J Biomol NMR (2011) 50:137–146
DOI 10.1007/s10858-011-9502-8

ARTICLE

Exclusively NOESY-based automated NMR assignment and structure determination of proteins

Teppei Ikeya · Jun-Goo Jee · Yoshiki Shigemitsu · Junpei Hamatsu · Masaki Mishima · Yutaka Ito · Masatsune Kainosho · Peter Güntert

Received: 18 January 2011 / Accepted: 11 March 2011 / Published online: 30 March 2011

Abstract A fully automated method is presented for determining NMR solution structures of proteins using exclusively NOESY spectra as input, obviating the need to measure any spectra only for obtaining resonance assignments but devoid of structural information. Applied to two small proteins, the approach yielded structures that coincided closely with conventionally determined structures.

Keywords Automated assignment · CYANA · FLYA · SAIL · Protein structures

Introduction

All NMR data required for a protein structure determination based on nuclear Overhauser effect (NOE) derived distance restraints is contained in the NOESY spectrum (Wüthrich 1986). In practice, however, the evaluation of NOESY spectra requires sequence-specific resonance assignments that are determined using additional spectra to delineate through-bond connections. A considerable amount of measurement time and interactive work is needed for spectra that are used only for the chemical shift assignment but do not contribute structural information. To circumvent the chemical shift assignment step, "assignment-free" methods for NMR protein structure determination have been proposed (Atkinson and Saudek 2002; Grishaev and Llinás 2002; Kraulis 1994; Malliavin et al. 1992) but are rarely used because their requirements on the quality of the input NMR data are difficult to meet experimentally. Assignment algorithms that use NOESY data have been developed earlier, however not without requiring information from through-bond spectra or other sources. For instance, the ASCAN algorithm assigns side-chain chemical shifts, provided that the backbone assignments are already known (Fiorito et al. 2008). Several algorithms for the structure-based chemical shift assignment of proteins (Bailey-Kellogg et al. 2000; Dobson et al. 1984; Pristovšek et al. 2002; Stratmann et al. 2009) make use of NOEs but cannot be applied for structure determination because they require the three-dimensional (3D) structure as input. Here, we present a fully automated method for protein structure determination by NMR that relies exclusively on the NOESY spectra to simultaneously find the chemical shift assignments, conformational restraints, and the 3D structure. This is different from the commonly used automated algorithms for the assignment of NOESY cross peaks which require the sequence-specific resonance assignments to be known (Güntert 2009; Herrmann et al. 2002; Huang et al. 2006; Mumenthaler et al. 1997; Nilges 1995).

Electronic supplementary material The online version of this article (doi:10.1007/s10858-011-9502-8) contains supplementary material, which is available to authorized users.

T. Ikeya · P. Güntert (✉)
Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, and Frankfurt Institute for Advanced Studies, Goethe University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany
e-mail: guentert@em.uni-frankfurt.de

T. Ikeya · J.-G. Jee · Y. Shigemitsu · J. Hamatsu · M. Mishima · Y. Ito · M. Kainosho · P. Güntert
Graduate School of Science, Tokyo Metropolitan University, 1-1 Minami-Osawa, Hachioji, Tokyo 192-0397, Japan

M. Kainosho
Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8602, Japan

Materials and methods

NOESY-FLYA algorithm

Our approach is based on the fully automated NMR protein structure determination algorithm FLYA (López-Méndez and Güntert 2006), which has been adapted for use with only the protein sequence and 3D ^{13}C- or ^{15}N-resolved NOESY spectra as input (Fig. 1). Neither other spectra, nor manually prepared peak lists, nor structural information is needed as input.

Peak positions and intensities are identified using the automated peak picking algorithms of the programs NMRView (Johnson 2004) or AZARA (http://www.ccpn.ac.uk/azara). Since no manual corrections are made, the resulting raw peak lists may contain, in addition to the entries representing true signals, a significant number of artifacts. The following steps of the fully automated structure determination algorithm can tolerate the presence of such artifacts, as long as the majority of the true peaks have been identified. Based on the NOESY peak positions and peak volumes peak lists are prepared by CYANA (Güntert 2003; Güntert et al. 1997). Depending on the spectra, the preparation may include unfolding aliased signals, systematic correction of chemical shift referencing, and removal of peaks near the diagonal or water lines. The peak lists resulting from this step remain invariable throughout the rest of the procedure.

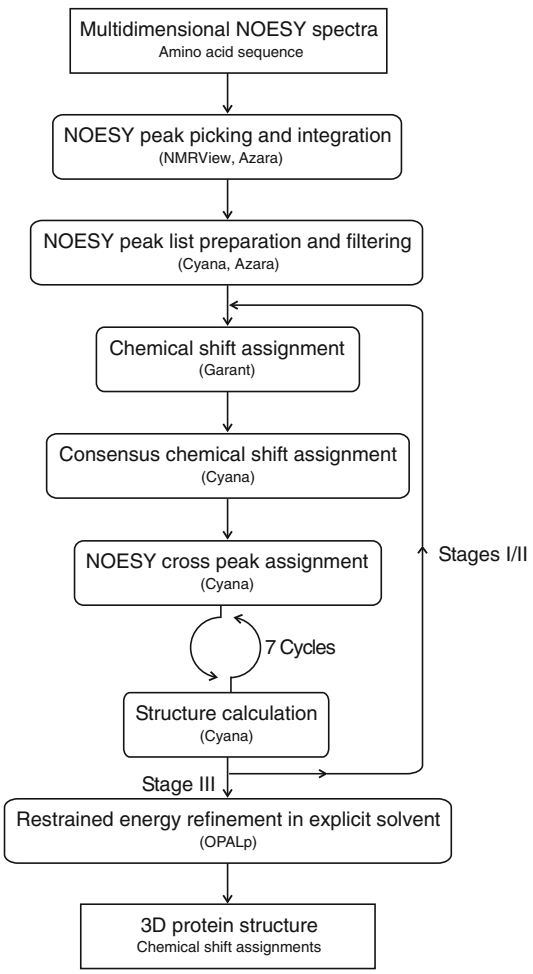


Fig. 1 Flowchart of the NOESY-FLYA algorithm

An ensemble of initial chemical shift assignments is obtained by multiple runs of a modified version of the GARANT algorithm (Bartels et al. 1996, 1997) with different seed values for the random number generator (Malmodin et al. 2003). The original GARANT algorithm was modified for the treatment of NOESY spectra when 3D structures are available (Ikeya et al. 2009). First a list of expected peaks is constructed, initially based only on the amino acid sequence. Only intraresidual and sequential connectivities can, with certain limitations, be predicted from the sequence alone. Subsequently, when a first preliminary structure has been obtained, the expected NOESY cross peaks are generated on the basis of spatial proximities in the preliminary structure, which allows to predict also medium-range and long-range NOEs. These expected peaks are then mapped onto the experimentally observed peaks. While the positions of the observed peaks are known precisely, their assignment is initially unknown. In contrast, the expected peaks are, by construction, always unambiguously assigned to atoms and characterized by approximate positions from a chemical shift database statistics and an a priori probability to be observed. The optimal mapping of the expected peaks onto the observed peaks implies the chemical shift assignments of the atoms involved in the mapped expected peaks. The search for this optimal mapping is complicated by the fact that both peak lists may be incomplete and may contain errors. Thus, NOE cross-peaks arising due to the close proximity of pairs of hydrogen atoms cannot be predicted from the sequence alone, or entries in the list of observed peaks may be missing, for example because of spectral overlap or limited signal to noise. Experimental lists of observed peaks may include spurious entries resulting, for example, from spectral noise. The GARANT algorithm includes a scheme for evaluating the quality of non-final assignments in order to drive the search for the optimal mapping of expected peaks onto the observed peaks. The optimization procedure is a combination of a genetic algorithm with a local optimization routine (Bartels et al. 1997).

In analogy to NMR structure calculation in which not a single structure but an ensemble of conformers is calculated using identical input data but different randomized start conformers, the initial chemical shift assignment produces an ensemble rather than a single chemical shift value for each ^{1}H, ^{13}C and ^{15}N nucleus. These initial chemical shift assignments are consolidated by CYANA into a single consensus chemical shift list, which is used for

generating angle restraints with the program TALOS (Cornilescu et al. 1999) and for the assignment of NOE distance restraints. The most highly populated chemical shift value in the ensemble is computed for each ^{1}H, ^{13}C and ^{15}N spin and selected as the consensus chemical shift value that will be used for the subsequent automated assignment of NOESY peaks. The consensus chemical shift for a given nucleus is the value ω that maximizes the function $\mu(\omega) = \sum_j \exp(-(\omega - \omega_j)^2/2\Delta\omega^2)$, where the sum runs over all chemical shift values ω_j for the given nucleus in the ensemble of initial chemical shift assignments, and $\Delta\omega$ denotes the chemical shift tolerance as it is used elsewhere in the algorithm for matching peak positions and chemical shift values. It is typically set to 0.03 ppm for the ^{1}H dimensions and to 0.3 ppm for the ^{13}C and ^{15}N dimensions. NOESY cross peaks are assigned automatically (Herrmann et al. 2002) on the basis of the consensus chemical shift assignments and the same peak lists and chemical shift tolerance values used already for the chemical shift assignment. The automated NOE assignment algorithm of the program CYANA is used. The overall probability for the correctness of possible NOE assignments is calculated as the product of three probabilities that reflect the agreement between the chemical shift values and the peak position, the consistency with a preliminary 3D structure (Güntert et al. 1993), and network-anchoring (Herrmann et al. 2002), i.e. the extent of embedding in the network formed by other NOEs. Restraints with multiple possible assignments are represented by ambiguous distance restraints (Nilges 1995). Seven cycles of combined automated NOE assignment and structure calculation by simulated annealing in torsion angle space and a final structure calculation using only unambiguously assigned distance restraints are performed. Constraint combination (Herrmann et al. 2002) is applied in the first two cycles to all NOE distance restraints spanning at least three residues in order to minimize distortions of the structures by erroneous distance restraints that may result from spurious entries in the peak lists and/or incorrect chemical shift assignments.

A complete FLYA calculation comprises three stages. In the first stage, the chemical shifts and protein structures are generated de novo (stage I). In the next stages (stages II and III), the structures generated by the preceding stage are used as additional input for the determination of chemical shift assignments. At the end of the third stage, the 20 final CYANA conformers with the lowest target function values are subjected to restrained energy minimization in explicit solvent against the AMBER force field (Cornell et al. 1995; Ponder and Case 2003) using the program OPALp (Koradi et al. 2000; Luginbühl et al. 1996). The complete procedure is driven by the NMR structure calculation program CYANA, which is also used for parallelization of all time-consuming steps.

Proteins and NMR spectroscopy

We applied the algorithm to the 3D structure determinations of *Chlorella* ubiquitin (76 residues) and the *Thermus thermophilus* HB8 protein TTHA1718 (66 residues) (Ikeya et al. 2010; Sakakibara et al. 2009). The ubiquitin sample was produced by *E. coli* cell-free protein synthesis optimized for the preparation of labeled NMR samples (Takeda et al. 2007; Torizawa et al. 2004) using 50 mg of stereo-array isotope labeled (SAIL) (Kainosho et al. 2006) amino acid mixture (SAIL Technologies) (Ikeya et al. 2009). Uniformly ^{13}C/^{15}N labeled TTHA1718 was expressed in *E. coli* cells (Ikeya et al. 2010; Sakakibara et al. 2009). The protein concentrations were 0.4 and 2.6 mM in sodium phosphate (pH 6.6 and 7.0) with 10% $^{2}H_2O$, respectively. The SAIL ubiquitin sample was the same as for our earlier study using NOESY in conjunction with different sets of through-bond spectra (Ikeya et al. 2009). This allowed for a comparison between the NOESY-only approach and FLYA calculations with conventional input data sets.

3D ^{13}C- or ^{15}N-resolved NOESY spectra with a mixing time of 100 ms and 1,024 × 218 × 50/36 (^{1}H × ^{1}H × ^{13}C/^{15}N) data points for ubiquitin and a mixing time of 80 ms and 1,024 × 256 × 64/64, data points for TTHA1718 were recorded at 310 K on Bruker DRX and Avance 600 spectrometers equipped with cryogenic probes. Additional ^{13}C-resolved NOESY spectra were recorded for the region of the aromatic ^{13}C resonances. 2D maximum entropy processing with AZARA was used for the NOESY spectra of TTHA1718 as described previously (Ikeya et al. 2010; Sakakibara et al. 2009). The maximum entropy processing resulted in better resolution and a higher signal-to-noise ratio than conventional Fourier transformation, thus facilitating the peak identification in overlapped and complicated regions, particularly for the ^{13}C-resolved NOESY of uniformly ^{13}C/^{15}N-labeled proteins. The ubiquitin spectra exhibited the sharp lines as typical for SAIL, and were processed by conventional Fourier transformation. The chemical shift assignments and the solution structure were compared to those from the conventional approach based on manual peak picking, manual chemical shift assignment, and automated NOESY assignment (Güntert 2009; Herrmann et al. 2002), referred to as the "reference" (Ikeya et al. 2009; Sakakibara et al. 2009). For ubiquitin, a crystal structure is available of the human protein (Vijay-Kumar et al. 1987), which differs in two sequence positions from *Chlorella* ubiquitin. It has a backbone RMSD to the reference structure of 1.29 Å for residues 1–72.

Automated peak picking

Automatic peak picking was performed over the entire spectra, excluding only two narrow bands along the diagonal and the water line. Previous manual or automatic chemical shift assignment approaches required at least one 2D or 3D spectrum for delineating through-bond connections besides the NOESY spectra. These additional through-bond spectra contain fewer peaks than NOESY, which renders them also useful for filtering noise and artifacts by checking the consistency of peak positions among different spectra. This approach cannot be adopted when using exclusively NOESY, and the potentially large number of noise and artifact entries in automatically prepared NOESY peak lists, which comprise only position and intensity information, could lead to errors in the FLYA calculations. To minimize this possible problem, we exploited the fact that noise and artifacts are generally not uniformly distributed in a spectrum, so that it makes sense to determine the noise level for peak picking locally using a similar approach as in the peak picking program AUTOPSY (Koradi et al. 1998). A local noise level determination and a diagonal filter were implemented into AZARA version 2.8 as follows. For a given (potential) peak in an n-dimensional spectrum we consider a region of $L_1 \times \cdots \times L_n$ data points centered at the position of the peak. For each of the n one-dimensional slices through the peak a sliding average of the squared spectral intensities is computed over all stretches of length $l_1, \ldots, l_n$ within the region, and its maximal values stored as $a_1^2, \ldots, a_n^2$. A local noise level for the position of the peak is then defined by $D = w\left(1/n\sum_{k=1}^{n} a_k^2\right)^{1/2} + D_g$, where w is a weighting factor for the locally determined noise level and D_g is a user-defined global noise level for the whole spectrum. For all applications in this paper, $l_1, \ldots, l_n$ corresponded to 5% of all data points in the given dimension, $L_k = 2l_k$ $(k = 1, \ldots, n)$, and $w = 3.0$.

In most 2D and 3D NOESY spectra the non-diagonal cross peaks are accompanied by corresponding diagonal peaks with equal sign of the intensity. The cross peaks can therefore be filtered with respect to the corresponding diagonal intensities, and potential cross peaks with a diagonal intensity below a user-defined threshold are discarded. Although such a check could in principle be applied to a preliminary peak list it was incorporated into the peak picking to resolve diagonal peaks.

Assignment and structure calculation

Peak lists were not edited interactively. The peak position tolerance for the GARANT and CYANA assignment algorithms was set to 0.03 ppm for ^{1}H, and 0.3 ppm for ^{13}C and ^{15}N. Upper distance limits were derived from the NOESY peaks according to a r^{-6} intensity-to-distance relationship and confined to the range 2.4–5.2 Å. Restraints on the ϕ and ψ torsion angles were produced with TALOS (Cornilescu et al. 1999). No hydrogen bond restraints were applied. Structure calculations started from 500 conformers with random torsion angles, five times more than the default, to minimize the possible influence of erroneous chemical shift assignments. Ten runs of the entire procedure were conducted with identical input data using different seed values for the random number generator (Supporting Information Tables S1 and S2). The run that yielded the structure bundle with the lowest average AMBER energy was selected and analyzed.

Structure analysis

The program MOLMOL (Koradi et al. 1996) was used to visualize 3D structures. CYANA was used to obtain statistics on target function values and restraint violations, and to compute RMSD values to the mean coordinates of a structure bundle for superpositions of the backbone atoms N, C and C′ or the heavy atoms for the structured regions of the proteins. Conformational energies were calculated with OPALp (Koradi et al. 2000; Luginbühl et al. 1996) using the AMBER force field (Cornell et al. 1995; Ponder and Case 2003).

Results

NOESY-FLYA assignment and structure determination

The local noise level determination and the diagonal filter reduced the number of artifact entries in the peak list substantially (Fig. 2). For instance, while the original algorithm in AZARA picked 5,735 peaks in the 3D ^{13}C-edited NOESY spectrum of TTHA1718, the use of local noise level determination and diagonal filtering excluded many artifact peaks and reduced the number of picked peaks to 2,694. This is not much more than the 2,299 peaks that were identified by a human expert. Nevertheless, a significant number of artifacts must have remained in the automatically picked peak list because 1,212 of these peaks were left unassigned by the subsequent automated NOESY assignment for the structure calculation.

Using 5 Intel Xeon 2.8 GHz Quad-core processors, a NOESY-FLYA run was completed in less than 6 h for ubiquitin. The results of the calculations are summarized in Table 1. Further details are given in Supplementary Tables S1 and S2. Of all chemical shift assignments, 9 and 13% were wrong for ubiquitin and TTHA1718, respectively. These numbers decreased slightly (by less than 1%) for the

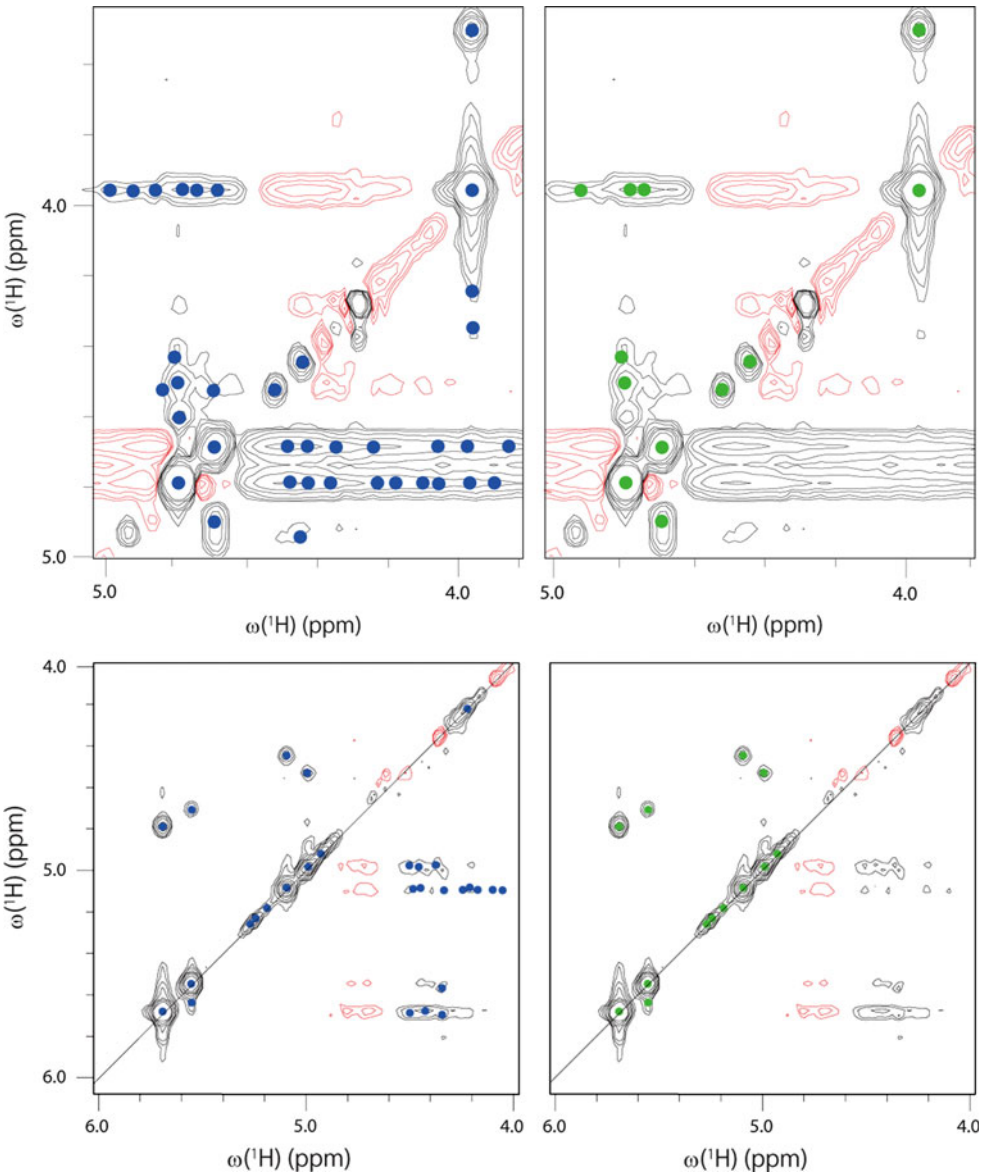


Fig. 2 Automated peak picking of the 3D ^{13}C-edited NOESY spectrum of the protein TTHA1718 with the program AZARA using the standard algorithm based on a constant noise level (*left*; *picked peaks in blue*) and an improved algorithm that employs a local noise level determination and diagonal filter functions (*right*; *picked peaks in green*). Excerpts from the spectral planes at $\omega(^{13}C) = 27.8$ ppm (*top*) and 16.9 ppm (*bottom*)

residues in regular secondary structure elements. The NOESY-FLYA calculations yielded structures in good agreement with the reference structures (Fig. 3). The backbone RMSDs to the reference structures were 0.87 and 1.07 Å for ubiquitin and TTHA1718, respectively. This indicates that the automated NOE assignment algorithm in CYANA produced self-consistent sets of correct distance restraints on the basis of the imperfect sequence-specific resonance assignments obtained in the preceding step of the FLYA algorithm. The accuracy of the resulting structures in terms of the RMSD to the conventionally determined reference structure was about the same for both proteins. The SAIL labeling of ubiquitin simplifies the task of the NOESY-FLYA algorithm by yielding spectra with sharper and fewer lines and reduces the number of assignment possibilities of the remaining peaks. On the other hand, the ubiquitin measurements were made with a ~6 times more dilute sample and on a lower-field spectrometer than those for the slightly smaller TTHA1718 protein.

In the case of ubiquitin, fully automated structure determinations with FLYA had been performed earlier using different selections of through-bond spectra in addition to the NOESYs (Ikeya et al. 2009). For instance, when the NOESY spectra were complemented by 12 through-bond spectra (Supplementary Table S3), the percentage of chemical shifts that agreed with the reference was about 10% higher than with the exclusively NOESY-based approach (Ikeya et al. 2009). Nevertheless, the subsequent CYANA calculation yielded with both sets of input spectra a similar number of long-range distance restraints, and structures in equally good agreement with the

Table 1 Statistics of fully automated FLYA structure determinations of ubiquitin and TTHA1718 using as input only NOESY spectra

Quantity	ubiquitin	TTHA1718
Equal $^{1}H/^{13}C/^{15}N$ assignments (%)[a]	85.8	78.7
Different $^{1}H/^{13}C/^{15}N$ assignments (%)[a]	13.1	20.7
Wrong $^{1}H/^{13}C/^{15}N$ assignments (%)[a]	9.3	13.0
Assigned NOESY cross peaks	1,701	2,343
Long-range ($\|i-j\| \geq 5$) distance restraints	264	441
CYANA target function (Å2)[b]	0.287 ± 0.001	2.780 ± 0.046
AMBER energy (kcal/mol)[c]	−3,119 ± 60	−1,928 ± 73
Backbone RMSD to mean (Å)[d]	0.28 ± 0.05	0.20 ± 0.03
Heavy atom RMSD to mean (Å)[d]	0.66 ± 0.07	0.65 ± 0.05
Backbone RMSD to reference (Å)[e]	0.87	1.07
Heavy atom RMSD to reference (Å)[e]	1.32	1.72

[a] Chemical shift assignments were classified as 'equal' if they coincided, within tolerances of 0.03 ppm for ^{1}H and 0.3 ppm for $^{13}C/^{15}N$, with the corresponding reference assignment, as 'different' if they differed by more than the tolerance from the reference assignment, and as 'wrong' if they did not match any conventionally assigned shift within the same residue. The wrong assignments are a subset of the different ones. In addition, the algorithm yielded assignments for a small number of 0.6–1.1% of the chemical shifts for which no reference assignment was available, and that could thus not be classified. In the case of ubiquitin, the unstructured C-terminal region of residues 73–76 was excluded

[b] For the 20 CYANA conformers with lowest target function values

[c] For the 20 energy-refined CYANA conformers that represent the solution structure

[d] Between the 20 energy-refined conformers and their mean coordinates for the backbone atoms N, C^{α} and C' or for all heavy atoms in the structured regions of residues 1–72 in ubiquitin and 1–66 in TTHA1718

[e] Between the mean coordinates of the 20 energy-refined conformers and the conventionally determined reference structure

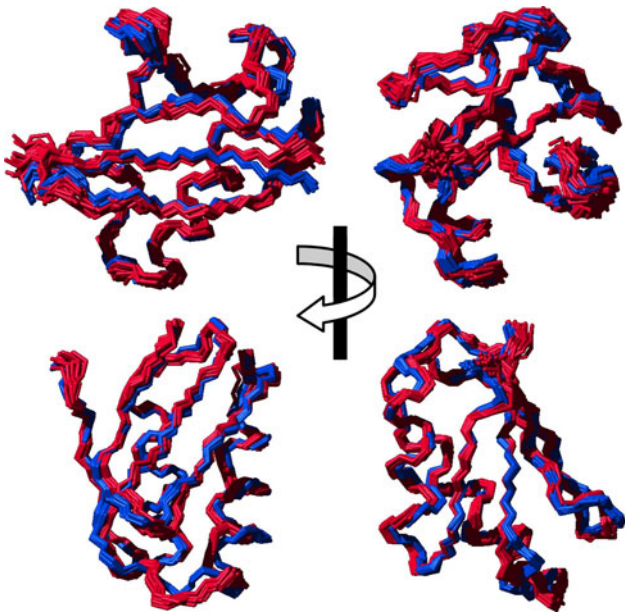


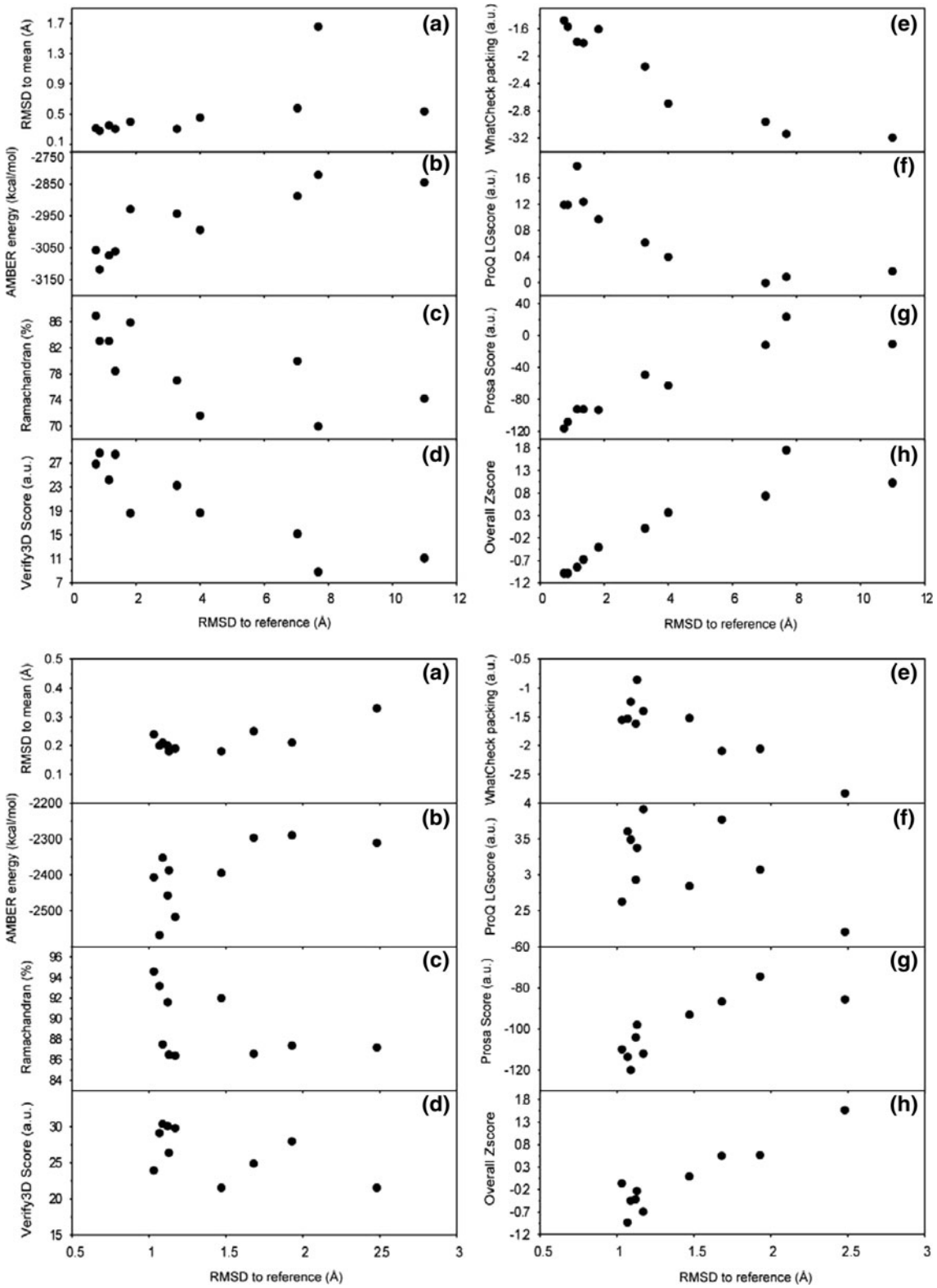
Fig. 3 Structures of ubiquitin (*top*) and TTHA1718 (*bottom*) obtained using exclusively NOESY spectra for chemical shift assignment and structure calculation (*blue*) superimposed on the conventionally determined NMR solution structures (*red*)

reference structure. The backbone RMSDs to the reference structure were 0.87 and 0.93 Å for the lowest average AMBER energy structures, which is comparable to the estimated accuracy of 0.8 Å of the reference structure (Ikeya et al. 2009).

It has been shown earlier that about 90% correct chemical shift assignments are required for successful combined automated NOESY assignment and structure calculation with CYANA (Jee and Güntert 2003). The NOESY-only approach with 9–13% wrong assignments is at the limit of this rule. In fact, not all of the 10 NOESY-FLYA calculations converged to an accurate structure. The backbone RMSDs to the reference structure were 0.75, 0.87, 1.15, 1.35, 1.82, 3.27, 4.00, 7.03, 7.68, and 11.0 Å in the case of ubiquitin (Supporting Information Table S1) and 1.03, 1.07, 1.09, 1.12, 1.13, 1.17, 1.47, 1.68, 1.93, and 2.48 Å in the case of TTHA1718 (Supporting Information Table S2). Recognizing correct structures is thus crucial for the reliability of the NOESY-only approach.

Structure validation

The 4 out of 10 ubiquitin structures with less than 1.5 Å backbone RMSD to the reference structure could be distinguished correctly from the others either by the number of assigned NOESY cross peaks, the number of long-range distance restraints (Supporting Information Table S1), the AMBER potential energy (Ponder and Case 2003), the Verify3D score (Lüthy et al. 1992), the packing score of the Whatcheck program (Hooft et al. 1996), the LGscore of the ProQ program (Wallner and Elofsson 2003), or the score of the ProSa 2003 program (Sippl 1993). An overall *Z*-score (Ikeya et al. 2009) that



combines the latter five validation scores, the RMSD to the mean coordinates, and the percentage of residues in the most favored region of the Ramachandran plot, as defined by the program Procheck (Laskowski et al. 1993), shows for both proteins a particularly clear correlation with the RMSD to the reference structure (Fig. 4). The

◀**Fig. 4** Validation scores for NOESY-FLYA structures of ubiquitin (*top*) and TTHA1718 (*bottom*) plotted against the backbone RMSD deviation from the reference structure. For both proteins 10 runs of the entire procedure were conducted with identical input data using different seed values for the random number generator. Each run produced a bundle of 20 conformers and is represented by one dot in the panels. **a** Backbone RMSD to the mean coordinates. **b** AMBER potential energy. **c** Percentage of residues in the most favored region of the Ramachandran plot, as defined by the program Procheck. **d** Verify3D score. **e** Packing score of the Whatcheck program. **f** LGscore of the ProQ program. **g** Score of the ProSa 2003 program. **h** Overall *Z*-score, computed from the seven individual validation parameters

overall *Z*-score was computed from the seven individual validation parameters as $Z = 1/7 \sum_{i=1}^{7} (S_i - \bar{S}_i)/\sigma(S_i)$, where S_i is the value of an individual validation score with the sign chosen such that lower values represent better scores, $\bar{S}_i$ its mean value, and $\sigma(S_i)$ its standard deviation. Correct structures can thus be discriminated reliably from seriously wrong ones. For both proteins the overall *Z*-scores were about -1.2 in the best case and below 0.0 for the good structures. These absolute values may not be universally valid for other proteins. However, the overall *Z*-score was used here only for a relative ranking of the structures obtained by multiple runs of the same algorithm for the same protein, a less difficult task for which the overall *Z*-score gave consistent results.

Discussion

Our results show that NOESY spectra alone can yield sufficiently accurate chemical shift assignments to obtain high-quality solution structures of proteins. This constitutes a significant conceptual advance by concentrating the whole NMR measurement effort on the spectrum type that provides the structural data. In practice, it enables faster structure determination. The NOESY-only chemical shift assignments are still less reliable than those of the conventional approach, but improving the NOESY spectra and assignment algorithms may in the future close the gap and make the approach applicable to larger proteins, for which the algorithm in its present form failed to yield correct structures. It would be difficult to give a clear-cut size limit because the performance of the algorithm depends as much on the quality of the NMR data, which determines how reliably peaks can be identified, as on the size of the protein, which determines the number of ambiguous assignment possibilities for NOEs. Higher effective dimensionality, non-linear sampling, and alternatives to Fourier transformation can be expected to provide NOESY spectra with better resolution and less overlap (Luan et al. 2005; Malmodin and Billeter 2005), thus helping the algorithm to cope with its main challenge, namely that a priori peak assignments are more ambiguous in NOESY than in through-bond spectra. The observations that in the present calculations still significant fractions of the automatically picked peaks remain unassigned and that despite relatively similar chemical shift assignments the accuracy of the structures resulting from different NOESY-FLYA runs varied considerably suggest two promising directions for future developments of the FLYA algorithm. One may target the automated peak picking with the (competing) aims of reducing the number of artifacts picked and enhancing the identification of weak peaks, which include many long-range NOEs with strong impact on the quality of the resulting structure. The other direction of development may lead to a more robust combined NOESY assignment and structure calculation algorithm that is less susceptible to erroneous chemical shift assignments than the CANDID algorithm (Herrmann et al. 2002) and the current algorithm in CYANA (Güntert 2009), which require about 90% correct chemical shift assignments (Jee and Güntert 2003). Improvements in the automated chemical shift assignment part (Bartels et al. 1997) of the FLYA algorithm are also conceivable.

Although already the original homonuclear sequence-specific resonance assignment method for proteins relied on NOESY data (in conjunction with COSY and TOCSY) (Wüthrich 1986), so far all manual or automated assignment strategies required low ambiguity through-bond data. Only the present NOESY-FLYA approach uses exclusively NOESY data. Since reliable algorithms for the automated assignment of distance restraints (Güntert 2009; Herrmann et al. 2002) and the structure calculation (Güntert et al. 1997) were already available in CYANA, the principal obstacle that had to be overcome by the present NOESY-FLYA method was the exclusively NOESY-based automated sequence-specific resonance assignment. The reason for its success can be rationalized as follows. Existing automated resonance assignment algorithms generally work by unambiguously assembling small groups of resonances into "spin systems" which are subsequently mapped to the primary structure of the protein (Baran et al. 2004; Gronwald and Kalbitzer 2004). Characteristically, the assignment process is divided into distinct steps, e.g. spin system identification, backbone assignment, and side-chain assignment, each of which assumes the results from the preceding steps to be given and correct. In fact, these algorithms work best for the backbone assignment of uniformly $^{13}C/^{15}N$-labeled proteins on the basis of triple resonance spectra (Moseley et al. 2001), which have a low inherent ambiguity. This "build-up approach" cannot be followed with NOESY spectra alone because the simultaneous presence of short-range, medium-range and long-range NOEs makes it impossible to unambiguously group resonances into spin systems that remain fixed for the remainder of the algorithm. Instead, the assignment can

only be established by simultaneously considering correlations for the backbone and the side-chains of spatially neighboring residues. This is achieved by a general formulation of the resonance assignment problem as finding an optimal match between the experimental NOESY cross peaks and those that are expected given the protein sequence (Bartels et al. 1997; López-Méndez and Güntert 2006). In the case of NOESY-FLYA these expected peaks are generated in stage I from the protein sequence by applying empirical rules that describe which atom pairs commonly give rise to identifiable NOESY cross peaks (Bartels et al. 1997; Billeter et al. 1982; Wüthrich 1986). At this stage, all expected cross peaks are intraresidual or sequential. In the following NOESY-FLYA stages II and III, the expected NOESY cross peaks are generated on the basis of spatial proximity in preliminary structures, whereby NOEs of any sequence range are included naturally into the assignment process. This approach takes into account all data—backbone and side-chain, short-, medium- and long-range—simultaneously and does not require any particular peak to be present in the input experimental peak list. A second crucial feature of our method is an effective algorithm (Bartels et al. 1996, 1997) to score and optimize the agreement between the experimental peaks, whose position and intensity but not their assignment are known, and the expected peaks, which are assigned but whose precise location is unknown, under the high-ambiguity condition of NOESY spectra.

In this paper we have given a proof of principle that NMR structures can be solved exclusively from NOESY spectra. To our knowledge, this had never been achieved so far, and is thus of interest even though the NOESY-only approach remains at present less robust than the conventional one and recording additional through-bond spectra for the backbone and side-chain chemical shift assignment would not be unfeasible.

Acknowledgments We thank Drs. H. Yoshida, T. Terauchi, and A. M. Ono for preparing the SAIL ubiquitin sample. Financial support by a Grant-in-Aid for Scientific Research of the Japan Society for the Promotion of Science, the Targeted Proteins Research Program of the Ministry of Education, Culture, Sports, Science and Technology of Japan, and the Lichtenberg program of the Volkswagen Foundation is gratefully acknowledged.

References

Atkinson RA, Saudek V (2002) The direct determination of protein structure by NMR without assignment. FEBS Lett 510:1–4

Bailey-Kellogg C, Widge A, Kelley JJ, Berardi MJ, Bushweller JH, Donald BR (2000) The NOESY JIGSAW: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. J Comput Biol 7:537–558

Baran MC, Huang YJ, Moseley HNB, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. Chem Rev 104:3541–3555

Bartels C, Billeter M, Güntert P, Wüthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. J Biomol NMR 7:207–213

Bartels C, Güntert P, Billeter M, Wüthrich K (1997) GARANT–a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J Comput Chem 18:139–149

Billeter M, Braun W, Wüthrich K (1982) Sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra: computation of sterically allowed proton proton distances and statistical-analysis of proton proton distances in single-crystal protein conformations. J Mol Biol 155:321–346

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302

Dobson CM, Howarth MA, Redfield C (1984) Nuclear Overhauser effects and the assignment of the proton NMR spectra of proteins. FEBS Lett 176:307–312

Fiorito F, Herrmann T, Damberger FF, Wüthrich K (2008) 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 42:23–33

Grishaev A, Llinás M (2002) CLOUDS, a protocol for deriving a molecular proton density via NMR. Proc Natl Acad Sci USA 99:6707–6712

Gronwald W, Kalbitzer HR (2004) Automated structure determination of proteins by NMR spectroscopy. Prog Nucl Magn Reson Spectrosc 44:33–96

Güntert P (2003) Automated NMR protein structure calculation. Prog Nucl Magn Reson Spectrosc 43:105–125

Güntert P (2009) Automated structure determination from NMR spectra. Eur Biophys J 38:129–143

Güntert P, Berndt KD, Wüthrich K (1993) The program ASNO for computer-supported collection of NOE upper distance constraints as input for protein structure determination. J Biomol NMR 3:601–606

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

Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319:209–227

Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381:272

Huang YJ, Tejero R, Powers R, Montelione GT (2006) A topology-constrained distance network algorithm for protein structure determination from NOESY data. Proteins 62:587–603

Ikeya T, Takeda M, Yoshida H, Terauchi T, Jee J, Kainosho M, Güntert P (2009) Automated NMR structure determination of stereo-array isotope labeled ubiquitin from minimal sets of spectra using the SAIL-FLYA system. J Biomol NMR 44:261–272

Ikeya T, Sasaki A, Sakakibara D, Shigemitsu Y, Hamatsu J, Hanashima T, Mishima M, Yoshimasu M, Hayashi N, Mikawa T, Nietlispach D, Wälchli M, Smith BO, Shirakawa M, Güntert P, Ito Y (2010) NMR protein structure determination in living *E. coli* cells using nonlinear sampling. Nat Protoc 5:1051–1060

Jee J, Güntert P (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. J Struct Funct Genom 4:179–189

Johnson BA (2004) Using NMRView to visualize and analyze the NMR spectra of macromolecules. Meth Mol Biol 278:313–352

Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Ono AM, Güntert P (2006) Optimal isotope labelling for NMR protein structure determinations. Nature 440:52–57

Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 14:51–55

Koradi R, Billeter M, Engeli M, Güntert P, Wüthrich K (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. J Magn Reson 135:288–297

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

Kraulis PJ (1994) Protein three-dimensional structure determination and sequence-specific assignment of ^{13}C-separated and ^{15}N-separated NOE data–a novel real-space ab initio approach. J Mol Biol 243:696–718

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

López-Méndez B, Güntert P (2006) Automated protein structure determination from NMR spectra. J Am Chem Soc 128:13112–13122

Luan T, Jaravine V, Yee A, Arrowsmith CH, Orekhov VY (2005) Optimization of resolution and sensitivity of 4D NOESY using multi-dimensional decomposition. J Biomol NMR 33:1–14

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

Lüthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with 3-dimensional profiles. Nature 356:83–85

Malliavin TE, Rouh A, Delsuc MA, Lallemand JY (1992) Approche directe de la détermination de structures moléculaires à partir de l'effet Overhauser nucléaire. C R Acad Sci II 315:653–659

Malmodin D, Billeter M (2005) High-throughput analysis of protein NMR spectra. Prog Nucl Magn Reson Spectrosc 46:109–129

Malmodin D, Papavoine CHM, Billeter M (2003) Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. J Biomol NMR 27:69–79

Moseley HNB, Monleon D, Montelione GT (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Meth Enzymol 339:91–108

Mumenthaler C, Güntert P, Braun W, Wüthrich K (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. J Biomol NMR 10:351–362

Nilges M (1995) Calculation of protein structures with ambiguous distance restraints–automated assignment of ambiguous NOE crosspeaks and disulfide connectivities. J Mol Biol 245:645–660

Ponder JW, Case DA (2003) Force fields for protein simulations. Adv Prot Chem 66:27–85

Pristovšek P, Rüterjans H, Jerala R (2002) Semiautomatic sequence-specific assignment of proteins based on the tertiary structure–the program st2nmr. J Comput Chem 23:335–340

Sakakibara D, Sasaki A, Ikeya T, Hamatsu J, Hanashima T, Mishima M, Yoshimasu M, Hayashi N, Mikawa T, Wälchli M, Smith BO, Shirakawa M, Güntert P, Ito Y (2009) Protein structure determination in living cells by in-cell NMR spectroscopy. Nature 458:102–105

Sippl MJ (1993) Recognition of errors in 3-dimensional structures of proteins. Proteins 17:355–362

Stratmann D, van Heijenoort C, Guittet E (2009) NOEnet-use of NOE networks for NMR resonance assignment of proteins with known 3D structure. Bioinformatics 25:474–481

Takeda M, Ikeya T, Güntert P, Kainosho M (2007) Automated structure determination of proteins with the SAIL-FLYA NMR method. Nat Protoc 2:2896–2902

Torizawa T, Shimizu M, Taoka M, Miyano H, Kainosho M (2004) Efficient production of isotopically labeled proteins by cell-free synthesis: a practical protocol. J Biomol NMR 30:311–325

Vijay-Kumar S, Bugg CE, Cook WJ (1987) Structure of ubiquitin refined at 1.8 Å resolution. J Mol Biol 194:531–544

Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073–1086

Wüthrich K (1986) NMR of proteins and nucleic acids. Wiley, New York