J Biomol NMR (2013) 57:193–204
DOI 10.1007/s10858-013-9779-x

ARTICLE

Reliability of exclusively NOESY-based automated resonance assignment and structure determination of proteins

Elena Schmidt · Peter Güntert

Received: 9 July 2013 / Accepted: 2 September 2013 / Published online: 15 September 2013

Abstract Protein structure determination by NMR can in principle be speeded up both by reducing the measurement time on the NMR spectrometer and by a more efficient analysis of the spectra. Here we study the reliability of protein structure determination based on a single type of spectra, namely nuclear Overhauser effect spectroscopy (NOESY), using a fully automated procedure for the sequence-specific resonance assignment with the recently introduced FLYA algorithm, followed by combined automated NOE distance restraint assignment and structure calculation with CYANA. This NOESY-FLYA method was applied to eight proteins with 63–160 residues for which resonance assignments and solution structures had previously been determined by the Northeast Structural Genomics Consortium (NESG), and unrefined and refined NOESY data sets have been made available for the Critical Assessment of Automated Structure Determination of Proteins by NMR project. Using only peak lists from three-dimensional ^{13}C- or ^{15}N-resolved NOESY spectra as input, the FLYA algorithm yielded for the eight proteins 91–98 % correct backbone and side-chain assignments if manually refined peak lists are used, and 64–96 % correct assignments based on raw peak lists. Subsequent structure calculations with CYANA then produced structures with root-mean-square deviation (RMSD) values to the manually determined reference structures of 0.8–2.0 Å if refined peak lists are used. With raw peak lists, calculations for 4 proteins converged resulting in RMSDs to the reference structure of 0.8–2.8 Å, whereas no convergence was obtained for the four other proteins (two of which did already not converge with the correct manual resonance assignments given as input). These results show that, given high-quality experimental NOESY peak lists, the chemical shift assignments can be uncovered, without any recourse to traditional through-bond type assignment experiments, to an extent that is sufficient for calculating accurate three-dimensional structures.

Keywords Automated resonance assignment · Automated NOESY assignment · Protein structure determination · CASD-NMR · FLYA · CYANA

Introduction

NMR is considered as a slow method for protein structure determination compared to X-ray crystallography (once suitable crystals have been obtained) due to long measurement times and the time-consuming spectral analysis, which is necessary to relate the measured signals to the respective protein atoms. Much of the required time is spent on the chemical shift assignment of the protein atoms. In general, a set of through-bond spectra that reveal the atom connectivities in the protein backbone and side-chains is specifically recorded for the purpose of determining the resonance assignments. The resonance

Electronic supplementary material The online version of this article (doi:10.1007/s10858-013-9779-x) contains supplementary material, which is available to authorized users.

E. Schmidt · P. Güntert (✉)
Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, 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

P. Güntert
Graduate School of Science and Engineering, Tokyo Metropolitan University, 1-1 Minami-ohsawa, Hachioji, Tokyo 192-0397, Japan

assignment step is time-consuming because it is usually still done manually, even though many automated approaches have been developed and discussed in reviews (Altieri and Byrd 2004; Baran et al. 2004; Guerry and Herrmann 2011; Güntert 2009; Williamson and Craven 2009). Assignment-specific spectra make the resonance assignment feasible and robust, but provide in general no conformational information that is relevant for protein structure determination. On the other hand, in principle all information required for the atom assignment is also contained in ^{15}N-resolved NOESY and ^{13}C-resolved NOESY spectra, but the a priori large number of assignment possibilities, missing peaks, and spectral overlap make it difficult to assign proteins solely based on NOESY spectra. Therefore, this “NOESY-only” approach has hardly or never been used although it would make the measurement of assignment-specific spectra dispensable and, if combined with automated chemical shift assignment, the time for data measurement and analysis could be reduced considerably. The principal applicability of automated NOESY based chemical shift assignment for NMR structure determination was demonstrated on two small proteins, i.e. stereo-array isotope labeled (SAIL) (Kainosho et al. 2006) ubiquitin (76 residues) and the *Thermus thermophilus* HB8 protein TTHA1718 (66 residues) (Ikeya et al. 2011). Ikeya et al. used automatically picked NOESY peak lists and the GARANT algorithm (Bartels et al. 1996, 1997) for the automated resonance assignment. The results showed that it is in principle possible to obtain sufficiently complete assignments exclusively from NOESY spectra, but the approach lacked robustness: even when applied to two small, well-behaved proteins, not every structure calculation converged to a well-defined and correct structure. This is presumably due to the fact that the completeness of correct resonance assignments remained slightly below the threshold considered necessary for reliable automated NOESY distance restraint assignment for the structure calculation (Herrmann et al. 2002a; Jee and Güntert 2003).

Recently, we introduced the new FLYA automated resonance assignment algorithm (Schmidt and Güntert 2012), which is able to assign the chemical shifts to the respective atoms based on any combination of NMR spectra that contains the complete connectivity information of the atoms, and that yields a higher degree of correct resonance assignments than the former GARANT algorithm. We demonstrated the use of the FLYA resonance assignment, which is part of the CYANA (Güntert 2009; Güntert et al. 1997) software package, by assigning proteins based on a set of 15 or 16 through-bond and NOESY spectra for resonance assignment. In addition the FLYA algorithm was applied to proteins for which data was recorded with solid state NMR (Schmidt et al. 2013) and for the assignment of RNAs (Aeschbacher et al. 2013). Here we present the automated chemical shift assignment and structure calculation of eight proteins with sequence lengths of 63–160 residues. We show that with the new FLYA algorithm and carefully prepared peak lists it is possible for all these proteins to obtain almost complete resonance assignments and high-quality structures from NOESY spectra alone.

Materials and methods

Algorithm

NMR resonance assignment is based on several experiments that couple atom signals such that they can be measured as multidimensional peaks in the corresponding spectra. Assignment experiments are chosen to complement each other in such a way that the connectivity of the atoms in a protein can be represented by a network of peaks that are expected to be observed. Mapping this network of expected peaks with unknown positions to the unassigned measured peaks with known positions provides an assignment of the frequencies to the atoms (Bartels et al. 1996, 1997). The FLYA resonance assignment algorithm (Schmidt and Güntert 2012) that has been implemented in the CYANA software package (Güntert 2009; Güntert et al. 1997), uses this general approach to assign all types of NMR spectra, those which are based on scalar couplings as well as experiments that take advantage of the nuclear Overhauser effect or corresponding solid state NMR experiments. The CYANA software package is available to other users; details and documentation are given at www.cyana.org.

The FLYA algorithm starts by deducing the expected peak network from the protein sequence and the experiment specifications. For NOE-based experiments expected peaks can in general only be predicted for pairs of atoms that are close in sequence. Expected peaks resulting from long-range contacts can only be obtained if the 3D structure of the protein is available. To determine sequence-based contacts, 20 random structures with the sequence of the respective protein are calculated without using experimental restraints and expected NOESY peaks are generated for ^{1}H-^{1}H contacts with a user defined maximal distance in all 20 structures. Assuming that the probability of a peak to be actually measured in the experiment decreases with the distance between the respective atoms, distance-dependent probabilities are assigned to the expected peaks.

The mapping of expected peaks to measured peaks is done using an evolutionary optimization routine that works with a population of individuals, each representing an assignment solution (Schmidt and Güntert 2012). The evolutionary optimization is complemented by a local optimization routine. Solutions that are produced during the optimization are created such that the search space of an expected peak for a mapping is consistent with general chemical shift statistics (by default from the BMRB (Ulrich et al. 2008), or user defined), the deviation of the measured frequencies of different measured peaks that are assigned to the same atom remain within a given tolerance, and an expected peak can be mapped to only one measured peak. The first generation of solutions is generated randomly, but fulfilling these criteria. In each generation a local optimization algorithm takes small parts of a mapping back and reassigns the expected peaks for a defined number of iterations, 15,000 is default. Afterwards the different solutions of one generation are recombined to a new generation. The individuals and the specific parts of an individual that contribute to a new individual are selected via a scoring function that takes into account the distribution of chemical shift values with respect to the given shift statistics, the alignment of peaks assigned to the same atom, the completeness of the assignment, and a penalty for chemical shift degeneracy. The solution that maximizes this function is given as final assignment at the end of the calculation.

To increase the accuracy of the assignment, and to obtain a reliability measure for each assigned atom, several independent runs of the algorithm, 20 for all calculations in this paper, with different random seeds are performed. From the resulting 20 chemical shift values for each atom a consensus chemical shift value and a measure of the self-consistency of the assignment are computed. The self-consistency measure equals the fraction of runs yielding a chemical shift value that is, within user-defined tolerances, in agreement with the consensus chemical shift value of the atom. Experience has shown (López-Méndez and Güntert 2006; Malmodin et al. 2003; Schmidt and Güntert 2012) that assignments with high self-consistency ("strong" assignments) are more reliable than others ("weak" assignments).

Data sets from CASD-NMR

To avoid a possible source of bias, we based this study exclusively on input data that had been prepared independently by other researchers before the present study was conceived. Automated chemical shift assignment and structure calculation were therefore performed with input data from eight different proteins that were provided in 2011–2012 as test data sets in the CASD-NMR project (Rosato et al. 2009, 2012), i.e. the human NFU1 iron-sulfur cluster scaffold homolog, Northeast Structural Genomics Consortium (NESG) target HR2876B (PDB accession code 2LTM, 107 amino acid residues); the human mitotic checkpoint serine/threonine-protein kinase BUB1 N-terminal domain, HR5460A (2LAH, 160 aa); the RRM domain of RNA-binding protein FUS, HR6430A (2LA6, 99 aa), the homeobox domain of the human homeobox protein Nkx-3.1, HR6470A (2L9R, 69 aa), a de novo designed protein with IF3-like fold, OR135 (2LN3, 83 aa) (Koga et al. 2012), a de novo designed protein with P-loop NTPase fold, OR36 (2LCI, 134 aa), TSTM1273 from *Salmonella typhimurium* LT2, StT322 (2LOJ, 63 aa), and the NifU-like protein from *Saccharomyces cerevisiae* YR313A (2LTL, 119 aa). The corresponding structures, which were manually refined by experts, have already been released to the PDB and were used as reference structures when evaluating results of the calculations in this paper. In principle, it would be desirable to compare the results of our structure calculations to independently determined X-ray structures but these are not available for the proteins of the CASD-NMR project.

The NMR data were prepared and analyzed according to standard procedures of the NESG structural genomics consortium (Liu et al. 2005, 2010), using mainly the program XEASY (Bartels et al. 1995) for peak picking and interactive spectra analysis. For every protein two different datasets were used, each containing a ^{15}N-resolved NOESY and a ^{13}C-resolved NOESY peak list (Table 1). The ^{13}C-resolved NOESY peak list was obtained by combining the two original peak lists for the aliphatic and aromatic ^{1}H–^{13}C groups from CASD-NMR. One data set contained the "refined" NOESY peak lists that were used as input for the final structure calculations that lead to the reference structures, i.e. the large majority of cross peaks in the refined peak lists resulted in conformational restraints for the final structure calculations and there are only few noise peaks. The second data set contained "raw" NOESY peak lists from an early stage of the spectra analysis before the structure calculations. The NESG researchers generated initial NOESY peak lists, which were then manually edited by limited visual inspection of the NOESY spectra (Liu et al. 2005, 2010). The percentage of real peaks in the peak lists (column 'Assigned' in Table 1) was estimated by checking whether a peak is present in a specific peak list for every expected peak. The respective search range was defined by the reference chemical shifts and the tolerance that was also used for the assignment calculations. The refined and raw data sets differ mainly in the number of artifact peaks which amount to 1–12 % for the refined peak lists but 4–68 % for the raw peak lists (Table 1). The root-mean-square deviation between the reference chemical shift value and the chemical shift position coordinates of the measured peak is higher for the raw peak lists than for the refined peak lists (Table 1). In some cases it is nearly twofold higher than the corresponding value for the refined peak lists. Table 1 thus shows that the raw peak lists are in general far from being perfect.

Assignment calculations

The FLYA resonance assignment was used in the same way and with the same parameters for all data sets. 20 random structures fulfilling steric restraints were calculated. Expected

Table 1 Experimental peak lists

Protein	Spectrum	Expected peaks	Measured peaks	Assigned (%)	Complete (%)	Deviation (ppm)
HR2876B refined	15N-resolved NOESY	3,791	1,436	99	47	0.32
	13C-resolved NOESY	17,647	5,618	97	44	0.40
HR2876B raw	15N-resolved NOESY	3,791	2,060	72	51	0.43
	13C-resolved NOESY	17,647	12,042	39	40	0.73
HR5460A refined	15N-resolved NOESY	5,478	2,964	94	61	0.30
	13C-resolved NOESY	19,975	9,051	92	55	0.40
HR5460A raw	15N-resolved NOESY	5,478	4,172	58	60	0.47
	13C-resolved NOESY	19,975	13,078	40	40	0.78
HR6430A refined	15N-resolved NOESY	2,835	1,501	97	60	0.30
	13C-resolved NOESY	11,078	5,142	96	56	0.38
HR6430A raw	15N-resolved NOESY	2,835	1,628	95	62	0.31
	13C-resolved NOESY	11,078	5,197	93	55	0.40
HR6470A refined	15N-resolved NOESY	2,008	949	99	53	0.26
	13C-resolved NOESY	7,240	3,266	96	56	0.34
HR6470A raw	15N-resolved NOESY	2,008	988	96	53	0.28
	13C-resolved NOESY	7,240	3,267	88	48	0.43
OR135 refined	15N-resolved NOESY	3,424	1,529	99	57	0.36
	13C-resolved NOESY	13,928	4,830	97	50	0.46
OR135 raw	15N-resolved NOESY	3,424	2,937	50	48	0.47
	13C-resolved NOESY	13,928	4,812	54	34	0.71
OR36 refined	15N-resolved NOESY	3,984	2,125	97	64	0.32
	13C-resolved NOESY	16,187	7,334	95	59	0.45
OR36 raw	15N-resolved NOESY	3,984	2,634	59	56	0.51
	13C-resolved NOESY	16,187	11,160	34	41	0.78
StT322 refined	15N-resolved NOESY	1,922	835	94	53	0.28
	13C-resolved NOESY	7,846	1,622	97	33	0.33
StT322 raw	15N-resolved NOESY	1,922	1,793	38	48	0.47
	13C-resolved NOESY	7,846	3,190	45	31	0.63
YR313A refined	15N-resolved NOESY	4,770	1,605	99	42	0.33
	13C-resolved NOESY	18,982	4,987	98	37	0.42
YR313A raw	15N-resolved NOESY	4,770	1,984	64	38	0.49
	13C-resolved NOESY	18,982	10,319	32	29	0.74

Expected peaks Number of expected peaks by FLYA using the reference structure for the generation of distance dependent expected peaks. *Measured peaks* Number of measured peaks. *Assigned* Percentage of measured peaks that can be assigned within a tolerance of 0.03 ppm for ^{1}H and 0.3 ppm for ^{13}C and ^{15}N, based on the reference chemical shift assignments. The theoretical maximum of 100 % corresponds to having all measured peaks assigned. Note that several expected peaks can be mapped to the same measured peak, i.e. the assignments of measured peaks can be unambiguous or ambiguous. Remaining unassigned measured peaks are likely to be artifacts. *Complete* Percentage of expected peaks that can be mapped to a measured peak based on the reference chemical shift assignments. The theoretical maximum of 100 % corresponds to the situation that the spectra "explain" all expected peaks. Each expected peak can be mapped to at most one measured peak. Remaining expected peaks correspond to missing peaks in the measured peak list. *Deviation* Root-mean-square deviation between the chemical shift position coordinates of the measured peaks to which an expected peak can be mapped and the corresponding reference chemical shift value

NOESY peaks were generated for ^{1}H–^{1}H distances up to 6 Å that could be observed in all 20 random structures. Expected peak probabilities were set to 0.9, 0.8, 0.7, 0.6, and 0.5 for distances up to 4.0, 4.5, 5.0, 5.5, and 6.0 Å, respectively. The tolerance for the chemical shift assignment calculations was 0.03 ppm for ^{1}H and 0.3 ppm for ^{13}C and ^{15}N. The same tolerances were used for the determination of the assignments and their evaluation by comparison with the manually determined reference assignments. The population size of the evolutionary algorithm was 200, and 15,000 iterations of the local optimization were allowed per generation of the evolutionary algorithm. Chemical shift assignments were consolidated from 20 independent runs with different random number generator seeds. The assignment of an atom was classified as "strong" if at least 80 % of the 20 chemical shift values deviated by less than the tolerance of 0.03 ppm for ^{1}H and

0.3 ppm for ^{13}C and ^{15}N from the consensus value. Other assignments were classified as “weak”.

Structure calculations

All (strong and weak) chemical shift assignments from the FLYA algorithm were used as input for the automated assignment of NOESY cross peaks, which were converted to distance restraints for the structure calculation. The combined automated NOESY assignment and structure calculation by the standard CYANA method (Güntert 2009) used as input the protein sequence, the list of assigned chemical shifts from the FLYA algorithm, and the unassigned NOESY peak lists. The tolerance for chemical shift and peak position matching was again set to 0.03 ppm for ^{1}H and 0.3 ppm for ^{13}C and ^{15}N. Torsion angle restraints were generated on the basis of the chemical shift values with the program TALOS+ (Shen et al. 2009) for the backbone torsion angles φ and ψ of all non-proline residues with a prediction classified as “Good” by TALOS+. The torsion angle restraints were centered at the predicted average value and their full width was set to four times the predicted standard deviation or 20°, whichever was larger. The program CYANA was used for seven cycles of combined automated NOE assignment (Herrmann et al. 2002a) and structure calculation by torsion angle dynamics (Güntert et al. 1997). Peak intensities were converted into upper distance bounds according to a $1/r^6$-relationship. Each structure calculation was started from 200 conformers with random torsion angle values, the standard CYANA simulated annealing schedule was applied with 15,000 torsion angle dynamics steps, and the 20 conformers with lowest CYANA target function values were analyzed. NOE distance restraints involving ^{1}H atoms with degenerate chemical shifts, e.g. methyl groups, were treated as ambiguous distance restraints using $1/r^6$-summation over the distances to the individual ^{1}H atoms. Non-stereospecifically assigned methyls and methylene protons were treated by automatic swapping of restraints between diastereotopic partners (Folmer et al. 1997) during the seven cycles of automated NOE assignment, and by pseudoatom correction and symmetrization (Güntert 1998; Güntert et al. 1991) for the final structure calculation.

The entire procedure was driven by the program CYANA, which was also used for the parallelization of the FLYA resonance assignment and the CYANA structure calculations on 20 processors of a Linux cluster system with Intel quad-core 2.8 GHz processors.

Results and discussion

The fully automated method for resonance assignment, distance restraint assignment, and structure calculation was applied to the refined and raw NOESY data sets of eight proteins. The computation times varied between 32 min for the refined data set of the protein HR6470A and 105 min for the raw data set of the protein HR6470A. About 70 % of the computation time was required for the FLYA automated resonance assignment, and the remaining 30 % for the combined distance restraint assignment and structure calculation.

Automated chemical shift assignment results for the eight proteins are shown in Table 2. For the refined peak lists more than 95 % of the assignments were correct in all cases, except for StT322 for which only 90.6 % of the assignments were in agreement with the reference. The best result with 98.1 % correctness was obtained for the refined lists of YR313A. This shows that with high-quality NOESY data it is possible to determine the resonance assignments of proteins almost as completely and reliably as from a set of specialized through-bond assignment spectra. For the raw peak lists the correctness of the automated chemical shift assignment was in the range of 64.1–95.8, or 1.2–31.1 % points lower than with the corresponding refined NOESY data set. In all cases the correctness of the backbone assignment was higher than the correctness of the side-chain assignment. The results for the backbone atoms were between 0.5 and 5.2 % points better than for all atoms.

Strong assignments that are self-consistent over at least 80 % of 20 independent runs of the algorithm are considered more reliable than others (Schmidt and Güntert 2012). In all 16 calculations the percentage of correct assignments was always slightly higher than the percentage of strong assignments, i.e. the percentage of strong assignments is a lower bound for the correctness of the assignments. All calculations with more than 80 % strong assignments yielded more than 90 % correct assignments, and those with more than 70 % strong assignments yielded more than 80 % correct assignments. For HR5460A and OR36, for which there are fewer than 50 % strong assignments with raw peak lists, the correctness is below 70 %.

The results for the individual assignments are shown in Figs. 1 and S1 for the refined data sets and in Figs. 2 and S2 for the raw data sets. In most of the successful calculations the few wrong assignments are distributed over the sequence without an obvious pattern. Overall Phe and charged residues exhibit a comparatively high number of erroneous assignments. Two or more erroneous assignments accumulate in several of these residues. In most of the calculations with the raw data sets (Figs. 2 and S2) some erroneous assignments neighbor the regions that have no reference assignments at the beginning or end of the sequence. In case of the raw data sets of HR5460A and OR36 (Fig. 2b and S2b), which have the lowest percentage of correct assignments, wrong assignments and assignments without reference cluster in groups of up to

Table 2 Resonance assignments by the FLYA algorithm

	Assignments (%)		
	Correct (backbone)	Correct (all atoms)	Strong (all atoms)
HR2876B refined	98.3	95.3	84.0
HR2876B raw	87.3	82.1	71.3
HR5460A refined	97.7	95.2	83.3
HR5460A raw	66.5	64.1	30.4
HR6430A refined	98.8	97.0	86.2
HR6430A raw	98.8	95.8	86.3
HR6470A refined	99.1	97.5	85.5
HR6470A raw	94.0	91.2	76.8
OR135 refined	97.6	97.0	87.9
OR135 raw	84.8	80.7	77.7
OR36 refined	97.4	95.9	87.6
OR36 raw	71.7	68.8	41.6
StT322 refined	92.0	90.6	83.6
StT322 raw	86.6	82.5	75.1
YR313A refined	98.6	98.1	84.1
YR313A raw	86.5	84.9	75.2

Assignments are considered as correct if they agree with the manually determined reference assignment within the chemical shift tolerance of 0.3 ppm for ^{13}C and ^{15}N and 0.03 for ^{1}H. *Correct (backbone)* The same for the atoms H, N, C$^{\alpha}$ and C$^{\beta}$. 100 % corresponds to all backbone reference assignments. *Correct (all atoms)* Comparison of all FLYA assignments for which a reference assignment is present. 100 % corresponds to all reference assignments. *Strong (all atoms)* means that the assignment is consistent within at least 80 % of 20 independent runs of the algorithm. 100 % corresponds to all assignments

13 residues. In case of the refined lists of OR36 and StT322 (Fig. S1b and S1c) several wrong assignments cluster in residues 18–22 and 58–62, respectively.

Strong assignments are more likely to agree with the reference. Since it can be assumed that for atoms without a manually determined reference assignment the measured data do not contain sufficient information for determining the respective assignment, these atoms should not receive strong assignments. Except for the refined peak lists of HR6470A and OR36 the His-tag regions at the beginning or the end of the sequence, which have no reference assignments, are nearly completely classified as weak. Regions of several connected residues with erroneous assignments are correctly marked as weak in all cases. Only few single atoms in these regions form exceptions. The results vary in the different calculations with erroneous assignments that are distributed over the sequence. In the result from the refined lists of YR313A only 2 assignments are erroneously marked as strong. Overall, the correctness of the strong assignments is 98.2–99.8 % using refined peak lists, and 86.3–98.5 % using raw peak lists. In all cases at least about half of the erroneous assignments are correctly marked as weak. Weak, but nevertheless correct assignments occur often around wrong assignments or regions without reference assignments (Figs. 1b, S1b, S1c, S1d, 2b, S2b).

The NOESY cross peak assignment for generating distance restraints for the structure calculation was performed with the standard CYANA algorithm for this purpose (Güntert 2009; Herrmann et al. 2002a), using the FLYA resonance assignments as input. Of all NOESY cross peaks 89.0–95.3 % could be assigned using the refined peak lists, and 35.7–93.7 % using the raw peak lists (Table 3). This shows that the refined NOESY peak lists contain almost only peaks which can be assigned and converted to distance restraints that can be fulfilled in the CYANA structure calculation. With raw peak lists the extent of NOEs that can be assigned and interpreted as consistent distance restraints is in 5 out of 8 cases much lower, i.e. for the proteins HR2876B, HR5460A, OR36, StT322, and YR313A only between 35 and 42 % of the NOESY cross peaks can be used (Table 3), indicating that the raw peak lists for these proteins contain a large number of artifacts.

The NOE assignment and structure calculations progress iteratively over 7 cycles (Güntert 2009; Herrmann et al. 2002a). It has been shown that achieving convergence in the structure calculation of the first cycle is crucial for reliably obtaining a correct structure, and that an Root-mean-square deviation (RMSD) value to the mean coordinates of less than 3 Å in cycle 1 (not in the final structure calculation) is a sufficient criterion for convergence (Herrmann et al. 2002a; Jee and Güntert 2003). This

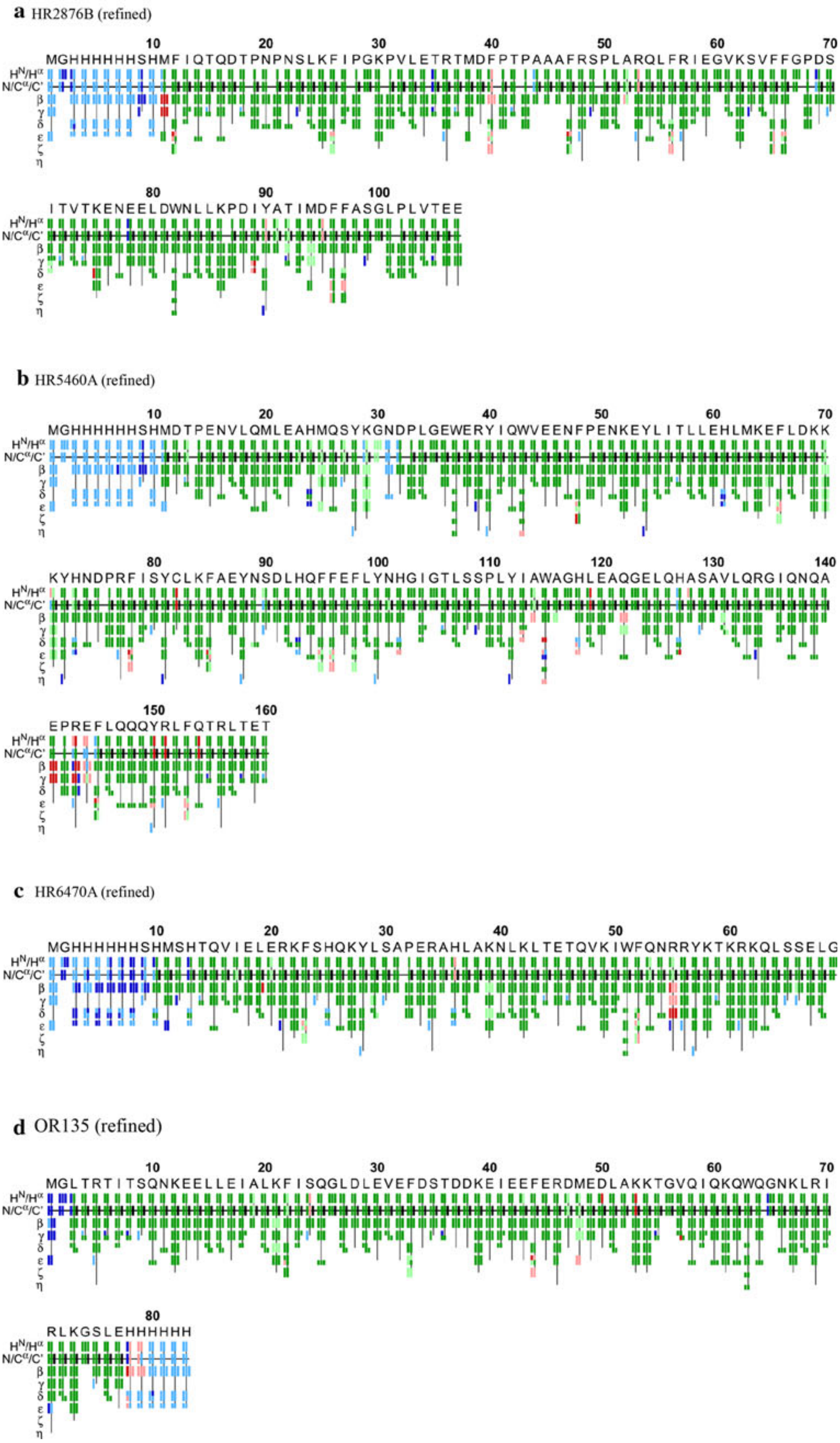

Fig. 1 Extent, correctness, and reliability of individual assignments obtained with the FLYA automated resonance assignment algorithm using refined peak lists for **a** HR2876B, **b** HR5460A, **c** HR6470A, and **d** OR135. Each assignment for an atom is represented by a *colored rectangle*. *Green*, assignment by FLYA agrees with the manually determined reference assignment within a tolerance of 0.03 ppm for ^{1}H and 0.3 ppm for ^{13}C and ^{15}N; *red*, assignment differs from reference; *blue*, assigned by FLYA but no reference available; *black*, with reference assignment but not assigned by FLYA. Respective *light colors* indicate assignments not classified as strong by the chemical shift consolidation. The *row* labeled H^{N}/H$^{\alpha}$ shows for each residue HN on the *left* H$^{\alpha}$ in the *center*. The N/C$^{\alpha}$/C′ *row* shows for each residue the N, C$^{\alpha}$, and C′ assignments from *left* to *right*. The *rows* β–η show the side-chain assignments for the heavy atoms in the *center* and hydrogen atoms to the *left* and *right*. In the case of branched side-chains, the corresponding *row* is split into an upper part for one branch and a lower part for the other branch

Fig. 2 Extent, correctness, and reliability of individual assignments obtained with the FLYA automated resonance assignment algorithm using raw peak lists for **a** HR2876B, **b** HR5460A, **c** HR6470A, and **d** OR135. See Fig. 1 for details

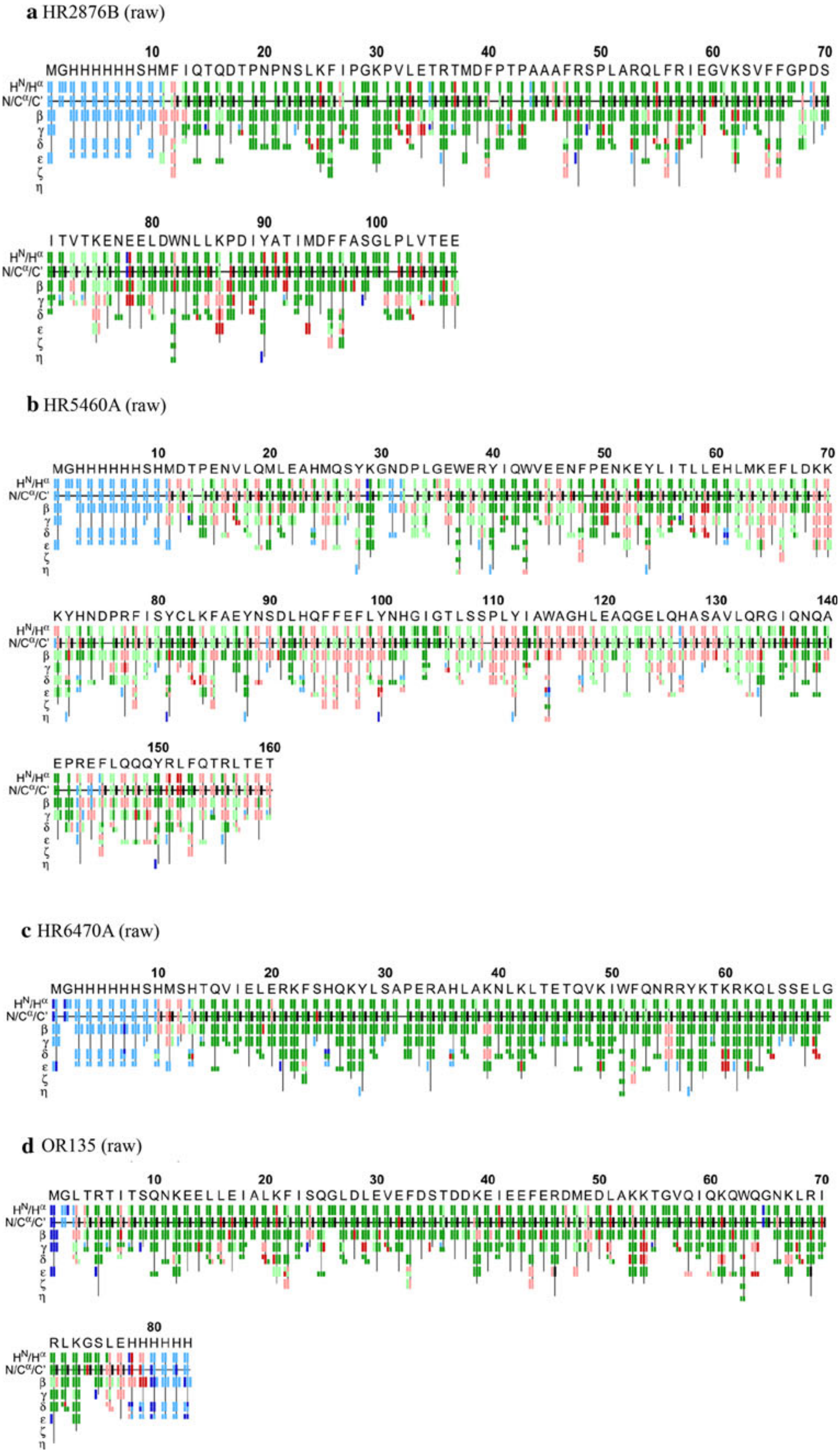

Table 3 Structure calculation

Data set	Backbone RMSD to mean, average ± s.d. (Å)		Assigned peaks in NOE assignment (%)	RMSD to reference (Å)	
	In cycle 1	Final		Using automatically assigned shifts	Using reference shifts
HR2876B refined	0.58 ± 0.10	0.22 ± 0.05	93.1	1.04	0.79
HR2876B raw	2.42 ± 0.54	0.12 ± 0.04	38.9	1.56	0.95
HR5460A refined	1.56 ± 0.34	0.34 ± 0.08	89.7	1.73	1.38
HR5460A raw	6.59 ± 0.84	0.46 ± 0.17	37.1	14.27	3.38
HR6430A refined	0.48 ± 0.08	0.11 ± 0.03	95.3	0.80	0.92
HR6430A raw	0.52 ± 0.08	0.20 ± 0.04	93.7	0.98	1.15
HR6470A refined	0.51 ± 0.11	0.27 ± 0.08	93.3	1.02	0.37
HR6470A raw	0.84 ± 0.16	0.26 ± 0.06	91.4	0.77	0.61
OR135 refined	0.49 ± 0.10	0.11 ± 0.03	94.9	0.89	0.89
OR135 raw	1.83 ± 0.27	0.15 ± 0.04	58.2	2.81	1.13
OR36 refined	1.51 ± 0.37	0.45 ± 0.10	94.9	1.48	0.98
OR36 raw	5.58 ± 1.50	0.38 ± 0.07	35.9	4.28	1.03
StT322 refined	1.48 ± 0.49	0.19 ± 0.05	89.0	2.04	1.49
StT322 raw	4.50 ± 0.80	0.17 ± 0.05	41.6	8.92	6.73
YR313A refined	1.15 ± 0.20	0.46 ± 0.13	94.7	1.62	1.59
YR313A raw	4.38 ± 0.66	0.52 ± 0.25	35.7	6.37	1.64

Average backbone RMSDs to the mean were calculated by first superimposing all conformers in the ensemble onto the first one for calculating the mean coordinates. Then the RMSD between each individual conformer and the mean coordinates was computed and averaged. RMSDs to reference are the (single) RMSD values between the mean coordinates of the current and reference ensembles. All RMSDs were calculated for the well-defined regions comprising residues 13–104 for HR2876B, 12–160 for HR5460A, 12–99 for HR6430A, 12–55 for HR6470A, 5–75 for OR135, 3–125 for OR36, 23–63 for StT322, and 16–116 for YR313A

condition is fulfilled by all calculations with refined NOESY peak lists, and for 4 out of 8 calculations with raw NOESY peak lists (Table 3). The four calculations with raw data sets that do not converge in cycle 1 (HR5460A, OR36, StT322, and YR313A) are exactly those that result in erroneous final structures with more than 4 Å RMSD to the corresponding reference structure. In contrast to the RMSDs of cycle 1, the RMSDs to the mean coordinates for the final structures are small for all calculations (Table 3) and can therefore not serve as an indicator of the correctness of the structure.

The accuracy of the final structures obtained by the fully automated, exclusively NOESY based method is quantified by the RMSD values to the reference structure in Table 3, and the corresponding final and reference structures are superimposed in Figs. 3 and S3. Structure quality factors calculated with PSVS (Bhattacharya et al. 2007) are given in Tables S3-S18. The refined peak lists yielded structures with accuracies of 0.80–2.04 Å, whereas the accuracies obtained from the raw peak lists are 0.77–2.81 Å for the four calculations that converged in cycle 1, and 4.28–14.27 Å for those four that did not converge in cycle 1 (Table 3). With refined peak lists, it is thus possible to obtain almost complete assignments and correct structures for all eight proteins in this study. On the other hand, with raw peak lists, the method yields correct structures in four cases, which can be identified clearly on the basis of their cycle 1 RMSD. For comparison, Table 3 also shows the RMSD values to the reference structure that were obtained in calculations with the same NOESY peak lists but using as input the reference chemical shift assignments instead of those from FLYA. The results show structure accuracies of 0.37–1.64 Å except for two cases, namely the calculations with the raw peak lists for HR5460A (RMSD to reference 3.38 Å) and StT322 (6.73 Å), i.e. even with the almost complete reference resonance assignments as input, it is not possible to obtain a correct structure from the raw peak lists for these two proteins. On the other hand, the raw peak lists for OR36 and YR313A yield accurate structures only when using the reference chemical shift assignments as input.

With the two raw data sets that yield more than 90 % correct resonance assignments, i.e. HR6430A and HR6470A, high structural accuracies of 0.98 and 0.77 Å, respectively, were obtained (Table 3, Fig. S3a and 3c). Considering the total number of peaks in the lists (Table 1) and the percentage of peaks that were assigned during the NOE assignment (Table 3), the difference between raw and refined lists is smaller for these two proteins than for the remaining six proteins. For these, the assignment correctness was below 90 % where the raw lists contained up to 2.15 times the number of peaks in the refined list (Table 1), and 35.7–58.2 % of the peaks were assigned during the NOE assignment

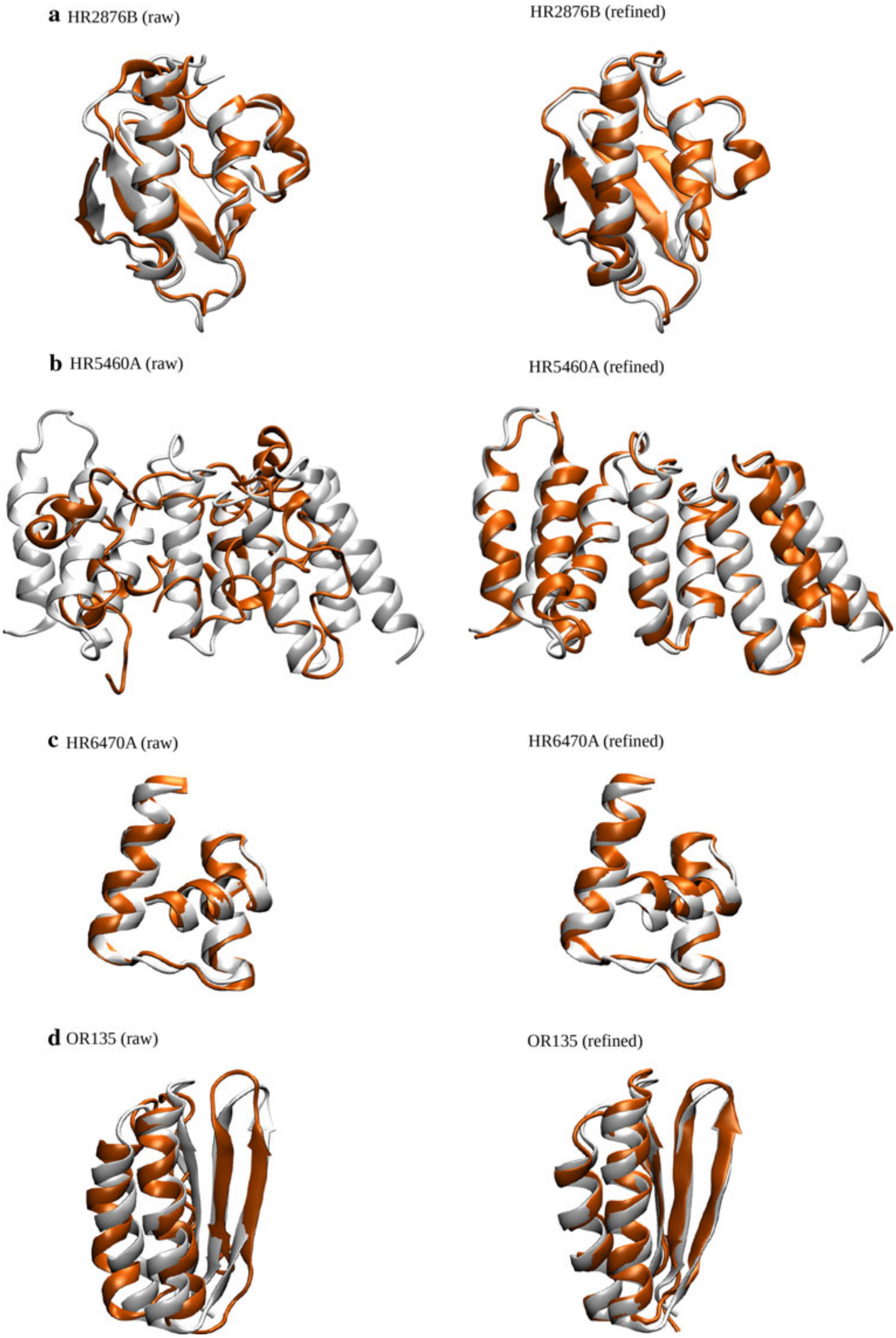

Fig. 3 Superposition of the structure obtained by structure calculation using automatically assigned chemical shifts (*orange*) and the reference structure (*white*) for **a** HR2876B, **b** HR5460A, **c** HR6470A, and **d** OR135. Only the well-defined regions comprising residues 13–104 for HR2876B, 12–160 for HR5460A, 12–99 for HR6430A, 12–55 for HR6470A, 5–75 for OR135, 3–125 for OR36, 23–63 for StT322, and 16–116 for YR313A are shown

compared to 89.0–94.9 % in the respective refined lists. Two calculations with resonance assignment correctness between 80 and 90 % led to RMSDs to reference of 1.56 and 2.81 Å. In the remaining four cases, i.e. the raw lists of HR5460A, OR36, StT322 and YR313A, that are based on resonance assignments with a correctness of 64.1, 68.8, 82.5 and 84.9 %, the correct global fold could not be obtained (Figs. 3 and S3). In these cases the quality of the assignments was not sufficient to achieve convergence in cycle 1 of the combined NOESY assignment and structure calculation. This is in agreement with the earlier finding that an assignment correctness of about 90 % is necessary for successful structure determinations with automated NOESY cross peak assignment (Herrmann et al. 2002a; Jee and Güntert 2003).

Success in the structure calculation is also strongly correlated with the extent of strong resonance assignments: For the four data sets for which the structure calculation was not successful, the percentage of strong assignments was between 30.4 and 75.2 %. On the other hand, the structure calculation was successful in all cases with more than 80 % strong assignments.

Conclusions

The results of this paper show that, in contrast to common belief based on experience with manual resonance assignment methods, NOESY spectra alone can be sufficient to determine the resonance assignments of proteins to a similarly high extent and correctness as the conventional method based on triple resonance experiments, provided that high-quality peak lists are prepared. The refined experimental NOESY peak lists for the eight proteins from the CASD-NMR project contained in general less than 5 % artifacts and 33–67 % of the expected short-range NOEs (Table 1), which was sufficient to obtain correct structures. Nevertheless, the results with the unfiltered "raw" data may be of most practical interest, as these are what is available before chemical shift and NOE assignment for an unknown structure. With raw peak lists that contained on average 40 % artifacts structure calculations did in some cases not converge. These cases can be identified readily (without comparison to reference chemical shift assignments and/or reference structures) based on the extent of strong assignments and the RMSD in the first cycle of the structure calculation. NOESY spectra reveal simultaneously short-range interactions, which are crucial for the resonance assignments, and medium- and long-range contacts, which are crucial for calculating the three-dimensional structure. The latter, however, increase also significantly the number of initial resonance assignment possibilities. The large number of potential assignment possibilities would in general overwhelm a spectroscopist who searches for the correct resonance assignments manually. For this reason, the NOESY-only approach is in general only feasible in conjunction with a powerful automated resonance assignment algorithm that can rapidly evaluate a large number of possible assignment patterns. The new FLYA resonance assignment algorithm fills this gap. It yields significantly better results than, for instance, the earlier GARANT algorithm (Bartels et al. 1996, 1997) with which the principal feasibility of NOESY-only resonance assignment and structure determination had been demonstrated for the first time (Ikeya et al. 2011).

The NOESY-only approach is "extreme" in the sense that it works completely without information from through-bond spectra. The opposite extreme is the use of an extensive set of about 15 through-bond and through-space spectra as input for FLYA (Schmidt and Güntert 2012), which corresponds to the traditional approach. Because of its generality, the FLYA algorithm can also be applied to any intermediate input. For example, it can be attractive to complement the NOESY spectra with 1–2 through-bond spectra in order to increase the reliability of the assignment without significantly increasing the total NMR measurement time. For instance, we repeated the calculations of this paper using in addition a CBCANH peak list as input. The results are given in Tables S1–S2. Overall, the correctness of the FLYA resonance assignments increased significantly by using the CBCANH data (Table S1 vs. Table 2). There was in general also a slight improvement of the structure accuracy although for three proteins the RMSD to the reference structure remained above 3 Å when using the raw peak lists as input (Table S2 vs. Table 3), which is presumably due to impact of the large imperfections of the raw NOESY peak lists on the CYANA algorithm for automated NOE distance restraint assignment.

Future improvements of the FLYA chemical shift assignment algorithm are conceivable. At present, the algorithm relies exclusively on peak positions. It ignores peak intensities or volumes, which are implicitly taken into consideration during manual assignment. Their use, especially in the case of NOESY spectra, could help the resonance assignment. Considering that even the refined NOESY peak lists do by far not contain all the peaks that are expected based on the sequence, a further improvement could be envisaged by enabling the FLYA algorithm to directly access the spectra (rather than only the peak lists), e.g. to check during the assignment calculation whether a certain expected peak is likely to be present at the predicted position in the spectrum, even if it has not (yet) been picked, or not (Herrmann et al. 2002b).

Acknowledgments We thank Profs. G. Montelione and C. Arrowsmith of the Northeast Structural Genomics (NESG) consortium and the CASD-NMR project for providing the experimental data. We gratefully acknowledge financial support by the Lichtenberg program of the Volkswagen Foundation and the Deutsche Forschungsgemeinschaft (DFG).

References

Aeschbacher T, Schmidt E, Blatter M, Maris C, Duss O, Allain FH-T, Güntert P, Schubert M (2013) Automated and assisted RNA resonance assignment using NMR chemical shift statistics. Nucleic Acids Res. doi:10.1093/nar/gkt665

Altieri AS, Byrd RA (2004) Automation of NMR structure determination of proteins. Curr Opin Struct Biol 14:547–553

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

Bartels C, Xia TH, Billeter M, Güntert P, Wüthrich K (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 6:1–10

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

Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. Proteins 66:778–795

Folmer RHA, Hilbers CW, Konings RNH, Nilges M (1997) Floating stereospecific assignment revisited: application to an 18 kDa protein and comparison with J-coupling data. J Biomol NMR 9:245–258

Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. Q Rev Biophys 44:257–309

Güntert P (1998) Structure calculation of biological macromolecules from NMR data. Q Rev Biophys 31:145–237

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

Güntert P, Braun W, Wüthrich K (1991) Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. J Mol Biol 217:517–530

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 (2002a) 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

Herrmann T, Güntert P, Wüthrich K (2002b) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. J Biomol NMR 24:171–189

Ikeya T, Jee J-G, Shigemitsu Y, Hamatsu J, Mishima M, Ito Y, Kainosho M, Güntert P (2011) Exclusively NOESY-based automated NMR assignment and structure determination of proteins. J Biomol NMR 50:137–146

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

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

Koga N, Tatsumi-Koga R, Liu GH, Xiao R, Acton TB, Montelione GT, Baker D (2012) Principles for designing ideal protein structures. Nature 491:222–227

Liu GH, Shen Y, Atreya HS, Parish D, Shao Y, Sukumaran DK, Xiao R, Yee A, Lemak A, Bhattacharya A, Acton TA, Arrowsmith CH, Montelione GT, Szyperski T (2005) NMR data collection and analysis protocol for high-throughput protein structure determination. Proc Natl Acad Sci USA 102:10487–10492

Liu GOH, Huang YJ, Xiao R, Wang DY, Acton TB, Montelione GT (2010) Solution NMR structure of the ARID domain of human AT-rich interactive domain-containing protein 3A: a human cancer protein interaction network target. Protein Struct Funct Bioinfo 78:2170–2175

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

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

Rosato A, Bagaria A, Baker D, Bardiaux B, Cavalli A, Doreleijers JF, Giachetti A, Guerry P, Güntert P, Herrmann T, Huang YJ, Jonker HRA, Mao B, Malliavin TE, Montelione GT, Nilges M, Raman S, van der Schot G, Vranken WF, Vuister GW, Bonvin AMJJ (2009) CASD-NMR: critical assessment of automated structure determination by NMR. Nat Methods 6:625–626

Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, Eletsky A, Giachetti A, Guerry P, Gutmanas A, Güntert P, He YF, Herrmann T, Huang YPJ, Jaravine V, Jonker HRA, Kennedy MA, Lange OF, Liu GH, Malliavin TE, Mani R, Mao BC, Montelione GT, Nilges M, Rossi P, van der Schot G, Schwalbe H, Szyperski TA, Vendruscolo M, Vernon R, Vranken WF, de Vries S, Vuister GW, Wu B, Yang YH, Bonvin AMJJ (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 20: 227–236

Schmidt E, Güntert P (2012) A new algorithm for reliable and general NMR resonance assignment. J Am Chem Soc 134:12817–12829

Schmidt E, Gath J, Habenstein B, Ravotti F, Székely K, Huber M, Buchner L, Böckmann A, Meier BH, Güntert P (2013) Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. J Biomol NMR 56:243–254

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao HY, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36:D402–D408

Williamson MP, Craven CJ (2009) Automated protein structure calculation from NMR data. J Biomol NMR 43:131–143