Automated Resonance Assignment of Proteins Using Heteronuclear 3D NMR. 2. Side Chain and Sequence-Specific Assignment

Kuo-Bin Li[†] and B. C. Sanctuary*

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

Received April 26, 1996[⊗]

A sequential assignment protocol for proteins was developed using heteronuclear 3D NMR. The protocol consists of an amino acid type recognition algorithm and a primary sequence mapping algorithm. The former measures the similarity between each detected spin pattern and 20 standard amino acid coupling patterns. Both chemical shift and topologically likeness are considered. The mapping algorithm uses the amino acid type information to direct detected polypeptides to proper position onto protein primary sequence. The assignment protocol can be applied to spin systems generated by many different approaches. We designed a few computer programs to derive a protein's backbone and side chain spin systems using heteronuclear 3D NMR. The results was then input to the sequential assignment protocol. All of the algorithms were tested on NMR data of a 90-residue N-domain of chicken skeletal troponin-C.

INTRODUCTION

Resonance assignment is a tedious work in protein structure determination from NMR. To develop a computer-assisted resonance assignment package, several steps have to be accomplished: (1) Individual amino acid residues' spin coupling systems must be extracted. (2) Sequential connectivities between these spin systems must be established based on available interresidue correlations. (3) Spin system identification, i.e., which amino acid each extracted spin system actually is, must be conducted. (4) Sequence-specific mapping between spin systems and a protein's primary sequence must be created. In a previous paper,[1] we presented a computer algorithm to extract the protein backbone spin systems. This paper reports a complete resonance assignment protocol covering the above four steps using heteronuclear 3D NMR. Initially an algorithm was developed to merge data from the protein backbone and aliphatic side chain spin systems. Secondly, a spin system pattern recognition algorithm[2] was revised to automatically determine all the possible amino acids each spin system may correspond to. Finally, a mapping algorithm maps spin systems to their proper positions on the protein primary sequence. The sequence-specific assignment protocol and the implementation of the algorithms is described in this paper. Application of all the proposed computer algorithms to a 90-residue protein is reported. The heteronuclear 3D NMR experiments involved in the application include 3D HNCO, HNCA, HCACO, HN(CO)CA, ^{15}N TOCSY-HMQC, HCCH-COSY, and HCCH-TOCSY.

TOWARD THE SEQUENTIAL ASSIGNMENT

As mentioned in our previous paper,[1] spin patterns of the individual amino acid residues and the sequential connectivities between these patterns can be derived from heteronuclear 3D NMR. The remaining problem of the protein resonance assignment is to match the derived polypeptides onto the known protein primary sequence. This task can be done manually based on human expertise. For example, a spectrscopist may notice that one of the spin systems in a polypeptide might be a leucine. Moreover, another spin system three residues away from the leucine may be identified as a glycine. Provided that the leucine-X-X-glycine pattern occurs only once in the primary sequence, it is straightforward to match the target polypeptide to the correct primary sequence.

To automate this "polypeptide to primary sequence" mapping, it is necessary to have sufficient information about each of the spin-coupling patterns, i.e., one must know the possible amino acids each spin pattern could be. Suppose a polypeptide is composed of five spin systems, $S1-S2-S3-S4-S5$. Spin system $S1$ is identified to be one of the following amino acids: leucine, isoleucine, and valine. Similarly, $S2$ can be one of serine, phenylalanine, ..., etc.; see Figure 1. Having known which amino acid each spin pattern might be, it is possible to construct a set of possible primary sequence combinations. In Figure 1 these combinations include Leu-Phe-Arg-Gly-Glu, Leu-Ser-Arg-Gly-Glu, Leu-Asp-Arg-Gly-Glu, ..., etc. If the polypeptide is long enough and the number of possible amino acids each spin systems might be assigned to is not too large, a unique mapping between polypeptide and primary sequence can be achieved. This is shown Figure 1, where only Leu-Asp-Arg-Gly-Glu has a matching position, say residue 14 to residue 18, on the protein's primary sequence, while all the other combinations fail to match. Thus it is reasonable to assign polypeptides $S1-S2-S3-S4-S5$ to residue 14−15−16−17−18. In the case that a unique mapping is not possible, a ranking parameter can be calculated on the basis of the similarities between each spin system of the polypeptide and its possible amino acid identities.

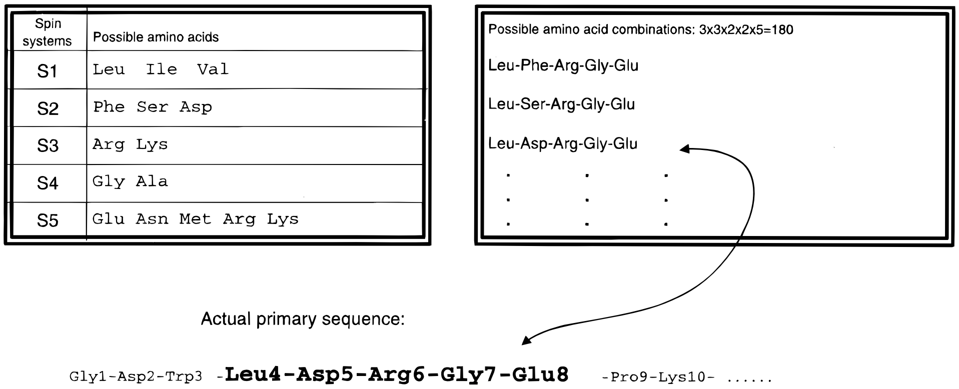
An amino acid pattern recognition algorithm (AAPR)[2] was designed to achieve the goal of mapping individual spin pattern to possible amino acids residues. AAPR gives all possible amino acids each of the spin patterns might be assigned to. Every possible assignment has an associated

* To whom correspondence should be addressed.

[†] Current address: Rowland Institute for Science, 100 Edwin H. Land Boulevard, Cambridge, MA 02142.

[⊗] Abstract published in *Advance ACS Abstracts,* March 15, 1997.

S0095-2338(96)00372-1 CCC: $14.00



Spin systems	Possible amino acids
S1	Leu Ile Val
S2	Phe Ser Asp
S3	Arg Lys
S4	Gly Ala
S5	Glu Asn Met Arg Lys

Possible amino acid combinations: 3x3x2x2x5=180
Leu-Phe-Arg-Gly-Glu
Leu-Ser-Arg-Gly-Glu
Leu-Asp-Arg-Gly-Glu
Actual primary sequence:
Gly1-Asp2-Trp3 -Leu4-Asp5-Arg6-Gly7-Glu8 -Pro9-Lys10- ......

Figure 1. Schematic representation of the mapping of a polypeptide $P1-P2-P3-P4-P5$ to Leu4-Asp5-Arg6-Gly7-Glu8. Residue $P1$ could be assigned to one of Leu, Ile, and Val. $P2$ can be either Phe, Ser, or Asp. There are 180 possible combinations of amino acid sequences for this polypeptide. In this example, the sequence Leu-Asp-Arg-Gly-Glu is the correct mapping on the actual primary sequence.

similarity value measuring the likeness between the amino acid and the spin pattern. In general, it is not easy for a computer algorithm to determine amino acid types for deduced spin patterns based only on backbone frequencies. Although there are published chemical shift database[3,4] by using which one can classify backbone spin systems, the accuracy of amino acid type recognition will be higher if the side chain information of each spin pattern is also available. The more details available of a spin pattern the more accurate the spin pattern recognition. For this reason, an algorithm ASPA[5] (aliphatic side chain partitioning algorithm) was designed to retrieve protein aliphatic side chain resonances from heteronuclear 3D NMR. Combining the protein backbone with this side chain information, an amino acid pattern recognition procedure can provide sufficient information about each spin pattern, thereby making it possible to automate the mapping between polypeptides and primary sequence.

In summary, the DBPA was developed to retrieve a protein's backbone resonances and establish partial sequential connectivities in the forms of dipeptides. PGA is then responsible for merging retrieved dipeptides to polypeptides. ASPA was designed to extact a protein's aliphatic side chain information. Having the information of backbone and side chain spin systems, AAPR gives knowledge about the amino acid types of each spin pattern. PBSMA (protein backbone side chain merging algorithm) then is required to merge both backbone and side chain frequencies. The final step involves an algorithm called PMA (polypeptide mapping algorithm) which is able to perform the mapping task where polypeptides are actually mapped to the protein's primary sequence. Figure 2 shows the relationships between these algorithms.

Integration of Backbone and Aliphatic Side Chains. Many 3D NMR experiments have been proposed for protein side chain resonance assignment, such as 3D HCCH-COSY,[6−8] HCCH-TOCSY,[9] HCC(CO) NH-TOCSY,[10,11] and HCCNH-TOCSY.[10,12] These experiments resolve the crowded aliphatic side chain proton regions in traditional 2D DQF-COSY and TOCSY by introducing another dimension. Therefore a crowded and overlapped 2D spectrum can be split into a series of less overlapped 2D planes in such a 3D NMR experiment. For example, the $^{1}H-^{1}H$ planes in 3D HCCH-COSY experiment resemble a 2D $^{1}H-^{1}H$ COSY spectrum except that these planes are edited by the chemical shifts of the ^{13}C nuclei bonded to ^{1}H resonance observed in F1 dimension of 3D HCCH spectrum. The algorithm ASPA[5] was proposed to automatically extract amino acid spin systems from three-dimensional HCCH-COSY and TOCSY experiments. ASPA was designed based on the concepts of a 2D constrained partitioning algorithm (CPA)[13,14] whose main feature is that all the mergings of cross peaks are accomplished by imposing various constraints which reduce the complexity caused by spectral overlap. ASPA produces aliphatic side chain spin systems as a series of graphs represented as adjacency lists[15] which can be processed by a subsequent graph pattern recognition algorithm to complete amino acid identification. Details about the procedures of recognition will be discussed in next section.

Side chain spin coupling systems are usually investigated after the backbone spins are successfully assigned provided that the $^{15}N/^{13}C$ labeled protein samples are available. Hence triple resonance 3D NMR data can be acquired. The backbone αH and C_{α} frequencies can then be taken into consideration in creating side chain spin systems. For example, algorithm DBPA produces backbone spin systems; the αH and C_{α} chemical shifts of these spin systems can be taken as starting points for side chain resonance assignment using ASPA. Thus a more efficient searching can be accomplished due to a resulting smaller search space.

The side chain resonance frequencies can also be revealed by experiments recording long-range couplings between protons, such as 2D TOCSY and 3D ^{15}N TOCSY-HMQC. In principle a sole 2D TOCSY or 3D ^{15}N TOCSY-HMQC spectrum has sufficient information to assign a protein's entire side chain and backbone spins. In practice, however, not all spin systems can be identified in a TOCSY experiment, especially in the case of α-helix-based proteins which have small $^{3}J_{NH-\alpha H}$ coupling constants.

Despite the fact that a sole TOCSY experiment sometimes fails to provide sufficient information for long-spin systems, it is still useful to examine these TOCSY experiments as they have simpler cross peak patterns compared with DQF-COSY. NCPA (nitrogen constraint partitioning algorithm) was proposed to extract amino acid spin-coupling systems from 2D TOCSY or 3D ^{15}N TOCSY-HMQC experiment. NCPA is complimentary to ASPA as they both provide side chain information but using different approaches (see Figure 3).

The actual procedures to merge backbone and side chain spin systems are described in the pseudocodes given in Chart 1. To merge a backbone and a side chain spin systems, PBSMA requires that they share several common frequencies. Suppose a backbone amino acid contains five frequen-

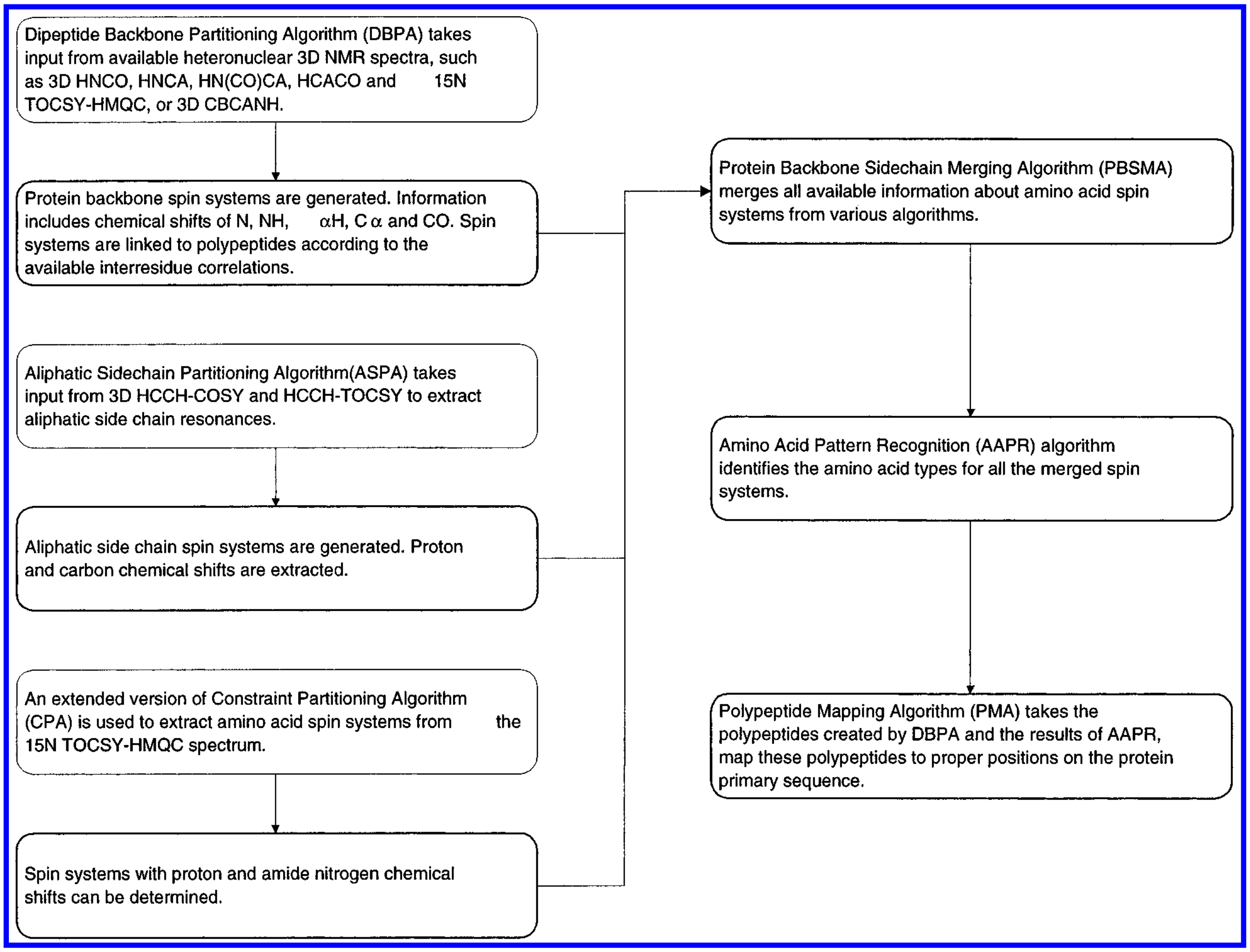

Dipeptide Backbone Partitioning Algorithm (DBPA) takes input from available heteronuclear 3D NMR spectra, such as 3D HNCO, HNCA, HN(CO)CA, HCACO and 15N TOCSY-HMQC, or 3D CBCANH.
Protein backbone spin systems are generated. Information includes chemical shifts of N, NH, αH, C α and CO. Spin systems are linked to polypeptides according to the available interresidue correlations.
Aliphatic Sidechain Partitioning Algorithm(ASPA) takes input from 3D HCCH-COSY and HCCH-TOCSY to extract aliphatic side chain resonances.
Aliphatic side chain spin systems are generated. Proton and carbon chemical shifts are extracted.
An extended version of Constraint Partitioning Algorithm (CPA) is used to extract amino acid spin systems from the 15N TOCSY-HMQC spectrum.
Spin systems with proton and amide nitrogen chemical shifts can be determined.
Protein Backbone Sidechain Merging Algorithm (PBSMA) merges all available information about amino acid spin systems from various algorithms.
Amino Acid Pattern Recognition (AAPR) algorithm identifies the amino acid types for all the merged spin systems.
Polypeptide Mapping Algorithm (PMA) takes the polypeptides created by DBPA and the results of AAPR, map these polypeptides to proper positions on the protein primary sequence.

Figure 2. A flow diagram of our sequential assignment protocol using heteronuclear 3D NMR.

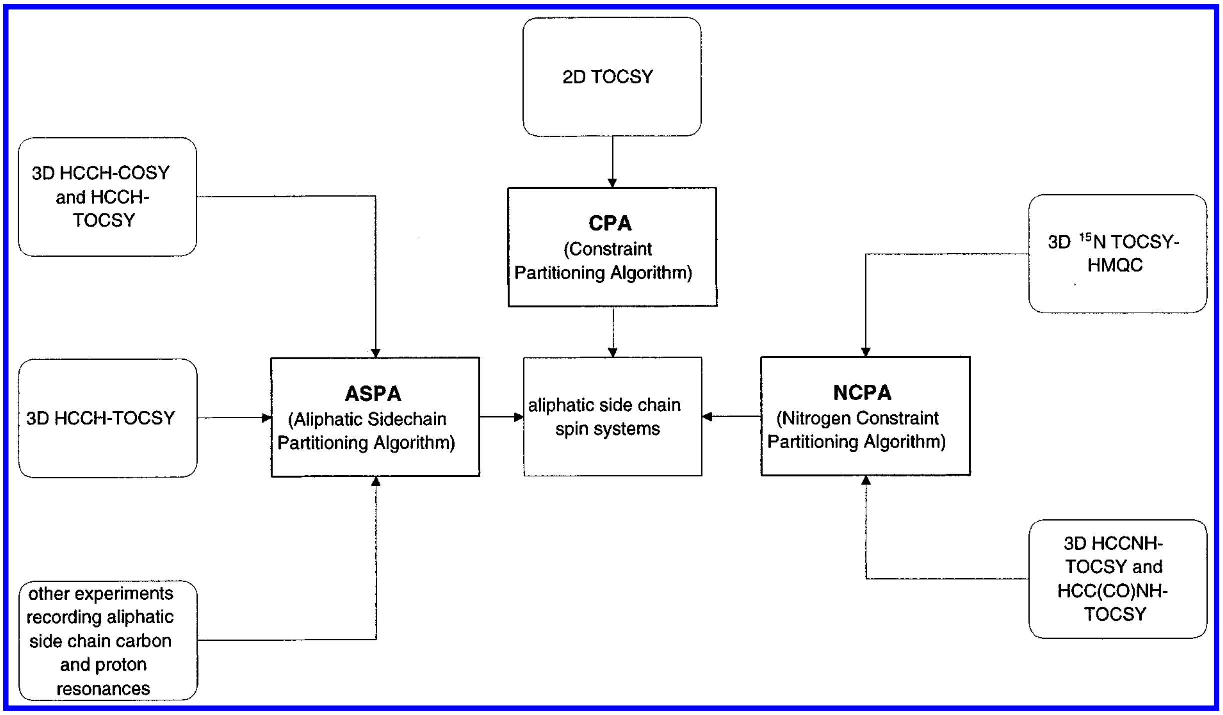

2D TOCSY
3D HCCH-COSY and HCCH-TOCSY
CPA (Constraint Partitioning Algorithm)
3D 15N TOCSY-HMQC
3D HCCH-TOCSY
ASPA (Aliphatic Sidechain Partitioning Algorithm)
aliphatic side chain spin systems
NCPA (Nitrogen Constraint Partitioning Algorithm)
3D HCCNH-TOCSY and HCC(CO)NH-TOCSY
other experiments recording aliphatic side chain carbon and proton resonances

Figure 3. There are many approaches to obtain the protein's side chain resonances. In this example, three algorithms were designed to extract side chain spin systems from 2D and 3D NMR spectra.

cies (NH, N, αH, C_α, CO), and a side chain spin system is composed of four spins (NH, αH, βH_1, βH_2). Depending on the NMR experiments used to construct these spin systems, some resonances may be present in both backbone and side chain spin systems. In the above example, NH and αH are the two overlapped resonances. The more overlapped resonances found, the more reliable the merge. In some cases, another experimental data set provides additional information which can be used as extra constraints to confirm the merge of a backbone and a side chain spin system. A

Chart 1

```
void  MergeBackboneSidechain(BackboneSpinsystem_type,...,
                             SidechainSpinsystem_type,...)
{
   //Input: 1. a set of backbone spin systems $B_1, B_2, B_3,...$
   //       2. a set of side chain spin systems $S_1, S_2, S_3,...$
   //       3. if available, another set of side chain spin
   //          systems $T_1, T_2,...$
   //Examples: $B_i$ were derived from algorithm **UPA**, $B_i$ contains
   //                     (N,NH,$\alpha$H,C$\alpha$,CO).
   //          $S_j$ were derived from algorithm **NCPA**, $S_j$ contains
   //                     (N,NH,$\alpha$H,$\beta$H,...).
   //          $T_k$ were derived from algorithm **CCPA**, $T_k$ contains
   //                     ($\alpha$H,$\beta$H,$\gamma$H,C$\alpha$,C$\beta$,...).
   //
   //Output: a set of amino acid spin systems $A_1, A_2...A_i$ composed of
   //          backbone and side chain information.

   for each of the backbone spin systems $B_i$ {
     for each of the side chain spin system $S_j$ {
       compare $B_i$ and $S_j$;
         if $B_i$ and $S_j$ share several
              common resonances, e.g., $\alpha$H,NH,N  {
              if another set of side chain spin
              systems $T_k$ are available {
                if((one or more resonances in $B_i$ can be found in $T_k$) &&
                   (one or more resonances in $S_j$ can be found in $T_k$)) {
                       $A_l = B_j + S_j + T_k$;
                }
              } else
                 $A_l = B_i + S_j$;
         }
     }
   }

}
```

3D HCCH-COSY/TOCSY data set provides aliphatic side chain resonances including αH, Cα, βH, Cβ, ..., etc.; these spins can be treated as additional constraints for merging backbone and side chain resonances. In other words, to merge to (NH, N, αH, Cα, CO) backbone with a (NH, αH, βH_1, βH_2) side chain, one can check the spin systems output from 3D HCCH-COSY/TOCSY to seek evidences such as a spin system (αH, Cα, βH_1, Cβ, βH_2, ...) where two frequencies (αH and Cα) can be found in the backbone candidate while two others (αH and βH_1) can be found in the side chain candidate.

Once the backbone and side chain spin systems are properly merged, it is possible to perform the amino acid identification process, i.e., to recognize these spin systems according to their spin-coupling patterns and chemical shifts. The aim of spin pattern recognition is to obtain all possible amino acids that a spin system might be assigned to. A spin pattern recognition algorithm was developed by Xu[2] to handle the amino acid identification task. This algorithm makes use of fuzzy mathematics to recognize each amino acid's distinct pattern. Most spin systems recognition algorithms (e.g., the one by Kleywegt[16]) utilize chemical shift information exclusively. However, Xu's algorithm is able to recognize amino acids' spin topologies based on the fact that each topology has different connectivities between its components. Along with the chemical shift information, the above graph theory and fuzzy mathematics based pattern recognition algorithm provides more accurate results in terms of determining possible amino acids that a spin system corresponds to.

As described above, the backbone and side chain spin systems can be extracted from various NMR experiments. Backbone spin patterns may come from 3D HNCO, HNCA, HCACO, HN(CO)CA, and ^{15}N TOCSY-HMQC. They may also come from 3D CBCANH experiment. Similarly, side chain spin systems may be derived from 3D HCCH type experiments as well as from HCC(CO)NH-TOCSY. Even 2D DQF-COSY and TOCSY NMR spectra provide valuable information. The spin pattern candidates therefore may consist of various information. Those spin patterns from 2D COSY/TOCSY may contain proton frequencies whereas those spin patterns derived from 3D HCCH COSY/TOCSY may be composed of carbon and proton frequencies. Moreover, the spin patterns may differ from each other in terms of connectivity relationships. Patterns from TOCSY type experiments may not contain detailed connectivity information. For example, TOCSY type experiments may not be able to distinguish spin pattern 4.53 (αH), 2.25 (βH), 1.93 (βH), from pattern 4.53 (αH), 1.93 (βH), 2.25 (γH) as it is not generally easy to determine whether a specific peak is arising from ^3J or long-range couplings. Figure 4 provides a summary of the three different kinds of spin patterns described above, and several experimentally observed spin patterns are given as examples.

Figure 5 illustrates how an experimentally observed amino acid pattern is mapped to various amino acid residue. The chemical shift values of standard amino acid patterns may contain protons only; proton and nitrogen; proton and carbon; or proton, carbon, and nitrogen, depending on available NMR experiments. The proton database of standard 20 amino acid was adopted from Gross,[4] a nitrogen database was adopted from Choy,[17] and the carbon chemical shift database was adopted from Wishart.[3] Note that in Figure 5 there might be more than one mapping from a candidate spin pattern to

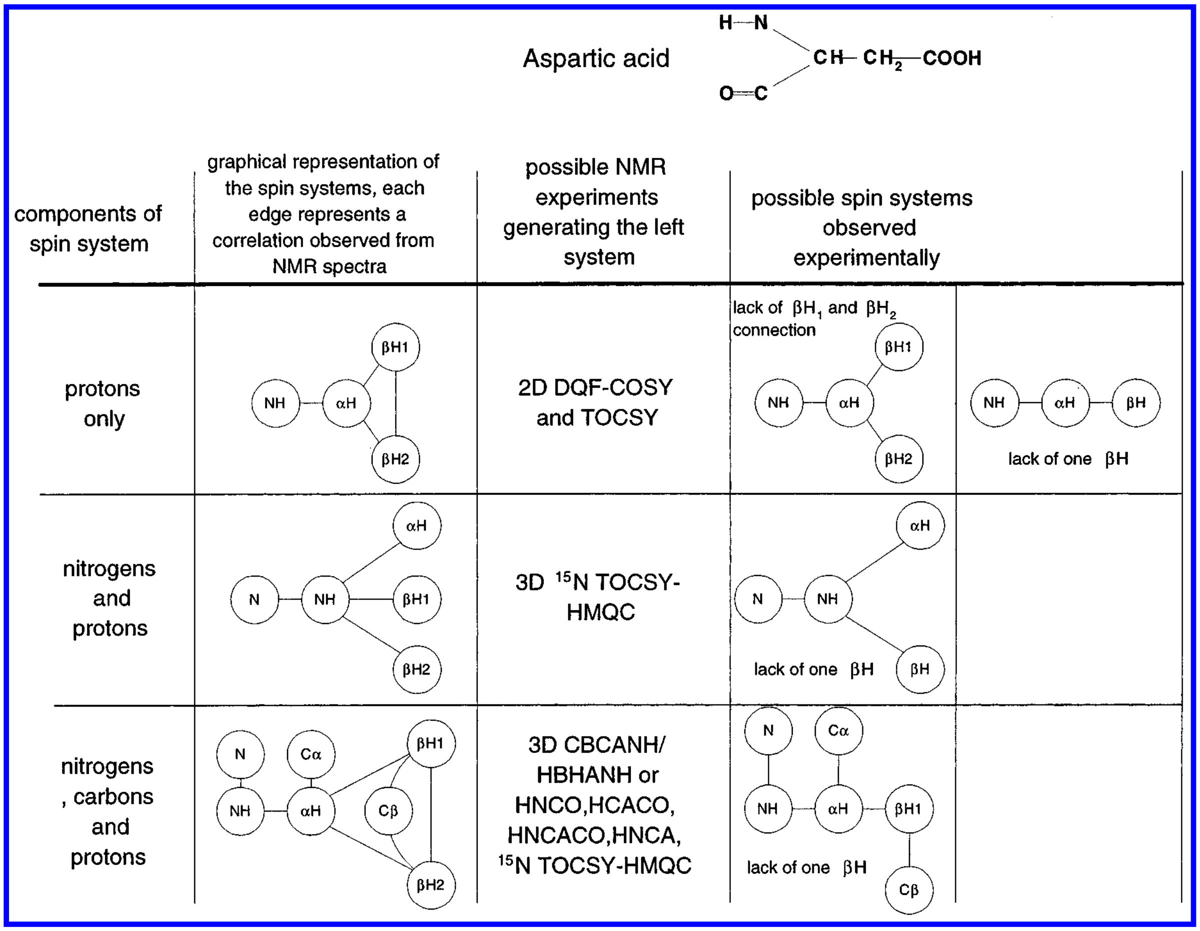

Aspartic acid
H—N
CH—CH2—COOH
O=C
components of spin system
graphical representation of the spin systems, each edge represents a correlation observed from NMR spectra
possible NMR experiments generating the left system
possible spin systems observed experimentally
protons only
NH
αH
βH1
βH2
2D DQF-COSY and TOCSY
lack of βH1 and βH2 connection
βH
lack of one βH
nitrogens and protons
N
3D 15N TOCSY-HMQC
nitrogens, carbons and protons
Cα
Cβ
3D CBCANH/ HBHANH or HNCO,HCACO, HNCACO,HNCA, 15N TOCSY-HMQC

Figure 4. Aspartic acid is illustrated to show spin-coupling patterns composed of various nuclei. Possible experiments generating these patterns are also listed.

a standard spin pattern. For each of the mappings there is an associated value which represents the similarity between candidate and standard spin patterns. Details about the similarity values were published by Xu.[2] After performing the pattern recognition on all of the extracted amino acid spin patterns, a "spin pattern to residue" table can be created where one can locate all possible amino acids that each spin pattern can be assigned to. Figure 6 shows a small segment of such a table; note that amino acids with low similarity values were eliminated to shorten the table.

We now summarize the tasks having been described to this point. A number of amino acid spin systems with backbone and all the available side chain information are derived. These spin systems' identities have been examined; that is, a table such as the one shown in Figure 6 gives all possible amino acids that a spin system can be assigned to. The sequential assignment problems was partially solved since heteronuclear 3D NMR provides interresidue correlations from which polypeptides can be built. The rest of the resonance assignment task is to map these polypeptides to their actual positions within the primary sequence with the help of the "spin patterns to amino acid residue table". This task can be achieved manually since spectroscopists usually have additional information at hand to guide them through the mappings of the polypeptides. Here a general purpose sequential assignment protocol was proposed to automate the mapping. This protocol aims at giving an additional tool to help spectroscopists handle tedious assignment tasks. The first step of our sequential assignment protocol involves a conversion of the "spin systems to amino acids" table to an "amino acid residues to spin systems" table. Figure 7 illustrates such a conversion. Once done, the remaining work is to check each of the polypeptides against the "amino acid residue to spin systems" table. If a polypeptide can be located in the table, the corresponding assignment is immediately determined. In Figure 8 a nine-residues polypeptide is used to explain the assignment procedure. Algorithm PMA was designed to carry out the mapping. The pseudocode is listed in Chart 2.

In the pseudocode there is a function *check* which is called recursively to compare each element of a polypeptide with a residue of the primary sequence. If the function check reaches the end of the polypeptide, a proper mapping is located as shown in Figure 8.

RESULTS

A sequential assignment protocol is describe in the previous section. The protocol involves two major steps. In the first step amino acid spin systems are extracted from NMR spectra and then linked to form polypeptides. In the second step, all amino acid spin systems are identified according to their spin topological patterns. Thereafter polypeptides can be mapped to the primary sequence. Each of these tasks can be achieved through various strategies, both manually and automatically, using computer algorithms. To illustrate the effectiveness of our sequential assignment protocol, several computer algorithms were implemented to accomplish all of the stages. The details of these algorithms have already been described while this section presents the application of these computer programs to a real case.

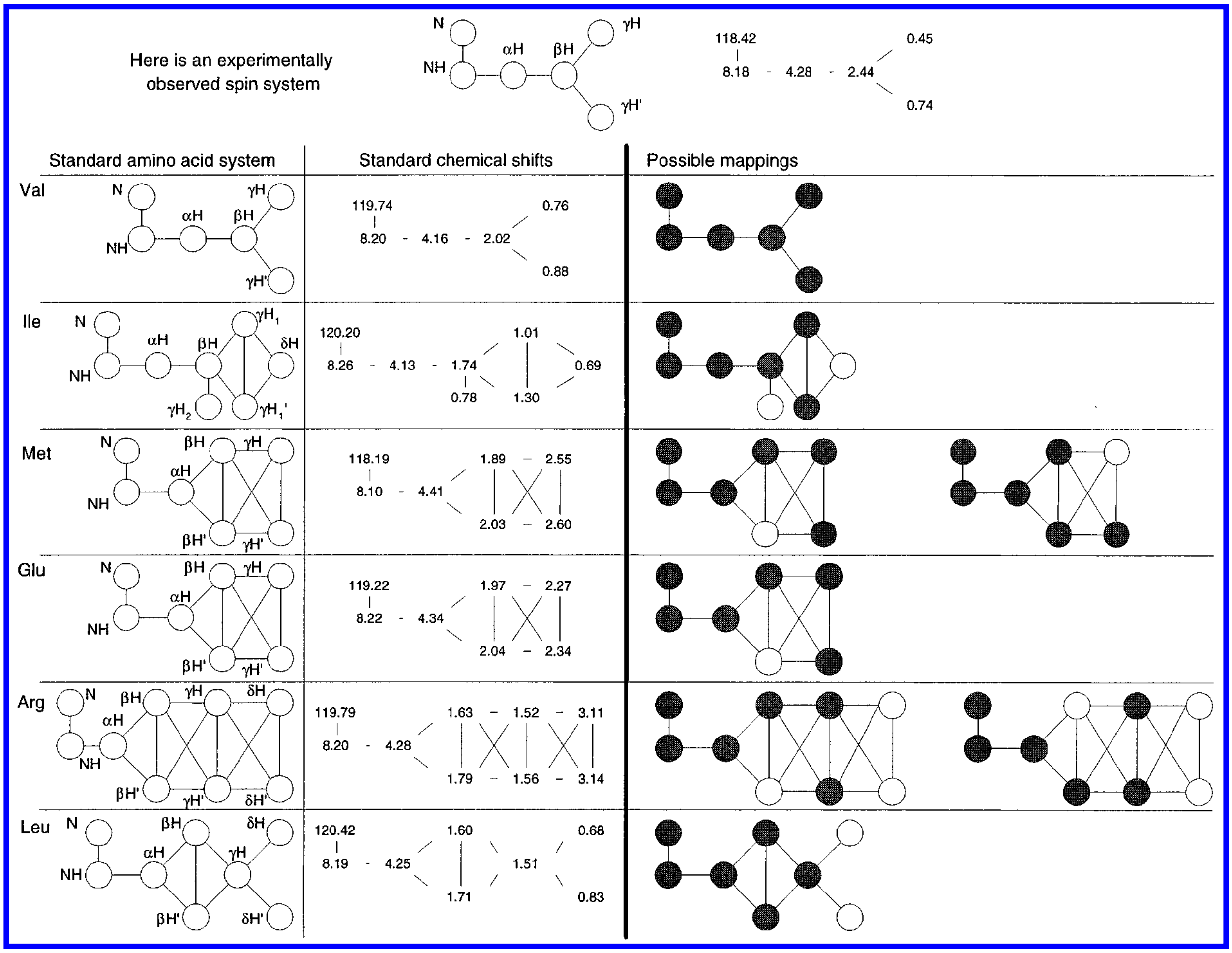


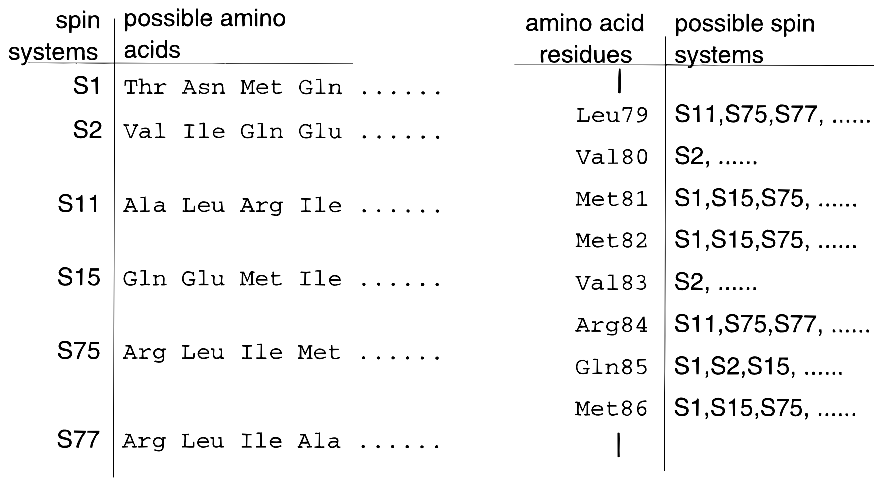
Figure 5. Schematic representation of mappings between an observed spin pattern and its possible amino acids: Val, Ile, Met, Glu, Arg, and Leu. Note that there could be more than one mapping for the same amino acid, such as the case of Met and Arg.

15	Gln 0.877	[✓ Glu 0.854]	Met 0.738	Ile 0.627	Arg 0.615	Lys 0.615	Leu 0.595					
11	[✓ Ala 0.898]	Leu 0.819	Arg 0.794	Ile 0.781	Val 0.725	Met 0.620	Glu 0.616	Gln 0.610	Thr 0.545	Ser 0.535	Phe 0.488	Asn 0.415
75	[✓ Arg 0.820]	Leu 0.799	Ile 0.702	Met 0.702	Val 0.665	Gln 0.636	Glu 0.584					
77	Arg 0.931	Leu 0.908	Ile 0.868	[✓ Ala 0.841]	Val 0.779	Glu 0.747	Gln 0.738	Met 0.692	Phe 0.654	Thr 0.619	Asp 0.566	Gly 0.565
81	Arg 0.856	Lys 0.856	[✓ Phe 0.726]	Ser 0.617	Glu 0.591	Thr 0.591	Leu 0.585	Met 0.585	Gln 0.581	Val 0.566		
82	Ile 0.650	Arg 0.636	Lys 0.636	[✓ Leu 0.589]	Gln 0.586	Met 0.570	Glu 0.570	Pro 0.441				
88	Thr 0.872	Asn 0.774	Met 0.771	Gln 0.730	Phe 0.715	Val 0.685	Glu 0.671	Asp 0.666	Arg 0.636	Leu 0.629	Ile 0.621	[✓ Ser 0.595]
66	Val 0.819	Ile 0.763	Gln 0.705	[✓ Glu 0.671]	Leu 0.666	Arg 0.662	Ala 0.657	Met 0.632	Phe 0.547	Gly 0.468	Ser 0.436	Thr 0.420
32	[✓ Glu 0.813]	Gln 0.804	Val 0.728	Ile 0.728	Thr 0.702	Ser 0.697	Lys 0.673	Arg 0.673	Gly 0.669	Leu 0.631		

Figure 6. A "spin systems to amino acids" table. Spin system no. 15 can be either Gln, Glu, Met, Ile, Arg, Lys, or Leu. This table was generated by the amino acid pattern recognition algorithm. The number below each amino acid denotes the similarity between that amino acid and the spin pattern on the very left. A higher similarity value indicates a closer match. The values range from 0 to 1.

Sample protein is a calcium-loaded regulatory N-domain of chicken skeletal troponin-C (NTnC) residue 1−90. Uniformly enriched ^{15}N and ^{13}C NTnC were also prepared. Available heteronuclear 3D NMR experiments include 3D HNCA,[18] 3D HNCO,[18] 3D HNCOCA,[18] 3D HCACO,[19] 3D ^{15}N TOCSY-HMQC, and NOESY.[20] Peak lists of the above NMR experiments were given to the authors by the University of Alberta.[21] Peaks were picked using the CAPP peak-picking program[22] and then processed by a filter program to remove some of the false peaks.[21]

The amino acid spin systems can be derived from three separated algorithms each using a different set of NMR



spin systems	possible amino acids
S1	Thr Asn Met Gln ......
S2	Val Ile Gln Glu ......
S11	Ala Leu Arg Ile ......
S15	Gln Glu Met Ile ......
S75	Arg Leu Ile Met ......
S77	Arg Leu Ile Ala ......

amino acid residues	possible spin systems
|	
Leu79	S11,S75,S77, ...
Val80	S2, ...
Met81	S1,S15,S75, ...
Met82	S1,S15,S75, ...
Val83	S2, ...
Arg84	S11,S75,S77, ...
Gln85	S1,S2,S15, ...
Met86	S1,S15,S75, ...
|	

Figure 7. The conversion between a "spin systems to amino acids" table to "amino acid residues to spin systems" table.

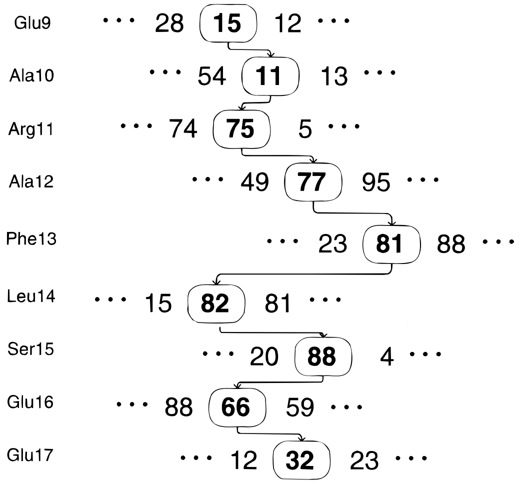


Figure 8. Illustration of a possible assignment of polypeptide 15−11−75−77−81−82−88−66 to Glu9-Ala10-Arg11-Ala12-Phe13-Leu14-Ser15-Glu16-Glu17. The numbers on the right are the spin system numbers.

experiments. Algorithm DBPA involves several triple resonance heteronuclear 3D NMR experiments and is able to deduce the backbone spin systems. In addition, polypeptides can be created since interresidue information can also be observed from some triple resonance NMR experiments. The details of DBPA algorithm are presented in a previous paper.[1] DBPA gave 98 output protein backbone spin systems, 58 of which can be verified manually against separately done manual assignments.[21] Using the interresidue information embedded in the NMR cross peaks, 161 dipeptides can be created based on the 98 spin systems. Further, a total of 5432 polypeptides with length from 3 to 26 were built from these 161 dipeptides.

Besides triple resonance NMR experiments, spin systems can also be determined by TOCSY type experiment exclusively as long as there are sufficient long-range couplings observed. Algorithm NCPA was used to extract spin systems composed of amide nitrogen and protons from ^{15}N TOCSY-HMQC. Application of NCPA to 90-residue NTnC gives a total of 83 spin systems of which 73 can be verified against structural assignment result.[21] The tolerance value for comparing proton chemical shifts was chosen to be 0.02 ppm, and 0.20 ppm was chosen for nitrogen.

Side chain resonances occur in the crowded aliphatic regions of NMR spectra. Therefore complete assignment of side chain resonances is a challenging undertaking especially for a large protein. A side chain spin pattern extraction algorithm ASPA[5] was designed for 3D HCCH-COSY/TOCSY NMR spectra. For protein NTnC, 915 HCCH-COSY peaks and 710 HCCH-TOCSY peaks were automatically picked by CAPP.[22] The output of ASPA includes 60 spin systems, among which 55 can be verified against the manual assignment. However there are 395 unpartitioned cross peaks which may arise from the false picked peaks by the automatical peak-picking program. Figure 9 summarizes the spin systems information retrieved so far.

The remaining tasks, that is, the second part of our sequential assignment protocol, requires integration of available spin system information, recognition of amino acid types, and mapping of polypeptides to their anticipated position on the protein primary sequence.

There are three kinds of spin system information available: (1) backbone spin systems containing sequential information from triple resonance NMR, (2) spin systems derived from TOCSY type correlations, and (3) side chain spin systems determined from 3D HCCH type experiments. Algorithm PBSMA combines these data and results in 40 spin systems with detailed side chain correlations and 32 spin systems with TOCSY correlations on the side chain. Figure 10 is the schematic representations of these two types of spin systems and their corresponding building blocks. Once the complete amino acid spin systems, i.e., backbone and side chain, are constructed as shown in Figure 10, they can be identified using algorithm AAPR. Figure 6 shows part of the output of the pattern recognition program. In the final stage, the algorithm PMA is responsible for mapping the 5432 candidate polypeptides to the primary sequence based on information shown in Figure 6. PMA gave a total 2161 mappings. Of these, many are redundant mappings. For example, polypeptide 8−9−49−15−11 (where numbers denote spin system number) was assigned to Gln6-Gln7-Ala8-Glu9-Ala10, while simultaneously the polypeptide 8−9−49−15−11−75 was assigned to Gln6-Gln7-Ala8-Glu9-Ala10-Arg11. It is obvious that the former is a redundant mapping. A set of rules were designed based on systematic relationships to remove such redundancies. In addition, human expertise and intuition can also be applied to reduce the number of mappings. Details about these rules will be described in the Discussion.

Chart 2

```
void MapPolypeptide(primary_sequence, polypeptides,
                                        SpinSystemToAminoAcid_table )
{
  //Input: 1. protein's primary sequence $R_1 - R_2 - R_3 - \ldots - R_m$.
  //       e.g.: Glu9-Ala10-Arg11-Ala12-Phe13-Leu14-Ser15-Gly16-Glu17-...
  //
  //       2. a set of polypeptides: $P_1$, $P_2$, $P_3$, ......
  //       e.g.: $P_1$ =S15-S11-S75-S77-S81-S82-S88-S66-S32
  //             where S stands for spin systems.
  //
  //       3. spin-systems to amino-acids table which maps each spin
  //          system to possible amino acids.
  //       e.g.:
  //       Spin system             Possible amino acids
  //     ____________________________________________________________
  //          S15                  Gln,Glu,Met,Ile,......
  //          S11                  Ala,Leu,Arg,Ile,......
  //          S75                  Arg,Leu,Ile,Met,......
  //          S77                  Arg,Leu,Ile,Ala,......
  //
  //
  //
     Known the protein's primary sequence, it is possible to convert
       the above table to "amino-acid-residue to spin-system" table ;
  //       e.g.:
  //       Residue                 Possible spin system candidate
  //       __________________________________________________________
  //            ≀
  //          Glu9                 ......,S25,S15,S12,......
  //          Ala10                ...,S54,S11,S13,......
  //          Arg11                .....,S74,S75,S5,.......
  //          Ala12                ........,S49,S77,S95,......
  //            ≀
  //

      for each of the polypeptide $P_i = S_{i_1} - S_{i_2} - S_{i_3} - \ldots - S_{i_n}$  {
         for each of the amino acid residue $R_j$ in the primary sequence {
            check(1,$j$);   // to see if $S_{i_1}$ can be found in the candidate
                                  // list of $R_j$ ;
         }
     }
}
void check(integer $p$,integer $q$)
{
      if spin system $S_{i_p}$ can be found in the candidate list
       of residue $R_q$
      {
          if ($p \le n$)
                                            // spin system $S_{i_q}$ is
                                            //  the end of polypeptide $P_i$

          and $(q + (n - p)) \le m$   {
                                            // assure there are enough number
                                            // of residues remaining
                                            // in the primary sequence to be
                                            // mapped to polypeptide $P_i$

                check($p+1$,$q+1$);
                                            // call itself recursively

          } else if ($p == n$) {
                     a mapping is found; // $S_{i_1}$--->$R_j$
                                          // $S_{i_2}$--->$R_{j+1}$
                                          // $S_{i_3}$--->$R_{j+2}$
                                                  ≀          ≀
                                                  ≀          ≀
                                          // $S_{i_n}$--->$R_{j+n}$
                   }
      }
}
```

The final assignment includes mappings of a 14-residue polypeptide to "Gln7 Ala8 Glu9 Ala10 Arg11 Ala12 Phe13 Leu14 Ser15 Glu16 Glu17 Met18 Ile19 Ala20", a seven-residue polypeptide to "Ile19 Ala20 Glu21 Phe22 Lys23 Ala24 Ala25", a seven-residue polypeptide to "Met28 Phe29, Asp30 Ala31 Asp32 Gly33 Gly34", and a seven-residue polypeptide to "Arg84 Gln85 Met86 Lys87 Glu88 Asp89 Ala90". Figure 9 lists the summary of the results.

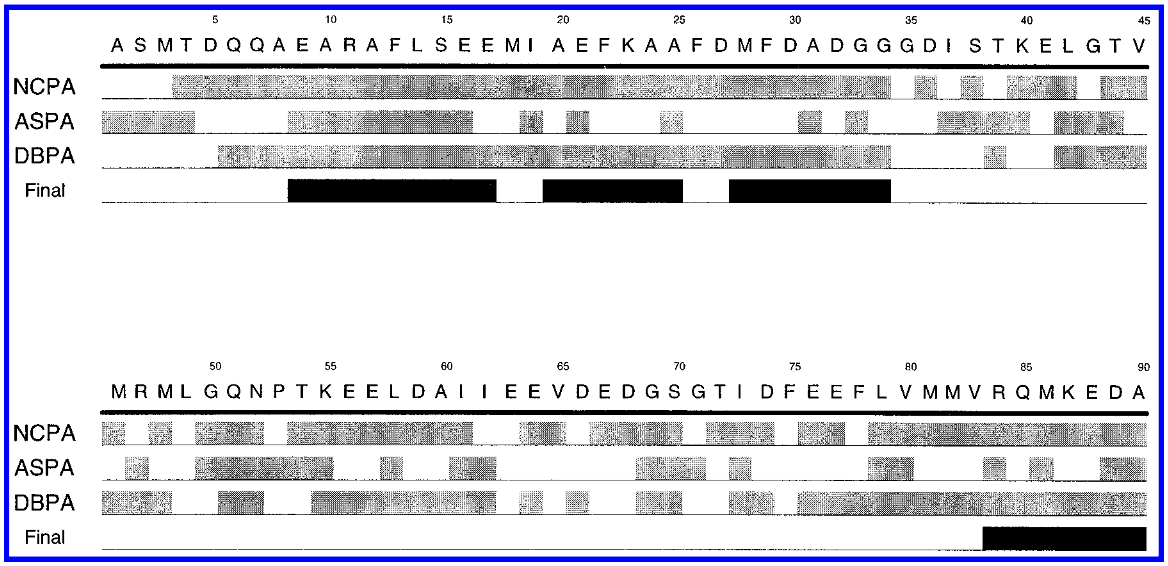


Figure 9. The results of our sequential assignment protocol for a 90-residue protein NTnC. NCPA represents the extracted residues using 3D ^{15}N TOCSY-HMQC and nitrogen constraint partitioning algorithm. ASPA represents the extracted side chain spin systems using 3D HCCH-COSY, HCCH-TOCSY, and aliphatic side chain partitioning algorithm. DBPA represents the extracted backbone spin systems using 3D HNCO, HCACO, HNCO, HN(CO)CA, ^{15}N TOCSY-HMQC, and dipeptide backbone partitioning algorithm. "Final" represents the sequence-specific assigned residues. Lack of sufficiently long backbone polypeptides between residues 35 and 80 prevents automated sequence-specific assignment in that region. However, individual residues' resonance assignments are still obtained.

DISCUSSION

Algorithm PBSMA provides a way to integrate a protein's backbone and side chain data. The detailed information of the backbone and side chain can be determined independently using different NMR data. PBSMA does not limit itself to certain types of experiments. On the contrary, PBSMA accepts a wide variety of spin systems including spin systems composed of protons and spin systems composed of protons and carbons, in addition to spin systems composed of protons, carbons, and nitrogens. As examples to illustrate the effectiveness of PBSMA, two sets of experimental data were used. The first set of NMR data includes 3D HNCO, HNCA, HCACO, HN(CO)CA, and ^{15}N TOCSY-HMQC. These five experiments are able to provide backbone spin systems and partial sequential connectivities. Furthermore, ^{15}N TOCSY-HMQC alone provides another set of spin systems based on long-range scalar couplings between protons. PBSMA merges the backbone and side chain data by overlapping each backbone amino acid spin systems with side chain counterparts. They can be merged if a reasonable overlapping between these two can be verified. The second set of NMR data to test PBSMA includes two more experiments, 3D HCCH-COSY and HCCH-TOCSY. These two NMR experiments give an additional set of amino acid side chain spin systems which in turn are as constraints to increase the accuracy of PBSMA. The more experimental data availabe, the more accurate backbone and side chain merging can be obtained.

The second algorithm discussed in this paper is an amino acid pattern recognition algorithm (AAPR). Originally this pattern recognition algorithm was designed for spin systems containing protons only.[2] A revised version was presented where other atoms can be included in the spin patterns. The availability of heteroatoms (carbon and nitrogen) mainly depends on experimental data. Spin patterns with carbon resonances can be derived provided that an NMR data set which correlates carbon and proton frequencies is available. Here the flexibility of the resonance assignment protocol is evident.

The third and the most important algorithm is PMA (polypeptide mapping algorithm). Its responsibility for mapping all polypeptides to their proper positions on the protein's primary sequence. In principle, a unique mapping can be determined provided that the polypeptide is sufficiently long. For example, a 10-residue polypeptide could end up being mapped uniquely to residues 18−27 on the primary sequence. However, in practice, this kind of uniqueness is not likely since each component residue of a polypeptide could be assigned to many amino acids (although only one of them can be correct). This usually leads to multiple mappings. A set of rules were designed to manipulate this kind of multiple mappings. The first rule is the simplest and depends heavily upon the human's experience. Recall in conducting amino acid pattern recognition, each spin patterns is assigned a similarity value with respect to each amino acid. This value is calculated according to a mathematical similarity between the query spin pattern and a standard one. Both topological and chemical shift similarities are considered during the process. The similarity values range from 0 to 1, the higher value indicating a closer match. Having obtained each residue's similarity value, an overall score of each mapping can be given. Suppose a polypeptide $S_1-S_2-S_3-...-S_n$ is to be mapped to the primary sequence between residue R_p and $R_{(p+n-1)}$. The similarity value between S_i and $R_{(p+i-1)}$ is denoted as r_i. The overall score of this mapping is defined as

$$(\prod_{i=1}^{n} r_i)^{1/n}$$

Because all r_i's range between 0 and 1, the overall score also ranges from 0 to 1. A higher score indicates a more likely mapping. The first rule to reduce the number of mappings is to simply set a threshold for the overall scores from all the mappings. Only those mappings with a score higher than this threshold remain. A typical threshold value is between 0.6 and 0.7 and is determined by the quality of all spectra and individual user's experience. This threshold

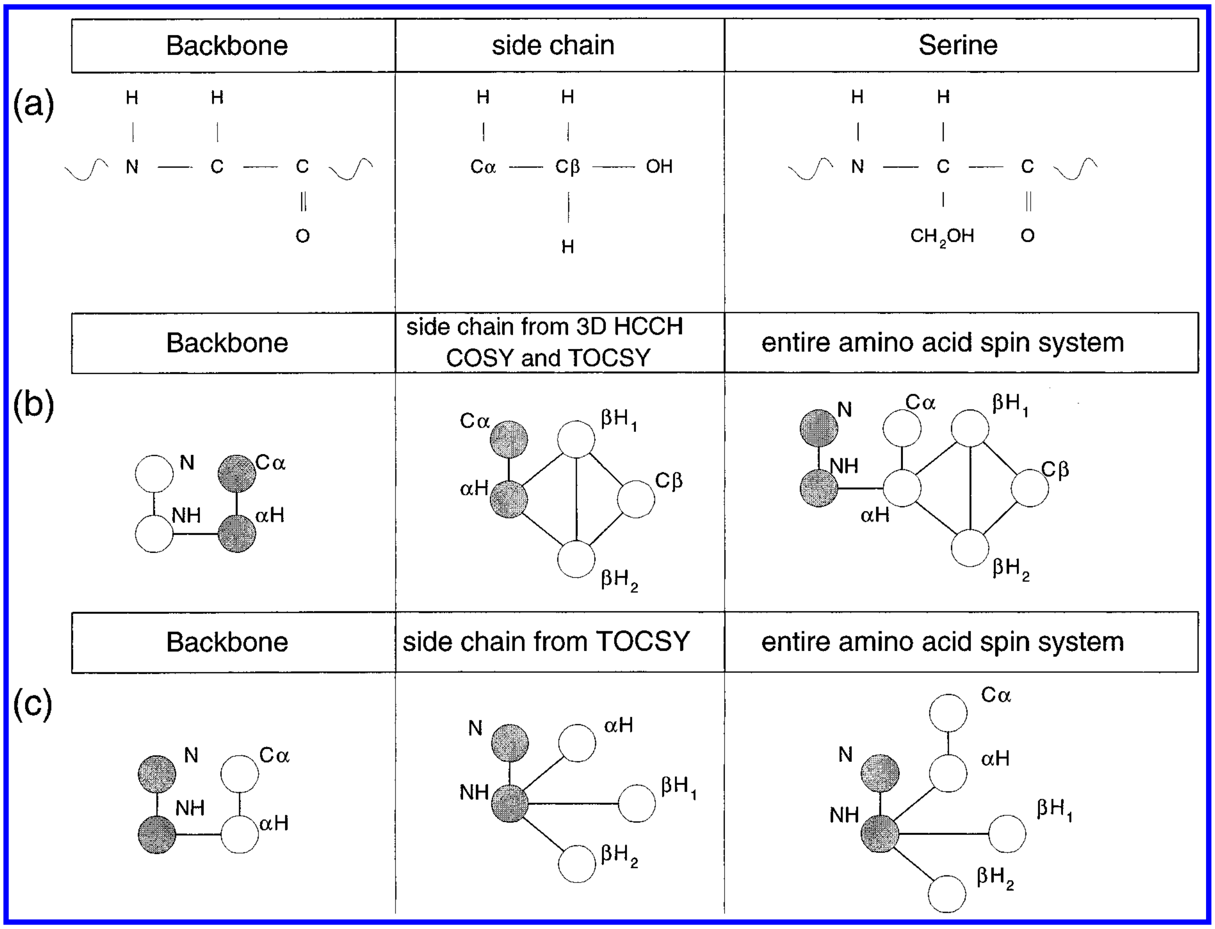


Figure 10. Illustration of the merging of backbone and side chain spin systems. Filled circles represent overlapped resonances. (a) Chemical structure of serine's backbone and side chain. (b) Using 3D HCCH-COSY and HCCH-TOCSY, it is possible to obtain side chain spin system's carbon frequencies. Thus the merged spin system contains proton and carbon frequencies. (c) Using 3D ^{15}N TOCSY-HMQC, the side chain spin pattern contains a nitrogen frequency.

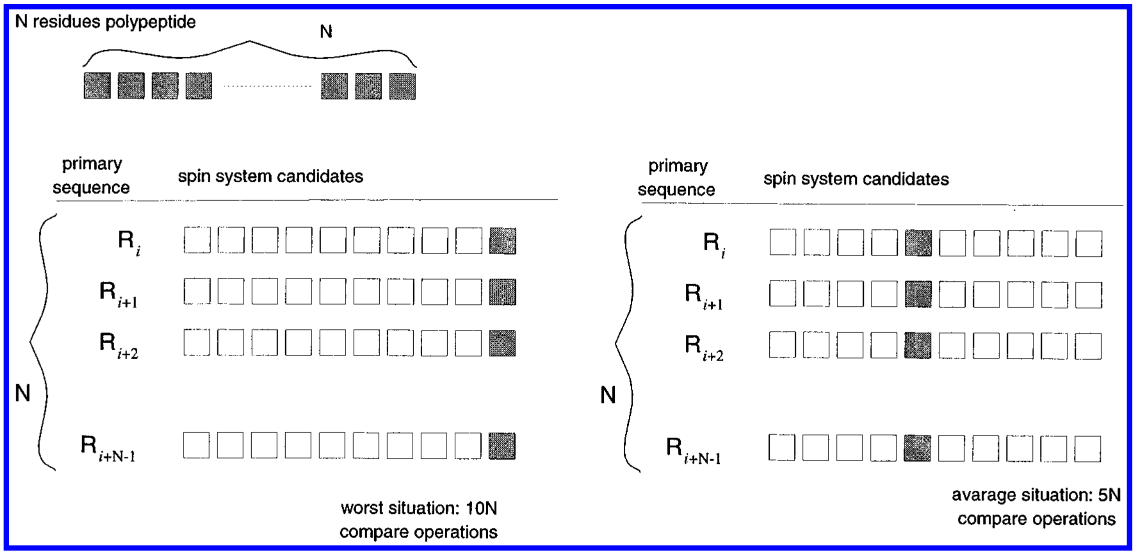


Figure 11. Performance analysis of polypeptide mapping algorithm. An *N*-residue polypeptide is to be assigned. In the worst situation, the correct spin systems all occur at the end of the spin system candidate lists. 10*N* comparisons are expected in this case. In the average situation, the correct spin systems occur in the middle of the spin system candidate list; thus a total of 5*N* comparisons can be expected.

of mapping score can eliminate a large number of multiple mappings.

The second rule deals with redundant mappings. Suppose polypeptide P_i can be mapped to S_i, and another polypeptide P_j can be mapped to S_j, where S are segments on the primary sequence. Suppose P_i is a subset of P_j and S_i is a subset of S_j. Mapping P_i−S_i is discarded since this mapping is a subset of mapping P_j−S_j. For example, polypeptide (*S*5−*S*4−*S*91−*S*94−*S*95) is mapped to residue 30−34 while polypeptidie (*S*21−*S*78−*S*5−*S*4−*S*91−*S*94−*S*95) is mapped to residue 28−34. It is obvious that the former is a redundant mapping. In cases that more than one polypeptide can be mapped to residue 28−34, a third rule is used which suggests that the polypeptide with the highest mapping score will be picked. Similarly, if a polypeptide can be mapped to more than one positions, the mapping with highest score will be kept.

By employing these rules, the number of mappings, can be reduced to a reasonable value whereby users are able to manually select the final assignments.

The efficiency of the polypeptide mapping algorithm is a great improvement over its predecessor, the tree search algorithm (TSA).[2] Consider the following example. A polypeptide with N spin systems is to be assigned. In Figure 11, suppose each amino acid residue has 10 possible spin system candidates, only one of them can be assigned to the corresponding residue. In the worst situation the correct mappings occur at the last spin systems of each residue, thus a total of $10N$ comparison operations must be done to assign this N-residue polypeptide. In the average situation, $5N$ comparing operations can be anticipated.

CONCLUSION

The sequential assignment protocol presented in this paper is the first one using amino acid pattern recognition and heteronuclear 3D NMR. Detected spin patterns are compared with 20 standard amino acid patterns to determine their amino acid types. The comparison is 2-fold. First, the similarities of chemical shifts are calculated. Second, the topological consistency between query pattern and standard pattern is checked. Using heteronuclear 3D NMR, the chemical shifts can be nitrogen, carbon, and proton nuclei. DBPA (dipeptide backbone partitioning algorithm), ASPA (aliphatic side chain partitioning algorithm), and NCPA (nitrogen constraint partitioning algorithm) were designed to extract both backbone and side chain spin systems from heteronuclear 3D NMR spectra. PBSMA (protein backbone side chain merging algorithm) was designed to incorporate all the spin system information and prepare spin patterns for amino acid type determination. These “amino acid type determined” spin systems then become input into PMA (polypeptide mapping algorithm) along with the sequential connectivities extracted in DBPA to complete the final assignment.

A complete resonance assignment protocol is presented. It is fully automated and general, i.e., not limited to a particular NMR experiments. However, the automated assignment protocol are not designed to entirely replace manual assignment. Proper human intervention still plays an important role in the computer-assisted protein resonance assignment.

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

ACKNOWLEDGMENT

The authors thank Stéphane Gagne for providing the NMR data of the protein data of chicken skeletal troponin-C. The authors also 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 1994. This work is supported by NSERC operating and collaborating grants.

REFERENCES AND NOTES

(1) Li, K.-B.; Sanctuary, B. C. Automated assignment of proteins using 3D heteronuclear NMR. Part I: Backbone spin systems extraction and creation of polypeptides. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 359−366.

(2) 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.

(3) 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.

(4) Gross, K.-H.; Kalbitzer, H. R. Distribution of chemical shifts in ^{1}H nuclear magnetic resonance spectra of proteins. *J. Magn. Reson.* **1988**, *76*, 87−99.

(5) Li, K.-B.; Sanctuary, B. C. Automated extracting of amino acid spin systems in protein using 3D HCCH-COSY/TOSCY spectroscopy and Constrained Partitioning Algorithm(CPA). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 585−593.

(6) Bax, A.; Clore, G. M.; Driscoll, P. C.; Gronenborn, M. A.; 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.

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

(8) 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.

(9) 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.

(10) 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.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) Gersting, J. L. *Mathematical structures for computer science*; Computer Science Press: New York, 1993.

(16) 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.

(17) Choy, W.-Y.; Sanctuary, B. C. Protein ^{15}N chemical shift data base. Private communication, 1995.

(18) Grzesiek, S.; Bax, A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. *J. Magn. Reson.* **1992**, *96*, 432−440.

(19) Powers, R.; Gronenborn, A. M.; Clore, G.; Bax, A. Three-dimensional triple-resonance NMR of ^{13}C/^{15}N-enriched proteins using constant-time evolution. *J. Magn. Reson.* **1991**, *94*, 209−213.

(20) Marion, D.; Driscoll, P. C.; Kay, L. E.; Wingfield, P.; Bax, A.; Gronenborn, A. M.; Clore, G. M. Overcoming the overlap problem in the assignment of ^{1}H NMR spectra of larger proteins by use of three-dimensional heteronuclear ^{1}H-^{15}N Hartmann-Hahn-multiple quantum coherence and nuclear overhauser-multiple quantum coherence spectroscopy: Application to interleukin 1 β. *Biochemistry* **1989**, *28*, 6150−6156.

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

(22) 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.

CI960372K