Journal of Biomolecular NMR, 4 (1994) 241–256
ESCOM

J-Bio NMR 164

Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence

Diane Zimmerman[a,b], Casimir Kulikowski[a,*], Lingze Wang[b,c], Barbara Lyons[b] and Gaetano T. Montelione[b,c,*]

[a]*Department of Computer Science,* [b]*Center for Advanced Biotechnology and Medicine and* [c]*Graduate Program in Chemistry, Rutgers University, Piscataway, NJ 08854, U.S.A.*

Received 25 May 1993
Accepted 24 August 1993

Keywords: Triple-resonance NMR; Constraint propagation; Constraint reasoning; Automated resonance assignments; Staphylococcal Protein A; Human type-α transforming growth factor

SUMMARY

We have developed an automated approach for determining the sequential order of amino acid spin systems in small proteins. A key step in this procedure is the analysis of multidimensional HCC(CO)NH-TOCSY spectra that provide connections from the aliphatic resonances of residue i to the amide resonances of residue i + 1. These data, combined with information about the amino acid spin systems, provide sufficient constraints to assign most proton and nitrogen resonances of small proteins. Constraint propagation methods progressively narrow the set of possible assignments of amino acid spin systems to sequence-specific positions in the process of NMR data analysis. The constraint satisfaction paradigm provides a framework in which the necessary constraint-based reasoning can be expressed, while an object-oriented representation structures and facilitates the extensive list processing and indexing involved in matching. A prototype expert system, AUTOASSIGN, provides correct and nearly complete resonance assignments with one real and 31 simulated 3D NMR data sets for a 72-amino acid domain, derived from the Protein A of *Staphylococcus aureus*, and with 31 simulated NMR data sets for the 50-amino acid human type-α transforming growth factor.

INTRODUCTION

The process of protein structure determination by NMR involves four principal steps (Wüthrich, 1986). First, the spin systems of the various amino acid residues in the protein are identified (Wüthrich et al., 1982; Wüthrich, 1983). Next, *sequence-specific assignments* for these

*To whom correspondence should be addressed.

spin-system resonances are determined by establishing their positions in the polypeptide sequence (Dubs et al., 1979; Billeter et al., 1982). These resonance assignments are then used to interpret through-space nuclear Overhauser effects (NOEs) and through-bond internuclear interactions as *sequence-specific conformational constraints.* Finally, structure generation programs are used to compute 3D models of the protein, satisfying these conformational constraints.

Resonance assignments are crucial to the structure determination process. In the classical approach (Dubs et al., 1979; Billeter et al., 1982; Wüthrich et al., 1982; Wüthrich, 1983,1986), these assignments are determined using NOE interactions between atoms of residues that are consecutive in the protein sequence. This has worked well for many proteins. However, attempts to automate this analysis process (Billeter et al., 1988; Eads and Kuntz, 1989; Kraulis, 1989; Cieslar et al., 1990; Nelson et al., 1991; Oschkinat et al., 1991) have had limited success, as the identification of sequential interactions from NOE data is complicated by the presence of many long-range interactions between nuclei that are far apart in sequence but nearby in 3D space. Alternate strategies using heteronuclear ^{1}H-^{13}C-^{15}N coherence transfer pathways to identify sequential relationships between amino acid spin systems (Ikura et al., 1990; Kay et al., 1990; Montelione and Wagner, 1990; Boucher et al., 1992; Grzesiek and Bax, 1992a,b; Kay et al., 1992; Logan et al., 1992; Montelione et al., 1992; Grzesiek et al., 1993; Lyons and Montelione, 1993; Lyons et al., 1993) are better suited to automated analysis. In this paper we describe a constraint-based object-oriented expert system (AUTOASSIGN) for automated analysis of triple-resonance NMR data. We also demonstrate that recently developed HCC(CO)NH-TOCSY triple-resonance experiments (Logan et al., 1992; Montelione et al., 1992; Grzesiek et al., 1993) provide a redundant set of sequential connectivity information which can be analyzed automatically by the AUTOASSIGN system.

MATERIALS AND METHODS

NMR measurements

Multidimensional HCCNH-TOCSY (Lyons and Montelione, 1993; Lyons et al., 1993) and HCC(CO)NH-TOCSY (Montelione et al., 1992; Lyons et al., 1993) were carried out as described previously, except that pulsed-field gradients (PFGs) were used to select for pathways passing through heteronuclear single-quantum states and to suppress H_2O solvent (Wang and Montelione, in preparation). NMR spectra were obtained on a Varian Unity 500 spectrometer system with the sample temperature thermostated at 30 °C. Natural abundance and uniformly ^{13}C,^{15}N-enriched samples of the Z-Domain of Protein A from *Staphylococcus aureus* were prepared as described elsewhere (Lyons et al., 1993). NMR samples were prepared in either 100% ^2H_2O or 90% H_2O/10% ^2H_2O at 2 mM protein concentration and pH 6.5.

Proton spin systems were identified using 2D 2QF-COSY (Piatini et al., 1982) and TOCSY (Braunschweiler and Ernst, 1983) data, recorded on a natural abundance Z-Domain sample dissolved in ^2H_2O. Nitrogen resonances were then connected to these spin systems using ^{13}C-decoupled ^{15}N-HSQC (Bodenhausen and Ruben, 1980) and ^{15}N-HSQC-TOCSY (Moy et al., 1993) together with 2D and 3D HCCNH-TOCSY data using a sample of ^{15}N,^{13}C-enriched Z-domain. Identification of spin-system types generally followed the classic method (Wüthrich, 1986). However, it was also possible to use HCCNH-TOCSY and HCC(CO)NH-TOCSY to uniquely identify some AMX spin systems as asparagines and some long-type spin systems as

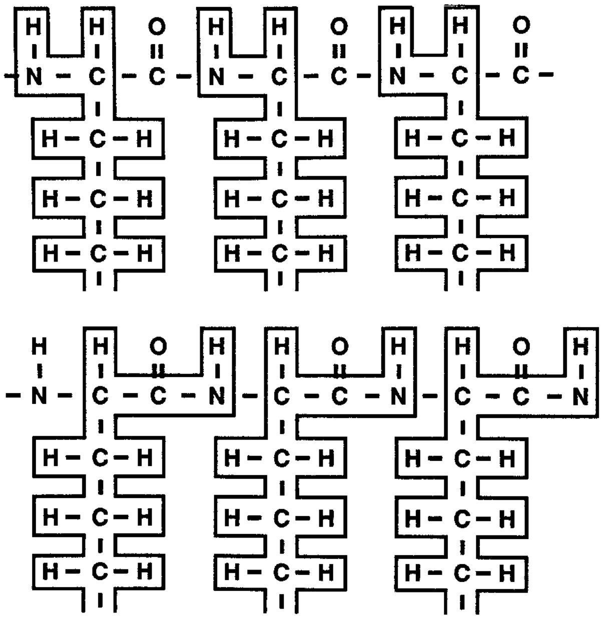

Fig. 1. Symbolic representation of the input data for AUTOASSIGN. The atoms enclosed in solid lines define generic spin systems (GSs) (top) and strings of cross peaks (CO-LDDRs) identified in HCC(CO)NH-TOCSY spectra (bottom). These same objects are defined in Table 1. The GSs provide intraresidue correlations from backbone and side-chain aliphatic resonances to the amide nitrogen and proton resonances of the same residue, while CO-LDDRs provide sequential correlations from the backbone and aliphatic resonances of residue i to the amide nitrogen and proton resonances of residue i + 1. AUTOASSIGN considers the aliphatic resonance frequencies of each generic spin system GS (residue i) and searches the set of unassigned CO-LDDRs to find the one which has the best match of aliphatic resonance frequencies. This CO-LDDR then provides information about the proton and ^{15}N resonance frequencies of the next amino acid in the sequence (residue i + 1). The GS corresponding to these amide resonance frequencies (GS (i + 1)) is then selected and a search is made for the CO-LDDR which best matches the aliphatic resonances of this GS. This CO-LDDR (i + 1) then allows identification of the amide resonances of residue i + 2, and identification of GS (i + 2). The matching procedure is further constrained by the requirement that the resulting sequence of GSs corresponds to a sequence of amino acid spin systems (SSs) in the primary structure of the protein.

glutamines (Montelione et al., 1992; Lyons et al., 1993). Modified versions of these 3D HCCNH-TOCSY and HCC(CO)NH-TOCSY experiments, in which aliphatic ^{13}C frequencies are measured during a constant t_1 evolution period, were also used to determine side-chain ^{13}C chemical shifts (Montelione, unpublished results). Using these data, spin systems of serine residues were identified by their characteristic $^{13}C^{\beta}$ chemical shifts (Szyperski et al., 1992).

3D cross peaks in the PFG HCC(CO)NH-TOCSY spectrum were picked automatically using NMR Compass software (Molecular Simulations, Inc.) on a Silicon Graphics 4D-120 computer system. This set of peaks was filtered to remove stray noise and artifacts in the spectra by comparing their amide ^{15}N and H^N resonance frequencies with the resonances in the spin-system list. Finally, the proton and nitrogen spin systems of each amino acid were assigned to specific positions in the sequence through automated analysis of these filtered 3D PFG HCC(CO)NH-TOCSY data by AUTOASSIGN. A symbolic representation of the input data for AUTOASSIGN is outlined in Fig. 1. As proline residues do not have an amide proton, Xxx-Pro dipeptides cannot show connectivities in these triple-resonance experiments and the corresponding H^{α}-H^{δ} and H^{β}-H^{δ} links are established manually (at present) using 2D NOESY data (Wüthrich et al., 1984). Sequence-specific proton and nitrogen resonance assignments for Z-Domain, based on manual analysis of 3D HCC(CO)NH-TOCSY data recorded with solvent suppression by preirradiation, are reported elsewhere (Lyons et al., 1993).

Implementation of AUTOASSIGN

The principal data structures used by AUTOASSIGN and their relationships are summarized in Table 1. Generic spin-system objects (GSs) correspond to entries in the 'spin-system list', while sequence-specific spin-system objects (SSs) correspond to specific residues in the sequence. The 'spin-system list', which is input for AUTOASSIGN, is a list of amino acid spin systems derived from manual analysis of 2QF-COSY, TOCSY, ^{15}N-HSQC, ^{15}N-edited TOCSY, and HCCNH-TOCSY data, as described above. This information is used to create the GS objects (Fig. 1, top). The HCC(CO)NH-TOCSY data are represented as 'ladders' (CO-LDDRs) of peaks (CO-PEAKs) whose coordinates correspond to the aliphatic resonance frequencies of residue i and the amide proton and nitrogen frequencies of residue i + 1 (Fig. 1, bottom). The CO-LDDR objects used by the algorithm are derived from and represent the multidimensional HCC(CO)NH-TOCSY data. AUTOASSIGN seeks to associate (i.e. assign) a single GS with each SS using the CO-LDDR information. In the course of the analysis, both GSs and SSs keep a list of their possible assignments (poss). GSs are characterized by their backbone amide proton (H^N value), backbone amide nitrogen (N value), and side-chain aliphatic proton (H^C) and aliphatic carbon (C) resonance frequencies (Table 1). Initially, GSs which have observable amide nitrogen (N value) and amide proton (H^N value) resonance frequencies in the HCC(CO)NH-TOCSY data are associated with a CO-LDDR. GSs with degenerate N-H^N frequencies are initially associated with separate identical CO-LDDRs. The task of inferring adjacencies requires establishing a connection (MTCH) between the aliphatic resonance frequencies of a given GS (H^C or C values) and the CO-LDDR frequencies of another GS. In general, each GS will have a list of MTCHs reflecting possible adjacencies in the C-terminal direction, and a single CO-LDDR whose MTCHs reflect possible adjacencies in the N-terminal direction. The heuristic score of an MTCH (mscore) is computed as:

$$\text{mscore} = \frac{\#\ \text{GS H}^{\text{C}}\ \text{or C values matched}}{\text{total}\ \#\ \text{GS H}^{\text{C}}\ \text{or C values}} \times \frac{\#\ \text{rungs matched}}{\text{total}\ \#\ \text{rungs}} \times \left[\#\ \text{matches} - \frac{\text{sum of errors}}{\text{match tolerance}}\right]$$

Each CO-LDDR also has an associated heuristic score (lscore), which is a linear function of the number of overlapping ladders, the number of peaks shared with those ladders, and the scatter of H^N and N values of CO-PEAKs within the ladder. A link-score between two GSs is computed as the product of the mscore and lscore values. This heuristic scoring scheme was settled on as one which best captures the judgments of an expert in the analysis of NMR assignments. Higher link-scores initially occur in less crowded regions of the spectrum. As matches are confirmed, it becomes possible to 'pull apart' overlapping ladders by eliminating shared (degenerate) rungs, and the link-scores associated with these ladders can thus be incrementally improved.

The sequential assignment process is guided by an interplay between two mutually constraining subgoals: (1) establishing links between GSs and (2) making a definite GS assignment for each SS. These subgoals can be accomplished by interleaving constraint propagation with goal-driven reasoning (Fig. 2). The constraint-satisfaction problem paradigm (Kumar, 1992) provides a

TABLE 1
OBJECTS USED BY AUTOASSIGN IN THE AUTOMATED ANALYSIS OF HCC(CO)NH-TOCSY DATA

Object type	Attributes	Description
GS: Generic spin system A spin system specified in the spin-system list without reference to the amino acid sequence. In addition to nitrogen and proton resonance frequencies, each GS also keeps track of possible N-terminal and C-terminal links (N-link and C-link) with other GSs and possible assignments (poss) to SSs.	H^N value	The backbone amide proton resonance frequency of GS
	H^C values	The aliphatic proton resonance frequencies of GS
	C values	The aliphatic ^{13}C resonance frequencies of GS
	N value	The backbone ^{15}N resonance frequency of GS
	MTCHs	A list of objects indicating the CO-LDDR whose H^C or C values match those of this GS
	N-link	A GS which is thought to precede this one in the sequence
	C-link	A GS which is thought to follow this one in the sequence
	poss	A list of possible SS assignments for this GS
	def	A single SS representing a definite assignment
SS: Sequence-Specific spin system A specific amino acid in the protein sequence. Its residue type directly constrains the GSs which might be assigned to it.	N-link	The preceding SS in the sequence
	C-link	The following SS in the sequence
	poss	A list of possible GS assignments for this SS
	def	A single GS representing a definite assignment
CO-LDDR A cluster of CO-PEAKs that fall within an empirically determined radius of the N and H^N values of a GS. If two GSs have identical N-H^N values, two identical CO-LDDRs will be initially created.	NH SPINSYS	A single GS uniquely associated with this ladder's N-H^N values
	CO-PEAKs	The HCC(CO)NH-TOCSY peaks which define this ladder's rungs
	MTCH	A list of objects indicating the GSs whose H^C (or C) values match this ladder's peaks
	lscore	A heuristic score measuring intra-ladder scatter (noise) and inter-ladder separation (overlap)
CO-PEAK A cross peak in the multidimensional HCC(CO)NH-TOCSY spectrum. The N and H^N values correspond to the backbone amide resonances of residue i, while the H^C (or C) value corresponds to one of the aliphatic resonances of residue $i - 1$.	CO-LDDR	The CO-LDDR(s) to which this peak was mapped
	H^N value	The backbone amide proton resonance frequency of the peak
	H^C values	The aliphatic proton resonance frequency of the peak
	C values	The aliphatic carbon resonance frequency of the peak
	N value	The backbone nitrogen resonance frequency of the peak
MTCH A match between H^C or C values of a GS and a rung (CO-PEAK) of a CO-LDDR.	status	Indicates whether or not this match is confirmed
	GS	The GS involved in this match
	CO-LDDR	The CO-LDDR involved in this match
	mscore	A heuristic score for the quality of the match

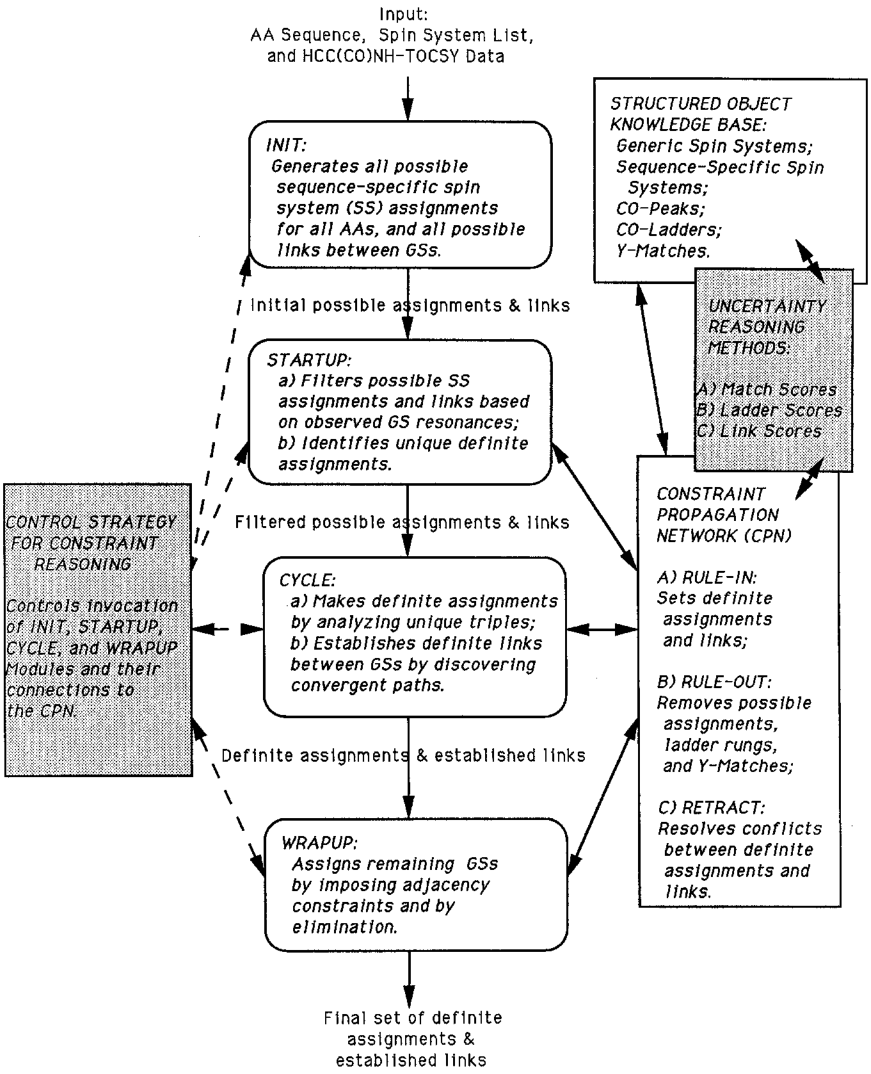
Input:
AA Sequence, Spin System List, and HCC(CO)NH-TOCSY Data
INIT:
Generates all possible sequence-specific spin system (SS) assignments for all AAs, and all possible links between GSs.
Initial possible assignments & links
STARTUP:
a) Filters possible SS assignments and links based on observed GS resonances;
b) Identifies unique definite assignments.
Filtered possible assignments & links
CYCLE:
a) Makes definite assignments by analyzing unique triples;
b) Establishes definite links between GSs by discovering convergent paths.
Definite assignments & established links
WRAPUP:
Assigns remaining GSs by imposing adjacency constraints and by elimination.
Final set of definite assignments & established links
CONTROL STRATEGY FOR CONSTRAINT REASONING
Controls invocation of INIT, STARTUP, CYCLE, and WRAPUP Modules and their connections to the CPN.
STRUCTURED OBJECT KNOWLEDGE BASE:
Generic Spin Systems;
Sequence-Specific Spin Systems;
CO-Peaks;
CO-Ladders;
Y-Matches.
UNCERTAINTY REASONING METHODS:
A) Match Scores
B) Ladder Scores
C) Link Scores
CONSTRAINT PROPAGATION NETWORK (CPN)
A) RULE-IN:
Sets definite assignments and links;
B) RULE-OUT:
Removes possible assignments, ladder rungs, and Y-Matches;
C) RETRACT:
Resolves conflicts between definite assignments and links.

Fig. 2. Overall control flow and the various modules of AUTOASSIGN.

convenient framework for sequential assignment: given a set of variables, a discrete set of possible values for each variable (a domain), and a set of constraints over the variables, the goal is to find a complete assignment of all variables which can simultaneously satisfy all constraints. Our variables are the definite assignments (defs) of SSs to GSs and the links which must be established between GSs (N-link and C-link relationships). These variables and their domains are represented as attributes of the objects listed in Table 1.

The list of possible assignments associated with a GS defines the domain of that GS's def variable. Similarly, the poss attribute of an SS defines the GSs which its def attribute can assume. Each MTCH, being a match between the H^C (or C) values of a GS and the H^C (or C) values of a CO-LDDR, represents an adjacency hypothesis. Accordingly, the MTCHs emanating from a GS define the possible links that can be established in the C-terminal direction (C-links), while MTCHs impinging on the CO-LDDR of the GS define the possible values of its N-link. Each SS also has N-link and C-link attributes which are fixed by the position of the SS in the sequence.

A key to finding a consistent sequential assignment is understanding how these attributes constrain each other. For example, consider an MTCH indicating a possible C-link from an ALA-type GS to an ILE-type GS. If none of the ALA-type SSs are followed by an ILE in the sequence, this MTCH can be eliminated. This reasoning assumes that these spin-system types are correctly identified in the spin-system list. On the other hand, when a possible SS assignment to a GS is not supported by MTCHs in either the N-terminal or C-terminal directions, this assignment can be tentatively excluded. This last constraint can rule out a correct assignment when the expected HCC(CO)NH-TOCSY cross peaks are too weak to be identified in the spectrum. However, the correct assignment may be recovered later in the analysis.

Other simple constraints involve mutual exclusion; for example, an SS cannot be simultaneously assigned to two GSs (and vice versa), and each GS can have at most one link in either direction. As definite assignments and links are established, the domains of as yet unassigned variables are also progressively narrowed. More subtle types of constraint propagation involve reasoning about domains of unassigned variables. For example, if two GSs have a link relation, then their possible assignments can be filtered according to positions in the sequence where they can be mapped as a pair (pairwise consistency). These mechanisms for pruning the domains of unassigned variables correspond to constraint-satisfaction methods (Mackworth, 1977), which reduce the amount of search required to find a solution by first filtering all mutually constrained domains for pairwise consistency.

In order for these methods of 'ruling out' to be useful, mechanisms for 'ruling in' definite assignments and links must also be present (Fig. 2). Links can be established on the basis of very high link-scores, or alternatively, by discovering 'convergent paths' generated from previously assigned GSs. Using a branching factor B, all possible paths from the N- and C-termini of all previously assigned segments are generated. When two such paths moving in opposite directions cross each other and the only way to reach otherwise unlinked GSs is via these paths, the implied 'mutually exclusive' links are committed.

Definite assignments (defs) are made in three ways. First, AUTOASSIGN identifies unique triplets of SSs whose central residue has not yet been identified. If there is only one GS that has very high link scores in both directions consistent with the central SS, that GS-to-SS assignment is made. A second way of using SS triplets to make definite assignments is based on the adjacency constraints imposed by GSs associated with surrounding SSs (i.e. SS_{i-1} and SS_{i+1}). A definite assignment is made when the possible links to these GSs reduce the possible assignments of SS_i to a single GS. Finally, additional assignments are made by elimination. For example, if in considering all possible assignments for a segment of linked GSs all but one contiguous segment of SSs in the sequence are eliminated, then those assignments of GSs to SSs are made definite.

Modules used to rule in variable assignments are an integral part of a constraint-propagation network (CPN) which ensures that the domains of all variables are kept pairwise consistent. Figure 2 gives an abstract representation of the overall control flow and how various modules interact with the CPN. Each of the rule-in modules triggers the rule-out modules which remove possible MTCHs between GSs and CO-LDDRs and possible assignments (poss) between GSs and SSs, as described above.

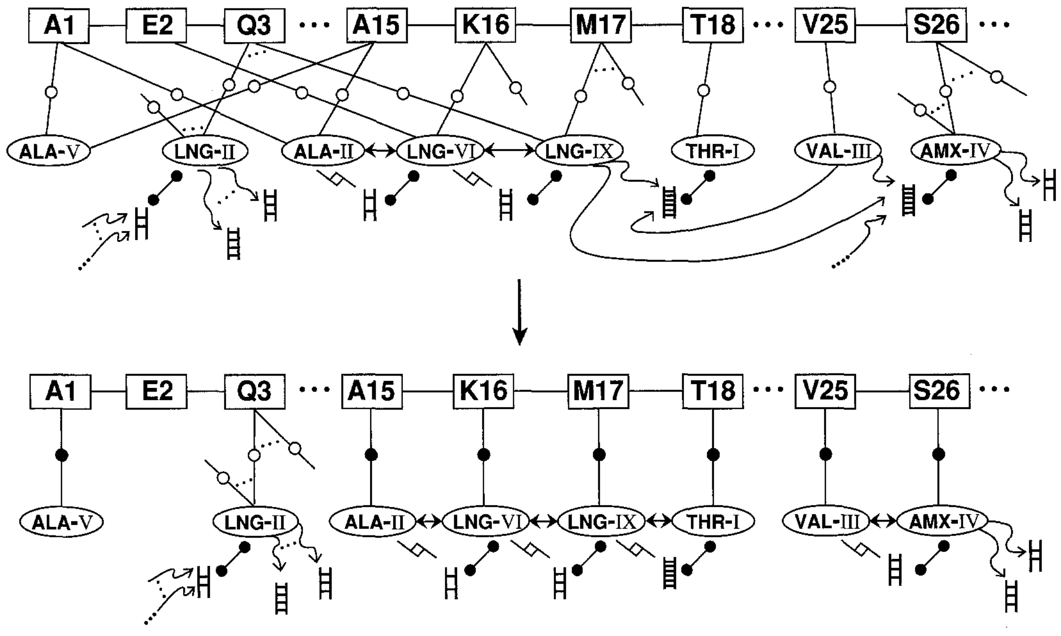

Fig. 3. Schematic view of relationships established between objects by AUTOASSIGN. These objects include: sequence-specific spin-system objects (SSs) associated with each position in the protein sequence (e.g. A15, K16, M17, etc.); generic spin-system objects (GSs) created from the spin-system list (e.g. ALA-II, LNG-VI, LNG-IX, etc.); CO-LDDRs (☷) and their unique associations (⁄●) with GSs via their common H^N and N values; unconfirmed (⤳) and confirmed (⤲) MTCHs between GSs and CO-LDDRs; definite links between GSs (↔) indicating that they are adjacent in the sequence; and the possible (○) and definite (●) assignments of GSs to SSs. This hypothetical example shows 'before' and 'after' views of the internal representation within the CYCLE phase of analysis, as described in the text.

A hypothetical example presented in Fig. 3 shows how several of the reasoning mechanisms of AUTOASSIGN interact during processing to arrive at self-consistent definite assignments. The following sequence of events demonstrates how decisions are propagated, starting at an arbitrary point within the CYCLE stage. (1) As a result of previous processing, the possible SS assignments for THR-I and VAL-III have just been reduced to Thr^{18} and Val^{25}, respectively. The number of GSs with MTCHs to the CO-LDDR of THR-I has been reduced to two (i.e., LNG-IX and VAL-III), and the CO-LDDR of VAL-III is not observed. (2) As none of the possible SS assignments for VAL-III (i.e., Val^{25}) precedes a THR spin system in the sequence, the unconfirmed MTCH between VAL-III and THR-I's CO-LDDR is now inconsistent. As a result, the rungs of THR-I's CO-LDDR are pruned to eliminate CO-PEAKs which match only the H^C values of VAL-III. This pruning enhances the link-score between GSs LNG-IX and THR-I. (3) In the next attempt to seed unique triples, the link-scores surrounding LNG-IX now exceed a threshold T, and a definite assignment is made for LNG-IX to Met^{17} as the central GS in a triplet of GSs which uniquely match the protein sequence. As Met^{17} is not followed by an AMX-type spin system in the protein sequence, the unconfirmed MTCH between LNG-IX and AMX-IV's CO-LDDR is also deleted and a definite link is established between LNG-IX and Thr-I (bottom panel). (4) SSs Thr^{18} and Val^{25} are then assigned (def) as these are the only SSs left on the 'possible assignment' lists of GSs THR-I and VAL-III, respectively. (5) The assignments of Met^{17} and

Thr^{18} are now propagated to other linked GSs, leading to assignments of LNG-VI to Lys^{16} (based on the definite link LNG-VI to LNG-IX (Met^{17})) and ALA-II to Ala^{15} (based on the definite link ALA-II to LNG-VI (Lys^{16})). The rungs of AMX-IV's CO-LDDR are also pruned as the possible MTCH from LNG- IX is not consistent with its assignment to Met^{17}. (6) The link-scores surrounding AMX-IV now exceed the threshold T, indicating a definite link to Val-III. As the GS following VAL-III (Val^{25}), AMX-IV is therefore assigned to Ser^{26}. (7) The one remaining ALA GS, ALA-V, is now assigned to the one remaining alanine SS, Ala^{1}. The assignments for GS LNG-II and SSs Glu^{2} and Gln^{3} remain ambiguous.

The pruning of MTCHs also offers the opportunity to re-examine peaks which occur on overlapping CO-LDDRs. CO-PEAKs are eliminated from a CO-LDDR when they no longer match any H^{C} (or C) values of the GSs specified by the remaining MTCHs of that ladder. This has the effect of raising both match scores and ladder scores, allowing new links to be established. In this way the system can resolve pairs (or more) of GSs with overlapping H^{N}-^{15}N values. Specifically, even spin systems with identical amide H^{N} and ^{15}N resonance frequencies can be automatically assigned, provided that: (1) the spin systems are of different types; or (2) the connectivity constraints imposed by the sequence are different for the two spin systems.

Situations can arise where an attempt to establish a definite assignment conflicts with the current MTCHs of GSs which have already been assigned. AUTOASSIGN handles these by determining the point at which the decision was made which gave rise to this inconsistency and then retracting any commitments which have propagated from that erroneous decision. Having backtracked in this way, AUTOASSIGN can then proceed to consider other assignments.

The current prototype version of AUTOASSIGN is written in the LISP programming language. Execution times for the 63 data sets described below were 4–9 min each on a Sun Sparc I workstation. Source code for the version of the program described in this paper is available upon request from the authors.

RESULTS

AUTOASSIGN was tested using both real and simulated 3D HCC(CO)NH-TOCSY data for the 72-amino acid Z-Domain derived from the staphylococcal Protein A, and with simulated data for the 50-amino acid human type-α transforming growth factor (hTGFα) protein. These simulated data were designed as part of a sensitivity analysis to test the effects of noise in the measured frequency values and of missing cross peaks on the robustness of the assignment process. For Z-Domain, the experimental spin-system list (Lyons et al., 1993) was used with both the real and simulated HCC(CO)NH-TOCSY data. For hTGFα, a simulated spin-system list was generated from the published resonance assignments (Moy et al., 1993). Resonance assignments were determined from one real (REAL) data set for Z-Domain, and 62 simulated data sets for Z-Domain and hTGFα.

Automated analysis of real HCC(CO)NH-TOCSY data for Z-Domain

At pH 6.5 and 30 °C, approximately 65% of the expected sequential cross peaks were observed in the REAL PFG HCC(CO)NH-TOCSY data. Z-Domain presents an especially challenging test case. It consists of a three-helical bundle conformation plus a disordered 14-amino acid leader sequence (Lyons et al., 1993), resulting in a highly degenerate proton frequency distribution.

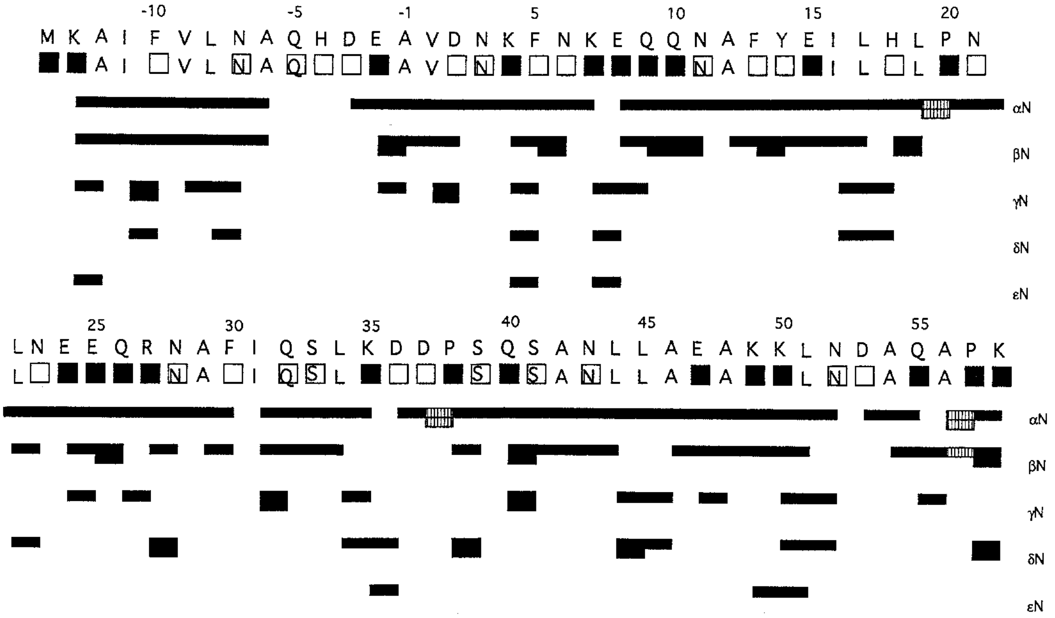
-10 -5 -1 5 10 15 20
M K A I F V L N A Q H D E A V D N K F N K E Q Q N A F Y E I L H L P N
αN
βN
γN
δN
εN
25 30 35 40 45 50 55
L N E E Q R N A F I Q S L K D D P S Q S A N L L A E A K K L N D A Q A P K
αN
βN
γN
δN
εN

Fig. 4. Survey of sequential connections identified by analysis with AUTOASSIGN of pulsed-field gradient HCC(CO)NH-TOCSY data for Z-Domain. The amino acid sequence is shown along with the corresponding sequence of spin-system types which are either unique, AMX-type (□) or LONG-type (■). Some asparagine and glutamine spin systems were identified with triple-resonance data, and serine spin systems were identified by their characteristic C^{β} resonance frequencies. Sequential cross peaks from the aliphatic proton H^{α}, H^{β}, H^{γ}, H^{δ} and H^{ϵ} resonances of residue i to the ^{15}N and H^{N} resonances of residue i + 1, identified by AUTOASSIGN, are indicated by horizontal bars labeled αN, βN, γN, δN, and εN, respectively. Xxx-Pro dipeptide connections were established using 2D NOESY data, indicated by vertically hatched H^{α}-H^{δ} and H^{β}-H^{δ} links.

Several spin-system pairs have degenerate, or nearly degenerate, ^{15}N, H^{N} and H^{α} resonances. In addition, some of the surface amide protons exhibit significant attenuation of signal intensity, due to solvent-saturation transfer effects.

Results of the automated analysis of the REAL data are summarized in Table 2. Using the peak-picked 3D PFG HCC(CO)NH-TOCSY data, 70 of the 72 spin systems of Z-Domain were assigned to their correct positions in the amino acid sequence. No errors were made in the assignment procedure. The sequential connectivities identified by AUTOASSIGN and used to establish these resonance assignments are summarized in Fig. 4. Sixty-nine spin systems were assigned directly, and one (His^{-4}) was identified by the program at the end of the analysis as the single remaining AMX spin system. For the remaining two spin systems, the 3D PFG-HCC(CO)NH-TOCSY data did not provide sequential connections, even when analyzed manually (Lyons et al., 1993), as the spin system of residue Met^{-14} could not be found in the spectra and the amide protons of residues His^{-4} and Asp^{-3} are attenuated by saturation transfer effects, even when using pulsed-field gradients for H_2O solvent suppression (Li and Montelione, 1993).

Automated sensitivity analysis of simulated HCC(CO)NH-TOCSY data for Z-Domain

A single complete and exact simulated data set (COMPLETE-EXACT) was created for Z-Domain by generating all of the expected HCC(CO)NH-TOCSY cross peaks from the assigned

TABLE 2
RESULTS FOR AUTOASSIGN PROCESSING OF REAL AND SIMULATED NMR DATA

HCC(CO)NH-TOCSY data set[a]	Number of data sets	Number of definite assignments	Assignment errors
Z-Domain			
REAL	1	70	0
COMPLETE-EXACT	1	72	0
COMPLETE-NOISY	10	72.0 (0.0)[b]	1.0 (1.1)
DELETED-EXACT	10	67.8 (2.3)	3.9 (3.0)
DELETED-NOISY	10	68.1 (5.2)	4.3 (3.7)
hTGFα			
COMPLETE-EXACT	1	50	0
COMPLETE-NOISY	10	50.0 (0.0)[b]	0.4 (1.3)
DELETED-EXACT	10	48.5 (1.6)	1.7 (2.7)
DELETED-NOISY	10	47.7 (2.3)	2.3 (2.5)

[a] The one real and 62 simulated data sets are described in the text.

[b] Average value (and standard deviation between brackets).

proton and nitrogen resonance frequencies (Lyons et al., 1993). Thirty additional simulated HCC(CO)NH-TOCSY data sets were then generated as follows. First, 10 different data sets (COMPLETE-NOISY) were created in which random Gaussian noise (Forsythe et al., 1977) was added to all resonance frequencies. This frequency noise had standard deviations of 4, 20 and 90 ppb in the H^N, H^C, and N dimensions, respectively*. These values were determined by the average difference between resonance frequencies in the spin-system list and their associated sequential cross peaks in the 3D HCC(CO)NH-TOCSY data. To simulate the effects of missing cross peaks, a second group of 10 different data sets (DELETED-EXACT) was generated in which 35% of the cross peaks were randomly deleted from the COMPLETE-EXACT data set. A different random set of peaks was deleted to generate each of these 10 DELETED-EXACT data sets. In a third set of 10 simulated data sets (DELETED-NOISY), the same Gaussian frequency noise used in creating the COMPLETE-NOISY data sets was applied to the DELETED-EXACT data set. These DELETED-NOISY data correspond most closely to the REAL data obtained from NMR measurements.

Results of the automated analysis of these 31 simulated data sets of Z-Domain are also summarized in Table 2. Using the COMPLETE-EXACT data set, AUTOASSIGN determined 72 correct spin-system assignments with no errors. For the 10 simulated COMPLETE-NOISY data sets of Z-Domain, AUTOASSIGN determined an average of 72.0 spin-system assignments

*Perturbations in the H^N and N dimensions of the actual HCC(CO)NH-TOCSY data compared to values in the spin-system list are not independent but highly correlated. This suggests that small temperature differences between the HCC(CO)NH-TOCSY data and the data used to generate the spin-system list are the primary source of H^N and N frequency noise. The differential sample heating probably results from different decoupler duty cycles used in different NMR experiments. In general, all peaks on a given CO-LDDR are shifted in the same direction and by the same magnitude in the H^N-N plane. In order to reproduce this effect in the simulated data, identical H^N and N perturbations were applied to all peaks in a given ladder. In contrast, the noise added to aliphatic H^C values was computed independently for each cross peak.

with 1.0 wrong assignments for each data set. For the DELETED-EXACT data sets, an average of 67.8 spin systems was assigned with 3.9 errors per data set, and for the DELETED-NOISY data sets AUTOASSIGN identified 68.1 spin-system assignments with an average of 4.3 errors. Hence with DELETED-NOISY simulated data, similar in quality to that obtained experimentally, there was an average of 63.8 correct assignments (89%) for the 72-amino acid Z-Domain. In three of the 10 runs with DELETED-NOISY data, 70–72 spin systems were assigned with no errors.

The dynamic performances of AUTOASSIGN on the REAL and simulated data sets are summarized in Table 3 and in Fig. 5. Comparing the spin-system list with the amino acid sequence without consideration of the HCC(CO)NH-TOCSY data leads to 1378 possible assignments of GSs to SSs. This number represents the 'spin-system degeneracy' of this amino acid sequence. With REAL data, 70% of these possible assignments are eliminated in the first phase of the analysis (STARTUP) and 42% of the GSs are assigned to SSs (Table 3). By the end of the second phase (CYCLE), the number of possible assignments of SSs to unassigned GSs is 5% of its initial value, and 82% of the GSs have been assigned to SSs. In the final phase (WRAPUP), most of the remaining GSs are assigned to SSs. In the case of the COMPLETE-EXACT data, nearly all of the assignments are made in the STARTUP stage, while for the COMPLETE-NOISY data only 63% of the assignments are made in STARTUP and the remainder are completed by the end of CYCLE. Both the DELETED-EXACT and DELETED-NOISY data sets exhibit performance

TABLE 3
PERCENTAGE OF POSSIBLE AND DEFINITE ASSIGNMENTS AT DIFFERENT STAGES OF PROCESSING OF SIMULATED AND REAL HCC(CO)NH-TOCSY DATA FOR Z-DOMAIN

Stage	Data sets									
	REAL		COMPLETE-EXACT		COMPLETE-NOISY		DELETED-EXACT		DELETED-NOISY	
	#P[a] (%P[c])	#A[b] (%A[d])	#P (%P)	#A (%A)	#P (%P)	#A (%A)	#P (%P)	#A (%A)	#P (%P)	#A (%A)
1 INIT	1378 (100)	0 (0)	1378 (100)	0 (0)	1378 (100)	0 (0)	1378 (100)	0 (0)	1378 (100)	0 (0)
2 STARTUP	413 (30)	30 (42)	4 (0)	69 (96)	140 (10)	45 (63)	338 (25)	31 (43)	482 (35)	25 (35)
3 CYCLE	69 (5)	59 (82)	2 (0)	71 (99)	1 (0)	71 (99)	33 (2)	61 (84)	42 (3)	60 (83)
4 WRAPUP	2 (0)	70 (97)	0 (0)	72 (100)	0 (0)	72 (100)	7 (1)	68 (94)	13 (1)	68 (95)

[a] #P: the number of remaining possible assignments of GSs to SSs.
[b] #A: the number of definite assignments of GSs to SSs made by the program.
[c] %P: the percent of remaining possible assignments of GSs to SSs.
[d] %A: the percent of definite assignments of GSs to SSs made by the program.

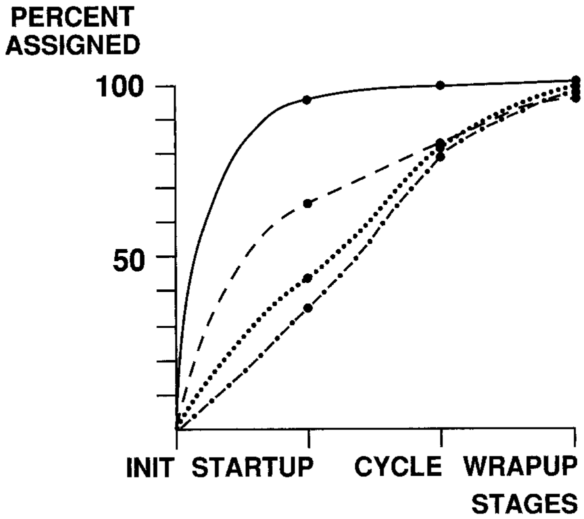

Fig. 5. Comparative dynamic performance of AUTOASSIGN. Curves are shown for the COMPLETE-EXACT (solid line), COMPLETE-NOISY (dashed line), REAL (dotted line), and DELETED-NOISY (dot-dashed line) real and simulated HCC(CO)NH-TOCSY data sets of Z-Domain, described in the text. The dynamic performance curve for the DELETED-EXACT simulated data set (not shown) was almost identical to that for REAL data.

curves similar to that of the REAL data (Fig. 5). These results indicate that while AUTOASSIGN is robust with respect to the frequency noise observed in the experimental data, nonsystematic absences of HCC(CO)NH-TOCSY cross peaks greatly affect its intermediate performance. However, the final results are on the whole remarkably accurate.

Automated sensitivity analysis of simulated HCC(CO)NH-TOCSY data for hTGFα

For hTGFα, 31 simulated HCC(CO)NH-TOCSY data sets were created using the same noise distributions and fractional deletions as for the simulated Z-Domain data sets. One COMPLETE-EXACT, 10 COMPLETE-NOISY, 10 DELETED-EXACT, and 10 DELETED-NOISY data sets were generated from nearly complete ^{1}H and ^{15}N resonance assignments which have been determined at pH 6.5 and a temperature of 10 °C (Moy et al., 1993). No degenerate N-H^{N} pairs occur in the simulated data sets for hTGFα. A real HCC(CO)NH-TOCSY data set is not yet available for hTGFα.

Results of the automated analysis of these 31 simulated data sets for hTGFα are also summarized in Table 2. Using the COMPLETE-EXACT data set, AUTOASSIGN determined all 50 correct spin-system assignments with no errors. For the 10 simulated COMPLETE-NOISY data sets of hTGFα, AUTOASSIGN determined an average of 50.0 spin-system assignments with 0.4 wrong assignments for each data set. For the DELETED-EXACT data sets, an average of 48.5 spin systems was assigned with 1.7 errors per data set, and for the DELETED-NOISY data sets AUTOASSIGN identified 47.7 spin-system assignments with an average of 2.3 errors. Hence with DELETED-NOISY simulated data, similar in quality to that obtained experimentally, there was an average of 45.4 correct assignments (91%) for the 50-amino acid hTGFα. In two of the 10 runs with DELETED-NOISY data all 50 spin systems were assigned with no errors. The dynamic performances of AUTOASSIGN on these simulated hTGFα data were similar to those described for the corresponding simulation groups for Z-Domain.

DISCUSSION

The 63 separate analyses of resonance assignments for Z-Domain and hTGFα demonstrate that the data generated from HCC(CO)NH-TOCSY experiments are highly amenable to automated analysis by AUTOASSIGN. The system provides an analysis in minutes, which normally takes several weeks or longer to complete by manual methods. For Z-Domain at pH 6.5, some 35% of the total number of expected cross peaks are missing in the REAL data, while over 90% of the sequential residue pairs share at least one expected cross peak. Accordingly, the 3D PFG HCC(CO)NH-TOCSY data, combined with information about the amino acid spin systems and Xxx-Pro links established from NOESY data, provide sufficient constraints to automatically assign most proton and nitrogen resonances of Z-Domain.

The prototype system described in this paper determines the sequence of spin systems with the highest heuristic score. It works by attempting to mimic human reasoning processes in the analysis of HCC(CO)NH-TOCSY data. For all of the data sets examined, AUTOASSIGN provides a reasonable set of resonance assignments with high reliability, and it makes no errors on the real data set with which it was trained. However, with many of the simulated data sets used to test the program, some erroneous assignments were made. For this reason, we consider the current version of the program an aid to interactive manual analysis of HCC(CO)NH-TOCSY data, rather than a replacement for human analysis. The user must evaluate the assignments proposed by the program and accept or reject them, based on human judgment.

AUTOASSIGN is a flexible expert system which can be used in real NMR assignment problems. The constraint-satisfaction paradigm provides a framework in which the necessary constraint-based reasoning can be easily expressed, while an object-oriented representation facilitates the extensive list processing and indexing involved in matching. While the current application of AUTOASSIGN has focused on analysis of 3D HCC(CO)NH-TOCSY data, with minor modifications the system could also make use of other higher dimensional (4D or 5D) data sets or of other triple-resonance experiments for establishing sequential connections between amino acid spin systems.

In the design of AUTOASSIGN, we were faced with the unusual challenge of developing robust methods, based on a single real data set. While it is difficult to fully capture the characteristics of real NMR data by simulation, our results with simulated data suggest that additional methods of error detection and recovery are still needed in AUTOASSIGN, especially when the set of HCC(CO)NH-TOCSY cross peaks is not complete. In the future, AUTOASSIGN could be extended to include backtracking mechanisms, which would allow multiple complete assignments, the delineation of the set of assignments which are equally consistent with the experimental data, and explanation facilities which would allow the user to ask the system how certain assignments were made.

In the real and simulated data sets described in this work, sequential HCC(CO)NH-TOCSY data involving ^{13}C aliphatic resonance frequencies were not considered. The overall assignment procedure will be even more robust and reliable once it includes these additional sequential connections between aliphatic ^{13}C frequencies of residue i and the amide and nitrogen frequencies of residue i + 1. Information about ^{13}C chemical shifts would also further narrow the possible assignment space (Grzesiek and Bax, 1993). This information could also be used in combination with intraresidue HCCNH-TOCSY data (Logan et al., 1992; Lyons and Montelione, 1993) to automatically generate the spin-system lists, allowing full automation of the assignment process. We are presently engaged in work along these lines.

NOTE ADDED IN PROOF

Improvements in AUTOASSIGN, made after submitting this paper for publication, now result in correct assignments for all 71 spin systems in the real data set for Z-Domain and significantly improved performance with simulated data sets.

ACKNOWLEDGEMENTS

We thank Dr. K. Newkirk for assistance in processing 3D NMR data sets and Dr. E. Bradley and Ms. L. Snyder for helpful comments on the manuscript. We also thank Dr. B. Nilsson and Ms. L. Cedergren for providing the sample of ^{15}N,^{13}C-enriched Z-Domain. This work was supported by grants from The National Institutes of Health (GM-47014) and The National Science Foundation (DIR-9019313). D.Z. is supported by an NIH Biotechnology Training Grant (GM-08330).

REFERENCES

Billeter, M., Basus, V.J. and Kuntz, I.D. (1988) *J. Magn. Reson.*, **76**, 400–415.
Billeter, M., Braun, W. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 321–346.
Bodenhausen, G. and Ruben, D.J. (1980) *Chem. Phys. Lett.*, **69**, 185–189.
Boucher, W., Laue, E.D., Campbell-Burk, S. and Domaille, P.J. (1992) *J. Am. Chem. Soc.*, **114**, 2262–2264.
Braunschweiler, L. and Ernst, R.R. (1983) *J. Magn. Reson.*, **53**, 521–528.
Cieslar, C., Holak, T.A. and Oschkinat, H. (1990) *J. Magn. Reson.*, **87**, 400–407.
Dubs, A., Wagner, G. and Wüthrich, K. (1979) *Biochim. Biophys. Acta*, **577**, 177–194.
Eads, C.D. and Kuntz, I.D. (1989) *J. Magn. Reson.*, **82**, 467–482.
Forsythe, G.E., Malcolm, M.A. and Moler, C.B. (1977) *Computer Methods for Mathematical Computations*, Prentice-Hall, Englewood Cliffs, NJ.
Grzesiek, S. and Bax, A. (1992a) *J. Am. Chem. Soc.*, **114**, 6291–6293.
Grzesiek, S. and Bax, A. (1992b) *J. Magn. Reson.*, **99**, 201–207.
Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.
Grzesiek, S., Anglister, J. and Bax, A. (1993) *J. Magn. Reson. Ser. B*, **101**, 114–119.
Ikura, M., Kay, L.E. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.
Kay, L.E., Ikura, M., Tschudin, R. and Bax, A. (1990) *J. Magn. Reson.*, **89**, 496–514.
Kay, L.E., Wittekind, M., McCoy, M.A., Friedrichs, M.S. and Müller, L. (1992) *J. Magn. Reson.*, **98**, 443–450.
Kraulis, P.J. (1989) *J. Magn. Reson.*, **84**, 627–633.
Kumar, V. (1992) *AI Magazine*, **13**, 32–44.
Li, Y.-C. and Montelione, G.T. (1993) *J. Magn. Reson. Ser. B*, **101**, 315–319.
Logan, T.M., Olejniczak, E.T., Xu, R.X. and Fesik, S.W. (1992) *FEBS Lett.*, **314**, 413–418.
Lyons, B.A. and Montelione, G.T. (1993) *J. Magn. Reson. Ser. B*, **101**, 206–209.
Lyons, B.A., Tashiro, M., Cedergren, L., Nilsson, B. and Montelione, G.T. (1993) *Biochemistry*, **32**, 7839–7845.
Mackworth, A.K. (1977) *Artif. Intell.*, **8**, 99–118.
Montelione, G.T. and Wagner, G. (1990) *J. Magn. Reson.*, **87**, 183–188.

Montelione, G.T., Lyons, B.A., Emerson, S.D. and Tashiro, M. (1992) *J. Am. Chem. Soc.*, **114**, 10974–10975.

Moy, F.J., Li, Y.-C., Rauenbuehler, P., Winkler, M.J., Scheraga, H.A. and Montelione, G.T. (1993) *Biochemistry*, **32**, 7334–7353.

Nelson, S.J., Schneider, D.M. and Wand, A.J. (1991) *Biophys. J.*, **59**, 1113–1122.

Oschkinat, H., Holak, T.A. and Ciesler, C. (1991) *Biopolymers*, **31**, 699–712.

Piatini, U., Sørensen, O.W. and Ernst, R.R. (1982) *J. Am. Chem. Soc.*, **104**, 6800–6801.

Szyperski, T., Neri, D., Leiting, B., Otting, G. and Wüthrich, K. (1992) *J. Biomol. NMR*, **2**, 323–334.

Wüthrich, K. (1983) *Biopolymers*, **22**, 131–138.

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

Wüthrich, K., Wider, G., Wagner, G. and Braun, W. (1982) *J. Mol. Biol.*, **155**, 311–319.

Wüthrich, K., Billeter, M. and Braun, W. (1984) *J. Mol. Biol.*, **180**, 715–740.