Journal of Biomolecular NMR **29:** 403–406, 2004.

Letter to the Editor: NMR structure determination of the hypothetical protein TM1290 from *Thermotoga maritima* using automated NOESY analysis

Touraj Etezady-Esfarjani[a,*], Torsten Herrmann[a], Wolfgang Peti[a], Heath E. Klock[b], Scott A. Lesley[b] & Kurt Wüthrich[a]
[a]*The Scripps Research Institute, Department of Molecular Biology and Joint Center for Structural Genomics, 10550 North Torrey Pines Road, La Jolla, CA 92037, U.S.A.;* [b]*Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, CA 92121, U.S.A.*

Received 3 March 2004; Accepted 18 March 2004

Key words: Automation with ATNOS and CANDID, NMR structure determination, structural proteomics, *Thermotoga maritima*

Biological context

The 115-residue hypothetical protein TM1290 is a member of the COG1433 protein family, and it has sequence similarity to NifX and the C-terminal domain of NifB (Shah et al., 1999). These proteins are involved in the biosynthesis of the iron-molybdenum cofactor (FeMo-co) of dinitrogenase, an enzyme which catalyzes the reduction of nitrogen gas to ammonium in an ATP- and reductant-dependent reaction (Rubio et al., 2002).

This paper documents the first *de novo* protein structure determination using the algorithms ATNOS (Herrmann et al., 2002a) for automated NOESY peak picking and NOE identification, and CANDID (Herrmann et al., 2002b) for automated NOE assignment, which enables direct feedback between the 3D protein structure and the raw NMR data during protein structure refinement.

Methods and results

TM1290 was selected as a pilot NMR project in the Joint Center for Structural Genomics (JCSG). Its expression and purification was reported previously (Etezady-Esfarjani et al., 2003).

For the NMR data acquisition, we used a 2.6 mM D_2O solution of unlabeled TM1290, and solutions in 95% H_2O/5% D_2O (v/v) of 2.3 mM uniformly ^{15}N-labeled and 3.8 mM uniformly $^{13}C/^{15}N$-labeled protein. The pH was 6.0.

Three NOESY spectra were recorded on a Bruker Avance900 spectrometer with a mixing time t_m = 70 ms at T = 313 K: 3D ^{15}N-resolved [^1H,^1H]-NOESY in H_2O (Zuiderweg and Fesik, 1989); 3D ^{13}C-resolved [^1H,^1H]-NOESY (Ikura et al., 1990) in H_2O with the ^{13}C carrier frequency in the aliphatic region; 2D [^1H,^1H]-NOESY in D_2O. These spectra were automatically analyzed with the software ATNOS/CANDID (Herrmann et al., 2002a,b) incorporated in the torsion angle dynamics program DYANA (Güntert et al., 1997). The input consisted of the chemical shifts obtained from the previous sequence-specific resonance assignment (Etezady-Esfarjani et al., 2003) and the three NOESY spectra. The standard protocol with seven cycles of ATNOS peak picking, CANDID NOE assignment and DYANA 3D structure calculation was applied (Herrmann et al., 2002a, b). Stereospecific assignments of 40 valine and leucine isopropyl methyl groups were determined experimentally by biosynthetic fractional ^{13}C-labeling. For each cycle of structure calculation, these 40 stereospecific assignments, and 142 constraints for the backbone dihedral angles derived from C^α chemical shifts (Spera and Bax, 1991) were added to the NOESY input. For the final structure calculation in cycle 7, only distance constraints were retained that could be unambiguously assigned based on the protein three-dimensional structure from cycle 6. A total of 2,444

*To whom correspondence should be addressed. E-mail: touraj@scripps.edu

Table 1. Input for the structure calculation and characterization of the energy-minimized NMR structures of TM1290

Quantity	Value[a]
NOE upper distance limits	2444
Dihedral angle constraints	142
Residual target function [$Å^2$]	1.91 ± 0.51
Residual NOE violations	
Number ≥ 0.1 Å	34 ± 4 (27–42)
Maximum [Å]	0.15 ± 0.01 (0.14–0.16)
Residual angle violations	
Number ≥ 2.5 deg	7 ± 1 (4–10)
Maximum [deg]	4.76 ± 0.72 (3.54–7.01)
Amber energies [kcal/mol]	
Total	−4354.95 ± 62.88
van der Waals	−279.04 ± 15.93
Electrostatic	−5008.90 ± 70.16
rmsd from ideal geometry	
Bond lengths [Å]	0.0079 ± 0.0002
Bond angles [deg]	2.08 ± 0.04
rmsd to the mean coordinates [Å][b]	
bb (4–43, 55–110)	0.31 ± 0.03 (0.26–0.37)
ha (4–43, 55–110)	0.75 ± 0.07 (0.65–0.91)
Ramachandran plot statistics	
Most favored region (%)	66.2
Additional allowed region (%)	29.9
Generously allowed region (%)	2.6
Disallowed region (%)	1.3

[a]Except for the top two entries, the average value for the 20 energy-minimized conformers with the lowest residual DYANA target function values and the standard deviation among them are listed, with the minimum and maximum values given in parentheses.
[b]bb indicates the backbone atoms N, C^{α}, C'; ha stands for 'all heavy atoms'. The numbers in parentheses indicate the residues for which the rmsd was calculated.

meaningful NOE upper distance constraints, extracted from a total of 5,309 assigned NOESY cross peaks, were used as input for the final structure calculation (Table 1, Figure 1). The 20 conformers with the lowest residual DYANA target function values obtained from cycle 7 were energy-refined in a water shell with the program OPALp (Luginbühl et al., 1996), using the AMBER force field (Cornell et al., 1995). The program MOLMOL (Koradi et al., 1996) was used to analyze the protein structure and to prepare the figures.

The solution structure of TM1290 contains a five-stranded β-sheet with residues 2–7, 25–30, 35–40, 64–66 and 84–86, three α-helices with residues 53–58, 75–79 and 93–101, and a 3_{10}-helical turn 20–22 (Figure 1). The five β-strands are combined together in the order β3, β2, β1, β4 and β5, where β2 is antiparallel to the other four β-strands, and the β-sheet has a pronounced right-handed twist. The β-sheet is sandwiched between the helices α1 and α2 on one side, and the helix α3 on the other side. A well-structured loop with residues 8–19 connects β1 and the 3_{10}-helical turn. A poorly defined loop with residues 41–52 is located between β3 and α1 (Figure 1). For the five residues 47–50 and 52, no correlation peaks in the 2D [^{1}H,^{15}N]-HSQC spectrum could be identified (data not shown), indicating either rapid exchange of the corresponding amide protons with the solvent or slow conformational exchange.

Discussion and conclusions

The TM1290 α/β motif is present in seven protein superfamilies, where the members of the ribonuclease HI family have the closest fold similarity to TM1290.

Using an upper distance limit of 4.2 Å for the distances between carboxyl oxygen atoms of Asp or Glu and side-chain nitrogen atoms of Arg, Lys or His, each of the salt bridges Arg 2–Asp 30, Lys 84–Glu 107 and Arg 87–Glu 89 (Figure 2) is present in at least 10 out of the 20 energy-minimized conformers of TM1290 (Figure 1). With an upper distance limit of 4.5 Å, an additional salt bridge between Glu 37 and Lys 60 was identified (Figure 2). The spatial arrangement of these ion pair interactions appears to be an important factor for the stabilization of a large surface cleft between the central β-sheet and the helix α3 (Figure 2). On one side of the cleft, the orientation of the flanking helix α3 is constrained through the Lys 84–Glu 107 and Arg 87–Glu 89 ion pairs. On the other side, the arrangement of the strands β1–β4 is stabilized by the salt bridges Arg 2–Asp 30 and Glu 37–Lys 60.

MTH1175 (Cort et al., 2001) and TM1290 are two proteins in the COG1433 family of 15 proteins for which NMR structures have been determined. These two proteins have 45% sequence identity, and correspondingly very similar folds. In both proteins the central segment connecting the β strand to helix α1 (residues 45–53 in MTH1175 and 45–51 in TM1290) is structurally only poorly defined (Figure 1). Multiple sequence alignments of COG1433 proteins (data not shown) did not reveal any conserved residues in this segment, but showed a conserved XXEN (X =

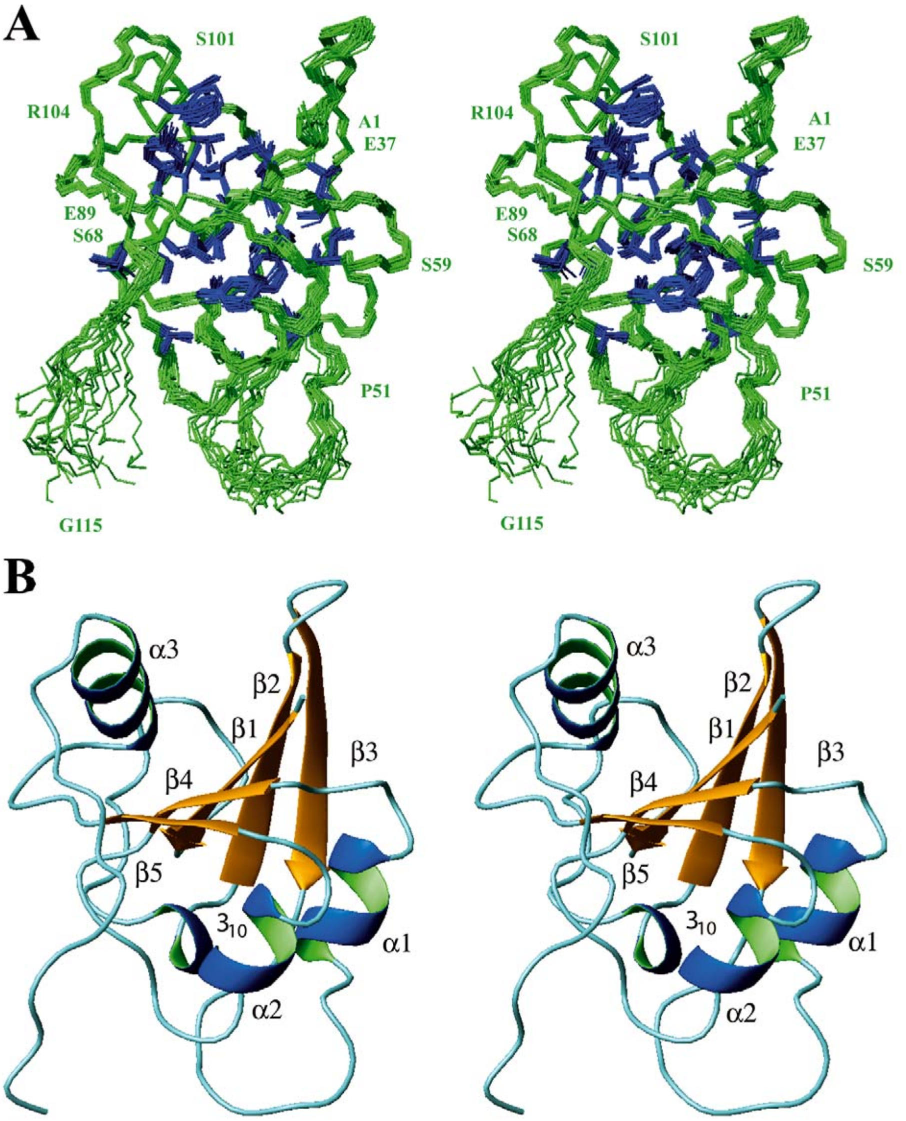

Figure 1. (A) Stereoview of the bundle of 20 energy-minimized DYANA conformers representing the NMR structure of the hypothetical protein TM1290. The superposition is for best fit of the backbone atoms N, C^{α} and C' of the residues 4–43 and 55–110. The entire polypeptide backbone with residues 1–115 is shown in green, and to ease access to the structure some sequence positions are indicated with green letters. The side chains with local displacement values up to 1.0 Å and solvent accessibility $< 30\%$, which form the core of the protein, are shown in blue. (B) Stereoview of a ribbon representation of the closest conformer to the mean coordinates of the bundle in (A). Same orientation as in (A). The regular secondary structures α1 to α3, 3_{10}, and β1 to β5 are identified.

Val, Ile) motif preceding this disordered region. These observations indicate that the disordered central loop might be an interaction site between TM1290 and a physiological binding partner, which possibly will become structured in the complex with this hypothetical partner molecule. However, a functional assignment for TM1290 remains open at this point, and no evidence could be found for a nucleic acid binding site in TM1290 that would correspond to the basic protrusion in Rnase H (Katayanagi et al., 1992).

In view of the demand for high-throughput NMR studies of proteins in current structural proteomics initiatives, any further automation of the NMR structure determination process will be attractive (Güntert, 2003). The presently used ATNOS/CANDID approach for automated NOESY spectral analysis follows the prime objectives of increasing the efficiency

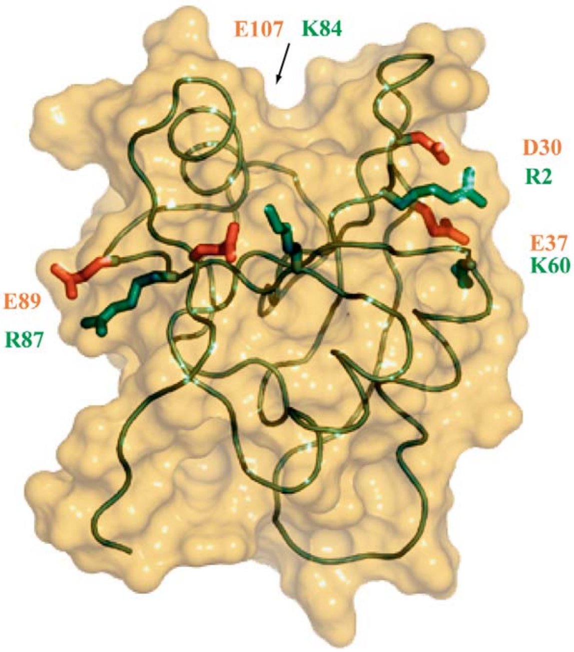

Figure 2. Transparent surface presentation of TM1290 superimposed with a spline function through the C^{α} positions. A cleft bounded by the helix $\alpha3$ and the β-sheet is visible at the top (see text). The side chains of four pairs of salt bridge-forming amino acids are shown with stick diagrams and identified with the one-letter amino acid code and the sequence numbers. Same orientation of TM1290 as in Figure 1, with the disordered loop 41–52 in the lower right.

as well as the reliability of protein structure determination by NMR. This paper confirms that this approach is capable of meeting the aforementioned goals. Furthermore, during the application of ATNOS/CANDID for the *de novo* structure determination of TM1290, previous conclusions on the proper performance of the automated approach derived from model calculations (Herrmann et al., 2002a,b) could be confirmed. In this *de novo* structure determination of TM1290, ATNOS/CANDID dramatically enhanced the efficiency of the NOESY spectral analysis when compared to conventional, interactive procedures.

Atomic coordinates for a bundle of 20 conformers representing the NMR solution structure of TM1290 have been deposited in the Protein Data Bank (http://www.rcsb.org; PDB ID code 1RDU).

Acknowledgements

K.W. is the Cecil H. and Ida M. Green Visiting Professor of Structural Biology at TSRI. Further financial support was obtained from the Schweizerischer Nationalfonds (project 31-66427-01). The JCSG consortium is funded by NIGMS GM0622411. W.P. is recipient of an E. Schrödinger Fellowship (J2145). We acknowledge the use of the high-performance computing facilities of the ETH Zürich and The Scripps Research Institute.

References

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T. et al. (1995) *J. Am. Chem. Soc.*, **117**, 5179–5197.

Cort, J.R., Yee, A., Edwards, A.M., Arrowsmith, C.H. and Kennedy, M.A. (2001) *J. Struct. Func. Genomics*, **1**, 15–25.

Etezady-Esfarjani, T., Peti, W. and Wüthrich, K. (2003) *J. Biomol. NMR*, **25**, 167–168.

Güntert, P. (2003) *Prog. NMR. Spectrosc.*, **43**, 105–125.

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

Herrmann, T., Güntert, P. and Wüthrich, K. (2002a) *J. Biomol. NMR*, **24**, 171–189.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002b) *J. Mol. Biol.*, **319**, 209–227.

Holm, L. and Sander, C. (1997) *Nucl. Acids Res.*, **25**, 231–234.

Ikura, M., Bax, A., Clore, G.M. and Gronenborn, A.M. (1990) *J. Am. Chem. Soc.*, **112**, 9020–9022.

Katayanagi, K., Miyagawa, M., Matsushima, M., Ishikawa, M., Kanaya, S., Nakamura, H., Ikehara, M., Matsuzaki, T. and Morikawa, K. (1992) *J. Mol. Biol.*, **223**, 1029–1052.

Koradi, R., Billeter, M. and Wüthrich, K. (1996) *J. Mol. Graph.*, **14**, 51–55.

Luginbühl, P., Güntert, P., Billeter, M. and Wüthrich, K. (1996) *J. Biomol. NMR*, **8**, 136–146.

Rubio, L.M., Rangaraj, P., Homer, M.J., Roberts, G.P. and Ludden, P.W. (2002) *J. Biol. Chem.*, **277**, 14299–14305.

Shah, V.K., Rangaraj, P., Chatterjee, R., Allen, R.M., Roll, J.T., Roberts, G.P. and Ludden, P.W. (1999) *J. Bacteriol.*, **181**, 2797–2801.

Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.

Zuiderweg, E.R.P. and Fesik, S.W. (1989) *Biochemistry*, **28**, 2387–2391.