Automated Resonance Assignment of Proteins Using Heteronuclear 3D NMR. 1. Backbone Spin Systems Extraction and Creation of Polypeptides

Kuo-Bin Li and B. C. Sanctuary*

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

Received May 15, 1996[⊗]

Heteronuclear 3D NMR has been used to determine protein solution structures for many years. Scalar magnetization transfer through peptide bonds can be observed in triple resonance 3D NMR and thus makes the sequential assignment of a protein backbone more straightforward. In this paper a generic algorithm is proposed to automate protein backbone resonance assignments. The algorithm searches and merges cross peaks among all available NMR spectra. Individual spin systems can be extracted and linked to form polypeptide chains based on observed interresidue correlations. The algorithm is not restricted to a particular type of experiments and is shown to be applicable to two sets of NMR spectra. The first set of NMR data includes five experiments: 3D HNCO, HNCA, HN(CO)CA, HCACO, and ^{15}N TOCSY-HMQC. The second set of data is a 3D CBCANH spectrum. The implemented computer program was tested on the first NMR data set of a 90 residue protein N-domain of chicken skeletal troponin-C.

INTRODUCTION

Numerous approaches[3] have been applied to the automated assignment problem using multidimensional NMR. Vuister *et al*.[4] proposed an assignment strategy for homonuclear 3D NOE-HOHAHA spectrum, and Kleywegt *et al*.[5] implemented and expanded the strategy for homonuclear 3D [J,NOE]- and [NOE-J]-type NMR spectra of proteins. Oschkinat *et al*.[6] presented an automated strategy making use of homonuclear 3D TOCSY-TOCSY and TOCSY-NOESY. Among the attempts using heteronuclear 3D NMR, Zimmerman *et al*.[7] developed an approach for determining the sequential order of amino acid spin systems using 3D HCC(CO)NH-TOCSY and constraint propagation methods. Bernstein *et al*.[8] applied the technique of combinatorial minimization to achieve sequence-specific assignment of protein using 3D ^{15}N-HMQC-TOCSY and ^{15}N-HMQC-NOESY. Two complete protein automated resonance assignment protocols were proposed: one was done by Meadows *et al*.[9] the other by Morelle *et al*.[10] The first makes use of 4D HNCAHA, HN(CO)CAHA, HC(CO)NH-TOCSY, 3D HNCA, and HN(CO)CA, while the second protocol uses a set of 2D triple resonance NMR spectra to assign the protein's backbone resonances. Some of these computer programs, for example, Zimmerman's and Berstein's, automate sequential assignment only. The amino acid spin systems have to be created and identified using other approaches. Meadow's and Morelle's protocols are able to extract amino acid spin systems, but an automated amino acid type recognition routine is lacking. In addition, many of these programs put emphasis on particular types of NMR experiments.

This study reports a computer algorithm that can extract a protein's backbone spin system using 3D heteronuclear NMR. Because many heteronuclear 3D NMR experiments record both intra- and interresidue correlations, the sequential information embedded in the spectra can also be derived at the same time. The algorithms presented in this study are not designed for any specific NMR experiment, which means any general data set can be used. The only information required by the algorithm is sufficient inter- and intraresidue correlations. The algorithm is generic, because these correlations need not come from certain particular experiments. Two sets of 3D NMR experiments are used as examples to demonstrate how the the protein backbone is extracted by the algorithm. The first set of NMR data consists of 3D HNCO, HNCA, HN(CO)CA, HCACO, and ^{15}N TOCSY-HMQC. The second set of NMR data is 3D CBCANH. Experimental data from the first set of NMR experiments were used to test the implemented algorithm. The target protein is a calcium loaded N-domain of chicken skeletal troponin-C (residue 1−90). Along with a sequence-specific resonance assignment protocol presented in a companion paper,[11] it is possible to achieve the goal of developing a nearly fully automated resonance assignment package. This package is able to extract backbone spin systems; create dipeptide links from interresidue correlations observed in 3D heteronuclear NMR; obtain protein side chain spin systems; merge backbone and side chains; identify amino acid types; and, finally, achieve sequence-specific assignment.

IDENTIFICATION OF BACKBONE SPIN PATTERNS

Many heteronuclear 3D NMR experiments[1] have been designed for assigning backbone resonances of ^{15}N/^{13}C isotope enriched proteins. These experiments usually observe correlations between three or more nuclei on a protein's backbone. Both inter- and intraresidue correlations can be recorded therefore making it possible to assign the backbone resonances, along with their sequential connectivities, by applying heteronuclear 3D NMR exclusively.

Before illustrating how to make use of the information provided by 3D NMR experiments, a general description of using computer algorithms to assign NMR cross peaks is discussed here. In general NMR cross peaks from 3D spectra can be represented as (δ_i, δ_j, δ_k) where the three coordinates denote the three chemical shift values. For homonuclear 3D

* To whom correspondence should be addressed.
[⊗] Abstract published in *Advance ACS Abstracts,* October 15, 1996.

S0095-2338(96)00045-5 CCC: $14.00

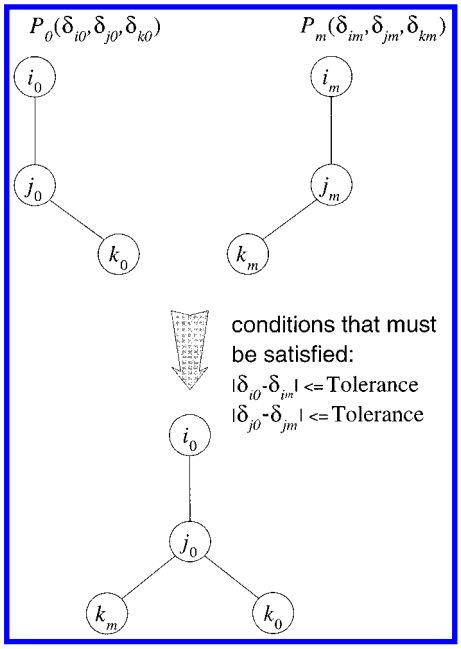


Figure 1. A 3D NMR cross peak $P_0(\delta_{i_0}, \delta_{j_0}, \delta_{k_0})$ can be merged with another peak $P_m(\delta_{i_m}, \delta_{j_m}, \delta_{k_m})$ provided that the two conditions shown are satisfied. The merge results in a spin system with four spins $\{i_0, j_0, k_0, k_m\}$.

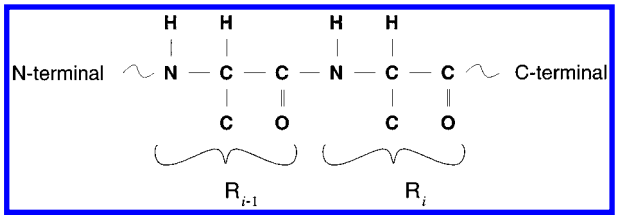


Figure 2. The chemical structure of a dipeptide with only the backbone atoms shown.

NMR all three coordinates represent proton chemical shifts. For heteronuclear 3D NMR, δ_i, δ_j, and δ_k can be proton, carbon, or nitrogen chemical shifts. To make use of the 3D NMR data, computer algorithms usually perform the following steps: for a starting peak $P_0(\delta_{i_0}, \delta_{j_0}, \delta_{k_0})$, a search is conducted on the same spectrum or other spectra to find one or more peak $P_1(\delta_{i_1}, \delta_{j_1}, \delta_{k_1})$, $P_2(\delta_{i_2}, \delta_{j_2}, \delta_{k_2})$, ..., $P_n(\delta_{i_n}, \delta_{j_n}, \delta_{k_n})$ from which two resonances are in common with P_0. For example, P_0 and P_1 may have the same resonances in the first two coordinates. That is, two resonances satisfy the relationships of $|\delta_{i_0} - \delta_{i_1}| \leq$ (a predefined tolerance) and $|\delta_{j_0} - \delta_{j_1}| \leq$ (another predefined tolerance). The next step involves the implementation of a ranking system to distinguish peaks P_1, P_2, ..., P_n in such a way that a peak P_m is picked which is the most likely peak to be in the same spin coupling system with P_0. At this stage the target spin system expands its size from three resonances to four. This operation is shown in Figure 1. The ranking system usually involves searching for evidence in the way of peaks to confirm the merging of P_0 and P_m. In summary, to extract spin coupling systems out of 3D NMR peaks, computer algorithms must have the following features: (1) the algorithms must be able to merge cross peaks, (2) in order to merge two cross peaks, two of the three coordinates should overlap, (3) to verify the merge, other spectral evidence in the form of cross peaks is required.

The application of heteronuclear 3D NMR to protein backbone assignment is now discussed. Figure 2 shows a protein backbone segment.

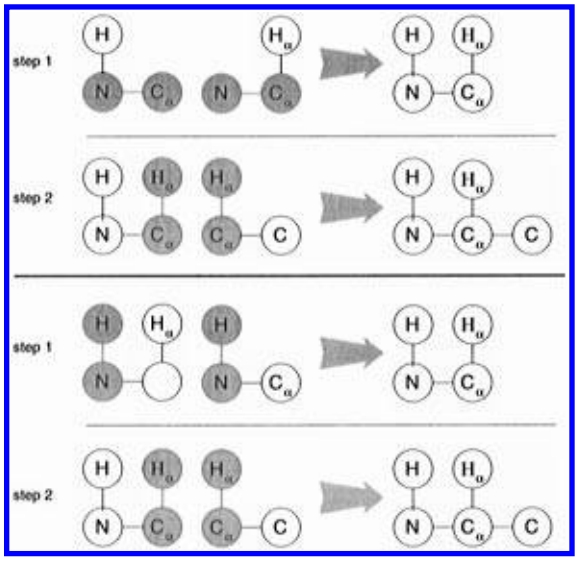


Figure 3. The construction of a backbone spin system is shown. Two possible approaches are listed. In the upper one, an HNCO peak, an HN(CO)CA peak and an HCACO peak are merged to form a spin system. In the lower one, a TOCSY-HMQC peak, an HNCA and an HCACO peaks are merged. The filled circles represent the overlapped resonances discovered by the computer algorithm in order to merge peaks.

A typical triple resonance heteronuclear 3D NMR spectrum observes correlations of three resonances, a proton, a carbon, and a nitrogen. For example, the 3D HNCA[12] experiment gives inter- and intraresidue correlations between NH, N, and C_α. Some experiments can even observe correlations spanning more than three spins such as CBCANH,[13] where inter- and intraresidue C_β, C_α, NH, and N correlations are extracted in one single experiment. A 3D CBCANH peak may have four interpretations: NH−N−C_α(interresidue), NH−N−C_β(interresidue), NH−N−C_α(intraresidue), and NH−N−C_β(intraresidue). C_α resonances of glycine and C_β resonances of all other residues are opposite in phase relative to the other C_α correlations.[13] To resolve the ambiguities between the inter- and intraresidue CBCANH peaks, an extra 3D CBCA(CO)NH[14] experiment may be helpful. Since both inter- and intraresidue correlations are available in heteronuclear 3D NMR, individual amino acid residues and sequential connectivities can be obtained simultaneously. Suppose the general merging algorithm described above is applied, which means there must be at least three correlations available to construct the complete backbone spin system of an amino acid. Here complete backbone spin systems are the ones having their N, NH, αH, C_α, and CO resonances assigned. Figure 3 shows two of the possible combinations from which the backbone spin systems can be constructed. Note that these three correlations may come from three different experiments. However it is also possible that they all come from the same experiment which combines multiple information into one spectrum.

Recall in Figure 2 that the minimum peptide unit having inter- and intraresidue correlations is a dipeptide, i.e., two adjacent amino acid residues. It has been demonstrated that three NMR correlations are required to create an amino acid residue. To create a dipeptide, however, eight instead of six NMR correlations must be observed. The additional two correlations are necessary for establishing the interresidue connectivity. See Figure 4 for the pictorial illustration. In

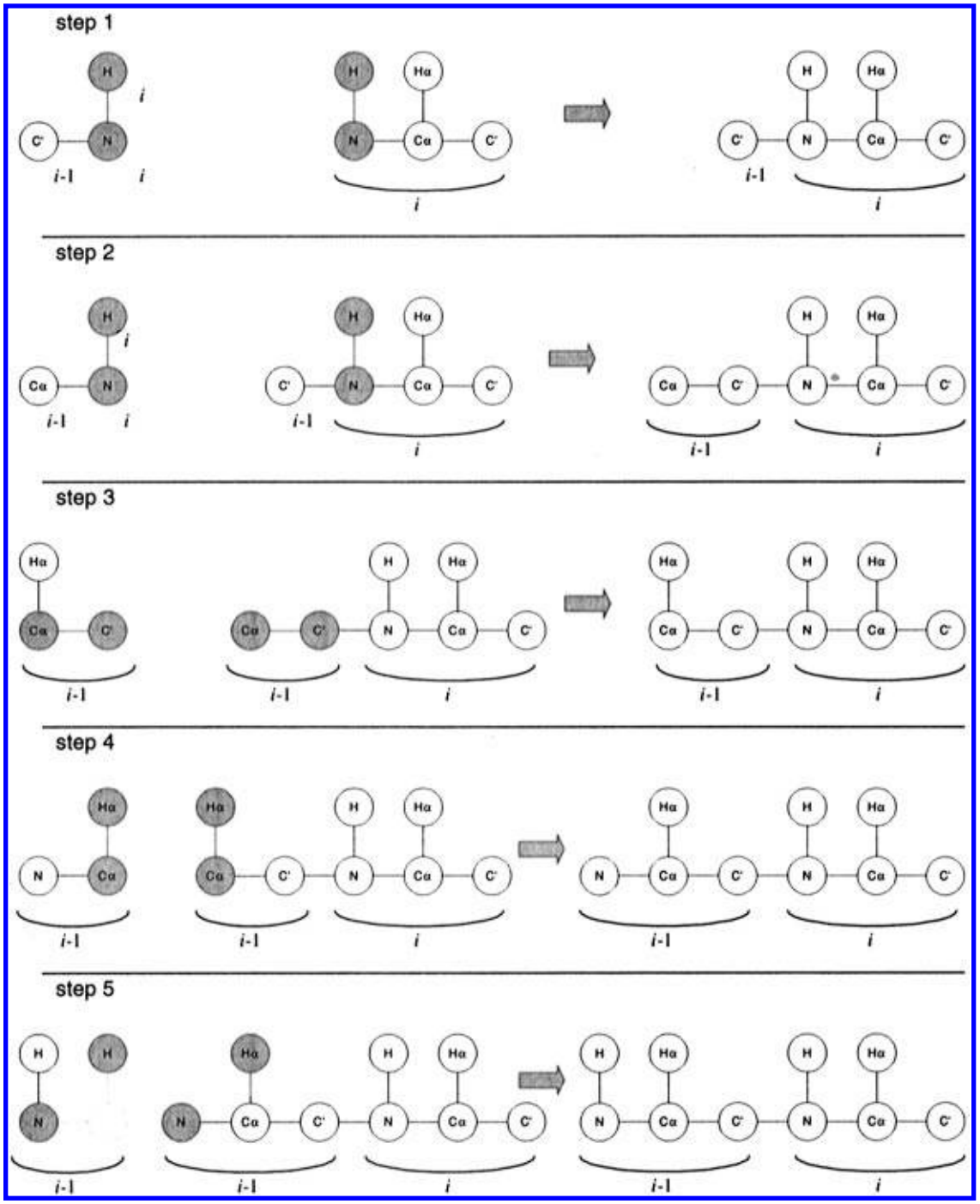


Figure 4. The formation of a dipeptide unit. In step 1, residue (*i*) is a determined spin system. A total of five peaks are required to extend the assignment from residue (*i*) to residue ($i - 1$). Steps 1 and 2 involve the interresidue correlations while steps 3−5 use intraresidue correlations. Note that residue (*i*) needs three correlations to construct itself. Hence a total of eight correlations are required for the construction of a dipeptide unit.

the next section the implementation of these ideas is described.

Description of Backbone Assignment Strategy. In this section examples from two sets of heteronuclear 3D NMR

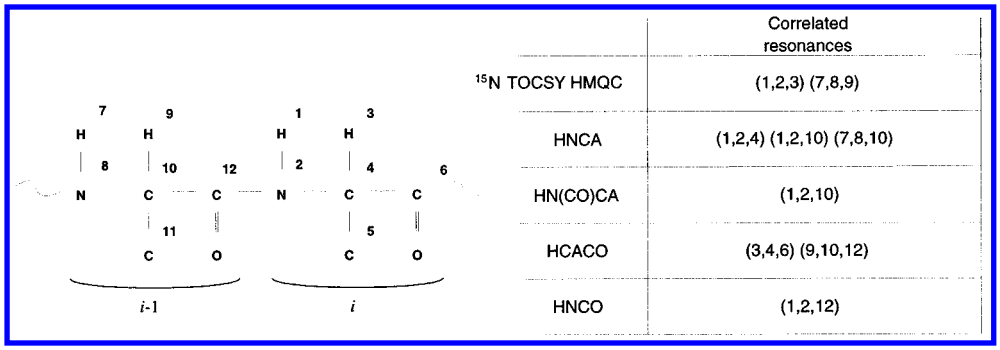


Figure 5. Five triple resonance NMR experiments and the nuclei they correlate.

	N	Cα	C'	NH	αH
Residue i-1	9	7	3	10	8
Residue i	1	5	6	2	4

Steps	Experiment involved	cross peak	Results
1	HNCO	(1,2,3)	Identify three resonances, 1,2 and 3
2	^{15}N TOCSY-HMQC	(1,2,4)	from 1,2, get resonance 4
3	HNCA	(1,2,5)	from 1,2, get resonance 5
4	HCACO	(4,5,6)	from 4,5, get resonance 6
5	HN(CO)CA	(1,2,7)	from 1,2, get resonance 7
6	HCACO	(3,7,8)	from 3,7, get resonance 8
7	^{15}N TOCSY-HMQC	(8,9,10)	from 7,8, get resonance 9 and 10
8	HNCA	(7,9,10)	same as above

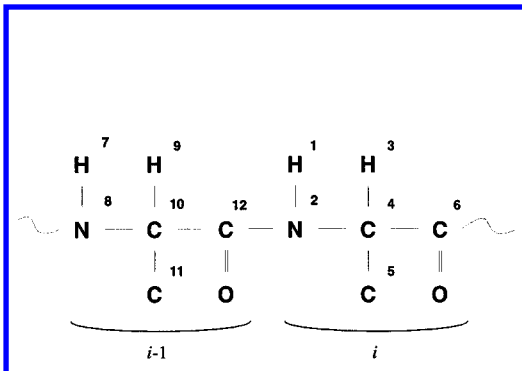
Figure 6. The eight steps are listed for assigning the 10 resonances of a dipeptide. Starting from the 3D HNCO cross peak (1, 2, 3), each subsequent step adds one more resonance to the dipeptide, making a 10 resonance spin system.

spectra are adopted to illustrate the general algorithm discussed in the previous section. Figure 5 shows the five 3D NMR experiments used in the first set of spectra.

The algorithm for assigning protein backbone was designed in such a way to start the searching from any of the input NMR experiments. The advantage of choosing a specific experiment may sometimes be obvious. For example, spectroscopists may notice that a certain experiment is more sensitive, hence it is reasonable to start the assignment procedure from that experiment. However, it is emphasized that the complete assignment of a dipeptide can be achieved through more than one path. Figure 6 describes an eight-step scenario of assigning a dipeptide where cross peaks of 3D HNCO were chosen as the starting experiment. Each of the eight steps involved in the assignment procedure has an associated NMR cross peak. In step 1, the HNCO peak (1,2,3) is selected as the initial spin system. In step 2, the ^{15}N-HMQC-TOCSY peak (1, 2, 4), where the first two frequencies are in common with the previous HNCO peak (1, 2, 3), is added to the spin system. Similarly, by repeating the eight steps, the ten resonance dipeptide (N, NH, αH, C_α, CO)$_{i-1}$ − (N, NH, αH, C_α, CO)$_i$ can be constructed.

In the second example, a single 3D CBCANH experiment was chosen as the input data to illustrate how backbone assignment can be achieved through various approaches. CBCANH has several advantages over the traditional heteronuclear 3D NMR experiments, for example, HNCA, in that CBCANH is able to distinguish inter- and intraresidue peaks in terms of the peak intensities.[13] Moreover, aliphatic C_α and C_β frequencies appear in opposite phases in CBCANH[13] making it possible to separate the C_α from the C_β in aliphatic region. Figure 7 shows a typical dipeptide and its corresponding cross peaks from 3D CBCANH spectrum. Figure 8 shows how the assignment procedure using CBCANH is accomplished. Note that additional spectra may be necessary in order to obtain the frequencies of αH, βH, and CO.

Implementation of the Algorithm. Our algorithm, Dipeptide Backbone Partitioning Algorithm (DBPA), is composed of two parts. In the first part all possible dipeptides are extracted from available spectra. Following this, the individual dipeptides are merged to form polypeptides in the second stage. The algorithm used in the



	Correlated resonances	
CBCANH	(1,2,10)	NH_i, N_i, $C\alpha_{i-1}$
	(1,2,11)	NH_i, N_i, $C\beta_{i-1}$
	(1,2,4)	NH_i, N_i, $C\alpha_i$
	(1,2,5)	NH_i, N_i, $C\beta_i$
	(7,8,10)	NH_{i-1}, N_{i-1}, $C\alpha_{i-1}$
	(7,8,11)	NH_{i-1}, N_{i-1}, $C\beta_{i-1}$

Figure 7. A 3D CBCANH experiment provides three inter- and three intraresidue correlations of a dipeptide.

	NH	N	Cα	Cβ
Residue i-1	7	8	10	11
Residue i	1	2	4	5

Steps	cross peak	Results
1	(1,2,10)	Identify three resonances, 1,2 and 10
2	(1,2,11)	from 1,2, get resonance 11
3	(1,2,4)	from 1,2, get resonance 4
4	(1,2,5)	from 1,2, get resonance 5
5	(7,8,10)	from 10,11, get resonance 7
6	(7,8,11)	from 10,11, get resonance 8

Figure 8. The six correlations provided by the 3D CBCANH experiment can be used to create a dipeptide with eight resonances.

extraction of backbone spin systems and creation of dipeptides is listed in the following pseudo codes:

```
void CreateDipeptide(PeakList_type, ...)
{
  StartingSpectrum=SelectStartingSpectrum(all of the input spectra);
  for each of the peak in StartingSpectrum {

    dipeptide=AddSpinsToDipeptide(the peak);

    for every possible two spin pair (i,j) combination in above dipeptide
    {
         In the entire spectrum database excluding the starting spectrum,
         look for peaks (i',j',k), (i',j,k') and (k,i',j')
         which have two frequencies in common with the
         initial spin pair (i,j);

         if many peaks satisfy the above condition
            BestPeak=RankingProcedure(all of the peaks
                     (i',j',k), (i',j,k') and (k,i',j'));

         dipeptide=AddSpinsToDipeptide(BestPeak);

    }
    if the number of spins in this dipeptide has reached ten
                              // (N,NH,αH,Cα,CO) for two peptides
         keep this dipeptide;
  }
}
```

The pseudo code is self-explanatory except for the ranking procedure which is responsible for choosing the most probable peak to be merged into the existing spin system out of many possible candidate peaks. The pseudo codes for this ranking procedure is outlined in the following:

```
peak_type RankingProcedure(const Peak_type *, ...)
{
  //Input: 1. two resonances i0 and j0
  //       2. all 3D NMR peaks with two frequencies in common with
  //          i0 and j0
  //Example:
  // peak 1 (i1, j1, k1) where |i0 − i1| ≤ tolerance, |j0 − j1| ≤ tolerance
  // peak 2 (i2, j2, k2) where |i0 − i2| ≤ tolerance, |j0 − j2| ≤ tolerance
  // peak 3 (i3, j3, k3) where |i0 − i3| ≤ tolerance, |j0 − j3| ≤ tolerance

  //Output: The most likely peak that can be merged with i0 and j0
        define a ranking parameter:
        for peak 1: A1≡1 − √(|i0 − i1| * |j0 − j1|)
        for peak 2: A2≡1 − √(|i0 − i2| * |j0 − j2|)
        for peak 3: A3≡1 − √(|i0 − i3| * |j0 − j3|)

        return peak n (in, jn, kn) with greatest A value;
}
```

The geometric mean $\sqrt{|i_0 - i_n| * |j_0 - j_n|}$ was adopted as the measure of the average deviation between peak n and peak 0. The geometric mean was chosen over the arithmetic mean, because the former tends to reduce effects from extremes of large and small values.

Once the dipeptide database has been created, it is possible to merge these dipeptides into longer chains such as tripeptide, tetrapeptide, ..., etc. For example, a dipeptide $R10-R28$ can be merged with $R28-R35$ to make a tripeptide $R10-R28-R35$ where R_i simply indicates this is the ith residue retrieved by DBPA. The aim of constructing these polypeptides is to identify the amino acid type information of their component residues thereby mapping them to the primary sequence of the protein. The probability that an "amino-acid-type-recognized" polypeptide occurs only once in a protein depends on the length of the polypeptide.[15] A longer polypeptide has a higher probability of being mapped uniquely to its corresponding primary sequence. The algorithm PGA(Polypeptide Generating Algorithm) listed below shows how dipeptides can be merged together to form polypeptides. Details concerning the amino acid type recognition and primary sequence mapping are discussed in a companion paper.[11]

```
void CreatePolypeptide(Dipeptide_type, ...)
{
  //Input: a set of dipeptides
  //Output: polypeptides
  //Examples:
  //          R3 − R5
  //          R5 − R29
  //          R29 − R18
  //          R18 − R16
  //          R18 − 38
  //          produce output R3 − R5 − R29 − R18 − R16 and R3 − R5 − R29 − R18 − R38

      for each dipeptide in the input {
          copy this dipeptide into the polypeptide chain P;
          for each dipeptide in the input {
              if this dipeptide can be merged with chain P {
                   push this dipeptide into stack S;
              }
          }
          append(P,S);     // append() function will increase the length
                           //  polypeptide P
      }
}
```

RESULTS

All of the algorithms are implemented in C computer language and were tested on a 90 residue globular protein. Figure 9 is a brief flowchart illustrating the relationships between the input data and various algorithms. The experimental data were provided by University of Alberta. All spectra were obtained on a Varian Unity 600 NMR spectrometer operating at 30 °C.[16] The sample protein is the 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, 3D HNCO, 3D HNCOCA, 3D HCACO, 3D ^{15}N TOCSY-HMQC, and NOESY-HMQC.

Cross peaks were automatically picked from the transformed 3D spectra using the CAPP peak picking program.[17]

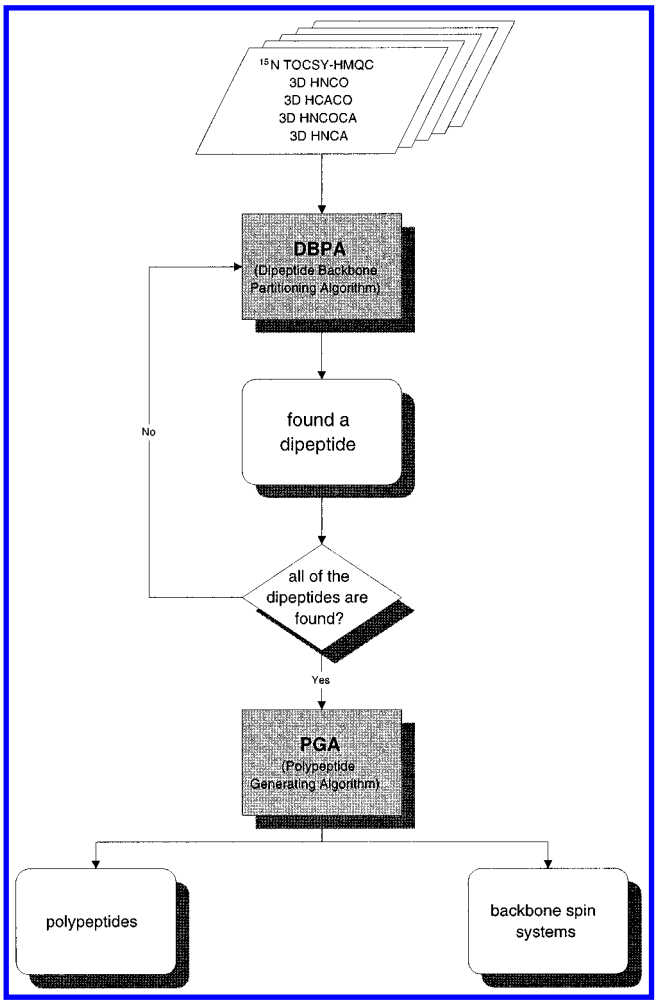


Figure 9. The flow diagram of the algorithms mentioned in this paper.

Table 1. Extracted Residues of the 90 Amino Acid Protein NTnC Using the Dipeptide Backbone Partitioning Algorithm[a]

obsd residues in the first run of DBPA	obsd residues in the second run of DBPA	obsd residues in the third run of DBPA	residues unable to assign without human inspection of data	obsd residues in the first run of DBPA	obsd residues in the second run of DBPA	obsd residues in the third run of DBPA	residues unable to assign without human inspection of data
			D5				L49
Q6							G50
Q7				Q51			
A8				N52			
E9							P53
A10							T54
R11				K55			
A12				E56			
F13					E57		
L14					L58		
S15				D59			
E16				A60			
E17				I61			
M18						I62	
I19							E63
A20						E64	
E21							V65
F22							D66
K23							E67
A24				D68			
A25					G69		
		F26			S70		
		D27					G71
F29							T72
D30				I73			
A31				D74			
D32							F75
G33						E76	
G34				E77			
			G35			F78	
			D36			L79	
			I37	V80			
			S38	M81			
		T39		M82			
			K40	V83			
			E41	R84			
L42				Q85			
		G43		M86			
		T44		K87			
V45				E88			
M46				D89			
		R47		A90			
		M48					

[a] See text for the definition of various runs of DBPA.

The CAPP program is run at the noise level; therefore, a number of false peaks are unavoidably picked. Many of these false peaks can be removed by filtering the peak lists of the 3D spectra through high signal-to-noise 2D spectra.[16] There was no further human inspection of the peak lists. The final peak lists were given to the authors by B. Sykes at the University of Alberta.[16]

The 3D HNCO peak list contains 135 cross peaks compared with about 90 peaks predicted for a 90 residue protein. The 3D HCACO peak list has 125 peaks and 3D HNCA has 242 peaks, which include both interresidue $NH_i-N_i-C\alpha_{i-1}$ and intraresidue $NH_i-N_i-C\alpha_i$ peaks. 3D HN(CO)CA has 135 peaks and ^{15}N TOCSY-HMQC consists of 141 peaks. All peak lists were input into DBPA as shown in Figure 9.

To process peaks coming from different spectra, various tolerance values are introduced since the spectra were not perfectly aligned. The tolerance value for comparing proton frequencies was chosen to be 0.05 ppm. For the rest of the nuclei, tolerance values are 0.40 ppm for nitrogen, 0.30 ppm for CO, and 0.47 ppm for C_α. These tolerance values are adjustable based on user's experience and spectra quality.

According to Figure 6, eight 3D NMR cross peaks are required to construct a dipeptide. However, it is unlikely to receive such a good data set without missing peaks. Hence, the ability of handling missing peaks becomes an important criterion for automated assignment tools. In the Troponin-C spectral data, 34 out of the 86 amino acid residues have at least one missing peak. In the first run of DBPA we define a successfully assigned dipeptide as the one having all of the 10 resonance identified. This is a strict condition. As a result, the above 34 residues are not assigned in the first run of DBPA. The successful assignment percentage is approximately 60% (see Table 1).

Perhaps the best way to demonstrate how DBPA can overcome the peak missing problem is by the example shown below. Residue E57 of Troponin-C misses a 3D NMR peak, the HCACO peak (αH, C_α, CO). HCACO and HNCO are the two experiments observing CO frequencies. While HNCO peaks, (CO($i-1$), HN(i), N(i)), in general determine the CO resonance of the first residue of a dipeptide, lack of HCACO peak makes DBPA unable to determine the CO frequency of E57 in dipeptide E56-E57. As a result, E56-E57 remains in the category of unassigned dipeptides in the

first run of DBPA on Troponin-C data set because its CO frequency has not been determined yet. In order to identify E57, users have an option to relax the 10-resonance definition of a dipeptide. In other words, DBPA now can recognize E57 as the second residue of dipeptide E56-E57 even though E57 has one undetermined resonance. The relaxation of the definition of dipeptides must be conducted carefully, because the possibility of receiving multiple assignments for a dipeptide is increasing due to the fact that only seven instead of eight peaks are required for identifying a dipeptide. A compromised approach is to take out all the used peaks, peaks that have been used by DBPA to construct dipeptides in the first run, after the first run of DBPA. Using this approach, DBPA successfully assigns additional four dipeptides, E56-E57, E57-E58, D68-G69, and G69-S70 in the second run. Note that the CO frequencies of these residues are absent. Proper human assistance could help to retrieve the absent frequencies.

Sometimes a single missing peak may lead to two unassigned resonance within a dipeptide. Using Troponin-C as an example, the missing ^{15}N TOCSY-HMQC peak, (N, NH, αH), of F78 makes DBPA fail to determine the αH frequency of F78 in dipeptide E77-F78. The missing αH results in a missing CO of F78 because the CO frequency is supposed to comes from peak (αH, C_α, CO). To extract a dipeptide with two missing frequencies, in this example CO and αH, one needs to further relax the definition of a dipeptide. Therefore, now eight assigned resonance can be considered as an assigned dipeptide in the third run of DBPA. Using the Troponin-C data, additional 12 dipeptides can be determined after the third run of DBPA. This makes the percentage of assigned residues to about 79% (67 of 85).

Eighteen residues remain unassigned after three runs of DBPA. They all miss two or more peaks. Before appropriate manual inspection on the data set is conducted, it is difficult to assign more residue at this stage.

The algorithm DBPA produced 161 dipeptides which in turns was input into the algorithm PGA. In PGA, the 161 dipeptides were compared against each other to eliminate redundant spin systems, finally resulting in 98 unique backbone spin systems. Redundant spin systems are those residues having very close resonance frequencies to each other. For example, the following two dipeptides, {(120.39, 58.98, 178.74, 8.11, 4.04) − (124.44, 55.41, 180.06, 7.89, 4.12)} and {(117.07, 56.20, 175.12, 7.81, 4.61) − (124.42, 55.04, 180.24, 7.89, 4.14)} have a similar, in terms of chemical shifts, C-terminal residue. DBPA considers these two residues as the same spin system. Theoretically 90 spin systems should be observed for the 90-residue NTnC. Table 1 summarizes the results of the backbone spin system extraction. Once the unique backbone spin systems have been extracted, PGA merges the dipeptides into polypeptide chains.

DISCUSSION

Computer algorithms are presented to automate the resonance assignment of protein backbone using heteronuclear NMR. The principle and implementation of the algorithm DBPA (Dipeptide Backbone Partitioning Algorithm) is described. In this section, a number of program options are discussed.

DBPA has an option to handle two searching operations. Both operations are used in the construction of a dipeptide. They are described in the following:

1. Given a dipeptide with m assigned frequencies, DBPA takes two frequencies, δ_i, δ_j, where $i, j \in \{1, 2, ..., m\}$ and $i \neq j$, then searches a candidate peak having two frequencies overlapped with δ_i, δ_j in the spectrum database. Suppose the third frequency of the candidate peak is δ_k. δ_k will be merged into the dipeptide and result in a dipeptide with $m + 1$ assigned resonance. If many candidate peaks are found, a ranking system is implemented in DBPA to select a peak from the many candidates and merge this peak to the dipeptide. Alternatively, a user can tell DBPA to make a replication of the dipeptide for each of the candidate peaks then merge that peak to the replicated dipeptide. This procedure is described later.
2. Given a dipeptide with m assigned frequencies, DBPA takes two frequencies, δ_i, δ_j, where $i, j \in \{1, 2, ..., m\}$ and $i \neq j$, then searches two candidate peaks in the input spectrum database. The first candidate has frequency δ_i and two other frequencies, suppose they are denoted as δ_k and δ_l. The second candidate peak has frequency δ_j, δ_k, and δ_l. Note that two frequencies are overlapped between the two candidate peaks. DBPA will then merge resonance δ_k and δ_l into the dipeptide. This procedure results in a dipeptide with $m + 2$ frequencies.

These operations can both be seen in Figure 6. The first operation is used in steps 1−6, while the second operation is used in steps 7 and 8.

DBPA is not designed for certain types of NMR experiments. It can process many combinations of triple resonance heteronuclear 3D NMR experiments. The only requirement is to give DBPA sufficient information in order to accomplish complete dipeptide assignments. For example, a single 3D HNCO spectrum does not provide enough information to assign a dipeptide because only three resonances, NH_i, N_i, and CO_{i-1}, can be determined. Similarly, a 3D HNCO and a HNCA, giving four resonances, NH_i, N_i, $C\alpha_i$ and CO_{i-1}, do not provide enough information, either. Apparently an approach is needed to determine whether an NMR data set is sufficient to assign the 10 resonances of a dipeptide or not. A simple algorithm was designed to verify the completeness of the input NMR data set. The algorithm is listed as follows:

```
void VerifyCompleteness(Heteronuclear3DNMR_type, ...)
{
//
// Input : All available heteronuclear 3D NMR spectra. Required
//         information includes the resonances observed in the experiments
//         and correlations between the resonances.
//
// Example: For 3D HNCO spectrum, the input information is
//          ( NH_i, N_i, CO_{i-1}).
//
// Output: All possible permutations of the input NMR experiments leading
//         to a complete dipeptide assignment, i.e.,
//         ( N, NH,α H, Cα,CO)_{i-1}--( N,  NH,α H, Cα,CO)_i

  suppose the number of input NMR experiments is N;
  compute all possible N! permutations for the N NMR experiments;
  for each of the permutation {
     fill the three observed resonances of the first experiment into an
     empty dipeptide;

     for each of the remaining N-1 experiments in this permutation {
        if two and only two of the three observed resonances overlap
        with any two resonances in the dipeptide
            add the third observed resonance of this experiment into the
            dipeptide;
        if the 10 resonances of the dipeptide are filled
            a complete permutation is found, break the inner loop;
     }

     if the dipeptide are filled with 10 resonances
        output this permutation;
     else
        this permutation does not provide sufficient information to assign
        10 backbone resonances;
  }
}
```

Essentially this approach follows the same concept of DBPA, namely, two overlapped resonances of two 3D NMR cross peaks confirm the merge of these two peaks. In the beginning all possible permutations of the supplied NMR experiments are computed. For a data set containing N spectra, there are $N!$ permutations. Here a permutation means a sequence of using the peaks of the NMR experiments. This $N!$ permutations are then examined to determine whether they provide sufficient resonance correlations to construct a dipeptide. Consider the following five NMR spectra: 3D HNCO, HNCA, HN(CO)CA, HCACO, ^{15}N TOCSY-HMQC. There are a total of $5! = 120$ sequences to apply those five spectra. Not all the permutations result in a complete assignment of a dipeptide. It is possible that none of them provide sufficient information. Given a data set containing N NMR spectra, the above short codes extract all the permutations that produce complete dipeptides.

In this paper we introduced the procedure that requires a minimum of eight correlations to assign the backbone resonances of a dipeptide. The minimum number is determined based on the fact that each residue's backbone has five resonances (N, αH, Cα, NH, CO), thus a dipeptide is composed of 10 resonances. Suppose these 10 resonances are denoted as (a_1, b_1, c_1, d_1, e_1) and (a_2, b_2, c_2, d_2, e_2) where the first five numbers represent the resonances of residue 1, while the last five numbers are the resonances of residue 2. One of the possible combinations of the eight necessary correlations are $\{a_1, b_1, c_1\}$, $\{b_1, c_1, d_1\}$, $\{c_1, d_1, e_1\}$, $\{d_1, e_1, a_2\}$, $\{c_1, d_1, b_2\}$, $\{a_2, b_2, c_2\}$, $\{b_2, c_2, d_2\}$, and $\{c_2, d_2, e_2\}$. In this case, correlation $\{a_1, b_1, c_1\}$ and $\{b_1, c_1, d_1\}$ give rise to four resonances, a_1, b_1, c_1, and d_1. Similarly, resonance e_1 can be determined by merging $\{b_1, c_1, d_1\}$ and $\{c_1, d_1, e_1\}$. Repeating this procedure, all the 10 resonance can be determined. It is generally not easy to declare a minimum set of required NMR experiments for automated assignment strategy like the one discussed here, nor is it necessary. There are many different NMR experiments, each provides one or more inter- or intraresidue correlations. What is relevant here is the minimum number of correlations between the nuclei, not the number of NMR spectra.

CONCLUSION

A set of algorithms is proposed to automate protein backbone resonance assignment using through peptide bond interresidue correlations. The DBPA (Dipeptide Backbone Partitioning Algorithm) merges cross peaks among available NMR spectra to extract the backbone spin systems. Every merge is confirmed by two pieces of evidences, i.e., two overlapped frequencies of a 3D cross peak. To fulfill this requirement, six intraresidue and two interresidue correlations are needed to construct a dipeptide spin systems. Once all the possible dipeptides are obtained, PGA (Polypeptide Generating Algorithm) links dipeptides to form polypeptides each of which in turn can be manually or automatically assigned to the protein's primary sequence. DBPA can be applied to many different sets of NMR experiments. The five experiments (3D HNCO, HNCA, HN(CO)CA, HCACO, and ^{15}N TOCSY-HMQC) along with 3D CBCANH were chosen to demonstrate the generality of the DBPA algorithm.

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

ACKNOWLEDGMENT

The authors thanks Stéphane Gagne for providing the NMR data of the protein data of chicken skeletal troponin-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) Bax, A.; Grzesiek, S. Methodological advances in protein NMR. *Acc. Chem. Res.* **1993**, *26*, 131−138.
(2) 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.
(3) Oschkinat, H.; Croft, D. Automated assignment of multidmensional nuclear magnetic resonance spectra. *Methods Enzymol.* **1994**, *239*, 308−318.
(4) Vuister, G. W.; Boelens, R.; Padilla, A.; Kleywegt, G. J.; Kaptein, R. Assignment strategies in homonuclear three-dimensional ^{1}H NMR spectra of proteins. *Biochemistry* **1990**, *29*, 1829−1839.
(5) 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.
(6) 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.
(7) 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.
(8) 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.
(9) 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.
(10) Morelle, N.; Brutscher, B.; J.-P., S.; Marion. D. Computer assignment of the backbone resonances of labelled proteins using two-dimensional correlation experiments. *J. Biomol. NMR* **1995**, *5*, 154−160.
(11) Li, K.-B.; Sanctuary, B. C. Automated assignment of proteins using 3D heteronuclear NMR. Part 2: Side Chain and Sequence-specific Assignment. *J. Chem. Inf. Comput. Sci.* **1996**, submitted.
(12) Grzesiek, S.; Bax, A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. *J. Magn. Reson.* **1992**, *96*, 432−440.
(13) Grzesiek, S.; Bax, A. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. *J. Magn. Reson.* **1992**, *99*, 201−207.
(14) Grzesiek, S.; Bax, A. Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J. Am. Chem. Soc.* **1992**, *114*, 6291−6293.
(15) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; Wiley: New York, NY, 1986.
(16) 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 Science* **1994**, *3*, 1961−1974.
(17) Garret, 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.

CI960045C