J. Mol. Biol. (1997) **269**, 592–610

JMB

Automated Analysis of Protein NMR Assignments Using Methods from Artificial Intelligence

Diane E. Zimmerman[1], Casimir A. Kulikowski[2], Yuanpeng Huang[1,2] Wenqing Feng[1], Mitsuru Tashiro[1], Sakurako Shimotakahara[1] Chen-ya Chien[1], Robert Powers[3] and Gaetano T. Montelione[1]*

[1]*Center for Advanced Biotechnology and Medicine and Department of Molecular Biology and Biochemistry Rutgers University 679 Hoes Lane, Piscataway NJ 08854-5638, USA*

[2]*Department of Computer Science, Rutgers University Piscataway NJ 08854-5638, USA*

[3]*Department of Structural Biology, Wyeth-Ayerst Research, Pearl River NY 10965, USA*

**Corresponding author*

An expert system for determining resonance assignments from NMR spectra of proteins is described. Given the amino acid sequence, a two-dimensional ^{15}N-^{1}H heteronuclear correlation spectrum and seven to eight three-dimensional triple-resonance NMR spectra for seven proteins, AUTOASSIGN obtained an average of 98% of sequence-specific spin-system assignments with an error rate of less than 0.5%. Execution times on a Sparc 10 workstation varied from 16 seconds for smaller proteins with simple spectra to one to nine minutes for medium size proteins exhibiting numerous extra spin systems attributed to conformational isomerization. AUTOASSIGN combines symbolic constraint satisfaction methods with a domain-specific knowledge base to exploit the logical structure of the sequential assignment problem, the specific features of the various NMR experiments, and the expected chemical shift frequencies of different amino acids. The current implementation specializes in the analysis of data derived from the most sensitive of the currently available triple-resonance experiments. Potential extensions of the system for analysis of additional types of protein NMR data are also discussed.

Keywords: constraint satisfaction; expert system; heteronuclear triple-resonance experiments; isotope enrichment; knowledge-based data structure

Introduction

Resonance assignments form the basis for analysis of protein structure and dynamics by NMR (Wüthrich, 1986) and their determination represents a primary bottleneck in protein solution structure analysis. In many cases, the sequence-specific assignment of backbone resonances is sufficient to allow immediate interpretation of chemical shift, NOESY, and scalar coupling data in terms of the protein's secondary structure and chain fold. The introduction of multi-dimensional triple-resonance NMR (Montelione & Wagner, 1989, 1990; Ikura *et al.*, 1990; Kay *et al.*, 1990) has dramatically improved the speed and reliability of the protein assignment process. Interpretation of these triple-resonance data is greatly facilitated by computer-assisted analysis (Zimmerman *et al.*, 1993, 1994; Friedrichs *et al.*, 1994; Hare & Prestergard, 1994; Meadows *et al.*, 1994; Olsen & Markley, 1994; Morelle *et al.*, 1995; Zimmerman & Montelione, 1995; Bartels *et al.*, 1996). AUTOASSIGN† (Zimmerman *et al.*, 1993, 1994) is a prototype expert system that determines backbone ^{15}N, ^{13}C, and ^{1}H and side-chain ^{13}C$^{\beta}$ resonance assignments from a set of three-dimensional triple-resonance protein NMR spectra in conjunction with a two-dimensional ^{1}H-^{15}N heteronuclear correlation spectrum. The software

Abbreviations used: AI, artificial intelligence; Csp A, major cold shock protein A; FGF-2, human basic fibroblast growth factor; GS, a generic amino acid spin-system object derived from NMR spectral data; NMR, nuclear magnetic resonance; NOESY, nuclear Overhauser spectroscopy; NS-1, influenza A virus non-structural protein 1; RNase A, ribonuclease A; SS, sequence-specific spin-system object corresponding to an amino acid residue in the protein sequence; p.p.m., parts per million; 2D, 3D, two and three-dimensional.

† AUTOASSIGN and the input peak lists used for the results summarized in this paper are available by request. The software is implemented in the Allegro Common Lisp Object System (CLOS) and requires a lisp compiler (available from Franz Inc.) for execution.

0022–2836/97/240592–19 $25.00/0/mb971052

utilizes many of the analytical processes employed by NMR spectroscopists, including constraint-based reasoning (Fox, 1986; Nadel, 1986; Kumar, 1992) and domain-specific knowledge-based methods, exploiting known characteristics of the specific NMR experiments and unique features of amino acids in the sequence.

In this paper we report the performance of AUTOASSIGN on seven triple-resonance NMR data sets. These proteins contain between 69 and 154 amino acid residues and the spectra vary widely with respect to completeness, resolution, degeneracy, and noise perturbations. With these data, approximate execution times† varied from 16 seconds to nine minutes on a Sun Sparc 10 workstation, depending on the size of the protein, the quality of the spectra, the number of spin systems present in excess of the number expected from the protein sequence, and other more general measures of complexity of the interpretation. For proteins that yield reasonably good quality triple-resonance NMR data sets, AUTOASSIGN provides almost complete automated analysis of backbone resonance assignments in minutes, reducing the analysis process to the 7 to 21 days of NMR instrument time needed for recording the requisite data. For poorer quality or incomplete data sets, AUTOASSIGN provides partial assignments along with interactive tools to support further exploratory analysis.

Results

Overview of the problem-solving strategies and experimental input

Figure 1 summarizes the experimental NMR data used as standard input for AUTOASSIGN. The ^{15}N-H^{N} resonance frequencies of cross-peaks detected in these various spectra are used to define groups of cross-peaks associated with common amide N-H atoms. Situations in which ^{15}N-H^{N} cross-peaks of two or more amino acid residues overlap in these spectra are handled specially. Although the term "spin system" can be used most generally to refer to any set of atoms that interact through a defined set of nuclear magnetic interactions, in this paper we use it to refer specifically to scalar-coupled heteronuclear spin systems associated with specific N-H sites in the protein sequence. A generic spin system (GS) is one that has been identified from the NMR data but not yet assigned to a specific site in the amino acid sequence. The root of each GS is defined by its backbone amide ^{15}N-H^{N} resonance frequencies.

In addition to these root frequencies, each GS has two lists of designated ^{13}C and ^{1}H chemical shifts, derived from triple-resonance experiments that detect interactions of additional nuclei with the amide N-H group. Four of these (i.e. H(CA)(CO)NH, CA(CO)NH, CBCA(CO)NH, and HNCO) are considered "sequential" triple-resonance experiments, as they correlate H$^{\alpha}$, C$^{\alpha}$, C$^{\beta}$, or C′ nuclei of residue ($i-1$) with the backbone N-H atoms of residue i. Complementary to these, four "intraresidue" experiments (i.e. H(CA)NH, CANH, CBCANH, and HN(CA)CO) detect interactions of the H$^{\alpha}$, C$^{\alpha}$, C$^{\beta}$, or C′ nuclei of residue i with its

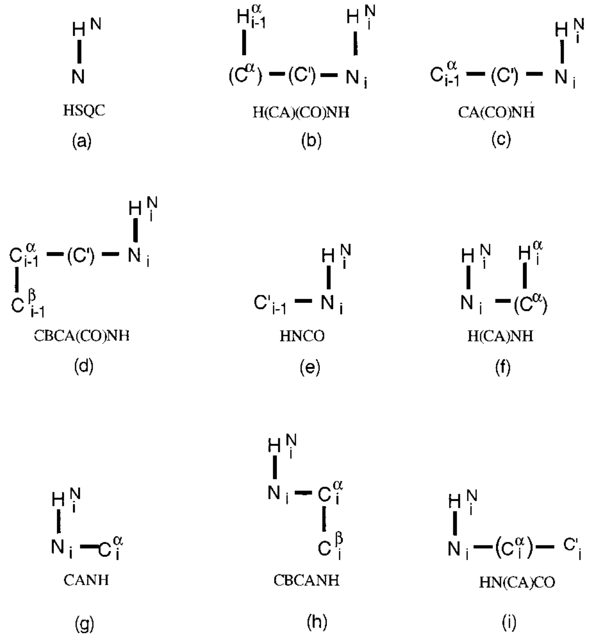


Figure 1. Schematic representation of experimental input to AUTOASSIGN. Each NMR experiment used in creating input for AUTOASSIGN is depicted as a non-directed graph whose edges reflect the transfer of magnetization through the participating nuclei. Those nuclei that occur along a given path but that are not detected in a given experiment are shown in parentheses. Peak-picked 2D ^{15}N-H^{N} heteronuclear correlation data (a) define the backbone amide resonances used to identify the roots of generic spin systems (GSs). Experiments (b) through (e) correlate the H$^{\alpha}$, C$^{\alpha}$, C$^{\beta}$, and C′ frequencies of residue $i-1$ with the ^{15}N-H^{N} frequencies of residue i and are used to define CO-ladders. Experiments (f) through (i) correlate the H$^{\alpha}$, C$^{\alpha}$, C$^{\beta}$, and C′ frequencies of residue i with its own ^{15}N-H^{N} frequencies and are used to define CA-ladders. With the exception of FGF-2, experiments (a) through (h) were collected for all of the proteins tested. For the RNase A data sets, HNCACO data (i) were also collected. For FGF-2, experiments equivalent to (a) and (d) through (h) were carried out. In addition, for the FGF-2 data manual analysis was used to extract C$^{\alpha}$ cross-peaks from a CBCA(CO)NH-type spectrum and H$^{\alpha}$ cross-peaks from an HBHA (CO)NH spectrum, and a separate pre-processing module was used to infer intraresidue C′ resonances from an HCACO spectrum. Simulated peak lists corresponding to data that would be provided by experiments (b), (c), and (i) were thus defined for FGF-2 and included in the input for AUTOASSIGN.

† Execution times reported here are elapsed real times for runs carried out on a Sun Sparc 10 workstation and varied depending on the workstation's work load.

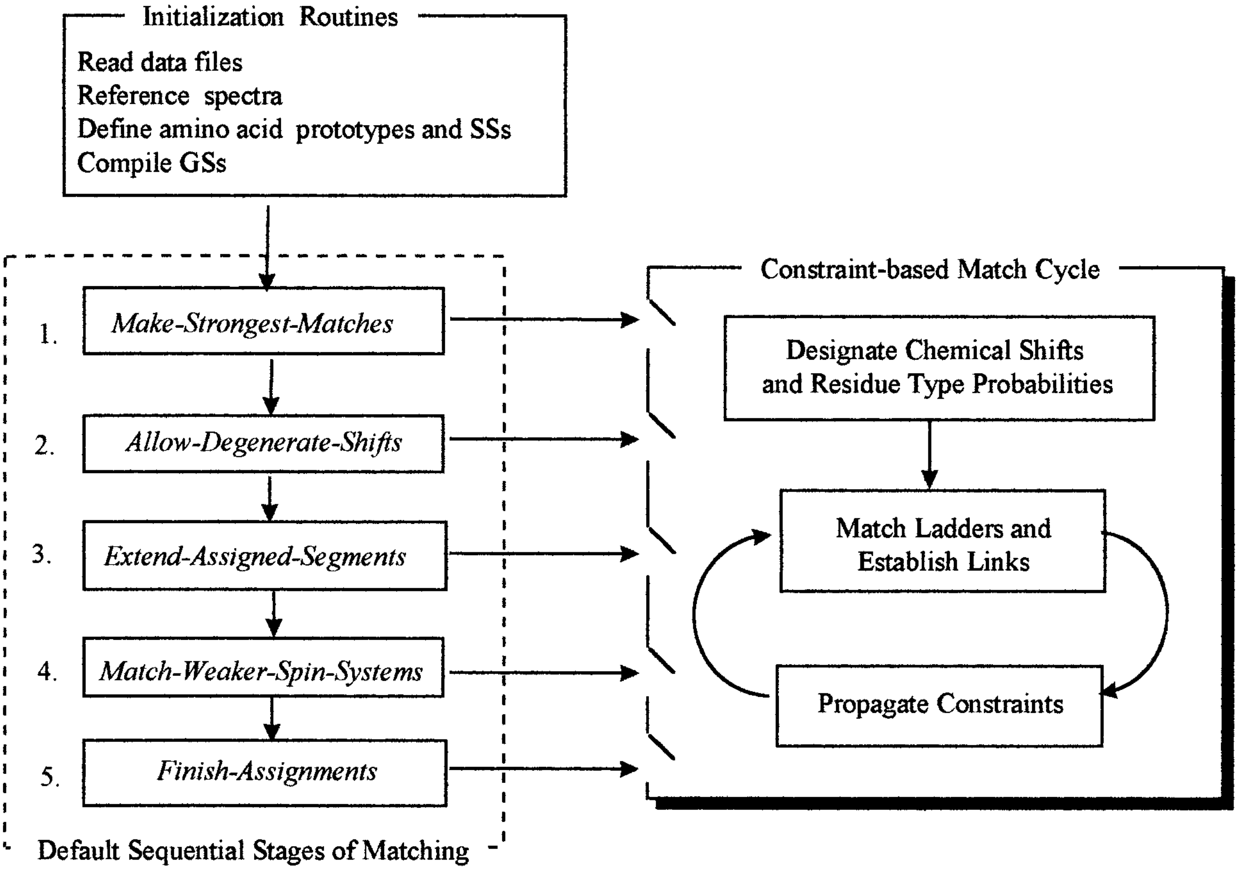

Initialization Routines
Read data files
Reference spectra
Define amino acid prototypes and SSs
Compile GSs
1. Make-Strongest-Matches
2. Allow-Degenerate-Shifts
3. Extend-Assigned-Segments
4. Match-Weaker-Spin-Systems
5. Finish-Assignments
Default Sequential Stages of Matching
Constraint-based Match Cycle
Designate Chemical Shifts and Residue Type Probabilities
Match Ladders and Establish Links
Propagate Constraints

Figure 2. Schematic overview of AUTOASSIGN's default execution sequence. Five sequential stages of analysis (see the text) follow the initialization routines that process the input files and set up AUTOASSIGN's internal representations. Depending on the execution stage, different methods and/or criteria are used to designate chemical shifts and establish sequential links between spin systems in the constraint-based match cycle.

own backbone amide N-H atoms†. Adjacency relations between GSs are inferred by matching the designated intraresidue shifts of one GS to the sequential shifts of another.

There is an important distinction between the designated shifts of a GS and the many cross-peaks that may be associated with the corresponding N-H root, since cross-peaks from multiple experiments may imply several candidate frequencies for a given resonance. A designated chemical shift is a single numerical value that is assigned as that atom's resonance frequency. To emphasize this distinction, the lists of designated intraresidue and sequential chemical shifts associated with each N-H root are referred to as the CA-ladder and CO-ladder, respectively, of the GS.

The software can be run interactively or in "batch mode". Figure 2 shows a schematic overview of AUTOASSIGN's default execution sequence when run in batch mode. First, a set of initialization routines (top of Figure 2) is executed to process the input files and create AUTOASSIGN's internal representations. Unless otherwise directed by the user, the software next enters a sequence of stages of "constraint-based matching", which progressively relax or otherwise redefine the criteria used to designate chemical shifts and establish sequential matches between the CO and CA-ladders. In each of these stages, AUTOASSIGN begins by initializing or redefining the currently designated chemical shifts. The designated chemical shifts are then used to establish iteratively the best matches between the CO and CA-ladders; matches that can be confirmed as the "best possible" are established as sequential links between GSs. Any constraints entailed by these links are then propagated to establish sequence-specific assignments or further constrain the remaining possible matches. Although the actual methods used to designate shifts and establish links vary depending on the particular stage of analysis, the basic strategy of designating chemical shifts followed by iterative sequential matching and constraint propagation is common to all stages, and is abstractly depicted in Figure 2 as a Constraint-based Match Cycle.

† Sequential cross-peaks are also observed in these "intraresidue" triple-resonance spectra (Montelione & Wagner, 1989, 1990; Ikura *et al.*, 1990), but can be identified through comparisons with the corresponding "sequential" triple-resonance data.

In the first stage of analysis (Make-Strongest-Matches), only those chemical shifts that can be unambiguously inferred are designated on ladders and only those ladders having complete specifications participate in constraint-based matching. The second stage of matching (Allow-Degenerate-Shifts) refines some of the incompletely specified ladders and allows matching of the remaining unmatched (and possibly less complete) ladders. The third stage (Extend-Assigned-Segments) uses the currently established assignments and links to guide the specification of incompletely designated ladders, and focuses on extending the currently assigned stretches of the sequence. The fourth stage of constraint-based matching is used only in data sets involving extraneous GSs; i.e. data sets in which many more GSs are identified in the spectra than are expected from the amino acid sequence. In these cases, the weakest GSs (in terms of cross-peak intensities) are initially set aside. Stage 4 (Match-Weaker-Spin-Systems) then refines the designated ladders of these GSs and reinstates them in the general pool for another round of constraint-based matching. The final stage (Complete-Assignments) examines the currently designated chemical shifts and sequential assignments for possible discrepancies, making corrections and refinements where possible, and concludes with a final cycle of constraint-based matching.

Figure 3 shows the X-windows interface implemented in Tk/Tcl (Ousterhout, 1993). The default stages of execution define entry points to the constraint-based match routines corresponding to the numbered boxes in Figure 2, and are provided as options in the Assignment Tools submenu of Figure 3. Additional tools are also provided that allow the user interactively to designate chemical shifts, establish sequential assignments and/or links, search for matches to arbitrarily selected ladders, and provide statistical analyses of AUTOASSIGN results. Although the software can be run interactively or in fully automated mode, the results described here were obtained without user intervention, except where specifically indicated.

AUTOASSIGN was developed and tested using triple-resonance data sets obtained for five distinctly different proteins: the Z domain of staphylococcal nuclease protein A (7.5 kDa: Lyons *et al.*, 1993; Tashiro *et al.*, unpublished results), the single-stranded-RNA-binding cold-shock protein A from *Escherichia coli* (Csp A, 7.3 kDa: Newkirk *et al.*, 1994; Feng *et al.*, unpublished results), a homodimeric double-stranded RNA-binding domain from the influenza A virus non-structural protein 1 (NS-1(1-73), 16.6 kDa: Chien *et al.*, unpublished results), human basic fibroblast growth factor (FGF-2, 17.2 kDa: Moy *et al.*, 1995), and bovine pancreatic ribonuclease A (RNase A, 13.5 kDa: Shimotakahara *et al.*, 1997). Further testing was carried out using additional triple-resonance data obtained for two disulfide mutants of RNase A: [C65S, C72S]-RNase A (Shimotakahara *et al.*, 1997) and [C40A, C95A]-RNase A. The three RNase A data sets provide a useful case study; while their amino acid sequences are 98% identical, about one-third of the N-H resonance frequencies are significantly different in the spectra of the wild-type (wt) and mutant proteins. In addition, the spectra differ in terms of the extent of spurious peaks, N-H degeneracy, and extra GSs. All of the data sets for RNase A were analyzed independently; the results of one analysis were not used to guide the other analyses. In the discussion that follows, we treat these as two different study groups, first comparing the analyses of the five distinct proteins (group I) and subsequently, comparing the results for the three RNase A data sets (group II).

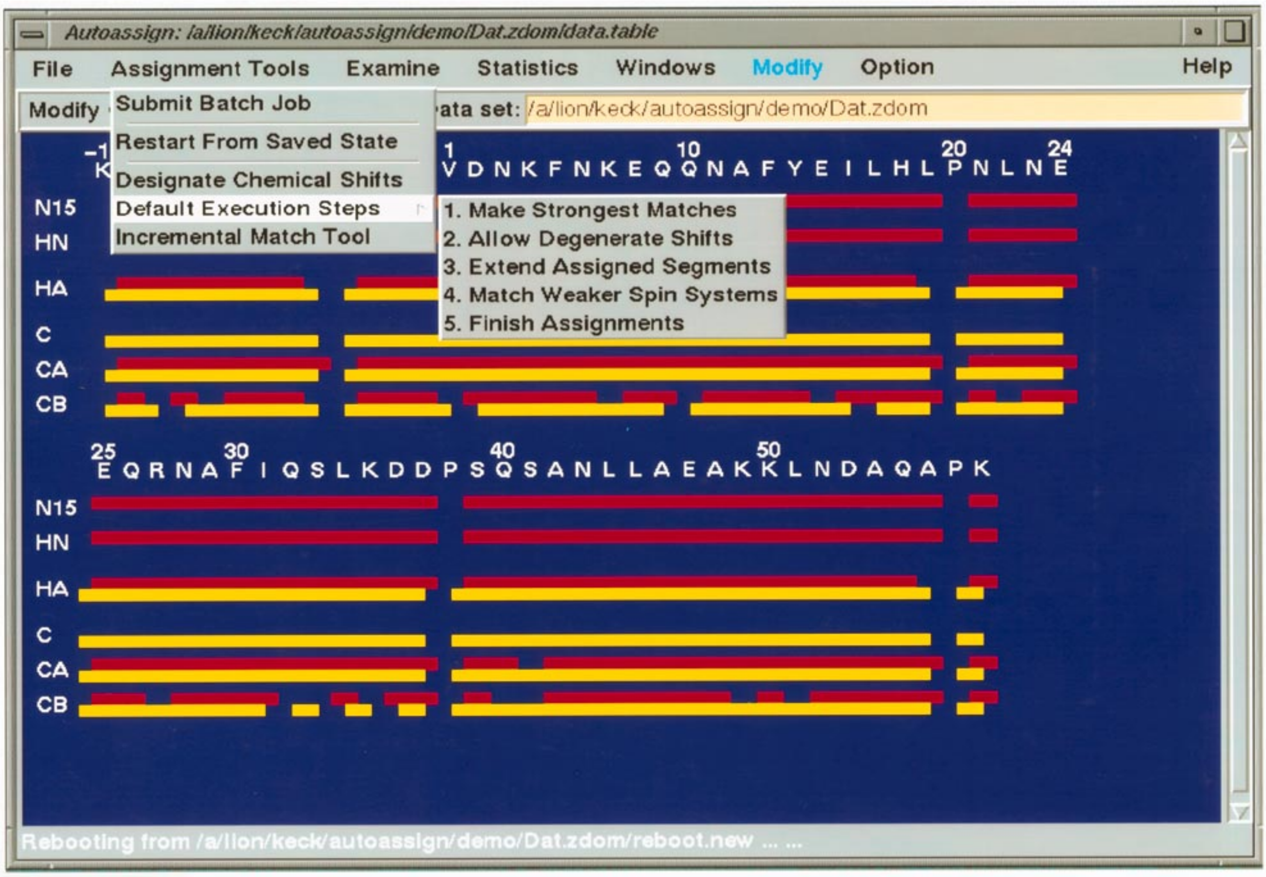
In considering AUTOASSIGN's performance, it is useful to describe the number of GS assignments obtained in terms of the number expected from the amino acid sequence. Given AUTOASSIGN's definition of a GS, only those amino acid residues that have a backbone amide N-H group are "directly" assignable to a GS. In this sense, neither the N-terminal residue nor any proline residue is directly assignable to a GS. However, the assignment of a GS to residue i also yields information about the chemical shifts of some atoms in the preceding residue. Thus, the designated chemical shifts of a GS pertain to a sequence-specific site of interacting nuclei involving a pair of sequential residues. With this understanding, GSs are assigned to sequence-specific sites that are labeled with the name of the residue containing the corresponding backbone N-H group. The percentage of complete GS assignments obtained is computed as the total number of GS assignments divided by the total number of assignable sites in the sequence. As proline and N-terminal residues do not have backbone amide N-H groups, for a sequence of n residues containing p prolines and one N-terminal residue, there are only $n - p - 1$ assignable sites.

Analysis of the group I proteins

For the proteins included in group I, AUTOASSIGN assigned an average of 97.6% of the assignable sites in the amino acid sequence to GSs identified from triple-resonance spectra, with an error rate of 0.002 (one error† out of 457 assigned sites). The execution traces in Figure 4(a) plot the fraction of GSs assigned over time, using the default stages of constraint-based matching (Figure 2) as time intervals. Figure 4(b) summarizes the fraction of sequence-specific C′, C^{α}, C^{β}, and H^{α} resonance assignments determined by AUTOASSIGN for each of these five proteins.

The Z domain is exceptional, in that 88% of the GS assignments (59 out of 67 assignable sites)

† Because the manual analysis is generally considered more reliable, automated assignments that are inconsistent with subsequent or prior manual analysis that was performed for that protein are considered errors.



Autoassign: /a/lion/keck/autoassign/demo/Dat.zdom/data.table

ata set: /a/lion/keck/autoassign/demo/Dat.zdom

Rebooting from /a/lion/keck/autoassign/demo/Dat.zdom/reboot.new ... ...

File
Assignment Tools
Examine
Statistics
Windows
Modify
Option
Help
Modify
Submit Batch Job
Restart From Saved State
Designate Chemical Shifts
Default Execution Steps
Incremental Match Tool
1. Make Strongest Matches
2. Allow Degenerate Shifts
3. Extend Assigned Segments
4. Match Weaker Spin Systems
5. Finish Assignments
-1
K
1
10
20
24
V D N K F N K E Q Q N A F Y E I L H L P N L N E
N15
HN
HA
C
CA
CB
25
30
40
50
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
N15
HN
HA
C
CA
CB

Figure 3. The X-windows interface to AUTOASSIGN. Implemented in the Tk/Tcl programming language, the main window features a default graphical connectivity map depicting, schematically, assignments and links as they are established. As indicated by the submenu for Assignment Tools, the software can be run either incrementally or in fully automated batch mode. The submenu of numbered incremental steps mimics the default sequence of problem-solving methods executed in batch mode. An additional option under the Assignment Tools includes restarting the software from a previously defined state of execution. Other options on the main menubar include tools for examining and/or modifying the currently defined spin systems, their sequential links, designated chemical shifts, and possible (or established) sequential assignments. The Statistics menu provides tools for assessing the quality of the spectra and performance of the software and generates tabulated reports that can be saved to a file. The Modify Tools allows for interactive modification of cross-peak or spin-system lists.

are made in the first stage by directly establishing the best sequential matches. The second stage relaxes the requirement that the designated CA-ladder intraresidue shifts must be distinct from corresponding chemical shifts on the CO-ladder and achieves an additional five GS assignments. Finally, two additional GS assignments are obtained in AUTOASSIGN's fifth stage of analysis, for a total of 66/67 sites assigned, leaving one assignable site and one GS unassigned. Execution time for the analysis of Z domain was about 16 seconds.

Each assignment of a GS to a site in the sequence yields atom-specific resonance assignments for those nuclei associated with the GS's CA and CO-ladders. As can be seen from Figure 4(b), the resulting resonance assignments for Z domain were also quite complete. Thus, with 66 GSs assigned to specific sites in the sequence, 71/71 H^{α}, 71/71 C^{α}, 63/71 C^{β}, and 66/67 C' resonance assignments were obtained†. This is particularly impressive considering that Z domain is highly α-helical, exhibits significant chemical shift degeneracy in the C^{α} and C^{β} dimensions, and has 19% of its GS-root frequencies partially or fully overlapped in the H-N dimensions. These assignments were verified by independent manual analysis of these (and other) triple-resonance spectra, analysis of NOESY spectra, and self-consistent structure generation calculations using these assignments (Tashiro *et al.*, unpublished results).

Like Z domain, NS-1(1-73) is a relatively small, highly α-helical domain containing 73 amino acid

† Although all of the amino acid residues include a carbonyl group in the peptide moiety, several of the protein data sets included only sequential (i.e. HNCO-type) and not intraresidue (i.e. HNCACO-type) connectivity information. In these cases, the expected number of assigned backbone carbonyl frequencies is equal to the number of assignable sites in the sequence, i.e. $n - p - 1$.

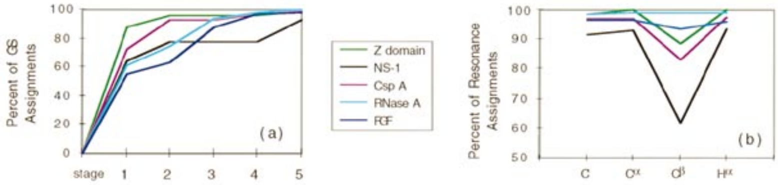


Figure 4. (a) Execution traces for the group I proteins. The percentage of GS assignments at each stage of execution is computed as the number of assigned GSs divided by the total number of assignable residue sites. Intervals on the horizontal axis correspond to the enumerated stages in Figure 2. (b) The percentages of assigned resonances for the proteins in group I. Each residue that is either itself an assignable site or is followed by an assignable site is expected to have a complete set of C′ (C), C$^\alpha$, C$^\beta$, and H$^\alpha$ resonance assignments. The percentage of assigned resonances for each atom type is calculated as the number assigned divided by the total number expected.

residues. However, NS-1(1-73) forms a homodimer with a molecular mass of 16.6 kDa, and the coherence-transfer efficiency in these triple-resonance experiments is significantly lower. The resulting spectra were therefore less complete than those obtained for Z domain. In addition, the extent of backbone amide N-H overlap is more severe (i.e. ~30% of N-H cross-peaks overlap in the 2D HSQC spectrum) and many of the chemical shifts of adjacent residues are very similar to one another. For example, resonances subsequently assigned to residues Arg37 and Arg38 have the following chemical shifts:

	H^N	N	C′	C$^\alpha$	C$^\beta$	H$^\alpha$
R37	8.48	118.0	180.2	59.5	29.8	4.42
R38	8.48	118.3	180.3	59.5	29.8	4.20

In this case two types of ambiguity arise due to: (1) the complexity of allocating peaks to GSs on the basis of N-H distinctions, and (2) the difficulty of distinguishing sequential from intraresidue triple-resonance cross-peaks in the C′, C$^\alpha$, and C$^\beta$ dimensions for Arg38. This second type of ambiguity is due to the fact that the intraresidue experiments (e.g. HNCA, CBCANH, etc.) detect sequential, as well as intraresidue, interactions† and occasionally only the sequential connections are actually observed. Given this intrinsic ambiguity, AUTOASSIGN initially defers the interpretation of cross-peaks in the intraresidue spectra that closely resemble cross-peaks in other spectra that are known to be sequential. As a result of this ''intraresidue/sequential degeneracy'', only 65% of the assignable sites are assigned for NS-1 during the first stage of matching. In the second stage, however, these ambiguous intraresidue cross-peaks are reconsidered as possible intraresidue resonances and an additional 12% of the assignable sites are assigned. By the end of execution, ten additional GS assignments have been obtained, for a total of 65/71 (92%) assignable sites. Six assignable sites and five GSs are left unassigned. The 65 GS assignments provided 71/78 H$^\alpha$, 68/73 C$^\alpha$, 44/68 C$^\beta$, and 64/71 C′ resonance assignments *via* intraresidue and/or sequential connectivities. Only about 65% of the expected C$^\beta$ resonance assignments could be derived from these triple-resonance data. Subsequent manual analyses of these triple-resonance spectra, analysis of NOESY spectra, and self-consistent structure generation calculations (Chien *et al*., unpublished results) confirmed all of the assignments made by AUTOASSIGN. This was the first example in which extensive resonance assignments were made by AUTOASSIGN using data from a protein for which assignments were not previously known, demonstrating the reliability of the software for simple systems. The execution time for automated analysis of these NS-1(1-73) triple-resonance data was about 22 seconds.

Csp A is a small β-sheet protein composed of 69 residues. In the 3D triple-resonance spectra obtained for CspA, the peaks are relatively well dispersed in the amide ^{15}N and H^N dimensions. However, some of these spectra are relatively incomplete. In addition, a few extra spin systems attributed to minor species of Csp A were also observed. As the number of observed backbone GSs (69) exceeds the number expected (66), AUTOASSIGN begins by setting aside the weakest of these for subsequent analysis (as described in Methods). As can be seen in the progress curve of Figure 4(a), stage 1 of AUTOASSIGN assigns 73% of the assignable residue sites to GSs by directly establishing the highest quality sequential matches between GSs. An additional 19% of GS assignments are obtained during stage 2, but no further progress is made until the weaker GSs are reinstated in stage 4. At that point an additional 5% of the assignable sites are assigned to GSs. The final stage obtains one additional GS assignment for a total of 65 out of 66 (98%) residue sites assigned. This leaves one assignable site (Gly3) and four GSs unassigned, but as none of these remaining GSs is consistent with the Gly3 site, no further assignments are made. The resulting 65 GS assignments provide 77/79 H$^\alpha$, 67/69 C$^\alpha$, 49/59 C$^\beta$, and 64/66 C′ resonance assignments *via* intraresidue and/or sequential connections. These assignments for Csp A have been verified by manual analysis of these (and other) triple-resonance spectra, manual analysis of NOESY spectra, and self-consistent structure generation calculations (Feng *et al*., unpublished results). The execution time for this

† Sequential cross-peaks are also observed in these ''intraresidue'' triple-resonance spectra (Montelione & Wagner, 1989, 1990; Ikura *et al*., 1990), but can be identified through comparisons with the corresponding ''sequential'' triple-resonance data.

analysis of assignments by AUTOASSIGN was approximately 16 seconds.

AUTOASSIGN has also been tested on α/β proteins; bovine pancreatic RNase A (RNase, 124 residues) and basic fibroblast growth factor (FGF-2; 154 residues). Both proteins exhibit a significant number of extra spin systems which have been attributed to minor conformations that are in slow exchange on the NMR timescale (Moy *et al.*, 1995; Shimotakahara *et al.*, 1997). As with Csp A, the analysis of the weakest GSs identified by AUTOASSIGN is deferred until the fourth stage of matching.

For FGF-2, 141/144 assignable sites were assigned to GSs, leaving three assignable sites and 23 GSs unassigned. The resulting 141 GS assignments provide 162/169 H^{α}, 148/153 C^{α}, 128/137 C^{β}, and 148/153 C' resonance assignments. One of these spin system assignments did not agree with the independent manual analysis (Moy *et al.*, 1995), as AUTOASSIGN assigned a GS believed to correspond to a minor conformation of the protein to position Ala20 of the major protein species. However, for these two spin systems corresponding to major and minor environments of Ala20, all but the backbone N-H shifts are indistinguishable. The execution time for the analysis of FGF-2 was just under four minutes.

For RNase A, 119/119 assignable sites were assigned to GSs, with 29 extraneous GSs remaining unassigned. The resulting 119 GS assignments provide 125/127 H^{α}, 122/124 C^{α}, 120/121 C^{β}, and 122/124 C' resonance assignments. These assignments for RNase A have been verified by subsequent manual analysis of the triple-resonance spectra and the interpretation of NOESY data based on these assignments identifies secondary structure that is completely consistent with its X-ray crystal structure (Shimotakahara *et al.*, 1997). The execution time for this analysis was also just under four minutes.

The triple-resonance data sets used to generate input for AUTOASSIGN were automatically peak-picked and then edited by interactive graphics to remove obvious artifactual peaks. The resulting peak lists were far from ideal, as they were incomplete and and still contained many artifactual peaks. For example, as shown in Figure 4(b) the percentage of C^{β} resonances that could be assigned from these data is consistently low relative to the other three nuclei (i.e. H^{α}, C^{α}, and C'), due to incomplete coherence transfer in the CBCANH and/or CBCA(CO)NH experiments. In general, these two spectra have significantly lower sensitivity and are often incomplete, of lower resolution, and/or include a relatively large number of spurious peaks that are not identified by the simple editing methods used in preparing the input files for AUTOASSIGN. NS-1(1-73) and wt-RNase A exhibited an especially large number of (generally weak) extraneous peaks (50% and 30%, respectively) in the CBCA(CO)NH spectrum; i.e. peaks that could not be attributed to intraresidue or sequential C^{α} or C^{β} connections for any GS. FGF-2 showed the largest number (24%) of such ''unaccounted for'' peaks in the CBCANH spectrum. In contrast, Csp A had relatively few unaccounted for CBCANH peaks, but only about half of the expected peaks were actually observed.

Triple-resonance NMR spectra of proteins sometimes reveal more spin systems than are expected from the amino acid sequence, and analysis tools must be able to handle these. Figure 5 compares the number of unassigned GSs to the number that could either be assigned to residues in the sequence or classified by AUTOASSIGN as ''side-chain'' NH groups. For all but NS-1(1-73), the number of assigned GSs is within 1 to 2% of the total number of assignable-residue sites (red line). All of these protein data sets exhibit one or more GSs that could neither be assigned to backbone NH sites nor classified as arising from side-chain NH groups (black portions of histograms in Figure 5). In RNase A and FGF-2, many unassigned GSs arise from proposed minor conformational states (Moy *et al.*, 1995; Shimotakahara *et al.*, 1997); in the Csp A data the extra spin systems result from chemical heterogeneity that is due to slow sample decomposition during the NMR measurements (Feng *et al.*, unpublished results). As indicated by the black line, the number of GSs with degenerate roots (i.e. those with overlap in the ^{15}N-H^{N} dimensions) is also significant in all of these data sets, particularly for RNase A. The num-

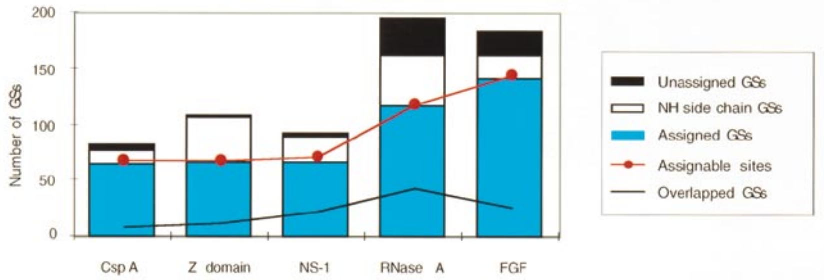


Figure 5. Summary of observed GSs for group I. A GS may be assigned to a residue-specific site, classified as a side-chain NH, or left unassigned. The number of assignable residue sites is calculated as the number of amino acid residues in the sequence excluding the N-terminal residue, less the number of proline residues. Two GSs are considered overlapped if their peaks in the HSQC and/or HNCO spectra are within the user-specified match tolerances in the H^{N} and ^{15}N dimensions. For Csp A, RNase A, NS-1, and Z domain, these tolerances were 0.025 p.p.m. and 0.35 p.p.m. respectively; for FGF-2, tolerances of 0.02 p.p.m. and 0.25 p.p.m. were used.

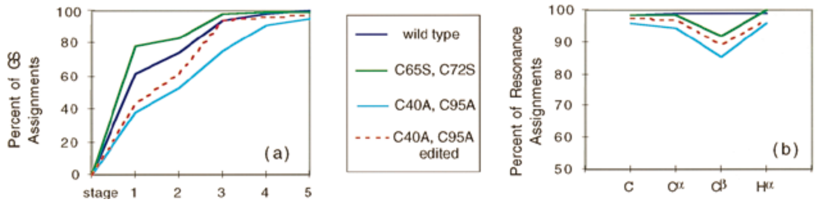


Figure 6. (a) Execution traces for the group II proteins. (b) Percentages of assigned resonances for the group II proteins. As described in the text, two different sets of results are reported for [C40A, C95A]-RNase A. The first of these (continuous blue lines) corresponds to the unedited spectra; the second (broken red lines) corresponds to the edited spectra.

ber of GSs with degenerate N-H roots observed for NS-1(1-73) is comparable to the number for FGF-2, which has more than twice the sequence length of NS-1(1-73).

Analysis of the group II proteins

Having compared the performance of AUTOASSIGN on a set of five different proteins, we next compared the performance on a set of three very similar proteins, wild-type (wt) RNase A and two disulfide mutants, [C65S, C72S]-RNase A and [C40A, C95A]-RNase A. The spectra for [C65S, C72S]-RNase A were comparable with those observed for wt RNaseA, but showed about 30% fewer extraneous spin systems (20 *versus* 29) and slightly less overlap in the ^{15}N-H^N dimensions. These differences led to significantly reduced execution time (one minute, 40 seconds for [C65S, C72S]-RNase A, *versus* three minutes, 53 seconds for wt RNase A). As can be seen from Figure 6(b), however, about 10% fewer C^β chemical shifts were assigned for [C65S, C72S]-RNase A than for the wt protein, due to the increased occurrence of noise peaks in the CBCANH and CBCA(CO)NH spectra of the mutant protein. Overall, for [C65S, C72S]-RNase A AUTOASSIGN obtained 118/119 assignments of assignable sites to GSs, providing 125/127 H^α, 122/124 C^α, 111/121 C^β, and 122/124 C' resonance assignments.

While the performance results for the [C40A, C95A]-RNase A data set (blue traces in Figure 6) are not significantly different from those achieved for wt- and [C65S, C72S]-RNase A, the execution time increased to just under nine minutes, three assignment errors occurred, and only 113/119 residue sites were assigned. The number of unassigned GSs also increased to 52, compared with 29 unassigned GSs in wt and 21 unassigned GSs in [C65S, C72S]-RNase A. As discussed below however, about 24% of the extraneous spin systems identified for [C40A, C95A]-RNase A could be attributed to peak picking artifacts rather than conformational heterogeneity.

AUTOASSIGN also includes interactive features to allow editing of poorly peak-picked or otherwise problematic peak lists. These editing features were used to improve the performance of AUTOASSIGN in the analysis of [C40A, C95A]-RNase A assignments. For example, the following two peaks were picked for a single spin system with an isolated upfield H^N resonance frequency:

Peak no.	H^N	C'	^{15}N	Intensity	Spectrum
132	6.74	178.7	112.8	18635030	HNCO
135	6.74	178.4	112.8	204198896	HNCO

In such cases where a weaker peak (peak 132) could be unambiguously identified as a ‘‘shoulder’’ of a stronger peak (peak 135) resulting from processing or peak-picking artifacts, the weaker peak was manually deleted from the file. Using the software interactively to locate such problems, 16 weak peaks in the HNCO spectrum and five weak peaks in the HSQC spectrum were identified as artifactual duplicates of other stronger peaks, and deleted. In addition, the H^N chemical shift of one GS derived from the [C40A, C95A]-RNase A spectra (later assigned to Met30) was modified by 0.01 p.p.m. to obtain a better alignment with the corresponding peaks in the remaining spectra. When the software was run on the resulting edited [C40A, C95A]-RNase A data the performance was improved significantly (broken red lines in Figure 6); the number of extraneous GSs was reduced from 46 to 35 and AUTOASSIGN assigned 115/119 (97%) of assignable sites to GSs and provided 123/127 H^α, 120/124 C^α, 108/121 C^β, and 121/124 C' resonances. All of these assignments appear to be correct, and are fully consistent with an independent manual analysis that was made using these same triple-resonance data (J. Laity, H. A. Scheraga & G. T. Montelione, unpublished results). In addition, the AUTOASSIGN execution time for the edited data was reduced to approximately six minutes.

The data files input to AUTOASSIGN were created by automatic peak picking with commercial NMR software followed by manual editing procedures that were somewhat different for each data set. This peak-picking and subsequent editing process used no assumptions about the assignments or structure of the protein, and in most cases was done before the assignments were even available. AUTOASSIGN can tolerate a good deal of incompleteness, spurious peaks, peak frequency perturbations, chemical shift degeneracy, and conformational and/or chemical heterogeneity that results in extraneous spin systems in the spectra. However, as demonstrated by the data for [C40A, C95A]-RNase A, as the quality of the peak-picked data deteriorates, there is a point at which performance is compromised. In particular, the quality and completeness of the backbone assignments critically depend on the quality of the HNCO and HSQC spectra and the reliability with which these spectra are peak-picked.

Despite this sensitivity to peak-picking reliability, our results demonstrate that AUTOASSIGN can obtain almost complete and highly reliable assignments of backbone N, C^{α}, C′, H^{N}, H^{α} and side-chain C^{β} resonances from reasonably good quality triple-resonance NMR data. For six of the seven data sets studied, nearly complete backbone resonance assignments were obtained with an error rate of 0.2% from the peak-picked cross-peak files without any user interaction. With the unedited [C40A, C95A]-RNase A peak list, 95% of assignable sites were assigned to GSs, with an error rate of 2.7% (three errors out of 113 sites assigned) and at the expense of significantly increased execution time. Minor interactive editing of two of the [C40A, C95A]-RNase A peak files with tools available within AUTOASSIGN reduced execution time by 30%, yielded assignments for over 97% of assignable sites to GSs, and incurred no errors. These results suggest that while the current implementation of AUTOASSIGN is reasonably robust with respect to peak-picking artifacts, even better performance can be anticipated once software is developed to provide more consistent and reliable peak-list input files.

Discussion

Reliability

For most protein NMR data sets, an exhaustive comparison of all possible sequential assignments is not feasible, as the number of possible solutions increases exponentially with the length of the protein sequence. To reduce the effective search space, AUTOASSIGN combines best-first search with constraint satisfaction methods (Mackworth, 1977; Fox, 1986; Nadel, 1986; Kumar, 1992). The basic idea is that at any given point of execution, the search engine considers only those candidate solutions that are still logically consistent with the current partial solution. The risk in this approach is that initial errors in the partial solution may be propagated through their logical consequences. The most challenging aspects of applying constraint propagation to the general problem of data interpretation involve: (1) ensuring that errors are rarely, if ever, introduced; and (2) minimizing the propagation of errors once they have occurred. In particular, the latter requires careful definition of what constitutes logical inconsistencies. For example, the assumption that the correct solution corresponds to a simple one-to-one mapping of GSs to assignable sites is not, in general, a sound one, as there may be extra and/or missing spin systems.

The results reported here were obtained for data sets used in development as well as testing. Although it is quite impressive that AUTOASSIGN works well on these seven proteins, new data sets may present problems that are not yet addressed by the software. In particular, the situations arising in these seven protein data sets (e.g. severe N-H degeneracy and conformational/chemical heterogeneity) were used to define the conditions under which certain constraints are applicable. Robust methods of handling overlapped and/or extraneous spin systems were developed for these data, and should generalize well to new data sets. However, it is conceivable that unforeseen complications may arise in other proteins that contradict various underlying assumptions.

Reliability also depends to some extent on the user-specified match tolerances in the backbone amide ^{1}H-^{15}N dimensions (for compiling spin systems) and in the ^{1}H-^{13}C dimensions (for establishing sequential links). These in turn depend on the quality and resolution of the experimental data. Default tolerances of 0.025 p.p.m. and 0.35 p.p.m. in the H^{N} and ^{15}N dimensions worked well on all of the data sets tested to date. However, in our experience, regions of some proteins exhibit resonance frequencies that are very sensitive to sample temperature and other conditions. Larger tolerances are less likely to omit critical peaks for spin systems occurring in these regions of the protein, but may introduce errors and/or increase execution time. On the other hand, the use of smaller H^{N} and ^{15}N match tolerances may significantly reduce the number of assignments obtained. Related problems occur with the match tolerances used to establish sequential links. For the test cases reported here, typical match tolerances were 0.5 p.p.m. for C^{α}, 0.5 p.p.m. for C^{β}, 0.25 p.p.m. for C′, and 0.05 p.p.m. for H^{α} dimensions, respectively.

Comparison with other methods

Several alternative implementations for automated assignment of protein NMR spectra have employed optimization algorithms that attempt to minimize a pseudo-energy function or to maximize some measure of ‘‘goodness of fit’’. These have included neural networks (Hare & Prestegard, 1994), simulated annealing (Bernstein *et al.*, 1993; Kraulis, 1994; Morelle *et al.*, 1995), and genetic algorithms (Wehrens *et al.*, 1993). The assignment process is thus mapped to a global optimization problem with the potential of becoming trapped in local objective-function minima. Additional disadvantages of global optimization methods such as these stem from their reliance on a global objective function that assesses only alternative complete assignments. As a result, the solution is biased towards a complete set of ''acceptable'' assignments in preference to obtaining an incomplete set that may be of higher quality. Also, exploratory tools for incremental analysis of partial assignments are not easily supported. Finally, in cases where the complete solution is under-constrained, it may be difficult to extract a reliable partial assignment from a complete assignment obtained by global optimization. While it is true that individual residue-specific scores may be used to interpret the results, these methods may not distinguish ''weak'' assignments that can be obtained reliably by logical processes of elimination from those that are truly unreliable.

Conversely, an advantage of numerical optimization methods over best-first search strategies (Nilsson, 1980), like those used in AUTOASSIGN, is that the former may be more tolerant of spurious, contradictory data, while the latter may be somewhat ''brittle'' in its interpretations. AUTOASSIGN, for example, requires that rigorous uniqueness and matching criteria be satisfied in order to progress towards a solution in the earliest stages of analysis. Thus, if the data are so severely compromised that no possible links or assignments can satisfy the initial criteria, AUTOASSIGN may never progress to the later stages where some of these requirements are relaxed. A second problem is that most best-first search implementations progressively refine the single most promising partial assignment and early errors may be difficult to recover from. In contrast, numerical optimization methods by definition explore numerous complete, alternative assignments. For these reasons, it is possible that with increasing noise and degradation of the data, numerical optimization methods that allow more ''probabilistic interpretations'' may prove more effective than best-first strategies like those used in the current implementation of AUTOASSIGN.

Other implementations of systems for automated analysis of triple-resonance spectra (Friedrichs *et al*., 1994; Meadows *et al*., 1994; Olsen & Markley, 1994) more closely resemble AUTOASSIGN's approach, in that a form of best-first matching is utilized to establish sequential links. None of these, however, combines its search strategies with constraint satisfaction, and spin-system type information plays a considerably less important role. In contrast, the software described by Billeter *et al*. (1988) for the analysis of homonuclear spectra has a strong constraint satisfaction component but lacks any heuristic search strategy. It is not possible, however, to compare the completeness and reliability of results generated by these various implementations, as each has been tested on considerably different experimental data and input options.

Generality and possible extensions

While it would be ideal to develop a system that could accommodate whatever experimental data are available, this goal is probably too ill-defined for an effective computational implementation. In the artificial intelligence literature, systems that attempt to model general problem-solving behavior such as this are said to use ''weak methods'' (Laird & Newell, 1983) , as they impose minimal assumptions on the types of problems to which they may be applied. At the other end of the spectrum, problem-specific ''strong AI'' systems place stringent restrictions on the types of problems they can solve and are generally more efficient, as their methods of problem-solving have been tailored to those problem types. The current version of AUTOASSIGN uses ''strong AI'', as it has been designed to utilize a specific resonance assignment strategy based on the specific types of experimental data described in Figure 1.

Although it may be possible to obtain reliable assignments using various subsets of the triple-resonance spectra indicated in Figure 1, using the current implementation of AUTOASSIGN the solution will often be underconstrained by a lack of sufficient connectivity information. Rather than force the system to infer unreliable assignments from such data, our approach has been to develop reliable inference methods that perform consistently with data from the eight or nine NMR spectra described here. However, AUTOASSIGN does provide the user with interactive analysis tools to support incremental matching and constraint propagation with less complete data sets.

The experienced spectroscopist brings to the analysis a profound understanding of the experimental data, and in general, uses a model of how the various spectra relate to one another to construct flexible, but effective, algorithms for the interpretation of these data. Attempts to implement the spectroscopist's most general problem-solving behavior, however, are likely to lead to more general but less effective problem solving. A more reasonable goal is to define a subset of the possible spectral inputs that can be composed in alternative ways to provide sufficient information for the sequential assignment problem. In order for the software to then properly interpret these data, a model of how the various spectra relate to each other as well as a deeper computational model of the sequential assignment problem itself is also required. One of our current focuses is on defining a more general set of spectral input that can support increased flexibility in AUTOASSIGN's interpretive power without significantly sacrificing the reliability and efficiency that has been achieved through specialization. Additional types of NMR data that could be handled by fairly simple extensions of AUTOASSIGN algorithms include: (1) phase data (i.e. positive and negative intensities) and other methods for editing triple-resonance spectra to provide classification of amino acid residue types (Grzesiek & Bax, 1993; Tashiro *et al.*, 1995; Dötsh & Wagner, 1996; Dötsch *et al.*, 1996; Feng *et al.*, 1996; Ríos *et al.*, 1996); (2) HCACO-type (Ikura *et al.*, 1990; Kay *et al.*, 1990; Dijkstra *et al.*, 1994) and CBCA-CO(CA)HA-type (Kay *et al.*, 1992a; Kay, 1993) spectra, as higher sensitivity replacements for HNCACO spectra (Clubb *et al.*, 1992), that currently provide source data for the identification of intraresidue carbonyl resonance frequencies; (3) HBHANH (Wang *et al.*, 1994), HBHA(CO)NH (Grzesiek & Bax, 1993), HCCNH-TOCSY (Logan *et al.*, 1992; Clowes *et al.*, 1993; Lyons & Montelione, 1993) and/or HCC(CO)NH-TOCSY (Logan *et al.*, 1992; Montelione *et al.*, 1992; Grzesiek *et al.*, 1993; Lyons *et al.*, 1993) spectra, providing sequential matching of chemical shifts associated with more peripheral nuclei; (4) 3D ^{15}N-edited NOESY data to confirm proposed backbone assignments.

Other generalizations to AUTOASSIGN's current methods that we are considering include: (1) the development of more robust tools for the unfolding of spectra recorded at narrower sweepwidths; and (2) methods of distinguishing C^{α} from C^{β} chemical shifts in the CBCANH and CBCA (CO)NH spectra directly (Grzesiek & Bax, 1992a,b) without recourse to comparisons with CANH or CA(CO)NH-type data as is done in the current implementation.

In addition to extending the software to accept alternative types of input, additional analysis tools being considered include: (1) automated analysis of complete side-chain resonance assignments; (2) applications involving uniformly ^{2}H, ^{13}C, ^{15}N-enriched proteins (Grzesiek *et al.*, 1995; Yamazaki *et al.*, 1995); (3) analysis tools that can exploit the known chemical shifts of proteins with homologous structures (Bartels *et al.*, 1996); (4) integration of alternative methods of assignment (for example, simulated annealing); (5) development of an interface between the table of chemical shift assignments output by AUTOASSIGN and commercially available software capable of generating spectral strip plots; (6) NOESY spectra assignment tools.

Conclusions

Our current approach is a generalization of our earlier work (Zimmerman *et al.*, 1993, 1994) which used a similar constraint propagation network together with 3D HCCNH-TOCSY and HCC(CO)NH-TOCSY data (Logan *et al.*, 1992; Montelione *et al.*, 1992; Lyons & Montelione, 1993; Lyons *et al.*, 1993) for determining sequence-specific assignments. The present implementation demonstrates the utility of the underlying model of constraint satisfaction for the sequential assignment problem and is the first step towards a more general model of expert problem solving in this domain. As implemented, the current prototype system provides extremely useful tools for analysis of the types of NMR data illustrated in Figure 1. For proteins containing as many as 154 amino acid residues that yield reasonably good quality triple-resonance NMR data sets, AUTOASSIGN provides almost complete backbone N, C, and H and many side-chain C^{β} resonance assignments in under ten minutes. The system handles many difficult situations that challenge the human expert, including extra spin systems due to chemical or conformational heterogeneity, severe overlap in the N-H and aliphatic carbon dimensions, and missing spectral information. Given minimal editing of automatically peak-picked data, AUTOASSIGN reduces the backbone resonance assignment process to the 7 to 21 days of NMR instrument time needed for recording the requisite data.

Methods

Input specifications

All of the peak lists analyzed by AUTOASSIGN are in the form of ASCI text files listing the peak coordinates (in p.p.m.) and intensities. Processing of the NMR spectra was done using VNMR (Varian Associates), NMRPipe (Delaglio *et al.*, 1995), or Felix (Molecular Simulations Inc.) programs, with automated peak picking using tools provided by NMRCompass (Molecular Simulations Inc.), Felix (Molecular Simulations Inc.), or PIPP (Garrett *et al.*, 1991). For most spectra, an initial list of automatically picked peaks was generated for each 2D and 3D spectrum using intensity and linewidth filters. This list was then edited manually to identify and eliminate extraneous peaks, using interactive graphics and various general features such as the approximate expected number of peaks, the visual quality of alignment across spectra, and peak shape criteria. However, as no general specifications for peak picking were given to the users, the user-defined criteria for manual editing of peak-picked spectra varied considerably. The interactive manual editing required about one hour per 3D NMR data set (i.e. about ten hours for the complete set of spectra), and can be carried out while data collection is in progress, adding little to the total time required for the complete process of determining backbone resonance assignments.

The types of spectral data used as input for AUTOASSIGN are described in Figure 1. In most cases, the required NMR data sets were generated using the following pulsed-field gradient (PFG) NMR pulse sequences: (1) PFG - ^{15}N - H^{N} - HSQC (Kay *et al.*, 1992b; Li & Montelione, 1993); (2) PFG-H(CA)(CO)NH (Boucher *et al.*, 1992; Feng *et al.*, 1996); (3) PFG-CA(CO)NH (Boucher *et al.*, 1992; Feng *et al.*, 1996); (4) PFG-CBCA (CO)NH (Grzesiek & Bax, 1992a; Ríos *et al.*, 1996); (5) PFG-HNCO (Muhandiram & Kay, 1994); (6) PFG-H(CA)NH (Montelione & Wagner, 1990; Feng *et al.*, 1996); (7) PFG-CANH (Montelione & Wagner, 1989; Feng *et al.*, 1996); (8) PFG-CBCANH (Grzesiek & Bax, 1992b; Ríos *et al.*, 1996); and (9) PFG-HNCACO (Clubb *et al.*, 1992). Additional flexibility and control of the interpretation of these peak lists is provided by a specifications table (Table 1). This table allows the user to specify for each spectrum important parameters such as the detected atom types and expected chemical shift ranges in each dimension, any absolute referencing corrections

Table 1. Table of user-defined spectral parameters for input to AUTOASSIGN

i	$i-1$	X	l	u	n	Y	l	u	n	Z	l	u	n	Ref.	File name
1	0	HN	4	10	0	nil[a]				N	100	135	0	nil[a]	hsqc.pks
0	1	HN	4	10	0	C′	170	180	0	N	100	135	0	hsqc	hnco.pks
1	1	HN	4	10	0	C^α	40	70	0	N	100	135	0	hsqc	hnca.pks
0	1	HN	4	10	0	C^α C^β	10	70	0	N	100	135	0	hsqc	cbcaconh.pks
0	1	HN	4	10	0	C^α	40	70	0	N	100	135	0	hsqc	hncoca.pks

The first two columns indicate whether or not the spectrum includes intraresidue (i) and/or sequential ($i-1$) information. For two- and three-dimensional spectra, the next three sets of four columns specify the nuclei detected in each dimension (X, Y, Z), the lower (l) and upper (u) bounds of the corresponding frequency axes (in p.p.m.), and a global reference correction (n, in p.p.m.) for each dimension of each spectrum. The column labeled Ref. specifies which spectrum should be used as the source spectrum in the N-H dimensions, and the last column indicates the file name of the peak list.

Each peak file that is to be used as input must have an entry in the specifications table; this example is a partial table in that only five of the eight or nine spectra that define AUTOASSIGN's standard input are specified.

[a] Values of nil indicate that the information associated with that field is not applicable to the spectrum occurring in that row. In this example, the 2D HSQC spectrum (hsqc.pks) is used as the source for referencing all of the remaining spectra in the N and HN dimensions, so the Y-dimension nucleus and Ref. fields are not applicable.

that may be required, the type of interactions detected (intraresidue and/or sequential), and a reference spectrum to be used for alignment in the ^{15}N and H^N dimensions.

Overview of AUTOASSIGN's strategies

The model of problem solving used by AUTOASSIGN can be viewed as a generalization of the basic procedure developed by Wüthrich (1986) and co-workers for the analysis of homonuclear spectra. First, generic spin systems (GSs) are identified by mapping the 3D cross-peaks in the various target spectra to the corresponding peaks in the [^{15}N-^{1}H]-HSQC source spectrum. Given these mappings, the CA and CO-ladders of these GSs are then defined by designating the corresponding intraresidue and sequential C^α, C^β, C′, and H^α chemical shifts, respectively. The characteristic C^α and C^β shifts associated with different residue types are then used to obtain residue-type probability scores, which are in turn used to define the set of residue types consistent with each CA and CO-ladder. Next, sequential connectivity information is used to establish adjacency relations between GSs i and $i+1$. As these sequential links are established, the sequence is scanned to determine if the type information for that pair of linked GSs defines a unique pair of residue sites in the sequence. If so, the assignments are made; otherwise, the possible assignments of the linked pair are noted. In particular, links between unassigned GSs are not confirmed until these GSs have been assigned to specific residue sites in the protein sequence. These last three steps form a general description of the constraint-based match cycle, depicted in Figure 2.

Prior to executing any of the constraint-based match stages, a set of initialization routines are invoked, which: (1) define a list of sequence-specific sites (SSs) corresponding to each residue in the protein sequence; (2) instantiate amino acid prototypes specifying the expected resonance frequencies and standard deviations for these SSs; (3) apply chemical shift reference corrections to improve the "between-spectra" alignments; and (4) compile a list of generic spin systems (GSs) inferred from the triple-resonance spectra. Figure 7 illustrates schematically the relationships between these data structures and the objects in AUTOASSIGN's internal representation.

Initialization of SSs and prototypes

Each residue in the sequence is initialized as a sequence-specific site (SS) with an (initially empty) list of possible GS assignments and pointers to the preceding and following residue sites in the sequence. For each newly encountered amino acid type, a type-specific prototype is also created, which the SS can then query to obtain its expected C^α and C^β resonances and standard deviations. In the event that the current residue is another instance of a previously defined prototype, the corresponding SS is simply added to that prototype's list of instances. Table 2 shows the internal structure of an Ile prototype generated during the analysis of Z domain. The expected values and standard deviations of the prototypes' C^α and C^β resonances are stored in AUTOASSIGN's class definitions of amino acid prototype objects.

Spectral referencing

Before individual peaks can be mapped to GSs, the various spectra must be referenced with respect to one another in order to minimize the match tolerances used in subsequent stages of analysis. The first step in aligning the spectra in the amide N-H dimensions requires defining reference peaks in the source spectra. Each such peak must occur in a relatively isolated region in the ^{1}H,^{15}N dimensions of the source spectrum, and corresponding peaks should be observed in all of the target spectra. Each target spectrum may specify its own source spectrum (see column 15 of Table 1); in most cases this is the HSQC spectrum. Referencing of a particular target spectrum is then achieved by uniquely mapping the peaks in that spectrum to the source's defined reference peaks, computing the average differences in the ^{15}N and H^N dimensions, and using these differences as global reference correction factors for the target spectrum.

Redundancies in the information contained in pairs of spectra can be used to align selected spectra in the ^{13}C and ^{1}H dimensions as well. For example, if spectrum A includes only sequential peaks carrying information about S_{i-1} for a selected atom type, while spectrum A′ includes both intraresidue (S'_i) and sequential (S'_{i-1}) information for that atom, then the average difference, $\langle S_{i-1} - S'_{i-1} \rangle$, can be applied as a global reference correction to the S dimension of all peaks in A′. Alternatively, if A includes only information about C^α frequencies while spectrum A′ includes information

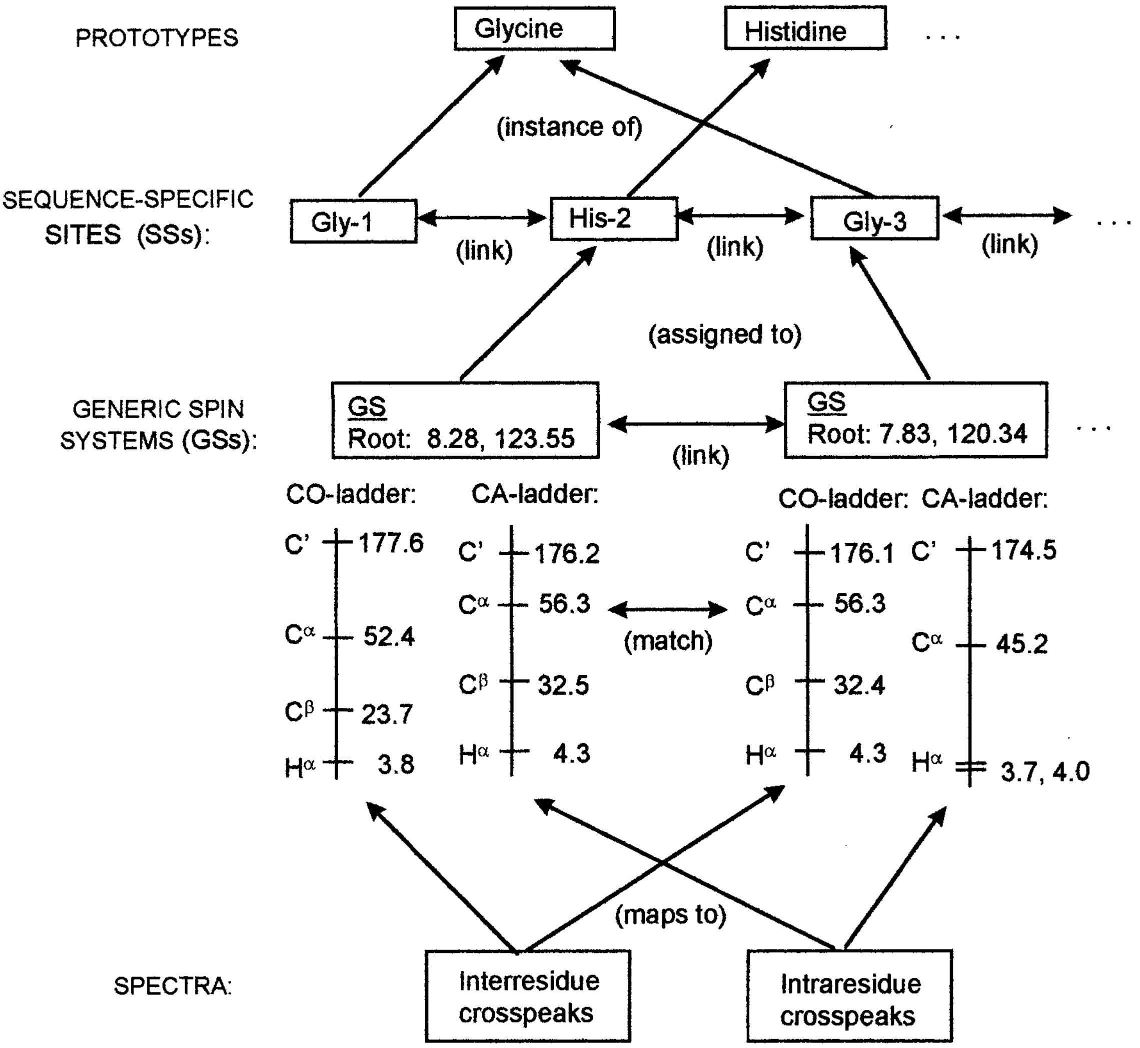


Figure 7. The relationships between objects in AUTOASSIGN's internal representations.

about both C^{α} and C^{β} frequencies, then spectrum A′ can again be referenced with respect to A, in this case using $\langle C^{\alpha} - C^{\alpha'} \rangle$. These relations between spectra are inferred by AUTOASSIGN from the specifications table (Table 1), using columns 1, 2, and 7.

Spin system compilation

Generic spin systems (GSs) are initialized using the HSQC and HNCO spectra. Each N-H cross-peak in the 2D HSQC spectrum is initially interpreted as the backbone root of a GS. All cross-peaks in the remaining 3D spectra whose ^{15}N and H^{N} chemical shift values fall within the specified match tolerances of these root coordinates are then added to each GS. The HNCO spectrum is treated specially, however, as it is of comparable sensitivity to the HSQC spectrum and provides separation in the carbonyl dimension of GSs that are overlapped in the N-H dimensions. Accordingly, AUTOASSIGN maps only the closest HNCO peak (using normalized Euclidean distance in the N-H dimensions) to each GS having an HSQC root. Subsequently, the HNCO spectrum is scanned to identify peaks that are not yet included in any GS and each of these is initialized as the root of a new "HNCO-rooted GS". Cross-peaks in the remaining 3D spectra whose ^{15}N and H^{N} chemical shift values fall within the tolerances are again mapped to these new GSs whose roots are defined in the HNCO spectrum. Because AUTOASSIGN considers only those peaks in the 3D spectra with ^{15}N-H^{N} frequencies similar to those of "root peaks" in the HSQC and HNCO spectra, many spurious artifactual peaks that are present in these data are effectively filtered out. By definition, GSs with very similar ^{15}N-H^{N} shifts may include some of the same peaks from the other 3D spectra. These GSs are considered "overlapped" or "degenerate".

Identification of side-chain NH resonances

The next step in spin-system compilation involves the identification of ^{1}H-^{15}N correlations associated with side-chains. In many cases, cross-peaks arising from indole and amide side-chain nuclei have attenuated intensities (in properly tuned triple-resonance spectra) and will not be included in the specified peak lists. When they are detected, GSs arising from side-chain NH groups usually have less than three peaks in the set of triple-resonance spectra, and AUTOASSIGN classifies these as side-chain

Table 2. An example Ile prototype for Z domain

Prior[a]	3/71
Instances[b]	(Ile −11, Ile16, Ile31)
$C^{\alpha}C^{\beta}$-stats[c]	(60.9, 2.6, 39.4, 2.5)
Symbol[d]	I

[a] Prior is the number of instances of this amino acid in the sequence, divided by the sequence length.

[b] Instances is a list of pointers to the SSs created for these residues.

[c] $C^{\alpha}C^{\beta}$-stats lists the expected C^{α} chemical shift, the standard deviation for C^{α}, the expected C^{β} resonance, and the C^{β} standard deviation for the specific prototype.

[d] The symbol associated with a prototype is its one-letter amino acid code.

GSs. While this method may occasionally misclassify some weaker spin systems associated with backbone NH groups as side-chains, the justification is that such backbone GSs contain insufficient information for generating reliable sequential assignments and are better left out of the backbone resonance assignment process. This method does not, however, consider spin systems associated with side-chain NH groups that do include a significant number of cross-peaks in these triple-resonance spectra; additional methods are used to identify specifically some of these guanido and amide side-chain NH groups.

One such method of identifying side-chain NH spin systems uses an HSQC spectrum of larger sweep width (e.g. 60 to 140 p.p.m. in the ^{15}N dimension) to identify arginine guanido N-H cross-peaks on the basis of their characteristically upfield ^{15}N chemical shifts. The corresponding peaks in the original HSQC spectrum are then located by calculating the folded frequency in the ^{15}N dimension. Additional side-chain NH_2 correlations are identified automatically using pattern matching to the expected profiles of Asn and Gln side-chain ^{13}C frequencies, and removed from the list of backbone GS roots.

Separation of overlapped GSs

Significant ambiguity can arise when two or more GSs have identical (or very similar) backbone amide N-H chemical shifts, and several methods are employed to determine how peaks should be associated with degenerate GS roots. The simplest method of crosspeak allocation for overlapped GSs uses a nearest neighbor approach. This method is most effective for GSs with only marginal N-H overlap and generally reduces the extent of ambiguously mapped peaks by a factor of only 2. In cases where "safe" decision boundaries cannot be defined due to more severe N-H overlap, the same cross-peaks may still be included in the specifications of two or more GSs. The deconvolution of these more severely overlapped GSs is achieved using designated shifts on CA- and CO-ladders of assigned GSs, as explained below (in Extend Assigned Segments, stage 3).

Identification of weak spin systems (minor conformers)

The final step in spin-system compilation assesses the number of backbone GSs (as opposed to spin systems associated with side-chain NH groups) in order to determine if there is an excess of GSs resulting from conformational and/or chemical heterogeneity. If so, the weakest of the backbone GSs are set aside for subsequent analysis. The number to be set aside is determined as a function of n, the number of assignable sites in the sequence. Specifically, AUTOASSIGN calculates $k = 0.9n$, and initially focuses on only the k strongest observed spin systems, with the remaining designated as "weak". For example, for wt RNase with 150 backbone GSs and 119 assignable sites, the 42 weakest GSs are initially set aside as "weaker" spin systems.

Clearly, the spin systems associated with major and minor protein species are not separable on the basis of intensity alone; this 90% cutoff was determined empirically in order to minimize false negatives without sacrificing too many false positives or overfitting the test data. In subsequent stages of analysis, however, GSs initially classified as "weak" may be assigned to the major chemical/conformational species, while others initially classified as "strong" may not be assigned.

Iterative constraint-based matching

The Constraint-based Match Cycle described in Figure 2 is not a concrete software entity, but rather refers to a collection of loosely coupled routines that are combined differently according to the stage of analysis. Specific parameters (or arguments) given to a general-purpose sequential "match" routine specify a set of CA-ladders, a set of CO-ladders, a minimum match threshold, an incremental step size, and a number of iterations. In "unrestricted" matching, all CA and CO-ladders are matched against one another. In the initial stages, however, only the most complete ladders (e.g. those containing C^{α}, H^{α}, C^{β}, and C' rungs) are allowed to participate. Only the final stages of analysis resort to unrestricted matching. The sequential match routine is invoked many times throughout execution and is coupled to a set of constraint propagation methods that are invoked with each match between ladders confirmed as a definite link between GSs. As these propagation methods may in turn lead to site-specific GS assignments, a second set of constraint propagation methods that are triggered by these assignments may also be invoked.

Three different methods of chemical shift designation are described below. The first of these is used in stages 1, 2, and 4 (Figure 2) initially to construct and subsequently to refine the CA and CO-ladders directly from the associated spectra. The second method is used in stage 3 to further refine incompletely specified ladders using the chemical shift profiles designated by assigned GSs lacking predecessors or successors. The third method is used in the last stage of constraint-based matching to correct possible discrepancies in the chemical shifts designated by the first two methods.

Initial construction and matching of CA- and CO-ladders (stage 1)

AUTOASSIGN uses the expected redundancies between the different spectra to guide the atom-specific designation of chemical shifts, which define the CA- and CO-ladders. The H^{α}_{i-1}, C^{α}_{i-1}, and C'_{i-1} chemical shifts of each GS's CO-ladder can usually be determined directly from the HA(CA)(CO)NH, CA(CO)NH, and HNCO data, respectively. For most residue types, the C^{β} frequency occurs in a range easily distinguished from the C^{α} frequencies. Thus for most GSs, C^{β}_{i-1} can also be determined unambiguously from the CBCA(CO)NH data. However, because several residue types (e.g. Thr, Ser, Leu) have characteristic C^{β} frequencies that are not easily distinguished from C^{α} frequencies, AUTOASSIGN uses the CA(CO)NH spectrum to filter out C^{α} peaks in the CBCA(CO)NH data and to designate the C^{β}_{i-1} frequencies.

Similarly, the designation of intraresidue shifts on a GS's CA-ladder involves filtering out sequential correlations that are sometimes observed in the intraresidue spectra†. In particular, the CBCANH spectrum must be "doubly filtered" to remove both sequential as well as C^{α} frequencies in the designation of intraresidue C^{β} resonances. For most spin systems, these methods of filtering uniquely define the chemical shift information required to construct the CA- and CO-ladders.

Three situations can arise, however, that lead to only partially specified CA- and/or CO-ladders. In instances where the H^{α}, C^{α}, C^{β}, or C' chemical shifts of residues i and $i+1$ are very similar, the true intraresidue correlations are not easily distinguished from the sequential correlations. In these cases the associated frequencies of the CA-ladder are left undesignated until later stages of processing where such "intraresidue-sequential degeneracy" is allowed. The second situation involves degenerate GSs with identical (or nearly identical) ^{15}N and H^{N} resonance frequencies, where several peaks may be candidates for the H^{α}, C^{α}, C^{β}, and/or C' frequencies of the GSs' ladders. In these cases, the corresponding resonance shifts on the CO and CA-ladders are also left undesignated. Similar complications arise with non-degenerate GSs that include spurious peaks of comparable intensity to the "real" peaks. Here again, incompletely specified CA and CO-ladders may be defined. Conversely, if the observed spurious peaks are of relatively low intensity, simple "intensity filters" are applied to extract the appropriate resonance frequencies. These methods of chemical shift designation are initially applied uniformly to all GSs. During stage 1, however, only the ladders associated with "stronger" GSs are matched against one another.

Refinement and further matching of CA and CO-ladders (stage 2)

After the first cycle of constraint-based matching has completed, AUTOASSIGN reevaluates the incompletely specified CA and CO-ladders. In this second pass (stage 2), the sequential filter is removed, allowing for overlap between intraresidue and sequential cross-peaks. This method of designating chemical shifts is essentially identical to that used for the initial construction of ladders, the only difference being that the "uniqueness constraint" is relaxed. On average, 5 to 10% more shifts are designated to CA and CO-ladders, and iterative constraint-based matching concludes the second stage of analysis.

Extending assigned segments (stage 3)

Stage 3 of AUTOASSIGN's default execution sequence (Extend-Assigned-Segments) is the only place where a truly exhaustive search of the remaining solution space is applied. At this point, 85% or more of the assignments and links have usually been established and the designated chemical shifts of assigned GSs are now used to guide the specifications of the CA and CO-ladders of the remaining unassigned GSs. Specifically, the CO-ladders of GSs occurring at the N termini of assigned segments are used to delineate the expected values of "missing" CA-ladders. All peaks associated with the unassigned GSs are examined for the possibility of construction of CA-ladders that match these expectations. If only one unassigned GS can be associated with an N-terminal CO-ladder, then these peaks are designated on the CA-ladder for that GS, the link is established, and the assignment is made. Similarly, the designated CA-ladder frequencies of C-terminal GSs occurring in assigned segments are used to define the chemical shifts for "missing" CO-ladders. All unassigned GSs are considered during this stage, including those previously set aside as weak and those identified as "overlapped" in the N-H dimensions. Thus it is possible to reinstate and assign a weaker spin system prior to stage 4 (below), but only in those cases where the weaker GS has associated peaks that provide a unique, high quality match to an expected profile generated from some assigned GS.

Matching weak spin systems (stage 4)

The first three stages of analysis exhaustively mine the data for all inferences that can be made with relatively high confidence, initially focusing on the "strongest" spin systems. The fourth stage begins by reinstating the remaining unassigned "weaker" GSs to the pool of unassigned GSs. A cycle of iterative constraint-based matching is again invoked, this time allowing the weaker GSs to compete with the stronger GSs for available links and sequential assignments. The same methods of refining the CA and CO-ladders of GSs used in stage 2 (above) are applied to the weaker GSs during stage 4 (Match-Weaker-Spins) prior to re-invoking the generic match routines. The CA and CO-ladders at the ends of linked segments of GSs also provide information that is used to pull apart the multiple CA (and CO-) ladders of overlapped GSs.

Completing the assignments (stage 5)

The last stage of constraint-based matching begins by scanning the current set of designated chemical shifts to identify any obvious errors. One kind of error detected at this stage derives from the fact that cross-peaks in the 3D CBCANH experiments arising from sequential connectivities can sometimes be stronger than those arising from intraresidue connectivities. An example scenario involves an unassigned GS_i whose correct, but unknown, assignment is to SS_i. Problems can arise if the intraresidue C^{β}_i–N_i–H^{N}_i cross-peak is too weak to be detected in the CBCANH experiment, while the sequential C^{β}_{i-1}–N_i–H^{N}_i cross-peak is detected in this same experiment. As a result, the chemical shift C^{β}_{i-1} may be incorrectly designated as the C^{β}_i chemical shift on the CA-ladder of residue i. If this erroneous C^{β} chemical shift differs significantly from the expected range of C^{β} resonance frequencies for residue i, the correct SS assignment may have been ruled out as being of incompatible type. In stage 5, AUTOASSIGN looks for potential sites of such errors. In such cases the C^{β} shift designation of the CA-ladder is retracted, the possible SS assignments for this GS are reinitialized, and the constraint-based match cycle is then reinvoked.

Finally, AUTOASSIGN considers additional assignments that may be obtained by a process of elimination. For example, at this point in processing the NMR data

† Sequential cross-peaks are also observed in these "intraresidue" triple-resonance spectra (Montelione & Wagner, 1989, 1990; Ikura *et al.*, 1990), but can be identified through comparisons with the corresponding "sequential" triple-resonance data.

for Z domain, all but two GSs and two assignable sites (His −4 and Lys58) have been assigned†. Neither of these GSs is consistent with the CO- and CA-ladders of GSs assigned to residues Asp −3 and Gln −5, respectively. The C-terminal residue of the sequence is adjacent to a proline, Pro57. As only one of the unassigned GSs has a CO-ladder consistent with a proline and a CA-ladder consistent with a lysine, the assignment of this GS to Lys58 is made.

Generic routines for all stages of analysis

The following descriptions apply to routines used throughout execution. These include methods of calculating and updating residue type probability scores, calculating match scores between ladders, establishing sequential links between GSs, and methods of constraint propagation.

Calculation of residue type probabilities

Whenever the C^{α} and/or C^{β} chemical shifts of a CA or CO-ladder are initialized or revised, the residue-type probability scores associated with that ladder are computed as follows. First, the Bayesian class posterior probability (Duda & Hart, 1973) is calculated using the expected C^{α} and C^{β} chemical shift values and standard deviations stored with the amino acid prototype. The class posterior probability calculated for an observed pair of C^{α}/C^{β} chemical shifts with respect to amino acid residue-type R is computed as:

$$p(R \mid C^{\alpha},\ C^{\beta}) = p(C^{\alpha},\ C^{\beta} \mid R)P(R)/\Sigma_R p(C^{\alpha}, C^{\beta} \mid R)P(R) \tag{1}$$

where $p\ (R \mid C^{\alpha}, C^{\beta})$ is the probability that residue-type R has occurred given the observed chemical shift values C^{α} and C^{β}, $p\ (C^{\alpha}, C^{\beta} \mid R)$ is the probability of observing chemical shift values C^{α} and C^{β}, given residue type R, and $P(R)$ is the frequency of occurrence of residue-type R in the protein sequence. The two variables (C^{α}, C^{β}) are assumed to be normally distributed and independent, with the means and standard deviations for each residue type calculated from a database of over 1400 residues (Seavey *et al.*, 1991; Wishart *et al.*, 1991). Class posterior probabilities can be thought of as class conditional probabilities weighted by the class prior probability, $P(R)$, and normalized by the unconditional probability of the variable(s). This approach differs from a similar classificaton scheme described by Grzesiek & Bax (1993), which assumed identical standard deviations in ^{13}C chemical shift values for all amino acid residue-types and did not take into consideration $P(R)$.

Although probability scores are initially computed for all residue-types occurring in the sequence, only as many as are required to achieve a sum of 0.99990 are included in considering the possible residue types associated with a CA or CO-ladder. With this criterion, a given (C^{α}, C^{β}) chemical shift pair is assigned to, on average, seven possible residue types. These methods of probabilistic classification were verified using eightfold cross-validation on the original database; the correct classification was never omitted.

Establishing sequential links and constraining assignments

Sequential links between GSs are derived from the best type-consistent matches that can be established between the CA-ladder of one GS and the CO-ladder of a second GS using tight match criteria. If the potential link is both consistent with the sequence and "unique", it is established. A link between CA-ladder (i) and CO-ladder (j) is "unique" if there are no other high quality matches of CA-ladder (i) to some other CO-ladder and no other high quality matches of CO-ladder (j) to some other CA-ladder. The match score between CA-ladder (i) and CO-ladder (j) is calculated as $M_{ij} = e^{-d(i,j)}$, where $d(i,j)$ is computed as a Euclidean distance. Specifically, for each ladder, a vector of normalized values is computed:

$$v_i = \left\langle \frac{C'_i - \mu(C')}{s(C')},\ \frac{C^{\alpha}_i - \mu(C^{\alpha})}{s(C^{\alpha})},\ \frac{C^{\beta}_i - \mu(C^{\beta})}{s(C^{\beta})},\ \frac{H^{\alpha}_i - \mu(H^{\alpha})}{s(H^{\alpha})} \right\rangle \tag{2}$$

where the means (μ) and standard deviations (s) for each chemical shift dimension are computed over all designated chemical shifts for that atom type. The "distance" between two ladders is then computed as the Euclidean distance between their associated vectors, where the dimension of each vector corresponds to the number of chemical shifts (or rungs) that are common to the two ladders for which the match score is calculated. Each such distance is greater than or equal to zero. As a result, $0 < M_{ij} \leqslant 1$ for all possible matches, with higher scores reflecting higher quality matches. In the event that a match involves incompletely designated chemical shifts for one or both of the two ladders, the missing dimensions are masked as zeros on both ladders. To minimize possible errors arising from such "incomplete" ladders, however, AUTOASSIGN focuses on matching the most complete ladders first, and only establishes a link when the corresponding match score is the best observed for both ladders. In addition, throughout the first four stages of execution, all links require a minimum of two-dimensional matching.

The highest scoring, most complete matches are established first and the sequence is scanned to identify all possible assignments of these new segments as they are defined. As each new GS is added to a linked segment, the set of possible assignments of all the participating GSs are constrained to be mutually consistent with their adjacencies in the segment and corresponding stretches in the sequence. If the number of possible mappings of a segment of linked GSs is reduced to one, the associated sequence-specific assignments are made. Alternatively, if no mappings are possible, the inferred sequential links are discarded. Finally, if a number of alternative mappings are defined for a segment, the possible assignments are noted and the segment is added to a list of unassigned segments.

The constraint propagation network

The "constraint propagation network" (Zimmerman *et al.*, 1994) is a set of tightly coupled routines that are triggered each time a sequential link or GS assignment is established. For example, when a stretch of residues is assigned to a particular segment of linked GSs, these

† The first 14 amino acid residues in Z domain are a leader sequence and are designated with negative residue numbers; for example, residue His −4 is the fourth residue from the C-terminal end of this leader sequence.

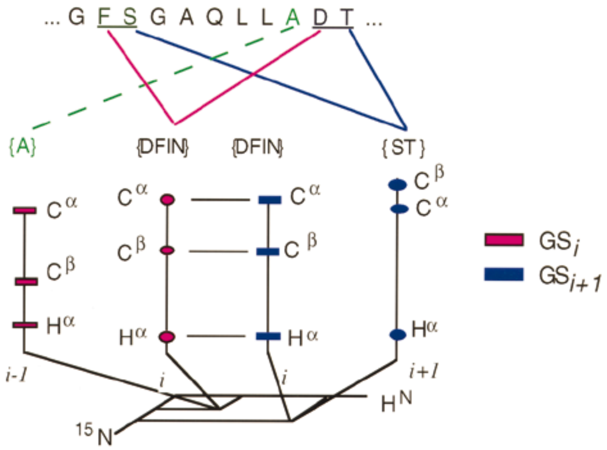


Figure 8. An example of constraint propagation in AUTOASSIGN. In this example, two spin systems (GS_i, red; and GS_{i+1}, blue) are shown whose CA-ladders have C^α and C^β chemical shifts consistent with residue types {Asp, Phe, Ile, Asn} and {Ser, Thr}, respectively. In addition, the CO-ladder values of GS_{i+1} indicate probable residue types {Asp, Phe, Ile, Asn} for the preceding residue with a good sequential match to the CA-ladder of GS_i. The CO-ladder of GS_{i-1} is consistent only with the Ala residue type. There are two dipeptide sites in the protein sequence (Phe-Ser and Asp-Thr, underlined) that are consistent with the hypothesis that these two GSs are adjacent. However, only one of these is also consistent with the probable residue type(s) {Ala} associated with the CO-ladder of GS_i. Consequently, only one assignment is possible for the GSs as a pair; i.e. the Asp-Thr segment. All other assignments will be ruled out if the link is established.

sites must be removed from the list of possible assignments that are stored for all other GSs. This narrowing of the possible assignments of other GSs may, in turn, lead to additional sequential assignments and/or further restrictions on possible links. More generally, constraints are derived and propagated from sequential links and assignments that can be ''ruled in'', as well as those that can be ''ruled out'' on the basis of inconsistencies or contradictions in the matching process. In this way, AUTOASSIGN works progressively toward a solution in a bootstrap fashion.

A simple example of the constraint propagation used in AUTOASSIGN is shown in Figure 8. In this hypothetical situation, there are two dipeptide sites in the sequence whose residue types are consistent with the CA-ladders of the two GSs and the CO-ladder of GS_{i+1}. However, only one of these dipeptide sites is also consistent with the CO-ladder of GS_i. Thus if the link is established, the assignments to that dipeptide site (Asp-Thr) will be established by the process of constraint propagation. Conversely, if other GSs are assigned to the Asp-Thr dipeptide site shown, the potential link between GS_i and GS_{i+1} will be ruled out.

Summary of AUTOASSIGN

In summary, AUTOASSIGN's best-first search strategies are applied to establish selectively the most reliable links and assignments first, with rigorous criteria used to filter out decisions involving uncertainty. The second stage of iterative constraint-based matching relaxes the requirement that all intraresidue cross-peaks must be distinct from sequential cross-peaks, and allows matching of less complete ladders. In the third cycle of constraint-based matching, additional sequential links and assignments may be obtained through exhaustive exploration of ways to extend the assigned segments. Stage 4 reinstates the weaker spin systems whose analysis was deferred and again invokes iterative constraint-based matching. In the final fifth stage, the designated shifts of assigned and unassigned GSs are scanned for obvious discrepancies, assignments and/or designations are retracted where appropriate, corrections are made where possible, and iterative constraint-based matching is repeated for the last time. Finally, the remaining unassigned SSs and GSs are examined to determine if any additional assignments can be inferred by a process of elimination.

Acknowledgments

We thank Drs C.B. Ríos, J. Laity, F. Moy, Z. Shang and G.V.T. Swapna for helpful discussions and for data used in this work. We also thank R. Watson for helpful comments on the manuscript, and Molecular Simulations Inc. for providing NMRCompass software. This work was supported by grants from the National Institutes of Health (GM-47014 and GM-50733, and NRSA award 1-F37-LM00036 to D. Z.), the National Science Foundation (MCB-9407569), a National Science Foundation Young Investigator Award (MCB-9357526), and an American Cyanamid Young Investigator Award in Analytical and Physical Chemistry. G.T.M. is a Camille Dreyfus Teacher Scholar.

References

Bartels, C., Billeter, M., Guntert, P. & Wüthrich, K. (1996). Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *J. Biomol. NMR*, **7**, 207–213.

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

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

Boucher, W., Laue, E. D., Campbell-Burk, S. & Domaille, P. J. (1992). Four-dimensional heteronuclear triple resonance NMR methods for the assignment of backbone nuclei in proteins. *J. Am. Chem. Soc.* **114**, 2262–2264.

Clowes, R. T., Boucher, W., Hardman, C. H., Domaille, P. J. & Laue, E. D. (1993). A 4D HCC(CO)NNH experiment for the correlation of aliphatic side-chain and backbone resonances in ^{13}C/^{15}N-labelled proteins. *J. Biomol. NMR*, **3**, 349–354.

Clubb, R. T., Thanabal, V. & Wagner, G. (1992). A constant-time 3D triple-resonance pulse scheme to correlate intraresidue ^{1}H, ^{15}N, and ^{13}C′ chemical shifts in ^{15}N-^{13}C-labeled proteins. *J. Magn. Reson.* **97**, 213–217.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 277–293.

Dijkstra, K., Kroon, G. J., van Nuland, N. A. & Sheek, R. M. (1994). The COCAH experiment to correlate carbonyl, C^{α}, and H^{α} resonances in proteins. *J. Magn. Reson. ser. A*, **117**, 102–105.

Dötsch, V. & Wagner, G. (1996). Editing for amino-acid type in CBCACONH experiments based on the $^{13}C\beta$-$^{13}C\gamma$ coupling. *J. Magn. Reson. ser. B*, **111**, 310–313.

Dötsch, V., Oswald, R. E. & Wagner, G. (1996). Amino-acid-type-selective triple-resonance experiments. *J. Magn. Reson. ser. B*, **110**, 107–111.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Anaylsis* John Wiley and Sons, New York, NY.

Feng, W., Ríos, C. B. & Montelione, G. T. (1996). Phase labeling of C-H and C-C spin system topologies: application in PFG-HACANH and PFG-HACA (CO)NH triple-resonance experiments for determining backbone resonance assignments in proteins. *J. Biomol. NMR*, **8**, 98–104.

Fox, M. S. (1986). Observations on the role of constraints in problem-solving. *The Sixth Canadian Proceedings in Artificial Intelligence*.

Friedrichs, M. S., Mueller, L. & Wittekind, M. (1994). An automated procedure for the assignment of protein $^{1}H^{N}$, ^{15}N, $^{13}C^{\alpha}$, $^{1}H^{\alpha}$, $^{13}C^{\beta}$, and $^{1}H^{\beta}$ resonances. *J. Biomol. NMR*, **4**, 703–726.

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

Grzesiek, S. & Bax, A. (1992a). Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J. Am. Chem. Soc.* **114**, 6291–6293.

Grzesiek, S. & Bax, A. (1992b). An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. *J. Magn. Reson.* **99**, 201–207.

Grzesiek, S. & Bax, A. (1993). Amino acid type determination in the sequential assignment procedure of uniformly $^{13}C/^{15}N$-enriched proteins. *J. Biomol. NMR*, **3**, 185–204.

Grzesiek, S., Anglister, J. & Bax, A. (1993). Correlation of backbone amide and aliphatic side-chain resonances in $^{13}C/^{15}N$-enriched proteins by isotropic mixing of ^{13}C magnetization. *J. Magn. Reson. ser. B*, **101**, 114–119.

Grzesiek, S., Kuboniwa, H. & Hinck, A. (1995). Multiple-quantum line narrowing for measurement of H^{α}-H^{β} J couplings in isotopically enriched proteins. *J. Am. Chem. Soc.* **117**, 5312–5315.

Hare, B. J. & Prestegard, J. H. (1994). Application of neural networks to automated assignment of NMR spectra of proteins. *J. Biomol. NMR*, **4**, 35–46.

Ikura, M., Kay, L. E. & Bax, A. (1990). A novel approach for sequential assignment of ^{1}H, ^{13}C, and ^{15}N spectra of larger proteins: heteronuclear triple-resonance three dimensional NMR spectroscopy. *Biochemistry*, **29**, 4659–4667.

Kay, L. E. (1993). A three-dimensional NMR experiment for the separation of aliphatic carbon chemical shifts via the carbonyl chemical shift in $^{15}N,^{13}C$-labeled proteins. *J. Magn. Reson. ser. B*, **101**, 110–113.

Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. (1990). Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J. Magn. Reson.* **89**, 496–514.

Kay, L. E., Ikura, M., Grey, A. A. & Muhandiram, D. R. (1992a). Three-dimensional NMR experiments for the separation of side-chain correlations in proteins via the carbonyl chemical shift. *J. Magn. Reson.* **99**, 652–659.

Kay, L. E., Keifer, P. & Saarinen, T. (1992b). Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *J. Am. Chem. Soc.* **114**, 10663–10665.

Kraulis, P. J. (1994). Protein three-dimensional structure determination and sequence-specific assignment of ^{13}C and ^{15}N-separated NOE data. A novel real space *ab initio* approach. *J. Mol. Biol.* **243**, 696–718.

Kumar, V. (1992). Constraint satisfaction methods in artificial intelligence. *Artificial Intelligence Magazine*, Spring, 32–44.

Laird, J. & Newell, A. (1983). A universal weak method. Technical Report, *CMU-CS-83-141* Department of Computer Science, Carnegie-Mellon University.

Li, Y. C. & Montelione, G. T. (1993). Solvent saturation-transfer effects in pulse-field gradient heteronuclear single-quantum-coherence spectra of polypeptides and proteins. *J. Magn. Reson. ser. B*, **101**, 315–319.

Logan, T. M., Olejniczak, E. T., Xu, R. X. & Fesik, S. W. (1992). Side chain and backbone assignments in isotopically labeled proteins from two heteronuclear triple resonance experiments. *FEBS Letters*, **314**, 413–418.

Lyons, B. A. & Montelione, G. T. (1993). 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. ser. B*, **101**, 206–209.

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

Mackworth, A. K. (1977). Consistency in networks of relations. *Artifici. Intell.* **8**, 99–118.

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

Montelione, G. T. & Wagner, G. (1989). Accurate measurements of homonuclear H^{N}-H^{α} coupling constants in polypeptides using heteronuclear 2D NMR experiments. *J. Am. Chem. Soc.* **111**, 5474–5475.

Montelione, G. T. & Wagner, G. (1990). Conformation-independent sequential NMR connections in isotope-enriched polypeptides by ^{1}H-^{13}C-^{15}N triple-resonance experiments. *J. Magn. Reson.* **83**, 183–188.

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

Morelle, N., Brutscher, B., Simorre, J. P. & Marion, D. (1995). Computer assignment of the backbone resonances of labeled proteins using two-dimensional correlation experiments. *J. Biomol. NMR*, **5**, 154–160.

Moy, F. J., Seddon, A. P., Campbell, E. B., Böhlen, P. & Powers, R. (1995). ^{1}H, ^{15}N, ^{13}C, and ^{13}CO assignments and secondary structure determination of basic fibroblast growth factor using 3D heteronuclear NMR spectroscopy. *J. Biomol. NMR,* **6**, 245–254.

Muhandiram, D. R. & Kay, L. E. (1994). Gradient-enhanced triple-resonance three-dimensional NMR experiments with improved sensitivity. *J. Magn. Reson. ser. B,* **103**, 203–216.

Nadel, B. A. (1986). The general consistent labeling (or constraint satisfaction) problem. Technical Report, DCS-TR-170. Computer Science Department, Rutgers University.

Newkirk, K., Feng, W., Jiang, W., Tejero, R., Emerson, S. D., Inouye, M. & Montelione, G. T. (1994). Solution NMR structure of the major cold shock protein (CspA) from *Escherichia coli*: identification of a binding epitope for DNA. *Proc. Natl Acad. Sci. USA,* **91**, 5114–5118.

Nilsson, N. J. (1980). *Principles of Artificial Intelligence* Morgan Kaufman Publishers, Inc., San Mateo, CA.

Olsen, J. B., Jr & Markley, J. L. (1994). Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances: a demonstration of the connectivity tracing assignment tools (CONTRAST) software package. *J. Biomol. NMR,* **4**, 385–410.

Ousterhout, J. K. (1993). *Tcl and the TK Toolkit* Addison-Wesley Publishing, Reading, MA.

Ríos, C. B., Feng, W., Tashiro, M., Shang, Z. & Montelione, G. T. (1996). Phase labeling of C-H and C-C spin-system topologies: application in constant-time PFG-CBCA(CO)NH experiments for discriminating amino-acid spin-sytem types. *J. Biomol. NMR,* **8**, 345–350.

Seavey, B. R., Farr, E. A., Westler, W. M. & Markley, J. (1991). A relational database for sequence-specific protein NMR data. *J. Biomol. NMR,* **1**, 217–236.

Shimotakahara, S., Ríos, C. B., Laity, J. H., Zimmerman, D. E., Scheraga, H. A. & Montelione, G. T. (1997). NMR structural analysis of an analog of an intermediate formed in the rate-determining step of one pathway in the oxidative folding of bovine pancreatic ribonuclease A: automated analysis of ^{1}H, ^{13}C and ^{15}N resonance assignments in the wild-type and [C65S, C72S] mutant forms. *Biochemistry,* in the press.

Tashiro, M., Ríos, C. B. & Montelione, G. T. (1995). Classification of amino acid spin systems using PFG HCC(CO)NH-TOCSY with constant-time aliphatic ^{13}C frequency labeling. *J. Biomol. NMR,* **6**, 211–216.

Wang, A. C., Lodi, P. J., Qin, J., Vuister, G. W., Gronenborn, A. M. & Clore, G. M. (1994). An efficient triple-resonance experiment for proton-directed sequential backbone assignment of medium-sized proteins. *J. Magn. Reson. ser. B,* **105**, 196–198.

Wehrens, R., Lucasius, C., Buydens, L. & Kateman, G. (1993). Sequential assignment of 2D-NMR spectra of proteins using genetic algorithms. *J. Chem. Inf. Comput. Sci.* **33**, 245–251.

Wishart, D. S., Sykes, B. D. & Richards, F. M. (1991). Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.* **222**, 311–333.

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

Yamazaki, T., Lee, W. & Revington, M. (1995). An HNCA pulse scheme for the backbone assignment of ^{15}N, ^{13}C, ^{2}H-labeled proteins: application to a 37-kDa Trp repressor-DNA complex. *J. Am. Chem. Soc.* **116**, 6464–6465.

Zimmerman, D. E. & Montelione, G. T. (1995). Automated analysis of nuclear magnetic resonance assignments for proteins. *Curr. Opin. Struct. Biol.* **5**, 664–673.

Zimmerman, D., Kulikowski, C. & Montelione, G. T. (1993). A constraint reasoning system for automating sequence-specific resonance assignments from multidimensional protein NMR spectra. *Proceedings of the First International Conference of Intelligent Systems for Molecular Biology,* **1**, 447–455.

Zimmerman, D., Kulikowski, C., Wang, L. L., Lyons, B. & Montelione, G. T. (1994). Automation of sequential resonance assignments in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation with artifical intelligence. *J. Biomol. NMR,* **4**, 241–256.

Edited by P. E. Wright

(Received 30 December 1996; received in revised form 14 March 1997; accepted 20 March 1997)