J Biomol NMR (2014) 59:75–86
DOI 10.1007/s10858-014-9828-0

ARTICLE

An automated framework for NMR resonance assignment through simultaneous slice picking and spin system forming

Ahmed Abbas · Xianrong Guo · Bing-Yi Jing · Xin Gao

Received: 3 February 2014 / Accepted: 5 April 2014 / Published online: 19 April 2014

Abstract Despite significant advances in automated nuclear magnetic resonance-based protein structure determination, the high numbers of false positives and false negatives among the peaks selected by fully automated methods remain a problem. These false positives and negatives impair the performance of resonance assignment methods. One of the main reasons for this problem is that the computational research community often considers peak picking and resonance assignment to be two separate problems, whereas spectroscopists use expert knowledge to pick peaks and assign their resonances at the same time. We propose a novel framework that simultaneously conducts slice picking and spin system forming, an essential step in resonance assignment. Our framework then employs a genetic algorithm, directed by both connectivity information and amino acid typing information from the spin systems, to assign the spin systems to residues. The inputs to our framework can be as few as two commonly used spectra, i.e., CBCA(CO)NH and HNCACB. Different from the existing peak picking and resonance assignment methods that treat peaks as the units, our method is based on 'slices', which are one-dimensional vectors in three-dimensional spectra that correspond to certain (N, H) values. Experimental results on both benchmark simulated data sets and four real protein data sets demonstrate that our method significantly outperforms the state-of-the-art methods while using a less number of spectra than those methods. Our method is freely available at http://sfb.kaust.edu.sa/Pages/Software.aspx.

Keywords Resonance assignment · Peak picking · Spin system · Wavelet

Electronic supplementary material The online version of this article (doi:10.1007/s10858-014-9828-0) contains supplementary material, which is available to authorized users.

A. Abbas · X. Gao (✉)
Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
e-mail: xin.gao@kaust.edu.sa

X. Guo
Imaging and Characterization Core Lab, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

B.-Y. Jing
Department of Mathematics, Hong Kong University of Science and Technology, Kowloon, Hong Kong

Introduction

Up to now, the structures of most known protein structures were delineated by two techniques, X-ray crystallography and nuclear magnetic resonance (NMR) (Berman et al. 2000). In contrast to X-ray crystallography that requires proteins to be crystallized, NMR can study protein structures in solvent and provide unique information about the dynamics of the proteins. NMR protein structure determination is a multi-step procedure (Wüthrich 1986), that consists of spectrum generation, which generates through-bond and through-space spectra, peak picking, which identifies coupling signals from the spectra, resonance assignment, which assigns chemical shifts from the peaks to corresponding atoms of the protein, nuclear Overhauser enhancement (NOE) assignment, which determines distance constraints, and structure calculation, which calculates the final structures (ensemble).

The traditional NMR protein structure determination process has mainly been accomplished by manual or semiautomated data processing, with the help of visualization and analysis tools (Johnson and Blevins 1994; Delaglio et al. 1995; Günther et al. 2000; Vranken et al. 2005; Goddard and Kneller 2007). This process is costly and time-consuming. Various computational methods have therefore been developed to automate each step of this tedious process (Herrmann et al. 2002; Altieri and Byrd 2004; Gronwald and Kalbitzer 2004; Dancea and Güntert 2005; Takeda et al. 2007; Güntert 2009; Alipanahi et al. 2009; Ikeya et al. 2009; Alipanahi et al. 2011; Jang et al. 2011; Gao 2012, 2013). Among the aforementioned five steps, peak picking and resonance assignment require advanced computational techniques. For well-structured small- and medium-size proteins with high signal-to-noise ratios, a number of automated peak picking (Garret et al. 1991; Carrara et al. 1993; Antz et al. 1995; Koradi et al. 1998; Korzhneva et al. 2001; Liu et al. 2012; Abbas et al. 2013; Cheng et al. 2014) and resonance assignment methods (Güntert et al. 2000; Coggins and Zhou 2003; Jung and Zweckstetter 2004; Lin et al. 2005; Wan and Lin 2007; Masse and Keller 2005; Lemak et al. 2008; Volk et al. 2008; Jang et al. 2010; Tycko and Hu 2010; Alipanahi et al. 2011; Zeng et al. 2011; Jang et al. 2011) have been shown to work well.

Peak picking and resonance assignment are often treated as two separate problems. AUTOPSY is among the pioneering works in automated peak picking (Koradi et al. 1998). AUTOPSY first estimates the noise level of the given spectrum by assuming that the noise is white Gaussian. It then eliminates all the data points that have intensity values lower than the estimated noise level. The spectrum is then approximated from the outer product of one-dimensional line shapes and peaks are predicted inside each line shape. PICKY was developed to improve the performance of AUTOPSY (Alipanahi et al. 2009). PICKY adopts a similar idea to estimate the noise level and removes most of the data points that do not contain signals. In contrast to AUTOPSY, PICKY decomposes the remaining spectrum into small ‘components’, each of which contains only one or a few peaks. Singular value decomposition is then applied to each component to identify the peaks. A multi-step refinement is developed to refine the predicted peaks. WaVPeak was recently proposed based on the ‘soft-thresholding’ idea (Liu et al. 2012). Different from AUTOPSY and PICKY, which remove many data points from the spectrum, WaVPeak smoothes the spectrum by wavelet decomposition and reconstructs the spectrum without removing any data point, thus eliminating the risk of removing true but weak peaks. WaVPeak further estimates the volume of each predicted peak to identify the true peaks. The peak volume was demonstrated to be more powerful than the intensity values. The peak selection problem was further cast as a multiple testing problem and a Benjamini–Hochberg-based algorithm was proposed to determine automatically how many predicted peaks to return to the user (Abbas et al. 2013).

The resonance assignment problem has long been a target of the computational research community due to its clear formulation. Given peaks identified from through-bond spectra that share the same ‘root’ *(N, H)*, such as ^{15}N-HSQC, CBCA(CO)NH, and HNCACB, the goal is to link and assign the peaks along the amino acid chain of the protein by using inter- and intra-residue information. There are various existing assignment methods, including search algorithms (Bartels et al. 1996; Zimmerman et al. 1997; Coggins and Zhou 2003; Volk et al. 2008; Lemak et al. 2008), maximum independent set algorithms (Wu et al. 2006), sequential algorithms (Wan and Lin 2007; Tycko and Hu 2010), logic algorithms (Masse and Keller 2005), fragment-based algorithms (Güntert et al. 2000; Jung and Zweckstetter 2004) and optimization algorithms (Zimmerman et al. 1997; Lin et al. 2005; Alipanahi et al. 2011; Jang et al. 2010, 2011, 2012). AUTOASSIGN (Zimmerman et al. 1997), one of the widely used assignment method, requires peak lists of the 2D ^{15}N-HSQC spectrum and a number of 3D spectra. It formulates the assignment problem as a constraint satisfaction problem in artificial intelligence and encodes information from a domain-knowledge base to enhance the accuracy. MARS was proposed as a consensus-based assignment method that makes both local and global assignments (Jung and Zweckstetter 2004). Local assignment is done through the identification of reliably connected fragments, whereas global assignment is done through the optimization of a pseudo-energy function. GANA, which uses a genetic algorithm to find good assignments, was later proposed (Lin et al. 2005). The key part in GANA is the fitness function, which determines how computer-simulated evolution proceeds. GANA uses connectivity information between spin systems to build the fitness function. IPASS was recently proposed. It formulates resonance assignment as an integer linear programming problem (Alipanahi et al. 2011). The spin systems are formed by peaks from ^{15}N-HSQC, CBCA(CO)NH and HNCACB and the connectivity information is extracted from both through-bond spectra and the N-NOESY spectrum. To reduce the search space, a probabilistic model based on BioMagResBank (BMRB) (Seavey et al. 1991) was developed to estimate how likely it is that a particular spin system is found for a certain residue position. Experimental results on automatically picked peaks by PICKY on four real protein NMR data sets demonstrated that IPASS was more error tolerant than were the other assignment methods and that it could work reasonably well with peaks chosen by automated methods.

Despite the significant advances in the development of peak picking and resonance assignment methods, such methods still do not reduce the required manual work nor can they be practically applied in NMR labs. The main reason is that these methods lose significant information. One source of this loss of information is that almost all the peak picking methods consider peaks as independent signals, whereas many of them are indeed coupled. For instance, in CBCA(CO)NH, if we pick a peak at (N_i, C_i, H_i), it is very likely that there is another peak in the (N_i, H_i) slice, unless the previous residue is a glycine or the signal is absent from the spectrum. Similar reasoning is valid for HNCACB as well. Another source of loss of information is that peak picking and resonance assignment are usually considered as two consecutive steps. Once peak picking is done, resonance assignment methods assign those peaks regardless of their quality. That is why most resonance assignment methods perform poorly with automatically picked peaks. However, the assignment step should provide feedback information to direct the peak picking or to improve the peak quality.

In this paper, we propose a framework that simultaneously picks slices and forms them into spin systems for assignment purposes. The inputs to our framework can be as few as two spectra, i.e., CBCA(CO)NH and HNCACB, in the UCSF format. If more spectra are available, the performance can be further boosted. Different from the existing methods that pick peaks, our method considers carbon 'slices' (denoted as C-slices) as the target in CBCA(CO)NH and HNCACB. Each C-slice is a one-dimensional (1D) vector that corresponds to a certain (N, H) pair. We use the slices chosen from the CBCA (CO)NH spectrum to guide the slice picking in HNCACB to yield better spin systems. In this way, the assignment sends feedback information to guide the slice picking and the dependency of peaks is also modeled. After forming spin systems, we feed them to a genetic algorithm to assign them to the residues of the protein. Here 'simultaneous' refers to the fact that information is propagated forwardly and backwardly between the peak picking and spin system forming steps in our method. Thus, our method is fully automatically from the two input NMR spectra to the final resonance assignment, without any human intervention. Although a number of automated resonance assignment methods are available, few of them work directly on automatically picked peaks that have low quality. IPASS (Alipanahi et al. 2011) is considered to be the state-of-the-art error-tolerant assignment method. We thus tested our framework on the four real protein raw spectra sets that were used to test IPASS, as well as 20 benchmark simulated data sets, ranging from 66 to 370 aa. We compared our method with five state-of-the-art methods, MARS, RIBRA, AUTOASSIGN, IPASS and GANA. Our results demonstrate that our framework significantly outperforms the existing methods in terms of sensitivity and it compares favorably on specificity. The improvement is further indicated by the fact that IPASS, AUTOASSIGN, MARS and RIBRA need more spectra than our method. Our results demonstrate that our proposed framework is able to extract much more information from the real data than the five state-of-the-art methods.

Materials and methods

Overview of our method

In this work, we assume the minimum inputs to be two types of commonly used spectra, CBCA(CO)NH and HNCACB. These two spectra contain adjacency information. Peaks in CBCA(CO)NH are in the forms (N_i, CA_{i-1}, H_i) and (N_i, CB_{i-1}, H_i), where i is the index of the residue and CA and CB stand for C_α and C_β, respectively. On the other hand, the HNCACB spectrum contains peaks in the forms (N_i, CA_{i-1}, H_i), (N_i, CB_{i-1}, H_i), (N_i, CA_i, H_i) and (N_i, CB_i, H_i), where CA peaks are positive and CB peaks are negative. Therefore, from these two spectra, in the ideal case, we can form spin systems of the form $(N_i, H_i, CA_{i-1}, CB_{i-1}, CA_i, CB_i)$. However, in reality, there are various sources of errors, including false positive peaks, false negative peaks, linkage errors, and ambiguities in forming spin systems. Our method is based on the observations that a signaling slice, i.e., a 1D vector corresponds to a certain (N, H) pair in CBCA(CO)NH, should contain two peaks (unless the previous residue is a glycine in which case it should contain one), whereas in HNCACB it should contain four peaks (unless the current or the previous residue is a glycine in which case it should contain three). Our method picks slices instead of peaks and uses the slice information in CBCA(CO)NH to assist the slice picking in HNCACB by considering assignment needs. In this manner, the slice picking and spin system forming are done simultaneously and better spin systems are formed. The spin systems are further assigned to protein residues by a genetic algorithm. Although both GANA and our method use genetic algorithms, the two methods are distinct in the sense that GANA takes spin system groups as inputs whereas our method takes spin systems as inputs. Furthermore, in contrast to GANA, the fitness function of our genetic algorithm contains both connectivity information between spin systems and amino acid typing information.

Our method does not depend on any specifically trained parameters. All the parameters in our method are either set to the rule-of-thumb values, such as error range for different atom types, or set according to previous studies, such as the number of individuals in the genetic algorithm that was set in the same way as in GANA. Therefore, all the results in this paper are based on test proteins that are not involved in the training process.

Forming spin systems by slice picking

Forming spin systems is the prerequisite to resonance assignment. A spin system is the set of chemical shift values from the same or neighboring residues. Previous studies have verified that better spin systems result in better resonance assignments (Jung and Zweckstetter 2004; Lin et al. 2005; Alipanahi et al. 2011). All existing spin system forming methods form spin systems from the lists of peaks provided by either manual or automated peak picking methods. Peak picking and spin system forming are thus considered as two consecutive steps. Our method, in contrast, depends on 1D C-slices and uses information from both steps to guide each other. To extract a signaling carbon slice from these two spectra or any (N, H)-rooted 3D spectrum, we fix a nitrogen value and a hydrogen value. The nitrogen value should be a local maximum in a nitrogen slice, and similarly the hydrogen value should be a local maximum in a hydrogen slice.

The extraction of carbon slices from CBCA(CO)NH is done in the following manner. For every single pair of carbon and hydrogen values (i.e., every point on the (C, H) space), (C_1, H_1), there is a corresponding nitrogen slice. We smooth the nitrogen slice using wavelet smoothing with a soft threshold to decrease the effect of noise. More technical details of wavelet smoothing can be found in *Supplementary Materials S1*. We then search for the highest local maximum in the smoothed nitrogen slice. Let the value of nitrogen at this point be N_1. The pair of (N_1, C_1) then represents a hydrogen slice, which is further smoothed using wavelet smoothing. We then search for the highest local maximum in this hydrogen slice and let its value be H_1'. If $H_1 = H_1'$, we save the two values, H_1 and N_1. These two values indicate a C-slice that our method then chooses. In this way, our method identifies all the carbon slices that have (N, H) values as local maxima.

For each C-slice we identify in CBCA(CO)NH, we search its surrounding regions (where the N chemical shift is within 0.5 ppm and the H chemical shift is within 0.05 ppm) in HNCACB for signaling slices in a similar way. Note here our search looks for similar (N, H) values only, rather than C values. If such a slice is found, we pick the local maximum with the highest intensity in the CBCA(CO)NH slice and up to three local maxima providing that the intensity is higher than 10 % of the highest intensity. The reason to select up to four peaks in each slice is to deal with the situations where spin systems overlap at the same (N, H) position or peaks in spin systems have unusual intensities, which are the main challenging issues in automated assignment. We then search in the corresponding HNCACB slice in a similar way for up to four positive and negative local maxima. We further search for the matches between the local maxima sets in the two slices and identify the largest positive and negative matches. If such matches are found, they are fixed to be CA_{i-1} and CB_{i-1} and the largest unmatched positive and negative peaks in the HNCACB slice are set to CA_i and CB_i, respectively. If only one match is found (without loss of generality, suppose that it is positive), it is set to be CA_{i-1} and the largest unmatched local maximum in CBCA(CO)NH whose chemical shift value falls in the theoretical ranges of C_β is set to be CB_{i-1}. The largest unmatched positive local maximum in HNCACB is set to be CA_i. For CB_i, we will take the largest two negative local maxima in HNCACB as candidates for it. In this case we will form two spin systems with the two values of CB_i. If no match is found, the largest CBCA(CO)NH local maximum in the theoretical range of C_α values is set to be CA_{i-1}. Similarly, the largest CBCA(CO)NH local maximum in the theoretical range of C_β values is set to be CB_{i-1}. Now, we will take the largest two positive local maxima in HNCACB as two candidates for CA_i and the largest two negative local maxima as two candidates for CB_i. Thus, we have four combinations forming four spin systems. The underlying idea behind this combinatorial way of spin system forming is to accommodate missing peaks and missing/wrongly-assigned chemical shifts that are key to connectivities. That is, this step tries to form spin systems as complete and correct as possible, whereas the next assignment step will assign imperfect spin systems through a stochastic heuristic algorithm.

Resonance assignment using a genetic algorithm

To assign the spin systems to corresponding residues, we developed a genetic algorithm. The idea was inspired by GANA (Lin et al. 2005), which is also based on a genetic algorithm. However, GANA uses spin system groups as assignment units and encodes connectivity information in the fitness function, whereas our method treats ambiguous spin systems as individual ones and encodes both connectivity information and amino acid typing information in the fitness function. These differences are further detailed below and experiments demonstrate these differences result in significant improvements in the output of our method compared with that of GANA.

Basics of genetic algorithms

Genetic algorithm, a popular method in artificial intelligence, is a heuristic search approach that mimics the process of natural evolution. In a genetic algorithm, the search states are often represented as a string of bits. The algorithm starts with a random 'population' of states, each of which is called an 'individual'. A fitness function is defined and applied to evaluate the quality of each individual. The individuals are selected proportionally to their fitness scores to derive the next 'generation' with the same number of individuals. There is a low chance for each individual in the new generation to have a mutation, which is a change to one bit in the state representation. The generation after a mutation is then used as the initial generation for the next round of evolution. This procedure continues until convergence or a maximum number of generations is reached. A genetic algorithm is conceptually analogous to hill climbing on the fitness function, but it is more advanced in the sense that it has a mechanism that allows it to jump out of local minima.

The fitness function

The fitness function is one of the most important features in genetic algorithms. It can significantly influence both the efficiency and effectiveness of evolutionary algorithms. GANA uses a fitness function based on connectivity information between the spin systems. The fitness function used in our method, however, consists of two parts, one for encoding connectivity information and the other for encoding amino acid typing information.

The connectivity part of our fitness function is defined in a similar way as in GANA (Lin et al. 2005). To define the fitness score of an individual, each residue i assigned by the spin system x_i is examined to determine whether or not x_i is connected with x_{i-1} and x_{i+1}. If residue i is not assigned, x_i is set to 0. Two variables, $D_L(i)$ and $D_R(i)$, are then defined as the sum of the absolute differences between the corresponding CA and CB chemical shifts, to reflect the degree of connectivity between (x_{i-1}, x_i) and (x_i, x_{i+1}), respectively. $D_L(i)$ and $D_R(i)$ are then binned into the ranges of [0.0, 0.1), [0.1, 0.3), [0.3, 0.5), [0.5, 0.7), [0.7, 1.0) and [1.0, ∞) and scored as 5, 4, 3, 2, 1 and -3, respectively. The binned scores are denoted as $S_L(i)$ and $S_R(i)$. The connectivity score for x_i is then defined as:

$$Score_{con}(x_i) = \begin{cases} 0, & \text{if } x_i = 0; \\ 1, & \text{if } x_i \neq 0, x_{i-1} = x_{i+1} = 0; \\ S_L(i), & \text{if } x_i, x_{i-1} \neq 0, x_{i+1} = 0; \\ S_R(i), & \text{if } x_i, x_{i+1} \neq 0, x_{i-1} = 0; \\ S_L(i) + S_R(i), & \text{otherwise.} \end{cases} \tag{1}$$

The connectivity score of the individual is then defined as: $SCORE_{CON}(ind) = \sum_{i=1}^{l} Score_{con}(x_i)$, where l is the length of the protein. Thus, the connectivity score is defined in the same way as in (Lin et al. 2005).

Our fitness function contains another term to encode the amino acid typing information to evaluate the score of assigning a spin system to an amino acid. This information is not included in GANA's fitness function because GANA works on spin system groups rather than on spin systems. We first built a database that contains CA and CB chemical shifts for each residue and its previous residue (if it exists) in the BMRB database (Seavey et al. 1991). Each entry in our database thus contains four chemical shift values and two corresponding amino acids, one for the current residue and the other for the preceding residue. To calculate the amino acid typing score that a spin system, x_i, assigns to a residue, we search in the database the entries that had the same amino acid types for the residue and its preceding one. The root mean square distance (rmsd) of carbon chemical shifts between each of these entries and the spin system is calculated and the smallest rmsd (denoted as d) is picked up. The amino acid typing score for assigning this spin system to the residue is then defined as $Score_{AA}(x_i) = 1/d$. The intuition is to prefer the amino acid typing with smaller chemical shift difference in terms of rmsd. The amino acid typing score for an individual is thus defined as $SCORE_{AA}(ind) = \sum_{i=1}^{l} Score_{AA}(x_i)$.

The fitness function used in our method is the weighted sum of these two terms,

$$SCORE(ind) = SCORE_{con}(ind) + w \cdot SCORE_{AA}(ind), \tag{2}$$

where w is set to 5 by default and can be customized by the user. The value of five was selected because $SCORE_{con}$ was often several times larger than $SCORE_{AA}$. In all of our experiments, the BMRB entries of the target protein and its homologs are removed to ensure a fair comparison. Our fitness function provides a tradeoff between amino acid typing and connectivity information, thus can naturally handle a number of issues in assignment, including missing chemical shifts in spin systems, wrongly grouped chemical shifts in spin systems, unusual chemical shifts for particular amino acids, and missing connectivity between spin systems.

The genetic algorithm model

The genetic algorithm model has the following components:

- Initialization: Each individual is represented by a 1D vector of length l, where l is the length of the protein to be assigned. Each entry in this vector can be assigned an integer between 0 and m, where m is the number of spin systems. If the entry is assigned 0, this means that the residue is not assigned. Otherwise, it is assigned by the corresponding spin system. To generate one individual randomly, our model starts with an unassigned vector, i.e., all entries are set to 0. We then randomly select a residue to assign a random spin system that has CA and CB values that fall into the expected chemical shift range of the amino acid type of the residue. The ranges are relaxed ranges of the theoretical chemical shift values such that noise and errors are tolerated (see Table S2). We then try to extend this assigned residue to both directions. To extend to the left, we randomly select an unassigned spin system that has CA and CB values within 0.5 ppm to the CA_{-1} and CB_{-1} values of this residue. Similarly, to extend to the right, we randomly selected an unassigned spin system that has CA_{-1} and CB_{-1} values within 0.5 ppm to the CA and CB values of this residue. This extension continues until no extension is available. We then randomly select another unassigned residue to repeat this assignment and extension procedure. The iteration ends when no more residue can be assigned. We generate 600 individuals for the initial population.
- Selection: Each individual is evaluated by our proposed fitness function, $SCORE(ind)$, and assigned a fitness score. To generate the new generation, each individual has a probability that is proportional to its fitness score. That is, the higher the fitness score, the more likely it is that this individual is in the next generation.
- Crossover: The individuals in each generation are ranked by their fitness scores and the top 50 % are kept in the generation after crossover. These top 50 % are also treated as parent candidates for crossover operation. Seventy percent of the remaining 50 % are generated using crossover. The crossover is done in the same way as in GANA (Lin et al. 2005). That is, two individuals are selected to generate one new individual by randomly selecting a position from either individual to copy to this new, empty assignment (i.e., new individual). If this position is assigned by a spin system in the selected individual, it is copied to the new assignment and extended to both directions by referencing the selected individual. That is, starting from this copied spin system, we extend to both directions according to the connected fragment in the selected individual that contains this spin system, until the end(s) of the fragment or until the spin system(s) in the copied fragment is already used in the newly formed individual, whichever happens earlier. Otherwise, we randomly select another position of either individual to copy to the new one. This procedure is repeated until no further assignment can be done for the new individual. The remaining individuals of the new generation after crossover are generated using the random chromosome initialization as in the initialization step.
- Mutation: Mutation is an important exploration step that can potentially help genetic algorithms to jump out of local maxima. In our method, each position in each individual has a probability of 0.2 % to be mutated. We started from the first position. Once a position is selected for mutation, we randomly select an unassigned spin system that has CA and CB values that fall into the expected ranges of the amino acid to replace the existing spin system. We then extend the spin system to the right for as far as possible by considering the connectivity information as explained in the initialization step. This extension is done to avoid splitting of connected fragments into pieces. Once the extension is finished, the next position is selected for a mutation with the same 0.2 % chance, until all the positions are traversed. It should be noted that although mutation provides a mechanism to possibly jump out of local maxima, it does not guarantee a search path to a better assignment, nor does it guarantee a complete assignment.
- Stopping condition: Our genetic algorithm stops if either the best fitness score among individuals does not improve over 100 consecutive generations or the maximum number of 500 generations is reached, whichever is sooner.

Results

Performance on raw NMR spectrum sets

We tested our method on the spectrum sets of four real proteins, TM1112 from *Thermotoga maritima*, CASKIN (the SH3 domain of the CASKIN neuronal signaling protein), VRAR (*S. aureus* VraR DNA binding domain), and HACS1 (the SH3 domain of the HCAS1 human myeloid/hemopoetic signaling protein). These four proteins are the same as those used to test IPASS (Alipanahi et al. 2011).

Although these four proteins are considered to be small- or medium-sized proteins in NMR (67–89 aa), (Alipanahi et al. 2011) showed that fully automated peak picking and resonance assignment of proteins of this size were still very challenging and most existing methods performed poorly. There are two main reasons for this poor performance. First, automatically picked peaks contain both false positives and false negatives, which introduce errors and noise in the formation of the spin systems. Second, although they perform perfectly on simulated data sets, existing resonance assignment methods are not error tolerant enough to handle peaks chosen by automated methods.

For evaluation purposes, we measure recall, precision and F1-score. Assume that the target protein has N_r manually assigned residues and that our resonance assignment method assigns N_o residues, where T_p of them are assigned correctly. Recall is then defined as T_p/N_r, while precision is defined as T_p/N_o. The F1-score is defined as the harmonic mean of recall and precision. For a non-proline residue, a spin system is considered to be assigned correctly if and only if at most one chemical shift is out of the error range (0.5 ppm for nitrogen and carbons, and 0.05 ppm for hydrogen) of the manual assignment. For a proline residue, a spin system is considered to be assigned correctly if and only if both its carbon chemical shifts are within the error range of the manual assignment.

We compared our method with four state-of-the-art methods, MARS (Jung and Zweckstetter 2004), IPASS (Alipanahi et al. 2011), AUTOASSIGN (Zimmerman et al. 1997) and GANA (Lin et al. 2005). All these methods are resonance assignment methods that require the inputs to be peak lists or spin systems. In (Alipanahi et al. 2011), PICKY (Alipanahi et al. 2009) was used to automatically pick peaks from four spectra, ^{15}N-HSQC, CBCA(CO)NH, HNCACB, and N-NOESY, as inputs to IPASS and MARS. Here, we followed the same procedure for these two methods. AUTOASSIGN requires peak lists from the 2D ^{15}N-HSQC as the root list, plus peak lists from CBCA(CO)NH and HNCACB. The peaks picked by PICKY on these three spectra were thus fed into AUTOASSIGN. GANA requires formed spin systems as inputs. For the sake of fair comparison, we gave two types of inputs to GANA, i.e., the spin systems formed by IPASS (denoted as GANA^a) and the spin systems formed by our method (denoted as GANA^b). The assignments from IPASS, MARS and GANA^a were therefore performed by using four spectra; that from AUTOASSIGN was done on three spectra; whereas those from GANA^b and our method were performed by using only two spectra.

Table 1 shows the performance of the five methods on the four real protein data sets. Our method significantly outperforms all other methods on recall and F1-score, while it compares favorably on precision. For IPASS, MARS, AUTOASSIGN and GANA^a, the spin systems are formed by the peaks picked by PICKY that does not take assignment needs into consideration. Our method, although using a less number of spectra, effectively identifies better spin systems by simultaneously picking peaks and forming spin systems, thus achieving significantly more accurate assignments on all the proteins. This conclusion is further supported by comparing the results from GANA^a and GANA^b. When GANA takes the spin systems formed by our method as input, it performs significantly better than when taking spin systems formed by IPASS as input. The F1-score is approximately 100 % better. Note that the only difference between GANA^a and GANA^b is the spin systems that are used as input. This suggests that our simultaneous slice picking and spin system forming framework is much more accurate than the traditional pipeline that treats these two steps separately. Our method also significantly outperforms GANA^b, which uses the same spin systems as input as our method uses. This implies that the novel fitness function in our genetic algorithm is more

Table 1 Performance comparison between MARS, IPASS, AUTOASSIGN, GANA and our method on four real protein data sets

Protein	Len	Man	MARS	IPASS	AUTOASSIGN	GANA^a	GANA^b	Our method	Proline
TM1112	89	89	55/63	71/72	67/76	42/64	86/87	86/87	5/5
CASKIN	67	58	23/25	29/39	9/18	24/34	36/56	40/56	1/3
VRAR	72	60	6/17	30/37	7/7	8/18	34/54	42/54	0/0
HACS1	74	61	15/16	37/50	18/24	6/21	39/60	42/59	1/3
REC_{ave}	–	–	0.34	0.60	0.33	0.28	0.69	0.76	–
PRE_{ave}	–	–	0.77	0.82	0.78	0.52	0.73	0.80	–
$F1_{ave}$	–	–	0.47	0.69	0.42	0.36	0.71	0.78	–

The second column gives the number of residues in the protein. The third column gives the number of manually assigned residues (including prolines). Starting from the fourth column, the performance of each method is shown in the format of “number of correctly assigned residues/ total number of assigned residues”. GANA^a lists the results of GANA by using the same spin systems formed by IPASS as input. GANA^b lists the results of GANA by using the same spin systems formed by our method as input. The last column gives the “number of correctly assigned proline residues/total number of manually assigned proline residues”. REC_{ave}, PRE_{ave} and $F1_{ave}$ stand for the average recall, precision and F1-score, respectively

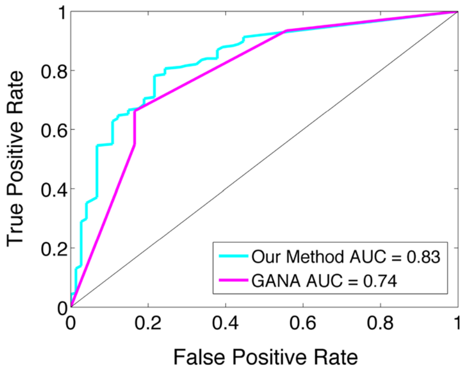


Fig. 1 ROC curves for our method and GANA^b on the four protein data sets. GANA^b denotes GANA by using the same spin systems formed by our method as input

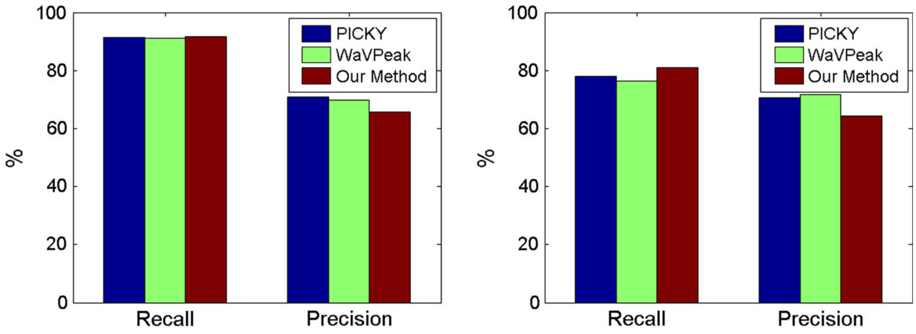
effective than GANA's fitness function. Figure 1 shows the receiver operating characteristic (ROC) curves for our method and GANA^b on these four protein data sets, which suggests that the improvements of our method over GANA is not just the result of a better tradeoff parameter, but the result of a consistently better method. Another noticeable advantage of our method over the existing methods is that our method does not depend on peaks from ^{15}N-HSQC, which is considered as a root spectrum that is widely used in computational peak refinement and spin system forming. Methods that depend on ^{15}N-HSQC, such as IPASS, cannot assign proline residues because prolines do not have peaks in ^{15}N-HSQC. Our method, on the other hand, can assign prolines through slices that correspond to the residues after prolines. As shown in Table 1, our method is able to assign most of the proline residues correctly in these proteins (seven correctly assigned prolines out of eleven).

Among these proteins, the performance of different assignment methods on TM1112 is significantly better than that on the other three proteins. This is due to the nature of CASKIN, VRAR and HACS1. CASKIN and HACS1 both have long, flexible regions that even manual assignments are not available. The proteins are thus partitioned into various segments by those flexible regions, which significantly reduced the connectivity information among spin systems. VRAR, on the other hand, is a helical protein that is known to be difficult to assign, even by manual assignment. Since chemical shift values depend on local geometric and environmental factors, residues of VRAR have very similar carbon chemical shift values, which introduced a large number of ambiguities in assignment. Therefore, future research is needed to improve the accuracy of our method on helical proteins and proteins with flexible regions.

Although the goal of our method is not to pick peaks, we evaluated the peaks identified in the slices picked by our method. We applied a naïve approach that picks the peak with the highest intensity in each of our slices in CBCA(CO)NH and up to three other peaks in the same slice providing that their intensities are higher than 10 % of the highest peak. For HNCACB, we did the same selection for peaks with both positive and negative intensities. This naïve method is able to identify most of the true C-slices (Figures S6–S9). Figure 2 shows the performance comparison of this approach with two state-of-the-art peak picking methods, PICKY (Alipanahi et al. 2009) and WaVPeak (Liu et al. 2012). Our approach compares favorably on recall but slightly worse on precision. This implies that statistically better peaks do not necessarily lead to better assignments. This explains why simply combining state-of-the-art methods in a sequential order often does not yield a good pipeline. Instead, peak picking and assignment steps should be considered simultaneously.

Performance on spectrum sets simulated from manual assignments

We further tested our method on a simulated data set that consists of 20 proteins, ranging from 66 to 370 aa. In contrast to previous studies that use simulated peaks to test assignment methods, we simulated spectra to evaluate our framework for simultaneous slice picking and resonance assignment. To simulate various sources of noise in real spectra, we first generated all the ideal spin systems from the manual chemical shift assignment of each of these proteins. We then randomly generated 20 % more spin systems to simulate false positive signals in real spectra. Note that each complete spin system can be decomposed into four HNCACB peaks (less than four if the residue or the preceding one is glycine), two CBCA(CO)NH peaks (one if the preceding residue is glycine), and one ^{15}N-HSQC peak. Expected peaks in CBCA(CO)NH, HNCACB, and ^{15}N-HSQC were extracted from all these spin systems. Each chemical shift value of each peak has a 10 % chance to be shifted by +0.2 or −0.2 ppm for N and C, and +0.02 or −0.02 ppm for H. These shifts simulated the linking errors in real spectra. Each peak then has a 1 % probability to be removed, which simulated false negatives in real spectra. For each (N, H) slice in CBCA(CO)NH and HNCACB, we randomly added a peak with 10 % probability to introduce ambiguities into the formation of spin systems. Peaks were constructed to have the same intensities. After all peaks were constructed, we added white Gaussian noise to all the data points in the spectra. The noise was set to have a zero mean and the standard deviation to be the root of the intensity of the peaks. The simulated spectra thus contained noise, false positive peaks, false negative peaks, linking errors and ambiguous



(a) CBCA(CO)NH (b) HNCACB

Fig. 2 Peak picking performance comparison on **a** CBCA(CO)NH and **b** HNCACB spectra of the four real proteins. The recall and precision values are averaged over the four proteins

Table 2 Performance comparison on the simulated date sets

Protein	Len	RIBRA		AUTOASSIGN		GANA		Our method	
		PRC	REC	PRC	REC	PRC	REC	PRC	REC
bmr4391	66	0.83	0.41	0.58	0.19	0.79	0.83	0.84	0.83
bmr4752	68	0.62	0.39	0.95	0.67	0.94	0.91	0.97	0.94
bmr4144	78	0.69	0.26	0.67	0.21	0.61	0.52	0.88	0.71
bmr4579	86	0.66	0.37	0.88	0.46	0.93	0.88	0.90	0.88
bmr4316	89	0.71	0.48	0.96	0.81	0.77	0.70	0.88	0.82
bmr4288	105	0.82	0.52	0.93	0.44	0.86	0.83	0.94	0.90
bmr4929	114	0.74	0.45	0.81	0.56	0.84	0.79	0.89	0.82
bmr4302	115	0.63	0.37	0.90	0.36	0.86	0.86	0.86	0.86
bmr4670	120	0.81	0.53	0.74	0.44	0.81	0.79	0.83	0.76
bmr4353	126	0.88	0.46	0.68	0.31	0.77	0.76	0.90	0.88
bmr4027	158	0.52	0.34	0.56	0.32	0.93	0.92	0.97	0.96
bmr4318	215	0.43	0.19	0.96	0.23	0.81	0.80	0.90	0.89
bmr4836	217	0.42	0.30	0.69	0.22	0.87	0.85	0.96	0.94
bmr4102	221	0.61	0.36	0.85	0.29	0.74	0.78	0.82	0.85
bmr4022	260	0.37	0.21	0.88	0.25	0.80	0.79	0.92	0.90
bmr6074	261	0.36	0.21	0.87	0.24	0.80	0.79	0.91	0.89
bmr4384	262	0.49	0.33	0.95	0.30	0.87	0.87	0.89	0.91
bmr5316	288	0.32	0.15	0.85	0.14	0.74	0.76	0.77	0.79
bmr6136	306	0.41	0.22	0.97	0.24	0.72	0.72	0.76	0.75
bmr4987	370	0.27	0.17	0.73	0.25	0.73	0.75	0.78	0.80
Average	–	0.58	0.34	0.82	0.35	0.80	0.78	0.86	0.84
F1-score	–	0.42		0.47		0.79		0.85	

The first column on the left presents the protein ID in the BMRB database, sorted according to the protein's length. The next column gives the length of the 20 proteins. Starting from the third column, the precision (PRC) and recall (REC) values for RIBRA, AUTOASSIGN, GANA and our method are listed

spin systems. It turned out that our method was not very sensitive to the settings of these errors (data not shown).

We compared our method with RIBRA (Wu et al. 2006), AUTOASSIGN (Zimmerman et al. 1997) and GANA (Lin et al. 2005) on these 20 simulated sets of spectra. WaVPeak (Liu et al. 2012) was used to pick peaks from the simulated spectra. RIBRA and AUTOASSIGN required the peaks from all three spectra as inputs, whereas GANA and our method required the peaks from only two. The spin systems formed by our method were used as inputs to GANA. Table 2 shows the performance comparison between the four methods. Both our method and GANA significantly outperformed RIBRA and AUTOASSIGN in terms of recall and F1-score, while our method also exhibited clear improvements over GANA. In terms of the F1-score, our method has almost twofold improvement over RIBRA and AUTOASSIGN. This comparison demonstrated the proposed simultaneous framework is much more effective and powerful than the traditional sequential pipeline. Again, the reason for the improvement in performance of our method over GANA is the fitness function proposed in our method. It should also be noted that AUTOASSIGN, GANA with our spin systems and our method all have a reasonably high precision value, which suggests these methods can provide reliable assignments to the users.

Table 3 Calculated structures of CS-ROSETTA for the four real proteins

	TM1112			CASKIN			VRAR			HACS1		
	Man	Ours	No_{CS}	Man	Ours	No_{CS}	Man	Ours	No_{CS}	Man	Ours	No_{CS}
$RMSD_{TopScore}$	1.32	1.34	17.02	11.85	3.36	8.80	2.27	2.77	14.08	1.13	2.96	7.88
$RMSD_{BestOfTop10}$	1.20	1.22	14.07	4.30	3.36	6.99	1.99	2.19	9.38	1.08	2.96	7.88
$RMSD_{BestOfPool}$	1.20	1.12	12.06	2.30	2.82	6.07	1.68	1.66	8.63	1.08	1.83	6.81

'Man', 'Our', and 'No_{CS}' stand for the calculated structure pool of CS-ROSETTA that takes the BMRB manual assignments, the assignments by our method, and only amino acid sequences without chemical shift assignments as inputs, respectively. '$RMSD_{TopScore}$', '$RMSD_{BestOfTop10}$', and '$RMSD_{BestOfPool}$' stand for the RMSD of the structural model that has the lowest the CS-ROSETTA scoring function, the lowest RMSD within the top 10 models by the CS-ROSETTA scoring function, and the lowest RMSD within the entire structural model pool generated by CS-ROSETTA, respectively. All the values in the table are in Å

When the protein size increases, the performance of different methods generally decreases. This suggests that for longer and more complex proteins, more spectra are needed to provide more information to assignment.

Structure calculation from chemical shift assignment

Although our method provides significantly more complete and accurate chemical shift assignment than other state-of-the-art methods, the assignment done by our method is still not complete. It is thus not clear whether such assignment could lead to accurate 3D structures of the target protein, which is the ultimate goal of the NMR protein structure determination process. To test this, we tried to calculate the final structures of the four target proteins solely from the chemical shift assignments and amino acid sequences by using the CS-ROSETTA server (Shen et al. 2008, 2009).

For each target protein, we used three types of inputs besides the amino acid sequence for CS-ROSETTA. The first one is the manual assignment downloaded from BMRB. The second one is the chemical shift assignment predicted by our method from the raw NMR spectrum set of the target. The third one is the empty assignment file, which requires CS-ROSETTA to calculate final structures solely based on sequence information. The third input was used to confirm sequence information alone could not lead to accurate 3D structures, even if close homologs of the target protein might be in the database of CS-ROSETTA. For each input, 3,000 structural models were generated by CS-ROSETTA.

As shown in Table 3, when the chemical shift assignments predicted by our method were used, accurate structural models could be generated and good models were always ranked in top, which made the model selection step much easier. In fact, the structures calculated based on our assignments are as good as those based on the manual assignments, and sometimes even better. It is also clear that sequence information alone is not sufficient to lead to accurate structural models. The top scored models were aligned well with the experimentally determined native structures (Fig. 3).

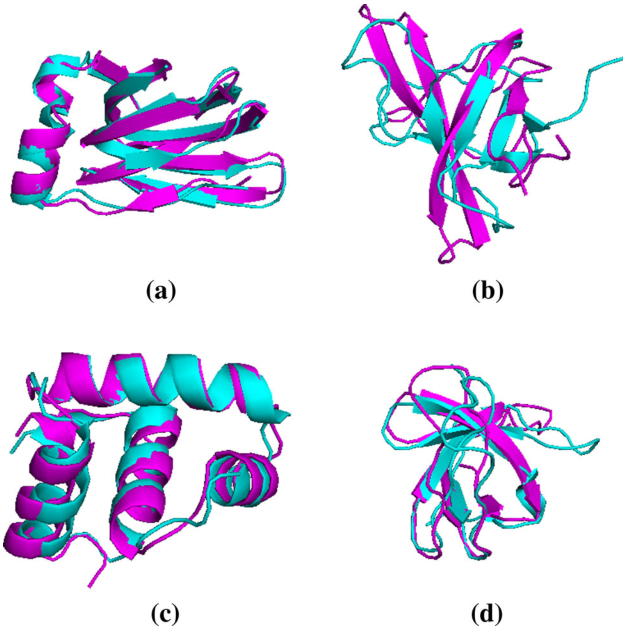


Fig. 3 Structural alignments between the top scored structural models by using CS-ROSETTA with our assignments and the experimentally determined native structures for the four real proteins. The structures in *cyan* are the native structures and those in *magenta* are the calculated structures. **a** TM1112. **b** CASKIN. **c** VRAR. **d** HACS1

Conclusion

In this paper, we proposed a framework for NMR resonance assignment that works significantly better than the state-of-the-art on real, noisy protein data sets. The framework is based on simultaneous slice picking and spin system forming. It requires as few as two through-bond spectra as inputs. If additional spectra are available, e.g., N-NOESY, our slice picking method can be straightforwardly applied to pick slices in those spectra and, at the same time, those spectra can provide feedback information to direct slice picking in CBCA(CO)NH and HNCACB. We demonstrated that the assignments done by our method can lead to accurate structural models. Our method is open source. The source code, README and sample data are freely available at http://sfb.kaust.edu.sa/Pages/Software.aspx.

Acknowledgments We thank Dr. Ad Bax's group for making CS-ROSETTA available. We are grateful to Dr. Yang Shen for answering our questions regarding CS-ROSETTA server. The spectra for TM1112 were generated by Cheryl Arrowsmith's Lab at the University of Toronto. The spectra for CASKIN, VRAR, and HACS1 were provided by Logan Donaldson's Lab at York University. We thank Virginia Unkefer for editorial assistance. This work was supported by Award No. GRP-CF-2011-19-P-Gao-Huang and a GMSV-OCRF award from King Abdullah University of Science and Technology (KAUST).

References

Abbas A, Liu Z, Jing B, Gao X (2013) Automatic peak selection by a Benjamini–Hochberg-based algorithm. PLOS One 8(1):e53112

Alipanahi B, Gao X, Karakoc E, Donaldson L, Li M (2009) PICKY: a novel SVD-based NMR spectra peak picking method. Bioinformatics 25(12):i268–i275

Alipanahi B, Gao X, Karakoc E, Li SC, Balbach F, Donaldson L, Li M (2011) Error tolerant NMR backbone resonance assignment and automated structure generation. J Bionform Comput Biol 9(1):15–41

Altieri A, Byrd R (2004) Automation of NMR structure determination of proteins. Curr Opin Struct Biol 14(5):547–553

Antz C, Neidig K, Kalbitzer H (1995) A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis. J Biomol NMR 5(3):287–296

Bartels C, Billeter M, Güntert P, Wthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program garant. J Biomol NMR 7(3):207–213. doi: 10.1007/BF00202037. 10.1007/BF00202037

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242

Carrara E, Pagliari F, Nicolini C (1993) Neural networks for the peak-picking of nuclear magnetic resonance spectra. J Neural Netw 6(7):1023–1032

Cheng Y, Gao X, Liang F (2014) Bayesian peak picking for NMR spectra. Genomics Proteomics Bioinform 12(1):39–47

Coggins B, Zhou P (2003) PACES: protein sequential assignment by computer aided exhaustive search. J Biomol NMR 26:93–111

Dancea F, Güntert U (2005) Automated protein NMR structure determination using wavelet de-noised NOESY spectra. J Biomol NMR 33(3):139–152

Delaglio F, Grzesiek S, Vuister G, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293

Gao X (2012) Mathematical approaches to the NMR peak-picking problem. J Appl Comput Math 1:1

Gao X (2013) Recent advances in computational methods for nuclear magnetic resonance data processing. Genomics Proteomics Bioinform 11(1):29–33

Garret D, Powers R, Gronenborn A, Clore G (1991) A common sense approach to peak picking in two-, three-, and four-dimensional spectra using automatic computer analysis of contour diagrams. J Magn Reson 95(1):214–220

Goddard T, Kneller D (2007) SPARKY 3. University of California, San Francisco

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

Güntert P, Salzmann M, Braun D, Wuthrich K (2000) Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. J Biomol NMR 18:129–137

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

Günther U, Ludwig C, Ruterjans H (2000) NMRLAB—advanced NMR data processing in matlab. J Magn Reson 145(2):201–208

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

Ikeya T, Takeda M, Yoshida H, Terauchi T, Jee JG, Kainosho M, Güntert P (2009) Automated NMR structure determination of stereo-array isotope labeled ubiquitin from minimal sets of spectra using the sail-flya system. J Biomol NMR 44(4):261–272. doi: 10.1007/s10858-009-9339-6

Jang R, Gao X, Li M (2010) Towards automated structure-based NMR resonance assignment. Lecture Notes Comput Sci 6044:189–207

Jang R, Gao X, Li M (2011) Towards fully automated structure-based NMR resonance assignment of 15N-labeled proteins from automatically picked peaks. J Comput Biol 18:347–363

Jang R, Gao X, Li M (2012) Combining automated peak tracking in SAR by NMR with structure-based backbone assignment from 15N-NOESY. BMC Bioinform 13(S3):S4

Johnson B, Blevins R (1994) NMR view: a computer program for the visualization and analysis of NMR data. J Biomol NMR 4:603–614

Jung Y, Zweckstetter M (2004) Mars-robust automatic backbone assignment of proteins. J Biomol NMR 30:11–23

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

Korzhneva D, Ibraghimov I, Billeter M, Orekhov V (2001) MUNIN: application of three-way decomposition to the analysis of heteronuclear NMR relaxation data. J Biomol NMR 21:263–268

Lemak A, Steren C, Arrowsmith C, Llinas M (2008) Sequence specific resonance assignment via Multicanonical Monte Carlo search using an ABACUS approach. J Biomol NMR 41:29–41

Lin HN, Wu KP, Chang JM, Sung TY, Hsu WL (2005) GANA: a genetic algorithm for NMR backbone resonance assignment. Nucleic Acids Research 33:4593–4601

Liu Z, Abbas A, Jing BY, Gao X (2012) WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering. Bioinformatics 28(7):914–920

Masse J, Keller R (2005) Autolink: automated sequential resonance assignment of biopolymers from NMR data by relative–hypothesis–prioritization-based simulated logic. J Magn Reson 174:133–151

Seavey BR, Farr EA, Westler WM, Markley JL (1991) A relational database for sequence-specific protein NMR data. J Biomol NMR 1(3):217–236

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105(12):4685–4690. doi:10.1073/pnas.0800256105

Shen Y, Vernon R, Baker D, Bax A (2009) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43(2):63–78. doi:10.1007/s10858-008-9288-5

Takeda M, Ikeya T, Güntert P, Kainosho M (2007) Automated structure determination of proteins with the SAIL-FLYA NMR method. Nat Protoc 2(11):2896–2902. doi:10.1038/nprot.2007.423

Tycko R, Hu K (2010) A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic angle spinning. J Magn Reson 205:304–314

Volk J, Herrmann T, Wüthrich K (2008) Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. J Biomol NMR 41:127–138

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59(4):687–696. doi:10.1002/prot.20449

Wan X, Lin G (2007) CISA: combined NMR resonance connectivity information determination and sequential assignment. IEEE/ACM Trans Comput Biol Bioinform 4:336–348

Wu K, Chang J, Chen J, Chang C, Wu W, Huang T et al (2006) RIBRA–an error-tolerant algorithm for the NMR backbone assignment problem. J Comput Biol 13:229–244

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

Zeng J, Zhou P, Donald BR (2011) Protein side-chain resonance assignment and NOE assignment using RDC-defined backbones without TOCSY data. J Biomol NMR 50(4):371–395. doi:10.1007/s10858-011-9522-4

Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien C, Powers R, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269(4):592–610