate-binding site. The indication of decreased structural order in these two segments motivated a detailed study of the structural dynamics in this molecular area and its possible functional role (to be published).

A recent determination of the structure of a 200-residue β-barrel protein with an integrative approach, "resolution-adapted structural recombination (RASREC) Rosetta", was considered to be a major technical advance (Lloyd and Wuttke 2014). This structure determination was based on the preparation of several differently isotope-labeled preparations of the protein and a large number of different NMR measurements (Sgourakis et al. 2014). In this context, it is remarkable that the present high-quality structure of a 191-residue α/β-protein was determined with the J-UNIO protocol, which uses a single, uniformly ^{13}C,^{15}N-labeled protein preparation and a total number of 7 NMR experiments, which were recorded with less than two weeks of instrument time.

Acknowledgments

References

DeMarco A, Wüthrich K. Digital filtering with a sinusoidal window function: An alternative technique for resolution enhancement in FT NMR. J Magn Reson. 1976; 24:201–204.

Fiorito F, Herrmann T, Damberger FF, Wüthrich K. Automated amino acid side-chain NMR assignment of proteins using ^{13}C- and ^{15}N-resolved 3D [^{1}H,^{1}H]-NOESY. J Biomol NMR. 2008; 42:23–33. [PubMed: 18709333]

Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol. 1997; 273:283–298. [PubMed: 9367762]

Herrmann T, Güntert P, Wüthrich K. Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. J Biomol NMR. 2002a; 24:171–189. [PubMed: 12522306]

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 DYANA. J Mol Biol. 2002b; 319:209–227. [PubMed: 12051947]

Hiller S, Fiorito F, Wüthrich K, Wider G. Automated projection spectroscopy (APSY). Proc Natl Acad Sci USA. 2005; 102:10876–10881. [PubMed: 16043707]

Hiller S, Wider G, Wüthrich K. APSY-NMR with proteins: practical aspects and backbone assignment. J Biomol NMR. 2008; 42:179–195. [PubMed: 18841481]

Holm L, Rosenstrom P. Dali server: conservation mapping in 3D. Nucl Acids Res. 2010; 38:W545–W549. [PubMed: 20457744]

Keller, R. CARA: computer aided resonance assignment. 2004. http://cara.nmr.ch/

Koradi R, Billeter M, Wüthrich K. MOLMOL: A program for display and analysis of macromolecular structures. J Mol Graph. 1996; 14:51–55. [PubMed: 8744573]

Koradi R, Billeter M, Güntert P. Point-centered domain decomposition for parallel molecular dynamics simulation. Comp Phys Commun. 2000; 124:139–147.

Kuchta K, Knizewski L, Wyrwicz LS, Rychlewski L, Ginalski K. Comprehensive classification of nucleotidyltransferase fold proteins: identification of novel families and their representatives in human. Nucl Acids Res. 2009; 37:7701–7714. [PubMed: 19833706]

Laskowski RA, Macarthur MW, Moss DS, Thornton JM. PROCHECK - a program to check the stereochemical quality of protein structures. J Appl Cryst. 1993; 26:283–291.

Lloyd NR, Wuttke DS. Less is more: Structures of difficult targets with minimal constrains. Structure. 2014; 22:1223–1224. [PubMed: 25185825]

Luginbühl P, Güntert P, Billeter M, Wüthrich K. The new program OPAL for molecular dynamics simulations and energy refinements of biological macromolecules. J Biomol NMR. 1996; 8:136–146. [PubMed: 8914272]

Pedrini B, Serrano P, Mohanty B, Geralt M, Wüthrich K. NMR-Profiles of protein solutions. Biopolymers. 2013; 99:825–831. [PubMed: 23839514]

Serrano P, Pedrini B, Mohanty B, Geralt M, Herrmann T, Wüthrich K. The J-UNIO protocol for automated protein structure determination by NMR in solution. J Biomol NMR. 2012; 53:341–354. [PubMed: 22752932]

Sgourakis NG, Natajaran K, Ying J, Vögeli B, Boyd LF, Margulies DH, Bax A. The structure of mouse cytomegalovirus m04 protein obtained from sparse NMR data reveals a conserved fold of the m02–m06 viral immune modulator family. Structure. 2014; 22:1263–1273. [PubMed: 25126960]

Toh Y, Takeshita D, Numata T, Fukai S, Nureki O, Tomita K. Mechanism for the definition of elongation and termination by the class II CCA-adding enzyme. Embo J. 2009; 28:3353–3365. [PubMed: 19745807]

Volk J, Herrmann T, Wüthrich K. Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. J Biomol NMR. 2008; 41:127–138. [PubMed: 18512031]
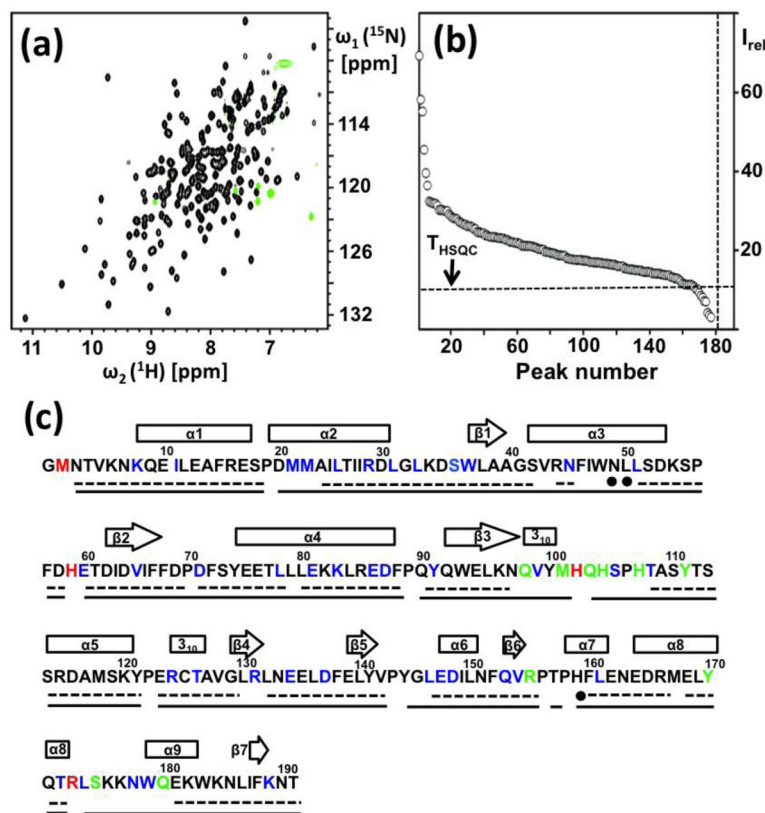
Fig. 1.
Characterization of the NP_344798.1 solution used for the NMR structure determination and survey of the polypeptide backbone assignments. (**a**) 2D [^{15}N,^{1}H]-HSQC spectrum at 600 MHz collected with a 1.2 mM solution of [u-^{13}C,^{15}N]-labeled NP_344798.1, pH = 6.0, T = 298K. (**b**) NMR-Profile. The number of backbone amide cross peaks expected from the amino acid sequence and the intensity threshold, T_{HSQC} (Pedrini et al. 2013), are indicated by dotted vertical and horizontal lines, respectively. (**c**) Amino acid sequence, with the following color-coded information: Residues for which the backbone amide cross peaks in the [^{15}N,^{1}H]-HSQC spectra are missing (red), partially overlapped (blue), or have intensity below T_{HSQC} (green). The dotted underline indicates correct assignments obtained automatically with UNIO-MATCH. Filled circles denote three residues for which UNIO-MATCH yielded erroneous assignments, which were indentified during the interactive validation and replaced with the correct chemical shifts prior to the next steps of the structure determination (see text). The solid horizontal line represents the final assignments obtained after interactive validation and extension of the UNIO-MATCH results using the NOESY data (see text). Above the sequence, the locations of regular secondary structures are indicated by rectangles for the helices and arrows for the β-strands.
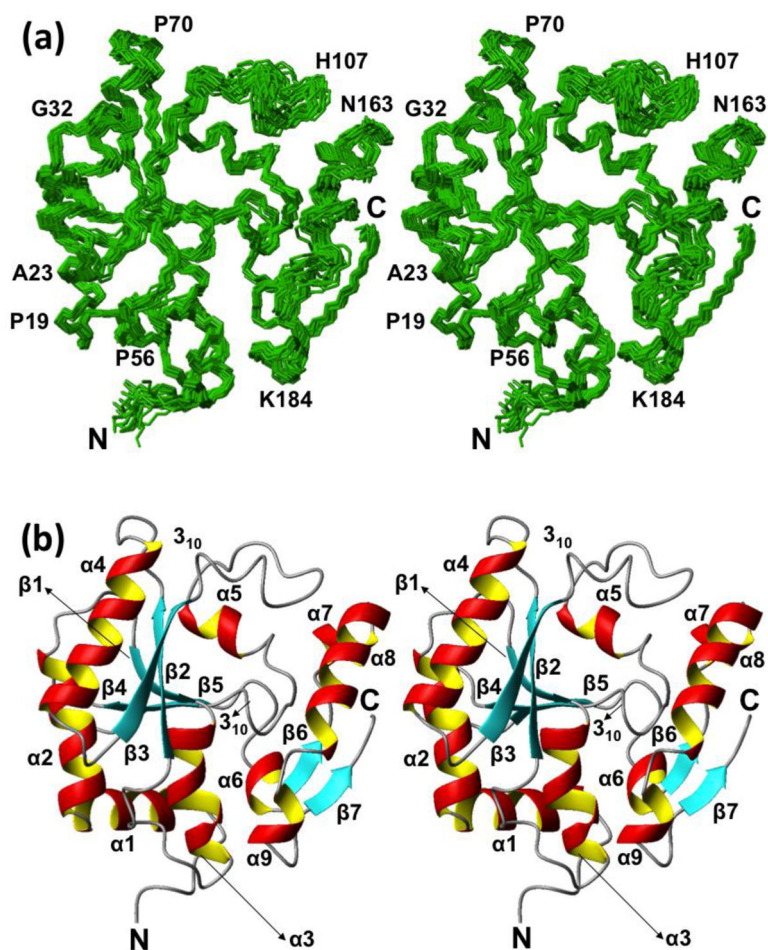
Fig. 2.
NMR structure of NP_344798.1. (**a**) Stereo view of a bundle of 20 NMR conformers representing the NMR structure. The chain ends and some sequence positions are identified to guide the eye. (**b**) Stereo ribbon representation of the conformer closest to the mean coordinates of the bundle shown in (**a**). β-strands are cyan, helices are red/yellow, and segments with non-regular secondary structure are grey. The chain ends and the regular secondary structures are identified. The figure was prepared using the program MOLMOL (Koradi et al. 1996).

Table 1

Input and calculation statistics for the NMR structure determination of the protein NP_344798.1 in aqueous solution at pH = 6.0 and T = 298K.

Quantity	Value[a]
NOE upper distance limits	4090
intraresidual	993
short-range	1055
medium-range	904
long-range	1138
Dihedral angle constraints	799
Residual target function value [Å^2]	3.9 ± 0.43
Residual NOE violations	
number 0.1 A	41 ± 6
maximum [Å]	0.2 ± 0.19
Residual dihedral angle violations	
number 2.5o	1 ± 1
maximum [°]	4.79 ± 1.5
AMBER energies [kcal/mol]	
total	−7843 ± 144
van der Waals	−705 ± 35
electrostatic	−8779 ± 112
RMSD from the mean coordinates [b][Å]	
backbone (2–191)	0.66 ± 0.09
all heavy atoms (2–191)	1.05 ± 0.09
Ramachandran plot statistics[c]	
most favoured regions [%]	75.1
additional allowed regions [%]	23.0
generously allowed regions [%]	1.6
disallowed regions [%]	0.3

[a]Except for the top six entries, which represent the input generated for the final cycle of structure calculation with UNIO-ATNOS/CANDID and CYANA 3.0, average values and standard deviations for the 20 energy-minimized conformers are given.

[b]The numbers in parentheses indicate the residues for which the RMSD was calculated.

[c]As determined by PROCHECK (Laskowski et al. 1993).