


Automated assignment of NMR chemical shifts based on a known structure and 4D spectra

Matthias Trautwein¹ · Kai Fredriksson¹ · Heiko M. Möller² · Thomas E. Exner¹ 

Received: 8 March 2016 / Accepted: 28 July 2016 / Published online: 2 August 2016
© Springer Science+Business Media Dordrecht 2016

Abstract Apart from their central role during 3D structure determination of proteins the backbone chemical shift assignment is the basis for a number of applications, like chemical shift perturbation mapping and studies on the dynamics of proteins. This assignment is not a trivial task even if a 3D protein structure is known and needs almost as much effort as the assignment for structure prediction if performed manually. We present here a new algorithm based solely on 4D [¹H,¹⁵N]-HSQC-NOESY-[¹H,¹⁵N]-HSQC spectra which is able to assign a large percentage of chemical shifts (73–82 %) unambiguously, demonstrated with proteins up to a size of 250 residues. For the remaining residues, a small number of possible assignments is filtered out. This is done by comparing distances in the 3D structure to restraints obtained from the peak volumes in the 4D spectrum. Using dead-end elimination, assignments are removed in which at least one of the restraints is violated. Including additional information from chemical shift predictions, a complete unambiguous assignment was obtained for Ubiquitin and 95 % of the residues were correctly assigned in the 251 residue-long N-terminal domain of enzyme I. The program including source code is available at <https://github.com/thomasexner/4Dassign>.

Keywords Chemical shift assignment · Protein · 3D structure · 4D NOESY

Introduction

The determination of NMR chemical shifts for all nuclei of a protein, i.e. NMR chemical shift assignment, is an essential and time-consuming part of 3D structure elucidation projects based on NMR spectroscopy (Herrmann et al. 2002). However, even if the structure was obtained by other means, mainly X-ray crystallography, assigning the chemical shifts of at least a subset of nuclei is still required for additional investigations. Interactions between proteins or proteins and small molecule ligands are, e.g., often screened by chemical shift perturbation mapping using [¹H,¹⁵N]-HSQC spectra and, when the assignment of the perturbed shifts is known, even the binding epitope can be determined (Fesik et al. 1997; Hajduk 2006; Jang et al. 2012; Shuker et al. 1996). Additionally, NMR opens unique possibilities to study protein flexibility and plasticity including the occurrence of multiple stable conformations (Berjanskii and Wishart 2006; Gobl and Tjandra 2012; Kay 2005; Kleckner and Foster 2011; Mittermaier and Kay 2006). Unfortunately, the assignment, which has to be performed for these applications, might be as time-consuming as for structure determination if it is performed manually on the same set of spectra (HNCA, CBCA(-CO)NH, HNCACB, HNCO,...). However, utilizing the information of the known 3D structure can extremely reduce the number of possible assignments and in this way speed up the process and reduce the amount of needed spectra. This idea has been translated into a number of computational approaches (Chao et al. 2012, 2014; Hus et al. 2002; Jung et al. 2004; Jung and Zweckstetter 2004;

Electronic supplementary material The online version of this article (doi:10.1007/s10858-016-0050-0) contains supplementary material, which is available to authorized users.

✉ Thomas E. Exner
thomas.exner@uni-konstanz.de

¹ Institute of Pharmacy, Eberhard Karls Universität Tübingen, Auf der Morgenstelle 8, 72076 Tübingen, Germany

² Institute of Chemistry, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam OT Golm, Germany

Stratmann et al. 2009, 2010; Venditti et al. 2011; Xu et al. 2009; Xu and Matthews 2013). NOE_{net} (Stratmann et al. 2009) and NOE_{net}2 (Stratmann et al. 2010) compare graphs, which are generated from NOE restraints obtained from HSQC and 3D ¹⁵N-NOESY spectra, with the ones based on a known X-ray structure. To further increase the assignment success rate, chemical shift predictions were used. Pristovsek et al. (2002) define in their *st2nmr* approach a scoring function comparing experimental and predicted NOE restraints from ¹⁵N- and ¹³C-NOESY spectra. This is then used in a Monte Carlo optimization to obtain sequence-specific assignments. While these first approaches try to predict backbone and especially amide proton assignments, like the ones presented here, assignments of specifically labeled groups become relevant for larger proteins. With FLAMEnGO (Chao et al. 2012, 2014), methyl groups can be assigned based on HMBC spectra, chemical shift predictions as well as paramagnetic relaxation enhancements (PREs). Venditti et al. (2011) took PREs and 4D NOESY spectra to determine the methyl groups chemical shifts of enzyme I. Finally, MAP-XSII (Xu and Matthews 2013), the successor of the approach of Xu et al. (2009), is able to reliably assign most of the methyl groups based on 4D NOESY spectra. For very large systems like the proteasome (Sprangers and Kay 2007a, b), PREs can additionally be used. Besides NOE restraints, residual dipolar couplings (RDC) can also guide the assignment process. The approaches of Jung and Zweckstetter (Jung et al. 2004; Jung and Zweckstetter 2004) as well as Hus et al. (2002) should be mentioned here explicitly.

In this paper, we will present a new approach to assign amide H^N and N^H chemical shifts which is based on 4D-NOESY spectra only. NOEs are still the most important NMR spectroscopic information used for structure determination of biomacromolecules. However, finding unique assignments for NOEs can be a challenging and tedious task. This applies primarily to 2D-[¹H,¹H]-NOESY experiments since the majority of 2D NOESY cross peaks will show ambiguous assignment possibilities even in moderately sized proteins. This problem can be circumvented by adding additional dimensions that characterize the chemical groups involved in an NOE contact. A 4D NOESY spectrum provides the maximum information in this regard. It characterizes the two ¹H-nuclei (origin and target nucleus, respectively) that are close in space supplemented by the resonance frequencies of their immediately attached heteronuclei. Thus, 4D [¹H,¹⁵N]-HSQC-NOESY-[¹H,¹⁵N]-HSQC experiments will correlate two amide protons being close in space, and, in addition, provide the ¹⁵N chemical shift of both the origin and the target amide nitrogen nucleus. The analysis of such spectra is illustrated in the upper part of Fig. 1 for the

example of Ubiquitin resulting in a list of distance restraints between groups specified by their ¹H/¹⁵N chemical shift combination. With this information at hand, testing our approach on three and two systems with artificial and real 4D spectra, respectively, resulted in high success rates of unambiguous assignments. For most other residues, only two to three assignment possibilities remained, which can be further reduced by scoring the assignment possibilities based on a quantitative comparison of NOE restraints or chemical shift predictions. These excellent results come with the guarantee that the correct assignment is definitely included in the remaining possible assignments (even with large uncertainties in the measured peak volumes) and the protein can be separated into well and badly defined regions according to the ambiguity of the chemical shifts in these regions.

Materials and methods

The starting point of the new approach is the 4D [¹H,¹⁵N]-HSQC-NOESY-[¹H,¹⁵N]-HSQC spectrum as well as a 3D structure of the protein. In real applications, the structures will come from X-ray crystallography. Here, we will mainly use structures determined by NMR spectroscopy since for these a manually generated full assignment is available for comparison. NOESY cross peaks in the spectrum are manually picked, integrated, and distance restraints are derived. The assignment problem can then be formalized in two different but equivalent ways:

1. Find the sets of residues (RS), each of which can be assigned to a specific NOESY diagonal peak (DP) defined by the corresponding ¹H and ¹⁵N chemical shifts in the HSQC plane without violating any restraint determined by the volumes of the NOESY cross peaks (CP).
2. Find the sets of NOESY diagonal peaks (DP), each of which can be assigned to a specific residue (RS) without violating any restraint determined by the volumes of the NOESY cross peaks (CP).

The discussion will mainly use the first. In the following, the term “peak” without additional specification will be used for diagonal peaks while cross peaks will always be specified as such. To solve the assignment problem, assignment possibilities, which violate at least one of the restraints, are removed in an iterative optimization procedure starting from an initial completely ambiguous assignment. The resulting final assignment consist of unambiguously assigned diagonal peaks, i.e. amid resonances or peaks with only a small number of assigned residues. To further reduce the amount of ambiguities, the size magnitude of violation expressed as

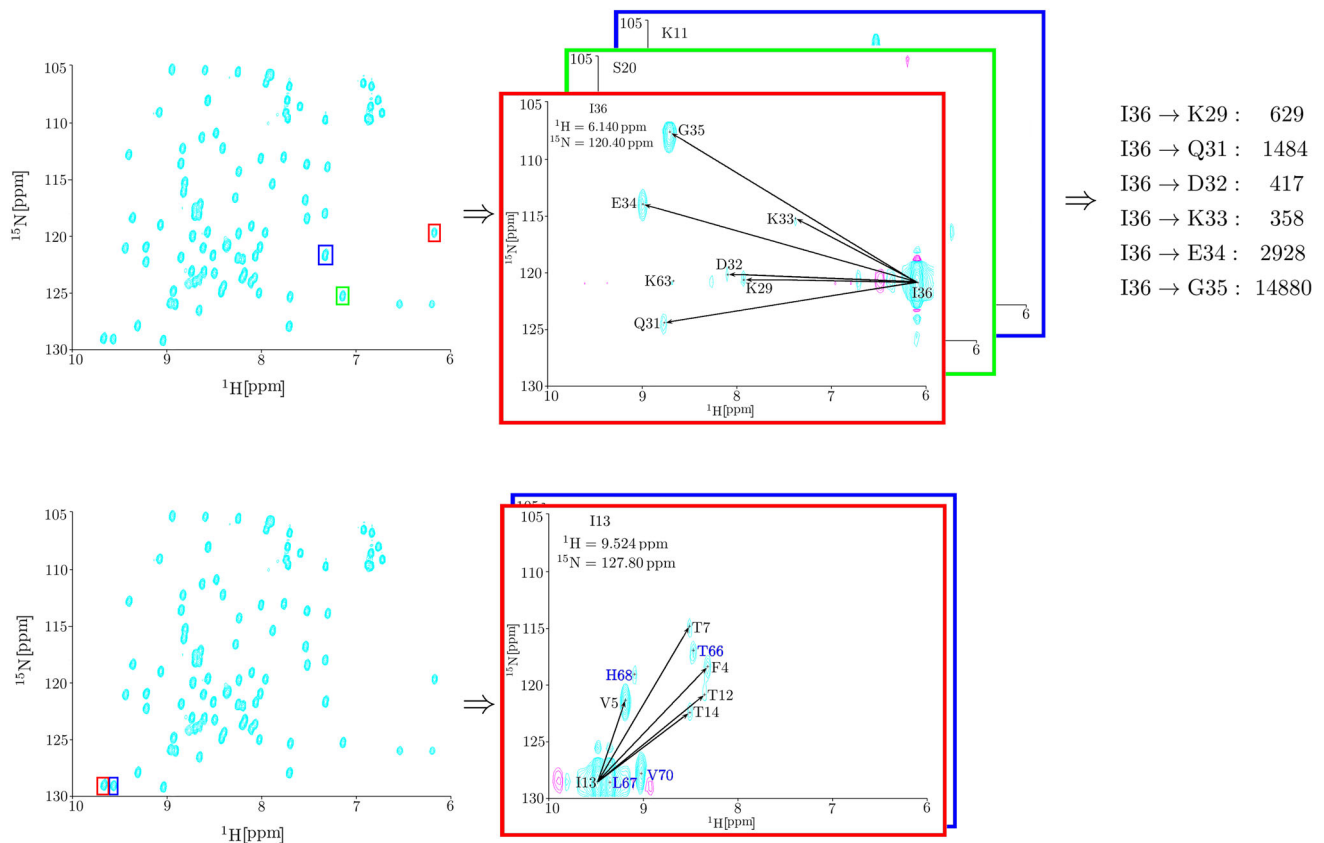


Fig. 1 To analyze 4D NOESY spectra it is convenient to work with 2D slices, here shown with a spectrum of Ubiquitin as an example. The full 4D spectrum can be handled as a series of 2D HSQC slices cut out of the 4D data set at all possible combinations of ^1H and ^{15}N chemical shifts. All peaks in the HSQC spectrum (first two dimensions, called origin in the following) are picked (*upper left*). Three peaks marked with *red*, *green*, and *blue* boxes with the $\delta^1\text{H}/\delta^{15}\text{N}$ coordinates 6.14 ppm/120.40 ppm (I36), 7.09 ppm/124.86 ppm (S20), and 7.26 ppm/121.98 ppm (K11) are examined exemplarily here. Since the assignment for Ubiquitin is known from the literature, we will use the knowledge about the residues belonging to these peaks to ease the description. The first two dimensions are then fixed to the coordinates of one peak after another. Each slice of the 4D NOESY extracted at these (origin) coordinates consists of an HSQC plane (dimension 3 and 4, called target) covering the full spectral width in both the ^1H and ^{15}N dimensions (*upper center*). The fixed values are shown in the *upper left* corner of each slice. The strong signal, where the target has the same coordinates as the origin, corresponds to the diagonal peak also known from the standard 2D NOESY. In addition, signals of further amide groups appear characterized by their $^1\text{H}/^{15}\text{N}$ frequencies and with the relative intensities as in the 2D NOESY, i.e. the volume is proportional to r^{-6}

with r being the distance between the two corresponding protons. In the case of I36, we see cross peaks to K29, Q31, D32, K33, E34, and G35. These cross peaks can be integrated and distance restraints can be derived to go e.g. into 3D structure calculation or as in our case into automatic assignment, with the goal to assign each peak to a specific amino acid pair (*upper right* listing the peak assignment and the peak volumes of the *red slice*) and, in this way, the chemical shift of each amid group. Signal overlap can complicate the situation. If the ^1H and ^{15}N chemical shifts of two amide groups are very similar, as shown for I13 (9.52 ppm/127.80 ppm, *red*) and L67 (9.43 ppm/127.88 ppm, *blue*) in the *lower left* part of the figure, the slices for these groups cannot be completely separated. Cross peaks of one group appear in the slice of the other group and vice versa (*lower center*). The closer the two peaks get in the first 2 dimension, the harder it will be to identify, which cross peak belongs to which group. In the worst case, when both amide groups have exactly the same chemical shifts, the distance restraints cannot be assigned to one amide group but the restraint has to be considered as violated only if the distance to both groups exceeds the upper limit of the restraint. To include such cases in the automatic assignment, peaks are defined as overlapped peaks if the difference in the chemical shifts is below a cutoff and two residues are assigned to this peak

sum of the squared errors (SSE) between distances in the 3D structure and restraints or predicted chemical shifts can be used to rank the remaining assignment possibilities. A schematic representation and a flow chart of the algorithm are shown in Figs. 2 and 3, respectively, and more details on each individual step are given in the following sections. A program implementing this algorithm, the source code and the Ubiquitin test example

described below are available at <https://github.com/thomasexner/4Dassign>.

4D [^1H , ^{15}N]-HSQC-NOESY- ^1H , ^{15}N]-HSQC Spectra

Experimental spectra for two protein systems, Ubiquitin and the N-terminal domain of enzyme I of the *Escherichia coli*

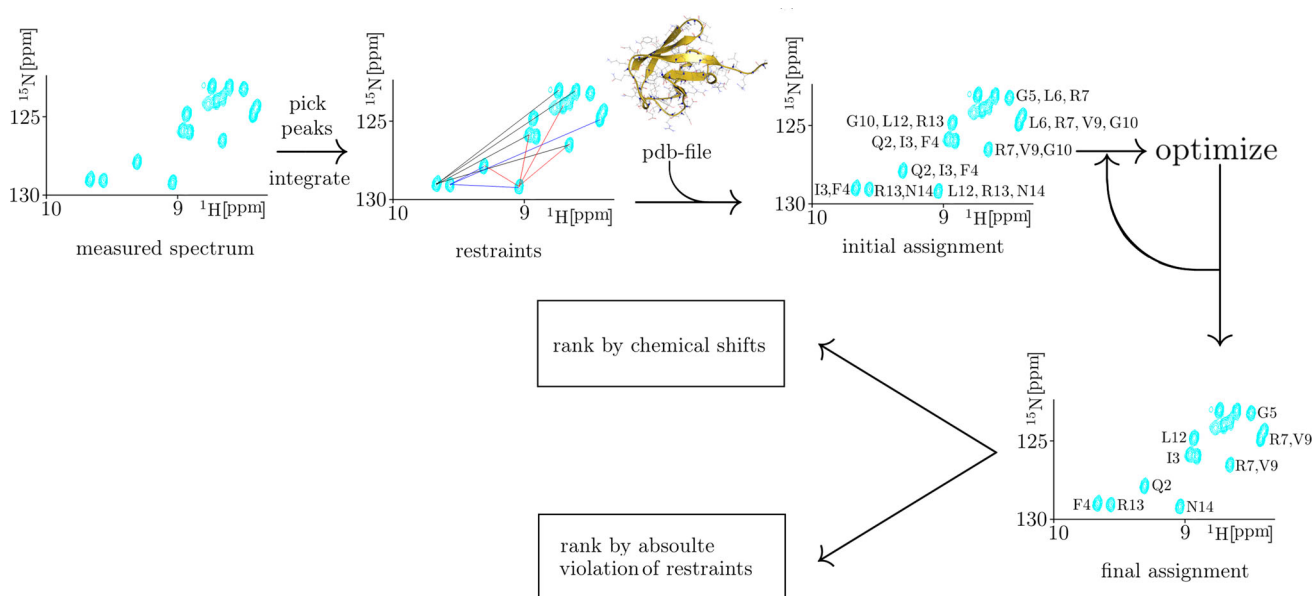


Fig. 2 Workflow of the automatic assignment of chemical shifts based on 4D spectra and a known 3D protein structure

phosphoenolpyruvate:sugar phosphotransferase, were used to test the approach. For the first, spectra were recorded on an 800 MHz Bruker spectrometer with room temperature probe (Remco Sprangers, private communication). The concentration was 2.5 mM in 40 mM PO_4^{3-} buffer at pH 5.8. A 120 ms NOE mixing time was used. The spectrum was recorded using uniform sampling with 1024, 28, 64 and 28 points in F4 (direct proton), F3 (indirect nitrogen, origin), F2 (indirect proton) and F1 (indirect nitrogen, target) dimensions. The resulting acquisition times are 40, 8.2, 3.3 and 8.2 ms. The spectral widths are 16, 21, 12 and 21 ppm. The data was processed with NMRPipe/NMRDraw (Delaglio et al. 1995). Peaks in the 2D HSQC slices of the 4D spectra were manually picked using CARA (Keller 2004a, b). The automatic procedure of CARA is then applied to integrate the picked NOE peaks, which are written out in CARA format and used as input for our procedure. To this input, a reference NOE peak is added with a peak volume corresponding to a distance restraint of 4 Å. A procedure to find an optimal value for the reference volume is described below. For comparison with the automatically generated assignment, a manual assignment was performed guided by the published data for PDB structure 1D3Z (BMRB entry 17,769). For the second protein, enzyme I, the chemical shifts of the amide groups as well as the NOE distance restraints were provided by G. Marius Clore (private communication) and taken as published (Garrett et al. 1997), respectively.

Overlapped amide resonances

In some cases, the differences between the ^1H and ^{15}N chemical shifts of different amide groups are too small to

assign NOE diagonal peaks unambiguously to either of these. In the lower part of Fig. 1, the 2D slices of origin peaks with very similar chemical shifts, I13 and L67 in this example, are shown. Besides cross peaks for the origin I13, which was used to fix the first 2 dimensions, cross peaks with similar or even higher intensity from the other origin L67 appear even if two separate peaks can be distinguished in the first two dimensions. In such cases, one can still identify the origin of each NOESY signal by finding the slice on which the maximum of each peak lies. This is, however, not possible anymore, when the chemical shifts get even more similar or are, in the worst case, exactly the same. At lower resolution, where such a separation might not be possible anymore, distance restraints to I13 would be defined erroneously even if the corresponding amide group is indeed close to L67. To circumvent such wrong distance restraints, origin peaks with very similar chemical shifts in the N and H dimension are combined and all NOE cross peaks are assigned to the averaged chemical shifts. Since the probability that three or more amide groups have the same chemical shifts is very small, a maximum of two residues will be later assigned to such diagonal peaks and we will call them overlapped peaks in the following. They are specifically marked in the input file to distinguish them from regular diagonal peaks with only one possible assignment. For the test calculation performed here, where the chemical shifts of all residues are known, two diagonal peaks with $\Delta\delta^1\text{H} < 0.01$ ppm and $\Delta\delta^{15}\text{N} < 0.1$ ppm were combined to an overlapped peak. For real applications, overlapped peaks can e.g. be identified relatively easily from 3D HNC0/TROSY-HNC0 spectra. Since these spectra require ^{13}C labeling of the protein, we also work on

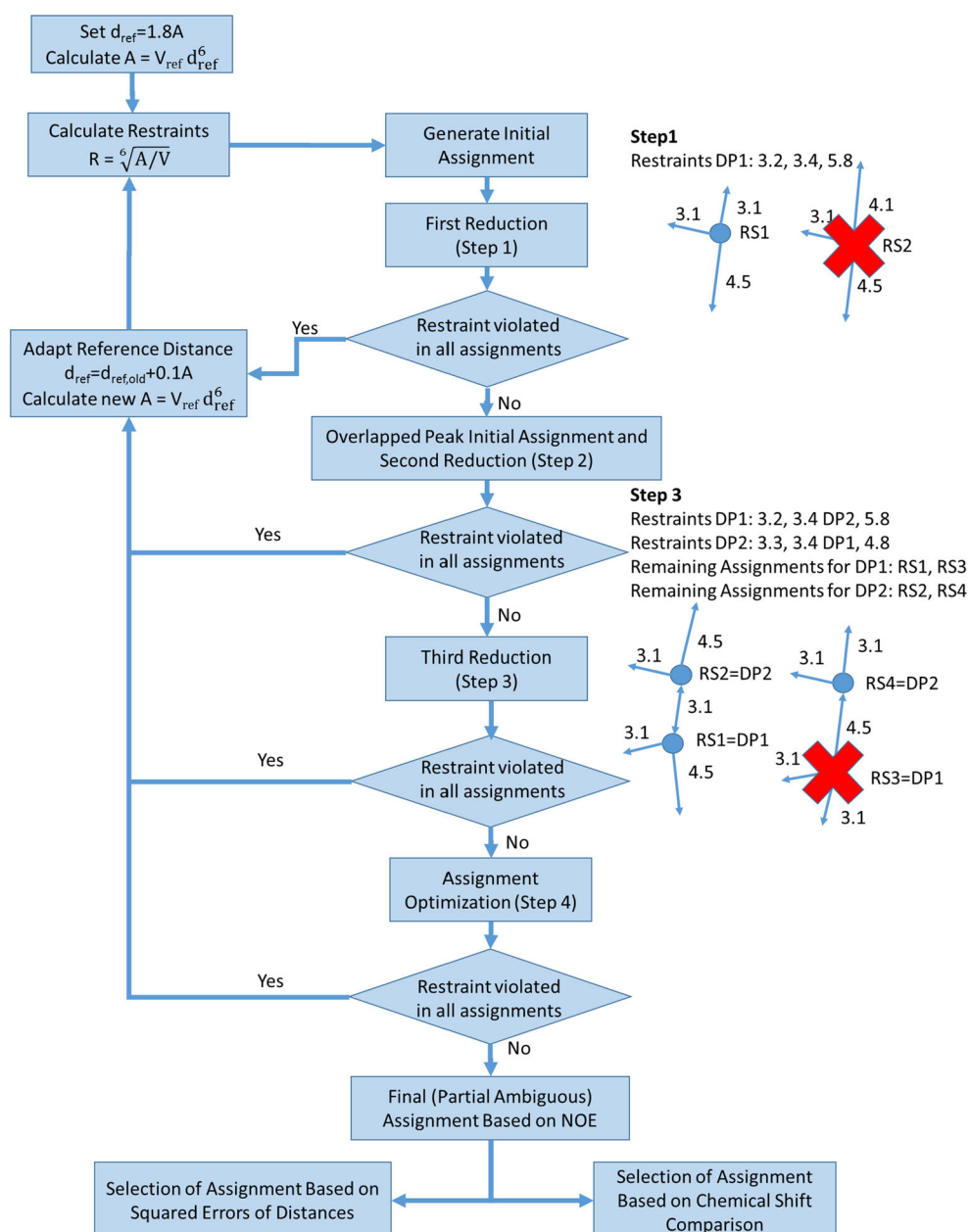


Fig. 3 Program flow chart. Four different steps are used to reduce the possible assignments starting from the completely ambiguous initial assignment. For step 1 and 3, schematic presentations of the selection criteria are shown on the right. The *arrows with numbers* correspond to distances found in the 3D structure. In step 1, all distances to the nearest neighbors of one residue are compared to the restraints of a diagonal peak (DP1) with volumes corresponding to restraints of 3.2, 3.4, and 5.8 Å, respectively. The *left side* of the schema for step 1 shows an assignment of the diagonal peak to a residue (RS1), in which all three distances are within the restraints while on the right side one restraint is violated when the diagonal peak is assigned to residue RS2. In step

2, the procedure is repeated including the overlapped peak, which were neglected in step 1. In step 3, both sides of a restraint are taken into account. With the origin peak assigned to DP1, the second restraint pointing to DP2 can be fulfilled by the assignment RS2 (*left side*). If the second option, RS3, is assigned to DP1, the restraint is violated with RS4 assigned to DP2 and every other assignment possibility so that RS3 can be removed from the assignment possibilities of DP1. In step 4, finally the complete network of restraints is considered using dead-end elimination. After each reduction step, a check is performed if any restraint is violated in every assignment and, if so, the volume of the reference distance is decreased and the procedure is repeated with longer distance restraints. For a detailed description see main text

an advanced variant of the approach, in which all diagonal peaks are initially considered as potentially overlapped and the optimization procedure selects the correct classification (regular or overlapped) simultaneously to the assignment.

Generation of an ambiguous initial assignment (step 1 and 2)

In the first step of the automatic assignment, every residue in the protein is assigned to every regular amide resonance, i.e. regular diagonal peak (the overlapped peaks are considered at a later stage), and vice versa to obtain a completely ambiguous initial assignment. Then assignments are removed, considering the information from one 2D slice at a time, i.e. only one origin with its nearest neighbors and the distances between them in the 3D structures are taken into account. The more stringent criterion, if a specific assignment of one diagonal peak fits consistently into the network formed by the NOESY restraints and the assignments of all other peaks, will only be used later in the optimization procedure described below. The restraints are calculated from the cross peak volumes based on the well-known equation $V = A \cdot r^{-6}$, with the constant A calculated according to the reference NOE peak for a distance of 4 Å provided in the input. To the so-defined distances a user-defined safety margin (standard value = 0.2 Å) is added. Please note that this safety margin was primarily introduced for the cases using artificial spectra to provide a more challenging test as when using the exact distances. However, even if the name suggests otherwise, increasing the safety margin can only be used to account for experimental uncertainty, i.e. too short restraints derived from some of the volumes, when the volume of the reference peak is known. In the case of real spectra, where this volume has to be determined in the iterative approach (see below), a larger (smaller) safety margin would be compensated by a larger (smaller) reference peak volume resulting in the same restraints and, finally, the same assignment. Thus, not the safety margin but this determination of the reference volume gives the algorithms some tolerance to deal with experimental uncertainties, since the restraints will be increased until an assignment without any violation can be produced. In the result, we will demonstrate that this approach can handle much larger experimental uncertainties (one order of magnitude) than the 0.2 Å added as safety margin for the theoretical examples. Also see the results and the conclusion section for detailed discussions on the error tolerance of the approach as well as on possible additional approaches to include larger experimental uncertainties. If the volume of a cross peak is close to zero, the distance is set to the maximum value of 7.5 Å. These modified values are used as upper bounds for the distance restraints. The restraints are then compared to the distances

seen in the 3D structure of the protein. If the 2D slice for a specific diagonal peak (origin) shows e.g. 3 cross peaks (see Fig. 3), the three nearest neighbors are searched for the amide group of each residue and the distances are calculated. If one distance is longer than allowed by the corresponding distance restraint (after sorting by length) as in the case of diagonal peak DP1 assigned to residue RS2 in Fig. 3, the residue is removed from the assignment possibilities of the peak. Beside the distances, no other information, like amino acid type and standard chemical shifts for this type, is used in this step. This procedure is then repeated for each diagonal peak. This way, unambiguous assignments of some of the peaks can already result. If only one possible residue remains for a specific peak (unique assignment), this residue can be removed as assignment possibility from all other peaks. Additionally, if two peaks share the same two residues as their only remaining assignment possibilities (pair assignment), these groups can also be removed from the assignment possibilities of all other peaks since the only remaining ambiguity is which of the two peak belongs to which of the two residues. This check is iterated over the list of diagonal peaks until no more assignments are removed.

In the second step, the overlapped peaks are included. Since two residues have to be assigned to these, a large number of assignments is in principle possible ($n \cdot (n - 1) / 2$ where n is the number of residues) if no additional information is available. However, due to the unambiguous single and pair assignments of regular peaks identified in the first step, these corresponding residues cannot be assigned to the overlapped peaks anymore, which reduces the number of possibilities. It is then tested for each pair of remaining assignments if the distance restraints are compatible with the distances seen in the structure. Every remaining assignment for the overlapped peak is then considered an ambiguous assignment of two residues. Then the NOE restraints are again compared to the distances seen in the 3D structure taking into account the neighbors of both amide groups simultaneously. To this end, the NOE-derived distances of all residues to either of the two residues from which the overlapped peak originates are checked against the distances found in the 3D structure.

Reduction of possible assignments (step 3)

In the third step, the assignment possibilities are further reduced by looking at individual restraints taking both sides of the restraint into account. Starting with one diagonal peak as origin and one as target of the restraint, all possible assignments remaining after step 1 and 2 are checked. Let's consider the example from Fig. 3 where peak DP1 has RS1 and RS3 and DP2 has RS2 and RS4 as possible assignments. RS1 has a shorter distance to RS2 of 3.1 Å than to RS4 and, thus, this distance is compared to

the restraint between DP1 and DP2 of 3.5 Å resulting in a valid assignment. If RS3 is assigned to DP1, the smallest distance in the 3D structure of 4.5 Å is to RS4. Even if both residues have the correct number of neighbors to fulfill all restraints and are, therefore, not removed in the first two steps, the distance between the two specific residues is longer than the restraint between DP1 and DP2. Thus, this restraint is violated by every possible combination of DP1 with a possible assignment of DP2 and RS3 can be removed from the assignment possibilities of DP1. This reduction will possibly produce additional unambiguous assigned residues, which can then be removed from every other peak. Therefore, this procedure is again executed repetitively until no more assignments can be removed.

Assignment optimization (step 4)

In the final fourth step, the full network of NOE restraints is considered to remove additional assignment possibilities. This is done using a dead-end elimination approach. We will exemplify the approach with four peaks as shown in Fig. 4. A tree structure is generated starting with the first ambiguous diagonal peak (DP1) in the list and then following all NOE distance restraints. Thus, the second layer of nodes of the tree (the first layer is just the starting peak DP1) consists of all assignments (RS1–RS4) of the first peak remaining after step 1–3. Then possible assignments for the peaks connected to the first one by an NOE restraint are identified. These are DP2 and DP3, to which the residues RS2, RS3 and RS4 are assigned at this stage. The third layer of nodes is formed by one of these possibilities for each NOE restraint. Residues, which are already

assigned in an earlier node of a specific branch (in this example RS1 is already assigned to DP1), are not considered again. This is continued until all NOE restraints and all diagonal peaks are included in each branch of the tree. After the full tree is generated, every path from the head node to the leaves (final node of a branch) corresponds to one possible unambiguous assignment of residues to diagonal peaks, called complete assignment from here on. However, such a complete assignment is only valid if none of the restraints represented by edges between nodes is violated. If, e.g., the distance between RS1 and RS3 is too long to fulfill the distance restraint between DP1 and DP2, the second and fifth branch from left marked in red can be removed. Such a restraint violation can be depicted by breaking the edge between the nodes producing a dead end in the tree giving the approach its name. Since all of the assignments produced by the subtree below the broken edge have at least one violated restraint, this complete subtree does not need to be generated and checked for further violations resulting in considerable computer time savings.

If all branches starting from a specific assignment of the starting peak are checked and in each of them one of the connections between two nodes is broken, the residue can be removed as a possible assignment from the peak since it cannot lead to a complete assignment in agreement with all NOE restraints. If this leads to an unambiguous assignment of this peak, the reduction of possible assignments as described in the previous section is called again (step 3). Step 4 is continued with the other diagonal peaks as head node and iterated since some branches of the tree of one head node might be removed due to the removal of possible

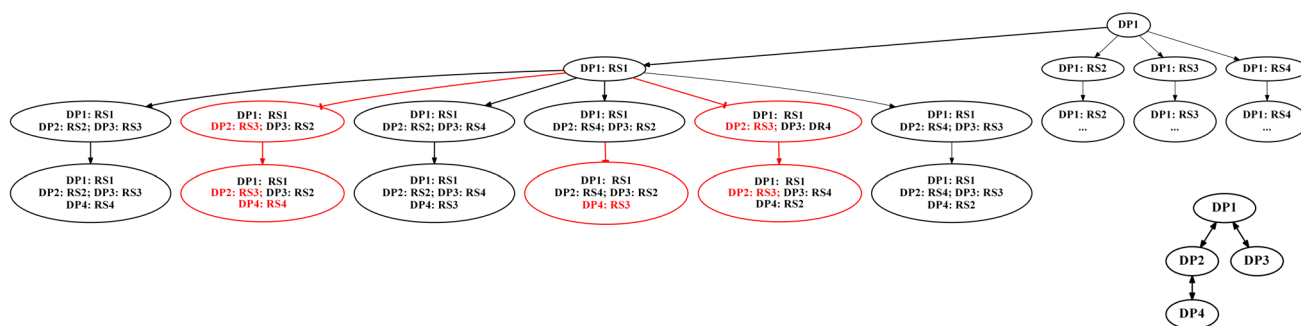


Fig. 4 Assignment tree for the optimization routine: In this example with only four amino acid residues, we start with a completely ambiguous assignment, where all 4 peaks (DP1 to DP4) can be assigned to the amid groups of all 4 residues (RS1 to RS4). We also assume that there are NOE cross peaks between residues corresponding to peaks DP1/DP2, DP1/DP3, and DP2/DP4 (see *small graph on the right*). Only the branch for the assignment of RS1 to DP1 is shown explicitly (*large graph*). As the head node in the first layer DP1 is chosen, which has four assignment possibilities forming the second layer. In the third layer, all NOE restraints originating at the residue of

the head node, i.e. DP1↔DP2 and DP1↔DP3 in this case, are considered and all possible combinations of assignments for DP2 and DP3 are compiled. Since all other peaks have already been assigned in each branch, only one assignment possibility is left for DP4, which is listed in the last (4th) layer. Finally, if the distance between RS1 and RS3 in the 3D structure is too long to fulfill the restraint between DP1 and DP2 derived from the corresponding cross peak volume and the distance between RS3 and RS4 is too long to fulfill the restraint between DP2 and DP4, all parts marked in *red* will be cut from the tree

assignments by the reduction procedure executed during the evaluation of the tree of other head nodes. If no more assignments can be removed, the remaining possibilities for each peak are written out and the percentage of unambiguously assigned peaks is calculated. Additionally, all possible complete assignments can be generated. Since in most cases, the assignment of one peak is not independent of the assignment of the remaining peaks, the number of these complete assignments is much lower than the combinatorial enumeration of all remaining assignment possibilities of all the peaks.

Calibration of the reference volume

In our approach of structure-based automated assignment as well as in NMR-based structure determination, the constant A has to be determined to be able to convert NOE peak volumes into upper bounds of the distance restraints. In automated structure-independent calibration used in programs like CANDID (Herrmann et al. 2002), this is done by setting the average distance of all observed restraints of a specific group of atoms to a predefined value. After the first preliminary structures are obtained the constant can be recalibrated such that no more than a predetermined percentage of the upper bounds are violated. In ARIA (Linge et al. 2003; Rieping et al. 2007), the ratio of the average of experimental to the average of the theoretical volumes is used as calibration constant. Theoretical volumes can be calculated using either the isolated spin pair approximation or relaxation matrix analysis (Linge et al. 2003; Rieping et al. 2007). Here we follow a different approach in order to be able to analyze the influence of the constant on the assignment success rate as well as avoiding problems related to the choice of a sub-optimal constant. The constant is determined by trial-and-error. It is first assumed that the peak with the largest volume in the NOE spectrum results from the shortest possible distance between two protons based on the van-der-Waals radius of approximately 1.8 Å. A corresponding constant A and the volume for the reference 4-Å peak can then be calculated. With this, the assignment procedure can be executed. Due to the large A , all experimental peaks are translated into very short distance restraints and many of them cannot be fulfilled, irrespective of the assignment used, and the procedure will terminate without generating a single valid complete assignment (all branches of the tree used in the optimization will be cut at one point of time). The distance restraint for the peak with the largest volume is then increased in 0.1 Å steps resulting in a larger and larger volume of the 4-Å peak (A is increased), until the procedure is able to generate at least one complete assignment.

Results and discussion

Test systems

Ubiquitin

Most of the tests regarding the performance of our approach were performed using Ubiquitin. Since it is the guinea pig of NMR method development, full assignments are available. Additionally, a large number of X-ray and NMR structures exist. Finally, due to the availability of protein samples, the complete procedure starting with the experimental 4D spectrum as depicted in Fig. 2 could be executed. Beside this experimental spectrum, artificial data composed of a list of peak volumes for the 4D spectrum was generated based on the published chemical shifts and the distances seen in the 3D structures. The structures were taken from the Protein Data Bank (PDB) (Bernstein et al. 1977). PDB entry 1D3Z (first model) (Cornilescu et al. 1998) and the corresponding BMRB (Ulrich et al. 2008) entry 17769 were used. For 72 of the 76 residues, assignments for the NH groups are available (missing: Pro19, Pro37, Pro38, and Met1). NOE contacts were defined for all pairs of protons of amide groups with distances below 5.5 Å in the reference structure resulting in 338 NOE contacts. Since our algorithm expects a list of peak volumes, the distances are translated into volumes using the r^{-6} relation and an arbitrary constant A . This artificial spectrum corresponds to the data set which could be obtained under optimal (experimentally not achievable) conditions especially regarding protein rigidity and spectral resolution.

YuaF from Bacillus subtilis

This protein is investigated in the group of one of the authors (Walker et al. 2008). It belongs to the NfeD-like proteins and plays a potential role in maintaining membrane integrity during conditions of cellular stress. Full backbone and side chain assignments of the soluble C-terminal domain (sYuaF, residues 97–174 of full-length YuaF) were obtained from triple-resonance spectra and the high-resolution solution structure was obtained based on this data and 3D NOESY spectra. Due to its particular structure composed of an extended N-terminal α -helix only weakly connected to a β -barrel formed by five β -strands, we chose this system as another example for which an artificial 4D spectrum was generated. The 3D structure and the NOE restraints were taken from the PDB entry 2K14 (Walker et al. 2008) (first model) and the chemical shifts from BMRB entry 15475 (Ulrich et al. 2008) (missing residues: Gly1, Ser2, His3, and Asn73 as well as Pro32, Pro80, and Pro83). For this system, a maximum distance of

7.0 Å was used to define the NOE contacts resulting in 366 NOE contacts.

N-terminal domain of enzyme I of the Escherichia coli phosphoenolpyruvate:sugar phosphotransferase system (enzyme I in the following)

As the last, most challenging test case, enzyme I was chosen since its structure was solved with the help of 4D NOESY spectra (Garrett et al. 1997). It is involved in the simultaneous phosphorylation and translocation of numerous sugars across the cytoplasmic membrane. The structure of the 259 residue-long protein was taken from PDB entry 1EZA (Garrett et al. 1997). It comprises two domains, an α/β domain (residues 1–20 and 148–230) consisting of six strands and three helices and an α -domain (residues 33–143) consisting of four helices, which are connected by two linkers (residues 21–32 and 144–147). The chemical shift assignments were provided by Dan Garrett and Marius Clore (personal communication, missing residues: Met1, Arg186, Thr187, and Ser191 as well as Pro9, Pro165, Pro200, and Pro231). 1106 experimental NOE contacts were used in the structure determination. For the first test described in the next chapter, an artificial spectrum was also generated from the known 3D protein structure using a maximum distance of 7.0 Å for the proton–proton distances. This resulted in 1276 artificial NOE contacts, 170 more than in the experimental spectrum.

Assignment performance using artificial spectra

For all three test cases, we ran our 3D-structure-guided algorithm using the artificial spectra, i.e. an optimal set of peak volumes and, this way, distance criteria. The results are summarized in Table 1. For Ubiquitin, the 338 peak volumes were internally retranslated into distances using the same constant A as during the generation of the artificial peak volumes. The resulting exact distances plus a safety margin of 0.2 Å were used as upper bounds for the NOE contacts. Since all amide groups have distinct $^1\text{H}/^{15}\text{N}$ chemical shift combinations given in the published data (Cornilescu et al. 1998), all peaks are nicely separated in the first two dimensions of the 4D spectrum and no amide groups had to be combined to overlapped peaks. With this

complete network of NOE contacts, the algorithm was able to correctly and unambiguously assign all peaks. In the following we will list the percentage of correctly and unambiguously assign peaks as unambiguous assignment success rate (UASR) so that the Ubiquitin results correspond to an UASR of 100 %. In the supporting information, the assignment possibilities after each step of the procedure are listed. For this small test case, step 1 and 2 already yield a number of unambiguous assignments and all peaks are assigned after the first reduction of possible assignments (step 3). The extremely short calculation time of 7.7 s on a single core of an Intel Core i7-3770 CPU is resulting from this, since the most time-consuming optimization (step 4) does not have to be called at all. As described above, NOE contacts up to 7.0 Å were taken for YuaF and enzyme I resulting in 366 and 1276 restraints, respectively. The pairs Tyr62/Leu67 in YuaF as well as Ile31/Glu109 and Arg131/Glu155 in enzyme I show very similar ^1H and ^{15}N chemical shifts and were combined to one overlapped peak each. Also for these additional artificial test sets, very good results were obtained. Only the peak belonging to the C-terminal residue Ile84 could not unambiguously be assigned in YuaF. This peak does not show any NOE contacts. Therefore, it can be assigned to all amide groups of the missing residues Ser2, His3, or Asn73, which do not have a visible amide resonance, in addition to the correct assignment. In the same way, the peak of His189 in enzyme I was also assigned to Ser191, for which the amide resonance is missing in the data. For this last example, the unambiguous assignment success rate is 0 % (no unambiguous peak) after the first three steps and the optimization has to generate and reduce large assignment trees resulting in a high increase in calculation time.

Adding uncertainties to the artificial spectra

To make the performance test more realistic, we added random noise and/or removed a defined number of distance restraints from the artificial data of Ubiquitin to model uncertainties in the experiment.

Due to molecular flexibility but also spectral noise and uncertainties in peak volume integration, the experimental distance restraints are much less exact than our artificial ones. To model this, we added a number randomly selected

Table 1 Assignment success of the 3D-structure-guided algorithm on artificial data for three proteins

Protein	Max. distance (Å)	No. of NOE contacts	UASR (%)*	Calculation time (s)**
Ubiquitin	5.5	338	100	2.8
YuaF	7.0	366	98	6.6
Enzyme I	7.0	1276	99	392

* Unambiguous assignment success rate

** On a single core of an Intel Core i7-3770 CPU running at 3.40 GHz

between 0 and a maximum value to each restraint and generated new, smaller peak volumes for the artificial spectrum with these. Larger volumes than the optimal ones were not used since termination of the algorithm with an error is unavoidable. Too large volumes lead to too small restraints, which cannot be fulfilled by any assignment. In real application, this problem will be circumvented by the trial-and-error determination of the reference peak leading to a larger constant A . The results for different maximum noise levels, i.e. maximum of the random values added to the upper bound of a restraint, are given in Table 2. They show clearly that the uncertainty in the distances can be up to 1 Å without reducing the number of unambiguous assignments. Even with random noise up to 2 Å, only single pairs or triples of peaks show ambiguities. These belong to short, solvent exposed turns showing NOE cross peaks only to the neighboring residues on both sides. Since no information to identify amino acids types is used and the differences in the distances of the amide group in the pairs/triples to the neighbors are in the range of 1–2 Å, all combinations of assignment are possible. Due to the independent random selection of the value added to each peak, the ambiguous peaks differ from one experiment to the other. Using different random seeds, ambiguities might occur also for maximum noise levels below 1.2 Å if the important restraints to distinguish the amide groups of the turns are elongated by values close to the maximum. Nevertheless, we can conclude that the approach can cope with random noise and the assignment success rate is only slightly reduced. This good performance has to be paid for by longer calculation times, which is not unexpected since steps 1–3 of the procedure generate less unambiguous assignments leaving more work for optimization in step 4.

The behavior of the approach when removing a given number of restraints was investigated in the second performance test. This simulates the effect that NOE cross

peaks can get hidden by other very intense peaks or in the random noise of the experimental spectrum especially for larger distances between the amide groups. First, we randomly selected cross peaks of the artificial spectrum and removed them from the input file. For the more challenging test, where complete restraints, i.e. both symmetry related cross peaks, are removed simultaneously, see below. Since it is expected that the results depend on the random selection even stronger than in the previous performance test, 100 independent runs with different random seed were performed. In Fig. 5, the unambiguous assignment success rates (UASR) for the best- and worst-performing run for each number of removed restraints are visualized. Even with 50 % of the restraints removed, the worst UASR is still almost 90 %. In the best cases, even a complete unambiguous assignment can be obtained. When removing more restraints, the UASR drops very quickly to 0.

Special note deserves the region between 100 and 150 remaining restraints of the worst-performing runs. Depending on the selected restraints, UASR of 0 or more than 70 % are seen. Note, that the removed cross peaks are chosen independently in each run leading to completely different restraints remaining in the worst-performing runs with e.g. 120 and 121 cross peaks. This extreme variation can be attributed to the specific structure of Ubiquitin (see Fig. 6). The first 33 residues form an α -helix, which is connected to the rest of the structure by a very limited number of restraints. Especially since it is followed by two Pro residues, restraints defining the sequential connection are missing. If, on the one hand, the random selection removes mainly restraints from this helix, the algorithm is not able to identify this specific secondary element and the unambiguous assignment of any peak becomes impossible. On the other hand, with enough restraints to assign the helix in the first steps of the approach, a large number of peaks do not have to be considered in the optimization procedure.

Table 2 Assignment success for Ubiquitin depending on the maximum value for random noise

Maximum noise (Å)	UASR (%)	Calculation time (s)	Ambiguous residues
0.0	100	2.8	
0.2	100	2.8	
0.4	100	5.1	
0.6	100	5.8	
0.8	100	4.6	
1.0	100	13.0	
1.2	97	17.9	Asp39/Gln40
1.4	100	23.8	
1.6	94	23.6	Asp39/Gln40 Asp52/Gly53
1.8	91	81.0	Glu34/Gly35/Ile36 Asp39/Gln40/Gln41
2.0	91	151.6	Glu34/Gly35 Asp39/Gln40 Asp52/Gly53

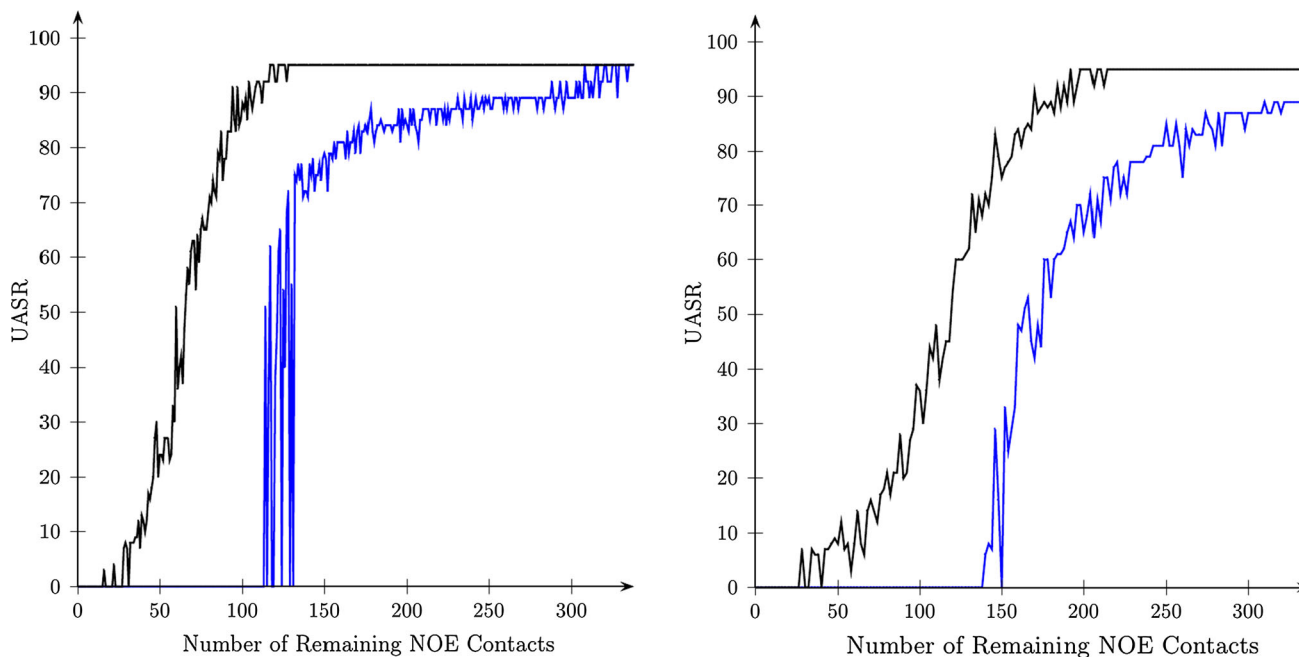


Fig. 5 Assignment success depending on the number of NOE cross peaks removed from the artificial spectrum of Ubiquitin: (*left*) Symmetry-related cross peaks are removed individually and (*right*) both symmetry-related cross peaks are removed at the same time. Black and blue: random selection leading to the best and worst assignment success, respectively. Please keep in mind, that

unambiguous assignment success rates (UASR) are reported. For many signals only very few assignment possibilities remain after the optimization even for UASR of 0 % and the correct assignment is always included. Reasons for the extreme variation of the worst assignment success in the range of 100–150 cross peaks are given in the main text

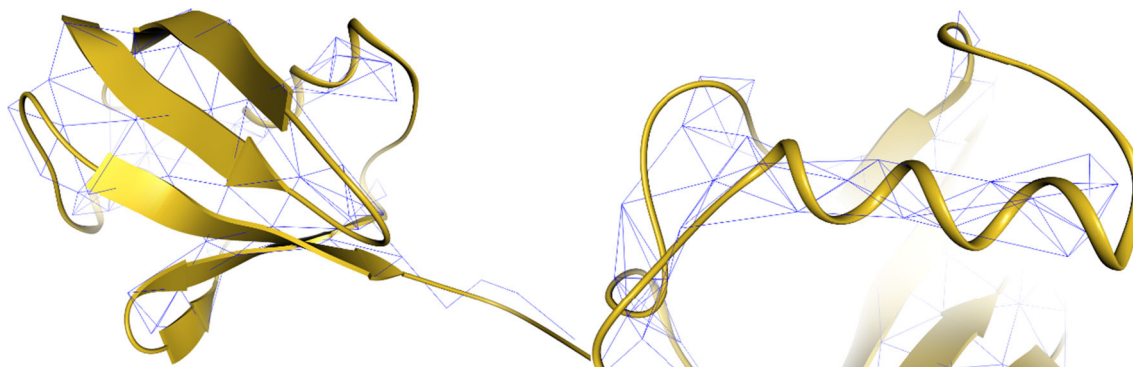


Fig. 6 H^N-H^N -NOE restraint network (*blue lines*) of Ubiquitin (*left*) and a close-up showing the α -helix (*right*). While there is only a limited amount of restraints between the N-terminal α -helix, the helix itself is well defined by restraints. However, when these helix-

defining restraints are removed, this secondary structure element cannot be identified anymore and the unambiguous assignment success rate degrades significantly

Even with a small number of remaining restraints to define the other parts of the molecule including the β -sheet, a partial assignment is possible due to the distinct structure motifs with their specific distribution of distances.

In the test just described, single peaks were removed. Since a short distance between two amide groups results in two symmetry-related cross peaks, a restraint between these groups can be defined even if only one of the peaks is seen in the spectrum. Thus, in the best case, where all peaks are removed from the upper or lower triangle of the

NOESY part of the 4D spectrum, half of the peaks can be removed without losing information. To account for this, we repeated the test but removed restraints instead of single peaks. This was done by removing both symmetry-related peaks simultaneously. It is clear that with this setting the UASR degrades faster with the number of removed peaks (see Fig. 5). Nevertheless, even in the worst case UASR of 80 % and more are obtainable when 50 restraints (100 cross peaks), i.e. almost 1/3 of the available information, are removed.

Table 3 Optimization of the volume of the standard peak for the test protein ubiquitin

Distance assumed for largest volume (Å)	Reference peak volume	UASR (%)	Number of violated restraints in correct assignment	Calculation time (s)
1.0	7453	–	228	1
1.5	84,901	–	225	1
2.0	477,031	–	209	1
2.5	1,819,729	–	105	1
2.6	2,302,538	–	69	1
2.7	2,887,682	–	43	10
2.8	3,591,823	–	24	16
2.9	4,433,584	–	8	24
3.0	5,433,684	–	3	30
3.1	6,615,108	82 (62 residues)	0	38
3.2	8,003,257	82 (62 residues)	0	49
3.3	9,626,102	82 (62 residues)	0	63
3.4	11,514,377	82 (62 residues)	0	76
3.5	13,701,721	71 (54 residues)	0	127
3.6	16,224,891	0	0	333
3.7	19,123,934	0	0	263
3.8	22,442,354	0	0	309
3.9	26,227,360	0	0	354
4.0	30,530,000	0	0	513

Cases, where the algorithm terminates without any assignment are marked by “–” in the unambiguous assignment success rate (UASR) column

Assignment performance using real spectra

Ubiquitin

For the following examples using real experimental data, the peak intensities have to be first converted into distance restraints by identifying a suitable constant *A*. As described in the section “[Materials and Methods](#)”, we assumed different distances for the peak in the NOE spectrum with the largest volume (NOE signal between residues Glu34 and Gly35 with a volume of 30530000) and determined the volume for the standard 4-Å peak based on these assumptions. The results, in which the atomic distances were extracted from PDB structure 1D3Z, are listed in Table 3. Setting the distance corresponding to this peak to ≤ 3.0 Å, the upper bounds, even after adding the safety margin of 0.2 Å, are shorter than the corresponding atomic distances based on the correct assignment. Since the algorithm removes an assignment if any one restraint is violated, it is clear that the correct assignment cannot be obtained. More interesting, all other wrong assignments are also rejected and the algorithm terminates with an error indication that no valid assignment is generated. When the upper bounds are increased and all become larger than the actual distances, assignments including the correct one are successfully generated. Further increasing the upper

bounds does not change the unambiguous assignment success rate (UASR) significantly. Only the calculation time increases up to ca. twofold. When the assumed distance for the peak with the largest distance reaches 3.5 Å, the UASR deteriorates. At 3.6 Å, the UASR drops to 0 %. The reason for this extreme decline is that the first three steps of the algorithm (generation of an ambiguous initial assignment for regular and overlapped peaks, reduction of possible assignments) are not able to generate an unambiguous assignment for any of the amide resonances. Due to the interdependence of the assignments, high UASR can be reached even if only a small number of peaks or even only one are assigned unambiguously after the first iteration of the optimization (step 4), as happens at distances up to 3.5 Å. In contrast, without any unambiguous assignment, no further optimization is possible with the proposed algorithm and a UASR of 0 % is unavoidable. This does, however, not mean that no useful information is produced. Fixing one or a small number of assignments by other experimental or computational means (see ranking by NMR chemical shifts or distance correlation below) could be enough to help the optimization to recover UASR comparable to values seen at a distance of 3.5 Å. To prove this, we ran the algorithm again fixing one of the amide resonance to the correct assignment. For all but one residue (Glu24, which has no NOE contact) UASR between 72 and

M1 Q2 I3 F4 V5 K6 T7 L8 T9 G10 K11 T12 I13 T14 L15 E16 V17 E18 P19 S20
 D21 T22 I23 E24 N25 V26 K27 A28 K29 I30 Q31 D32 K33 E34 G35 I36 P37 P38 D39
 Q40 Q41 R42 L43 I44 F45 A46 G47 K48 Q49 L50 E51 D52 G53 R54 T55 L56 S57 D58
 Y59 N60 I61 Q62 K63 E64 S65 T66 L67 H68 L69 V70 L71 R72 L73 R74 G75 G76

Fig. 7 Ambiguously (*red*) and unambiguously (*green*) assigned residues of Ubiquitin using the optimal reference peak volume. Proline residues, Met1, and Gly53, for which no amide resonances were detected in the 4D spectrum, are highlighted in *black*

80 % (51 and 61 unambiguous residues) were obtained outperforming the results for a distance of 3.5 Å and almost reaching optimal UASR. Similar improvements by pre-assigning single peaks should be obtainable for spectra with a large number of missing peaks as simulated in the previous section.

Based on this one example, the result justifies taking the smallest distance, at which an assignment is possible, as optimal value to calculate the volume of the reference peak. Since the runs using smaller distances very quickly terminate due to violated restraints, the reference peak determination only increases the calculation time by an acceptable three to four fold. Additionally, starting by a more reasonable distance of e.g. 2.8 Å would have almost no influence on the computational demand. The results also imply that the determination of the reference can tolerate some experimental uncertainties since larger distances also lead to acceptable USARs. Due to differences in the experimental setup of structure determination versus recording of the 4D spectra, e.g. X-ray crystallography versus solution NMR, distances in parts of the molecule could differ in the two samples and some of them could be larger than others even if they show the same peak volume. In such cases, the reference peak determination should lead to upper bounds which are large enough to describe all observed distances. Considering the observed error tolerance with respect to larger-than-optimal reference volumes (larger upper bounds), good UASR should be still obtainable. For our Ubiquitin test examples, such effects might be already relevant since the data used for the structure determination and the automatic assignment are from two different sources with different experimental conditions. Still, both are based on spectra from solution NMR. More differences are expected when an X-ray structure is used as reference. Repeating the calculation with the PDB entry 1UBQ (Vijaykumar et al. 1987) as reference, which is one of the earliest Ubiquitin X-ray structures, gave exactly the same results regarding the optimal reference peak volume and the UASR (data not shown). However, since the differences between the two Ubiquitin structures are only marginal (C_{α} root mean square deviation of 0.52 Å between 1UBQ and 1D3Z), the degree of dissimilarity tolerated by the algorithm must be determined by additional examples with multiple 3D structures when 4D spectral data becomes available for these.

126.97 7.63 (3) => T9 G10
 109.40 7.82 (4) => T9 G10
 121.26 9.42 (0) => E24 D52 G53
 111.91 8.96 (3) => A46 G47
 123.54 8.11 (2) => A46 G47
 120.45 8.15 (2) => D52 G53
 124.53 8.32 (2) => L73 R74
 121.99 8.41 (1) => L73 R74

Fig. 8 Ambiguously assigned amide resonances and the corresponding residues. The first two numbers are the nitrogen and hydrogen chemical shifts. Then the number of NOE contacts follows in *brackets*. The correct and the wrong assignments are given in *black* and *red*, respectively

For the optimal reference volume, i.e. smallest volume leading to a possible assignment, fourteen residues cannot be assigned unambiguously (see Figs. 7, 8) resulting in an overall USAR of ca. 82 %. These include the three proline residues Pro19, Pro37, and Pro38 as well as Met1, without amide protons or with a protonated, fast exchanging N-terminal amine group, respectively. Additionally, Gly53 does not show an amide resonance in the 4D spectrum and is assigned to the signal of Asp52 as a second possible assignment. No NOE contacts were detected for Glu24 in the 4D spectrum and, thus, also no useable information is available for it. The remaining 6 residues form independent pairs, for which the algorithm cannot determine which signal belongs to which residue. These are all part of solvent-exposed hairpins and only show NOE contacts to the neighboring residues. Due to the compact shape of the hairpin, the distances are all small and very comparable inevitably leading to ambiguities. As example, the local arrangement around residues Thr9 and Gly10 is shown in Fig. 9. The large volumes for these short restraints are error-prone due to overlap of the peak. Additionally, the experimental volumes could be too small and, thus, the restraints too long due to broadening by increased exchange or motion of the complete hairpin relative to the rest of the molecule.

Enzyme I

Since only distances and not volumes are given in the published data but our algorithm uses the latter as input, we

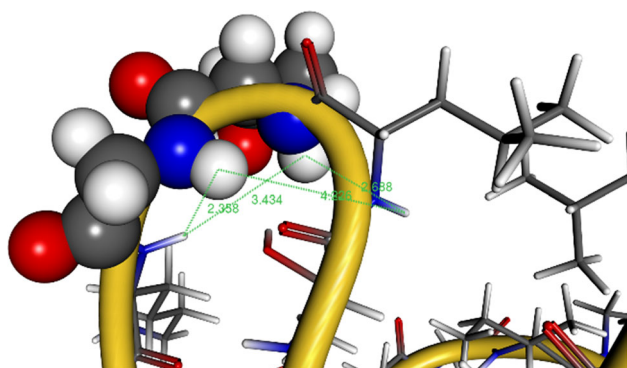


Fig. 9 Hairpin formed by the residues Leu8 to Lys11. The distances from Thr9 and Gly10 to their neighboring residues are all very small and the restraints can be fulfilled independent of the assignment of these central residues

chose an arbitrary constant A for the backtranslation in the same way as for the artificial spectra. The results using the experimental NOE contacts are given in Fig. 10 showing an overall unambiguous assignment success rate of 74 % (191 unambiguous, 60 ambiguous assignments, and 7 missing residues). The ambiguously assigned resonances and the corresponding residues are given in the supporting information. Additionally, the assignment possibilities after each step of the procedure are listed showing that the optimization step is more important for this complex, real-world problem compared to Ubiquitin. Similar to the experimental data for Ubiquitin, pairs and triples of ambiguous residues can be seen, which again are situated within solvent-exposed hairpins showing only a small number of NOE contacts. There are, however, two longer patches of ambiguously assigned residues. The first one belongs to the C-terminal α -helix. As shown in Fig. 11, there are no restraints connecting the helix with the rest of the molecule. This is caused by Pro231 connecting the helix to the rest of the molecule and breaking the NOE network along the sequence. In addition, this helix is

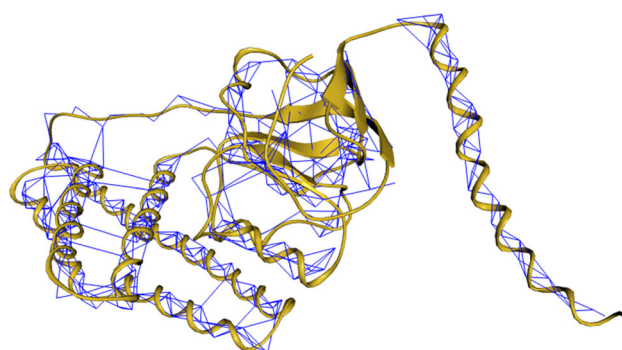


Fig. 11 Restraint network in enzyme I

apparently only loosely anchored to the core of the protein and increasing flexibility towards the C-terminus is expected to further reduce NOE intensities. Due to the point symmetry of a helix, the start and end cannot be determined by distances in the helix alone resulting in two possible assignments starting from Thr232 and Arg259 and then following the sequence in forward and reverse order, respectively. This opens the possibility to get an almost complete unambiguous assignment of the complete helix if one residue is assigned by other means since this rules out the wrong, reverse assignment e.g. by specific labeling/unlabeling. Similar reasons lead to the problems for residue Gly194 to Leu199 (see Fig. 12). Since the amide resonance of Ser191 was not observed in the 4D spectrum and residue 200 is a proline, the assignment of the last part of this patch can be moved one residue towards the N-terminus leading to an additional wrong assignment of the signal for residue n to residue number $n-1$. The assignment of the first part of the patch can be moved one residue towards the C-terminus due to the two missing amide resonances for Arg186 and Thr187 and the small number of restraints connecting the other residues. This ambiguity can also be resolved by including some additional information as demonstrated in the next section.

M1 I2 S3 G4 I5 L6 A7 S8 P9 G10 I11 A12 F13 G14 K15 A16 L17 L18 L19 K20 E21 D22 E23 I24 V25 I26 D27 R28
 K29 K30 I31 S32 A33 D34 Q35 V36 D37 Q38 E39 V40 E41 R42 F43 L44 S45 G46 R47 A48 K49 A50 S51 A52
 Q53 L54 E55 T56 I57 K58 T59 K60 A61 G62 G63 T64 F65 G66 E67 E68 K69 E70 A71 I72 F73 E74 G75 H76 I77
 M78 L79 L80 E81 D82 E83 E84 L85 E86 Q87 E88 I89 I90 A91 L92 I93 K94 D95 K96 H97 M98 T99 A100 D101
 A102 A103 A104 H105 E106 V107 I108 E109 G110 Q111 A112 S113 A114 L115 E116 E117 L118 D119 D120 E121
 Y122 L123 K124 E125 R126 A127 A128 D129 V130 R131 D132 I133 G134 K135 R136 L137 L138 R139 N140 I141
 L142 G143 L144 K145 I146 K147 D148 L149 S150 A151 I152 N153 D154 E155 V156 I157 L158 V159 A160 A161
 D162 L163 T164 P165 S166 E167 T168 A169 Q170 L171 N172 L173 K174 K175 V176 L177 G178 F179 I180 T181
 D182 A183 G184 G185 R186 T187 S188 H189 T190 S191 I192 M193 A194 R195 S196 L197 E198 L199 P200 A201
 I202 V203 G204 T205 G206 S207 V208 T209 S210 Q211 V212 K213 N214 D215 D216 Y217 L218 I219 L220 D221
 A222 V223 N224 N225 Q226 V227 Y228 V229 N230 P231 T232 N233 E234 V235 I236 D237 K238 M239 R240
 A241 V242 Q243 E244 Q245 V246 A247 A248 E249 K250 A251 E252 L253 A254 K255 L256 K257 D258 R259

Fig. 10 Ambiguously (*red*) and unambiguously (*green*) assigned residues of enzyme I. Proline residues, Met1, Arg186, Thr187, and Ser191, for which no amide resonances were given in the published data, are highlighted in *black*

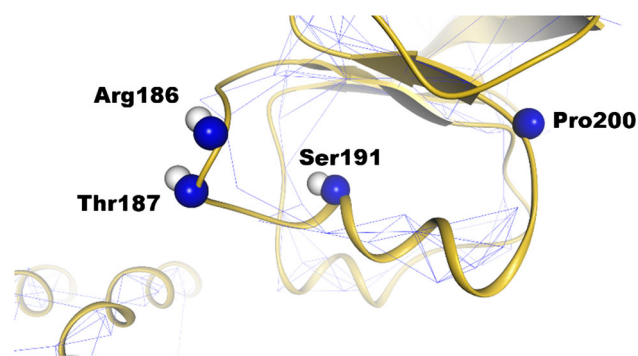


Fig. 12 Close-up of the restraint network for residues Arg186 to Pro200. The residues without amide resonance are highlighted by visualizing their amide nitrogen and hydrogen atoms

Evaluation of the algorithm's robustness against experimental uncertainties

Using the automatic determination of the volume of the reference peak leads to good assignment rates for both examples with experimental spectra. However, to estimate the tolerance with respect to experimental uncertainties and, in this way, its usefulness for other studies, the demands on the precision of the restraints derived from experiment has to be evaluated. For doing so, we compared the distances in the 3D structure of Ubiquitin with the distance restraints obtained with the optimal reference peak volume (distance for the peak with largest volume = 3.1 Å) and with a larger volume corresponding to a distance for the largest peak of 3.5 Å, which still leads to reasonable success rates. For the optimal volume, four restraints are shorter than the corresponding distances in the 3D structure but these are not considered as violated since the difference is well below the 0.2 Å safety margin. As already discussed above, changing the value for the safety margin or completely removing it would just slightly change the optimal volume of the reference peak without altering the assignment rates. More important for the estimation of the uncertainty tolerance is the distribution of differences of restraints and actual distances of the peaks, for which the restraint is longer than the real distance. Figure 13 shows that the differences are widely distributed and that even for very small distances in the 3D structure, the maximum value of 7.5 Å chosen for all peaks with a volume lower than a cutoff is assigned. Taking all peaks into account, a mean absolute deviation of 1.15 Å and a maximum deviation of 5.18 Å is obtained. With the larger volume of the reference peak, the mean absolute deviation changes to 1.68 Å. The maximum deviation does not change since it is obtained for a peak, for which the distance restraint is set to the maximum of 7.5 Å due to the very small volume.

The same analysis was repeated for enzyme I. Due to the lower quality of the spectra the authors of the original publication (Garrett et al. 1997) decided to group the restraints into four groups (strong, medium, weak and very weak) and one constant distance is used for each group. Figure 14 shows the comparison of the distance restraints with the actual distances and shows clearly the four groups (vertical line of points). For this larger protein, a slightly larger mean absolute deviation of 1.29 Å is obtained even if the maximum deviation is with 4.46 Å smaller than for Ubiquitin. In summary, the spectra of both proteins show a significant degree of uncertainty in the derived distance restraints and we expect similar or smaller deviations in similar and future studies profiting from technical improvements. It is interesting to see that the restraints derived for the larger protein enzyme I almost 20 years ago show very similar statistics as the ones obtained from the spectrum of Ubiquitin measured on state-of-the-art equipment.

Rescoring assignment based on real spectra by correlation between distances and upper bounds as well as chemical shifts

Until now, the NOE data was only used to define upper bounds for the interatomic distances, which led to the removal of many assignment possibilities. All remaining assignments fulfill all restraints and, since the algorithm does not further evaluate the agreement between the distances and the peak volumes, no ranking regarding the likelihood of the different possible assignments is available directly. In this way, an assignment in which a short and long distance is assigned to a small and large NOE peak volume, respectively, is regarded as equally likely as the reverse assignment, as long as both distances agree with both restraints. It is clear that the latter is the more reasonable assignment, a fact which can be translated into a ranking generated by comparing the exact values of the distances in the 3D structure with the distances calculated from the peak volumes. To this end, we evaluated all possible complete unambiguous assignments of Ubiquitin and enzyme I by generating the sum of the squared errors (SSE) between each distance retrieved from the 3D structure and from the NOE peak volumes. The two complete assignments with the smallest SSE for Ubiquitin are shown in Fig. 15 (additional ranks are shown in the supporting information) illustrating that this approach is indeed capable of identifying more complete and correct assignments. Since no NOE signals are seen between Glu24 and any other residue, there is no possibility to distinguish between the correct and incorrect assignment of the corresponding peak and both possibilities get the exact same score. The only other misassignment is seen for residues Leu73 and

Fig. 13 Comparison of the distance restraints obtained from the peak volumes in the 4D spectra and the corresponding distances in 3D structure of Ubiquitin. (*left*) and (*right*): The reference peak volume is calculated based on a distance for the most intense peak corresponding to 3.1 and 3.5 Å, respectively

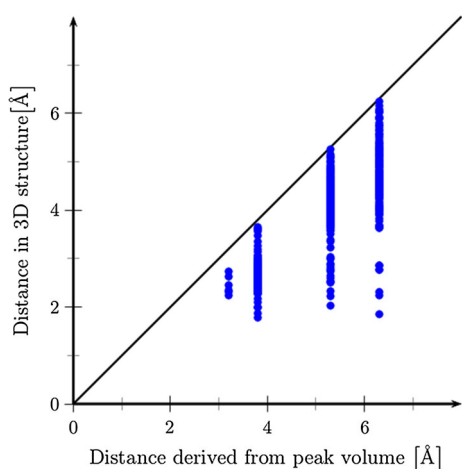
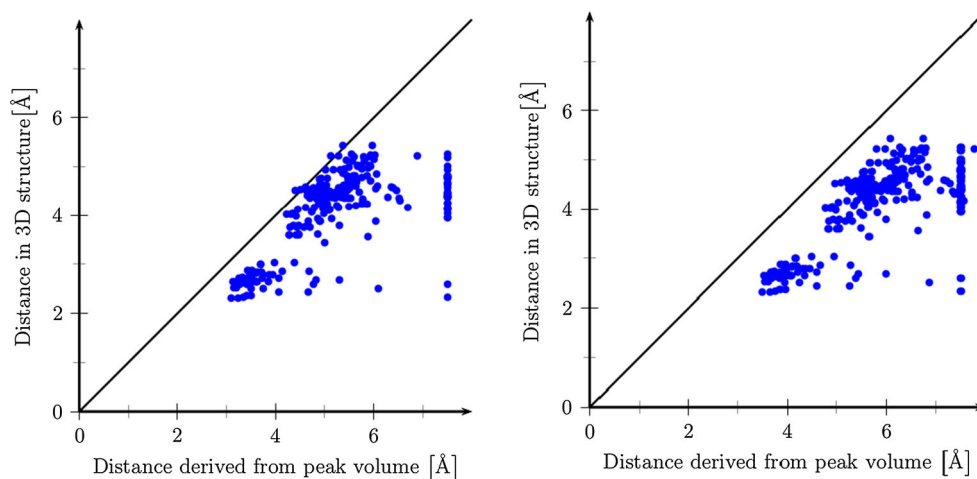


Fig. 14 Comparison of the distance restraints obtained from the peak volumes in the 4D spectra and the corresponding distances in 3D structure of enzyme I using the optimal reference peak volume

Arg74. As already discussed above, these residues belong to a hairpin with very small distances between all amide protons but overall to small peak volumes of the restraints connecting them probably due to flexibility.

For enzyme I, scoring by SSE is even more beneficial. Especially the directionality of the C-terminal α -helix (residues 234–258) is correctly predicted in all high-ranked complete unambiguous assignments (see Fig. 16 and supporting information). The first complete assignment with the wrong directionality is scored much worse and ends up on rank 792. Large parts of the poorly defined region

between residues 185 and 200 are also correctly assigned. The only remaining problems are the sequence from Gly185 to Ser191, where almost no NOE contacts are observed, as well as hairpins for the reasons discussed above.

Alternatively, the chemical shifts can be used as a criterion to re-evaluate possible complete unambiguous assignment generated by our “standard algorithm”. In contrast to the distance-based SSE, the evaluation by comparing experimental and predicted chemical shifts is orthogonal to the first assignment since the actual values were never used in the optimization procedure (except for defining groups of restraints belonging to the same nucleus). Even if the accuracy of state-of-the-art chemical shift prediction tools, either empirical (Han et al. 2011; Neal et al. 2003; Osapay and Case 1991; Shen and Bax 2010; Wishart et al. 1997; Xu and Case 2001) or quantum chemical (Exner et al. 2012; Frank et al. 2011, 2012; He et al. 2009; Jacob and Visscher 2006; Lee and Bettens 2007; Oldfield 2002), is not high enough for a complete automated assignment, they should be helpful for the largely simplified task of ranking a small number of pre-defined assignments. In most cases, the algorithm has only to distinguish between pairs or triples of residues (see e.g. Fig. 8) and there is a high degree of dependence among neighboring pairs/triples. For the C-terminal α -helix of enzyme I, e.g., it should be enough to fix the assignment for one residues since the neighboring residues will then be defined unambiguously by the NOE network. For proving

527.0 M1 Q2 I3 F4 V5 K6 T7 L8 T9 G10 K11 T12 I13 T14 L15 E16 V17 E18 P19 S20 D21 T22 I23 G53 N25 V26 K27 A28 K29 I30 Q31 D32 K33 E34 G35 I36 P37 P38 D39 Q40 Q41 R42 L43 I44 F45 A46 G47 K48 Q49 L50 E51 D52 R54 T55 L56 S57 D58 Y59 N60 I61 Q62 K63 E64 S65 T66 L67 H68 L69 V70 L71 R72 R74 L73 G75 G76
527.0 M1 Q2 I3 F4 V5 K6 T7 L8 T9 G10 K11 T12 I13 T14 L15 E16 V17 E18 P19 S20 D21 T22 I23 E24 N25 V26 K27 A28 K29 I30 Q31 D32 K33 E34 G35 I36 P37 P38 D39 Q40 Q41 R42 L43 I44 F45 A46 G47 K48 Q49 L50 E51 D52 R54 T55 L56 S57 D58 Y59 N60 I61 Q62 K63 E64 S65 T66 L67 H68 L69 V70 L71 R72 R74 L73 G75 G76

Fig. 15 Correctly (*green*) and incorrectly (*red*) assigned residues of Ubiquitin in the two highest ranked complete assignments according to the sum of the squared errors between distances retrieved from the 3D structure and from the NOE peak volumes

1989.3 M1 T187 S3 G4 L6 I5 A7 S8 P9 G10 I11 A12 F13 G14 K15 A16 L17 L18 L19 K20 E21 D22 E23 I24 V25 I26 D27 R28 K29 K30 I31 S32 A33 D34 Q35 V36 D37 Q38 E39 V40 E41 R42 F43 L44 S45 G46 R47 A48 K49 A50 S51 A52 Q53 L54 E55 T56 I57 K58 T59 K60 A61 G62 E63 T64 F65 G66 E67 E68 K69 E70 A71 I72 F73 E74 G75 H76 I77 M78 L79 L80 E81 D82 L85 E83 E84 E86 Q87 E88 I89 I90 A91 L92 I93 K94 D95 K96 H97 M98 T99 A100 D101 A102 A103 A104 H105 E106 V107 I108 E109 G110 Q111 A112 S113 A114 L115 E116 E117 L118 D119 D120 E121 Y122 L123 K124 E125 R126 A127 D129 V130 A128 I131 D132 I133 G134 K135 R136 L137 L138 R139 N140 I141 L142 G143 L144 K145 I146 I147 D148 L149 S150 A151 I152 Q153 D154 E155 V156 I157 L158 V159 A160 A161 D162 L163 T164 P165 S166 E167 T168 A169 Q170 L171 N172 L173 K174 K175 V176 L177 G178 F179 I180 T181 D182 A183 G185 R186 T187 S189 I189 S188 T190 S191 I192 M193 A194 R195 S196 L197 E198 L199 P200 A201 I202 V203 G204 T205 G206 S207 V208 T209 S210 Q211 V212 K213 N214 D215 D216 Y217 L218 I219 L220 D221 A222 V223 N225 N224 Q226 V227 Y228 V229 N230 P231 T232 N233 E234 V235 I236 D237 K238 M239 R240 A241 V242 Q243 E244 Q245 V246 S248 A247 E249 K250 A251 E252 L253 A254 K255 L256 K257 R259 D258

1989.3 M1 R186 S3 G4 L6 I5 A7 S8 P9 G10 I11 A12 F13 G14 K15 A16 L17 L18 L19 K20 E21 D22 E23 I24 V25 I26 D27 R28 K29 K30 I31 S32 A33 D34 Q35 V36 D37 Q38 E39 V40 E41 R42 F43 L44 S45 G46 R47 A48 K49 A50 S51 A52 Q53 L54 E55 T56 I57 K58 T59 K60 A61 G62 E63 T64 F65 G66 E67 E68 K69 E70 A71 I72 F73 E74 G75 H76 I77 M78 L79 L80 E81 D82 L85 E83 E84 E86 Q87 E88 I89 I90 A91 L92 I93 K94 D95 K96 H97 M98 T99 A100 D101 A102 A103 A104 H105 E106 V107 I108 E109 G110 Q111 A112 S113 A114 L115 E116 E117 L118 D119 D120 E121 Y122 L123 K124 E125 R126 A127 D129 V130 A128 I131 D132 I133 G134 K135 R136 L137 L138 R139 N140 I141 L142 G143 L144 K145 I146 I147 D148 L149 S150 A151 I152 Q153 D154 E155 V156 I157 L158 V159 A160 A161 D162 L163 T164 P165 S166 E167 T168 A169 Q170 L171 N172 L173 K174 K175 V176 L177 G178 F179 I180 T181 D182 A183 G185 R186 T187 S189 I189 S188 T190 S191 I192 M193 A194 R195 S196 L197 E198 L199 P200 A201 I202 V203 G204 T205 G206 S207 V208 T209 S210 Q211 V212 K213 N214 D215 D216 Y217 L218 I219 L220 D221 A222 V223 N225 N224 Q226 V227 Y228 V229 N230 P231 T232 N233 E234 V235 I236 D237 K238 M239 R240 A241 V242 Q243 E244 Q245 V246 S248 A247 E249 K250 A251 E252 L253 A254 K255 L256 K257 R259 D258

Fig. 16 Correctly (green) and incorrectly (red) assigned residues of enzyme I in the two highest ranked complete assignments according to the sum of the squared errors between distances retrieved from the 3D structure and from the NOE peak volumes. The two complete

assignments only differ in the assignment of the first resonance, which does not show any NOE contacts and, therefore, cannot be predicted correctly

1230.3 M1 Q2 I3 F4 V5 K6 T7 L8 T9 G10 K11 T12 I13 T14 L15 E16 V17 E18 P19 S20 D21 T22 I23 E24 N25 V26 K27 A28 K29 I30 Q31 D32 K33 E34 G35 I36 P37 P38 D39 Q40 Q41 R42 L43 I44 F45 A46 G47 K48 Q49 L50 E51 D52 R54 T55 L56 S57 D58 Y59 N60 I61 Q62 K63 E64 S65 T66 L67 H68 L69 V70 L71 R72 L73 R74 G75 G76

1240.2 M1 Q2 I3 F4 V5 K6 T7 L8 T9 G10 K11 T12 I13 T14 L15 E16 V17 E18 P19 S20 D21 T22 I23 E24 N25 V26 K27 A28 K29 I30 Q31 D32 K33 E34 G35 I36 P37 P38 D39 Q40 Q41 R42 L43 I44 F45 A46 G47 K48 Q49 L50 E51 D52 R54 T55 L56 S57 D58 Y59 N60 I61 Q62 K63 E64 S65 T66 L67 H68 L69 V70 L71 R72 R74 L73 G75 G76

Fig. 17 Correctly (green) and incorrectly (red) assigned residues of Ubiquitin in the two highest ranked assignments according to comparison of experimental and predicted chemical shifts

4933.0 M1 H189 S3 G4 I5 L6 A7 S8 P9 G10 I11 A12 F13 G14 K15 A16 L17 L18 L19 K20 E21 D22 E23 I24 V25 I26 D27 R28 K29 K30 I31 S32 A33 D34 Q35 V36 D37 Q38 E39 V40 E41 R42 F43 L44 S45 G46 R47 A48 K49 A50 S51 A52 Q53 L54 E55 T56 I57 K58 T59 K60 A61 G62 E63 T64 F65 G66 E67 E68 K69 E70 A71 I72 F73 E74 G75 H76 I77 M78 L79 L80 E81 D82 E83 E84 L85 E86 Q87 E88 I89 I90 A91 L92 I93 K94 D95 K96 H97 M98 T99 A100 D101 A102 A103 A104 H105 E106 V107 I108 E109 G110 Q111 A112 S113 A114 L115 E116 E117 L118 D119 D120 E121 Y122 L123 K124 E125 R126 A127 D129 V130 A128 I131 D132 I133 G134 K135 R136 L137 L138 R139 N140 I141 L142 G143 L144 K145 I146 I147 D148 L149 S150 A151 I152 Q153 D154 E155 V156 I157 L158 V159 A160 A161 D162 L163 T164 P165 S166 E167 T168 A169 Q170 L171 N172 L173 K174 K175 V176 L177 G178 F179 I180 T181 D182 A183 G184 G185 R186 T187 S188 S191 T190 S191 I192 M193 A194 R195 S196 L197 E198 L199 P200 A201 I202 V203 G204 T205 G206 S207 V208 T209 S210 Q211 V212 K213 N214 D215 D216 Y217 L218 I219 L220 D221 A222 V223 N224 N225 Q226 V227 Y228 V229 N230 P231 T232 N233 E234 V235 I236 D237 K238 M239 R240 A241 V242 Q243 E244 Q245 V246 A247 S248 E249 K250 A251 E252 L253 A254 K255 L256 K257 D258 R259

4933.0 M1 H189 S3 G4 I5 L6 A7 S8 P9 G10 I11 A12 F13 G14 K15 A16 L17 L18 L19 K20 E21 D22 E23 I24 V25 I26 D27 R28 K29 K30 I31 S32 A33 D34 Q35 V36 D37 Q38 E39 V40 E41 R42 F43 L44 S45 G46 R47 A48 K49 A50 S51 A52 Q53 L54 E55 T56 I57 K58 T59 K60 A61 G62 E63 T64 F65 G66 E67 E68 K69 E70 A71 I72 F73 E74 G75 H76 I77 M78 L79 L80 E81 D82 E83 E84 L85 E86 Q87 E88 I89 I90 A91 L92 I93 K94 D95 K96 H97 M98 T99 A100 D101 A102 A103 A104 H105 E106 V107 I108 E109 G110 Q111 A112 S113 A114 L115 E116 E117 L118 D119 D120 E121 Y122 L123 K124 E125 R126 A127 D129 V130 A128 I131 D132 I133 G134 K135 R136 L137 L138 R139 N140 I141 L142 G143 L144 K145 I146 I147 D148 L149 S150 A151 I152 Q153 D154 E155 V156 I157 L158 V159 A160 A161 D162 L163 T164 P165 S166 E167 T168 A169 Q170 L171 N172 L173 K174 K175 V176 L177 G178 F179 I180 T181 D182 A183 G184 G185 R186 T187 S188 S191 T190 S191 I192 M193 A194 R195 S196 L197 E198 L199 P200 A201 I202 V203 G204 T205 G206 S207 V208 T209 S210 Q211 V212 K213 N214 D215 D216 Y217 L218 I219 L220 D221 A222 V223 N224 N225 Q226 V227 Y228 V229 N230 P231 T232 N233 E234 V235 I236 D237 K238 M239 R240 A241 V242 Q243 E244 Q245 V246 A247 S248 E249 K250 A251 E252 L253 A254 K255 L256 K257 D258 R259

Fig. 18 Correctly (green) and incorrectly (red) assigned residues of enzyme I in the highest ranked assignment according to comparison of experimental and predicted chemical shifts

this assumption, we used the structure-based component (SHIFTX +) of the empirical SHIFTX2 method (Han et al. 2011) to predict the chemical shifts of Ubiquitin and enzyme I (the knowledge-based component SHIFTY (Wishart et al. 1997) was not applied since Ubiquitin was part of the training dataset of the method). The score for an assignment is then calculated as the sum of ^{15}N deviations plus ten-times the sum of the ^1H deviations to account for the different chemical shift ranges of these two types of nuclei. The best scored assignments are shown for Ubiquitin and enzyme I in Figs. 17 and 18, respectively. The mean absolute error of the predicted chemical shifts of Ubiquitin compared to the correct assignment are 1.428 and 0.2514 ppm for ^{15}N and ^1H , respectively. This clearly demonstrates that distinguishing two residues of the same or similar type, as it is needed when trying a complete chemical-shift-based assignment, is not possible. In

contrast, all ambiguities remaining after the NOE-based assignment can be resolved completely. Thus, the ranking criterion based on chemical shifts is even superior to the NOE-distance-based measure, for which the assignment of residues Leu73 and Arg74 are flipped in the best rank. The same flip is seen on rank 2 according to chemical shifts but with a significant worse score compared to rank 1. The comparison of chemical shifts also leads to an improvement for enzyme I even compared to the NOE-distance-based criterion. Only three groups of nuclei (Ile5/Leu6, Ala128/Asp129/Val130, Ile2/His189/Ser191) are incorrect in the best four, almost identically ranked assignments. Since these are short independent patches composed of similar residues and, especially in the last case, still highly ambiguous after the NOE-based assignment due to the sparse NOE network, a correct and unambiguous assignment based on 4D [^{15}N , ^1H]-HSQC-NOESY- ^{15}N , ^1H]-

Table 4 Ranking of individual complete unambiguous assignments according to the sum of the squared errors (SSE) between distances in the 3D structure and restraints as well as differences between predicted and experimental chemical shifts and the number of wrongly assignment residues

Rank	Score	No. of wrong assignments
<i>Ubiquitin—SSE between distances and restraints</i>		
1	527.0	3
2	527.0	2
3	531.6	5
4	531.6	4
5	539.4	3
6	539.4	4
7	541.5	4
8	541.5	5
9	544.0	5
10	544.0	6
<i>Ubiquitin—difference in chemical shifts</i>		
1	1230.3	0
2	1240.2	2
3	1270.4	2
4	1271.2	1
5	1280.4	4
<i>Enzyme I—SSE between distances and restraints</i>		
1	1989.3	18
2	1989.3	18
3	1989.3	18
4	1989.3	17
5	1990.7	16
6	1990.7	20
7	1990.7	20
8	1990.7	19
9	1992.7	17
<i>Enzyme I—difference in chemical shifts</i>		
1	4933.0	4
2	4933.0	5
3	4933.0	6
4	4933.0	7
5	4933.1	6
6	4933.1	6
7	4933.1	7
8	4933.1	7
9	4933.8	7
10	4933.8	7

HSQC spectra alone is probably impossible. Nevertheless, with an overall unambiguous-assignment success rate of 95 % the results can still be regarded as perfect when taking the size of the protein into account. One important fact is that the wrong ordering of the C-terminal α -helix is

now completely ruled out illustrated by its first occurrence on rank 315,454.

All rescoring results are summarized in Table 4 giving the highest ranked assignments with the corresponding scores (sum of the squared errors (SSE) between distances in the 3D structure and restraints or differences between predicted and experimental chemical shifts) and numbers of wrongly assigned residues.

Conclusion

In this paper we showed that 4D [^1H , ^{15}N]-HSQC-NOESY- [^1H , ^{15}N]-HSQC spectra contain the information needed for the automatic assignment of chemical shifts of amide groups based on a 3D protein structure. Starting from a complete ambiguous assignment, assignments of a specific group can be removed by comparing distances in the structures with distance restraints generated from the peak volumes in the spectrum. With experiments based on artificial datasets we could demonstrate that the algorithm can tolerate some uncertainties in the volumes of the NOESY peaks and, even more importantly, a large number of missing peaks. With the two real-world examples, Ubiquitin and enzyme I, high percentages of unambiguously assigned residues were obtained. Considering the experimental uncertainty in the spectra, i.e. for enzyme I only less than 90 % of the theoretically possible restraints are seen in the spectra and the restraints are on average 1.2 Å longer than the corresponding distances in the 3D structure and the maximum deviation is up to almost 4.5 Å, our results document the usefulness of the approach for real applications. Again taking enzyme I as an example, 191 of the 259 residues were assigned unambiguously. For the remaining 60 residues, for which HSQC peaks are expected, 2–11 possible assignments remain. To get a feeling of the reduction of possibilities, we calculated the number of possible complete unambiguous assignments at different stages. At the beginning, $251! = 8.1 \times 10^{495}$ unambiguous assignment possibilities are included in the completely ambiguous initial assignment. This is reduced to 8.7×10^{22} if one takes all combinatorial possible combinations of the remaining ambiguous residues after the optimization (step 4). However, because of the interdependence of the individual assignments, only around 4.6×10^6 complete assignments fulfil all distance restraints, from which many can be removed by using just a little bit of additional information. This information can come, on the one hand, from experiments, in which the chemical shifts of a small number of selected residues is determined e.g. by mutations. Selecting the most beneficial residues can be done on the basis of the automatic assignment. On the other hand, chemical shift predictions,

even with the high uncertainty of state-of-the-art empirical or ab initio methods, can e.g. remove the uncertainty in the directionality of the long, structurally separated α -helix and other longer ambiguous patches like Arg186 to Pro200 in enzyme I.

Irrespective of the good performance just described, a reduction of the success rates can be expected, when lower-quality spectra, i.e. because of limited concentration of the protein, are used, the spectra show extensive chemical shift overlap and/or exchange broadening. In these cases, it might be possible to support the algorithm by providing single pre-assigned residues, which could be obtained e.g. by mutation experiments. As shown above, this was very successful in the case of adding large noise to the artificial spectra. Additionally, we also plan to integrate the rescoring options more effectively in the workflow. For example, chemical shift calculations could be used at different stages to exclude possible assignments starting with identifying residues with extreme chemical shifts or group the residues according to their amino acid type based on chemical shifts even before the reduction of assignment option based on NOE cross peaks.

There are also other ways for further improvements. (1) Using additional spectra like ^{13}C NOESY or CB spectra, it should be possible to arrive at higher unambiguous assignment rates. After assigning the amide chemical shifts with the approach described here, these spectra could be included to optimize the amide assignments and get the carbon assignments of the backbone and side chains. However, for this ^{13}C labeled protein is needed making the measurements more expensive. Additionally, a similar approach can be developed for specifically labeled proteins, in which e.g. only methyl groups are ^{13}C labeled. In this way, also very large proteins could be analyzed for which ^{15}N NOESY spectra are not interpretable anymore. (2) At the moment, overlapped peaks have to be identified before performing the assignment. As described above, these can be identified using 3D HNCO or TROSY-HNCO spectra but ^{13}C labeling is needed. One other possibility is to give only the number of overlapped peaks to the algorithm and let it decide which peak is an overlapped peak using the distance restraints. (3) Finally, one point not addressed explicitly until now are disagreements between the predicted structure and the 4D spectrum resulting from the different sample preparation. E.g. the distances in the X-ray structure of a flexible loop in a protein might be larger than those in the ensemble measured by the NMR experiment. Since the distance restraints are violated even in the correct assignment, our approach cannot generate any assignment and will stop with an error. To circumvent such failures, we will try if reasonable assignment success rates can still be obtained when a number of violations are allowed. Additionally, methods to identify such

problematic regions will be developed, which will then be removed from the list of residues to be assigned. All these extensions will be made available as soon as possible on the project's website (<https://github.com/thomasexner/4Dassign>) for testing and additional contributions from the scientific community.

The current algorithm can directly handle 4D spectra of perdeuterated proteins extending its applicability to larger proteins. In perdeuterated proteins NOEs corresponding to larger amide amide distances ($>7 \text{ \AA}$) are visible. This should lead to a better performance since larger, more unique NOE networks can be analyzed. It is furthermore expected that recent developments for obtaining more precise NOEs (Vogeli et al. 2009) and higher resolution from non-uniform sampling (e.g. Luan et al. 2005) will have a positive impact on automatic assignment success rates. Needless to say that NOESY spectra collected at higher fields leading to increased dispersion (and TROSY effects) will lead to improved performance of the assignment approach described here.

Supporting information

Assignment possibilities after each step of the automated procedure based on the back-calculated spectra of Ubiquitin, ambiguously assigned HSQC signals for enzyme I and results after rescoring by the agreement of the squared distances and the chemical shifts are provided as supporting information.

Acknowledgments We thank Dr. Remco Sprangers for providing the 4D spectra of Ubiquitin and Prof. G. Marius Clore for the chemical shifts of the amide groups of enzyme I. This work was supported by the German Research Foundation (DFG) [EX15/17-1 to T.E.E.].

References

- Berjanskii M, Wishart DS (2006) NMR: prediction of protein flexibility. *Nat Protoc* 1:683–688
- Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112:535–542
- Chao FA, Shi L, Masterson LR, Veglia G (2012) FLAMEnGO: a fuzzy logic approach for methyl group assignment using NOESY and paramagnetic relaxation enhancement data. *J Magn Reson* 214:103–110
- Chao FA, Kim JG, Xia YL, Milligan M, Rowe N, Veglia G (2014) FLAMEnGO 2.0: an enhanced fuzzy logic algorithm for structure-based assignment of methyl group resonances. *J Magn Reson* 245:17–23
- Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J Am Chem Soc* 120:6836–6837

- Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6:277–293
- Exner TE, Frank A, Onila I, Moller HM (2012) Toward the quantum chemical calculation of NMR chemical shifts of proteins. 3. Conformational sampling and explicit solvents model. *J Chem Theory Comput* 8:4818–4827
- Fesik SW, Shuker SB, Hajduk PJ, Meadows RP (1997) SAR by NMR: an NMR-based approach for drug discovery. *Protein Eng* 10:73
- Frank A, Onila I, Moller HM, Exner TE (2011) Toward the quantum chemical calculation of nuclear magnetic resonance chemical shifts of proteins. *Proteins: Struct. Funct. Bioinf* 79:2189–2202
- Frank A, Moller HM, Exner TE (2012) Toward the quantum chemical calculation of NMR chemical shifts of proteins. 2. Level of theory, basis set, and solvents model dependence. *J Chem Theory Comput* 8:1480–1492
- Garrett DS, Seok YJ, Liao DI, Peterkofsky A, Gronenborn AM, Clore GM (1997) Solution structure of the 30 kDa N-terminal domain of enzyme I of the *Escherichia coli* phosphoenolpyruvate:sugar phosphotransferase system by multidimensional NMR. *Biochemistry* 36:2517–2530
- Gobl C, Tjandra N (2012) Application of solution NMR spectroscopy to study protein dynamics. *Entropy* 14:581–598
- Hajduk PJ (2006) SAR by NMR: putting the pieces together. *Mol Interv* 6:266–272
- Han B, Liu YF, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. *J Biomol NMR* 50:43–57
- He X, Wang B, Merz KM (2009) Protein NMR chemical shift calculations based on the automated fragmentation QM/MM approach. *J Phys Chem B* 113:10380–10388
- Herrmann T, Guntert P, Wuthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol* 319:209–227
- Hus JC, Prompers JJ, Bruschiweiler R (2002) Assignment strategy for proteins with known structure. *J Magn Reson* 157:119–123
- Jacob CR, Visscher L (2006) Calculation of nuclear magnetic resonance shieldings using frozen-density embedding. *J Chem Phys* 125:194104
- Jang R, Gao X, Li M (2012) Combining automated peak tracking in SAR by NMR with structure-based backbone assignment from N-15-NOESY. *BMC Bioinf* 13(Suppl 3):S4
- Jung YS, Zweckstetter M (2004) Backbone assignment of proteins with known structure using residual dipolar couplings. *J Biomol NMR* 30:25–35
- Jung YS, Sharma M, Zweckstetter M (2004) Simultaneous assignment and structure determination of protein backbones by using NMR dipolar couplings. *Angew Chem Int Ed Engl* 43:3479–3481
- Kay LE (2005) NMR studies of protein structure and dynamics. *J Magn Reson* 173:193–207
- Keller RLJ (2004a) The computer aided resonance assignment tutorial. Cantina Verlag, Goldau
- Keller RLJ (2004b) Optimizing the process of nuclear magnetic resonance spectrum analysis and computer aided resonance assignment. Ph.D., ETH Zürich
- Kleckner IR, Foster MP (2011) An introduction to NMR-based approaches for measuring protein dynamics. *Biochim Biophys Acta Proteins Proteomics* 1814:942–968
- Lee AM, Bettens RPA (2007) First principles NMR calculations by fragmentation. *J Phys Chem A* 111:5111–5115
- Linge JP, Habeck M, Rieping W, Nilges M (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* 19:315–316
- Luan T, Jaravine V, Yee A, Arrowsmith CH, Orekhov VY (2005) Optimization of resolution and sensitivity of 4D NOESY using multi-dimensional decomposition. *J Biomol NMR* 33:1–14
- Mittermaier A, Kay LE (2006) Review—New tools provide new insights in NMR studies of protein dynamics. *Science* 312:224–228
- Neal S, Nip AM, Zhang HY, Wishart DS (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. *J Biomol NMR* 26:215–240
- Oldfield E (2002) Chemical shifts in amino acids, peptides, and proteins: from quantum chemistry to drug design. *Annu Rev Phys Chem* 53:349–378
- Osapay K, Case DA (1991) A new analysis of proton chemical-shifts in proteins. *J Am Chem Soc* 113:9436–9444
- Pristovsek P, Ruterjans H, Jerala R (2002) Semiautomatic sequence-specific assignment of proteins based on the tertiary structure—the program st2nmr. *J Comput Chem* 23:335–340
- Rieping W, Habeck M, Bardiaux B, Bernard A, Malliavin TE, Nilges M (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics* 23:381–382
- Shen Y, Bax A (2010) SPARTA plus: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR* 48:13–22
- Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274:1531–1534
- Sprangers R, Kay LE (2007a) Probing supramolecular structure from measurement of methyl H-1-C-13 residual dipolar couplings. *J Am Chem Soc* 129:12668–12669
- Sprangers R, Kay LE (2007b) Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature* 445:618–622
- Stratmann D, van Heijenoort C, Guittet E (2009) NOE-net—Use of NOE networks for NMR resonance assignment of proteins with known 3D structure. *Bioinformatics* 25:474–481
- Stratmann D, Guittet E, van Heijenoort C (2010) Robust structure-based resonance assignment for functional protein studies by NMR. *J Biomol NMR* 46:157–173
- 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, Kent Wenger R, Yao H, Markley JL (2008) BioMagResBank. *Nucl Acids Res* 36:D402–D408
- Venditti V, Fawzi NL, Clore GM (2011) Automated sequence- and stereo-specific assignment of methyl-labeled proteins by paramagnetic relaxation and methyl-methyl nuclear overhauser enhancement spectroscopy. *J Biomol NMR* 51:319–328
- Vijaykumar S, Bugg CE, Cook WJ (1987) Structure of ubiquitin refined at 1.8 Å resolution. *J Mol Biol* 194:531–544
- Vogeli B, Segawa TF, Leitz D, Sobol A, Choutko A, Trzesniak D, van Gunsteren W, Riek R (2009) Exact distances and internal dynamics of perdeuterated ubiquitin from NOE buildups. *J Am Chem Soc* 131:17215–17225
- Walker CA, Hinderhofer M, Witte DJ, Boos W, Moller HM (2008) Solution structure of the soluble domain of the NfeD protein YuaF from *Bacillus subtilis*. *J Biomol NMR* 42:69–76
- Wishart DS, Watson MS, Boyko RF, Sykes BD (1997) Automated H-1 and C-13 chemical shift prediction using the BioMagResBank. *J Biomol NMR* 10:329–336
- Xu XP, Case DA (2001) Automated prediction of (15)N, (13)C(alpha), (13)C(beta) and (13)C ' chemical shifts in proteins using a density functional database. *J Biomol NMR* 21:321–333
- Xu YQ, Matthews S (2013) MAP-XSII: an improved program for the automatic assignment of methyl resonances in large proteins. *J Biomol NMR* 55:179–187
- Xu YQ, Liu MH, Simpson PJ, Isaacson R, Cota E, Marchant J, Yang DW, Zhang XD, Freemont P, Matthews S (2009) Automated assignment in selectively methyl-labeled proteins. *J Am Chem Soc* 131:9480–9481