IPASS: Error Tolerant NMR Backbone Resonance Assignment by Linear Programming

Babak Alipanahi[1⋆], Xin Gao[1⋆], Emre Karakoc[1⋆],
Frank Balbach[1], Logan Donaldson[2], and Ming Li[1⋆⋆]

[1] David R. Cheriton School of Computer Science,
University of Waterloo, Waterloo, ON, Canada N2L 6P7
[2] Department of Biology, York University, Toronto, ON, Canada M3J 1P3

Technical Report CS-2009-16

Abstract. The automation of the entire NMR protein structure determination process requires a superior error tolerant backbone resonance assignment method. Although a variety of assignment approaches have been developed, none works well on noisy automatically picked peaks. IPASS is proposed as a novel integer linear programming (ILP) based assignment method. In order to reduce size of the problem, IPASS employs probabilistic spin system typing based on chemical shifts and secondary structure predictions. Furthermore, IPASS extracts connectivity information from the inter-residue information and the ^{15}N-edited NOESY peaks which are then used to fix reliable fragments. The experimental results demonstrate that IPASS significantly outperforms the previous assignment methods on the synthetic data sets. It achieves an average of 99% precision and 96% recall on the synthesized spin systems, and an average of 96% precision and 90% recall on the synthesized peak lists. When applied on automatically picked peaks from experimentally derived data sets, it achieves an average precision and recall of 78% and 67%, respectively. In contrast, the next best method, MARS, achieved an average precision and recall of 50% and 40%, respectively.
Availability: IPASS is available upon request, and the web server for IPASS is under construction.
Contact:mli@uwaterloo.ca

1 Introduction

The backbone resonance assignment also known as chemical shift assignment plays a vital role in the entire NMR protein structure determination process. Here, the goal is to assign the picked peaks from NMR spectra to their corresponding nuclei of the target protein. Furthermore, backbone resonance assignment acts as an indispensable prerequisite for the NOE assignment. In fact, backbone resonance assignment is the part of the entire NMR process that has attracted the most computational attention for the last ten years [1–9].

Typically, the backbone resonance assignment is divided into three sub-problems: forming spin systems, linking spin systems into fragments, and mapping the fragments to the target sequence. A "spin system" denotes a group of coupled nuclei that can be observed as cross-peaks in one or more spectra. Usually spin systems contain both inter-residue and intra-residue information. The existing methods can be classified into two groups: assignment methods that require spin systems [3–5, 8] and assignment methods that do not require spin systems [1, 6, 7, 9]. However, the latter assignment methods always require high quality peak lists with a very small number of missing or false peaks and little difference in the chemical shift of the same nucleus in different spectra. Therefore, for most cases, the experiments carried out in such studies, are based on either manually picked and refined peak lists by spectroscopists, or on synthetic peak lists formed by assigned chemical shifts in a known protein database such as BioMagResBank (BMRB) [10].

Also, according to whether or not an assignment method needs human intervention, existing methods can be classified as "semi-automated" assignment methods [2, 3] or "fully-automated" assignment methods [1, 4–9]. AUTOASSIGN [1] is a fully-automated multi-stage expert system. The idea of AUTOASSIGN is the best first search, which assigns the strongest fragment matches first, and then gradually relaxes restrictions to assign weaker

⋆ The first three authors contributed equally to this paper.
⋆⋆ All correspondence should be addressed to mli@cs.uwaterloo.ca

matches. MAPPER [2] and PACES [3] are semi-automated methods that are also based on the best first search concept. Both of them employ exhaustive search strategy to map the fragments to target proteins. AUTOLINK [5] is an attempt to mimic human logic by a fuzzy logic and relative hypothesis prioritization method. To the authors' best knowledge, AUTOLINK is the first assignment method that extracts spin system connectivity information from the NOESY data. [6] later proposed a weighted maximum independent set formulation for the assignment problem. They provided a comprehensive summary of the different sources of the spectra errors in the lab experiments, and further simulated these errors on perfect datasets, extracted from BMRB.

MARS [4], one of the widely acknowledged assignment methods, is different from its ancestors in that it applies the consensus idea to multiple runs of assignments, where each run is carried out to optimize different objective functions. For the local assignment, MARS uses the best first search to find the local fit of the fragments, comprising as many as five spin systems. For global assignment, however, MARS optimizes the global pseudo-energy function, which measures how well a spin system matches a residue in the target protein. The pseudo-energy is based on the likelihood of observing a certain chemical shift for an amino acid type in the BMRB database.

Recently, [8] and [9] proposed two sophisticated methods to solve the resonance assignment problem on the most up-to-date NMR spectra. ABACUS [8] takes unassigned peaks from NOESY, COSY (correlation spectroscopy), and TOCSY (total correlation spectroscopy), as well as database-derived likelihoods, as the input. A multi-canonical Monte Carlo procedure, Fragment Monte Carlo (FMC), is used to perform sequence-specific assignments. In MATCH method [9], both the global and local optimization strategies merge where 6D APSY spectrum [11, 12] is the input.

In this paper, the goal is to develop a superior error-tolerant assignment method for automated peak-picking results. Therefore, IPASS is not developed with manually picked or synthesized peaks in mind. Thus, the proposed assignment algorithm should be appropriate for low quality input. Most of the previously designed assignment methods are designed to deal with high quality data sets. Therefore, none of these methods work well on the real data set. Consequently, a novel Integer Linear Programming (ILP) based assignment method, which combines a new spin system forming, an improved probabilistic spin system typing, and a novel connectivity extraction method is proposed.

2 Methods

2.1 Problem Formulation

Given an amino acid sequence of a protein with n residues as $r_1 r_2 \dots r_n$, define $R = \{r_1, r_2, \dots, r_n\}$. Spin systems are given as $S = \{\mathbf{s}_1, \dots, \mathbf{s}_m\}$, where $\mathbf{s}_\mathrm{j}$ is a vector of the chemical shifts. Then, the assignment problem is finding the correct mapping between spin systems and residues, expressed as $f : S \rightarrow R$. Due to the imperfect NMR spectra, peak picking, and spin systems forming, the number of spin systems can be smaller, larger, or equal to the number of residues and some spin systems. However, some spin systems might not be assigned. Each spin system contains N, $\mathrm{H^N}$, C^α and C^β chemical shifts such that

$$\mathbf{s}_\mathrm{j} = (\mathrm{N_j}, \mathrm{H_j^N}, \mathrm{C_j^\alpha}, \mathrm{C_j^\beta}, \tilde{\mathrm{C}}_\mathrm{j}^\alpha, \tilde{\mathrm{C}}_\mathrm{j}^\beta). \tag{1}$$

If $\mathbf{s}_\mathrm{j}$ is mapped to residue i, then $\tilde{\mathrm{C}}$ denotes Carbon chemical shifts of residue $i-1$, the preceding residue.

2.2 The General Strategy

The big picture of IPASS can be summarized as follows:

Forming Spin Systems: This is a pre-processing step for resonance assignment. A new graph-based method is developed to group chemical shifts from the peaks of different spectra into spin systems. The input to spin system-forming module is the peak lists of ^{15}N-HSQC, HNCA, CBCA(CO)NH, and HNCACB spectra, and the output is a set of spin systems.

Typing Spin Systems: Estimating the potential amino acids, which can generate the observed chemical shifts in a spin system, is called typing. Different amino acids exhibit different chemical shift statistics. Spin systems are typed in a probabilistic framework by using the collected statistics. The set of possible spin systems for each

residue is determined by combining the typing information with the secondary structure information, provided by PSIPRED [13]. The input to the typing module is the amino acid sequence, spin systems, and the secondary structure prediction. The output is a set of potential spin systems and their probabilities, associated with each residue.

Connectivity Information Extraction: Two spin systems are connected if they can be mapped to two consecutive residues. The connections are detected by inter-residue and intra-residue information. Chemical shifts within spin systems are noisy, such that a low threshold results in many undetected true connections. However, a large threshold results in many false connections, making the ILP problem intractable. In IPASS, two sets of connections are defined: a set of highly reliable connections based on the the C^{α} and C^{β} chemical shifts and the information extracted from the ^{15}N-edited NOESY peaks. Furthermore, a set of less-reliable connections are detected by a larger threshold. By using reliable connections, a set of fragments is determined and the combinations of them are enumerated. Fixing the fragments eliminates many false connections and makes the ILP problem feasible.

Integer Linear Programming: At this step, there are some spin system candidates, and their probabilities for each residue. Then, the assignment is formulated as an ILP problem to find the globally optimal assignment. The ILP is solved for all combinations of the fragments, and the one with the best score is picked as the final assignment.

2.3 Forming Spin Systems

The goal in this step is to group the chemical shifts that are determined from the different NMR spectra into spin systems. Each spin system corresponds to nuclei within a small vicinity, usually associated with a residue of the target protein. During the spin system forming process, the chemical shifts are grouped in relation to their local environment and are not assigned to a certain residue in the protein sequence. Here, spin systems are viewed as the building blocks of the backbone assignment process.

The NMR spectra used here are 2D ^{15}N-HSQC and triple resonance experiments CBCA(CO)NH, HNCA, and HNCACB. For the two consecutive residues as illustrated in Figure 1, these experiments provide the following information.

- The ^{15}N-HSQC experiment detects the H^N and N chemical shift pair, i.e. a peak at (N_i, H_i^N) for residue i and is referred to as the "root pair". It is noteworthy that ^{15}N-HSQC has the highest resolution and sensitivity.
- The HNCACB experiment detects H^N, N, C^{α}, and C^{β} chemical shifts. In the ideal case, it generates four peaks for each residue: $(N_i, C_i^{\alpha}, H_i^N)$, $(N_i, C_i^{\beta}, H_i^N)$, $(N_i, C_{i-1}^{\alpha}, H_i^N)$, and $(N_i, C_{i-1}^{\beta}, H_i^N)$. Two peaks are associated with C^{α} and C^{β} of residue i, and two with those of i-1. The sign of the intensity values can be used to differentiate between C^{α}s and C^{β}s, because they exhibit opposite signs.
- The CBCA(CO)NH experiment detects H^N, N, C^{α}, and C^{β} chemical shifts. In the ideal case, two peaks are generated for each residue: $(N_i, C_{i-1}^{\alpha}, H_i^N)$ and $(N_i, C_{i-1}^{\beta}, H_i^N)$.
- The HNCA experiment detects H^N, N, C^{α} chemical shifts. Ideally, it generates two peaks for each residue: $(N_i, C_i^{\alpha}, H_i^N)$ and $(N_i, C_{i-1}^{\alpha}, H_i^N)$.

The problem of finding spin systems is modeled as a graph theoretical problem. A solution, based on a simple clustering method is provided, by using the connected components of the graph. Ideally, the chemical shifts should be the same for each atom in the NMR spectra. In practice, a perfect peak set is not available due to experimental errors, artifacts, biases, and the resolution differences. Typically, a tolerance as high as 0.5 ppm is expected to exist in the ^{15}N and ^{13}C chemical shifts, and a tolerance as high as 0.05 ppm in the ^{1}H chemical shifts. Therefore, an exact match algorithm is not possible for comparing the different experimental NMR data. To overcome this problem, each peak is represented as a point, a node in the graph, in the multidimensional space, where each dimension corresponds to a certain type of nuclei such as ^{15}N, ^{1}H, or ^{13}C. Initially, the ^{15}N-HSQC, CBCA(CO)NH, HNCA and HNCACB peaks are represented in the atom space. Conceptually, the peaks that belong to the same residue should coincide at the same H^N and N position. In reality, they usually do not coincide, but are clustered nearby.

First, the peaks within each 3D spectrum are connected according to their N and H^N chemical shifts. Each spectrum provides multiple peaks for the same residue, and these peaks should be in the small vicinity of each

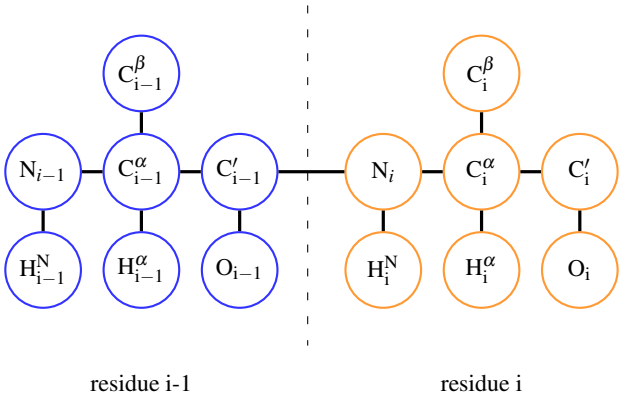

Fig. 1. Abridged diagram of atoms of two consecutive residues (note that all the side chain atoms are not shown).

other. The peaks that have similar root pairs are grouped by using an Euclidian distance function. Given two peaks with root pairs $P_x = (N_x, H^N_x)$ and $P_y = (N_y, H^N_y)$, the distance between them is defined as

$$d_{P_x,P_y} = \sqrt{(N_x - N_y)^2 + \omega^2 (H^N_x - H^N_y)^2}, \tag{2}$$

where ω is the scaling factor for the compensation of the difference in the resolution between ^{1}H and ^{15}N. Usually, ^{1}H chemical shifts are 10 times more sensitive than the ^{15}N chemical shifts, and so the default value of ω is 10. According to the distance defined in equation (2), each peak, P, in a given spectrum is associated with its nearest neighbor, P_{NN}. An edge is created between P, and all the peaks that are closer to P than $2 \times d_{P,P_{NN}}$. The edges between the peaks are directional, and the source is the reference peak, P. The peaks which are connected to each other represent the peaks from the same root pair.

The second step of generating a peak graph is to connect the peaks from different spectra. For example, the distance between $P_x = (N_x, C_x, H^N_x)$ in CBCA(CO)NH spectrum and $Q_y = (N_y, C_y, H^N_y)$ in HNCA is defined as

$$D_{P_x,Q_y} = \sqrt{(N_x - N_y)^2 + (C_x - C_y)^2 + \omega^2 (H^N_x - H^N_y)^2} \tag{3}$$

Similar to the aforementioned process, the edges can be created between P and its close vicinity peaks in other spectra, which are closer to P than $2 \times D_{P,P_{NN}}$. All of the created edges are directional. If there are two edges in both directions between two nodes, two edges are replaced by a non-directional edge.

After these two steps, each connected component represents a cluster that corresponds to a spin system in the resulting general peak graph. The primary advantage of this approach is its generalization. It can be applied to any set of available NMR spectra. After the connected components are found, each cluster contains similar H^N and N values such that these values are taken from the ^{15}N-HSQC spectrum. The important challenge is to detect C^{α}, C^{β}, $\tilde{C}^{\alpha}$, and $\tilde{C}^{\beta}$. The clusters are incomplete as a result of the missing peaks, and over-crowded as a result of the very similar spin systems.

A brute force approach that searches all the possible combinations of the chemical shift values for different C^{α} and C^{β} nuclei in each cluster is chosen. If a unique combination of the chemical shifts exists and does not conflict with the peaks in the cluster, a spin system is generated. After C^{α} and C^{β} are identified, $\tilde{C}^{\alpha}$, $\tilde{C}^{\beta}$ can be easily identified.

2.4 Typing Spin Systems

After the spin systems are formed, the next step is to type spin systems. Initially, any of the m spin systems can be mapped to any of the n residues. The objective of this step is to reduce the number of candidate spin systems for each residue, based on the chemical shift information. A statical analysis of the deposited chemical shifts in

the BMRB database reveals correlation among the chemical shifts, and amino acid type, and secondary structure. These statistics are used to find the probability that one spin system is mapped to a certain residue.

Collecting Statistics

All the BMRB entries with a matched PDB entry were downloaded as of December 15, 2008. Then, 1168 protein sequences were clustered by using CD-HIT [14] with a 40% sequence identity level. From each cluster, only the longest sequence was retained, resulting in a data set of 805 non-redundant proteins. DSSP [15] was selected to compute the secondary structure types for all the residues. From 88,436 collected residues, for Gly (which does not have a C^{β} chemical shift) 6,577 C^{α} chemical shifts, and for all the other amino acids, 68,028 C^{α} and C^{β} chemical shift pairs were extracted. The mean and covariance matrices were estimated for each amino acid and secondary structure type.

Probabilistic Typing

In this section, the task is to compute the probability that spin system $\mathbf{s}_\mathrm{j}$ can be mapped to residue r_i or $\Pr\{r_i \,|\, \mathbf{s}_\mathrm{j}\}$ for the n residues and m spin systems. Two vectors are defined for spin system $\mathbf{s}_\mathrm{j}$: $\mathbf{c}_\mathrm{j} = (\mathrm{C}_\mathrm{j}^{\alpha}, \mathrm{C}_\mathrm{j}^{\beta})^T$ and $\tilde{\mathbf{c}}_\mathrm{j} = (\tilde{\mathrm{C}}_\mathrm{j}^{\alpha}, \tilde{\mathrm{C}}_\mathrm{j}^{\beta})^T$. They contain the chemical shift information about the residue which $\mathbf{s}_\mathrm{j}$ is mapped to, r_i, and its preceding residue, r_{i-1}, respectively. Furthermore, since the N and H^{N} chemical shifts exhibit similar statistics for all amino acids, N and H^{N} are discarded, and only $\mathbf{c}_\mathrm{j}$ and $\tilde{\mathbf{c}}_\mathrm{j}$ from each spin system are considered. Therefore, $\Pr\{r_i \,|\, \mathbf{s}_\mathrm{j}\}$, the probability that $\mathbf{c}_\mathrm{j}$ and $\tilde{\mathbf{c}}_\mathrm{j}$ are mapped to r_i and r_{i-1}, respectively, can be written as in equation (4). If it is assumed that $\mathbf{c}_\mathrm{j}$ and $\tilde{\mathbf{c}}_\mathrm{j}$ are independent, equation (4) can be simplified to (5). By using the Bayes' rule, equation (5) is rewritten as (6) where $a_p, a_q \in A$, and A is the set of twenty amino acids.

$$\Pr\{r_i \,|\, \mathbf{s}_\mathrm{j}\} = \Pr\{r_i = a_p, r_{i-1} = a_q \,|\, \mathbf{c}_\mathrm{j}, \tilde{\mathbf{c}}_\mathrm{j}\} \tag{4}$$

$$= \Pr\{r_i = a_p \,|\, \mathbf{c}_\mathrm{j}\} \times \Pr\{r_{i-1} = a_q \,|\, \tilde{\mathbf{c}}_\mathrm{j}\} \tag{5}$$

$$= \frac{\Pr\{\mathbf{c}_\mathrm{j} \,|\, r_i = a_p\}\Pr\{r_i = a_p\}}{\Pr\{\mathbf{c}_\mathrm{j}\}} \times \frac{\Pr\{\tilde{\mathbf{c}}_\mathrm{j} \,|\, r_{i-1} = a_q\}\Pr\{r_{i-1} = a_q\}}{\Pr\{\tilde{\mathbf{c}}_\mathrm{j}\}} \tag{6}$$

In equation (6), $\Pr\{r_i = a_p\}$ only depends on a_p and not the position i. Therefore, it can be easily estimated by the proportional abundance of amino acid a_p. In addition, by using the total probability law,

$$\Pr\{\mathbf{c}_\mathrm{j}\} = \sum_{a_\ell \in A,\, a_\ell \neq \mathrm{Pro}} \Pr\{\mathbf{c}_\mathrm{j} \,|\, r_i = a_\ell\}\Pr\{r_i = a_\ell\} \tag{7}$$

$$\Pr\{\tilde{\mathbf{c}}_\mathrm{j}\} = \sum_{a_\ell \in A} \Pr\{\tilde{\mathbf{c}}_\mathrm{j} \,|\, r_{i-1} = a_\ell\}\Pr\{r_{i-1} = a_\ell\}$$

It should be noted that in Pro, $^{15}\mathrm{N}$ and $^{1}\mathrm{H}$ do not resonate and no spin system is mapped to it. In other words, $\Pr\{\mathbf{c}_\mathrm{j} \,|\, r_i = \mathrm{Pro}\} = 0$. Furthermore, the chemical shifts depend on both the amino acid and the secondary structure type. To incorporate the secondary structure information, the total probability law is used again and $\Pr\{\mathbf{c}_\mathrm{j} \,|\, r_i\}$ is reformulated as

$$\Pr\{\mathbf{c}_\mathrm{j} \mid r_i = a_\ell\} = \sum_{k=1}^{3} \Pr\{\mathbf{c}_\mathrm{j} \mid r_i = a_\ell, \gamma_i = \sigma_k\} \Pr\{\gamma_i = \sigma_k\}, \tag{8}$$

where γ_i denotes the secondary structure of r_i. For $k = 1, 2$, and 3, σ_k denotes random coil, β-strand, and α-helix, respectively. PSIPRED is used to estimate $\Pr\{\gamma_i = \sigma_k\}$ values [13].

It is assumed that $\Pr\{\mathbf{c}_\mathrm{j} | r_i,\, \gamma_i\}$ exhibits a joint Gaussian distribution due to the observed strong correlation between the C^{α} and C^{β} chemical shifts. By using the estimated covariance matrices ($\Sigma_{\ell,k}$) and mean vectors ($\mu_{\ell,k}$),

$$\Pr\{\mathbf{c}_\mathrm{j} \,|\, r_i = a_\ell,\, \gamma_i = \sigma_k\} = \frac{1}{2\pi|\Sigma_{\ell,k}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{c}_\mathrm{j} - \mu_{\ell,k})^T \Sigma_{\ell,k}^{-1} (\mathbf{c}_\mathrm{j} - \mu_{\ell,k})\right) \tag{9}$$

Therefore, when one of $^{13}\mathrm{C}$ chemical shifts is missing, the one-dimensional version of Gaussian distribution is used. By substituting equation (9) in (8), $\Pr\{\mathbf{c}_\mathrm{j} \,|\, r_i = a_\ell\}$ is computed. After computing $\Pr\{\mathbf{c}_\mathrm{j}\}$, the mappings that

are very unlikely are discarded. Therefore, if the condition in equation (10) holds for a_ℓ, $\Pr\{\mathbf{c}_\mathrm{j} \mid r_i = a_\ell\}$ is set to zero.

$$\frac{\Pr\{\mathbf{c}_\mathrm{j} \mid r_i = a_\ell\}\Pr\{r_i = a_\ell\}}{\Pr\{\mathbf{c}_\mathrm{j}\}} < \varepsilon \Rightarrow \Pr\{\mathbf{c}_\mathrm{j} \mid r_i = a_\ell\} = 0 \tag{10}$$

The omission threshold ε is chosen as 0.001 like the similar approaches in the literature [16]. This helps to reduce the number of candidate spin systems for each residue.

After this step, the $\Pr\{r_i \mid \mathbf{s}_\mathrm{j}\}$ values for $i = 1,\ldots,n$ and $j = 1,\ldots,m$ are established. The next step is to find the connections among spin systems.

2.5 Connectivity Information Extraction

The connectivity information is extracted from the C^α and C^β chemical shifts, as well as the ^{15}N-edited NOESY peaks information. Unlike previous studies, in this paper two sets of connections, reliable and loose, are defined. Although the spin system typing step can significantly reduce the number of candidate spin systems for each residue, this number is still large. Therefore, some highly reliable fragments are assigned and fixed. When spin system $\mathbf{s}_\mathrm{j}$ is fixed on reside r_i, $\mathbf{s}_\mathrm{j}$ should be removed from all other candidate sets, because the assignment is a one to one mapping. From another point of view, this step can be interpreted as performing a local optimization to make the global optimization feasible. If $\mathbf{s}_\mathrm{j}$ and $\mathbf{s}_\mathrm{k}$ satisfy two of the three following conditions, they are *reliably* connected

1. $|\mathrm{C}^\alpha_\mathrm{j} - \tilde{\mathrm{C}}^\alpha_\mathrm{k}| \le \delta_\alpha$
2. $|\mathrm{C}^\beta_\mathrm{j} - \tilde{\mathrm{C}}^\beta_\mathrm{k}| \le \delta_\beta$
3. $(\mathrm{N}_\mathrm{j}, \mathrm{H}^\mathrm{N}_\mathrm{k}, \mathrm{H}^\mathrm{N}_\mathrm{j})$ *and* $(\mathrm{N}_\mathrm{k}, \mathrm{H}^\mathrm{N}_\mathrm{j}, \mathrm{H}^\mathrm{N}_\mathrm{k})$ peaks exist in the ^{15}N-edited NOESY spectrum

Since these connections are crucial, $\delta_\alpha = \delta_\beta = 0.05$ ppm, are chosen and are one tenth of the maximum acceptable tolerance. If two spin systems are assigned to two adjacent residues on the target protein sequence, their hydrogen atoms of amide groups should be close in 3D space, providing a peak in the ^{15}N-edited NOESY spectrum.

For the loose connections, we set $\delta_\alpha = \delta_\beta = 0.5$. $\mathbf{s}_\mathrm{j}$ and $\mathbf{s}_\mathrm{k}$ are loosely connected if they can satisfy condition one or two without violating the other one. Due to the nature of the ^{15}N-edited NOESY peaks, condition three, alone, is not enough, because, for example $\mathrm{H}^\mathrm{N}_\mathrm{j}$ can be from a residue that is far from residue k in the sequence, but but close in the space.

Enumerating reliable fragments

In this step, it is assumed that p fragments, $F_1,\ldots,F_p$, are found, with lengths' $l_1,\ldots,l_p$, respectively. Each fragment is shown as $F_q = (\mathbf{s}_{e_1}, \mathbf{s}_{e_2}, \ldots, \mathbf{s}_{e_{l_q}})$, where $\mathbf{s}_{e_j}$ is connected to $\mathbf{s}_{e_{j+1}}$ for $j = 1,\ldots,l_q - 1$. Fragments shorter than three spin systems, or fragments that are the substrings of other fragments are discarded. For fragment F_q, a score is defined for the i-th position in the target sequence such that

$$T_i^{(q)} = -\sum_{k=1}^{l_q} \log\left(1 - \Pr\{r_{i+k-1} \mid \mathbf{s}_{e_k}\}\right), \quad 1 \le i \le n - l_q + 1. \tag{11}$$

It can be seen that $T_i^{(q)} \in [0,\infty)$. If $T_i^{(q)} = 0$, F_q cannot be mapped to position i. If $T_i^{(q)} > l_q\varepsilon$, then i is added to the set of possible mappings of F_q, and this set is shown as M_q. It should be noted that here the $\log(1-\varepsilon) \approx -\varepsilon$ approximation is used. If M_q is empty, F_q is discarded. After all the possible mappings are found, all combinations are enumerated. In each combination, no two fragments should be in conflict, i.e., they should not share any spin systems, and their mapped positions in the sequence should not overlap. Then, all the fragments within the combination are fixed. For example, if F_q's mapping starts from the i-th position, then

$$\Pr\{r_\ell \mid \mathbf{s}_{e_k}\} = \begin{cases} 1 & \text{if } i \le \ell < i + l_q \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

In other words, $\mathbf{s}_{e_k}$ is assigned to r_ℓ and removed from the candidate set of the other residues. The number of combinations is limited to 20000. In the experiments in this paper, no more than 200 combinations are discovered,

because a strict threshold is used for finding the reliable fragments. If the number of combinations exceeds the upper bound, the fragments of length four are discarded and so on. This process is continued until the number of combinations becomes fewer than the predefined upper bound. After ILP is solved for each combination, the one with the highest score is chosen as the final assignment.

2.6 IPASS: Integer Linear Programming-based Assignment

Originally, all the m spin systems can be mapped to any residue r_i. However, in spin system typing and the fragment fixing step, many $\Pr\{r_i \,|\, \mathbf{s}_\mathrm{j}\}$ values are set to zero.

Then, the backbone resonance assignment problem can be represented by a graph $G(V,E)$. Here, each node in V corresponds to a spin system, and the edges in E represent the connections in the spin systems. Notice that a spin system can be mapped to multiple locations with different probabilities, and multiple copies of the spin systems, which are differentiated according to their mapped location, exist. If $\Pr\{r_i \,|\, \mathbf{s}_\mathrm{j}\} \neq 0$, then $v_{i,j} \in V$, and variable $v_{i,j}$ is created in the ILP. Furthermore, if $v_{i,j}, v_{i+1,k} \in V$ and $\mathbf{s}_\mathrm{j}$ is connected to $\mathbf{s}_\mathrm{k}$, then $e_{i,j,k} \in E$, and variable $e_{i,j,k}$ is created in the ILP. Figure 2 illustrates the setup of the assignment problem. The two defined sets of variables are

- $v_{i,j}$: it is 1 if and only if $\mathbf{s}_\mathrm{j}$ is assigned to r_i and 0 otherwise.
- $e_{i,j,k}$: it is 1 if and only if $\mathbf{s}_\mathrm{j}$ is assigned to r_i and $\mathbf{s}_\mathrm{k}$ is assigned to r_{i+1}, and 0 otherwise.

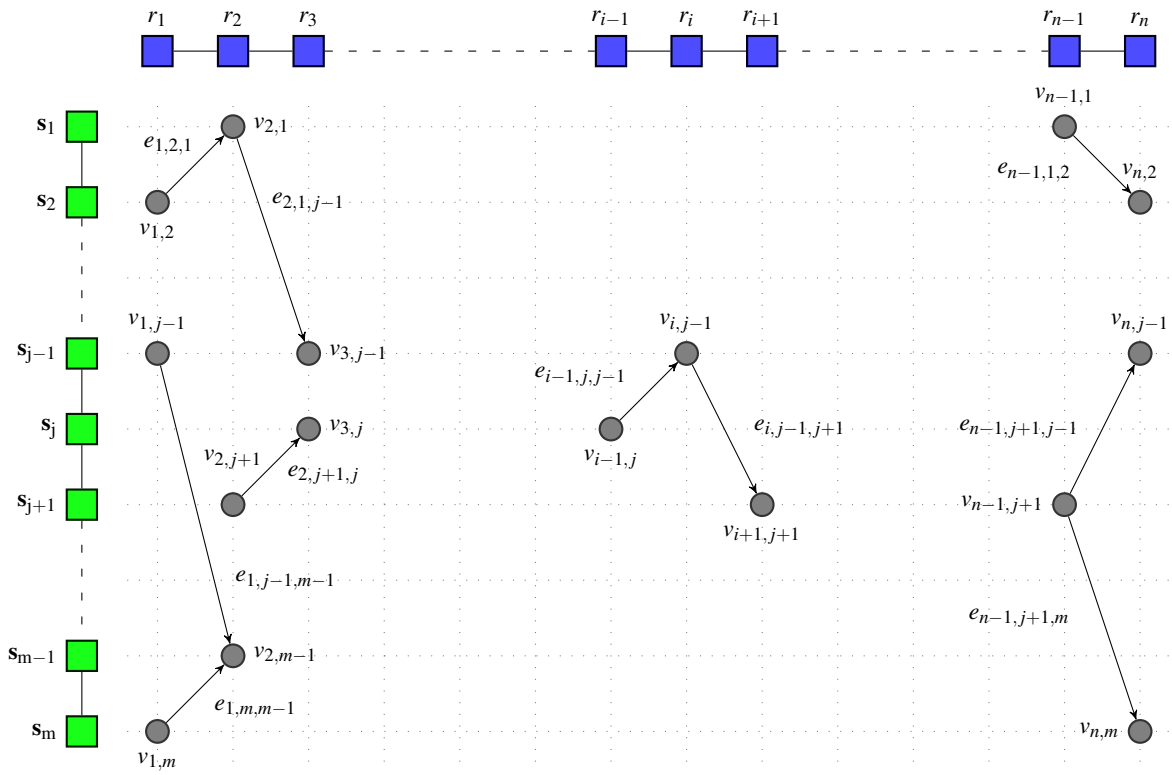

Fig. 2. Illustration of the problem setup of the assignment problem. There is a node $v_{i,j}$ (shown by the gray circles), corresponding to r_i and $\mathbf{s}_\mathrm{j}$, only if $\Pr\{r_i \,|\, \mathbf{s}_\mathrm{j}\} \neq 0$.

For each edge, a corresponding weight is defined as

$$\begin{aligned} w_{i,j,k} &= \log\left(\Pr\{r_i;\, r_{i+1} \,|\, \mathbf{s}_\mathrm{j},\, \mathbf{s}_\mathrm{k}\}\right) \\ &= \log\left(\Pr\{r_i \,|\, \mathbf{s}_\mathrm{j}\}\right) + \log\left(\Pr\{r_{i+1} \,|\, \mathbf{s}_\mathrm{k}\}\right), \end{aligned} \tag{13}$$

where$\{r_i \mid \mathbf{s}_\mathrm{j}\}$ and $\{r_{i+1} \mid \mathbf{s}_\mathrm{k}\}$ are assumed independent. $w_{i,j,k}$ corresponds to the probability of mapping two spin systems to two adjacent residues. Now, the task is to find the assignment which maximizes the total weight of the chosen edges. Inherently, each spin system can be assigned to, at most, one residue in the protein sequence. For each residue, there can be, at most, one spin system assigned. After the backbone assignment problem, which is an NP-hard problem (see Theorem 1), is formulated, the ILP formulation is

$$\max_{e_{i,j,k}} \sum_{e_{i,j,k} \in E} \left(w_{i,j,k} + \lambda\right) e_{i,j,k}, \tag{14}$$

$$\text{subj. to } \forall e_{i,j,k} \in E \quad e_{i,j,k} \leq v_{i,j};\ e_{i,j,k} \leq v_{i+1,k}, \tag{15}$$

$$\forall i \in \{1,\ldots,n\},\ \textstyle\sum_{j=1}^{m} v_{i,j} \leq 1, \tag{16}$$

$$\forall j \in \{1,\ldots,m\},\ \textstyle\sum_{i=1}^{n} v_{i,j} \leq 1, \tag{17}$$

$$\text{and } v_{i,j} \in \{0,1\},\quad e_{i,j,k} \in \{0,1\}. \tag{18}$$

Since the logarithm values are negative, the objective function adjusts all the edge weights to positive values by adding the $\lambda = -\min_{i,j,k}(w_{i,j,k})$ term. Then, the maximization is meaningful. Constraint (15) ensures that an edge can be selected, only if both of its ends are selected. Constraint (16) ensures that a spin system can be assigned to, at most, one residue, and constraint (17) ensures that a residue can be assigned to, at most, one spin system.

As the result of the fragment fixing step, the size of the problem, i.e., $|V|+|E|$ plus number of constraints, is substantially reduced, which makes the ILP problem tractable. CPLEX is used for solving the aforementioned ILP. For each enumerated combination, an ILP is generated and the solution is attained. The total cost function of the assignment represents the score of that configuration. The assignment with the highest score is reported as the final assignment.

In the end, the NP-hardness of the backbone resonance assignment problem is proven.

Theorem 1. *Backbone resonance assignment problem, under the proposed graphical representation is NP-hard.*

proof. The NP-hardness of the backbone resonance assignment problem is established under the graph representation through a reduction from the *Hamiltonian path* problem which is known to be NP-hard. The Hamiltonian path problem is defined as follows: Given a graph, $G(V,E)$, decide whether there exists a path in $G(V,E)$ that visits each vertex exactly once. For an instance of the Hamiltonian path problem, a new graph $G'(V',E')$, which is a product of the $\{1,2,...,n\} \times G$, where $n = |V|$ is constructed. Thus the new graph, $G'(V',E')$, has nodes of (i,v), where $v \in V$ and $1 \leq i \leq n$, and edges between (i,v) and (j,w). This occurs if an edge between v and w exists in G. Here, the edge weights are defined as 1 for all edges in G'.

G has a hamiltonian path, if and only if there exists a perfect assignment for the backbone resonance assignment problem. For each i, the vertices are connected to their adjacent vertices in the graph with the weight 1. A perfect backbone resonance assignment corresponds to a mapping, where each spin system is used once, and each residue is assigned to a single spin system with a total cost of $n-1$. Each spin system corresponds to a vertex in G and the residues correspond to the $\{1,2,...,n\}$ set. As a result, the perfect assignment visits each vertex once which corresponds to a Hamiltonian path. Similarly, if there is a Hamiltonian path visiting vertices $v_1, v_2, ..., v_n$, it corresponds to an assignment between v_i and residue i. Consequently, this problem is NP-hard. □

3 Experimental Results

To evaluate the performance of the proposed method, several experiments are conducted. Two performance measures are used in the following parts: *precision* and *recall*. *Precision* measures the ability to reject false assignments, whereas *recall* measures the ability to discover true assignments. Assume that for the target protein, there are N_r manually assigned residues, and a resonance assignment program assigns N_o residues, where T_p of them are assigned correctly. Then, recall (RCL) and precision (PRC) are defined as T_P/N_r and T_P/N_o, respectively.

Performance on Real NMR Lab Data Sets

In practice, the input for resonance assignment is not "perfect". Instead, the input peak lists contain various sources of error, such as the chemical shift differences of the same nucleus in different spectra and false peaks, picked

during the peak picking step. Therefore, an assignment method is practical only if it works on "low quality" real noisy input data sets.

In the NMR lab experiments, the spectroscopists usually conduct the whole NMR process altogether, i.e., the resonance assignment, NOE assignment, structure calculation information, as well as information from the various kinds of other spectra, which are used as feedback to refine the peak lists. Thus, the final peak lists provided by NMR labs are always "almost perfect", and do not represent the original peaks picked by spectroscopists. Recently, an automated peak picking method, PICKY [17], has been developed, which can automatically pick peaks from any NMR spectrum. PICKY is tested on 19 noisy spectra, provided by the collaborators. The average lower bound of RCL and PRC are 81% and 87%, respectively. The detailed performance measures of PICKY are listed in Table 1. Compared to the refined peak lists after the whole structure determination process, the peak lists, generated by PICKY, are closer to the original peaks picked by spectroscopists. As a result, the peak lists generated by PICKY are used to evaluate the performance of IPASS on real data sets.

Table 2 summarizes the performance of RIBRA, MARS, and IPASS for three real data sets. Specifically, protein TM1112 from *Thermotoga maritima* is provided by the Arrowsmith Lab at the University of Toronto [18] whereas CASKIN, the SH3 domain of the CASKIN neuronal signaling protein, and VRAR, *S. Aureus* VraR DNA binding domain [19], are provided by the Donaldson Lab at York University.

Since MARS cannot use the peak lists, it takes IPASS spin systems as the input. The performance of MARS and IPASS are compared on the same set of spin systems. RIBRA takes the peak lists of ^{15}N-HSQC, CBCA(CO)NH, and HNCACB as the input, so the performance of RIBRA and IPASS are compared on the same peak lists. Table 2 clearly shows that IPASS performs significantly better than RIBRA and MARS on all of the three real data sets. One thing to notice is that when the input peak list quality is as good as TM1112, IPASS can generate assignments, which are almost as good as the manual assignment. In Table 2, the number of Gly and Pro residues are shown. The Pro residues cut the fragments and make the assignment more challenging. The Gly residues are favorable in a way that can be typed very easily due to their distinct C^{α} values. However, The Gly residues shorten the fragments, because they do not have any C^{β} chemical shifts, and hence, no reliable connections.

Table 1. Performance of PICKY on TM1112, CASKIN, COILIN, and VRAR.

Protein	Length	^{15}N-HSQC	HNCA	CBCA(CO)NH	HNCACB	*Average*
TM1112	89	96 / 89	93 / 88	98 / 88	91 / 83	95 / 87
CASKIN	67	100 / 93	-	91 / 68	70 / 75	87 / 77
VRAR	72	87 / 93	-	83 / 71	69 / 72	80 / 79

TM1112 is provided by the Arrowsmith lab, while CASKIN, and VRAR are provided by the Donaldson lab. The first and second columns show the target protein names and lengths, respectively. Starting from the third column, for each spectrum of each protein, the *recall*/*precision* values are listed in the table in percentile.

Performance on Other Data Sets

Although the goal is to develop a backbone resonance assignment method which works on realistic data sets of automatically picked peak lists, comparison between IPASS and other programs is provided by using some previously used benchmark data.

Simulated Spin Systems as Input: First, the IPASS performance is evaluated on a simulated data set, used by [7], which contains 12 proteins. For each protein, the spin systems are simulated, based on the BMRB deposited chemical shift assignments of proteins, and used as the input for all of these programs. Each spin system contains N, H^{N}, C^{α}, C^{β}, $\tilde{C}^{\alpha}$, and $\tilde{C}^{\beta}$ chemical shifts. Since RANDOM and CISA are not available, the PRC and RCL values are selected from [7]. The accuracy of RANDOM, MARS, and CISA is calculated according to two different sets of threshold values, because these programs are sensitive to different threshold values. Note that in these experiments, the input for IPASS is simulated spin systems, so the spin system forming step is not tested here. Furthermore, it should be noted that AUTOASSIGN was not available at the time to be included in our experiments.

Table 2. Performance (PRC/RCL[1]) of RIBRA, MARS, and IPASS on target proteins TM1112, CASKIN, and VRAR.

Protein	Length	Manually Assigned	Spin Systems	Gly/Pro	C^{β}	RIBRA[2]	MARS[3]	MARS[4]	IPASS
TM1112	89	83	81 / 85	4 / 5	78	40 / 54	6 / 45	55 / 63	71 / 73
CASKIN	67	54	47 / 48	7 / 4	42	12 / 21	23 / 25	23 / 25	31 / 41
VRAR	72	60	47 / 47	1 / 0	41	4 / 13	6 / 17	6 / 17	34 / 42

The first and second column show the target protein name and length, respectively. The third column shows the number of manually assigned residues by the Arrowsmith and the Donaldson labs, which is considered the upper bound for an automated method. The Fourth column shows the number of correct/total spin systems discovered by the spin system forming module. The Fifth column denotes the number of Pro/Gly in the sequence and the sixth column denotes number of available C^{β} values in the spin systems. Starting from the seventh column, for each protein, the performance of each method is shown in "number of correctly assigned residues/total number of assigned residues" format.

[1] PRC and RCL stand for *precision* and *recall*, respectively.

[2] RIBRA's performance with ^{15}N, ^{13}C threshold values of 0.5 and ^{1}H threshold value of 0.05. No residue can be assigned if the default values are used. The parameters are set according to IPASS, which makes the comparison fair.

[3] MARS with the first set of default parameters: $\delta_{\alpha} = 0.5ppm$ and $\delta_{\beta} = 0.5ppm$.

[4] MARS with the second set of default parameters: $\delta_{\alpha} = 0.2ppm$ and $\delta_{\beta} = 0.4ppm$.

As it is shown in Table 3, IPASS performs very well and significantly better than any other program regardless of the set of threshold settings. The average PRC of IPASS is 99%, and IPASS achieves a 100% PRC on seven out of 12 target proteins. Meanwhile, IPASS can also achieve a high RCL value of 96%. It is noteworthy that MARS performs well on this data set. However, MARS has a relatively low RCL value, compared to that of IPASS. On the other hand, Table 3 demonstrates that RANDOM, MARS, or CISA are sensitive to the threshold settings. For this simulated data set, in which all actually the connected spin systems should yield perfect connectivity information, a smaller threshold value can give a much better accuracy. However, in practice, researchers do not know the quality of the spin systems. The potential difference in the chemical shift of the same nuclei in different spectra renders the selection of the proper threshold values challenging. In contrast, IPASS does not rely on any parameter settings and its parameters are chosen without using any special data set.

Simulated Peak Lists as Input: The IPASS performance is tested on the same data set, but with simulated peak lists. All the four steps of IPASS are tested in these experiments. However, the CISA paper [7] does not provide such a comparison on RANDOM, MARS, and CISA. Furthermore, RANDOM and CISA are not available. As a result, IPASS is compared with two available programs: MARS and RIBRA. MARS takes only formed spin systems as the input and RIBRA takes the peak lists as the input. RIBRA is used directly, and IPASS's spin system forming method is applied to form spin systems for MARS.

Table 4 shows that both MARS and IPASS perform well on the simulated peak lists, and are better than RIBRA. MARS achieves higher PRC and lower RCL values than IPASS.

4 Discussion

The new method, IPASS, outperforms other assignment methods on automatically picked peaks, and performs better or as well as others on simulated data sets. There is still one last question to answer: even if IPASS works better than other methods, is the accuracy enough for the ultimate goal, i.e., automatically determining the high resolution structures of proteins? IPASS and PICKY are combined with other available programs to calculate the structures of the aforementioned target proteins. First, PICKY is applied to ^{15}N-HSQC, HNCA, CBCA(CO)NH, and HNCACB spectra (see Table 1 for performance). Then, IPASS is used to conduct the backbone resonance assignment. The assignment of IPASS is then fed into SPARTA for fragment generation [20]. SPARTA takes protein sequence and resonance assignment as input, and selects 3-mer and 9-mer fragments, based on the backbone chemical shifts. Then, FALCON [21] is used for the structure calculation, based on the fragments selected by SPARTA. FALCON generates structural decoys by fragment Hidden Markov Model (HMM). To fairly evaluate

Table 3. Accuracy (PRC/RCL) of RANDOM, MARS, CISA, and IPASS for 12 protein data set (simulated spin systems) in percentile[1].

			$\delta_\alpha = 0.2ppm,\ \delta_\beta = 0.4ppm$			$\delta_\alpha = 0.4ppm,\ \delta_\beta = 0.8ppm$			
Protein ID	Length	Assignable[2]	RANDOM	MARS	CISA	RANDOM	MARS	CISA	**IPASS**
bmr4391	66	59	67 / 63	100 / 76	97 / 97	58 / 55	100 / 75	91 / 91	93 / 90
bmr4752	68	66	40 / 35	100 / 97	96 / 94	36 / 30	100 / 97	90 / 88	100 / 94
bmr4144	78	68	36 / 33	100 / 91	100 / 99	33 / 31	100 / 69	100 / 99	98 / 85
bmr4579	86	83	54 / 51	99 / 98	98 / 98	34 / 32	96 / 90	80 / 80	100 / 98
bmr4316	89	85	42 / 36	100 / 100	100 / 99	35 / 30	99 / 91	83 / 83	99 / 98
bmr4288	105	94	62 / 55	100 / 99	98 / 98	42 / 38	98 / 97	91 / 91	100 / 98
bmr4929	114	110	68 / 63	100 / 100	93 / 91	46 / 43	100 / 99	96 / 94	100 / 100
bmr4302	115	107	66 / 64	100 / 100	96 / 95	47 / 45	100 / 100	91 / 91	100 / 99
bmr4670	120	102	67 / 62	100 / 100	96 / 95	43 / 39	100 / 100	88 / 87	98 / 97
bmr4353	126	98	48 / 43	95 / 55	96 / 95	47 / 43	95 / 55	90 / 90	99 / 93
bmr4027	158	148	43 / 32	100 / 99	100 / 99	40 / 30	100 / 99	88 / 85	100 / 97
bmr4318	215	191	40 / 38	99 / 99	87 / 84	25 / 22	100 / 95	74 / 70	100 / 98
Average	112	101	53 / 48	99 / 93	96 / 95	41 / 37	99 / 89	88 / 87	**99 / 96**

[1] These 12 proteins are selected by CISA paper [7]. The spin systems are simulated based on BMRB deposited chemical shift assignment of these proteins and used as input for all of these programs. Since RANDOM and CISA are not available, here we have used precision and recall values from [7]. The accuracy of RANDOM, MARS, and CISA is calculated based on two sets of thresholds.
[2] This indicates number of residues that are manually assigned in the BMRB file.

Table 4. Accuracy (PRC/RCL) of RIBRA, MARS, and IPASS on 12 protein data set (simulated peak lists) in percentile[1].

Protein ID	Length	Assignable	Spin systems[2]	Gly/Pro	RIBRA[3]	MARS[4]	MARS[5]	**IPASS**
bmr4391	66	59	55	6/1	91 / 76	93 / 43	94 / 46	91 / 85
bmr4752	68	66	65	6/1	91 / 90	100 / 94	100 / 94	100 / 92
bmr4144	78	68	63	3/5	62 / 45	100 / 58	100 / 41	98 / 85
bmr4579	86	83	80	5/2	87 / 67	99 / 87	99 / 83	100 / 94
bmr4316	89	85	80	13/3	99 / 88	99 / 83	99 / 73	88 / 79
bmr4288	105	94	93	5/10	100 / 97	99 / 95	100 / 97	99 / 97
bmr4929	114	110	108	10/2	82 / 78	100 / 83	99 / 68	99 / 98
bmr4302	115	107	107	5/2	100 / 92	100 / 96	99 / 97	96 / 95
bmr4670	120	102	92	9/5	98 / 86	99 / 87	100 / 87	93 / 79
bmr4353	126	98	97	8/10	98 / 93	99 / 90	100 / 91	97 / 90
bmr4027	158	148	146	11/8	90 / 82	99 / 94	99 / 92	97 / 94
bmr4318	215	191	188	9/12	74 / 63	99 / 93	99 / 86	98 / 90
Average	112	101	98	5/8	89 / 80	99 / 84	99 / 80	**96 / 90**

[1] These 12 proteins are selected by CISA paper [7]. The peak lists are simulated based on BMRB deposited chemical shift assignment of these proteins. RIBRA directly accepts peak list whereas IPASS spin system forming module was used to generate spin systems for MARS and IPASS.

[2] This indicates number of correct spin systems discovered by the proposed spin system forming module.
[3] RIBRA's performance with ^{15}N and ^{13}C threshold values of 0.5 and ^{1}H threshold value of 0.05. Those parameters are set according to IPASS for the sake of fair comparison.
[4] MARS with the first set of default parameters $\delta_\alpha = 0.2ppm$, and $\delta_\beta = 0.4ppm$.
[5] MARS with the second set of default parameters, $\delta_\alpha = 0.5ppm$, and $\delta_\beta = 0.5ppm$ which is the same as IPASS.

the performance, the homologs of target proteins are removed from the FALCON database. This process results in high resolution final structures for TM1112, CASKIN, and VRAR, i.e., for these proteins, backbone RMSD to the native structure is below 1.6 Å. More comprehensive experiments are underway.

IPASS is implemented in C++. It takes IPASS fewer than 5 minutes to achieve its results for a practical noisy data set of a medium size protein (100-150 residues in length). In addition, the whole process requires only five seconds for a simulated data set. The difference in speed stems from the fact that for the simulated data set, most of the fragments are fixed. Consequently, ILP problem size is very small. The next step is to incorporate more NMR spectra in the assignment process and make the method iterative.

References

1. Zimmerman, D., Kulikowski, C., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R., Montelione, G.: Automated analysis of protein NMR assignments using methods from artificial intelligence. Journal of Molecular Biology **269** (1997) 592–610
2. Güntert, P., Salzmann, M., Braun, D., Wüthrich, K.: Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. Journal of Biomolecular NMR **18** (2000) 129–137
3. Coggins, B., Zhou, P.: PACES: protein sequential assignment by computer-assisted exhaustive search. Journal of Biomolecular NMR **26** (2003) 93–111
4. Jung, Y., Zweckstetter, M.: Mars - robust automatic backbone assignment of proteins. Journal of Biomolecular NMR **30** (2004) 11–23
5. Masse, J., Keller, R.: Autolink: automated sequential resonance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic. Journal of Magnetic Resonance **174** (2005) 133–151
6. Wu, K., Chang, J., Chen, J., Chang, C., Wu, W., Huang, T., Sung, T., Hsu, W.: Mars - robust automatic backbone assignment of proteins. Journal of Computational Biology **13** (2006) 229–244
7. Wan, X., Lin, G.: CISA: combined NMR resonance connectivity information determination and sequential assignment. IEEE/ACM Transactions on Computational Biology and Bioinformatics **4** (2007) 336–348
8. Lemak, A., Steren, C., Arrowsmith, C., Llinás, M.: Sequence specific resonance assignment via Multicanonical Monte Carlo search using an ABACUS approach. Journal of Biomolecular NMR **41** (2008) 29–41
9. Volk, J., Herrmann, T., Wüthrich, K.: Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. Journal of Biomolecular NMR **41** (2008) 127–138
10. Seavey, B., Farr, E., Westler, W., Markley, J.: A relational database for sequence-specific protein NMR data. Journal of Biomolecular NMR **1** (1991) 217–236
11. Hiller, S., Fiorito, F., Wüthrich, K., Wider, G.: Automated projection spectroscopy (APSY). Proceedings of the National Academy of Sciences **102** (2005) 10876–10881
12. Fiorito, F., Hiller, S., Wider, G., Wüthrich, K.: Automated resonance assignment of protein: 6D APSY-NMR. Journal of Biomolecular NMR **35** (2006) 27–37
13. McGuffin, L.J., Bryson, K., Jones, D.T.: The psipred protein structure prediction server. Bioinformatics **16**(4) (2000) 404–405
14. Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein for nucleotide sequences. Bioinformatics **22** (2006) 1658–1659
15. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**(12) (1993) 2577–2637
16. Grishaev, A., Steren, C.A.A., Wu, B., Pineda-Lucena, A., Arrowsmith, C., Llinás, M.: Abacus, a direct method for protein nmr structure computation via assembly of fragments. Proteins **61** (2005) 36–43
17. Alipanahi, B., Gao, X., Karakoc, E., Donaldson, L., Li, M.: PICKY: A Novel SVD-Based NMR Spectra Peak Picking Method. To appear in ISMB2009 (2009)
18. Xia, Y., Yee, A., Semesi, A., Arrowsmith, C.: Solution structure of hypothetical protein TM1112. PDB Database (2002)
19. Donaldson, L.W.: The nmr structure of the *Staphylococcus aureus* response regulator vrar dna binding domain reveals a dynamic relationship between it and its associated receiver domain. Biochemistry **47**(11) (2008) 3379–3388
20. Shen, Y., Bax, A.: Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. Journal of Biomolecular NMR **38** (2007) 289–302
21. Li, S., Bu, D., Xu, J., Li, M.: Fragment-HMM: a new approach to protein structure prediction. Protein Science (2008)