

# Automated Protein NMR Resonance Assignments

Xiang Wan

Protein Engineering Network Centers of Excellence  
Department of Computing Science, University of Alberta  
Edmonton, Alberta T6G 2E8, Canada  
xiangwan@cs.ualberta.ca

Dong Xu

Protein Informatics Group, Life Sciences Division, Oak Ridge National Laboratory  
Oak Ridge, TN 37831-6480, USA  
xud@ornl.gov

Carolyn M. Slupsky

Protein Engineering Network Centers of Excellence  
713 Heritage Medical Research Center, University of Alberta  
Edmonton, Alberta T6G 2S2, Canada

Guohui Lin

Protein Engineering Network Centers of Excellence  
Department of Computing Science, University of Alberta  
Edmonton, Alberta T6G 2E8, Canada  
ghlin@cs.ualberta.ca

## Abstract

*NMR resonance peak assignment is one of the key steps in solving an NMR protein structure. The assignment process links resonance peaks to individual residues of the target protein sequence, providing the prerequisite for establishing intra- and inter-residue spatial relationships between atoms. The assignment process is tedious and time-consuming, which could take many weeks. Though there exist a number of computer programs to assist the assignment process, many NMR labs are still doing the assignments manually to ensure quality. This paper presents (1) a new scoring system for mapping spin systems to residues, (2) an automated adjacency information extraction procedure from NMR spectra, and (3) a very fast assignment algorithm based on our previous proposed greedy filtering method and a maximum matching algorithm to automate the assignment process. The computational tests on 70 instances of (pseudo) experimental NMR data of 14 proteins demonstrate that the new score scheme has much better discerning power with the aid of adjacency information between spin systems simulated across various NMR spectra.*

*Typically, with automated extraction of adjacency information, our method achieves nearly complete assignments for most of the proteins. The experiment shows very promising perspective that the fast automated assignment algorithm together with the new score scheme and automated adjacency extraction may be ready for practical use.*

## 1 Introduction

Proteins act as the most basic working units in life, and determining protein structures is often crucial to understand their functions. Along with X-ray crystallography, Nuclear Magnetic Resonance (NMR) is another key technique used in experimental structure determination. NMR protein structure determination typically involves the following steps:

1. NMR spectral data generation, which produces the following:
  - Resonance peaks, each of which indicates a group of atoms in the target protein that have a

particular magnetic interaction; Resonance peak values are the resonance frequencies, or *chemical shifts*, of the interacting atoms.

- Certain geometric relationships (*e.g.* distances and angles) among resonance peaks.
2. Peak picking, which identifies “real” resonance peaks (peaks generated from protein atoms rather than noise) from NMR spectral maps.
  3. Peak assignment, which groups resonance peaks corresponding to the chemical shifts of atoms from a common amino acid into a *spin system*<sup>1</sup>, and assigns the spin system to a residue of the target protein sequence.
  4. Structural restraint extraction, which extracts intra- and inter-residue distances, dihedral angles, etc., based on the peak assignment.
  5. Structure calculation, which calculates the protein structure, using molecular simulation and energy minimization, under the identified structural restraints.

Among the above five steps, the third one (namely, NMR peak assignment) is very time consuming. This assignment process usually takes weeks or sometimes even months of manual work in order to produce a nearly complete assignment. The automation of the process is a challenging problem in NMR protein structure determination. Manually, the assignment process is typically done in an iterative fashion through establishing relationships between different NMR spectra [21] (we describe the process using HSQC, CBCA(CO)NH, and HNCACB spectra in Section 3, as an example). It involves (a) mapping peaks of chemical shifts from different NMR spectra to spin systems; (b) identifying the possible amino acid types of each spin system; (c) identifying adjacency relationships between spin systems and linking them into contiguous chains, which is done through recognizing overlapping tuples of chemical shifts across different NMR spectra (in our example, those are triples from CBCA(CO)NH and HNCACB); and (d) assigning the chains and the remaining isolated spin systems to the residues of the protein sequence. To summarize, two key pieces of information form the foundation of NMR resonance assignment after the spin systems are identified:

- The likelihood (or score, or confidence) of the mapping between a spin system and an amino acid on the protein sequence.

<sup>1</sup>Mathematically we can think a spin system as a multi-dimensional array of chemical shifts for the atoms, which are ( $H_\alpha$ ,  $C_\alpha$ ) and ( $H^N$ ,  $N$ ,  $C_\alpha$  and  $C_\beta$ ) chemical shifts in our experiments, of a particular residue, with each dimension corresponding to a type of atom. The establishment of a spin system is through combining a set of NMR experiments, as we will exemplify in Section 3 using HSQC, CBCA(CO)NH, HNCACB, each providing a two- or three-dimensional array of chemical shifts based on a peak for the coupling of two or three atoms in an NMR spectrum.

- The sequential adjacency (*i.e.*, connectivity) information of some subsets of spin systems (*i.e.*, each such subset of spin systems should map to a string of consecutive amino acids, or called a *polypeptide*, on the host protein sequence). Each maximal such subset is called a *chain* of spin systems.

## The Assignment Problem

There are a number of significant progress been made in automated methods for peak picking, adjacency information extraction, and resonance assignment [21]. Some most recently developed automated assignment algorithms include [2, 3, 4, 5, 10, 12, 13, 14, 17, 18, 20, 26]. The inputs to these algorithms are several NMR spectra, among which some are required and the others are optional. The assignment algorithms can be roughly classified into a few groups including simulated annealing, generic algorithms, and deterministic search.

Our automated assignment system is based on statistical significance and combinatorial optimization. Rather than requiring certain specific NMR spectra, our system can take in any combinations of NMR spectra, as long as they are sufficient for structure determination. Another difference between our system and the existing ones is that the proposed assignment algorithm is of combinatorial optimization nature and is guaranteed to run in seconds, while for the previous ones the exhaustive search nature may require unpredictably long time.

After all the spin systems, as well as their (partial) adjacency, have been identified, our goal is to automate the assignment of the chains and isolated spin systems to the residues in the protein, taking advantage of the statistical significance of the chemical shift data. The main differences between our automated assignment and the manual work are that efficient computational methods do the assignment at a global view rather than getting trapped into local maximal as in the manual process, and the automation produces an assignment within seconds on a Pentium III PC rather than a few weeks for a manual assignment. The global view helps avoid the tedious “undo – redo” which occurs fairly often throughout the manual work.

We follow the line of research that formulates the resonance assignment problem with connectivity information as a constrained weighted bipartite matching problem on two disjoint groups, one group containing spin systems and the other containing a sequence of amino acids with predicted secondary structures [25]. A weighted bipartite matching problem [16] is to find a one-to-one mapping between elements of two groups that maximizes the total weight, where each matched pair of elements has a pre-specified weight. The NMR resonance assignment can be naturally modeled as a weighted bipartite matching problem, where each

weighted edge represents a possible mapping of a spin system to an amino acid on the protein sequence with a quantitative confidence value. To incorporate the connectivity information, we can extend the above model to a *constrained* weighted bipartite matching problem, *i.e.* we define a (partial) neighboring relationship between the elements of each group and require that neighbors of a group be matched only with neighbors of the other group. Unfortunately, the constrained bipartite matching problem is NP-hard, even if the edges are unweighted [25]. Some heuristics have been proposed very recently, including a slow, exhaustive two-layer algorithm [25] and some fast approximation algorithms [8]. These heuristics and approximations attempt to find feasible matchings with (approximately) the largest weights, and may work well for the NMR peak assignment based on our observation that, in practice, a chain of connected spin systems typically have a better score at the “correct” assignment position (*i.e.* the matching between a spin system of NMR peaks and the residue that generates the peaks) than almost all other (incorrect) assignment positions, especially as the size of the chain increases. Later on, to overcome the drawbacks in the two-layer algorithm, another heuristics was developed for finding maximum-weight constrained bipartite matchings based on the branch-and-bound technique and a “greedy” filtering process [19]. The branch-and-bound algorithm uses an efficient (unconstrained) bipartite matching algorithm and the approximation algorithms in [8] to compute necessary lower and upper bounds on optimal solutions to help prune the search tree, and returns an optimal solution, *i.e.* a feasible matching with the largest weight. It runs much faster than the exhaustive search used in the two-layer algorithm. The comparison results for these heuristics and approximations can be found in [7, 19, 25].

In this paper, we develop a very fast heuristic two-phase assignment procedure, based upon our previously developed assignment algorithms: in the first phase, a greedy filtering is conducted to select some number of best possible mappings for spin systems residing in the identified chains; in the second phase, for every combination of chain mappings, an efficient maximum weight bipartite matching algorithm is used to complete the assignment by mapping isolated spin systems to the rest of residues. It outputs the best assignment from all combinations in terms of the assignment confidence. We choose to use this simple heuristic to evaluate our new scoring schemes and the automated extraction of the adjacency information.

A score scheme is used to weight the mapping of a spin system to an amino acid. In the ideal case, the score scheme is so powerful that it could make the correct assignment standing out against other possible assignments. However, because the chemical shifts generated from the same type of atoms in different types of amino acids may be close to each other, effective learning process is neces-

sary to score the preferences. In the previous score scheme used in [7, 19, 25], a statistics was done on a set of 220 proteins. The range of chemical shifts for every combination of amino acid and secondary structure is simply partitioned into 5 bins of equal size, and the frequency of chemical shift values appearing in a bin is taken to be the frequency of a chemical shift generated from that combination. The log odd of this frequency, multiplied by some constant and rounded into an integer, is taken to be the score of the mapping.

In Section 2, we describe an improved score scheme which is based on statistical learning on a larger set of proteins and a better (confirmed by the computational results shown in Section 4) way to estimate the frequency for a specific chemical shift value. In Section 3, we describe briefly, as an example, on how adjacency information among the spin systems can be extracted out of HSQC, CBCA(CO)NH, and HNCACB spectral maps. Comparison between the previous score scheme and the new score scheme, and the experiments using simulated adjacency information are presented in Section 4. Section 5 concludes the paper with some remarks and future work directions.

## 2 Score Scheme Learning

We have designed an improved statistics-based scoring scheme for assessing the likelihood of an array of chemical shifts that an amino acid in some type of second structure can generate. Our scoring scheme can take any combination of chemical shifts (the common ones are  $H^N$ , N,  $H_\alpha$ ,  $C_\alpha$ ,  $C_\beta$  and C). In this study, we derived the scoring scheme using two combinations: (1)  $H_\alpha$  and  $C_\alpha$ ; and (2)  $H^N$ , N,  $C_\alpha$ , and  $C_\beta$ . The derivation was done using the NMR spectra of proteins collected in BioMagResBank [22]. The derivation process has 4 steps. We use the second combination of chemical shifts of the score scheme as an example.

- (a) **Data Filtering:** The chemical shifts stored in the BioMagResBank database cannot be used directly to collect the statistics as we want, for two reasons. The first reason is that among the protein sequences collected in the BioMagResBank database, a lot of them are homologous. “b12seq” [1] is run on every pair of 863 sequences collected in the BioMagResBank and suitable clustering is done where each cluster contains sequences sharing at least 50% *modified* sequence identity (which is taken as the maximum number of matched amino acids over all local alignments returned by b12seq divided by the length of the shorter sequence). This homology filtering gives us at the end 463 clusters and we randomly pick one from each cluster for our score scheme training. Secondly, for a single protein, the chemical shifts collected in the

BioMagResBank for a few amino acids might be outliers. Since the abnormal behavior of a single outlier may disrupt our scoring scheme, an efficient statistical method, namely “boxplot” [9], is applied to remove the outliers. Such a treatment was not done in the previous scoring derivation [25]. After the treatment, the accepted range of  $C_\alpha$ 's chemical shifts in our training data set is from 39.93 to 69.80, substantially narrowed from the observed range of  $C_\alpha$ 's chemical shifts across all types of amino acids, *i.e.*, from 4.10 to 85.492. Figure 1(a) shows the distribution of chemical shifts of  $C_\alpha$  for all 20 types of amino acids. Figure 1(b) shows the distribution of all chemical shifts of  $C_\alpha$  for alanines only.

(b) **Chemical Shifts Based on Secondary Structure Prediction:** The chemical shift of a specific atom in a particular type of amino acid is not necessarily a constant. In fact it is affected by the local electronic-biochemical environment in which the amino acid lies. One important observation is that it depends on the type of secondary structures where the amino acid lies. For example, for alanines, the distributions of  $C_\alpha$  chemical shifts in  $\alpha$ -helices,  $\beta$ -sheets, and loop regions are significantly different, as shown in Figures 2(b), 2(c), and 2(d), respectively. Our score scheme accounts for the impact of structural information on chemical shifts, through incorporating predicted secondary structures by the PsiPred program [15], which is one of the best known secondary structure prediction programs with approximately 80% accuracy for assigning a residue to an  $\alpha$ -helix, a  $\beta$ -strand, or a loop.

(c) **Score Generation:** We also used a more sophisticated statistical method than the one in our previous to handle the scoring scheme derivation [25]. For every combination of amino acid  $aa$  and secondary structure  $ss$ , let  $\Pi(aa, ss)$  denote the chemical shift distribution we get out of the database given  $aa$  and  $ss$ . For example, Figure 2(b) shows the distribution  $\Pi(\text{Ala}, \text{helix})$ . We do not assume there is any specific pattern that the distribution should follow (usually a normal distribution is assumed in the literature, which can be seen not necessarily true from our training set), but use the chemical shift values directly. Let  $N(aa, ss)$  denote the total number of chemical shift values collected in the database given  $aa$  and  $ss$ . For every type of chemical shift, we associate it with an error bound  $\epsilon$  (which is different for different types of chemical shifts. As shown in Table 1, the third column contains all  $\epsilon_\alpha$  values we pick for  $C_\alpha$  chemical shifts in this score scheme). The error bound  $\epsilon$  is learned out of the training set such that 20 intervals of length  $\epsilon$  cover the range of the chemical shifts. It also maps well to the

reading error indicates in BioMagResBank. For every examined quadruple spin system  $(H^N, N, C_\alpha, C_\beta)$ , let  $(C_\alpha - \epsilon_\alpha, C_\alpha + \epsilon_\alpha)$  be the  $C_\alpha$  chemical shift *window*. Let  $N(aa, ss, C_\alpha)$  denote the number of  $C_\alpha$  chemical shift values collected in the database which fall into the window. The *probability* that this examined  $C_\alpha$  chemical shift is generated from amino acid  $aa$  in the secondary structure  $ss$  is

$$\text{Prob}(aa, ss, C_\alpha) = \frac{N(aa, ss, C_\alpha)}{N(aa, ss)}.$$

The 4th column in Table 1 shows the probabilities for a  $C_\alpha$  chemical shift value 56.74 (which is the mean value of  $C_\alpha$  chemical shifts) generated from the 60 combinations. The score for mapping the quadruple  $(H^N, N, C_\alpha, C_\beta)$  to combination  $(aa, ss)$  is

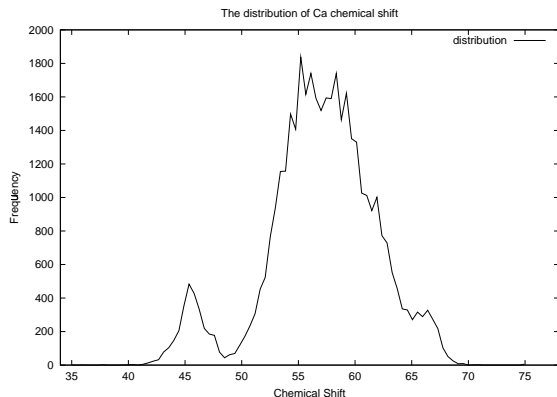
$$\begin{aligned} & \text{score}((H^N, N, C_\alpha, C_\beta) | (aa, ss)) \\ &= 10 \times \log \left( \text{Prob}(aa, ss, N) \times \text{Prob}(aa, ss, C_\alpha) \right. \\ & \quad \left. \times \text{Prob}(aa, ss, C_\beta) \right) \end{aligned}$$

(this number is truncated into an integer).

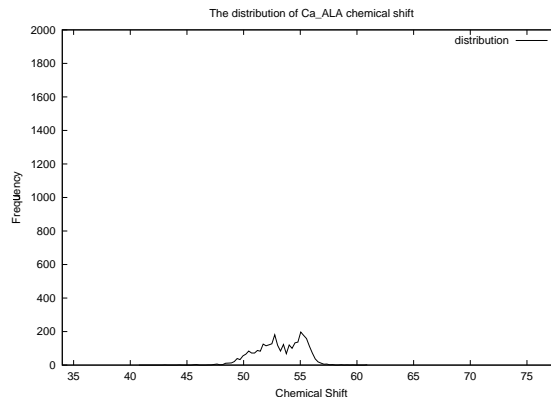
(d) **Score Scheme Enhancement:** One may notice that the distribution of  $H^N$  chemical shifts is not used in our score scheme. The reason is that across 60  $aa$  and  $ss$  combinations, no significant difference is found, and thus adding it into account does not give much useful information. On the positive aspects, there are quite a few special features of the chemical shifts which can be taken advantage to estimate the scores. We point out two in the following: (1) Since there is no  $C_\beta$  atom in a Glycine (GLY, G), no  $C_\beta$  chemical shift can be examined. If a quadruple does contain a  $C_\beta$  chemical shift, then it should not be generated from a GLY. Therefore, we can associate with the mapping a score minimum, which tells the assignment algorithm that such a mapping is *illegal*. (2) Similarly, since a Proline (PRO, P) doesn't have  $H^N$  atom, a quadruple containing a  $H^N$  chemical shift gets a score minimum when mapping to a PRO.

### 3 Adjacency Information Determination

We briefly describe how adjacency information is extracted/predicted out of the NMR spectra. As an example, we use three spectra: HSQC, CBCA(CO)NH, and HNCACB. Different NMR labs might be able to generate different combinations of other spectra, but the adjacency information is extracted in a similar fashion. This particular combination of NMR spectra is proposed in [23] and our description is based on the semi-automated approach referred



(a) for all 20 types of amino acids.



(b) for alanines only.

**Figure 1. Distribution of  $C_{\alpha}$  chemical shifts in training data set.**

to as “SMARTNOTEBOOK” [23]. The CBCA(CO)NH spectrum gives us triples of inter-residue chemical shifts  $(H_i^N, C_{\alpha i-1}, N_i)$  and  $(H_i^N, C_{\beta i-1}, N_i)$ , where  $i$  indexes the residue (but a GLY doesn’t have a  $C_{\beta}$  chemical shift). The HNCACB spectrum gives us triples of inter- and intra-residue chemical shifts  $(H_i^N, C_{\alpha i-1}, N_i)$ ,  $(H_i^N, C_{\beta i-1}, N_i)$ ,  $(H_i^N, C_{\alpha i}, N_i)$ , and  $(H_i^N, C_{\beta i}, N_i)$  (again a GLY doesn’t have a  $C_{\beta}$  chemical shift). The HSQC spectrum gives us pairs of intra-residue chemical shifts  $(H_i^N, N_i)$ . Therefore, from these 3 spectra, we can associate with residue  $i$  a quadruple chemical shifts  $(H_i^N, N_i, C_{\alpha i}, C_{\beta i})$ . Our assignment goal is to identify for each residue its true quadruple. In the ideal case, the quadruples can be read out of the 3 NMR spectra, and the number of quadruples is equal to the number of residues in the protein sequence. However, in general, the chemical shifts measured out of one NMR spectrum are different from those measured out of another NMR spectrum. Nonetheless, the difference is very small and we hope we can still be able to extract the quadruples, besides the existence of noise peaks and missing peaks.

In the following, we describe briefly on identifying the quadruples, together with the adjacency information between quadruples.

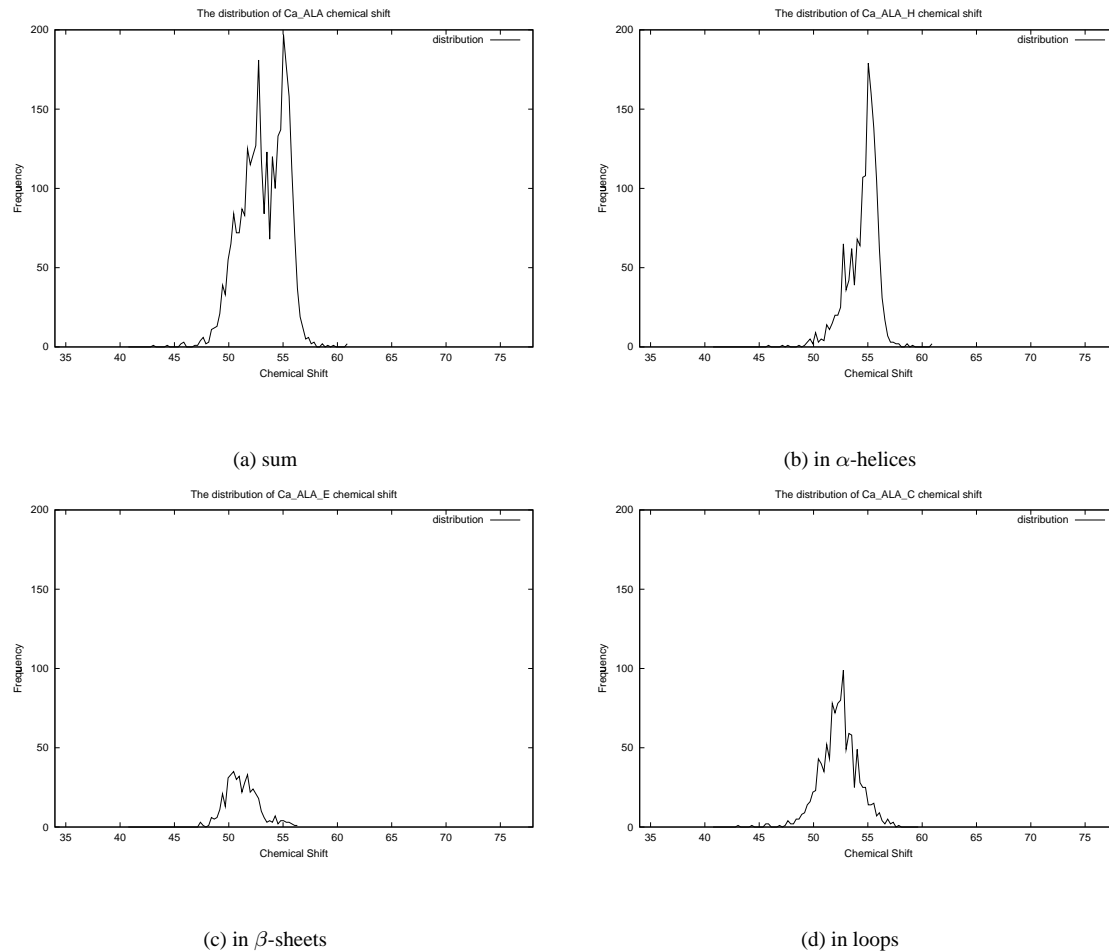
1. For a pair of chemical shifts (assuming it is associated with the  $i$ th residue) in HSQC spectrum, say  $(H_i^N, N_i)$ , search for 2 triples in the CBCA(CO)NH spectrum that share the  $H^N$  and  $N$  chemical shifts, keeping in mind that maybe only one can be found if the  $(i-1)$ th residue is a GLY. (In general, when there are more than 2 triples, we have to identify which one or two of them are true triples; when there is no triple, we might consider this pair  $(H_i^N, N_i)$  as noise.) Similarly we can

find 4 (or 2) triples from the HNCACB spectrum.

2. Suppose 2 triples in the CBCA(CO)NH spectrum are found. Then the C chemical shifts are for  $C_{\alpha}$  and  $C_{\beta}$  atoms from the  $(i-1)$ th residue. We can denote the two triples as  $(H_i^N, C_{\alpha i-1}, N_i)$  and  $(H_i^N, C_{\beta i-1}, N_i)$ .
3. Suppose 4 triples in the HNCACB spectrum are found. Then the C chemical shifts are for  $C_{\alpha}$  and  $C_{\beta}$  atoms from the  $(i-1)$ th residue, and for  $C_{\alpha}$  and  $C_{\beta}$  atoms from the  $i$ th residue itself. We can denote the 4 triples as  $(H_i^N, C_{\alpha i-1}, N_i)$ ,  $(H_i^N, C_{\beta i-1}, N_i)$ ,  $(H_i^N, C_{\alpha i}, N_i)$ , and  $(H_i^N, C_{\beta i}, N_i)$ .
4. Since triples  $(H_i^N, C_{\alpha i}, N_i)$  and  $(H_i^N, C_{\beta i}, N_i)$  appear in the HNCACB spectrum only, we identify the quadruple for  $i$ th residue at this moment to be  $(H_i^N, N_i, C_{\alpha i}, C_{\beta i})$ . To identify the quadruple for  $(i+1)$ th residue, we try to look for some pair of chemical shifts  $(H_{i'}^N, N_{i'})$  from HSQC spectrum such that:
  - $(H_{i'}^N, C_{\alpha i}, N_{i'})$  and  $(H_{i'}^N, C_{\beta i}, N_{i'})$  appear in the other two spectra,

which signify the sequential adjacency relationship between two quadruple spin systems,  $(H_i^N, N_i, C_{\alpha i}, C_{\beta i})$  and  $(H_{i'}^N, N_{i'}, C_{\alpha i'}, C_{\beta i'})$ . Subsequently, we can assign  $i' = i + 1$ .

This searching process is performed for every pair of chemical shifts in the HSQC spectrum to identify the quadruple and its adjacent quadruple. In the case that the chemical shift error is too large to be believed, the identifying process terminates and is re-started on another pair. The results are



**Figure 2. Distribution of  $C_{\alpha}$  chemical shifts from alanines in training data set.**

a set of quadruples, which are the spin systems, and the adjacency information among the quadruples. This adjacency information connects quadruples into chains, which will be mapped to non-overlapping polypeptides in the target protein sequence.

## 4 Assignment Algorithms

The assignment algorithm we implement in this paper is fairly intuitive, and very close to what is currently manually doing in an NMR laboratory. The main difference between the algorithm and manual work is that we employ efficient computational methods to automate the assignment process at a global view, which produce an assignment within seconds on a Pentium III PC. The global view also helps avoid the tedious “undo – redo” which occurs very often throughout the manual work. The assignment algorithm can be simply described as a two-phase procedure: in the first phase,

a greedy filtering is conducted to select out some number of best possible mappings for spin systems residing in the identified chains; in the second phase, for every combination of chain mappings, an efficient maximum weight bipartite matching algorithm is used to complete the assignment by mapping isolated spin systems to the rest of residues. It reports the best assignment from all combinations in terms of the assignment confidence.

### Greedy Filtering

The greedy filtering process is employed to evaluate the score scheme and take advantage of the discerning power of long chains. First, it sorts the chains into non-increasing length order, where the length of a chain is measured by the number of spin systems therein. Let  $k$  be a parameter to bound the number of the best combinations we want to put into the second phase. The filtering process starts with

		$\epsilon_\alpha$	Prob( $aa, ss, 56.74$ )
Ala	helix	0.680	0.191
	sheet	0.776	0.011
	loop	0.835	0.033
Arg	helix	1.004	0.096
	sheet	0.691	0.038
	loop	0.960	0.180
Asn	helix	0.873	0.168
	sheet	0.576	0.006
	loop	0.840	0.012
Asp	helix	0.816	0.219
	sheet	0.756	0.027
	loop	0.852	0.050
Cys	helix	2.156	0.102
	sheet	0.964	0.184
	loop	1.700	0.198
Gln	helix	0.816	0.093
	sheet	0.688	0.051
	loop	0.871	0.133
Glu	helix	0.684	0.066
	sheet	0.652	0.055
	loop	0.964	0.247
Gly	helix	0.656	0.000
	sheet	0.680	0.000
	loop	0.556	0.000
His	helix	1.080	0.112
	sheet	0.770	0.067
	loop	0.952	0.153
Ile	helix	1.275	0.005
	sheet	0.792	0.033
	loop	1.080	0.031

		$\epsilon_\alpha$	Prob( $aa, ss, 56.74$ )
Leu	helix	0.756	0.114
	sheet	0.664	0.027
	loop	0.888	0.083
Lys	helix	0.872	0.092
	sheet	0.640	0.062
	loop	0.940	0.210
Met	helix	1.160	0.206
	sheet	0.640	0.024
	loop	0.868	0.134
Phe	helix	1.320	0.060
	sheet	0.720	0.229
	loop	1.016	0.196
Pro	helix	0.760	0.000
	sheet	0.554	0.000
	loop	0.616	0.001
Ser	helix	1.064	0.048
	sheet	0.609	0.206
	loop	0.904	0.159
Thr	helix	1.648	0.007
	sheet	0.860	0.006
	loop	0.960	0.015
Trp	helix	1.320	0.098
	sheet	0.792	0.268
	loop	1.230	0.208
Tyr	helix	1.280	0.064
	sheet	0.760	0.225
	loop	1.112	0.183
Val	helix	1.040	0.002
	sheet	0.912	0.013
	loop	1.200	0.010

**Table 1. Error bounds for  $C_\alpha$  chemical shifts and a sample probability. Prob( $aa, ss, 56.74$ ) is the estimated probability of a chemical shift value 56.74 which is generated from atom  $C_\alpha$  in the amino acid  $aa$  residing in secondary structure  $ss$ .**

finding the first  $k$  best positions that the longest chain can map to. This gives the top  $k$  combinations, which involve only the longest chain at the moment. For every one of the  $k$  combinations, the process proceeds to find the first  $k$  best positions to which the second longest chain can map. In general this will give in total  $k^2$  combinations involving one more chain, and the process only keeps the top  $k$  ones and proceeds to find, for every one of the combinations, the  $k$  best positions to which the third longest chain can map; and so on.

The process is repeated for all chains of length at least  $L$ , which is a lower bound on the length of chains involved in combinations. The final output is a set of (at most)  $k$  combinations of positions to which the chains map. It worths pointing out that at some circumstances, the filtering process will terminate at some intermediate combinations, since it fails to find any position for the next chain, by the assumption that every identified chain should map to a polypeptide in the target sequence and no residue on this polypeptide can be mapped multiple times.

The parameters  $k$  and  $L$  can be tuned suitably to continue the assignment, depending on which algorithms employed in the second phase. In our assignment algorithm,

the second phase algorithm is a maximum weight bipartite matching algorithm. Therefore,  $L$  is set to 2, meaning that all identified chains should be mapped to non-overlapping polypeptides from the target protein sequence. (In the branch-and-bound heuristics,  $L$  is set to 3, meaning that those chains of length greater than 2 should be mapped, and length-2 chains and isolated spin systems is left to be handled by branch-and-bound. Of course, one could set  $L$  to be a larger number, whereby achieving better assignment confidence through sacrificing running time. The experimental study [19] shows that setting  $L$  to 3 is already a good trade-off.)  $k$  is set to 10 for all our experiments.

### Maximum Weight Bipartite Matching

An instance of the maximum weight bipartite matching problem consists of an edge-weighted bipartite graph  $G = (S, R, E)$ , where we assume without loss of generality that edge weights are non-negatives. Intuitively, the vertex set  $S$  contains all the isolated spin systems left after the greedy filtering ( $L = 2$ ); the vertex set  $R$  contains all the remaining residues in the target protein sequence. An edge indicates the mapping between a spin system and a

residue, where its weight records the confidence of the mapping. The goal of the problem is to compute a matching with maximum weight, corresponding to a partial assignment achieving the maximum confidence for those isolated spin systems. Since we require that all identified chains are mapped to non-overlapping polypeptides in any combination,  $|S| = |R|$  and the expected matching is perfect (meaning that every spin system is mapped to some residue).

There are various implementations based on efficient algorithms for the maximum weight bipartite matching problem. In our study, we choose the one by Goldberg and Kennedy [11].

## 5 Simulation Results

The score scheme we learned using  $H_\alpha$  and  $C_\alpha$  chemical shifts can be regarded as an improved version of our previous one [25]. The heuristics “greedy filtering + GoldbergKennedy [11] maximum matching” (GF-GK) algorithm is run on 14 protein sequences [7, 19, 25], using this new score scheme, similarly with various levels of adjacency information ranging from 50% to 90%. The result is shown in Table 2, where the comparison is made between the performance of GF-GK using this new score scheme and the performance of a few algorithms using the previous score scheme. These a few algorithms include GF-GK, “greedy filtering + Branch-and-Bound” (GF-BnB), and “greedy filtering +  $\frac{13}{7}$ -approximation” (GF- $\frac{13}{7}$ ) [6]. Define the mapping recoverage of an “algorithm and score scheme” combination to be the average percentage of correct mappings recovered by running the assignment algorithm using the specific score scheme. The mapping recoverage of these 4 “algorithm and score scheme” combinations are shown in Figure 3(a), where it can be seen that using the new score scheme enhances the recoverage significantly.

In another set of experiments, we test the score scheme learned using  $H^N$ ,  $N$ ,  $C_\alpha$ , and  $C_\beta$  chemical shifts. Two fast heuristics “GoldbergKennedy maximum matching” (GK) and GF-GK are run on the same set of instances with the same simulated adjacency information. Table 3 shows the results. The “score” column records the assignment score of the correct assignment; The “GK score” column is the optimal matching by ignoring all the adjacency information. It can be seen that these two scores are very close to each other, indicating that the score scheme is very promising.  $R^{GF-GK}$  ( $R^{GF}$ ) stores the number of correct mappings recovered by GF-GK (GF, respectively), where the number in the parentheses in the 8th column stores the number of spin systems in the identified chains. The 9th to 11th columns are the percentage of adjacency extracted from the simulated HSQC, HNCACB, and CBCA(CO)NH spectra, the score of the assignment output by GF-GK, and the number of correct mapping recovered, respectively. Figure 3(b) shows the

curves of the percentages of the number of correct mappings recovered by GK, GF, and GF-GK. It should be noted that the heuristics finishes an instance in at most a few seconds on a 1GHz Pentium III with 1GB memory. Under the high-throughput requirement, the fast GF-GK is very competitive. Another exciting notice is that with about 70% adjacency information identified, about 84% correct mappings are recovered; and with about 80% adjacency information identified, about 97.5% correct mappings are recovered.

For real NMR data, one experiment is conducted. For a cell cycle regulatory protein Cdc4p, a 141 residue protein studied in [24, 23] (BioMagResBank accession number bmr4851) with the identified 14 chains (out of HSQC, HNCACB and CBCA(CO)NH spectra), the GF-GK outputs an assignment in which the spin systems residing on the identified chains are 100% correctly mapped to their host residues. The same thing happens for the simulated adjacency data, where all spin systems in the identified chains are correctly mapped to their host residues. These results show that the new score schemes are much more effective than the previous one, in that with the aid of automated adjacency information extraction it gives more powerful discerning ability. This suite of computational techniques may be ready to assist spectroscopists to automate the NMR resonance assignment.

## 6 Discussion

In summary, we have developed a better score scheme learning system for the NMR resonance peak assignment problem. The advantages of this new system are demonstrated by running some simple and fast heuristics on simulated and real NMR data. From Table 3, we see that the assignment accuracy depends significantly on the adjacency information between spin systems. Good NMR instruments, particularly 800 or higher MHz NMR machines, can easily achieve 70% adjacency information (for example, 68% adjacency information is recovered for Cdc4p [23] and on average 71% is recovered using the simulated data). The ability to produce nearly complete assignments at such level of adjacency abundance implies that our new score scheme, automated adjacency extraction, and the fast assignment algorithms will be of potential employment in a lot of NMR labs. Nonetheless, from the experiment using real NMR spectra, we noticed that there is still some ambiguous places in the assignment process, which is mostly due to spectral data quality. What we can conclude from our work is that one feasible way is to use our automated assignment system to generate a set of a few very likely assignments and then plug in the structure determination pipeline to produce a set of a few candidate 3D structures for next phase investigation.

In this study, we only run some simple and fast heuris-

	length	$R_{old}^{GF-GK}$	$R_{old}^{GF-BaB}$	$R_{old}^{\frac{13}{7}}$	$R_{new}^{GF-GK}$
bmr4027.5	158	38	33	40	49
bmr4027.6		59	37	64	65
bmr4027.7		88	74	89	132
bmr4027.8		149	128	151	158
bmr4027.9		156	156	156	156
bmr4144.5	78	15	16	11	27
bmr4144.6		19	11	21	43
bmr4144.7		55	64	68	54
bmr4144.8		61	67	69	65
bmr4144.9		75	75	75	78
bmr4288.5	105	33	12	36	51
bmr4288.6		42	26	49	46
bmr4288.7		61	57	65	78
bmr4288.8		66	66	68	105
bmr4288.9		105	105	105	105
bmr4302.5	115	31	16	31	48
bmr4302.6		58	43	51	69
bmr4302.7		81	62	79	77
bmr4302.8		112	103	112	112
bmr4302.9		110	110	111	115
bmr4309.5	178	28	25	35	45
bmr4309.6		43	57	48	68
bmr4309.7		108	77	119	105
bmr4309.8		116	101	121	157
bmr4309.9		178	178	178	178
bmr4316.5	89	48	30	43	64
bmr4316.6		65	35	59	83
bmr4316.7		79	79	75	85
bmr4316.8		89	89	75	89
bmr4316.9		89	89	89	89
bmr4318.5	215	23	20	19	39
bmr4318.6		35	35	34	59
bmr4318.7		72	52	73	94
bmr4318.8		91	62	92	134
bmr4318.9		201	201	201	211

	length	$R_{old}^{GF-GK}$	$R_{old}^{GF-BaB}$	$R_{old}^{\frac{13}{7}}$	$R_{new}^{GF-GK}$
bmr4353.5	126	25	17	20	40
bmr4353.6		17	24	23	51
bmr4353.7		55	44	56	98
bmr4353.8		80	80	78	116
bmr4353.9		126	126	124	126
bmr4391.5	66	9	18	10	24
bmr4391.6		14	10	7	34
bmr4391.7		17	26	17	28
bmr4391.8		42	42	38	47
bmr4391.9		66	66	66	66
bmr4393.5	156	45	41	49	48
bmr4393.6		70	59	71	76
bmr4393.7		96	76	102	97
bmr4393.8		135	130	129	137
bmr4393.9		152	152	142	152
bmr4579.5	86	21	15	18	26
bmr4579.6		41	32	35	36
bmr4579.7		44	44	48	67
bmr4579.8		79	63	72	75
bmr4579.9		86	86	86	86
bmr4670.5	120	30	22	32	47
bmr4670.6		35	30	35	59
bmr4670.7		80	38	78	100
bmr4670.8		116	116	116	120
bmr4670.9		116	116	120	120
bmr4752.5	68	27	21	21	39
bmr4752.6		34	32	32	45
bmr4752.7		41	41	43	44
bmr4752.8		68	68	68	68
bmr4752.9		68	68	68	68
bmr4929.5	114	24	23	17	45
bmr4929.6		38	32	36	61
bmr4929.7		70	56	69	70
bmr4929.8		88	88	82	99
bmr4929.9		114	114	114	114

**Table 2. The numbers of corrected mappings recovered by various heuristics under the old and new score schemes. In this table, bmr#### is the protein entry in the BioMagResBank [22]. The digit following the underscore indicates the percentage of adjacency information assumed. For example, \_5 indicates that 50% adjacency is used. The “length” column records the length of the protein sequence, measured by the number of amino acids therein.  $R^{alg}$  denotes the number of corrected mappings recovered by algorithm “alg”, where “old” indicates using the previous score scheme and “new” indicates using the new score scheme.**

tics to do the assignment. Since the branch-and-bound will produce the best assignments consuming potentially exponential time, its results on the same set of instances are not included. Testing the branch-and-bound algorithms, as well as designing better approximations to provide better lower/upper bounds, would be our next task.

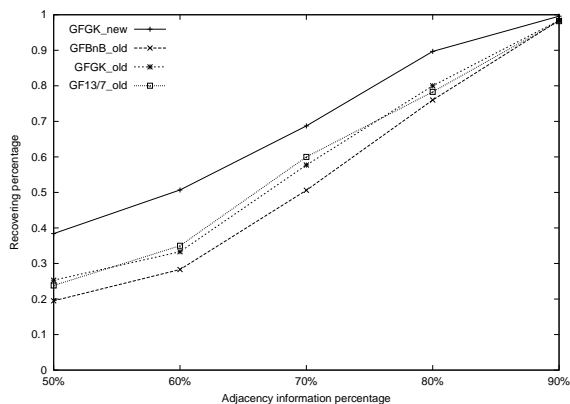
Another to-do task is to conduct a fair comparison between the performance of our system and the performance of the existing assignment procedures [21]. However, since most of the existing procedures use many more NMR spectra and/or additional spectral data other than chemical shifts, we need to design a fair comparison model first. Last but not the least, an evaluation model which can assess the quality of the output assignments should be built to provide quantitative confidence for the next phase of structure construc-

tion.

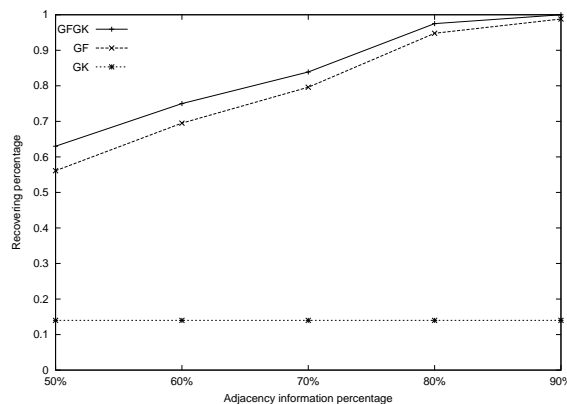
## Acknowledgments

XW is partially supported by PENCE and AICML. DX is supported by the Office of Biological and Environmental Research, U.S. Department of Energy, under Contract DE-AC05-00OR22725, managed by UT-Battelle, LLC. CMS is supported by PENCE. GL is supported by PENCE, AICML, NSERC, and a startup REE Grant from University of Alberta.

We thank Robert Boyko (PENCE), Zhi-Zhong Chen (Denki), Tao Jiang (UCR), Jianjun Wen (UCR), and Ying Xu (ORNL), for many helpful discussions.



(a) Comparison between the old and new score schemes using  $H_{\alpha}$  and  $C_{\alpha}$  chemical shifts. The curve marked with '+' is produced by GF-GK using the new score scheme based on  $H_{\alpha}$  and  $C_{\alpha}$  chemical shifts. The other three are using the previous score scheme based on  $H_{\alpha}$  and  $C_{\alpha}$  chemical shifts.



(b) Performance of the new score scheme using  $H^N$ ,  $N$ ,  $C_{\alpha}$  and  $C_{\beta}$  chemical shifts. The curve marked with '+' is produced by GF-GK using the new score scheme, which is significantly higher than the curve marked with '\*' which is produced without using adjacency information.

**Figure 3. Curves of the percentage of correct mapping recovery.**

## References

- [1] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.
- [2] H. S. Atreya, S. C. Sahu, K. V. R. Chary, and G. Govil. A tracked approach for automated NMR assignments in proteins (TATAPRO). *Journal of Biomolecular NMR*, 17:125–136, 2000.
- [3] C. Bailey-Kellogg, A. Widge, J. J. K. III, M. J. Berardi, J. H. Bushweller, and B. R. Donald. The NOESY Jigsaw: automated protein secondary structure and main-main assignment from sparse, unassigned NMR data. *Journal of Computational Biology*, 7:537–558, 2000.
- [4] C. Bartels, P. Güntert, M. Billeter, and K. Wüthrich. GARANT-A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *Journal of Computational Chemistry*, 18:139–149, 1997.
- [5] N. E. G. Buchler, E. P. R. Zuiderweg, H. Wang, and R. A. Goldstein. Protein heteronuclear NMR assignments using mean-field simulated annealing. *Journal of Magnetic Resonance*, 125:34–42, 1997.
- [6] Z.-Z. Chen, T. Jiang, G.-H. Lin, R. Rizzi, J. J. Wen, D. Xu, and Y. Xu. More reliable protein NMR peak assignment via improved 2-interval scheduling. Technical Report TR03-07, Department of Computing Science, University of Alberta, 2003.
- [7] Z.-Z. Chen, T. Jiang, G.-H. Lin, J. J. Wen, D. Xu, J. Xu, and Y. Xu. Approximation algorithms for NMR spectral peak assignment. *Theoretical Computer Science*, 299:211–229, 2003.
- [8] Z.-Z. Chen, T. Jiang, G.-H. Lin, J. J. Wen, D. Xu, and Y. Xu. Improved approximation algorithms for NMR spectral peak assignment. In *Proceedings of the 2nd Workshop on Algorithms in Bioinformatics (WABI 2002)*, volume 2452 of *Lecture Notes in Computer Science*, pages 82–96. Springer, 2002.
- [9] J. L. Devore. *Probability and Statistics for Engineering and the Science*. Duxbury Press, December 1999. Fifth Edition.
- [10] B. M. Duggan, G. B. Legge, H. J. Dyson, and P. E. Wright. SANE (structure assisted NOE evaluation): an automated model-based approach for NOE assignment. *Journal of Biomolecular NMR*, 19:321–329, 2001.
- [11] A. V. Goldberg and R. Kennedy. An efficient cost scaling algorithm for the assignment problem. *Mathematical Programming*, 71, 1995.
- [12] P. Güntert, M. Salzmann, D. Braun, and K. Wüthrich. Sequence-specific NMR assignment of proteins by global fragment mapping with the program mapper. *Journal of Biomolecular NMR*, 18:129–137, 2000.
- [13] T. Herrmann, P. Güntert, and K. Wüthrich. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of Molecular Biology*, 319:209–227, 2002.
- [14] T. K. Hitchens, J. A. Lukin, Y. Zhan, S. A. McCallum, and G. S. Rule. MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. *Journal of Biomolecular NMR*, 25:1–9, 2003.
- [15] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292:195–202, 1999.
- [16] E. L. Lawler. *Combinatorial Optimization: Networks and Matroids*. Holt, Rinehart and Winston, New York, 1976.

- [17] M. Leutner, R. M. Gschwind, J. Liermann, C. Schwarz, G. Gemmecker, and H. Kessler. Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *Journal of Biomolecular NMR*, 11:31–43, 1998.
- [18] K. B. Li and B. C. Sanctuary. Automated resonance assignment of protein using heteronuclear 3D NMR. 1. Backbone spin systems extraction and creation of polypeptides. *Journal of Chemical Information and Computational Science*, 37:359–366, 1997.
- [19] G.-H. Lin, D. Xu, Z. Z. Chen, T. Jiang, J. J. Wen, and Y. Xu. An efficient branch-and-bound algorithm for the assignment of protein backbone NMR peaks. In *Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB 2002)*, pages 165–174, 2002.
- [20] J. A. Lukin, A. P. Gove, S. N. Talukdar, and C. Ho. Automated probabilistic method for assigning backbone resonances of (13C, 15N)-labeled proteins. *Journal of Biomolecular NMR*, 9:151, 1997.
- [21] H. N. B. Moseley and G. T. Montelione. Automated analysis of NMR assignments and structures for proteins. *Current Opinion in Structural Biology*, 9:635–642, 1999.
- [22] U. of Wisconsin. BioMagResBank. 2001. <http://www.bmrb.wisc.edu>. University of Wisconsin, Madison, Wisconsin.
- [23] C. M. Slupsky, R. F. Boyko, V. K. Booth, and B. D. Sykes. SMARTNOTEBOOK: a semi-automated approach to protein sequential NMR resonance assignments. *Journal of Biomolecular NMR*, 2003. To appear.
- [24] C. M. Slupsky, M. Desautels, T. Huebert, R. Zhao, S. M. Hemmingsen, and L. P. McIntosh. Structure of Cdc4p, a contractile ring protein essential for cytokinesis in *Schizosaccharomyces pombe*. *Journal of Biological Chemistry*, 276:5943–5951, 2001.
- [25] Y. Xu, D. Xu, D. Kim, V. Olman, J. Razumovskaya, and T. Jiang. Automated assignment of backbone NMR peaks using constrained bipartite matching. *IEEE Computing in Science & Engineering*, 4:50–62, 2002.
- [26] D. Zimmerman, C. Kulikowski, Y. Huang, W. Tashiro, S. Shimotakahara, C. Chien, R. Powers, and G. Montelione. Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology*, 269:592–610, 1997.

	score	length	GK score	$R^{GK}$	GF-GK score	$R^{GF-GK}$	$R^{GF}$	sim. adj.	sim. score	sim. $R^{GF-GK}$
bmr 4027.5	1569792	158	1570527	22	1569790	107	98 (119)	88%	1569792	156
bmr 4027.6					1569546	126	118 (135)			
bmr 4027.7					1569590	142	135 (143)			
bmr 4027.8					1569792	158	152 (152)			
bmr 4027.9					1569792	158	155 (155)			
bmr 4144.5	773522	78	773987	13	773545	56	52 (58)	77%	773462	69
bmr 4144.6					773496	53	49 (65)			
bmr 4144.7					773435	63	60 (69)			
bmr 4144.8					773522	78	73 (73)			
bmr 4144.9					773522	78	75 (75)			
bmr 4288.5	1042182	105	1042432	12	1042137	79	71 (77)	75%	1042185	100
bmr 4288.6					1042184	97	85 (91)			
bmr 4288.7					1042184	99	88 (94)			
bmr 4288.8					1042182	105	101 (101)			
bmr 4288.9					1042182	105	104 (104)			
bmr 4302.5	1142379	115	1142856	23	1142322	84	75 (87)	72%	1142383	105
bmr 4302.6					1142379	97	92 (96)			
bmr 4302.7					1142318	101	99 (106)			
bmr 4302.8					1142421	112	112 (112)			
bmr 4302.9					1142379	115	114 (114)			
bmr 4309.5	1772122	178	1773531	14	1772541	56	53 (138)	56%	1772649	107
bmr 4309.6					1772287	87	87 (150)			
bmr 4309.7					1772031	107	105 (162)			
bmr 4309.8					1772140	172	168 (168)			
bmr 4309.9					1772122	178	176 (176)			
bmr 4316.5	885385	89	885398	16	885386	78	69 (69)	73%	885385	85
bmr 4316.6					885385	82	78 (78)			
bmr 4316.7					885385	85	83 (83)			
bmr 4316.8					885385	89	87 (87)			
bmr 4316.9					885385	89	88 (88)			
bmr 4318.5	2135541	215	2137107	16	2124939	88	84 (165)	83%	2135502	198
bmr 4318.6					2125425	147	141 (178)			
bmr 4318.7					2125003	159	155 (195)			
bmr 4318.8					2135435	203	200 (205)			
bmr 4318.9					2135541	215	215 (215)			
bmr 4353.5	1252058	126	1252609	22	1252105	84	74 (94)	75%	1252069	118
bmr 4353.6					1241711	75	69 (107)			
bmr 4353.7					1252060	121	115 (115)			
bmr 4353.8					1252058	126	123 (123)			
bmr 4353.9					1252058	126	124 (124)			
bmr 4391.5	655208	66	65523	11	645180	38	32 (50)	85%	655235	64
bmr 4391.6					655078	51	46 (56)			
bmr 4391.7					655085	53	49 (59)			
bmr 4391.8					655208	66	62 (62)			
bmr 4391.9					655208	66	65 (65)			
bmr 4393.5	1554375	156	1554971	9	1554470	48	48 (118)	67%	1554486	119
bmr 4393.6					1554281	62	62 (130)			
bmr 4393.7					1554156	96	95 (143)			
bmr 4393.8					1554245	125	125 (148)			
bmr 4393.9					1554375	156	154 (154)			
bmr 4579.5	854668	86	855151	10	854673	56	50 (69)	76%	854741	81
bmr 4579.6					844748	77	72 (76)			
bmr 4579.7					844575	71	65 (79)			
bmr 4579.8					854668	86	83 (83)			
bmr 4579.9					854668	86	85 (85)			
bmr 4670.5	1192228	120	1192920	13	1192068	65	58 (90)	69%	1192172	91
bmr 4670.6					1191926	89	84 (101)			
bmr 4670.7					1192137	110	103 (109)			
bmr 4670.8					1192228	120	117 (117)			
bmr 4670.9					1192228	120	120 (120)			
bmr 4752.5	675974	68	676109	15	675902	53	40 (48)	31%	676036	44
bmr 4752.6					675906	59	51 (56)			
bmr 4752.7					675805	58	55 (60)			
bmr 4752.8					675974	68	66 (66)			
bmr 4752.9					675974	68	67 (67)			
bmr 4929.5	1133001	114	1133518	18	1133086	93	84 (90)	66%	1133183	87
bmr 4929.6					1132999	101	91 (95)			
bmr 4929.7					1132950	108	101 (103)			
bmr 4929.8					1133001	114	109 (109)			
bmr 4929.9					1133001	114	113 (113)			
								71%		

**Table 3. Evaluation of the new score scheme. The “score” column records the assignment score of the correct assignment; The “GK score” column is the optimal matching by ignoring all the adjacency information. It can be seen that these two scores are very close to each other, indicating that the score scheme is very promising.  $R^{GF-GK}$  ( $R^{GF}$ ) stores the number of correct mappings recovered by GF-GK (GF, respectively), where the number in the parentheses in the 8th column stores the number of spin systems in the identified chains. The 9th to 11th columns are the percentage of adjacency extracted from the simulated HSQC, HNCACB, and CBCA(CO)NH spectra, the score of the assignment output by GF-GK, and the number of correct mapping recovered, respectively.**