Automated protein NMR structure determination in crude cell-extract

Journal of Biomolecular NMR (2006) 34: 3–11
DOI 10.1007/s10858-005-4519-5

Article

Automated protein NMR structure determination in crude cell-extract

Touraj Etezady-Esfarjani[†,*], Torsten Herrmann[†], Reto Horst & Kurt Wüthrich
Department of Molecular Biology and Joint Center for Structural Genomics, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California, 92037, USA

Received 8 August 2005; Accepted 21 October 2005

Key words: automated protein NMR structure determination with ATNOS and CANDID, network anchoring, structural proteomics, *Thermotoga maritima*

Abstract

A fully automated, NOE-based NMR structure determination of a uniformly ^{13}C,^{15}N-labeled protein was achieved in crude cell-extract, without purification of the overexpressed protein. Essentially complete sequence-specific assignments were obtained using triple resonance experiments, based on the high intensity of the resonances from the overexpressed protein relative to those of the background. For the collection of NOE distance constraints, efficient discrimination between NOE cross peaks from the target protein and background signals was achieved using the programs ATNOS and CANDID. In the iterative ATNOS/CANDID procedure, the identification of the desired protein NOEs is initially guided by the self-consistency of the protein NOE-network. Although the intensities of the signals in this network vary over a wide range, and are in many instances comparable to or smaller than those of the background, the first cycle of calculations resulted in the correct global polypeptide fold, and the structure was then refined in six subsequent cycles using the intermediate NMR structures for additional guidance. The experience gained with this work demonstrates that the ATNOS/CANDID procedure for automatic protein structure determination is highly robust and reliable in the presence of intense background signals, and might thus also represent a platform for future protein structure determinations in physiological fluids.

Introduction

In structural genomics centers, NMR-based screening using purified protein samples has proven to be a valuable and reliable approach for the selection of target proteins for structure determination by either NMR spectroscopy or X-ray crystallography (Yee et al., 2002; Peti et al., 2004; Page et al., 2005). Alternatively, NMR samples of crude cell-extracts, without the need of the cost- and time-consuming step of protein purification, have been used for probing the structural integrity of target proteins (Gronenborn and Clore, 1996; Galvao-Botton et al., 2003) and for the identification of amino acid types in proteins (Ou et al., 2001). The successful use of crude cell-extract samples for screening of globular proteins (Gronenborn and Clore, 1996; Hubbard et al., 2003; Oldfield et al., 2005; Ou et al., 2001; Serber and Dötsch, 2001) provided encouragement to explore this approach for *de novo* protein structure determination with the use of the software ATNOS/CANDID (Herrmann et al., 2002a, b) for automatic NOE spectral analysis and structure calculation.

Successful *de novo* structure determination of overexpressed recombinant proteins in cell extract

*To whom correspondence should be addressed. E-mail: touraj@mol.biol.ethz.ch
[†]Present address: Institut für Molekularbiologie und Biophysik, ETH Zürich, CH- 8093 Zürich, Switzerland

requires extension of previous NMR studies with crude cell lysate (Gronenborn and Clore, 1996; Ou et al., 2001; Serber and Dötsch, 2001; Hubbard et al., 2003; Oldfield et al., 2005) in two directions. First, a set of standard 3D NMR experiments for resonance assignment and collection of conformational constraints needs to be recorded, rather than correlation experiments with outstandingly high intrinsic sensitivity. This requirement will foreseeable limit this approach to systems with high protein overexpression. Second, algorithms for efficient and reliable automated spectral analysis of the [^{1}H,^{1}H]-NOESY data sets are needed. Since both the target protein and the background mass in the cell extract are uniformly isotope-labeled to a comparable extent, interactive analysis of the [^{1}H,^{1}H]-NOEs would be impractical, and in most situations actually impossible, so that the present work is a crucial test for the recently introduced software ATNOS/CANDID (Herrmann et al., 2002a, b).

In this paper we use automated NMR structure determination in crude cell-extract with the recombinant hypothetical protein TM1290 from *Thermotoga maritima*. The molecular weight of 12.5 kDa, high expression yields in *E. coli*, and thermostability make TM1290 a good target for NMR structure determination. Since the NMR structure of TM1290 has previously been determined with purified protein samples, using the same ATNOS/CANDID standard protocol for automated NOESY spectral analysis as employed here (Etezady-Esfarjani et al., 2004), we also had a reliable reference for the validation of the results obtained with the crude cell-extract.

Materials and methods

Protein expression

Expression of the target protein, TM1290, was carried out using the plasmid pET-25b(+), encoding TM1290 as the expression vector and with the *E. coli* strain BL21-CodonPlus (DE3)-IRL (Stratagene) as host cells. Expression of uniformly ^{13}C/^{15}N-labeled TM1290 was carried out by growing the cells in M9 minimal medium containing $^{15}NH_4Cl$ (1 g l^{-1}) and [$^{13}C_6$]-D-glucose (4 g l^{-1}) as the sole nitrogen and carbon sources, respectively. Cell cultures were grown at 37 °C with vigorous shaking to an OD_{600} of 0.6 before protein expression was induced with IPTG (final concentration 1 mM). The cells were grown for an additional 4 h before harvesting. After expression of the target protein and harvesting of the cells, the cell pellet was resuspended in 20 mM sodium phosphate buffer (1 g of cells (wet mass) in 7 ml of phosphate buffer, pH 6.0). For the preparation of NMR samples, the cells were lysed by sonication, the cell debris removed by centrifugation (10,000 *g*, 20 min) and the supernatant was stored at 193 K in 500 µl aliquots. Prior to the NMR measurements, the samples were supplemented with 5% D_2O, and with protease inhibitors (Complete, Roche) to a final concentration of 25 µg/ml.

NMR experiments

The NMR measurements were carried out at 313 K and pH 6.0 on either an Avance 600 or Avance 900 spectrometer (Brucker, Billerica, MA) equipped with four radio-frequency channels and triple resonance probeheads with shielded xyz-gradient coils. Proton chemical shifts were referenced to internal 3-(trimethyl-silyl)-propane-1,2,2,3,3-d_6-sulfonic acid, sodium salt (DSS). The ^{13}C and ^{15}N chemical shifts were referenced indirectly to DSS using absolute frequency ratios. The NMR spectra were processed with the software PROSA (Güntert et al., 1992), and interactive NMR data analysis was performed using the XEASY software (Bartels et al., 1995).

The sequence-specific resonance assignment for the TM1290 backbone atoms was based on the following experiments (reviewed in Sattler et al., 1999): 2D [^{15}N,^{1}H]-HSQC (time domain data size 256×1024 complex points, $t_{1,max}(^{15}N) = 79.3$ ms, $t_{2,max}(^1H) = 155$ ms); 3D HNCA and 3D HNCACB (time domain data size 128×64×1024 complex points, $t_{1,max}(^{13}C) = 7.0$ ms, $t_{2,max}(^{15}N) = 19.8$ ms, $t_{3,max}(^1H) = 77.5$ ms); 3D CBCA(CO)NH (time domain data size 64 × 56 × 2048 complex points, $t_{1,max}(^{13}C) = 3.5$ ms, $t_{2,max}(^{15}N) = 17.4$ ms, $t_{3,max}(^1H) = 155$ ms). The following NMR experiments were measured for obtaining the chemical shift assignments for the amino acid side chains (Sattler et al., 1999): 3D ^{15}N-resolved [^{1}H,^{1}H]-TOCSY (mixing time = 60 ms, time domain data size 256 × 80 × 2048 complex points, $t_{1,max}(^1H) = 19.4$ ms, $t_{2,max}(^{15}N) = 24.8$ ms, $t_{3,max}(^1H) =$

155 ms); 3D H(C)CH-TOCSY (mixing time = 60 ms, time domain data size 256×128×1024 complex points, $t_{1,max}(^1H) = 19.4$ ms, $t_{2,max}(^{13}C) = 6$ ms, $t_{3,max}$ (^1H) = 74.5 ms); 3D (H)CCH-TOCSY (mixing time = 14 ms, time domain data size 256×128×1024 complex points, $t_{1,max}(^1H) = 19.4$ ms, $t_{2,max}(^{13}C) = 8$ ms, $t_{3,max}(^1H) = 74.5$ ms).

Distance constraints for the structure calculation were derived from two NOESY spectra recorded with mixing times of $\tau_m = 70$ ms, i.e., 3D ^{15}N-resolved [^{1}H,^{1}H]-NOESY in H_2O (Zuiderweg and Fesik, 1989) (time domain data size 256 × 76 × 2048 complex points, $t_{1,max}(^1H) = 12.9$ ms, $t_{2,max}(^{15}N) = 15.7$ ms, $t_{3,max}(^1H) = 103.3$ ms) and 3D ^{13}C-resolved [^{1}H, ^{1}H]-NOESY in H_2O (Muhandiram et al., 1993) (carrier frequency = 68 ppm, time domain data size 256×100×2048 complex points, $t_{1,max}(^1H) = 12.9$ ms, $t_{2,max}(^{13}C) = 4.9$ ms, $t_{3,max}$ (^1H) = 103.3 ms). These spectra were recorded with a 900 MHz spectrometer and were also used for the chemical shift assignment of the side chain $^{15}NH_n$ groups of Asn, Gln, and Arg, and for the chemical shift assignment of the aromatic ring protons (Wüthrich, 1986).

Structure calculation

For combined collection of conformational constraints and three-dimensional protein structure calculation, the [^{1}H,^{1}H]-NOESY spectra were automatically peak picked and assigned with the software ATNOS/CANDID (Herrmann et al., 2002a, b) incorporated into the torsion angle dynamics program DYANA (Güntert et al., 1997). The input consisted of the chemical shift lists from the previous sequence-specific resonance assignment and the two aforementioned 3D heteronuclear-resolved [^{1}H,^{1}H]-NOESY spectra. The standard protocol with seven cycles of peak picking with ATNOS, NOE assignment with CANDID and structure calculation with DYANA was used. During the first six ATNOS/CANDID cycles, ambiguous distance constraints (Nilges, 1997) were used. In the second and subsequent cycles, the intermediate three-dimensional protein structures served as an additional guide for the interpretation of the NOESY spectra. 144 conformational constraints for the backbone dihedral angles derived from C^α chemical shifts (Luginbühl et al., 1995; Spera and Bax, 1991) were added to the input for each cycle of structure calculation. For the final structure calculation in cycle 7, only distance constraints were retained that could be unambiguously assigned based on the three-dimensional protein structure from cycle 6. 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 (Koradi et al., 2000; Luginbühl et al., 1996), using the AMBER force field (Cornell et al., 1995).

Results

For a successful NMR structure determination in crude cell-extract, long-term stability of the target protein is a critical requirement, since the NMR experiments currently used to obtain sequence-specific assignments and to collect NOE distance restraints typically require measurement times of several days. Even small changes in the sample composition due to protease activities or other degradation mechanisms tend to deteriorate the quality of the NMR spectra. In the present work, the cell extract was stabilized by addition of a protease inhibitor cocktail before the start of the NMR experiments (see Materials and methods). As a result, for example, two 2D [^{15}N,^{1}H]-HSQC spectra recorded before and after recording of a 3D ^{13}C-resolved [^{1}H,^{1}H]-NOESY experiment during 80 h at 40 °C showed only small, strictly limited differences (Figure 1). These consist of three peaks that were only seen before the NOESY experiment (2, 7 and 8 in Figure 1) and five weak cross-peaks in the HSQC spectrum recorded after the NOESY experiment (1 and 3–6 in Figure 1), The sequential resonance assignment then revealed that all these 8 peaks are background signals (see below). The close overlap of the strong peaks shows that the overexpressed TM1290 retained its structural integrity in the cell-extract during 80 h of NMR recording at 40 °C.

Using standard heteronuclear multidimensional NMR experiments based on through-bond scalar connectivities (Sattler et al., 1999), sequential resonance assignments for the backbone atoms were established (Figure 2a). The $C^{\alpha/\beta}$ chemical shifts were then used as a starting point for the resonance assignments of the amino-acid side chains, based on a 3D ^{15}N-resolved [^{1}H,^{1}H]-

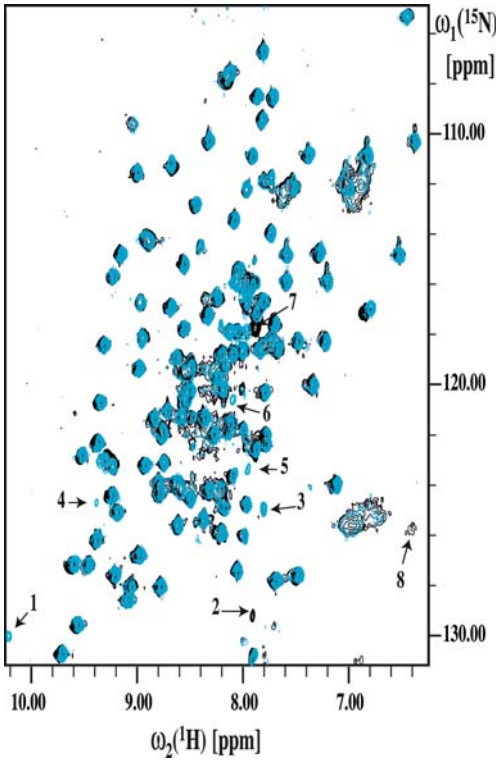

Figure 1. Long-time stability of crude cell-extract samples of TM1290. A 2D [^{15}N,^{1}H]-HSQC spectrum taken before a 3D ^{13}C-resolved [^{1}H,^{1}H]-NOESY experiment (black), which was recorded for 80 h at 40 °C, is superimposed on a 2D [^{15}N,^{1}H]-HSQC spectrum taken after this NOESY experiment (cyan). Eight resonance signals which are seen in only one of the two spectra, are indicated by arrows and arbitrary numbers (see text).

TOCSY spectrum (Figure 2b) to obtain proton chemical shifts of the amino-acid spin systems. Figure 2 clearly illustrates that the background signals in cell lysate interfere at most marginally with the interpretation of the NMR spectra used for resonance assignments in ^{13}C,^{15}N-labeled proteins. Complete resonance assignments for overexpressed, uniformly ^{13}C,^{15}N-labeled TM1290 in crude *E. coli* cell extract were thus obtained, except for all atoms of residues 48, 49, 50 and 73, the amide protons of residues 1, 46, 47, 51, 52 and 74, H^{α} and H^{β} of His 47, andδCH_2 of Arg 72 (Figure 3). This result coincides nearly identically with the corresponding data obtained with the purified protein (Etezady-Esfarjani et al., 2004). This coincidence of the extent of NMR assignments is readily rationalized by the fact that scalar coupling-based correlation experiments yield cross

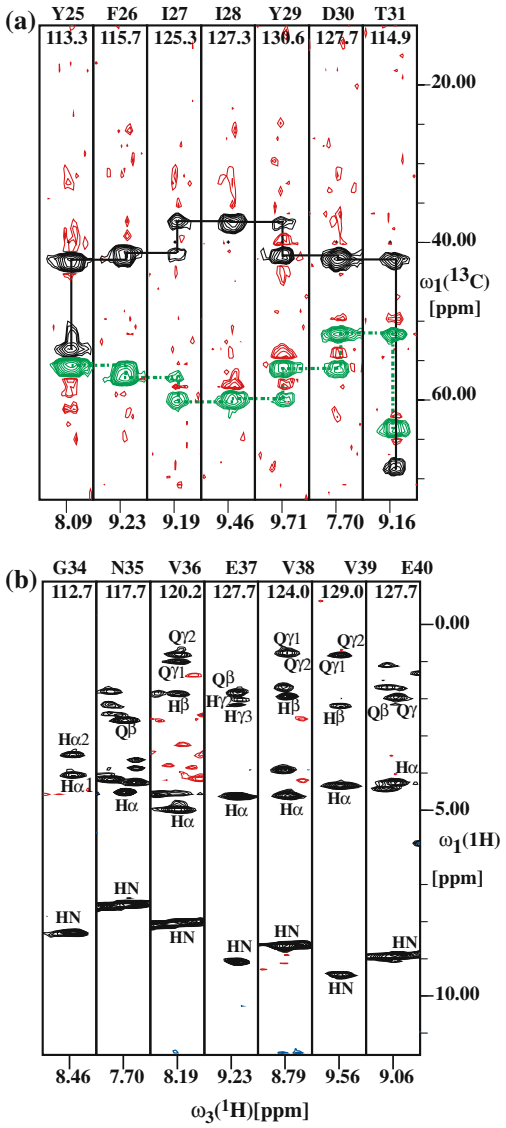

Figure 2. Resonance assignments of the backbone and side-chain atoms in TM1290. (a) [ω_1(^{13}C), ω_3(^{1}H)] strips for the polypeptide segment 25–31 from a 3D HNCACB spectrum used for obtaining sequential backbone assignments for ^{13}C/^{15}N-labeled TM1290 in cell-extract. The $C^{\beta}-H^{N}$ cross peaks are shown in black, and the $C^{\alpha}-H^{N}$ cross peaks in green. The sequential $C^{\beta}(i, i\text{-}1)$ and $C^{\alpha}(i, i\text{-}1)$ connectivities are presented by solid and dashed lines, respectively. (b) Chemical shift assignment of the side chain protons for the residues 34–40. Strips along the ω_1(^{1}H) dimension of a 3D ^{15}N-resolved [^{1}H,^{1}H]-TOCSY spectrum are shown. The side-chain ^{1}H chemical shift assignments are indicated by the atom names, where "Q" are pseudoatoms representing CH_n groups. In both spectra, the resonance assignments are indicated at the top by the one-letter amino acid code and the sequence number, the ^{15}N chemical shifts are given at the top of the strips, and background resonances from other components in the cell-extract are highlighted in red (see text).

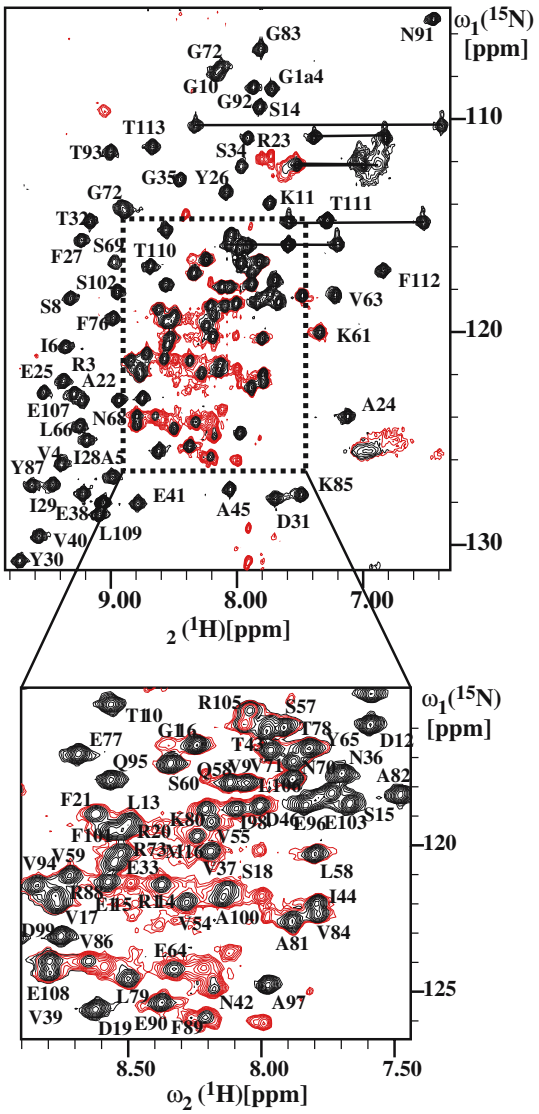

Figure 3. Mapping of the sequence-specific resonance assignments for recombinant TM1290 in a 2D [^{15}N, ^{1}H]-HSQC spectrum at 600 MHz, pH 6.0 and 313 K of uniformly ^{15}N-labeled crude cell-extract. Resonance assignments for the backbone ^{15}N–^{1}H moieties are indicated by the one-letter amino-acid code and the sequence number. The proton pairs of the Gln and Asn side chains amide groups are connected by horizontal lines. The background signals originating from other compounds in the cell-extract are highlighted in red.

peak intensities that relate to a good approximation linearly to the protein concentration. The overexpressed protein of interest is, therefore, represented in the experimental spectra by a pattern of intense peaks (Figures 1–3). Figure 3 further shows that strong background signals appear almost exclusively in the "random coil spectral region" (Wüthrich, 1986), with ^{1}H chemical shifts from 7.7. to 8.8 ppm.

The foremost challenge for *de novo* NMR structure determination in crude cell-extract is reliable discrimination between [^{1}H,^{1}H]-NOESY resonance signals that originate either from the protein of interest, or from all other cell components. Because of the r^{-6}-dependence of the NOEs, cross peaks arising from minor components in the crude cell-extract may have comparable or even higher intensities than a large percentage of the peaks originating from the target protein. Here, we tackled this challenge using the algorithms ATNOS (Herrmann et al., 2002b) for NOE cross peak identification and CANDID (Herrmann et al., 2002a) for NOE assignment and, in conjunction with DYANA (Güntert et al., 1997) structure calculation, performed spectral analysis of the 3D ^{13}C- and ^{15}N-resolved [^{1}H,^{1}H]-NOESY spectra, as illustrated in Figure 4.

Obtaining a defined and correct fold of the protein in the first cycle is a key feature for reliable and robust automated NOESY spectral analysis, since in subsequent cycles the analysis and the assignment of the NOESY spectra can predominantly be guided by reference to the intermediate protein structures. The three-dimensional structure obtained from ATNOS/CANDID in cycle 1 showed an average global root-mean-square deviation (rmsd) of 1.09 ± 0.07 Å for a bundle of 20 conformers relative to the mean coordinates. The rmsd between the mean coordinates of the structure bundle from cycle 1 and the mean coordinates of the reference protein structure determined using purified protein samples was 1.41 Å, as calculated for the backbone atoms of residues 4–43 and 55–110. These rmsd values show that the correct fold of the protein was obtained already after the first ATNOS/CANDID cycle, confirming that network-anchored NOE cross peak identification and assignment with these programs is highly efficient also under the difficult conditions in crude cell lysate (Figure 4, a and b). In the final ATNOS/CANDID cycle 7, a total of 3708 NOE cross peaks were assigned. A total of 1868 meaningful NOE upper distance limits comprised the input for the final structure calculation (Table 1), leading to a high-quality NMR structure (Figure 4), with a residual DYANA target function value of 1.54 ± 0.15 Å^2 and an average global root-mean-square deviation (rmsd) of 0.47 ± 0.05 Å for a

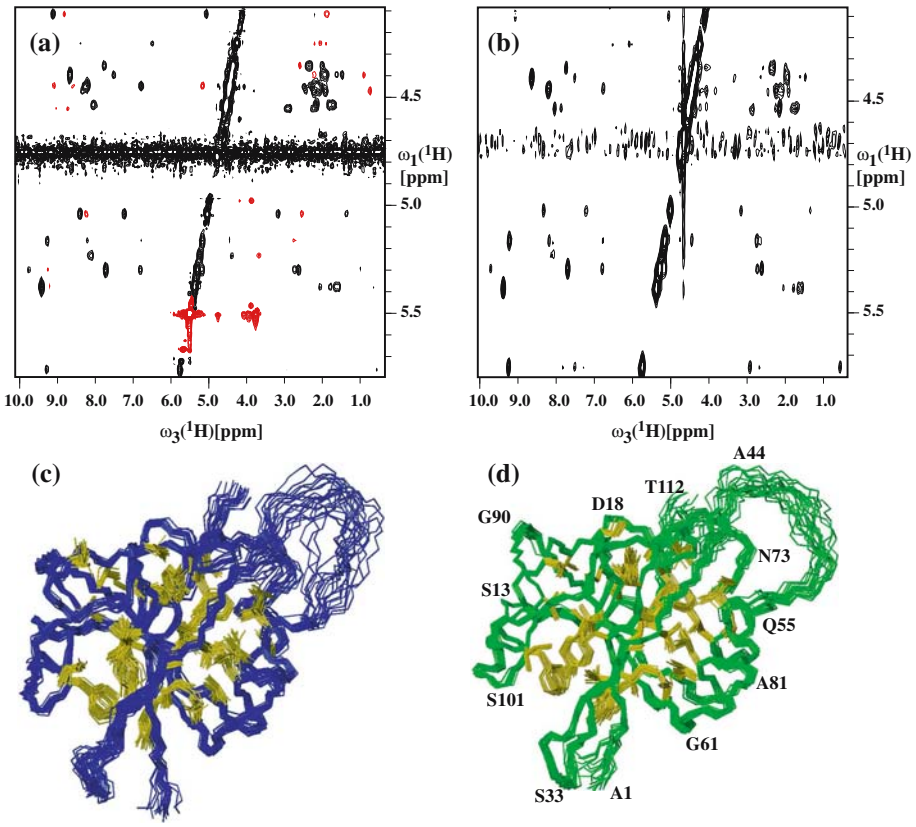

Figure 4. Comparison of the ^{13}C-plane at 53.6 ppm from 3D ^{13}C-resolved [^{1}H,^{1}H]-NOESY spectra of TM1290. (a) In cell-extract. Signals that had been picked and successively discarded during automated NOESY spectral analysis, due to missing network-anchored support, are highlighted in red (see text). (b) NMR sample of the purified protein in aqueous solution (Etezady-Esfarjani et al., 2004). (c) and (d) Bundles of 20 energy-minimized DYANA conformers representing the TM1290 solution structures obtained from crude cell-extract, and from a purified protein sample (data from Etezady-Esfarjani et al., 2004), respectively. The same viewing angle is used for the two figures, and for both proteins the superposition of the 20 conformers is for best fit of the backbone atoms N, C^{α} and C' of residues 4–43 and 55–110. Core side chains are shown in gold, and in (d) some sequence numbers are shown for reference.

bundle of 20 conformers (Figure 4c) relative to the mean coordinates (Table 1). For the backbone atoms of residues 4–43 and 55–110, a detailed comparison with the NMR structure determination of the purified protein (Etezady-Esfarjani et al., 2004) reveals close agreement between the mean coordinates of the two structure bundles, with a global rmsd of 1.08 Å calculated for the backbone heavy atoms. This value is only slightly larger than the sum of the rmsd values of the corresponding structure bundles. Differences between the two NMR structures are evenly distributed over the well-defined polypeptide segments of the protein, i.e., there are no specific regions of the protein showing outstandingly large structure variations. The close structural similarity of the two NMR structures extends also to the variation of the precision of the structure determination along the amino acid sequence (Figure 4, c and d). In particular, the amide ^{15}N- and ^{1}H-resonances of the central loop, Ala46-Lys52, could not be detected either in crude cell-extract or in a solution of purified TM1290. This observation shows for both environments, either that the amide protons of these residues undergo rapid exchange with the solvent, or that the amide groups are involved in slow conformational exchange.

Discussion

The present high-quality structure determination of the recombinant ^{13}C,^{15}N-labeled protein TM1290 in crude cell extract (Table 1, Figure 4c) was primarily based on exploiting the fact that the

ensemble of intra-protein ^{1}H–^{1}H NOEs forms a self-consistent data set. Using the ATNOS/CANDID algorithms (Herrmann et al., 2002a, b), background signals of comparable and even larger intensity than the intra-protein NOESY cross peaks (Figure 4a) were reliably eliminated during the iterative NOE assignment process. The generation of the correct polypeptide fold in the first cycle of the iterative structure determination confirmed previous conclusions about high robustness of the network-anchored concept for automated [^{1}H,^{1}H]-NOESY analysis. Inspection of the Figure 4a and b illustrates the challenge that was successfully met by ATNOS/CANDID. In the region of the spectrum recorded in cell extract that is shown in Figure 4a, 26 out of a total of 77 identified cross-peaks originate from the cell background (highlighted in red), i.e., 33% of all identified signals in the 3D ^{13}C-resolved[^{1}H,^{1}H]-NOESY data set had to be eliminated during the structure determination process. The somewhat lower precision of the structure determined in cell extract when compared with the corresponding data obtained with the purified protein (Table 1, Figure 4, c and d) is due to the fact that a higher number of NOE upper distance constraints could be identified in the more highly concentrated solution of purified TM1290 (3.8 mM) than in the

Table 1. Comparison of the input for the structure calculations and the characterization of the energy-minimized NMR structures of TM1290 from crude cell-extract (this paper) and from a solution of the purified protein (data from Etezady-Esfarjani et al., 2004)

Quantity	Value[a] (cell-extract)	Value[a] (purified)
NOE upper distance limits	1868	2444
Dihedral angle constraints	142	144
Residual target function [Å^2]	1.54 ± 0.15	1.91 ± 0.51
Residual NOE violations		
Number ≥0.1 Å	21 ± 4 (15–28)	34 ± 4 (27–42)
Maximum [Å]	0.14 ± 0.01 (0.13–0.16)	0.15 ± 0.01 (0.14–0.16)
Residual angle violations		
Number ≥2.5 deg	3 ± 1 (1–6)	7 ± 1 (4–10)
Maximum [deg]	6.39 ± 0.79 (4.44–7.55)	4.76 ± 0.72 (3.54–7.01)
Amber energies [kcal/mol]		
Total	−4573.02 ± 107.70	−4354.95 ± 62.88
van der Waals	−359.71 ± 14.50	−279.04 ± 15.93
Electrostatic	−5056.01 ± 102.92	−5008.90 ± 70.16
rmsd from ideal geometry		
Bond lengths [Å]	0.0075 ± 0.0001	0.0079 ± 0.0002
Bond angles [deg]	1.87 ± 0.05	2.08 ± 0.04
rmsd to the mean coordinates [Å][b,c]		
bb (4–43, 55–110)	0.47 ± 0.05 (0.33–0.55)	0.31 ± 0.03 (0.26–0.37)
ha (4–43, 55–110)	0.75 ± 0.07 (0.88–1.12)	0.75 ± 0.07 (0.65–0.91)
Ramachandran plot statistics		
Most favored region (%)	72.8	66.2
Additional allowed region (%)	24.7	29.9
Generously allowed region (%)	1.7	2.6
Disallowed region (%)	0.8	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.
[c]The bb (4–43, 55–110) rmsd between the TM1290 structures in the two environments, as calculated between the respective conformers with the smallest deviation from the average atom coordinates, is 1.09 Å.

crude cell-extract sample (0.8 mM). On average for the 111 assigned residues, an additional 1.0 sequential, 1.9 medium-range and 2.2. long-range NOEs were found in the more concentrated solution. Conversely, it is very encouraging that the NMR structure determination in crude cell-extract could be performed at a protein concentration below 1.0 mM.

Clearly, high overexpression of the target protein is needed to perform an automated structure determination with the protocol used here for TM1290. Broader use for less favorable proteins may be achieved with refined expression protocols, which reduce the background signals in cell extract samples. For example, in a two-step protein expression approach (Marley et al., 2001) the cells are initially grown in unlabeled medium, which is then replaced by the isotope-labeled high density growth medium during the induction phase. Alternatively, preferential labeling of the target protein has been achieved by inhibition of host protein synthesis through addition of rifampicin, which inhibits the bacterial RNA polymerase but not the T7 RNA polymerase, so that the target protein is more highly labeled with isotopes than the other cell components (Lee et al., 1995; Almeida et al., 2001). Since we were primarily focusing on a test of the ATNOS/CANDID algorithms under difficult solution conditions, we performed standard isotope-labeling, with the growth phase and the induction phase carried out in the same growth medium.

In conclusion, this paper demonstrates that NOE-based NMR structure determination of proteins under difficult solution conditions, such as in crude cell-extract, can be achieved with the use of automated, network-anchored [^{1}H,^{1}H]-NOESY analysis and structural interpretation, whereas collection of the conformational constraints with interactive approaches would be tedious and probably in most cases not practical. In view of commentaries received from the reviewers of this paper, we leave it up to our readers to hypothesize about practical applications of the experience gained in this paper. We merely want to indicate the promise emanating from the present work, that automatic procedures along the lines of the software ATNOS/CANDID (Hermann et al., 2002 a, b) might be able to support *de novo* protein structure determination *in situ*.

Acknowledgements

We thank Dr S. A. Lesley and H. E. Klock of the Genomics Institute of the Novartis Foundation in the JCSG consortium for providing us with the genomic materials. Financial support for T. E.-E. was obtained from the Schweizerischer Nationalfonds (project 31-66427-01); K.W. is the Cecil H. and Ida M. Green Professor of Structural Biology at TSRI. This work was sponsored by the National Institutes of Health, Protein Structure Initiative with Grant Nos. P50GM62411, U54GM07498.

References

Almeida, F.C., Amorim, G.C., Moreau, V.H., Sousa, V.O., Creazola, A.T., Americo, T.A., Pais, A.P., Leite, A., Netto, L.E., Giordano, R.J. and Valente, A.P. (2001) *J. Magn. Reson.*, **148**, 142–146.

Bartels, C., Xia, T.H., Billeter, M., Güntert, P. and Wüthrich, K. (1995) *J. Biomol. NMR*, **6**, 1–10.

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

Etezady-Esfarjani, T., Herrmann, T., Peti, W., Klock, H.E., Lesley, S.A. and Wüthrich, K. (2004) *J. Biomol. NMR*, **29**, 403–406.

Galvao-Botton, L.M., Katsuyama, A.M., Guzzo, C.R., Almeida, F.C., Farah, C.S. and Valente, A.P. (2003) *FEBS Lett.*, **552**, 207–213.

Gronenborn, A.M. and Clore, G.M. (1996) *Protein Sci.*, **5**, 174–177.

Güntert, P., Dötsch, V., Wider, G. and Wüthrich, K. (1992) *J. Biomol. NMR*, **2**, 619–629.

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

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

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

Hubbard, J.A., MacLachlan, L.K., King, G.W., Jones, J.J. and Fosberry, A.P. (2003) *Mol. Microbiol.*, **49**, 1191–1200.

Koradi, R., Billeter, M. and Güntert, P. (2000) *Comput. Phys. Commun.*, **124**, 139–147.

Lee, K.M., Androphy, E.J. and Baleja, J.D. (1995) *J. Biomol. NMR*, **5**, 93–96.

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

Luginbühl, P., Szyperski, T. and Wüthrich, K. (1995) *J. Magn. Reson. B*, **109**, 229–233.

Marley, J., Lu, M. and Bracken, C. (2001) *J. Biomol. NMR*, **20**, 71.

Muhandiram, D.R., Farrow, N.A., Xu, G.Y., Smallcombe, S.H. and Kay, L.E. (1993) *J. Magn. Reson.*, **102**, 317–321.

Nilges, M. (1997) *Fold. Des.*, **2**, S53–S57.

Oldfield, C.J., Ulrich, E.L., Cheng, Y., Dunker, A.K. and Markley, J.L. (2005) *Proteins*, **59**, 444–453.

Ou, H.D., Lai, H.C., Serber, Z. and Dötsch, V. (2001) *J. Biomol. NMR*, **21**, 269–273.
Page, R., Peti, W., Wilson, I.A., Stevens, R.C. and Wüthrich, K. (2005) *Proc. Natl. Acad. Sci. USA*, **102**, 1901–1905.
Peti, W., Etezady-Esfarjani, T., Herrmann, T., Klock, H.E., Lesley, S.A. and Wüthrich, K. (2004) *J. Struct. Funct. Genomics*, **5**, 205–215.
Sattler, M., Schleucher, J. and Griesinger, C. (1999) *Progr. Nucl. Magn. Reson. Spect.*, **34**, 93–158.
Serber, Z. and Dötsch, V. (2001) *Biochemistry*, **40**, 14317–14323.
Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids* Wiley, New York.
Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G.M., Bhattacharyya, S., Gutierrez, P., Denisov, A., Lee, C.H., Cort, J.R., Kozlov, G., Liao, J., Finak, G., Chen, L., Wishart, D., Lee, W., McIntosh, L.P., Gehring, K., Kennedy, M.A., Edwards, A.M. and Arrowsmith, C.H. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 1825–1830.
Zuiderweg, E.R.P. and Fesik, S.W. (1989) *Biochemistry*, **28**, 2387–2391.