

RIBRA—an Error-Tolerant Algorithm for the NMR Backbone Assignment Problem

Kun-Pin Wu¹, Jia-Ming Chang¹, Jun-Bo Chen¹,
Chi-Fon Chang², Wen-Jin Wu³, Tai-Huang Huang^{2,3},
Ting-Yi Sung¹ and Wen-Lian Hsu^{1*}

¹ Institute of Information Science, Academia Sinica, Taiwan

² Genomics Research Center, Academia Sinica, Taiwan

³ Institute of Biomedical Sciences, Academia Sinica, Taiwan

Abstract

We develop an iterative relaxation algorithm, called RIBRA, for NMR protein backbone assignment. RIBRA applies nearest neighbor and weighted maximum independent set algorithms to solve the problem. To deal with noisy NMR spectral data, RIBRA is executed in an iterative fashion based on the quality of spectral peaks. We first produce spin system pairs using the spectral data without missing peaks, then the data group with one missing peak, and finally, the data group with two missing peaks. We test RIBRA on two real NMR datasets: hbSBD and hbLBD, and perfect BMRB data (with 902 proteins) and four synthetic BMRB data which simulate four kinds of errors. The accuracy of RIBRA on hbSBD and hbLBD are 91.4% and 83.6%, respectively. The average accuracy of RIBRA on perfect BMRB datasets is 98.28%, and 98.28%, 95.61%, 98.16% and 96.28% on four kinds of synthetic datasets, respectively.

Keywords: NMR resonance assignment, iterative relaxation algorithm, nearest neighbor, weighted maximum independent set

1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography are the only two methods that can determine three-dimensional structures of proteins to atomic level. NMR spectroscopy has the additional power that it can also study the dynamics of proteins and screen for interacting partners such as drug screening in solution, also to the atomic/residue level. However, before any of the detailed studies can be carried out, sequence specific backbone resonance assignments must be completed. Multi-dimensional NMR spectra contain cross-peaks, which contain resonance frequencies (*chemical shifts*) and correlation information. The cross-peak represents a covalent-bond linkage (COSY type) or a spatial relation (NOESY type) among a set of nuclei depending on the types of NMR experiments performed. Different kinds of NMR experiments provide different partial resonance information of residues so that biologists can decide which experiments should be performed to best suit their needs. For example, the two dimensional HSQC experiment concerns whether there is a covalent bond between two atoms N and H^N; if there

*The corresponding author. Postal Address: 128, Section 2, Academia Road, Institute of Information Science, Academia Sinica, Nankang, Taipei 115, Taiwan. Email: hsu@iis.sinica.edu.tw, Tel: +886-2-27883799 ext. 1804, Fax: +886-2-27824814

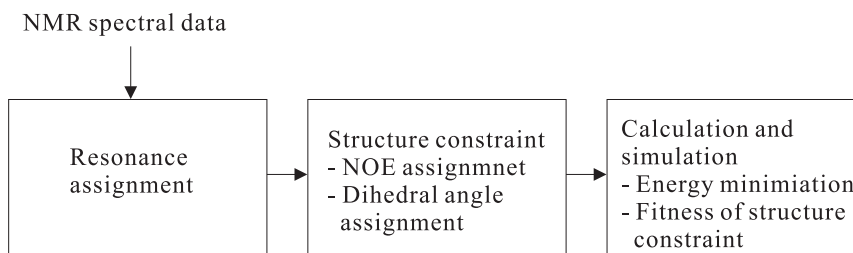


Figure 1: Workflow of protein 3D structure determination based on NMR data. The resonance assignment assigns chemical shifts to all atoms of target proteins. Then the spatial (via NOE assignment) and angular (via dihedral angle assignment) structural constraints can be determined. Finally, we determine protein 3D structures by finding the ones with minimum energy that satisfies the above structure constraints.

is, then a corresponding peak appears in the spectrum and its coordinate is composed of the chemical shifts of the two atoms. The general workflow of protein 3D structure determination based on NMR spectral data is depicted in Figure 1.

Once cross-peaks are available, we assign chemical shifts to the corresponding atoms within residues. This important stage is known as *resonance assignment* and is the main topic of this paper. Cross-peaks are usually extracted according to an intensity threshold. If the threshold is set too high, peaks with low intensities will be ignored and become missing peaks (false negatives). On the other hand, if it is set too low, noisy peaks with high intensity will be regarded as real peaks and become false positives. Even if spectral data contains no false negatives or false positives, there are still some other problems. The same kind of atoms may be located in similar environments. In such cases, these atoms may have nearly the same chemical shifts and be represented as a single cross-peak in a spectrum; namely, more than one peak may cluster together as a single point in the spectrum and become indistinguishable. Another problem is experimental errors. Theoretically, each atom in a residue has a unique chemical shift. However, different NMR experiments may generate slightly different chemical shifts for an atom due to differences in experimental conditions such as slight change in temperature or difference in digital resolution. Ambiguity arises when the results of different experiments are cross-referenced. In summary, there are mainly four kinds of problems in real spectral data: 1) false negatives, 2) false positives, 3) clustered peaks and 4) experimental errors. As these four kinds of data problems mix together in real data, resonance assignment becomes very challenging. For example, with respect to a residue, we cannot easily distinguish whether there is a missing peak, or the residue is a Glycine, which intrinsically has no C^β atom.

To overcome the above problems, researchers usually perform the following five procedures (Moseley and Montelione 1999):

1. Filter peaks and relate resonances from different spectra (filtering and referencing).
2. Group resonances into spin systems (grouping). A *spin system* contains the chemical shifts of atoms within a residue.
3. Identify the amino acid types of spin systems (typing).
4. Find and link sequential spin systems into segments (linking).
5. Map spin-system segments onto the primary sequence (mapping).

Note that different researchers may perform the above procedures in different order. Many works have been done to carry out part of these five procedures (Bailey-Kellogg et al.

2000, Bailey-Kellogg et al. 2004, Bartels et al. 1997, Chen et al. 2002, Chen et al. 2003, Coggins and Zhou 2003, Guntert et al. 2000, Hitchens et al. 2003, Hyberts and Wagner 2003, Langmead et al. 2003, Lin et al. 2002, Malmodin et al. 2003, Ou et al. 2001, Slupsky et al. 2003, Wang et al. 2003, Xu et al. 2002). Most models assume that the spin system grouping is already given (thus, eluding the possible ambiguity problem in grouping). Some model (such as the constrained bipartite graph approach of Xu et al. 2002) further prohibits linking ambiguity. Some other works perform all these five procedures (Atreya et al. 2000, Atreya et al. 2002, Buchler et al. 1997, Leutner et al. 1998, Li et al. 1997, Lukin et al. 1997, Zimmerman et al. 1997). Readers are referred to (Moseley and Montelione 1999) for a thorough survey.

The contribution of this paper is twofold. First, we design an error-tolerant algorithm, RIBRA (Relaxation and Iterative Backbone Resonance Assignment) based on iterative relaxation, to solve the backbone resonance assignment problem (*backbone assignment* for short) with good precision and recall. We use HSQC, CBCANH and CBCA(CO)NH spectral data to assign chemical shifts to atoms N, H^N , C^α and C^β along the backbone of a target protein. We use the rules of TATAPRO II (Atreya et al. 2002) to perform typing. Instead of performing the procedures as separate tasks, RIBRA adopts two operations *RGT* and *LM*, in which *RGT* does mixed referencing, grouping and typing, and *LM* does mixed linking and mapping. Based on data quality, we apply a relaxation technique to perform *RGT* and *LM* iteratively for noise filtering, and gradually carry out the entire backbone assignment.

An important source of NMR datasets is the BioMagResBank¹. Since the data in BMRB is normally error-free, researchers need to generate noises to test whether their algorithms can cope with these problems. However, in the past, these noises are generated in an ad-hoc fashion. Our second contribution is to create comprehensive synthetic datasets that reflect the following potential problems: false positives, false negatives and experimental errors. The original BMRB datasets can largely be regarded as “perfect standard datasets.” Note, however, there is no need to synthesize datasets containing clustered peaks since they are intrinsically embedded in the original BMRB data. These synthetic BMRB datasets are designed to serve as benchmark datasets for testing any future assignment algorithms. For current methods, none of them has been evaluated by such comprehensive datasets. Only PACES (Coggins and Zhou 2003) has tested a very small subset (21 proteins) of our dataset with linking errors. PACES is also a method that uses graph models and performs an exhausted search. However, there are significant differences between these two methods:

1. They accept different inputs. RIBRA accepts NMR spectral peaks, while PACES accepts spin systems and skips the grouping procedure.
2. Their goals are different. PACES aims at generating long segments; RIBRA aims at finding higher coverage of assignments.
3. They generate different numbers of segments. PACES performs linking before mapping. So to generate segments, PACES can only check all possible combinations of spin systems. It is possible that some of these generated segments never appear as subsequences of a target protein. RIBRA performs mapping before linking, so all segments generated by RIBRA must be subsequences of a target protein.
4. PACES requires human intervention, while RIBRA is fully automated.

¹BMRB, <http://www.bmrwisc.edu/index.html>

A performance comparison of RIBRA and PACES on this small dataset will be addressed in Section 3.2.2.

We use real experimental data from Academia Sinica, perfect BMRB data and synthetic BMRB data to evaluate RIBRA. Two real datasets are substrate binding domain of BCKD (hbSBD) and lipoic acid bearing domain of BCKD (hbLBD (Chang et al. 2002)). Each of them contains more than 50% false positives and false negatives. Define the *precision* and *recall* of an assignment as follows:

$$\text{precision} = \frac{\text{number of correctly assigned amino acids}}{\text{number of assigned amino acids}} \times 100\% \quad (1)$$

$$\text{recall} = \frac{\text{number of correctly assigned amino acids}}{\text{number of amino acids with known answers}} \times 100\% \quad (2)$$

Compared with the best manual solution, the precision and recall of RIBRA on the first dataset hbSBD are 91.43% and 76.19%, respectively; and those on the second dataset hbLBD are 83.58% and 70.00%, respectively. Such a performance is regarded as quite satisfactory in practice in the sense that the additional human postprocessing effort is quite minimal. We also test RIBRA on 902 perfect datasets in BMRB. The average precision and recall on these datasets are 98.28% and 92.33%, respectively. From these perfect datasets, we generate four kinds of synthetic datasets each with one type of errors from false positives, false negatives, grouping errors and linking errors. The average precision of RIBRA on false-positive, false-negative, grouping-error and linking-error datasets are 98.28%, 95.61%, 98.16% and 96.28%, respectively; the average recall of RIBRA on false positive, false negative, grouping errors and linking errors dataset are 92.35%, 77.36%, 88.57% and 89.15%, respectively.

The remainder of this paper is organized as follows. Section 2 describes the RIBRA algorithm. Section 3 presents experimental results and analysis. Finally, conclusions are given in Section 4.

2 RIBRA

The input of RIBRA is HSQC, CBCANH and CBCA(CO)NH spectral data. We first introduce the basic idea of forming spin systems using these three spectral data. To map spin-system segments onto the target protein sequence, we model it as a graph optimization problem, and provide a solution based on a heuristic maximum independent set algorithm. Finally, we introduce the iterative relaxation technique used by RIBRA to perform backbone assignment.

Our goal is to assign N, H^N, C^α and C^β chemical shifts along the backbone of target proteins. Figure 2a shows two consecutive residues, the $(i - 1)$ -th and the i -th residues, where only atoms along the backbone are depicted; we use R to denote all atoms of a side chain. For the i -th residue, the HSQC experiment detects H _{i} ^N and N _{i} chemical shifts (see Figure 2b). It generates one peak of the form (H _{i} ^N, N _{i} , +), where the first two elements are chemical shifts of H _{i} ^N and N _{i} , respectively. The “+” sign is used to denote that there is a positive peak intensity.

For the i -th residue, the CBCANH experiment detects H _{i} ^N, N _{i} , C _{$i-1$} ^α, C _{$i-1$} ^β, C _{i} ^α and C _{i} ^β chemical shifts (see Figure 2c). It generates four peaks of the form (H _{i} ^N, N _{i} , C _{$?$} , +/-), where the first three elements are chemical shifts of H _{i} ^N, N _{i} and C, respectively. The question mark of the third element is used to indicate that we do not know whether the carbon is located in the $(i - 1)$ -th residue or the i -th residue. The fourth element is the peak intensity in which C^α has a positive value and C^β has a negative one. Two of the

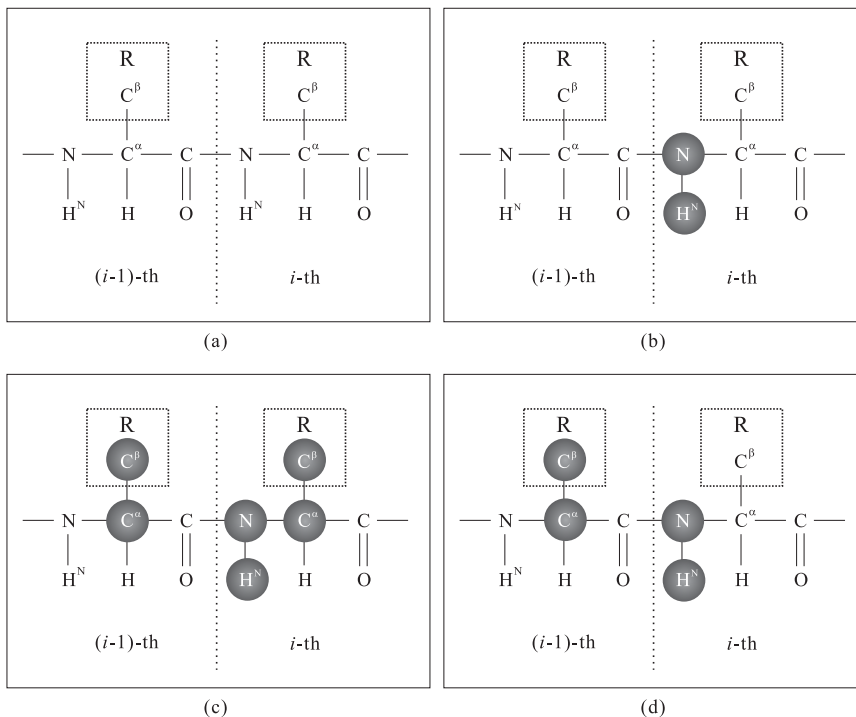


Figure 2: Different NMR experiments on two consecutive residues. The detected atoms with available chemical shifts are marked in black.

four peaks are associated with the α and β carbons of the $(i-1)$ -th residue, and the other two peaks are associated with those of the i -th residue.

For the i -th residue, the CBCA(CO)NH experiment detects H_i^N , N_i , C_{i-1}^α and C_{i-1}^β chemical shifts (see Figure 2d). It generates two peaks of the form $(H_i^N, N_i, C_{i-1}^?, +)$, where the first three elements are chemical shifts of H_i^N , N_i and C, respectively. The question mark of the third element indicates that it is not known whether the carbon is C^α or C^β , since the intensity of this experiment is always positive. Cross-referencing the HSQC, CBCANH and CBCA(CO)NH peaks for the i -th residue, we can generate two consecutive spin systems. That is, we use HSQC to select a spin system of a residue, say the i -th residue, use CBCA(CO)NH to distinguish carbons of the $(i-1)$ -th and the i -th residues, and use the “+” and “-” signs in CBCANH to distinguish α and β carbons. These groups of seven peaks are called a *spin system group* in the rest of the paper.

For example, consider the following seven peaks: a HSQC peaks (7.9, 113.3, +65920032); four CBCANH peaks (7.9, 113.3, 62.5, +79238811), (7.9, 113.3, 27.9, -65920032), (7.9, 113.3, 68.5, -51223894) and (7.9, 113.3, 56.3, +109928374); two CBCA(CO)NH peaks (7.9, 113.3, 56.3, +164325) and (7.9, 113.3, 27.9, +108099). These seven peaks have the same H^N and N chemical shifts, so they are associated with the same residue and their H^N and N have chemical shifts 7.9 and 113.3, respectively. By comparing CBCANH and CBCA(CO)NH peaks, we can infer that 56.3, 27.9 are chemical shifts of carbons of the $(i-1)$ -th residue and 62.5, 68.5 are those of the i -th residue. Moreover, 62.5, 56.3 are chemical shifts of α carbons since their CBCANH peak intensity values are positive; 27.9, 68.5 are chemical shifts of β carbons since their CBCANH peak intensity values are negative. Consequently, the chemical shifts of $(H_i^N, N_i, C_{i-1}^\alpha, C_{i-1}^\beta, C_i^\alpha, C_i^\beta)$ are (7.9, 113.3, 56.3, 27.9, 62.5, 68.5).

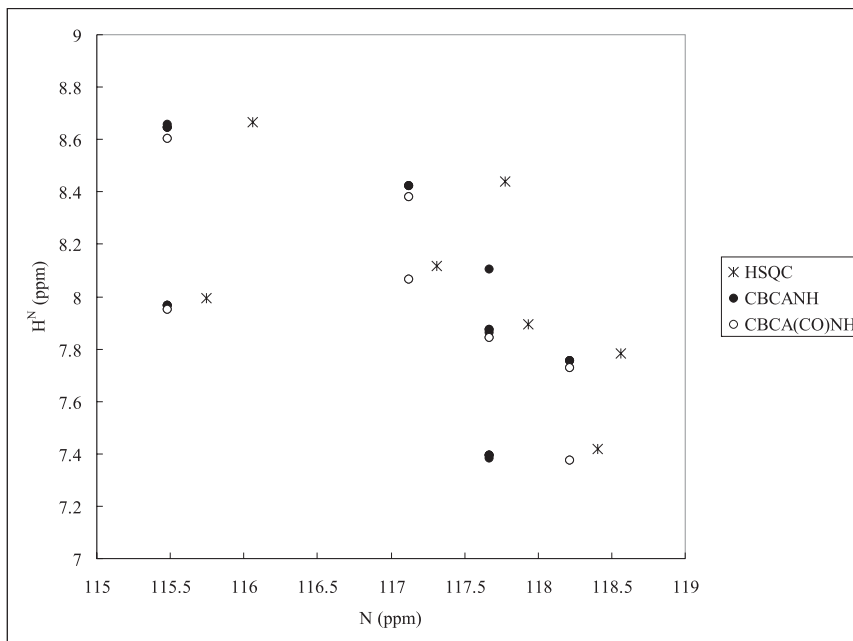


Figure 3: An H^N -N plane. The unit of chemical shifts is ppm (parts per million). Peaks with the same H^N and N chemical shifts are normally close to each other.

2.1 *RGT* Operation: Referencing, Grouping and Typing

In practice, perfect dataset is not available due to experimental errors and biases. So using an exact match algorithm is not feasible for comparing different experimental data. We design an operation *RGT* to group related peaks from different experiments, to generate two consecutive spin systems (or *spin system pair*), and to determine the amino acid types of the two spin systems. These three procedures, referencing, grouping and typing are all mixed together in *RGT*. Initially, we plot all HSQC, CBCANH and CBCA(CO)NH peaks onto an H^N -N plane (see Figure 3). Ideally, for each residue, there should be seven peaks mapped to a single point in the plane since they share the same H^N and N chemical shifts (i.e., these seven points coincide completely): one HSQC peak, two CBCA(CO)NH peaks and four CBCANH peaks. In reality, they usually do not coincide, but are clustered nearby. Since HSQC is more reliable than the other two experiments, we use its peaks as bases to identify different clusters. For each HSQC point in the plane, *RGT* finds the closest four CBCANH points and two CBCA(CO)NH points. If the distances between these points are within a certain threshold, *RGT* regards them as peaks associated with a specific residue, and forms the corresponding spin system pairs. If there are more than six peaks close enough to a HSQC peak, we generate all legal combinations (two CBCA(CO)NH peaks and four CBCANH peaks) to represent possible spin system pairs. Once spin system pairs are generated, we determine their residue types according to the ranges given by TATAPRO II (Atreya et al. 2002); see Table 1. According to Table 1, some typed spin systems are associated with a set of possible residues rather than a unique one. If a generated spin system cannot be typed according to the table, it is deleted. The typed spin systems are basic units to perform linking and mapping. Note that *RGT* only performs referencing, grouping and typing, and it does not handle false negatives or false positives. False positives are handled by the *LM* operation (Section 2.2); false negatives are handled by the iterative relaxation technique in Sections 2.3 and 2.4.

Table 1: Amino acid types based on carbon chemical shift characteristics (Atreya et al. 2002). Since Glycine has only a proton on its side chain, it has no C^β chemical shift. Proline intrinsically has no peaks appearing in NMR spectra, so it has no C^α or C^β chemical shifts. Cys^{red} and Cys^{oxd} represent reduced Cystein and oxidized Cystein, respectively.

Carbon chemical shift	Amino acid
Absence of C^β	Gly
$14 < C^\beta < 24$	Ala
$56 < C^\beta < 67$	Ser
$24 < C^\beta < 36$ and $C^\alpha < 64$	Lys, Arg, Gln, Glu, His, Trp, Cys^{red} , Val and Met
$24 < C^\beta < 36$ and $C^\alpha \geq 64$	Val
$36 < C^\beta < 52$ and $C^\alpha < 64$	Asp, Asn, Phe, Tyr, Cys^{oxd} , Ile and Leu
$36 < C^\beta < 52$ and $C^\alpha \geq 64$	Ile
–	Pro
$C^\beta > 67$	Thr

2.2 LM Operation: Linking and Mapping

Given a set of typed spin system pairs, we try to link them to form longer spin-system segments that can be mapped onto the target protein sequence. A typed spin system pair generated by *RGT* can be regarded as a segment of length 2. Initially, all segments are placed in possible positions with respect to the target sequence according to Table 1. A segment may be placed in more than one position. Two typed spin systems are matched only if their chemical-shift differences of C^α and C^β satisfy predefined thresholds. Any two segments (of length ≥ 2) can be linked to form a longer one if their overlapped typed spin systems are matched. Since segments have already been placed in all possible positions, there is no need to check all segment-pair combinations; it suffices to check consecutive segment pairs. Note that for each typed spin system pair/segment there may be more than one candidate to link to. In the *LM* operation, we generate all possible linked segments to prevent false negatives; this may generate false positives, which will be handled in the mapping mechanism below. The logic of segment extension is given in Figure 4, and an example is given in Figure 5.

Algorithm *Segment_Extension*

Input: A set of typed spin system pairs T from *RGT* and a protein sequence P

Output: A set of segments C

1. Add all spin system pairs in T into C ;
2. Place spin system pairs in possible positions of the target sequence according to the rules of Table 1;
3. **for** i from 3 **to** $\text{length}(P) - 1$
4. Generate all segments of length i by linking two consecutive matching segments of length $i - 1$;
5. **if** new segments of length i are generated
6. Add these new segments into C ;
7. **else**
8. **return** C ;
9. **return** C ;

Figure 4: The segment extension algorithm. In line 4, two consecutive segments of length k match if their overlapped $k - 1$ spin systems are matched. Two typed spin systems are matched only if they satisfy a set of predefined thresholds. The “**return**” command returns the result and halts the program.

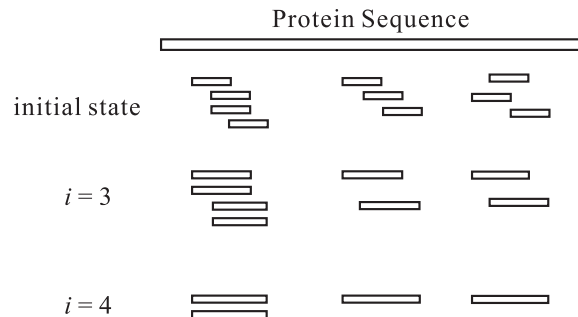


Figure 5: A segment extension example. Initially, there are 10 segments of length 2. After the first iteration ($i = 3$), 8 segments of length 3 are generated, and after the second iteration ($i = 4$), 4 segments of length 4 are generated.

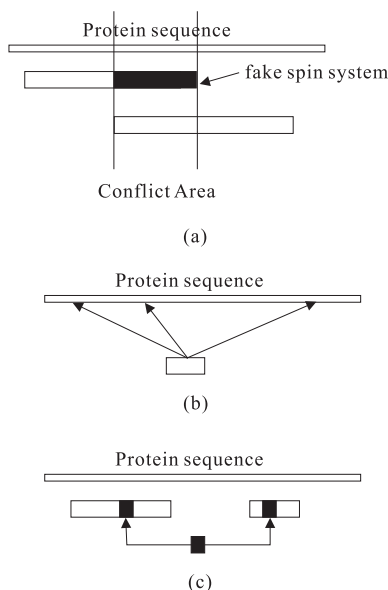


Figure 6: Problems in mapping. (a) Two segments overlap because one of them contains a fake spin system. (b) A segment maps to more than one position in the sequence. (c) Two segments that are far apart could contain an identical spin system.

If linked segments cannot be extended further and their lengths are greater than 3, some of them are then mapped to a target protein sequence. Even though segments can be mapped to the target sequence, there are some potential problems to consider. First, due to false positives of peaks, a segment may contain fake spin systems and should not be mapped to the target sequence; this causes more than one segment to be mapped to overlapped regions of the target protein, which is a contradiction (see Figure 6a). Second, since more than one residue may be associated with a typed spin system, a segment can probably be mapped to more than one position in the target sequence (see Figure 6b). Third, since we generate all possible segments, a spin system may be contained in more than one segment, and those segments cannot be mapped to the target protein sequence simultaneously. In fact, at most one of them can be correctly mapped (see Figure 6c). To resolve these mapping problems, we model them as a graph optimization problem.

Let $G(V, E)$ be an undirected graph, where V is a set of nodes and E is a set of edges. Each node in V represents a mapping from a segment to the target sequence. If a segment can be mapped to $n \geq 2$ positions, there will be n nodes to represent these n mappings.

There is an edge between two nodes if there is a conflict between them. Two nodes are in conflict if 1) they share the same spin system(s), or 2) they are mapped to overlapped regions in the target protein sequence. Our goal is to map segments to the target protein containing as many residues as possible without conflict. This can be formulated as the following weighted maximum independent set problem:

Given an undirected graph $G(V, E)$, find a set $S \subseteq V$ such that

1. For any two distinct nodes $p, q \in S$, $(p, q) \notin E$.
2. The weighted sum of all nodes in S is greater than or equal to that of any other subset of V satisfying (1).

The weight $w(v)$ of a node v is defined as follows:

$$w(v) = \frac{|v| + \sum_{x \in v} \frac{1}{N(x)}}{fre(v)}$$

where $|v|$ is the length of v in terms of mapped segment, x is a spin system of v , $N(x)$ is the number of spin systems having the same H^N and N chemical shifts as x , and $fre(v)$ is the number of positions on the target sequence to which v may map. This weight function satisfies

1. longer segments have higher weights;
2. segments that are more specific to the target sequence have higher weights;
3. segments containing less ambiguous spin systems have higher weights.

Nodes not included in S are regarded as false positives. This graph optimization problem is known to be NP-hard (Garey and Johnson 1979). Although there are algorithms for solving small size problems (such as 200 nodes) in reasonable amount of time, there are good heuristics with much better performance. In the *LM* operation, we adopt a modified heuristic proposed in (Boppana and Halldorsson 1992). The modified heuristic generates several independent sets (rather than a single maximum one) to include more candidates in the future extending and linking. *LM* is operated on the longest n segments to generate S ; in our implementation, n is 100. Figure 7 is an example of our mapping approach.

2.3 Classification of Spin System Group Data

To handle false negatives, we distinguish three types of spin system groups based on their data quality. The first type is *perfect spin system groups*. For a given HSQC peak, if we can find two CBCA(CO)NH peaks and four CBCANH peaks close enough to the peak in the H^N - N plane, we say that these seven peaks form a perfect spin system group, and we can easily generate a corresponding spin system pair.

The second type is *weak false-negative spin system groups*. For a given HSQC peak, if there are only five CBCA(CO)NH peaks and CBCANH peaks that are close enough to this HSQC peak in the H^N - N plane and the missing one can possibly be determined, we say that these six peaks form a weak false-negative spin system group. In such cases, we add a “pseudo peak” to enable cross-referencing and generate a corresponding spin system pair. Note that there may be more than one way to add the pseudo peak. For example, if only one CBCA(CO)NH peak (H^N , N , C , $+$) is available, we compare it with four CBCANH peaks to determine whether it corresponds to a C^α or a C^β . Suppose

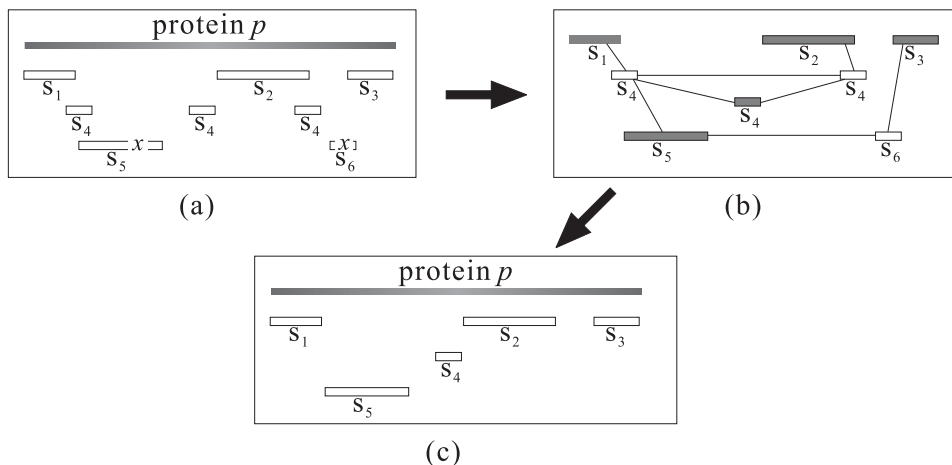


Figure 7: Use a maximum independent set algorithm to map segments. (a) There are six segments $S_1 \dots S_6$ to be mapped to the target protein p . Among them, S_4 can be mapped to three positions of p ; S_4 may overlap S_1 , S_2 and S_5 ; S_3 overlaps S_6 ; both S_5 and S_6 contain the same spin system x . (b) The corresponding undirected graph, in which nodes are represented by horizontal segments. The shaded nodes form the maximum weighted independent set S of the graph. (c) The final mapping.

it corresponds to a C^α . Then we know that the peak containing the C^β information is missing, which can usually be “copied” from CBCANH peaks. However, we only know which two peaks of CBCANH peaks correspond to the C^β , but not exactly which one corresponds to the missing peak. In this case, there are two ways to add pseudo peaks. In our implementation, we choose the one with lower intensity since the missing carbon is usually located in the $(i - 1)$ -th residue rather than the i -th residue. Note also that in some cases, the missing peak cannot be determined and thus the six peaks cannot form a weak false-negative spin system group. For example, if the missing peak is a CBCA(CO)NH peak which has no corresponding CBCA(CO)NH peak, then it would be impossible to determine the missing one.

The third type is *severe false-negative spin system groups*. For a given HSQC peak, if there are only four CBCA(CO)NH peaks and CBCANH peaks close enough to this HSQC peak in the H^N -N plane, and the two missing peaks can possibly be determined, we say that the four peaks form a severe false-negative spin system group. In such cases, we add two “pseudo peaks” to enable cross-referencing and generate a corresponding spin system pair; we may cross-reference CBCANH and CBCA(CO)NH peaks to recognize which peaks are missing and should be added. Similarly, the way to add the pseudo peaks is not unique, and we generate some possible candidates to create spin system pairs. In some cases, the two missing peaks cannot be determined and the five peaks cannot form a severe false negative spin system group.

Figure 8 gives three examples of different kinds of spin system groups.

2.4 Overview of RIBRA

RIBRA uses a relaxation approach to perform backbone assignment in an iterative fashion. The relaxation approach is controlled by the data quality. We first use peaks with top-level quality (perfect spin system groups) to generate a partial assignment of high confidence, and then use peaks with middle-level (weak false-negative spin system group) and low-level quality (severe false-negative spin system groups) to iteratively make more assignment.

<pre> CBCA(CO)NH (113, 8, 56, +1.6e008) (113, 8, 28, +1.1e008) CBCANH (113, 8, 63, +8.5e007) (113, 8, 56, +4.7e007) (113, 8, 68, -8.5e007) (113, 8, 28, -3.5e007) </pre>	→	<pre> N 113 H 8 C^α(i-1) 56 C^β(i-1) 28 C^α(i) 63 C^β(i) 68 </pre>
(a)		
<pre> CBCA(CO)NH (113, 8, 56, +1.6e008) (113, 8, 28, +1.1e008) CBCANH (113, 8, 63, +8.5e007) (113, 8, 56, +4.7e007) (113, 8, 68, -8.5e007) (113, 8, 28, -) </pre>	→	<pre> N 113 H 8 C^α(i-1) 56 C^β(i-1) 28 C^α(i) 63 C^β(i) 68 </pre>
(b)		
<pre> CBCA(CO)NH (113, 8, 56, +) (113, 8, 28, +1.1e008) CBCANH (113, 8, 63, +8.5e007) (113, 8, 56, +4.7e007) (113, 8, 68, -8.5e007) (113, 8, 28, -) </pre>	→	<pre> N 113 H 8 C^α(i-1) 56 C^β(i-1) 28 C^α(i) 63 C^β(i) 68 </pre>
(c)		

Figure 8: Three types of spin system groups. (a) A perfect spin system group that contains all 6 peaks and can determine chemical shifts of atoms easily through cross-referencing. (b) A weak false-negative spin system group that contains 5 peaks and misses a CBCANH peak. The missing peak can be recovered by copying it from CBCA(CO)NH peaks. (c) A severe false-negative spin system group that contains 4 peaks and misses one CBCA(CO)NH peak and one CBCANH peak. Since the CBCA(CO)NH peak does not match two CBCANH C^α peaks, so it must be a C^β peak and is copied to CBCANH peaks. At the same time, we know that CBCA(CO)NH lacks a C^α peak, so we copy the CBCANH C^α peak with smaller intensity to CBCA(CO)NH peaks.

The general steps of RIBRA is as follows:

1. Extract peaks that can form perfect spin system groups.
 - (a) Apply *RGT* to generate typed spin system pairs.
 - (b) Apply *LM* on the generated typed spin system pairs to carry out assignment.
 - (c) Delete used spin systems and related peaks.
2. Extract peaks that can form weak false-negative spin system groups.
 - (a) Add required pseudo peaks.
 - (b) Apply *RGT* to generate typed spin system pairs.
 - (c) Apply *LM* on the newly generated typed spin system pairs and existing segments to modify the assignment generated in Step 1b.
 - (d) Delete used spin systems and related peaks.
3. Extract peaks that can form severe false-negative spin system groups.
 - (a) Add required pseudo peaks.

- (b) Apply *RGT* to generate typed spin system pairs.
- (c) Apply *LM* on the newly generated typed spin system pairs and existing segments to modify the assignment generated in Step 2c.

In Steps 2 and 3, we may add new spin-system segments or extend those contained in the previous assignment. These segments can be used to modify the assignment by applying the weighted maximum independent set algorithm.

Note that the selection of our final assignment totally depends on scores of weighted maximum independent sets. So it is possible that RIBRA outputs more than one final assignment if all the corresponding maximum independent sets have the same top score.

3 Implementation

We implement RIBRA as a web service available at <http://ms.iis.sinica.edu.tw/ribra/> and test it on three kinds of data: BMRB perfect datasets, synthetic datasets based on BMRB data, and real wet-lab datasets. We evaluate RIBRA by calculating its recall and precision on different datasets. Note that Prolines are not taken into account because they have no NMR peaks in HSQC, CBCANH and CBCA(CO)NH experiments.

3.1 Experimental Results on BMRB Perfect Datasets

We downloaded the full BMRB datasets with 3129 proteins on September 10, 2004. Proteins satisfying the following two conditions are chosen for our experiments:

1. $50 \leq \text{protein length} \leq 400$;
2. each protein has at least half of the amino acids whose chemical shifts are available.

In the end, we have collected 902 proteins for testing. The average length of proteins in the collection is 128.12, and the average number of amino acids with available chemical shifts is 110.87. For each protein, we generate the corresponding HSQC, CBCANH and CBCA(CO)NH peaks for testing. The average precision and recall of RIBRA on these datasets are 98.28% and 92.33%, respectively.

Although our BMRB dataset contains proteins with lengths ≤ 400 , we also test RIBRA on a very challenging 723 residue Malate Synthase G, BMRB #5471, to demonstrate its ability to handle long proteins. In addition to the original data, we use synthetic data with false positives, false negatives, grouping errors and linking errors (explained in Section 3.2.1). The precision and recall rates are 98.42% and 95.99%, respectively, for the original data; 98.42% and 95.99%, respectively, for the data with false positives; 97.86% and 84.75%, respectively, for the data with false negatives; 93.32% and 79.66%, respectively, for the data with grouping errors; and 87.54% and 81.20%, respectively, for the data with linking errors.

3.2 Experimental Results on Synthetic Datasets

3.2.1 Synthetic Dataset Construction

We generate synthetic testing datasets to simulate real world noises. Three modifications on the 902 proteins used in Section 3.1 are considered: false negatives, false positives and errors. Since the problem of clustered peaks is an intrinsic property of NMR data, we do not synthesize such datasets.

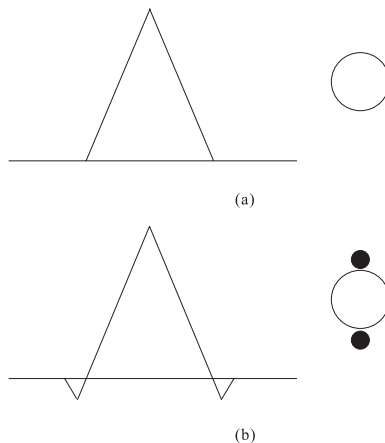


Figure 9: The formation of false positives. (a) An ideal case that one peak corresponds to one spot in a spectrum. (b) A phase distortion causes two small valleys, which forms false positive and produces two extra smaller spots (marked as black) adjacent to a real one.

To generate false-negative datasets, we create missing peaks by randomly deleting peaks. Since we use HSQC peaks as bases, we delete only CBCANH and CBCA(CO)NH peaks. For the i -th residue, the peaks associated with C_{i-1}^α and C_{i-1}^β in CBCANH are more likely missing than those associated with C_i^α and C_i^β due to their longer distances to H_i^N and N_i . And they are usually missing together. Note that we do not delete C_i^α and C_i^β peaks in CBCANH spectra since their peak intensities are usually high enough. Therefore, we delete peaks as follows:

1. Delete both C_{i-1}^α and C_{i-1}^β peaks in CBCANH randomly with probability 0.06.
2. Delete the C_{i-1}^α peak in CBCANH randomly with probability 0.02.
3. Delete the C_{i-1}^β peak in CBCANH randomly with probability 0.02.
4. Delete the C_{i-1}^α peak in CBCA(CO)NH randomly with probability 0.05.
5. Delete the C_{i-1}^β peak in CBCA(CO)NH randomly with probability 0.05.

We select deleted peaks based on a uniform distribution.

To generate false positive datasets, we generate fake peaks to simulate false positives. *Phase distortion* may generate noises near real peaks with opposite sign as illustrated in Figure 9. Such valleys in Figure 9 are distributed near real carbon peaks; their chemical shifts are usually off that of carbon by ± 1.5 . Noisy peaks are usually randomly distributed. Hence, we randomly generate fake peaks with carbon chemical shifts between 10-70. We follow the rules below to generate fake CBCANH peaks:

1. Randomly generate 5% of C_i^α and C_i^β peaks, respectively, to the CBCANH experiment.
2. Randomly generate 5% of carbon peaks with chemical shifts between 10-70 to the CBCA(CO)NH experiment.

We consider data errors arising from different experimental conditions. They are classified into *grouping errors* and *linking errors*:

Table 2: Experimental results of RIBRA on synthetic datasets

# of proteins	902	
# of available residues	100,001	
Testing Cases	Precision	Recall
False positives	98.28%	92.35%
False negatives	95.61%	77.36%
Grouping errors	96.84%	87.44%
Linking errors	94.58%	88.18%

- When grouping related peaks to form a spin system pair, one has to match chemical shifts of the CBCANH and CBCA(CO)NH peaks. The grouping errors complicate this matching since we do not know whether two peaks with similar chemical shifts are actually two peaks or identical. We simulate such errors as follows. When we generate two CBCA(CO)NH peaks for the i -th residue, alter the chemical shifts of H^N , N , C_{i-1}^α and C_{i-1}^β off their originals by ± 0.06 , ± 0.8 , ± 0.2 and ± 0.4 ppm, respectively. Furthermore, assume these chemical shift differences follow normal distributions with mean 0 and different standard deviations of 0.0024, 0.32, 0.08 and 0.16 ppm for H^N , H , α carbon and β carbon, respectively.
- When linking two consecutive segments, we match their overlapped spin systems. The match cannot be exact due to linking errors of C^α and C^β chemical shifts. We simulate such errors as follows. When we generate four CBCANH peaks for the i -th residue, we alter the chemical shifts of C_{i-1}^α and C_{i-1}^β off their originals by ± 0.2 and ± 0.4 ppm, respectively. Furthermore, these chemical shift differences all follow normal distributions with mean 0 and different standard deviations of 0.08 and 0.16 ppm for α carbon and β carbon, respectively.

3.2.2 Experimental Results

The precision and recall of RIBRA on false-positive, false-negative, linking-error and grouping-error datasets are listed in Table 2.

Among all 902 proteins with linking errors, PACES (Coggins and Zhou 2003) has tested 21 of them with a similar parameter setting (which is the only test on synthetic data that PACES performed in their paper). The attribute and experimental results of the 21 proteins are listed in Table 3. PACES can handle only 20 of them and gets 95.24% precision and 86.48% recall on average. RIBRA can handle all 21 proteins and gets 95.05% precision and 87.81% recall on average. Keep in mind that the final results of PACES are further edited by human experts, whereas RIBRA is fully automatic. Moreover, PACES assumes that the spin systems are already given.

3.3 Experimental results on real NMR spectral data

Real NMR spectra normally are mingled with false negatives, false positives, clustered peaks and errors. We test RIBRA on two datasets that correspond to substrate binding domain of BCKD (hbSBD) and lipolic acid bearing domain of BCKD (hbLBD (Chang et al. 2002)), respectively. The detailed properties of hbSBD and hbLBD are listed in Table 4. The precision and recall of RIBRA on the hbSBD dataset are 91.43% and 76.19%, respectively; and those on the hbLBD dataset are 83.58% and 70.00%, respectively.

Table 3: The 21 proteins tested by PACES and RIBRA. PACES cannot handle protein 4402 and produces no output. For the remaining 20 proteins, PACES gets 100% accuracy and 91.35% recall on average.

Protein	# of residues	RIBRA		PACES	
		Precision	Recall	Precision	Recall
4354	330	94.98%	91.82%	100%	94.22%
5316	257	92.86%	85.99%	100%	99.62%
5468	237	94.64%	89.45%	100%	98.33%
4384	211	99.00%	94.31%	100%	93.64%
4022	241	96.98%	93.36%	100%	92.95%
4102	189	90.34%	84.13%	100%	94.81%
4844	197	100.00%	94.92%	100%	98.98%
4836	204	97.46%	94.12%	100%	96.59%
4834	164	97.40%	91.46%	100%	99.39%
4094	127	92.31%	85.04%	100%	100.00%
5142	126	100.00%	99.21%	100%	100.00%
4444	105	94.95%	89.52%	100%	100.00%
4032	115	100.00%	95.65%	100%	100.00%
4152	189	95.93%	87.30%	100%	96.94%
4402	190	74.66%	57.37%	0%	0.00%
4082	132	100.00%	96.97%	100%	99.24%
4722	160	94.74%	90.00%	100%	97.26%
4769	65	100.00%	93.85%	100%	87.88%
4457	166	89.34%	65.66%	100%	17.39%
4341	117	100.00%	92.31%	100%	51.43%
4136	62	88.37%	61.29%	100%	81.94%
Average	170.67	95.05%	87.81%	95.24%	86.48%

3.4 Parameter Settings

To perform RIBAR, we have to specify several parameters that define thresholds in *RGT* and *LM*. Chemical shifts of two atoms are regarded as matched if their difference is lower than the corresponding threshold. The parameter settings of our experiments on all test datasets are listed in Table 5. We choose these settings based on assignment scores; the higher the top assignment score is, the better the parameter setting is. The settings listed in Table 5 are not guaranteed to be the best ones because we only try a certain number of manual settings. We will further develop a method that can automatically tune parameter settings without human intervention.

4 Conclusion

Many backbone assignment approaches have been developed. They usually perform well on perfect data but cannot achieve a good accuracy and recall simultaneously on real wet-lab spectral data containing missing peaks and noises. RIBRA resolves these problems in two ways by: 1) using weighted independent set algorithm to deal with false positives; and 2) using an iterative approach to deal with false negatives in a relaxation fashion. RIBRA uses the best quality NMR peaks to generate basic spin-system segments that serve as building blocks. Then use less confident peaks to extend previous results. The experimental results show that this approach achieves a high degree of precision.

In addition, we have also created four synthesized datasets from BMRB that can be

Table 4: The attributes of hbSBD and hbLBD datasets

Datasets	hbSBD	hbLBD
# of amino acids (A.A.)	53	85
# of A.A. manually assigned by biologists	42	80
# of HSQC peaks	58	78
# of CBCA(CO)NH peaks	258	271
# of CBCANH peaks	224	620
False positives (CBCA(CO)NH)	67.4%	41.0%
False positives (CBCANH)	25.0%	48.4%

Table 5: The parameter settings of all test cases.

Test Cases	<i>RGT</i>				<i>LM</i>	
	N	H ^N	C ^α	C ^β	C ^α	C ^β
Prefect data	0.01	0.01	0.01	0.01	0.01	0.01
False positives	0.01	0.01	0.01	0.01	0.01	0.01
False negatives	0.01	0.01	0.01	0.01	0.01	0.01
Grouping errors	0.64	0.048	0.16	0.32	0.01	0.01
Linking errors	0.01	0.01	0.24	0.48	0.24	0.48
hbSBD	0.884	0.064	0.626	0.626	0.626	0.626
hbLBD	0.936	0.064	0.983	0.983	0.983	0.983

used to perform comprehensive verification on automated backbone assignment algorithms in the future. They contain false negatives, false positives, linking errors and grouping errors, respectively, to simulate real NMR spectral data.

Acknowledgments

This work was supported in part by the thematic program of Academia Sinica under Grant AS91IIS1PP.

References

- Atreya, H.S., Sahu, S.C., Chary, K.V.R. and Govil, G. 2000. A tracked approach for automated NMR assignments in proteins (TATAPRO). *J Biomol Nmr*, 17, 125-136.
- Atreya, H.S., Chary, K.V.R. and Govil, G. 2002. Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II. *Curr Sci India*, 83, 1372-1376.
- Bailey-Kellogg, C., Widge, A., Kelley, J.J., Berardi, M.J., Bushweller, J.H. and Donald, B.R. 2000. The NOESY Jigsaw: Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J Comput Biol*, 7, 537-558.
- Bailey-Kellogg, C., Chainraj, S. and Pandurangan, G. 2004. A Random Graph Approach to NMR Sequential Assignment. *The Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB '04)*, 58-67.
- Bartels, C., Guntert, P., Billeter, M. and Wuthrich, K. 1997. GARANT—A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J Comput Chem*, 18, 139-149.

- Boppana, R. and Halldorsson, M.M. 1992. Approximating Maximum Independent Sets by Excluding Subgraphs. *Bit*, 32, 180-196.
- Buchler, N.E.G., Zuiderweg, E.R.P., Wang, H. and Goldstein, R.A. 1997. Protein heteronuclear NMR assignments using mean-field simulated annealing. *J Magn Reson*, 125, 34-42.
- Chang, C.F., Chou, H.T., Chuang, J.L., Chuang, D.T. and Huang, T.H. 2002. Solution structure and dynamics of the lipoic acid-bearing domain of human mitochondrial branched-chain alpha-keto acid dehydrogenase complex. *J Biol Chem*, 277, 15865-15873.
- Chen, Z.Z., Jiang, T., Lin, G.H., Wen, J.J., Xu, D. and Xu, Y. 2002. Improved approximation algorithms for NMR spectral peak assignment. *Lect Notes Comput Sc*, 2452, 82-96.
- Chen, Z.Z., Jiang, T., Lin, G.H., Wen, J.J., Xu, D., Xu, J.B. and Xu, Y. 2003. Approximation algorithms for NMR spectral peak assignment. *Theor Comput Sci*, 299, 211-229.
- Coggins, B.E. and Zhou, P. 2003. PACES: Protein sequential assignment by computer-assisted exhaustive search. *J Biomol Nmr*, 26, 93-111.
- Garey, M.R. and Johnson, D.S. 1979. *Computer and Intractability: A guide to the Theory of NP-completeness*. New York: W.H. Freeman and Co.
- Guntert, P., Salzmann, M., Braun, D. and Wuthrich, K. 2000. Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. *J Biomol Nmr*, 18, 129-137.
- Hitchens, T.K., Lukin, J.A., Zhan, Y.P., McCallum, S.A. and Rule, G.S. 2003. MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. *J Biomol Nmr*, 25, 1-9.
- Hyberts, S.G. and Wagner, G. 2003. IBIS—A tool for automated sequential assignment of protein spectra from triple resonance experiments. *J Biomol Nmr*, 26, 335-344.
- Langmead, C.J., Yan, A., Lilien, R., Wang, L. and Donald, B.R. 2003. Large a polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. *The Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB '03)*.
- Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. 1998. Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *J Biomol Nmr*, 11, 31-43.
- Li, K.B. and Sanctuary, B.C. 1997. Automated resonance assignment of proteins using heteronuclear 3D NMR. Backbone spin systems extraction and creation of polypeptides. *J Chem Inf Comp Sci*, 37, 359-366.
- Lin, G.H., Xu, D., Chen, Z.Z., Jiang, T., Wen, J.J. and Xu, Y. 2002. An efficient branch-and-bound algorithm for the assignment of protein backbone NMR peaks. *Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB '02)*, 165-174.

- Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. 1997. Automated probabilistic method for assigning backbone resonances of (^{13}C , ^{15}N)-labeled proteins. *J Biomol NMR*, 9, 151-166.
- Malmodin, D., Papavoine, C.H.M. and Billeter, M. 2003. Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. *J Biomol Nmr*, 27, 69-79.
- Moseley, H.N.B. and Montelione, G.T. 1999. Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struc Biol*, 9, 635-642.
- Ou, H.D., Lai, H.C., Serber, Z. and Dotsch, V. 2001. Efficient identification of amino acid types for fast protein backbone assignments. *J Biomol Nmr*, 21, 269-273.
- Slupsky, C.M., Boyko, R.F., Booth, V.K. and Sykes, B.D. 2003. Smartnotebook: A semi-automated approach to protein sequential NMR resonance assignments. *J Biomol Nmr*, 27, 313-321.
- Wang, X., Xu, D., Slupsky, C.M. and Lin, G.H. 2003. Automated Protein NMR Resonance Assignments. *Proceedings of the Second IEEE Computer Society Bioinformatics Conference (CSB '03)*, 197-208.
- Xu, Y., Xu, D., Kim, D., Olman, V., Razumovskaya, J. and Jiang, T. 2002. Automated assignment of backbone NMR peaks using constrained bipartite matching. *Comput Sci Eng*, 4, 50-62.
- Zimmerman, D.E., Kulikowski, C.A., Huang, Y.P., Feng, W.Q., Tashiro, M., Shimotakahara, S., Chien, C.Y., Powers, R. and Montelione, G.T. 1997. Automated analysis of protein NMR assignments using methods from artificial intelligence. *J Mol Biol*, 269, 592-610.