Archives of Biochemistry and Biophysics 628 (2017) 24–32

Contents lists available at ScienceDirect

Archives of Biochemistry and Biophysics

journal homepage: www.elsevier.com/locate/yabbi

NMR-based automated protein structure determination

Julia M. Würz [a], Sina Kazemi [a, b], Elena Schmidt [a], Anurag Bagaria [a], Peter Güntert [a, c, d, *]

[a] Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, Goethe University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany
[b] Frankfurt Institute for Advanced Studies, Goethe University Frankfurt am Main, Ruth-Moufang-Str. 1, 60438 Frankfurt am Main, Germany
[c] Laboratory of Physical Chemistry, ETH Zürich, 8093 Zürich, Switzerland
[d] Department of Chemistry, Graduate School of Science and Engineering, Tokyo Metropolitan University, Tokyo 192-0373, Japan

ARTICLE INFO

Article history:
Received 11 January 2017
Received in revised form
18 February 2017
Accepted 28 February 2017
Available online 2 March 2017

Keywords:
Peak picking
Resonance assignment
NOE assignment
Structure calculation
FLYA
CYANA

ABSTRACT

NMR spectra analysis for protein structure determination can now in many cases be performed by automated computational methods. This overview of the computational methods for NMR protein structure analysis presents recent automated methods for signal identification in multidimensional NMR spectra, sequence-specific resonance assignment, collection of conformational restraints, and structure calculation, as implemented in the CYANA software package. These algorithms are sufficiently reliable and integrated into one software package to enable the fully automated structure determination of proteins starting from NMR spectra without manual interventions or corrections at intermediate steps, with an accuracy of 1–2 Å backbone RMSD in comparison with manually solved reference structures.

1. Introduction

The standard procedure for protein structure determination by NMR [1–3] can be decomposed into two major parts. *Data acquisition* comprises sample preparation, usually with uniform or specific stable isotope labeling [4], the actual measurements at the NMR spectrometer [5], and data processing [6] to obtain a set of multidimensional NMR spectra [7]. *Data analysis* (Fig. 1, dark gray boxes) evaluates the acquired NMR spectra. It comprises the crucial steps of signal identification, chemical shift assignment, nuclear Overhauser effect (NOE) assignment, and structure calculation. The automation of data analysis will be discussed in this review.

Identifying signals in an NMR spectrum yields peak lists, and it is in this form that the information from the experimentally measured spectra enters the remaining steps of the procedure. In the next step, chemical shift assignment, the chemical shift values that are observed in the spectra are assigned to the corresponding protein atoms. This is followed by NOE assignment, where the cross peaks in NOESY spectra, which hold information about atom-atom distances in the 3D structure are assigned to the respective atoms based on the chemical shift assignment. Distance restraints are deduced from the volumes of these peaks. Finally, the 3D structure is calculated based on NOE distance restraints and possibly other conformational restraints, e.g. torsion angle restraints from chemical shifts or J-couplings, orientational restraints from residual dipolar couplings (RDCs), or hydrogen bond restraints. Once a preliminary 3D structure has been obtained, the structural information is used to improve the NOE assignment. This is done in several cycles. It is possible to refine the 3D structure using physical force fields, e.g. by molecular dynamics simulation in explicit solvent.

2. Automated signal identification

2.1. Peak picking principles

The identification of signals in an NMR spectrum, also known as peak picking, plays a central role in biomolecular NMR studies and is a prerequisite for sequence-specific resonance assignment and structure determination. Peak lists provide an abstraction of the multidimensional spectra that contains the most essential spectral

* Corresponding author. Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, Goethe University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany.
E-mail address: guentert@em.uni-frankfurt.de (P. Güntert).

http://dx.doi.org/10.1016/j.abb.2017.02.011
0003-9861/

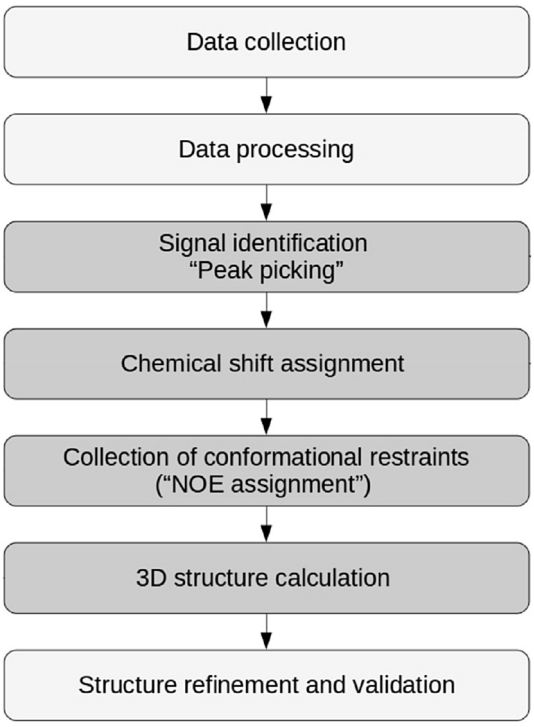

Fig. 1. Steps of a protein structure determination by NMR. Data analysis steps shown in dark gray are described in the following in more detail.

information—the position and intensity of the signals—in a form that is readily accessible by interactive or automated spectra analysis programs. The ease and reliability of spectrum assignment and the collection of conformational restraints relies on the quality of the peak lists, which in turn depends mainly on three factors: how many of the protein's real signals the peak lists contain, how few additional “artifact” peaks that do not correspond to true signals they contain, and how accurate they record the positions and intensities of the signals.

Peak lists do not have to be flawless to serve as a basis for chemical shift assignment and NOE assignment, followed by structure calculation. For instance, it has been shown [8] that the automated resonance assignment algorithm FLYA can yield more than 90% correct resonance assignments even if either 60% of the true peaks are missing or 5 times more artifacts than real peaks are present in the input peak lists. Automated NOE assignment and structure calculation with CYANA [9,10] can in many cases also tolerate 30–40% missing NOESY peaks without dramatic deterioration of the resulting structures [11,12].

Peak picking can be achieved by visual inspection of the spectra or automated methods. Along with algorithms for resonance assignment and structure calculation, the demand for automated peak picking is increasing, and various algorithms for the purpose have been proposed. Nevertheless, the task remains challenging. Reasons for this include low signal-to-noise ratios, peak overlap, and artifacts such as baseline distortions, intense solvent lines, ridges, or sinc wiggles.

2.2. Automated peak picking algorithms

Most of the existing peak picking algorithms can be classified as either threshold-based methods, methods that depend on symmetry criteria, peak-shape-based methods, methods that incorporate peak picking into NMR data processing, or a combination thereof. Threshold-based methods are the most straightforward and most commonly used automated peak picking approaches. Interactive spectrum analysis programs like XEASY [13], Sparky [14], NMRViewJ [15,16], or CcpNmr AnalysisAssign [17,18] (in the following abbreviated as CCPN) give the user the possibility to adjust a threshold manually and perform peak picking by finding local extrema above the threshold. These methods are particularly useful as a starting point for semi-automated peak identification, which is refined manually. WavPeak [19] employs wavelet-based smoothing of the spectrum prior to identifying peaks as local maxima. PICKY [20], is a singular value decomposition (SVD)-based automated peak picking method. Machine learning and computer vision methods have also been employed for peak picking, e.g. in the CV-Peak Picker program [21]. AUTOPSY [22] is a sophisticated automated peak picker that includes functions to determine a local noise level and to deconvolute clusters of overlapping peaks with the help of line shapes derived from non-overlapping peaks. ATNOS [23] is an automated peak picker specifically for NOESY spectra that is integrated into automated NOESY assignment and structure calculation and makes use of preliminary structural information to guide the peak picking. Peak picking can be part of NMR data processing, e.g. in the program MUNIN [24] that uses three-way decomposition to decompose a three-dimensional (3D) NMR spectrum into a sum of components defined as the direct product of three 1D shapes. The GAPRO peak identification algorithm [25] establishes peak lists for high-dimensional (e.g. 4D, 5D, 6D) APSY-type spectra by picking peaks in the experimentally recorded tilted 2D projections.

The human approach to peak picking can be characterized as the analysis of the shape and regularity of 2D contour lines. Real signals are manifested by concentric ellipses and have common properties which artifacts do not share, e.g. regarding peak width, convexity, or similarity. However, real signals can deviate from the proposed ideal shape for a number of reasons, such as, noise, spectral overlap, limited digital resolution, baseline instabilities, or phase distortions. An automated peak picking procedure should be able to handle these imperfections and shortcomings. A promising approach to automated peak picking is to mimic the human way of analyzing similarity and symmetry criteria of contour lines in 2D spectral planes. This approach has first been used in the CAPP algorithm [26].

2.3. The CYPICK algorithm

A recent example of an automated peak picking method that is based on analyzing geometric properties of contour lines is the CYPICK algorithm [27], which is implemented in the CYANA software package [28] and can be linked directly to automated chemical shift assignment and/or NOE assignment, followed by structure calculation, which are also available in CYANA. CYPICK follows, as far as possible, the manual approach taken by a spectroscopist who analyzes peak patterns in contour plots of the spectrum, but is fully automated. Human visual inspection is replaced by the evaluation of geometric criteria applied to contour lines, such as local extremality, approximate circularity (after appropriate scaling of the spectrum axes), and convexity. Fig. 2 shows a simplified flowchart of the CYPICK algorithm.

The first step is to read the processed NMR spectrum. Either a global noise level for the entire spectrum [29] or the local noise level at each data point is determined and used to set the intensity of the lowest (base) contour level. The global noise level is represented by a single number with obvious meaning that is straightforward to transfer to other algorithms. On the other hand, a noise level that is determined locally [22] permits the algorithm to better deal with noise bands, water lines, and similar artifacts, which

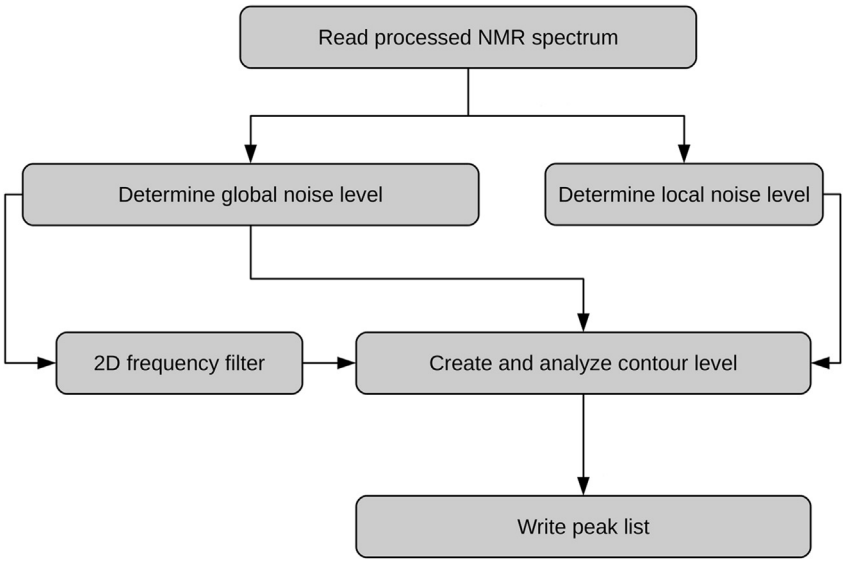

Fig. 2. Flowchart of the CYPICK peak picking algorithm implemented in CYANA.

automatically result in higher local noise levels and a reduced number of picked artifacts.

The next step is to find local extrema above the base level and to compute exponentially spaced contour lines in their vicinity. This can either be done over the entire spectrum or restricted by a frequency filter, defined by a 2D peak list, e.g. to pick peaks in a 3D spectrum, for instance a ^{15}N-resolved [^{1}H,^{1}H]-NOESY, based on a previously picked 2D spectrum, for instance a [^{1}H,^{15}N]-HSQC. The vertices of all contour lines that enclose the local extremum are stored. In order to achieve approximately circular contour lines for real peaks, the chemical shift coordinates of the points defining the contour lines are scaled according to the approximate line widths.

The contour lines belonging to local extrema are subsequently filtered and analyzed. A preliminary filtering process evaluates the following conditions: (i) The local extremum of interest has to be inside the contour line. (ii) No other local extremum except the local extremum of interest may be enclosed by the contour line. (iii) The contour line must have at least 5 vertices, because shape criteria (see below) cannot be evaluated meaningfully for contour lines with fewer points. (iv) At least two contour lines that fulfil all preceding conditions must enclose the local maximum.

After filtering, the remaining contour lines are further analyzed starting from the contour line with the highest absolute intensity. If the highest contour line does not fulfill the requirements, the next lower contour line is analyzed. At least two contour lines have to fulfill the two following conditions. The first condition is that its shape must be approximately circular. This is checked by computing the area-to-circumference-squared ratio, which equals $1/(4\pi)$ in case of a perfect circle. As a second condition, a contour line around an extremum is required to form an approximately convex polygon with all interior angles smaller than 180°. Nevertheless, for some of the real signals a slight deviation from perfect convexity should be tolerated.

The local extrema that fulfill these conditions correspond to the picked peaks. Their precise positions and intensities are determined by spline interpolation, and they are stored in a peak list.

The performance of CYPICK was evaluated for a variety of spectra from different proteins by systematic comparison with peak lists obtained by other, manual or automated, peak picking methods, as well as by analyzing the results of automated chemical shift assignment and structure calculation based on input peak lists from CYPICK [27]. The results show that CYPICK yielded peak lists that compare in most cases favorably to those obtained by other automated peak pickers with respect to the criteria of finding a maximal number of real signals, a minimal number of artifact peaks, and maximal correctness of the chemical shift assignments and the three-dimensional structure obtained by fully automated resonance assignment [8] and structure calculation [9,10] from the CYPICK peak lists.

3. Automated chemical shift assignment

3.1. Resonance assignment principles

Every NMR-detected nucleus in a macromolecule has a specific chemical shift value, which depends on its chemical environment. Revealing the relationship between atoms and chemical shifts is denoted as chemical shift or resonance assignment. Chemical shift assignment is not only necessary to exploit the distance information in NOESY spectra for structure determination, but in all cases in which atom-specific information has to be obtained from an NMR experiment. Examples include molecular interaction studies, alternative approaches for protein structure determination that are based on chemical shifts or RDCs, or investigations of protein dynamics.

To enable chemical shift assignment, several NMR experiments are performed that complement each other such that the connectivity of the atoms in a protein is represented. Based on the covalent structure that results from the protein sequence, it is possible to establish the relationship between chemical shifts and atoms. Usually, a set of standard experiments [7] is used to reveal the covalent atom connectivities (Fig. 3). The different spectra should be aligned as closely as possible; referencing offsets can be corrected automatically [30].

Since the general strategy for chemical shift assignment has been described in the 1980s [31], there have been many attempts to establish an automated procedure for this process, and reviews of these endeavours are available [32–34]. Some programs [8,35–38] perform the entire chemical shift assignment process starting from peak lists or NMR spectra as input data and ending with an (almost) complete assignment of backbone and side chain atoms, others are specialized in certain aspects of the assignment process, for

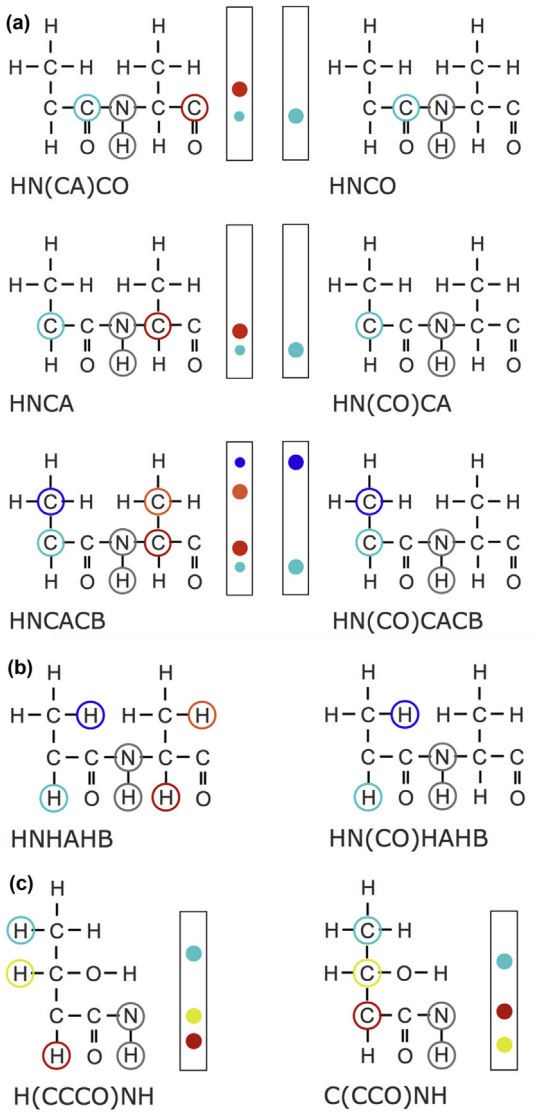

Fig. 3. Standard 3D triple resonance experiments for protein resonance assignment. (a) Experiments for the assignment of the backbone atoms N, H, C^{α}, C^{β}, and C'. (b) Experiments for the assignment of H^{α} and H^{β} nuclei. (c) Experiments for the assignment of side-chain ^{13}C and ^{1}H nuclei. Atoms that lead to peaks in the respective experiments are encircled. Peaks in the "strips", 2D rectangular regions taken from the 3D spectra at the backbone N/H positions, and the corresponding atoms are marked in the same color. Inter-residue peaks are colored in blue. Intra-residue peaks are colored in red. Backbone N and H resonances contribute to all peaks and are marked in gray. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

instance sequence-specific backbone assignment [39–42].

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 [35,36].

3.2. The FLYA algorithm

The FLYA resonance assignment algorithm [8] that has been implemented in the CYANA software package [28,43], 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 [44] or corresponding solid-state NMR experiments [45]. A scheme of the algorithm is shown in Fig. 4. FLYA 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, which is typically not the case in a structure determination. Therefore, in general only short-range expected NOESY peaks are generated. To this end, 20 random structures 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.

The mapping of expected peaks to measured peaks is done using an evolutionary optimization algorithm that works with a population of individuals, each representing an assignment solution [8]. The evolutionary optimization is complemented by local optimization that is applied to the individuals of each generation. 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 data bank [46], or user defined [47]), 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 iteration a local optimization algorithm takes small parts of a mapping back and reassigns the expected peaks for a predefined number of iterations (by default 15,000). The different solutions of one generation are then recombined into a new generation. The individuals and the specific parts of an individual that contribute to a new individual are selected based on a scoring function. The scoring function takes into account four conditions that should be fulfilled by correct assignments: 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 the 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 (typically 20) independent runs of the algorithm are performed with different random seeds. 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 [8,48,49] that assignments with high self-consistency ("strong" assignments) are more reliable than others ("weak" assignments).

4. Automated NOESY assignment and structure calculation

4.1. Automated NOE assignment principles

The structure determination of biological macromolecules by NMR in solution relies primarily on distance restraints derived from cross peaks in NOESY spectra. A large number of assigned NOESY cross peaks are necessary to compute an accurate 3D structure because many of the NOEs are short-range with respect to the sequence and thus carry little information about the tertiary

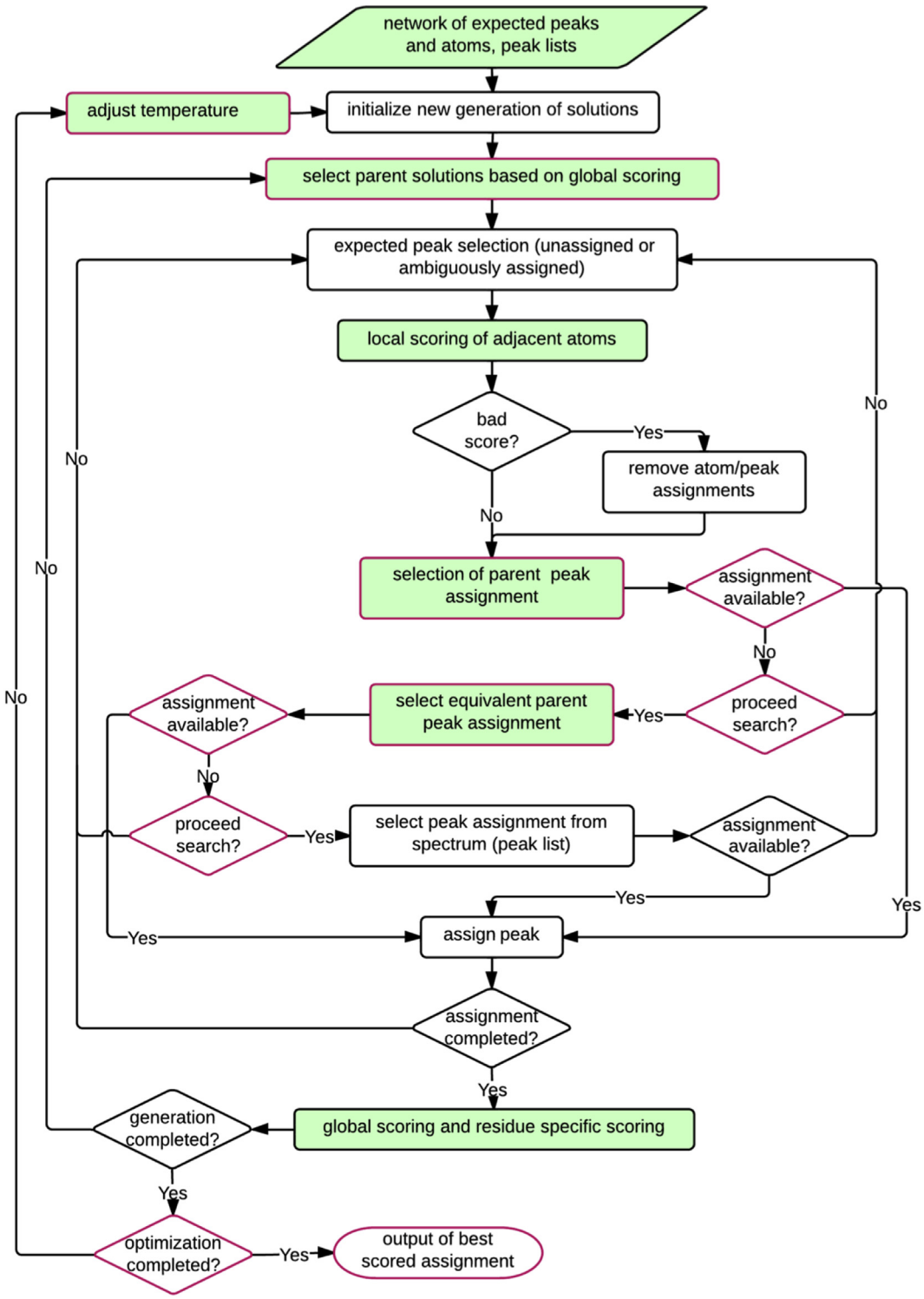

Fig. 4. Flowchart of the FLYA automated assignment algorithm. Evolutionary optimization specific steps, which are omitted in the initialization process, are in magenta. Steps for which new strategies were implemented in FLYA automated assignment are shown in green [8]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

structure and because NOEs are generally interpreted as loose upper bounds in order to implicitly account for internal motions and spin diffusion. Alternatively, accurate distance measurements have become available with eNOEs [50]. Obtaining a comprehensive set of distance restraints from NOESY spectra is in practice not straightforward. The large amount of data, as well as resonance and peak overlap, spectral artifacts and noise, and the absence of expected signals because of fast relaxation turn interactive NOESY cross peak assignment into a laborious and error-prone task, even if it is supported by semi-automated tools that propose and check assignment possibilities [51–53]. Therefore, the development of computer algorithms for automating this often most time-

consuming step of a protein structure determination by NMR has been pursued intensely [33]. Several algorithms have been developed for the automated analysis of NOESY spectra given the chemical shift assignments, e.g. NOAH [54,55], ARIA [56,57], ASDP [58], CANDID [10], PASD [59,60] and AutoNOE-Rosetta [61]. Automated NOESY peak picking guided by intermediate structures has also been integrated into ATNOS-CANDID method [23].

The basic problem of NOESY assignment is the ambiguity of cross peak assignments if only the match between cross peak positions and the chemical shift values of candidate resonances is considered. It has been shown that the number of assignment possibilities based on chemical shift matching increases exponentially with the uncertainty in the peak and resonance positions. As a consequence, there are in general not a sufficient number of unambiguously assigned distance restraints to obtain a structure [55]. Ambiguous distance restraints make it possible to use also NOEs with multiple assignment possibilities in a structure calculation [62]. Nevertheless, to minimize the information loss, additional criteria have to be applied to resolve these ambiguities as far as possible, such as using secondary structure information [58] or a preliminary structure that is refined iteratively in cycles of NOE assignment and structure calculation [54]. The CANDID automated NOESY assignment method [10] introduced the concepts of network anchoring to reduce the initial ambiguity of NOE assignments and constraint combination to reduce the impact of erroneous restraints.

4.2. Combined automated NOE assignment and structure calculation with CYANA

The algorithm for automated NOE assignment in CYANA [9] is a re-implementation of principles of the former CANDID procedure [10] on the basis of a probabilistic treatment of the NOE assignment process that is conceptually more consistent and better capable to handle situations of high chemical shift-based ambiguity of the NOE assignments. The key features of the algorithm are network anchoring to reduce the initial ambiguity of NOESY peak assignments, ambiguous distance restraints to generate conformational restraints from NOESY cross peaks with multiple possible assignments, and constraint combination to minimize the impact of erroneous distance restraints on the structure. Automated NOE assignment and the structure calculation are combined in an iterative process that comprises, typically, seven cycles of automated NOE assignment and structure calculation, followed by a final structure calculation using only unambiguously assigned distance restraints. Between subsequent cycles, information is transferred exclusively through the intermediary 3D structures. The molecular structure obtained in a given cycle is used to guide the NOE assignments in the following cycle. Otherwise, the same input data are used for all cycles, that is the amino acid sequence of the protein, one or several chemical shift lists from the sequence-specific resonance assignment, and one or several lists containing the positions and volumes of cross peaks in 2D, 3D, or 4D NOESY spectra. The input may further include previously assigned NOE upper distance bounds or other previously assigned conformational restraints for the structure calculation.

4.2.1. Assignment conditions and network anchoring

In each cycle, first all assignment possibilities of a peak are generated on the basis of the chemical shift values that match the peak position within given tolerance values, and the quality of the fit between the atomic chemical shifts and the peak position is expressed by a Gaussian probability, P_{shifts}. Second, the probability $P_{\text{structure}}$ for agreement with the preliminary structure from the preceding cycle (if available) is computed. Third, each assignment possibility is evaluated for its network anchoring, i.e., its embedding in the network formed by the assignment possibilities of all the other peaks and the covalently restrained short-range distances. The network anchoring probability P_{network} that the distance corresponding to an assignment possibility is shorter than the upper distance bound plus the acceptable violation is computed given the assignments of the other peaks but independent from knowledge of the three-dimensional structure. Only assignment possibilities for which the product of the three probabilities is above a threshold, $P_{\text{tot}} = P_{\text{shifts}} P_{\text{network}} P_{\text{structure}} \geq P_{\text{min}}$, are accepted (Fig. 5). Cross peaks with a single accepted assignment yield a conventional unambiguous distance restraint. Cross peaks with multiple accepted assignments result in an ambiguous distance restraint.

4.2.2. Ambiguous distance restraints

Ambiguous distance restraints [62] provide a powerful concept for handling ambiguities in NOESY cross peak assignments. When using ambiguous distance restraints, every NOESY cross peak is treated as the superposition of the signals from each of its possible assignments by applying relative weights proportional to the inverse sixth power of the corresponding interatomic distances. A NOESY cross peak with a unique assignment possibility gives rise to an upper bound b on the distance d (α,β) between two hydrogen atoms, α and β. A NOESY cross peak with $n > 1$ assignment possibilities can be interpreted as the superposition of n degenerate signals and interpreted as an ambiguous distance restraint, $d_{\text{eff}} \leq b$, with the "effective" or "r^{-6}-summed" distance

$$d_{\text{eff}} = \left(\sum_{k=1}^{n} d_k^{-6} \right)^{-1/6}$$

Each of the distances $d_k = d\,(\alpha_k,\beta_k)$ in the sum corresponds to one assignment possibility to a pair of hydrogen atoms, α_k and β_k. The effective distance d_{eff} is always shorter than any of the individual distances d_k. Thus, an ambiguous distance restraint will be fulfilled by the correct structure provided that the correct assignment is included among its assignment possibilities, regardless of the possible presence of other, incorrect assignment possibilities. Ambiguous distance restraints make it possible to interpret NOESY cross peaks as correct conformational restraints also if a unique assignment cannot be determined at the outset of a structure determination. Including multiple assignment possibilities, some but not all of which may later turn out to be incorrect, does not result in a distorted structure but only in a decrease of the information content of the ambiguous distance restraints.

4.2.3. Constraint combination

Spurious distance restraints may arise from the misinterpretation of noise and spectral artifacts, in particular at the outset of a structure determination before 3D structure-based filtering of the restraint assignments can be applied. CYANA uses "constraint combination" [9,10] to reduce structural distortions from erroneous distance restraints. Medium-range and long-range distance restraints are incorporated into "combined distance restraints", which are a generalization of ambiguous distance restraints with assignments taken from different, in general unrelated, cross peaks (Fig. 6). A basic property of ambiguous distance restraints is that the restraint will be fulfilled by the correct structure whenever at least one of its assignments is correct, regardless of the presence of additional, erroneous assignments. This implies that such combined restraints have a lower probability of being erroneous than the corresponding original restraints, provided that the fraction of erroneous original restraints is smaller than 50%. Constraint

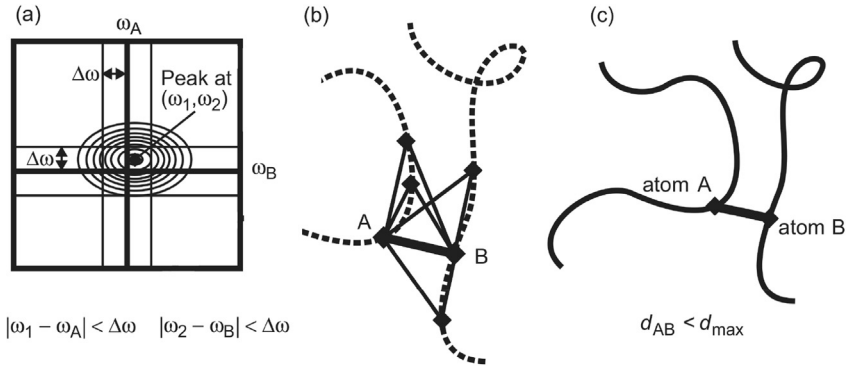

Fig. 5. Three conditions that must be fulfilled by a valid assignment of a NOESY cross peak to two protons A and B in the CYANA automated NOESY assignment algorithm. (a) Agreement between the proton chemical shifts ω_A and ω_B and the peak position (ω_1,ω_2) within a tolerance of $\Delta\omega$. (b) Network anchoring. The NOE between protons A and B must be part of a network of other NOEs or covalently restricted distances that connect the protons A and B indirectly through other protons. (c) Spatial proximity in a (preliminary) structure.

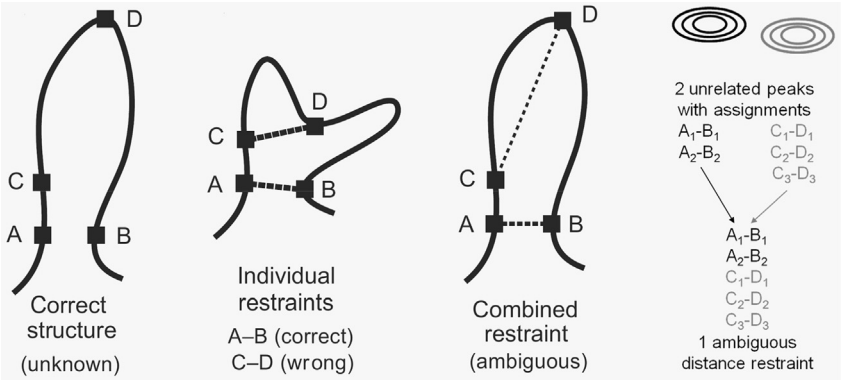

Fig. 6. Effect of constraint combination in the case of two distance restraints. A correct restraint connects atoms A and B, and a wrong atoms C and D. A structure calculation that uses these two restraints as individual restraints that have to be satisfied simultaneously will, instead of finding the correct structure (shown, schematically, in the first panel), result in a distorted conformation (second panel), whereas a combined restraint that will be fulfilled already if one of the two distances is sufficiently short leads to an almost undistorted solution (third panel). The formation of a combined restraint from the assignments of two peaks is shown in the right panel.

combination aims at minimizing the impact of erroneous NOE assignments on the resulting structure at the expense of a temporary loss of information. It is applied to medium- and long-range distance restraints, by default only in the first two cycles of combined automated NOE assignment and structure calculation with CYANA.

4.2.4. Structure calculation

The distance restraints are then included in the input for the structure calculation with simulated annealing by the fast CYANA torsion angle dynamics algorithm [28]. The structure calculations typically comprise seven cycles. The second and subsequent cycles differ from the first cycle by the use of additional selection criteria for cross peaks and NOE assignments that are based on assessments relative to the protein 3D structure from the preceding cycle. The precision of the structure determination normally improves with each subsequent cycle. Accordingly, the cutoff for acceptable distance restraint violations in the calculation of $P_{structure}$ is tightened from cycle to cycle. In the final structure calculation, an additional filtering step ensures that all NOEs have either unique assignments to a single pair of hydrogen atoms, or are eliminated from the input for the structure calculation. This facilitates the use of subsequent refinement and analysis programs that cannot handle ambiguous distance restraints.

5. Conclusions

The abovementioned computational tools are sufficiently reliable and integrated into one software package to enable the fully automated structure determination of proteins starting from NMR spectra without manual interventions or corrections at intermediate steps. It has been shown that the fully automated method can yield 3D structures of proteins with an accuracy of 1–2 Å backbone RMSD in comparison with manually solved reference structures of proteins [8,27,49].

For instance, the automated pipeline of CYPICK peak picking, FLYA chemical shift assignment, NOESY assignment and structure calculation with CYANA was applied to the 140-residue protein ENTH [27]. 16 through-bond and through-space spectra were available [49,63], and results could be compared with those obtained by manual peak picking and assignment. Over all peak lists, 75% of the manually picked peaks were identified also by CYPICK, whose peak lists contained 29% additional peaks. On the basis of the CYPICK peak lists, the chemical shift assignments by FLYA were correct for 95.4% of the backbone atoms and 89.4% of all atoms which had been assigned manually. Combined automated NOESY assignment and structure calculation based on the CYPICK peak lists and the FLYA chemical shift assignments yielded a structure bundle with 0.5 Å RMSD to the mean coordinates and 0.9 Å RMSD from the well-defined regions of the manually determined

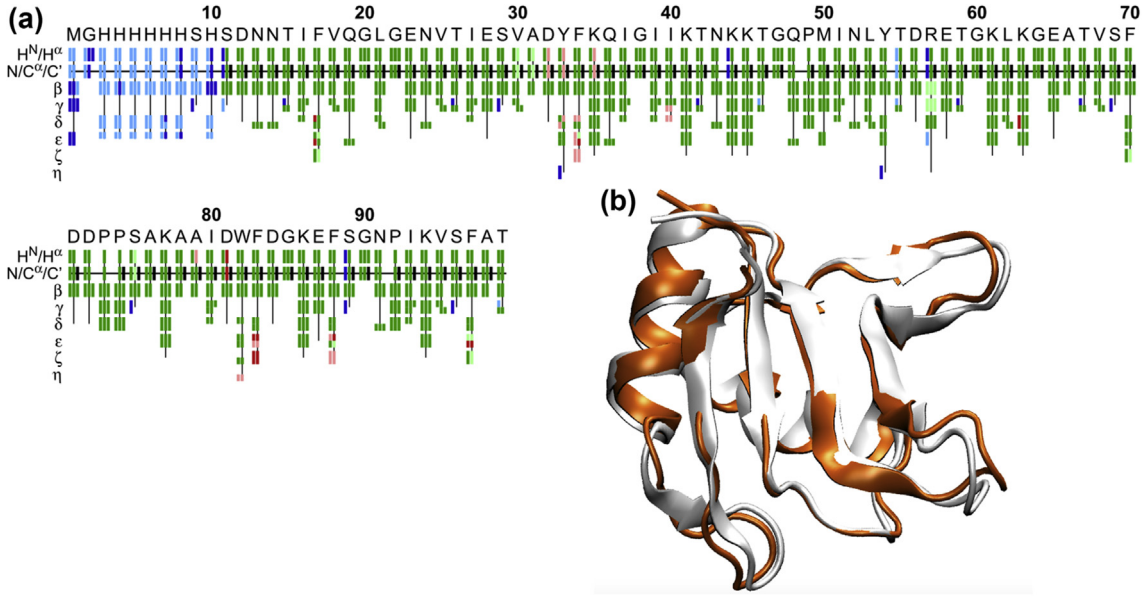

Fig. 7. Fully automated structure determination of the RRM domain of the RNA-binding protein FUS (PDB code 2LA6) using exclusively NOESY spectra as experimental input [44]. (a) Automated resonance assignment. Each assignment for an atom is represented by a rectangle; colored green, if the assignment by FLYA agrees with the manually determined reference chemical shifts within a tolerance of 0.03 ppm for ^{1}H and 0.3 ppm for ^{13}C/^{15}N; red, if the assignment differs from reference; blue, if assigned by FLYA but no reference available; black, if with reference assignment but not assigned by FLYA. Strong and weak colors represent "strong" (self-consistent) and "weak" (tentative) assignments as classified by chemical shift consolidation from multiple runs of the assignment algorithm. The row labeled H^N/H^α shows for each residue H^N on the left and H^α in the center. The $N/C^\alpha/C'$ row shows for each residue the N, C^α, 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. For branched side chains, the corresponding row is split into an upper part for one branch and a lower part for the other branch. (b) Structure determined by combined automated NOE assignment and structure calculation with CYANA, using the automatically assigned chemical shifts as input (orange). For comparison, the automatically determined structure is superimposed on the conventionally determined reference structure (white). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

reference structure.

In favorable cases, 3D protein structures can even be determined from NOESY spectra alone, without measuring any "through-bond" spectra for the resonance assignment (Fig. 7) [44].

5.1. Critical assessment of structure determination by NMR (CASD-NMR)

Recently, the critical assessment of structure determination by NMR (CASD-NMR) initiative [64] has evaluated several NMR structure determination methods by blind testing. Using high-quality data sets of small proteins from a structural genomics project, it was found that the NOESY-based methods included in the test yielded structures with an accuracy of 2 Å RMSD or better to the subsequently released reference structures [65,66].

5.2. Possible pitfalls of fully automated structure determination

Because experimental NMR spectra are never perfect, fully automated structure determination must be capable to cope with incomplete and partially erroneous input data. For instance, the algorithms should discard artifact peaks when making assignments or generating NOE distance restraints for the structure calculation. Under such circumstances, there is a potential danger that erroneous structures are generated. In principle, automated structure determination approaches can go wrong in two ways, especially with low-quality input data. Either the algorithm fails to ever assign enough NOE distance restraints to obtain a defined structure. This outcome, manifested by a divergent structure bundle with a high RMSD, is unfortunate but straightforward to detect. More problematic are failures of a second kind, where the algorithm, possibly gradually over several cycles, discards part of the NOE cross peaks and selects a self-consistent but incomplete subset of the data to compute a well-defined but erroneous structure, i.e. a tight bundle of conformers with low RMSD to its mean coordinates that, however, differs significantly from the (unknown) correct structure of the protein. If this outcome goes unnoticed, it may result in the publication or PDB deposition of erroneous structures that cannot be detected easily by common coordinate-based validation tools [67].

5.3. Realistic accuracy estimates by consensus structure bundles

To exclude such problems, the method of consensus structure bundles has been developed [68]. For this approach, one first performs 20 independent runs of combined automated NOESY assignment and structure calculation with CYANA using the same input data but different random start structures. Each run yields a structure bundle as well as the corresponding set of distance restraints. Because the NOESY peaks are assigned independently in each of the 20 runs, the sets of distance restraints from each run in general differ from each other. One now combines the individual sets of distance restraints in order to obtain a consensus set of distance restraints including assignments from all individual runs, which is then used to recalculate the final protein structure bundle, the *consensus bundle*. This new protocol for NMR structure determination produces, like the traditional method, bundles of conformers in agreement with a common set of conformational restraints, however with a realistic precision that has been shown, throughout a variety of proteins and NMR data sets, to be a much better estimate of structural accuracy than the precision of conventional structure bundles [68].

References

[1] K. Wüthrich, NMR of Proteins and Nucleic Acids, Wiley, New York, 1986.
[2] M. Billeter, G. Wagner, K. Wüthrich, J. Biomol. NMR 42 (2008) 155–158.
[3] P. Güntert, Prog. Nucl. Magn. Reson. Spectrosc. 43 (2003) 105–125.
[4] M. Kainosho, P. Güntert, Q. Rev. Biophys. 42 (2009) 247–300.
[5] J. Keeler, Understanding NMR Spectroscopy, Wiley, Chichester, UK, 2010.
[6] J.C. Hoch, A.S. Stern, NMR Data Processing, Wiley, New York, 1996.
[7] J. Cavanagh, W.J. Fairbrother, A.G. Palmer III, N.J. Skelton, M. Rance, Protein NMR Spectroscopy. Principles and Practice, Academic Press, San Diego, CA, 2007.
[8] E. Schmidt, P. Güntert, J. Am. Chem. Soc. 134 (2012) 12817–12829.
[9] P. Güntert, L. Buchner, J. Biomol. NMR 62 (2015) 453–471.
[10] T. Herrmann, P. Güntert, K. Wüthrich, J. Mol. Biol. 319 (2002) 209–227.
[11] J. Jee, P. Güntert, J. Struct. Funct. Genom 4 (2003) 179–189.
[12] L. Buchner, P. Güntert, J. Biomol. NMR 62 (2015) 81–95.
[13] C. Bartels, T.H. Xia, M. Billeter, P. Güntert, K. Wüthrich, J. Biomol. NMR 6 (1995) 1–10.
[14] T.D. Goddard, D.G. Kneller, Sparky 3, University of California, San Francisco, 2001.
[15] B.A. Johnson, Meth. Mol. Biol. 278 (2004) 313–352.
[16] B.A. Johnson, R.A. Blevins, J. Biomol. NMR 4 (1994) 603–614.
[17] W.F. Vranken, W. Boucher, T.J. Stevens, R.H. Fogh, A. Pajon, M. Llinás, E.L. Ulrich, J.L. Markley, J. Ionides, E.D. Laue, Proteins 59 (2005) 687–696.
[18] S.P. Skinner, R.H. Fogh, W. Boucher, T.J. Ragan, L.G. Mureddu, G.W. Vuister, J. Biomol. NMR 66 (2016) 111–124.
[19] Z. Liu, A. Abbas, B.Y. Jing, X. Gao, Bioinformatics 28 (2012) 914–920.
[20] B. Alipanahi, X. Gao, E. Karakoc, L. Donaldson, M. Li, Bioinformatics 25 (2009) i268–i275.
[21] P. Klukowski, M.J. Walczak, A. Gonczarek, J. Boudet, G. Wider, Bioinformatics 31 (2015) 2981–2988.
[22] R. Koradi, M. Billeter, M. Engeli, P. Güntert, K. Wüthrich, J. Magn, Reson 135 (1998) 288–297.
[23] T. Herrmann, P. Güntert, K. Wüthrich, J. Biomol. NMR 24 (2002) 171–189.
[24] V.Y. Orekhov, I.V. Ibraghimov, M. Billeter, J. Biomol. NMR 20 (2001) 49–60.
[25] S. Hiller, F. Fiorito, K. Wüthrich, G. Wider, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 10876–10881.
[26] D.S. Garrett, R. Powers, A.M. Gronenborn, G.M. Clore, J. Magn. Reson 95 (1991) 214–220.
[27] J.M. Würz, P. Güntert, J. Biomol. NMR 67 (2017) 63–76.
[28] P. Güntert, C. Mumenthaler, K. Wüthrich, J. Mol. Biol. 273 (1997) 283–298.
[29] P. Güntert, V. Dötsch, G. Wider, K. Wüthrich, J. Biomol. NMR 2 (1992) 619–629.
[30] L. Buchner, E. Schmidt, P. Güntert, J. Biomol. NMR 55 (2013) 267–277.
[31] K. Wüthrich, G. Wider, G. Wagner, W. Braun, J. Mol. Biol. 155 (1982) 311–319.
[32] W. Gronwald, H.R. Kalbitzer, Prog. Nucl. Magn. Reson. Spectrosc. 44 (2004) 33–96.
[33] P. Guerry, T. Herrmann, Q. Rev. Biophys. 44 (2011) 257–309.
[34] H.N.B. Moseley, G.T. Montelione, Curr. Opin. Struct. Biol. 9 (1999) 635–642.
[35] C. Bartels, P. Güntert, M. Billeter, K. Wüthrich, J. comput. Chem. 18 (1997) 139–149.
[36] C. Bartels, M. Billeter, P. Güntert, K. Wüthrich, J. Biomol. NMR 7 (1996) 207–213.
[37] A. Bahrami, A.H. Assadi, J.L. Markley, H.R. Eghbalnia, PLoS Comp. Biol. 5 (2009) e1000307.
[38] R. Schmucki, S. Yokoyama, P. Güntert, J. Biomol. NMR 43 (2009) 97–109.
[39] D.E. Zimmerman, C.A. Kulikowski, Y.P. Huang, W.Q. Feng, M. Tashiro, S. Shimotakahara, C.Y. Chien, R. Powers, G.T. Montelione, J. Mol. Biol. 269 (1997) 592–610.
[40] P. Güntert, M. Salzmann, D. Braun, K. Wüthrich, J. Biomol. NMR 18 (2000) 129–137.
[41] J. Volk, T. Herrmann, K. Wüthrich, J. Biomol. NMR 41 (2008) 127–138.
[42] Y.S. Jung, M. Zweckstetter, J. Biomol. NMR 30 (2004) 11–23.
[43] P. Güntert, Eur. Biophys. J. 38 (2009) 129–143.
[44] E. Schmidt, P. Güntert, J. Biomol. NMR 57 (2013) 193–204.
[45] E. Schmidt, J. Gath, B. Habenstein, F. Ravotti, K. Székely, M. Huber, L. Buchner, A. Böckmann, B.H. Meier, P. Güntert, J. Biomol. NMR 56 (2013) 243–254.
[46] E.L. Ulrich, H. Akutsu, J.F. Doreleijers, Y. Harano, Y.E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C.F. Schulte, D.E. Tolmie, R.K. Wenger, H.Y. Yao, J.L. Markley, Nucleic Acids Res. 36 (2008) D402–D408.
[47] T. Aeschbacher, E. Schmidt, M. Blatter, C. Maris, O. Duss, F.H.-T. Allain, P. Güntert, M. Schubert, Nucleic Acids Res. 41 (2013) e172.
[48] D. Malmodin, C.H.M. Papavoine, M. Billeter, J. Biomol. NMR 27 (2003) 69–79.
[49] B. López-Méndez, P. Güntert, J. Am. Chem. Soc. 128 (2006) 13112–13122.
[50] B. Vögeli, S. Kazemi, P. Güntert, R. Riek, Nat. Struct. Mol. Biol. 19 (2012) 1053–1057.
[51] P. Güntert, K.D. Berndt, K. Wüthrich, J. Biomol. NMR 3 (1993) 601–606.
[52] S.P. Skinner, B.T. Goult, R.H. Fogh, W. Boucher, T.J. Stevens, E.D. Laue, G.W. Vuister, Acta Crystallogr. D. 71 (2015) 154–161.
[53] N. Kobayashi, J. Iwahara, S. Koshiba, T. Tomizawa, N. Tochio, P. Güntert, T. Kigawa, S. Yokoyama, J. Biomol. NMR 39 (2007) 31–52.
[54] C. Mumenthaler, W. Braun, J. Mol. Biol. 254 (1995) 465–480.
[55] C. Mumenthaler, P. Güntert, W. Braun, K. Wüthrich, J. Biomol. NMR 10 (1997) 351–362.
[56] M. Nilges, M.J. Macias, S.I. ODonoghue, H. Oschkinat, J. Mol. Biol. 269 (1997) 408–422.
[57] W. Rieping, M. Habeck, B. Bardiaux, A. Bernard, T.E. Malliavin, M. Nilges, Bioinformatics 23 (2007) 381–382.
[58] Y.J. Huang, R. Tejero, R. Powers, G.T. Montelione, Proteins 62 (2006) 587–603.
[59] J.J. Kuszewski, R.A. Thottungal, G.M. Clore, C.D. Schwieters, J. Biomol. NMR 41 (2008) 221–239.
[60] J. Kuszewski, C.D. Schwieters, D.S. Garrett, R.A. Byrd, N. Tjandra, G.M. Clore, J. Am. Chem. Soc. 126 (2004) 6258–6273.
[61] Z. Zhang, J. Porter, K. Tripsianes, O.F. Lange, J. Biomol. NMR 59 (2014) 135–145.
[62] M. Nilges, J. Mol. Biol. 245 (1995) 645–660.
[63] B. López-Méndez, D. Pantoja-Uceda, T. Tomizawa, S. Koshiba, T. Kigawa, M. Shirouzu, T. Terada, M. Inoue, T. Yabuki, M. Aoki, E. Seki, T. Matsuda, H. Hirota, M. Yoshida, A. Tanaka, T. Osanai, M. Seki, K. Shinozaki, S. Yokoyama, P. Güntert, J. Biomol. NMR 29 (2004) 205–206.
[64] A. Rosato, A. Bagaria, D. Baker, B. Bardiaux, A. Cavalli, J.F. Doreleijers, A. Giachetti, P. Guerry, P. Güntert, T. Herrmann, Y.J. Huang, H.R.A. Jonker, B. Mao, T.E. Malliavin, G.T. Montelione, M. Nilges, S. Raman, G. van der Schot, W.F. Vranken, G.W. Vuister, A.M.J.J. Bonvin, Nat. Methods 6 (2009) 625–626.
[65] A. Rosato, J.M. Aramini, C. Arrowsmith, A. Bagaria, D. Baker, A. Cavalli, J.F. Doreleijers, A. Eletsky, A. Giachetti, P. Guerry, A. Gutmanas, P. Güntert, Y.F. He, T. Herrmann, Y.P.J. Huang, V. Jaravine, H.R.A. Jonker, M.A. Kennedy, O.F. Lange, G.H. Liu, T.E. Malliavin, R. Mani, B.C. Mao, G.T. Montelione, M. Nilges, P. Rossi, G. van der Schot, H. Schwalbe, T.A. Szyperski, M. Vendruscolo, R. Vernon, W.F. Vranken, S. de Vries, G.W. Vuister, B. Wu, Y.H. Yang, A.M.J.J. Bonvin, Structure 20 (2012) 227–236.
[66] A. Rosato, W. Vranken, R.H. Fogh, T.J. Ragan, R. Tejero, K. Pederson, H.W. Lee, J.H. Prestegard, A. Yee, B. Wu, A. Lemak, S. Houliston, C.H. Arrowsmith, M. Kennedy, T.B. Acton, R. Xiao, G.H. Liu, G.T. Montelione, G.W. Vuister, J. Biomol. NMR 62 (2015) 413–424.
[67] S.B. Nabuurs, C.A.E.M. Spronk, G.W. Vuister, G. Vriend, PLoS Comp. Biol. 2 (2006) 71–79.
[68] L. Buchner, P. Güntert, Structure 23 (2015) 425–434.