Automated Extracting of Amino Acid Spin Systems in Proteins Using 3D HCCH-COSY/TOCSY Spectroscopy and Constrained Partitioning Algorithm (CPA)

Kuo-Bin Li and B. C. Sanctuary*

Department of Chemistry, McGill University, 801 Sherbrooke St. West, Montréal, PQ, H3A 2K6 Canada

Received September 6, 1995[⊗]

An automated approach has been developed to extract amino acid spin systems from proteins or peptides by analyzing 3D HCCH-COSY/TOCSY spectra through a constrained partitioning algorithm. The key step in this approach is that two cross peaks are merged to form a spin system not only by sharing one frequency but also by observing additional contraining cross peaks, such as from COSY and/or TOCSY data. The extracted amino acid spin systems show the connectivity relationships between these spins as well as their chemical shifts. This feature differs from other automated spin system extraction programs making it easier to be applied to automatic amino acid recognition applications. The 90 amino acid protein N-domain of chicken skeletal troponin-C (1−90) was used to test the implementation of the algorithm with both simulated and real experimental data. Limitations of the algorithm are discussed.

INTRODUCTION

NMR based protein structure determination techniques have been widely used since early 1980s. The established procedure consists of several major steps.[1−4] First, the spin systems of all of the amino acid residues in the protein are identified, then a sequential assignment procedure attempts to map the extracted spin systems to the target protein's primary sequence. The results of the resonance assignments are then used to interpret through-space NOE cross peaks, from which a number of distance constraints can be derived from analysis of the NOESY data. Finally, these constraints are used to calculate the protein's 3D structure.

It is generally accepted that the resonance assignment of NMR data is tedious and time-consuming work, hence, there have been many attempts[5−17] to automate the resonance assignment part of the structure determination analysis. A self-contained automated assignment strategy should consist of three steps: (1) extraction of spin coupling systems, (2) mapping of the spin coupling systems to amino acid residues, and (3) searching for a most probable spin system sequence which matches the protein's primary sequence. All of these steps can be treated by a series of algorithms: constrained partitioning (CPA),[18,19] fuzzy pattern recognition,[20] and tree searching.[20,21] CPA can automatically extract and identify spin coupling networks from a combination of 2D COSY spectrum and TOCSY spectrum where the latter is used as partitioning constraints. The aim of this paper is to extend the CPA algorithm to 3D NMR and present a computer assisted spin systems extraction procedure based on 3D HCCH-COSY and HCCH-TOCSY NMR spectra.

The introduction of 3D NMR, combined with uniform ^{15}N and/or ^{13}C isotope labeling, significantly extends the molecular weight limit of proteins for which a solution state structure can be determined by NMR. Resonance assignment of a protein's backbone can be achieved by a combination of several triple resonance 3D NMR experiments.[22] Furthermore, to obtain a protein's detailed structures, the NOE side chain resonance cross peaks must be unambiguously assigned so that enough distance constraints can be produced to construct the protein side chain orientation. The NOE cross peak analysis usually requires the side chain resonance assignment to be completed. Several 3D NMR experiments have been proposed for protein side chain resonance assignment, such as 3D HCCH-COSY,[23−25] HCCH-TOCSY,[26] HCC(CO)NH-TOCSY,[27,28] and HCCNH-TOCSY.[27,29]

Using software to analyze 3D or 4D NMR data should be preferred over manual analysis since spectral data becomes complicated and difficult to handle manually. Among the several attempts for automated analysis of 3D NMR, two of them[7,14] are studies of the application of homonuclear 3D NMR on protein proton resonance assignments. The rest of the approaches use triple resonance heteronuclear 3D NMR to obtain protein backbone assignment[15] and to establish sequential connectivity of amino acid spin systems.[16,17] The availability of the information about the spin systems, including the backbone and side chain resonances as well as the amino acid types, is crucial in all these methods. However, in all of the heteronuclear 3D NMR approaches mentioned above, the spin system information has to be obtained elsewhere manually. This paper is directed in this regard to design an automatic strategy to obtain the information of protein amino acid spin systems. In this paper an algorithm is proposed to extract aliphatic side chain spin systems from heteronuclear 3D NMR data of proteins. The algorithm merges cross peaks from a 3D NMR spectrum, such as a 3D HCCH-COSY, to form spin coupling systems. At each merging step at least two constraints are required to assure the validity of the merging. Thus an additional NMR spectrum, such as a 3D HCCH-TOCSY, can be used to supply these constraints. The output spin coupling systems are given as a series of graphs represented as adjacency lists which can be processed by a subsequent graph pattern recognition algorithm[20] to perform the amino acid identification.

METHODS

The central ideal of the algorithm is to extract amino acid spin systems from NMR spectra. To illustrate how this approach works, a simple three-spin system is first considered

* To whom correspondence should be addressed.

[⊗] Abstract published in *Advance ACS Abstracts,* January 1, 1996.

0095-2338/96/1636-0585$12.00/0

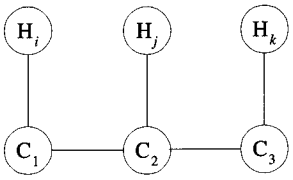

Figure 1. Example of a chemical structure fragment with three protons.

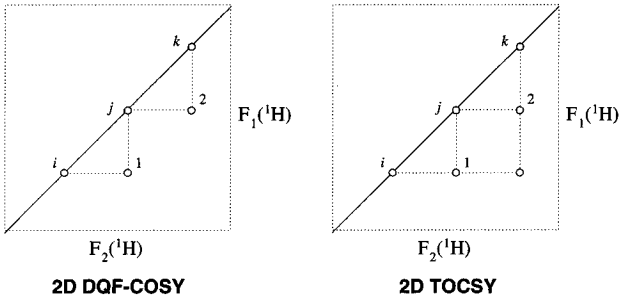

Figure 2. 2D DQF-COSY and TOCSY spectra of the chemical structure shown in Figure 1. The peaks on other sides are not displayed for convenience.

(see Figure 1). On a 2D DQF-COSY NMR spectrum, such a three-spin systems gives two cross peaks on each side of the diagonal, while in a 2D TOCSY spectrum, one extra peak is observed on each side (see Figure 2). To construct this three-spin system from the cross peaks, the usual manual assignment procedure probably starts from cross peak 1 (see Figures 1 and 2), then observes cross peak 2. In terms of an automated computer procedure, for peak 1 (δ_i, δ_j), and peak 2 ($\delta_{j'}$, δ_k), if δ_j and $\delta_{j'}$ are close enough (controlled by a predefined tolerance value), a three-spin system (*i*, *j*, *k*) can be constructed. Applying this procedure to the whole cross peak data set enables, in principle, all the amino acid spin systems to be extracted. However, in certain regions of the spectrum, heavy overlap makes this kind of merging process unreliable.

Suppose, for example, we have two three-spin systems, $\{\delta_i, \delta_j, \delta_k\}$, $\{\delta_l, \delta_m, \delta_n\}$, and coincidentally two spins, *j* and *m*, have resonance frequencies which are similar in values (see Figure 3). The COSY cross peaks they produced are (δ_i, δ_j), ($\delta_{j'}$, δ_k), (δ_l, δ_m), and ($\delta_{m'}$, δ_n) where δ_j, $\delta_{j'}$, δ_m, and $\delta_{m'}$ are difficult to distinguish in terms of chemical shifts. In analyzing these peaks to form unique spin systems, it is necessary to know whether cross peak (δ_i, δ_j) merges with ($\delta_{j'}$, δ_k) or ($\delta_{m'}$, δ_n). An extra constraint is needed to remove this ambiguity. One way is to look at the TOCSY spectrum. If spins *i*, *j*, *k* are indeed in the same spin system, i.e., δ_j and $\delta_{j'}$ come from the same spin, a TOCSY cross peak (δ_i, δ_k) should be observed. Similarly, if *i*, *j*, *n* are in the same spin system, δ_j and $\delta_{m'}$ come from the same spin, another TOCSY cross peak (δ_i, δ_n) should be observed. Hence by cross referencing with such TOCSY constraints, one can reduce the possibility of this kind of ambiguity, making it possible to design an automated spin system extraction procedure.

As the size of the target protein increase, the corresponding 2D NMR spectrum becomes more crowded. It is unlikely that one constraint alone can resolve the overlap situation when doing peak merging. One solution is to acquire another 2D NMR spectrum which may provide additional information to resolve the overlap. Another way is to introduce a third dimension in which another nucleus can be used as the additional constraint. The former was treated previously,[18,19] while in this paper we discuss the latter.

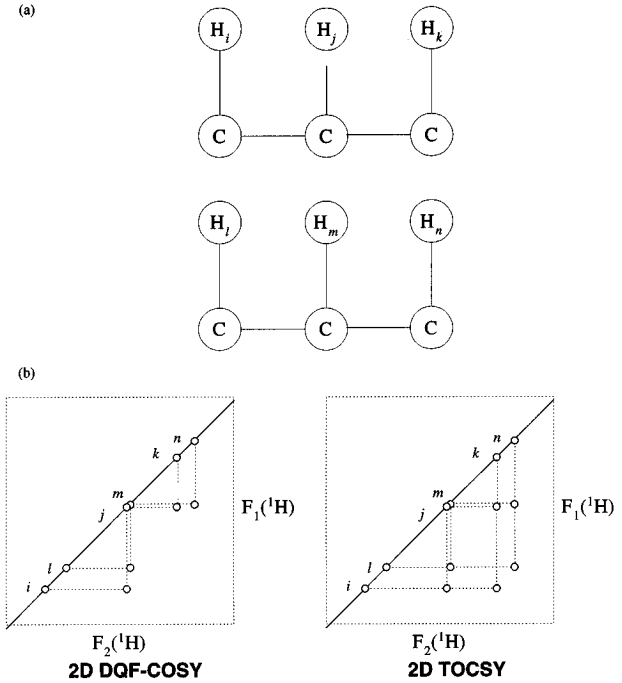

Figure 3. (a) Three-spin systems $\{\delta_i, \delta_j, \delta_k\}$ and $\{\delta_l, \delta_m, \delta_n\}$ where δ_j and δ_m are within a chemical shift tolerance. (b) 2D DQF-COSY and TOCSY spectra of the above two spin systems.

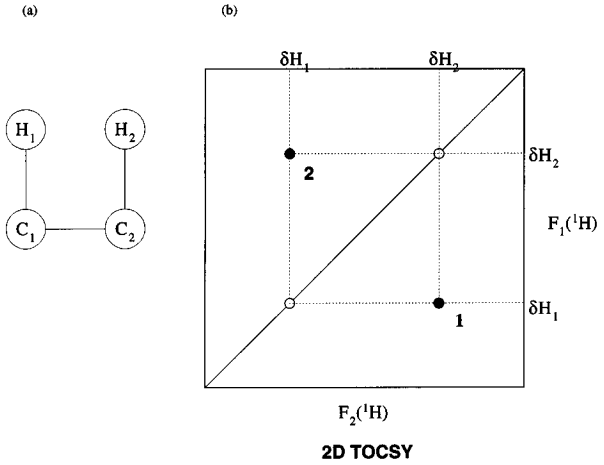

Figure 4. (a) Structure of a CH-CH fragment. (b) The corresponding 2D TOCSY spectrum. Cross peak 1 has chemical shifts (δH_1, δH_2), and cross peak 2 has chemical shifts (δH_2, δH_1).

The complete amino acid spin systems of a protein's side chain can be determined by 3D HCCH-COSY and HCCH-TOCSY experiments.[23,25,26,30] Both experiments make use of one bond $^1H-^{13}C$ ($\sim$140 Hz) and $^{13}C-^{13}C$ ($\sim$30$-$40 Hz) *J* couplings to transfer magnetization along the side chain via the pathway

$$^1H \xrightarrow{^1J_{CH}} {}^{13}C \xrightarrow{^1J_{CC}} {}^{13}C \xrightarrow{^1J_{CH}} {}^1H$$

To interpret 3D HCCH COSY/TOCSY spectra, consider first a 2D TOCSY segment. Figure 4 shows the spectrum that corresponds to the chemical structure shown on the left of the figure. Figure 5 shows the 3D HCCH-TOCSY spectrum of the same chemical structure as in Figure 4. Figure 5 shows the 3D HCCH-TOCSY spectrum of the same chemical

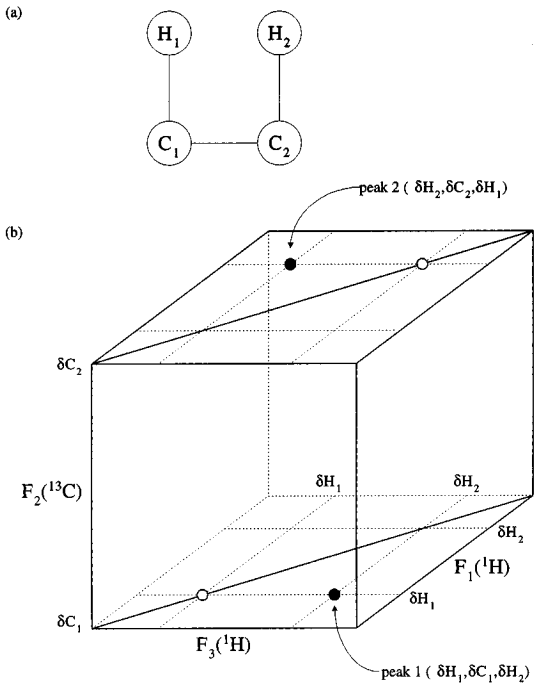

Figure 5. (a) The same structure as in Figure 4a. (b) The corresponding 3D HCCH-TOCSY spectrum. The $^1H(F_1)-^1H(F_3)$ planes are similar to that of 2D $^1H-^1H$ COSY or TOCSY experiment, except that $^1H(F_1)-^1H(F_3)$ are edited by the chemical shift of the ^{13}C nuclei. Note that peaks 1 and 2 do not occur symmetrically on both sides of the diagonal on the same plane.

structure as in Figure 4. The $^1H(F_1)-^1H(F_3)$ planes are similar to that of 2D $^1H-^1H$ COSY or TOCSY experiment, except that these planes are edited by the chemical shifts of the ^{13}C nuclei. Hence off-diagonal peaks in an $^1H-^1H$ plane at the ^{13}C frequency arise from protons directly bonded to the ^{13}C. For example, in Figure 5, the magnetization transfer pathway of cross peak 1 (δH_1, δC_1, δH_2) follows the path $H_1 \rightarrow C_1 \rightarrow C_2 \rightarrow H_2$, while the transfer pathway of cross peak 2 (δH_2, δC_2, δH_1) has path $H_2 \rightarrow C_2 \rightarrow C_1 \rightarrow H_1$. The cross peaks 1 and 2 in 3D HCCH experiments do not occur symmetrically on both sides of the diagonal of the same plane, but rather, occur on different F_1-F_3 planes as shown in Figure 5.

Algorithm. The NMR data sets used in the present algorithm are 3D HCCH-COSY and 3D HCCH-TOCSY. Currently the implemented computer program is designed to process cross peak lists. That is, cross peaks in the spectra must have been previously picked by a reliable peak picking procedure. In the peak list, cross peaks are represented by three chemical shift coordinate points, e.g., (3.52, 58.17, 1.46), where the first coordinate denotes the resonance frequency of the proton which is directly bonded to the carbon. The frequency of that carbon is the second coordinate, while the third coordinate is the frequency of another proton which can be reached by the transfer of magnetization along the side chain via the HCCH pathway. In the following context, a 3D cross peak is represented as (H_i, C_1, H_j). The corresponding chemical structures of a COSY cross peak (H_i, C_1, H_j) are shown in Figure 6.

The algorithm starts with the entire HCCH-COSY data set being searched to find pairs of cross peaks, (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k), which have one proton and one carbon resonance frequencies in common. In the algorithm, H_i, $H_{i'}$ and C_1, $C_{1'}$ are tested to determine whether they are within

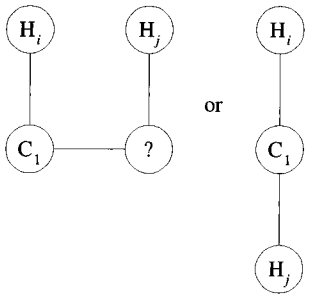

Figure 6. The possible chemical structures corresponding to a 3D HCCH-COSY cross peak (H_i, C_1, H_j). In the left one, the chemical shift of the carbon to which H_j bonds is undetermined.

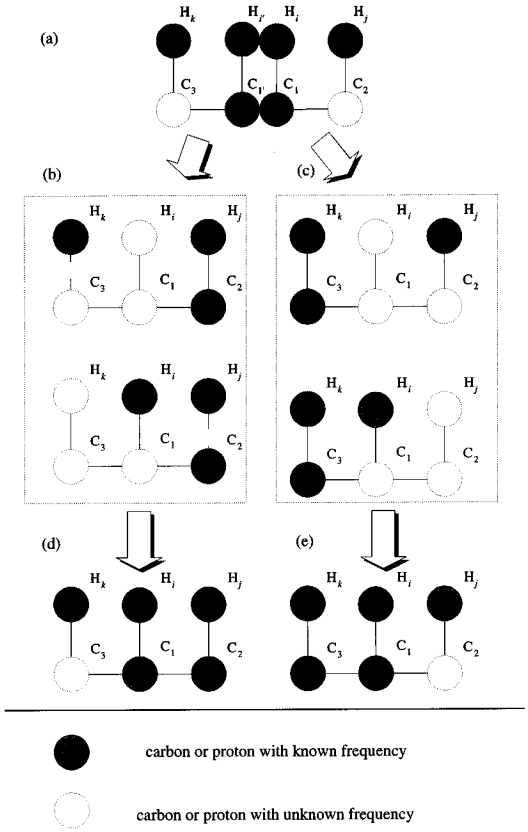

Figure 7. Schematic representation showing how two 3D HCCH-COSY cross peaks (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k) are merged to form a spin system. Each cross peak contains three frequencies depicted by filled circles, while the open circles indicate the frequencies are unknown from the cross peak data. (a) Two cross peaks (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k), where H_i, $H_{i'}$ and C_1, $C_{1'}$ are within the specified tolerances. (b) Two possible constraint cross peaks (H_j, C_2, H_k) and (H_j, C_2, H_i). (c) Another two possible constraint peaks (H_k, C_3, H_j) and (H_k, C_3, H_i). (d) A possible merged spin system with three protons, H_i, H_j, H_k, and three carbons, C_1, C_2, C_3. (e) Another possible merged spin system. Two peaks in (a) along with two constraint peaks in (b) lead to the spin system in (d). Two peaks in (a) along with two constraint peaks in (c) lead to the spin system shown in (e). In summary, (a)−(b)−(d) is one possible pathway to merge two cross peaks while (a)−(c)−(e) is another.

a user defined chemical shift tolerance, such as 0.02 ppm for proton and 0.20 ppm for carbon.

There are three different cases regarding connectivity between protons and carbons to be considered in merging cross peaks into spin systems. The first is that all three protons, H_i, H_j, and H_k, bond to different carbons. A schematic view shown in Figure 7 shows how two HCCH-COSY cross peaks, along with various constraint peaks, can

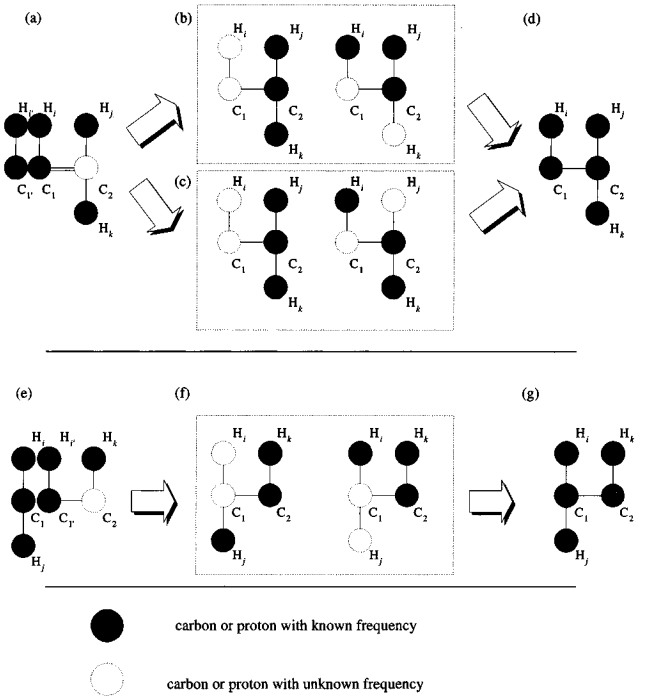

Figure 8. (a) Two 3D HCCH-COSY cross peaks (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k). (b) Two possible constraint peaks (H_j, C_2, H_k) and (H_j, C_2, H_i). (c) Another two possible constraint peaks (H_k, C_2, H_j) and (H_k, C_2, H_i). (d) A merged spin system with three protons, H_i, H_j, H_k, and two carbons, C_1 and C_2. Two peaks in (a) along with two constraint peaks either in (b) or (c) lead to the spin system in (d). (e) Two 3D HCCH-COSY cross peaks. (f) Two constraint peaks. (g) A merged spin system with three protons, H_i, H_j, H_k, and two carbons, C_1 and C_2. Two peaks in (e) along with two constraint peaks in (f) give rise to the spin system in (g).

arrive at a merged spin system. Figure 7d is the first possible merged spin system which is formed from Figure 7a along with two constraint peaks shown in Figure 7b. Similarly, the spin system in Figure 7c can be obtained from two cross peaks shown in Figure 7a along with two constraint peaks in Figure 7c.

A second case occurs when H_j and H_k bond to the same carbon as shown in Figure 8. One of two possible constraint peak sets, Figure 8b (part or c), is required to confirm that the spin system shown in Figure 8d can be constructed. A third case has H_i and H_j bonded to the same carbon as shown in Figure 8e. The presence of two constraint peaks, Figure 8f, confirms the spin system shown in Figure 8g. To summarize the above pictorial representations, Figure 9 shows the control flow of the partitioning algorithm.

Figure 10 is an example of part of the output for an amino acid spin system. Note that both protons' and carbons' resonance frequencies are recorded. The connectivity relationship between protons is also displayed numerically.

As the number of peaks and complexity of a spectrum increases, the uniqueness of the merging process is compromised. In other words, for a specific peak, it is common that more than one candidate peak can be merged to it. This is mainly due to spectral overlap, making it necessary to design a strategy to rank the candidate peaks, i.e., to select the most likely merging from the many possibilities.

In the partitioning algorithm a scoring parameter is introduced to rank every possible candidate peak. Consider a cross peak, (H_i, C_1, H_j), with which a candidate peak, ($H_{i'}$, $C_{1'}$, H_k), can be merged based on the presence of the

Step1 Search the HCCH-COSY cross peak list for pairs of (H_i, C_1, H_j) and ($H_{i'}, C_{1'}, H_k$), where H_i and $H_{i'}$ are within the ^1H chemical shift tolerance range, and C_1 and $C_{1'}$ are within the ^{13}C chemical shift tolerance range. Do the following steps to test if H_i, C_1, H_j and H_k can be added to a spin system.

Step2 If a HCCH-TOCSY (H_j, C_2, H_k) is found

and a HCCH-COSY (H_j, C_2, H_i) or HCCH-TOCSY (H_j, C_2, H_i) is found

then add H_i, C_1, H_j, H_k and C_2 to a spin system.

Step3 else if a HCCH-TOCSY (H_k, C_2, H_j) is found

and a HCCH-COSY (H_k, C_2, H_i) or HCCH-TOCSY (H_k, C_2, H_i) is found

then add H_i, C_1, H_j, H_k and C_2 to a spin system.

Step4 else if a HCCH-COSY (H_j, C_2, H_k) is found

and a HCCH-COSY (H_j, C_2, H_i) or HCCH-TOCSY (H_j, C_2, H_i) is found

then add H_i, C_1, H_j, H_k and C_2 to a spin system.

Step5 else if a HCCH-COSY (H_k, C_2, H_j) is found

and a HCCH-COSY (H_k, C_2, H_i) or HCCH-TOCSY (H_k, C_2, H_i) is found

then add H_i, C_1, H_j, H_k and C_2 to a spin system.

Step6 Back to **Step1** until no more COSY cross peak pair fulfilled the condition of **Step1** remain in the data set.

Figure 9. Control flow of the partition algorithm.

constraints already discussed (see Figure 9). Two constraints might be the presence of peaks ($H_{k'}$, C_2, $H_{j'}$) and ($H_{k''}$, $C_{2'}$, $H_{i''}$). The scoring parameter A is defined as

$$A = 1 - \sqrt[3]{\left(\frac{w_0}{T_H}\right)\left(\frac{D}{2T_H}\right)\left(\frac{w_1}{T_C}\right)}$$

where

$$w_1 = |\delta_{C_1} - \delta_{C_{1'}}|$$

$$w_0 = |\delta_{H_i} - \delta_{H_{i'}}|$$

$$T_H = \text{tolerance value of proton}$$

$$T_C = \text{tolerance value of carbon}$$

$$D = \frac{(D_1 + D_2)}{2}$$

with D_1 and D_2 depending on the constraining peaks as

$$D_1 = |\delta_{H_k} - \delta_{H_{k'}}| + |\delta_{H_j} - \delta_{H_{j'}}|$$

$$D_2 = |\delta_{H_k} - \delta_{H_{k''}}| + |\frac{(\delta_{H_i} + \delta_{H_{i'}})}{2} - \delta_{H_{i''}}|$$

w_0 measures the difference of the chemical shift value between the original and candidate peaks in the first coordinate of a 3D cross peak. T_H is the tolerance value, which the user chooses for the proton dimensions (F_1 and F_3). Candidate peaks which make w_0 greater than T_H are discarded, thus w_0 must be less than or equal to T_H, or $w_0/T_H \leq 1$.

w_1 measures the difference of the carbon chemical shift values between the original and candidate peaks, and T_C is the tolerance value of carbon, hence, $w_1/T_C \leq 1$.

D measures how well the two constraint peaks match the original and candidate peaks. A smaller D value corresponds to a better match.

The above three factors are used to decide how good a candidate peak is. In terms of the first factor w_0, a smaller proton chemical shift difference between H_i and $H_{i'}$ indicates a better match of the cross peak (H_i, C_1, H_j) and ($H_{i'}$, $C_{1'}$, H_k). Secondly, a smaller carbon chemical shift difference,

i.e., a smaller w_1, between C_1 and $C_{1'}$ also indicates a better match. Finally, *D* uses constraint peaks to evaluate this match.

The computer program calculates the scoring parameter for each of the merging pair giving a score from 0 to 1. A higher value of *A* is taken as a better match. Under such a scoring strategy, the candidate peak with a largest value of *A* is chosen to merge with the original peak.

RESULTS

The algorithm was implemented in C and C++ programming languages and tested on both real and simulated 3D HCCH-COSY/TOCSY data for a 90 amino acids protein N-domain of chicken skeletal troponin-C (1−90). The experimental spectra and manual assignments were provided by University of Alberta. The simulated data were generated based upon the manual assignments. Both exact and dispersive (with respect to chemical shifts, described later) simulations were used. The testing procedures and results are described below.

Analysis of Simulated HCCH-COSY/TOCSY Data for Troponin-C. Simulations were generated based upon manual assignments done previously.[31] Here an example is given to illustrate how the simulation were done. Figure 11 shows the manual assignment for Met3 and Thr4 which are used to generate the COSY and TOCSY peaks that must exist for these residues. The generated peaks are also shown. Resonance frequencies from C_β, H_β, C_γ, and H_γ are missing for Met3. Therefore no cross peaks can be simulated from the manual assignment for this residue. For Thr4, four HCCH-COSY cross peaks can be generated, among them two are symmetrical cross peaks. Similarly, six HCCH-TOCSY cross peaks can be generated as there are two additional peaks of (H_α, C_α, H_γ) and (H_γ, C_γ, H_α).

At the first stage of testing, no chemical shift dispersions were introduced in the simulated data set. That is, two cross peaks are allowed to be partitioned into a spin system as long as they share exactly the same chemical shift value. The chemical shift tolerance value is therefore zero. The purpose of using the exactly simulated data is to confirm that the algorithm works as designed. A total of 674 HCCH-COSY cross peaks and 1014 HCCH-TOCSY were simulated for the 90 amino acids which occur in troponin-C protein. Note that among all of the amino acid residues, glycine is considered to be a two-spin system. Each has two H_α protons, because the amide proton is not detectable in HCCH spectrum. Similarly, alanine, which contains one H_α and three methyl H_β, is also a two-spin system. The algorithm described above was designed to extract amino acid spin systems with three or more spins, so alanines and glycines are excluded in this particular test case. These cases will be considered during the real data testing stage presented later in this paper. Another point of note is that the chemical shifts data of aromatic carbons are not available since their resonance frequencies are much higher (∼130 ppm) than that of aliphatic carbons. As a consequence of the above, and due to several residues not being detected in the manual assignments, only 63 residues of the 90 were simulated.

The test results are summarized in Table 1. Note that the algorithm detects all the spin systems that were included in the simulated data of 63 residues which consisted of 674 HCCH-COSY and 1014 HCCH-TOCSY peaks. The execution time for this running is about 3 min on a Silicon Graphics workstation with a 133 MHz MIPS R4600SC CPU.

Table 1. Results for Test of Simulated Data I[c]

residues	no. of occurrence of a residue	no. of S.S. simulated as input	no. of S.S. obtained from output	remarks
Gly	7	N/A	N/A	spin systems with two spins are not tested[a]
Ala	10	N/A	N/A	spin systems with two spins are not tested[b]
Asp	10	10	10	
Glu	13	9	9	E41,57,67,77 were not simulated due to incomplete data
Lys	4	4	4	
Met	8	7	7	
Gln	4	3	3	
Arg	3	3	3	
Val	4	4	4	
Leu	5	4	4	
Phe	6	4	4	
Ile	5	5	5	
Thr	5	5	5	
Ser	4	3	3	
Pro	1	1	1	
Asn	1	1	1	
total	90	63	63	

[a] Gly has two H_α which produces only one cross peak pair. This is excluded from the simulation. [b] For the same reason as Gly. [c] See text for details.

```
/*8th G/       Total Peaks= 2
//Peak 19 (4.652 , 70.400 , 4.438)
//Peak 20 (4.652 , 70.400 , 1.176)
//Spin Coupling Topological Graph:
1H,4.652(70.400),2,3
2H,4.438(61.085),1
3H,1.176(21.310),1
```

Figure 10. Example of an extracted spin system represented by an adjacency list. In this case, two HCCH-COSY cross peaks (no. 19 and no. 20) were merged into a three-proton spin system. Proton 1 (4.652 ppm) bonds to a carbon (70.400 ppm) and couples to proton 2 (4.438 ppm) and proton 3 (1.176 ppm). Proton 2 (4.438 ppm) bonds to a carbon (61.085) and couples to proton 1 (4.652 ppm).

In a second test, the manual assignments which result in 63 spin patterns, were modified by the introduction of chemical shift dispersion. That is, to better simulate real experimental data, a systematic dispersion less than a predefined tolerance was introduced for every frequencies. The main aim of this test is to inspect the algorithm's capability of handling ill-aligned cross peaks. To better explain the dispersion, consider a three-spin system AMX. In principle there should be three cross peaks occurring on either side of the diagonal of a COSY or TOCSY spectrum. These three peaks are represented as (δ_A, δ_X), (δ_A, δ_M), and (δ_M, δ_X). The simulation of dispersion involves a pseudo-random number generator which gives random numbers R_i between −0.5 and +0.5. The simulated cross peaks are thus modified to ($\delta_A + R_1T$, $\delta_X + R_2T$), ($\delta_A + R_3T$, $\delta_M + R_4T$) and ($\delta_M + R_5T$, $\delta_X + R_6T$), where T is a tolerance value. For this particular testing case, T is set to 0.02 ppm for protons and 0.20 ppm for carbons.

An example of a spin system and its simulated COSY/TOCSY cross peaks are listed in Figure 12 which can be compared with Figure 11.

The result of applying the algorithm to the randomly distributed data set is listed in Table 2. Fifty-six of the 63

```
3 MET
N
HN
CA                55.950
HA                 3.840
CB
HB1
HB2
CG
HG1
HG2
CE                16.600
HE                 2.070
C                177.100

4 THR                                  ---> COSY
N                116.090                      4.438 61.085 4.652
HN                 8.013                      4.652 70.400 4.438
CA                61.085                      4.652 70.400 1.176
HA                 4.438                      1.176 21.310 4.652
CB                70.400
HB                 4.652                ---> TOCSY
CG2               21.310                      4.438 61.085 1.176
HG2                1.176                      1.176 21.310 4.438
C                175.000
```

Figure 11. Extract of the manual assignment listing of N-domain of chicken skeletal troponin-C (1−90). Met3 and Thr4 are shown here. Some resonances were not assigned, for example, C_β and H_β of Met3. For Met3, the assigned resonances are not sufficient to simulate a three bond coupling cross peak. The simulated cross peaks for Thr4 are shown on the right of Thr4's manual assignment.

```
4 THR                                  ---> COSY
N                116.090                      4.436 61.102 4.645
HN                 8.013                      4.645 70.403 4.433
CA                61.085                      4.644 70.338 1.168
HA                 4.438                      1.183 21.309 4.656
CB                70.400
HB                 4.652                ---> TOCSY
CG2               21.310                      4.445 61.106 1.171
HG2               1.176                       1.175 21.390 4.445
C                175.000
```

Figure 12. Extract of the manual assignment of N-domain troponin-C (1−90). Thr4 is shown. The simulated cross peaks for Thr4 are shown on the right. Note that a small chemical shift dispersion is introduced in the simulation, for example, 4.436 vs 4.438.

residues are successfully partitioned, and no missing assignment was found. Of the residues that are not successfully separated, four are glutamines, one is methionine, and two are isoleucines. These residues have highly overlapped resonance frequencies, for example, Figure 13 shows that E16 has four spins which are heavily overlapped with E21. The inability to resolve such overlapped spins is discussed in the next section.

Analysis of Experimental 3D HCCH-COSY/TOCSY Data. The success of the test on the simulated data indicates that the chemical shift degeneracy problem can be successfully resolved by the algorithm. The capability of handling missing peaks and spectrum artifacts is however inadequately tested by simulated data, and real experimental data are required for this purpose.

3D HCCH-COSY/TOCSY spectra of the test protein troponin-C were obtained from University of Alberta. Cross peaks in these spectra were picked automatically from a quick run of CAPP software.[32] No refinement in terms of peak picking were done since the original spectra and the peak picking program are not available to authors. A total of 915 HCCH-COSY and 710 TOCSY cross peaks were picked by the CAPP software. Three hundred twenty-one of the 915 COSY peaks and 225 of the 710 TOCSY peaks can be verified as real peaks by comparing with earlier done manual assignment result.

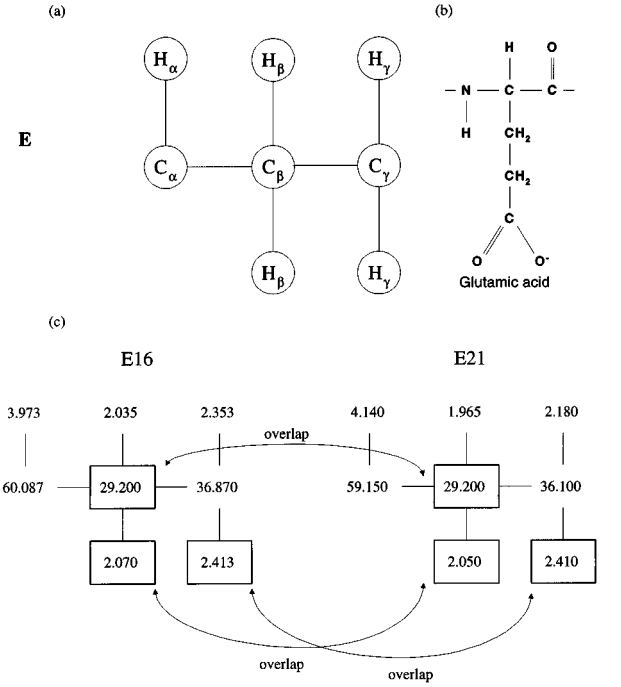

Figure 13. (a) The graph representation of a glutamic acid. (b) The chemical structure of a glutamic acid. (c) Glu16 and Glu21 are shown with their chemical shifts. Resonances in the boxes overlap.

Since extensive spectrum folding is employed in these multidimensional NMR experiments, the actual ^{13}C chemical shifts are given by $x \pm nSW$, where x is the ppm value of a carbon obtained from the spectrum, n is an integer, and SW is the spectral width. It is necessary to unfold the ^{13}C chemical shifts so that our spin system extracting program can work on real ^{13}C chemical shift data. A ^{13}C 2D HMQC spectrum is available from the same source for this unfolding purpose. The unfolding procedures are divided into two stages. First each of the HCCH-COSY and TOCSY cross peaks (H_i, C_i, H_j) are examined against the ^{13}C HMQC peak list. If a 2D ^{13}C HMQC cross peak (H_i, $C_i - SW$) were found, the 3D cross peak was corrected to (H_i, $C_i - SW$, H_j). The same procedure is also applied to HMQC peaks (H_i, C_i) and (H_i, $C_i + SW$). Secondly, for each 3D cross peak (H_i, C_i, H_j), if no corresponding 2D ^{13}C HMQC (H_i, $C_i \pm nSW$) were found, a statistical ^{13}C chemical shift database[33] was used to empirically determine the unfolded value of carbon. Following this the 915 HCCH-COSY peaks and 710 TOCSY peaks were used as input for our program. Various proton and carbon chemical shifts tolerances were checked to get good partitioning. Essentially, a small tolerance value generates more reliable results. In practice, however, small tolerance are not able to find enough spin systems due to the experimentally inconsistent chemical shift values, i.e., the same spin could have different chemical shift values in different spectra. On the other hand, a large tolerance value would incorrectly merge separate amino acid spin systems together. Compromise values of tolerance should be chosen carefully. Table 3 shows the partitioning results of the 915 COSY and 710 TOCSY peaks based upon the proton chemical shift tolerance 0.03 ppm and ^{13}C tolerance 0.40 ppm.

As can be seen from Table 3 some of the amino acid spin systems were incorrectly merged together, e.g., A20, A24, and A60 produced one large spin system. This is because

Table 2. Results for Test of Simulated Data II[c]

residues	no. of occurrence of a residue	no. of S.S. simulated as input	no. of S.S. obtained from output	remarks
Gly	7	N/A	N/A	spin systems with two spins are not tested[a]
Ala	10	N/A	N/A	spin systems with two spins are not tested[b]
Asp	10	10	10	
Glu	13	9	5	E9,16,21,63 were not separated
Lys	4	4	4	
Met	8	7	6	M46 were not separated with glutamine
Gln	4	3	3	
Arg	3	3	3	
Val	4	4	4	
Leu	5	4	4	
Phe	6	4	4	
Ile	5	5	3	I19 and I62 are not separated
Thr	5	5	5	
Ser	4	3	3	
Pro	1	1	1	
Asn	1	1	1	
total	90	63	56	

[a] Gly has two H_α which produces only one cross peak pair. This is excluded from the simulation. [b] For the same reason as Gly. [c] See text for details.

Table 3. Results for Testing of Real Data

residues	no. of occurrence of a residue	no. of A.A. obtained from output	remarks
Gly	7	5	G33,43,50,69,71
Ala	10	10	A1,8,10,12,25,31,90, (A20,24,60 not separated)
Asp	10	4	D89, (D5,27,59 not separated)
Glu	13	4	E9,16,21, (E17,M18,V65 not separated)
Lys	4	3	K40,55 (K87, Q85 not separated)
Met	8	3	M3,18,86
Gln	4	2	Q51,85
Arg	3	3	R11,47,84
Val	4	2	V65,80
Leu	5	4	L14,42,58,79
Phe	6	1	F13
Ile	5	5	I19,37,61,62,73
Thr	5	4	T4,39,44,54
Ser	4	3	S2,38,70
Pro	1	1	P53
Asn	1	1	N52
total	90	55	

all of their resonance frequencies overlap. By checking the result taken from manual assignment, these three alanine share common H_α, H_β, C_α, and C_β frequencies, see Figure 14.

Another point of note from Table 3 is that some spin systems are missing. For example, out of 10 aspartic acids, only four can be found. This is mainly due to the missing of crucial peaks in the experimental data. Aspartic acid is an AMX spin system and therefore should have one αH and two βH. According to our algorithm, all of the correlations between (H_α, $H_{\beta1}$), (H_α, $H_{\beta2}$), and ($H_{\beta1}$, $H_{\beta2}$) must be observed in order to place H_α, $H_{\beta1}$, and $H_{\beta2}$ into a spin system. The algorithm's condition is stricter than normal manual assignments procedure since avoiding incorrect merges is essential for computer assisted assignment tool. By carefully checking the peak lists, for D30, D32, D36, D59, D66, D68, and D89, it is found that the correlations between $H_{\beta1}$ and $H_{\beta2}$ are all missing, i.e., neither COSY ($H_{\beta1}$, C, $H_{\beta2}$) nor TOCSY ($H_{\beta1}$, C, $H_{\beta2}$) cross peaks were found in the peak lists. This is probably due to the fact that these βH cross peaks are too close to the diagonal to be unambiguously identified.

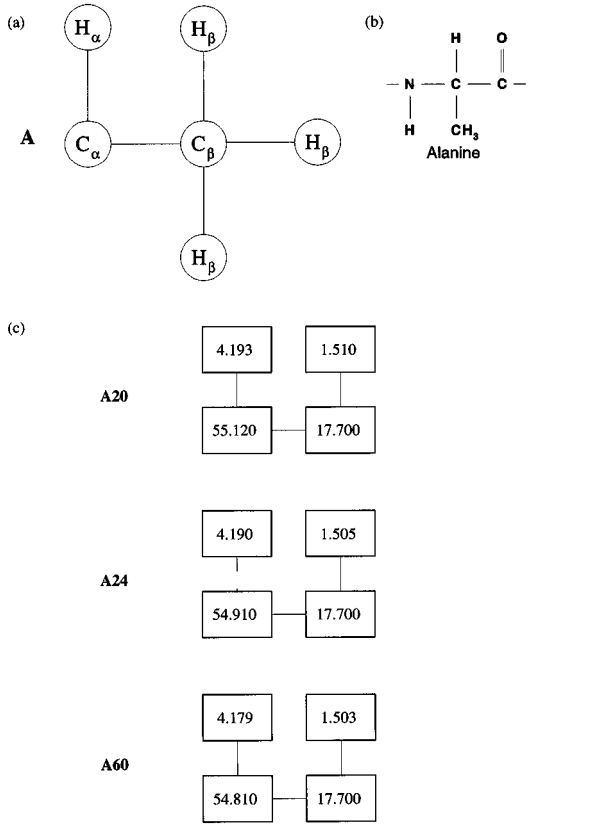

Figure 14. (a) The graph representation of an alanine. (b) The chemical structure of an alanine. (c) A20, A24, and A60 are shown with their chemical shifts. Resonances in the boxes overlap. It can be seen that these three alanines have nearly degenerated chemical shifts.

DISCUSSION

The advantage of using 3D HCCH-COSY/TOCSY over 2D COSY/TOCSY not only facilitates manual assignments but also computational analysis. This can be illustrated with an example of how 3D HCCH-COSY/TOCSY solves a chemical shift degeneracy problem whereas 2D COSY/TOCSY does not.

In Figure 15, two amino acid residue whose H_α and H_β have very close resonance frequencies are illustrated. In the traditional 2D COSY/TOCSY approach, two cross peaks can be merged into a single spin system as long as they share a common resonance frequency, and there is a constraint to prove these two cross peaks belong to a same spin system. In the above example, cross peak (4.073, 2.049) and (4.073, 2.175) belong to one spin system, while (4.062, 2.057) and (4.062, 2.943) belong to another spin system. The problem is 4.073 and 4.062 as well as 2.049 and 2.057 are too close to be distinguished computationally from 2D data alone. As a consequence, all four cross peaks (4.073, 2.049), (4.073, 2.175), (4.062, 2.057), and (4.062, 2.943) are incorrectly merged into a single large spin system, which is apparently incorrect because this large spin system contains three H_β and as many as four H_γ. In other words, from 2D NMR,

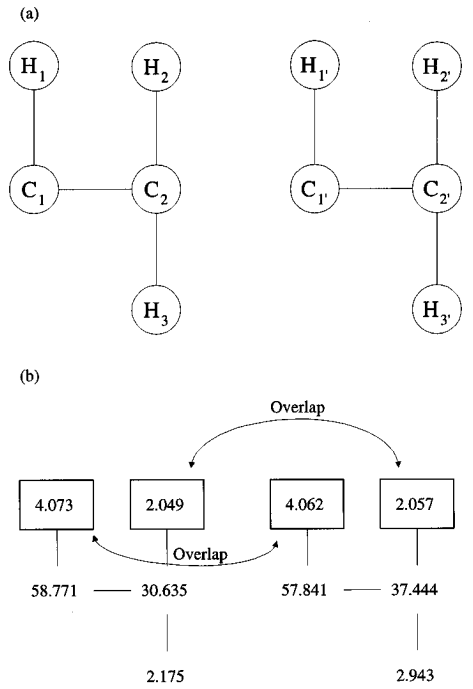

Figure 15. (a) Two fragments from two different molecules are shown. (b) The chemical shifts of the protons and carbons are displayed. Resonances in boxes are those having significantly overlapped chemical shifts.

Table 4. Summary of Overlap Resolution[a]

	3D	2D
H_1 overlaps with $H_{1'}$ H_2 overlaps with $H_{2'}$	resolved by checking C_1	unable to resolve
H_2 overlaps with $H_{2'}$ H_3 overlaps with $H_{3'}$	resolved by checking C_2	unable to resolve
H_1 overlaps with $H_{1'}$ H_2 overlaps with $H_{2'}$ C_1 overlaps with $C_{1'}$	unable to resolve	unable to resolve
H_1 overlaps with $H_{1'}$ H_2 overlaps with $H_{2'}$ H_3 overlaps with $H_{3'}$	unable to resolve	unable to resolve

[a] See Figure 15 for notation.

cross peaks (4.073, 2.049) and (4.062, 2.943) are put into the same spin system since they have one frequency in common, 4.073 vs 4.062. The presence of TOCSY peaks (2.049, 2.175) and/or (2.057, 2.943) does not resolve this problem. In contrast, if 3D NMR cross peaks are available, as described in the previous section, the partitioning algorithm will examine whether 4.073 and 4.062 bond to a same carbon. If not, these two cross peaks are put into different spin systems, and thus such a degeneracy problem is solved. In case that the carbon bonded to 4.073 overlaps with the carbon bonded to 4.062 (i.e., see Figure 15, if 58.771 and 57.841 cannot be distinguished), even 3D NMR algorithm cannot solve such triple degeneracy situations. Table 4 summarizes the limitations of the present algorithm to handle overlap ambiguities. It should be noted that Table 4 just lists the theoretical limits of the algorithm; in practice, certain overlaps can be resolved by using the ranking parameter mentioned in the previous section. Sometimes severe overlaps can also be resolved by using the protein's structure information.

In general, two factors effect the efficiency of our algorithm. They are chemical shift degeneracy and missing peaks. Degenerate chemical shift values usually generate large patterns which correspond to two or more spins systems coupled together. In addition, missing cross peaks can miss spin patterns altogether.

The test of this algorithm on both simulated data and experimental data show that if there are no missing peaks, the algorithm correctly produces all the desired spin systems that can be extracted from 3D data. Nevertheless, in the case where critical cross peaks are missing, correct spin systems may not be extracted. One can relax some merging conditions, described in Figure 9, but less stringent searching conditions may risk getting incorrect results.

Another feature of our algorithm is that the number of input experiments is flexible. To obtain a complete amino acid spin system which includes all the spins' resonance frequencies and their connectivity relationship, COSY type experiments, which record three bond scalar coupling, and TOCSY type experiments, which records long range relay coupling, are required. A sole COSY experiment, can still provide much information about resonance frequencies and connectivity between spins. The lack of long range relay coupling, however, makes the complete extraction of certain amino acids, such as threonine, impossible. A sole TOCSY type experiment, on the other hand, can provide enough information concerning all the spins' resonance frequencies but fails to provide complete connectivity.

Although this algorithm was designed for 3D HCCH-COSY/TOCSY NMR spectra, the ideas can be extended to other 3D NMR experiments. The basic concept behind this algorithm is to take advantage of the third dimension as an additional constraint so as to reduce the ambiguities causing by heavy overlap.

Under certain manual assignment situations, side chain spin systems are investigated after the backbone spins are successfully assigned. Therefore the backbone H_α, C_α frequencies can be taken as starting points for side chain assignment. In the design of this CPA algorithm, however, the traditional protein resonance assignment strategy was adopted, i.e., spin system identification is accomplished prior to sequential assignment. This implies that sequential information of amino acid residues is not incorporated into the CPA algorithm. Future improvement of this algorithm includes adding an option to supply H_α, C_α frequencies from earlier backbone assignments so that a more efficient searching can be achieved due to a resulting smaller search space. Further, an integrated computer assisted environment for protein resonance assignment using 3D heteronuclear NMR is developed. This environment includes complete identification of a protein backbone and side chain resonances; amino acid spin system pattern recognition, and establishment of sequential connectivity.

CONCLUSION

A new algorithm is proposed to automatically extract amino acid spin systems from three dimensional COSY and TOCSY type experiments. This algorithm is based on a 2D constrained partitioning algorithm,[18,19] whose main feature is that all the merging steps are accomplished by imposing various constraints. Another distinct feature is that by supplying both COSY and TOCSY type experiments not only the resonance frequencies of all the spin systems are available but also their connectivity relationships are extracted. This

makes the design of an amino acid pattern recognition procedure easier.

The extracted amino acid spin systems can be used for subsequent sequential assignment procedures. A number of sequential assignment strategies[5,7,16,20,34] can be applied to these spin systems. For example, through the use of triple resonance NMR experiments, it is straightforward to obtain the resonance assignments of a proteins' backbone.[22] The algorithm described in this paper provides strategy to obtain a protein's side chain resonance assignments. By properly incorporating the backbone and side-chain information, a virtually automatic procedure for protein resonance assignments could be achieved.

Those who are interested in the algorithms, please contact B.C.S.

ACKNOWLEDGMENT

The authors thanks Stéphane Gagne for providing the NMR data of the protein data of chicken skeletal triponin-c. The authors also wish to thank Brian Sykes and the Protein Engineering Network of Centers of Excellence (PENCE) for the hospitality to one of us (K.B.L.) during February of 1994. This work is supported by NSERC operating and collaborating grants.

REFERENCES AND NOTES

(1) Wider, G.; Lee, K.; Wüthrich, K. Sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra: Glucagon bound to perdeuterated dodecylphosphocoline micelles. *J. Mol. Biol.* **1982**, *155*, 367−388.

(2) Wüthrich, K. Sequential individual resonance assignments in the proton-NMR spectra of polypeptides and proteins. *Biopolymers* **1982**, *22*, 131−138.

(3) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; Wiley, New York, NY, 1986.

(4) Wüthrich, K.; Wider, G.; Wagner, G.; Braun, W. Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *J. Mol. Biol.* **1982**, *155*, 311−319.

(5) Cieslar, C.; Clore, A. M.; Gronenborn, A. M. Computer-aided sequential assignment of protein ^{1}H NMR spectra. *J. Magn. Reson.* **1988**, *80*, 119−127.

(6) Kleywegt, G. J.; Lamerichs, R. M. J. N.; Boelens, R.; Kaptein, R. Toward automatic assignment of protein ^{1}H NMR spectra. *J. Magn. Reson.* **1989**, *85*, 186−197.

(7) Kleywegt, G. J.; Vuister, G. W.; Padilla, A.; Knegtel, R. M. A.; Boelens, R.; Kaptein, R. Computer-assisted Assignment of Homonuclear 3D NMR Spectra of Proteins. Application to Pike Parvalbumin III. *J. Magn. Reson. B* **1993**, *102*, 166−176.

(8) Kleywegt, G. J.; Boelens, R.; Cox, M.; Llinás, M.; Kaptein, R. Computer-assisted assignment of 2D ^{1}H NMR spectra of proteins: Basic algorithms and application to phoratoxin B. *J. Biomol. NMR* **1991**, *1*, 23−47.

(9) Van de Ven, F. J. M. PROSPECT, a program for automated interpretation of 2D NMR spectra of Proteins. *J. Magn. Reson.* **1990**, *86*, 633−644.

(10) Yu, C.; Hwang, J.-F.; Chen, T.-B.; Soo, V.-W. RUBIDIUM, a program for computer-aided assignment of two-dimensional NMR spectra of polypeptides. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 183−187.

(11) Castasti, P.; Carrara, E.; Nicolini, C. Pepto: an expert system for automatic peak assignment of two-dimensional nuclear magnetic resonance spectra of proteins. *J. Comput. Chem.* **1990**, *11*, 805−818.

(12) Eads, C. D.; Kuntz, I. D. Programs for computer-assisted sequential assignment of proteins. *J. Magn. Reson.* **1989**, *82*, 467−482.

(13) Billeter, M.; Basus, V. J.; Kuntz, I. D. A program for semi-automatic sequential resonance assignments in protein ^{1}H nuclear magnetic resonance spectra. *J. Magn. Reson.* **1988**, *76*, 400−415.

(14) Oschkinat, H.; Holak, T. A.; Cieslar, C. Assignment of protein NMR spectra in the light of homonuclear 3D spectroscopy: An automatable procedure based on 3D TOCSY-TOCSY and 3D TOCSY-NOESY. *Biopolymers* **1991**, *31*, 699−712.

(15) Meadows, R. P.; Olejniczak, E. T.; Fesik, S. W. A computer-based protocol for semiautomated assignments and 3D structure determination of proteins. *J. Biomol. NMR* **1994**, *4*, 79−96.

(16) Bernstein, R.; Cieslar, C.; Ross, A.; Oschkinat, H.; Freund, J.; Holak, T. A. Computer-assisted assignment of multidimensional NMR spectra of proteins: Application to 3D NOESY-HMQC and TOCSY-HMQC spectra. *J. Biomol. NMR* **1993**, *3*, 245−251.

(17) Zimmerman, D.; Kulikowski, C.; Wang, L.; Lyons, B.; Montelione, G. T. Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence. *J. Biomol. NMR* **1994**, *4*, 241−256.

(18) Xu, J.; Sanctuary, B. C. CPA: Constrained Partitioning Algorithm for initial assignment of protein ^{1}H resonances from MQF-COSY. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 490−500.

(19) Xu, J.; Sanctuary, B. C.; Gray, B. N. Automated extraction of spin coupling topologies from 2D NMR correlation spectra for protein ^{1}H resonance assignment. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 475−489.

(20) Xu, J.; Straus, S. K.; Sanctuary, B. C.; Trimble, L. Use of fuzzy mathematics for complete automated assignment of peptide ^{1}H 2D NMR spectra. *J. Magn. Reson. B* **1994**, *103*, 53−58.

(21) Xu, J.; Weber, P. L.; Borer, P. N.; Computer-assisted assignment of peptides with nonstandard amino acids. *J. Biomol. NMR* **1995**, *5*, 183−192.

(22) Bax, A.; Grzesiek, S. Methodological advances in protein NMR. *Acc. Chem. Res.* **1993**, *26*, 131−138.

(23) Bax, A.; Clore, G. M.; Driscoll, P. C.; Gronenborn, A. M.; Ikura, M.; Kay, L. E. Practical aspect of proton-carbon-carbon-proton three dimensional correlation spectroscopy of ^{13}C-labeled proteins. *J. Magn. Reson.* **1990**, *87*, 620−627.

(24) Kay, L. E.; Ikura, M.; Bax, A. Proton-proton correlation via carbon-carbon couplings: a three-dimensional NMR approach for the assignment of aliphatic resonances in proteins labeled with carbon-13. *J. Am. Chem. Soc.* **1990**, *112*, 888−889.

(25) Clore, G. M.; Bax, A.; Driscoll, P. C.; Wingfield, P. T.; Gronenborn, A. M. Assignment of the side-chain ^{1}H and ^{13}C resonances of interleukin-1B using double and triple-resonance heteronuclear three-dimensional NMR spectroscopy. *Biochemistry* **1990**, *29*, 8172−8184.

(26) Bax, A.; Clore, G. M.; Gronenborn, A. M. ^{1}H−^{1}H correlation via isotropic mixing of ^{13}C magnetization, a new three-dimensional approach for assigning ^{1}H and ^{13}C spectra of ^{13}C-enriched proteins. *J. Magn. Reson.* **1990**, *88*, 425−431.

(27) Lyons, B. A.; Tashiro, M.; Cedergren, L.; Nilsson, B.; Montelione, G. T. An improved strategy for determining resonance assignments for isotopically enriched proteins and its application to an engineered domain of staphylococcal protein A. *Biochemistry* **1993**, *32*, 7839−7845.

(28) Montelione, G. T.; Lyons, B. A.; Emerson, S. D.; Tashiro, M. An efficient triple resonance experiment using carbon-13 isotropic mixing for determining sequence-specific resonance assignments of isotopically-enriched proteins. *J. Am. Chem. Soc.* **1992**, *114*, 10974−10975.

(29) Lyons, B. A.; Montelione, G. T. An HCCNH triple-resonance experiment using carbon-13 isotropic mixing for correlating backbone amide and side-chain aliphatic resonances in isotopically enriched proteins. *J. Magn. Reson. B* **1993**, *101*, 206−209.

(30) Clore, G. M.; Gronenborn, A. M. Application of three- and four-dimensional heteronuclear NMR spectroscopy to protein structure determination. *Prog. NMR Spectrosc.* **1991**, *23*, 43−92.

(31) Gagné, S. M.; Tsuda, S.; Li, M. X.; Chandra, M.; Smillie, L. B.; Sykes, B. D. Quantification of the calcium-induced secondary structural changes in the regulatory domain of troponin-C. *Protein Sci.* **1994**, *3*, 1961−1974.

(32) Garrett, D. S.; Powers, R.; Gronenborn, A. M.; Clore, G. M. A common sense approach to peak picking in two-, three-, and four-dimensional spectra using automatic computer analysis of contour diagrams. *J. Magn. Reson.* **1991**, *95*, 214−220.

(33) Wishart, D. S.; Bigam, C. G.; Holm, A.; Hodges, R. S.; Sykes, B. D. ^{1}H, ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *J. Biomol. NMR* **1995**, *5*, 67−81.

(34) Van de Ven, F. J. M.; Lycksell, P.-O.; Van Kammen, A.; Hilbers, C. W. Computer-aided assignment of the ^{1}H-NMR spectrum of the viral-protein-genome-linked polypeptide from cowpea mosaic virus. *Eur. J. Biochem.* **1990**, *109*, 583−591.

CI950103E