Journal of Biomolecular NMR, 4 (1994) 703–726
ESCOM

J-Bio NMR 200

An automated procedure for the assignment of protein ^{1}HN, ^{15}N, $^{13}C^{\alpha}$, $^{1}H^{\alpha}$, $^{13}C^{\beta}$ and $^{1}H^{\beta}$ resonances*

Mark S. Friedrichs, Luciano Mueller and Michael Wittekind**

Macromolecular NMR Department, Bristol-Myers Squibb Pharmaceutical Research Institute, P.O. Box 4000, Princeton, NJ 08543-4000, U.S.A.

Received 28 December 1993
Accepted 28 March 1994

Keywords: Protein; Automation; Resonance assignment

SUMMARY

A computer algorithm that determines the ^{1}HN, ^{15}N, $^{13}C^{\alpha}$, $^{1}H^{\alpha}$, $^{13}C^{\beta}$ and $^{1}H^{\beta}$ chemical-shift assignments of protein residues with minimal human intervention is described. The algorithm is implemented as a suite of macros that run under a modified version of the FELIX 1.0 program (Hare Research, Bothell, WA). The input to the algorithm is obtained from six multidimensional, triple-resonance experiments: 3D HNCACB, 3D CBCA(CO)HN, 4D HNCAHA, 4D HN(CO)CAHA, 3D HBHA(CO)NH and 3D HNHA(Gly). For small proteins, the two 4D spectra can be replaced by either the 3D HN(CA)HA, 3D H(CA)NNH, or the ^{15}N-edited TOCSY-HSQC experiments. The algorithm begins by identifying and collecting the intraresidue and sequential resonances of the backbone and $^{13}C^{\beta}$ atoms into groups. These groups are sequentially linked and then assigned to residues by matching the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift profiles of the linked groups to that of the protein's primary structure. A major strength of the algorithm is its ability to overcome imperfect data, e.g., missing or overlapping peaks. The viability of the procedure is demonstrated with two test cases. In the first, NMR data from the six experiments listed above were used to reassign the backbone resonances of the 93-residue human hnRNP C RNA-binding domain. In the second, a simulated cross-peak list, generated from the published NMR assignments of calmodulin, was used to test the ability of the algorithm to assign the backbone resonances of proteins containing internally homologous segments. Finally, the automated method was used to assign the backbone resonances of apokedarcidin, a previously unassigned, 114-residue protein.

INTRODUCTION

A major step in the process of determining protein structures by NMR is the assignment of the many nuclear resonances. This time-consuming task has been traditionally performed manually

*A preliminary account of the research presented in this manuscript was given on a poster at the Frontiers of NMR, Keystone Symposia on Molecular and Cellular Biology, Taos, NM, 1993. The macros and the subroutine code described in this paper are available to anyone who has written permission from Biosym to obtain our modified FELIX version.
**To whom correspondence should be addressed.

and remains a major bottleneck in protein structure determination projects. In efforts to streamline this procedure, a number of groups have proposed automated or semiautomated algorithms to assign homonuclear ^{1}H 2D spectra (examples include Cieslar et al., 1988; Weber et al., 1988; Eads and Kuntz, 1989; Catasti et al., 1990; Van de Ven, 1990; Eccles et al., 1991; Kleywegt et al., 1991; Nelson et al., 1991; Wehrens et al., 1993; Xu et al., 1993). However, the main obstacle to the successful implementation of automated assignment strategies has been the poor resolution of these spectra; severe overlap can be observed in 2D ^{1}H spectra of even small proteins.

The collection of 3D and 4D NMR spectra leads to a large increase in spectral resolution. Although higher dimensionality improves the resolution of 3D homonuclear ^{1}H spectra of proteins relative to their 2D counterparts (Griesinger et al., 1987; Vuister and Boelens, 1987), the most dramatic increases in spectral resolution are realized when the new dimensions are the covalent ^{15}N and/or ^{13}C resonances (Fesik and Zuiderweg, 1988; Marion et al., 1989). In particular, triple-resonance experiments allow correlation of the ^{1}H, ^{13}C and ^{15}N resonances of isotopically enriched proteins via large one-bond couplings (Ikura et al., 1990; Kay et al., 1990a; reviewed in Grzesiek and Bax, 1993b). Because the triple-resonance, multidimensional experiments are relatively sensitive and have high resolution, the data are well suited as input for automated assignment algorithms.

A number of protein assignment strategies exploiting multidimensional data have been proposed that could form the basis for automation algorithms (examples include Fesik and Zuiderweg, 1990; Ikura et al., 1990,1991a; Gao and Burkhart, 1991; Grzesiek et al., 1992; Bernstein et al., 1993; Kleywegt et al., 1993; Logan et al., 1993; Lyons et al., 1993). One approach (Domaille, 1991; Campbell-Burk et al., 1992) uses 4D HCANNH and HCA(CO)NNH (Boucher et al., 1992) or 4D HNCAHA and HN(CO)CAHA (Kay et al., 1992; Olejniczak et al., 1992a) spectra to correlate the intra- and interresidue backbone ^{1}HN, ^{15}N, ^{13}C$^\alpha$ and ^{1}H$^\alpha$ resonances. This method is particularly powerful when extended to use spectra from recently introduced pulse sequences that correlate complete side-chain ^{13}C and/or ^{1}H resonances with those of the backbone amide proton and nitrogen nuclei in a single 4D (Logan et al., 1992; Clowes et al., 1993) or 3D (Montelione et al., 1992; Grzesiek et al., 1993) experiment. These pulse sequences utilize ^{13}C isotropic mixing schemes to establish correlations among the side-chain nuclei. However, these side-chain correlation experiments have relatively low sensitivity, and as a result the use of these experiments with this approach is viable only for relatively small proteins.

A second proposed strategy utilizes the recently developed 3D experiments CBCA(CO)NNH (Grzesiek and Bax, 1992a), CBCANNH (Grzesiek and Bax, 1992b), and HNCACB (Wittekind and Mueller, 1993) that correlate the chemical shifts of the backbone amide groups with the ^{13}C$^\alpha$ and ^{13}C$^\beta$ resonances. These two carbon resonances are used to establish sequential connectivities between residues. In addition, the carbon chemical-shift information can be used both to assign residues by amino acid type (Richarz and Wüthrich, 1978; Oh et al., 1988) and to help align sequentially linked residues with the primary protein sequence (Grzesiek and Bax, 1993a). While these pulse sequences yield spectra with good signal-to-noise ratios for relatively large proteins, sequential linkages based on these spectra alone may be ambiguous due to the overlap of the ^{13}C$^\alpha$ and ^{13}C$^\beta$ resonances among the different residues.

Whereas a number of assignment strategies utilizing multidimensional protein spectra have been suggested, fewer accounts of actual automated implementations have been published. Descriptions of algorithms using 3D homonuclear proton data as input have been made (Cieslar

et al., 1990; Kleywegt et al., 1993). A simulated annealing assignment method using 3D ^{15}N-edited ^{1}H NOESY data to establish sequential connectivities has been reported (Bernstein et al., 1993). Montelione and co-workers have described an automated procedure using amide-detected ^{13}C side-chain TOCSY correlation spectra as input (Zimmerman et al., 1993).

In this report we describe an algorithm that automates the assignments of the ^{1}HN, ^{15}N, $^{13}C^{\alpha}$, $^{1}H^{\alpha}$, $^{13}C^{\beta}$ and $^{1}H^{\beta}$ resonances. The strategy employed is a synthesis of two approaches outlined above, combining the 4D backbone correlation approach of Laue, Domaille and co-workers with the $^{13}C^{\alpha}$-$^{13}C^{\beta}$ amide group correlation method of Grzesiek and Bax. The algorithm is conceptually simple, yet flexible enough to deal with incomplete or ambiguous data. It is implemented with a set of five macros that run under an enhanced, in-house version of the FELIX 1.0 program (Hare Research, Bothell, WA). The NMR input to the procedure is obtained from six multidimensional, triple-resonance experiments: 3D HNCACB (Wittekind and Mueller, 1993), 3D CBCA(CO)NH (Grzesiek and Bax, 1992a), 4D HNCAHA and HN(CO)CAHA (Kay et al., 1992; Olejniczak et al., 1992a), 3D HBHA(CO)NH (Grzesiek and Bax, 1993a,b), and 3D HNHA(Gly) (Wittekind et al., 1993) (see Fig. 1). These experiments were chosen because they provide relatively high sensitivity for proteins with molecular weights on the order of 20 kDa or more. Furthermore, they supply the α- and β-carbon chemical shifts which are used as signatures for the different amino acid types. For smaller proteins, the two 4D experiments can be substituted with either the 3D H(CA)NNH (Kay et al., 1991), HN(CA)HA (Clubb et al., 1992), or 3D ^{15}N-edited TOCSY-HSQC-type (Fesik and Zuiderweg, 1988; Marion et al., 1989) experiments, thereby reducing the data acquisition times.

The method was developed and refined for two previously assigned proteins: the human hnRNP C RNA-binding domain and calmodulin. Experimental NMR data was used as input for the assignment of the human hnRNP C RNA-binding domain, whereas simulated cross-peak lists were generated from the published assignments of calmodulin. Calmodulin was used as a test case for the method, because it possesses fourfold internal sequence homology. Finally, the automated method was applied to apokedarcidin, a previously unassigned protein. The assignment of this protein was also challenging because of the large number of glycines (16% of residues) and because the NMR sample was contaminated with a small amount of the holo-protein.

METHODS

Software environment

The algorithm is implemented with macros that operate under an in-house, enhanced version of the FELIX 1.0 program. The main advantages in using a macro language, as opposed to a formal computer language such as FORTRAN or C, are that a macro program is usually much more concise, and a macro command language provides an open interface between the user and the program. These features allow the user to write macros for a specific task relatively quickly and with considerably less programming effort than required for a FORTRAN or C program. The drawbacks to programming in a macro language are that macro languages often are not as rich in data structures and logic capabilities as the formal programming languages, and are generally slower.

We extended the already powerful macro facility in FELIX 1.0 by including multidimensional arrays with variable names, subroutine calls that allow arguments to be passed in either direction,

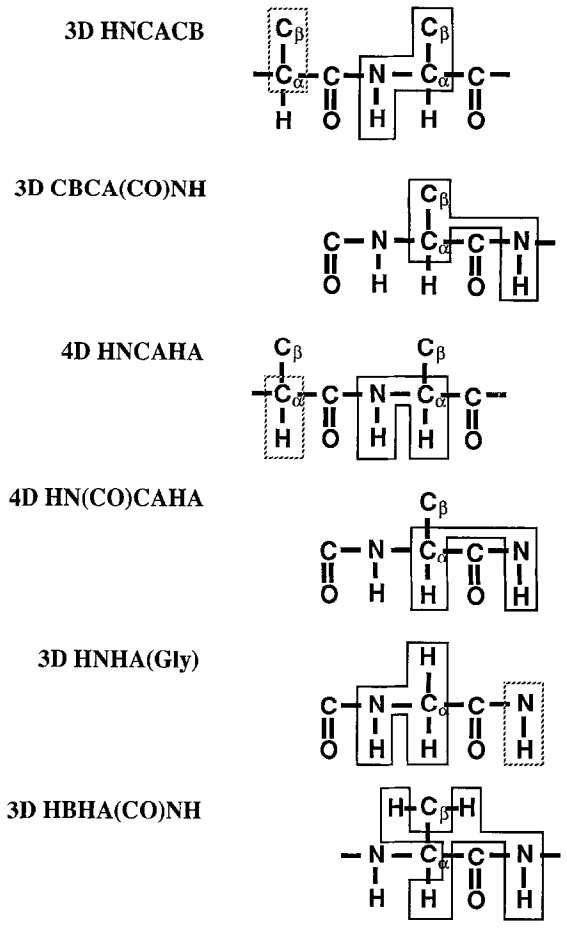

Fig. 1. Correlations defined by the six multidimensional triple-resonance NMR experiments used in this study. Solid boxes denote correlations that are defined by relatively strong one-bond coupling constants. Dotted boxes denote correlations mediated by the variable and relatively weak two-bond $^{2}J_{NC\alpha}$ coupling constants. Cross peaks arising from the latter correlations are used but not required by the algorithm to establish the sequential connectivities.

nested if-else-endif control structures with a primitive Boolean algebra, and simple list processing (e.g., intersection of lists of peaks based on chemical-shift values). In addition, to reduce the processing overhead, the macros are parsed only once upon their input into FELIX and pointers to the memory addresses of the macro variables (FELIX symbols) are stored. As a result, repeated parsing of the macro and searches through a table of macro variables are avoided. These modifications decrease the execution time by factors of up to eight times those of the unmodified FELIX.

A database facility, written in C and designed primarily for use in macros, was also incorporated into FELIX. This facility integrates information involving spectra, spin systems, residues and other NMR-related data under a common framework and alleviates much of the tedious bookkeeping. More importantly, user-defined relationships can be established among the different types of information. These relationships are typically recorded via memory address pointers

between data structures. For example, a user may set an address pointer between an atom and a peak data structure to register that the peak frequency in a particular spectral dimension arises from the atom's nuclear resonance. Memory-address pointers are particularly powerful in this context because they allow changes in database entries to be immediately propagated throughout the database; therefore, other entries that depend on the modified data structure are automatically updated. With the database and enhanced macro facilities, a wide range of applications can be written quickly and, most importantly, maintained easily via macros. These applications include the automated assignment of NOE spectra, analysis of NMR dynamical data, the automated assignment of side-chain atoms and, as discussed in this report, the automated assignment of backbone nuclei of a protein. Furthermore, even though the macros run more slowly than the equivalent C or FORTRAN code, the execution times are fast enough to prevent this from being a major drawback for most applications.

The peak-picking algorithm used here was an in-house implementation of a threshold pick. To increase the resolution, the centers and heights of the peaks were refined by fitting the local maxima/minima and the adjacent data points to a quadratic polynomial. The rectangular footprint of each peak was also calculated using a user-defined threshold. The length of the footprint along a given dimension is not required to be symmetric about the peak's center. In addition, the software attempts to deal with cases in which the footprints of two peaks overlap with varying degrees of success.

Data

The primary input to the algorithm is the data from the 3D HNCACB, 3D CBCA(CO)NH, 4D HNCAHA and 4D HN(CO)CAHA experiments. The 3D HBHA(CO)NH and HNHA(Gly) spectra are used for the assignment of glycine residues. In addition, the HBHA(CO)NH spectrum is used to assign the H^{β} resonances at the end of the assignment procedure. The correlations established by these experiments are diagrammed in Fig. 1. The published pulse sequences used to collect the data utilized in these studies were not modified*. The resolution of the carbon dimension in the HNCACB experiment was enhanced by multiplying the t_1 time domain data by an inverse cosine function tuned to 35 Hz prior to linear prediction (Wittekind and Mueller, 1993).

For non-glycine residues, the HNCACB spectrum contains four cross peaks that share $^{15}N_i$ and $^{1}HN_i$ frequencies: two peaks with positive intensities, corresponding to the ($^{13}C_i^{\alpha}$, $^{15}N_i$, $^{1}HN_i$) and ($^{13}C_{i-1}^{\alpha}$, $^{15}N_i$, $^{1}HN_i$) resonances and two peaks with negative intensities, corresponding to the ($^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}HN_i$) and ($^{13}C_{i-1}^{\beta}$, $^{15}N_i$, $^{1}HN_i$) resonances. For both the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ peaks, the integrated volumes of the intraresidue cross peaks are usually larger in absolute value than those of the sequential cross peaks. In the 3D CBCA(CO)NH spectrum, two cross peaks share $^{15}N_i$ and $^{1}HN_i$ frequencies: ($^{13}C_{i-1}^{\alpha}$, $^{15}N_i$, $^{1}HN_i$) and ($^{13}C_{i-1}^{\beta}$, $^{15}N_i$, $^{1}HN_i$); in this spectrum both peaks have positive intensities. The 3D HBHA(CO)NH is completely analogous, but with the $^{13}C_{i-1}^{\alpha}$ and $^{13}C_{i-1}^{\beta}$ frequencies replaced by those of $^{1}H_{i-1}^{\alpha}$ and $^{1}H_{i-1}^{\beta}$, respectively. The 4D HNCAHA spectrum contains two

*The published 3D HNCACB paper contains an error in the phase cycle. The phases of the three ^{13}C pulses ϕ_5, ϕ_6 and ϕ_8 should be incremented by 90° to obtain hypercomplex spectra in F1. The phase of the ^{15}N pulse ϕ_4 should be increased by 90° to obtain hypercomplex spectra in F2.

peaks at the ($^{13}C_i^{\alpha}$, $^{1}H_i^{\alpha}$, $^{15}N_i$, $^{1}HN_i$) and ($^{13}C_{i-1}^{\alpha}$, $^{1}H_{i-1}^{\alpha}$, $^{15}N_i$, $^{1}HN_i$) resonance positions, and the 4D HN(CO)CAHA contains a single peak at ($^{13}C_{i-1}^{\alpha}$, $^{1}H_{i-1}^{\alpha}$, $^{15}N_i$, $^{1}HN_i$). Finally, the 3D HNHA(Gly) experiment effectively selects amide nitrogens with couplings to methylene carbons. For glycines at residue position i, intraresidue backbone correlations give rise to the ($^{1}H_i^{\alpha 1}$, $^{15}N_i$, $^{1}HN_i$) and ($^{1}H_i^{\alpha 2}$, $^{15}N_i$, $^{1}HN_i$) cross peaks and sequential correlations give rise to the ($^{1}H_i^{\alpha 1}$, $^{15}N_{i+1}$, $^{1}HN_{i+1}$) and ($^{1}H_i^{\alpha 2}$, $^{15}N_{i+1}$, $^{1}HN_{i+1}$) cross peaks. For all of the experiments, no intraresidue resonances are present for prolines since they do not have an amide proton.

Assignment algorithm

Throughout the algorithm, the basic data structure used to store information about each residue is referred to as an RID (**R**esidue **I**nformation **D**ata structure). Each RID contains the ppm values for the different atom types ($^{13}C_i^{\alpha}$, $^{1}H_i^{\alpha}$, $^{13}C_{i-1}^{\alpha}$, $^{1}H_{i-1}^{\alpha}$, $^{15}N_i$, $^{1}HN_i$, ...), pointers linking the atoms to the dimensions of peaks, links between residues, and other data items. Before the residue-specific assignments are made, the RIDs serve as generic residues.

The assignment algorithm is divided into three major stages (see Fig. 2). In Stage 1, peaks from the different spectra (Fig. 1) that share amide ^{15}N and ^{1}HN chemical shifts are grouped into RIDs and then sorted according to the atom types that gave rise to the peaks. In the second stage, the RIDs are sequentially linked. In Stage 3, chains of sequentially linked RIDs are aligned with the protein's sequence. Stage 1 is carried out by a single macro, while the second and third stages are each implemented with two macros. Prior to the execution of the Stage 1 macro, the peaks from each of the spectra are automatically picked, and the chemical-shift positions and integrated volumes are stored in a database file.

The real challenge in automating the assignment process is dealing with chemical-shift degeneracies and incomplete or imperfect data. For proteins with molecular weights ranging from 10–20 kDa, our limited experience suggests that 10–30% of the RIDs have either chemical-shift degeneracies, or missing or spurious peaks. For these cases, our general approach is to try to resolve the ambiguities, if possible, during Stage 1. Any remaining problems are handled by the Stage 3 macros. The tactics employed during Stage 1 to deal with imperfect data include using minimal chemical-shift tolerances in searches for peaks within a spectrum or between two spectra, applying an inner product metric to gauge the alignment of peaks in two dimensions, and employing rules based on the peak volumes and the typical chemical-shift ranges of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ resonances to classify the peaks by atom type.

In situations where these strategies are unsuccessful, the ambiguities are recorded in the database. In general, two types of ambiguous situations can occur. First, there may be too few peaks grouped within an RID to allow the assignment of all the atom types. This situation typically arises from chemical-shift degeneracies or missing peaks. Such cases are handled by temporarily assigning a peak to multiple atom types, resulting in different atoms sharing the same chemical-shift value or by not making an assignment for an atom type. The second general case arises when too many peaks are grouped within an RID to allow all atom types to be uniquely assigned. Such cases may be due to the presence of spurious peaks or to overlap of the backbone amide $^{1}HN_i$ and $^{15}N_i$ frequency pairs of two or more residues. In these situations, an atom type within the RID is temporarily allowed to have multiple peak assignments, resulting in several different ppm assignments for that atom. In the special case of backbone amide ^{1}HN and ^{15}N overlap, it is important to note that a single RID will contain the chemical-shift information for

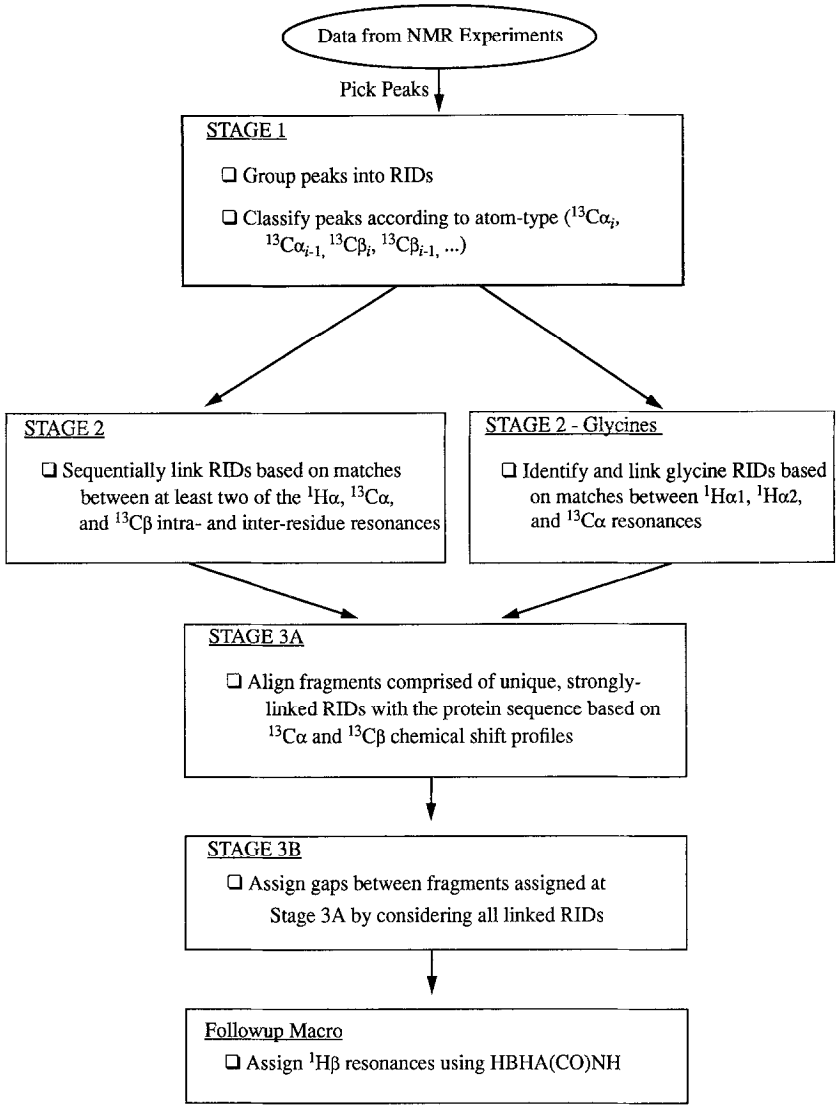

Fig. 2. Overview of the assignment algorithm. While both the Stage 2 and Stage 2-glycine macros must be run before proceeding to Stage 3, they are independent and can be applied in either order.

more than one residue. In both cases, these 'floating' assignments are dealt with in the later stages of the algorithm.

The Stage 3 macros are designed to cope with the ambiguities that remain unresolved after Stage 1. This capability of the Stage 3 macros originates from two sources. First, the information linking an RID to its sequential neighbor is overdetermined. The links between RIDs formed during Stage 2 are ideally based on the resonances of three atoms, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{1}H^{\alpha}$. However, links based on only two atoms are often sufficient to narrow the choices considerably. As a result, if the assignment for one atom is missing or misclassified, the program will still be able to link it to its correct neighbor. The second source of the ability of Stage 3 to decipher ambiguities arises

from the requirement that the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift profiles of the sequentially linked RIDs align with the profile derived from the protein's primary sequence. This constraint is very restrictive. By imposing the primary sequence information as a selection criterion in addition to the sequential linkage requirement, the number of possible assignments for a given residue position is drastically reduced. An overview of the design and flow of each macro is given below.

Stage 1. Classification of peaks – non-glycines

The Stage 1 macro groups the peaks from the four spectra – 3D HNCACB, 3D CBCA(CO)NH, 4D HNCAHA and 4D HN(CO)CAHA – into RIDs and classifies the peaks into the following four categories: $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$, $^{13}C_{i-1}^{\alpha}$ and $^{13}C_{i-1}^{\beta}$. As discussed above, a peak may belong to more than one category, if it represents degenerate resonances. Also, a category may have more than one peak assignment. This occurs, for example, when the ^{15}N and ^{1}HN chemical shifts of two residues are degenerate. An overview of the Stage 1 macro is given here; a more detailed account is given in the Appendix.

The macro begins with a peak p that has not been previously classified. It gathers all peaks from the four spectra with the same ^{15}N and ^{1}HN chemical-shift values as those of p, within user-specified tolerances. The chemical-shift tolerances are not fixed during the search. Instead, they are initially set to a lower bound and then are gradually increased until the expected number of peaks is found or an upper tolerance bound is reached. The lower and upper bounds are different for intra- and interspectra searches, with smaller ranges used for the intraspectra searches. This feature mitigates problems such as small but significant shifts in the measured resonance of a nucleus that arise from different spectral resolutions and experimental conditions. Extraneous outlier peaks from the same spectrum as peak p are identified next via an inner product metric that measures the alignment of the peaks in the ^{15}N and ^{1}HN dimensions (see Appendix). Peaks that are poorly aligned with p are removed from consideration for the current RID and, if they have not been previously assigned to another RID, placed back into the pool of unassigned peaks.

To classify the peaks into one of the four atom-type categories, their ^{13}C chemical shifts are matched across the spectra. Thus, for example, if the ^{13}C chemical shift of a negative peak from the HNCACB spectrum matches that of a peak in the CBCA(CO)HN spectrum, then the chemical shift must be the $^{13}C_{i-1}^{\beta}$ assignment. After the comparisons, if too many or no peaks are classified for a particular atom type, the macro attempts to resolve the problem by applying a number of tactics. These strategies include rule-based inferences based on the magnitude and sign of the peak volumes, the typical ppm ranges for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ atoms, and alignments of peaks in the ^{15}N and ^{1}HN dimensions as measured by the inner product metric (see Appendix). If the macro still cannot identify an atom resonance, it is omitted. Conversely, if several peaks remain as candidates for an atom assignment, the atom is temporarily assigned multiple resonances by the program.

As the Stage 1 macro is running, it prints out a log file containing essential information about each RID. A list of the RIDs with missing peaks, poorly aligned peaks within one of the spectra and inconsistent assignments from different spectra is also printed. Upon completion of the macro, the user can run the macro interactively to examine the anomalous RIDs in detail (visually inspect the spectra, compare the inner products among peaks within a spectrum, etc.) and modify the automated results if necessary.

Stage 2. Linkage of RIDs – non-glycines

The Stage 2 macro establishes links between pairs of sequential RIDs. For convenience in the discussion below, we use the subscript i to denote the RID associated with residue i of the protein, even though at this point in the algorithm, the residue(s) that a RID represents is unknown. Ideally, each RID_i has sequential chemical-shift entries corresponding to the $^1H^{\alpha}_{i-1}$, $^{13}C^{\alpha}_{i-1}$, and $^{13}C^{\beta}_{i-1}$ resonances, as well as its own corresponding intraresidue frequencies. To make links between RID_i and the other RIDs, the macro searches for all RIDs whose intraresidue chemical-shift values match the sequential chemical shifts of RID_i. A link is made if the chemical shifts of any two of the three atom types agree. By requiring that only two chemical shifts match, RIDs with a missing cross peak or a misassigned atom type are still included in the later alignment process. A minor drawback to this approach is that spurious links will often be made. However, as shown below, the alignment process in Stage 3 can almost always identify the correct link.

For cases in which a particular atom type has several chemical-shift assignments, all combinations of these assignments with those of the other atom types are used to establish links. For example, if both the $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ have two assignments and the $^1H^{\alpha}_{i-1}$ only one, then a total of four sets of chemical-shift values are used in the search for matching intraresidue resonances.

The link between a pair of RIDs is graded based on the number of atom types used to make the connection – links made with three atom-type matches receive a higher score than links based on only two matches. The residual differences between the sequential and intraresidue chemical shifts of the linked RIDs are also stored in the database. The grades and differences are later used by the Stage 3 macros to judge the fidelity of links between the RIDs and to derive overall scores for determining how well competing chains of sequentially linked RIDs align with a given protein segment.

Stage 2. Glycines

The cross peaks corresponding to intraresidue glycine residue correlations are missing in the 4D HNCAHA and HN(CO)CAHA experiments because the delay used to establish $^1H^{\alpha}_z$-$^{13}C^{\alpha}_x$ antiphase magnetization is tuned for amino acids having a single $^1H^{\alpha}$ (Kay et al., 1992; Olejniczak et al., 1992a). Because glycines have no $^{13}C^{\beta}$ resonance and their $^1H^{\alpha}$ resonances are missing, only the glycine $^{13}C^{\alpha}$ resonance is available for establishing sequential connectivities with the 3D HNCACB, 3D CBCA(CO)NH, 4D HNCAHA and 4D HN(CO)CAHA experiments. Since a single frequency is insufficient to unambiguously link RIDs, additional data is required. Sufficient information can be obtained from the glycine $^{13}C^{\alpha}$ and $^1H^{\alpha}$ resonances, since the two $^1H^{\alpha}$ atoms are usually nondegenerate.

The glycine $^1H^{\alpha 1}_i$ and $^1H^{\alpha 2}_i$ chemical-shift values are determined from the 3D HNHA(Gly) and 3D HBHA(CO)NH experiments using the Stage 2-glycine macro (see Fig. 2). The first part of the macro searches for RIDs with $^{13}C^{\alpha}_i$ chemical-shift values in the range 41–49 ppm. The $^{15}N_i$ and ^1HN_i chemical-shift values for these glycine RIDs are then matched to the F_2 and F_3 frequencies of the 3D HNHA(Gly) cross peaks to find the corresponding glycine $^1H^{\alpha 1}_i$ and $^1H^{\alpha 2}_i$ chemical-shift values. The second part of the macro then searches for RIDs that are C-terminal to the glycines, i.e., their $^{13}C^{\alpha}_{i-1}$ ppm values are within the range 41–49 ppm. The $^{15}N_i$ and ^1HN_i chemical-shift values for these RIDs are matched to the F_2 and F_3 frequencies of the 3D HBHA(CO)NH cross peaks to obtain the glycine $^1H^{\alpha 1}_{i-1}$ and $^1H^{\alpha 2}_{i-1}$ chemical-shift values. Using this information, the macro sets a link between RID_i and the glycine RID_{i-1}, based on the $^{13}C^{\alpha}$, $^1H^{\alpha 1}$ and $^1H^{\alpha 2}$ frequencies.

Stage 3A. Alignment with protein sequence – strong links only

In the final stage of the algorithm, the program assigns the RIDs to specific residues in the protein. This is accomplished by comparing the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift profiles of chains of linked RIDs to profiles, based on the protein sequence and the chemical-shift ranges of the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ atoms for each residue type. This strategy was originally proposed by Grzesiek and Bax (1993a). The $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift ranges for each residue type were derived from the published assignments of several proteins and are given in Table 1. An analysis of the distribution of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts as a function of residue type has recently been presented (Grzesiek and Bax, 1993a), and the bounds listed there are consistent with those in Table 1. In addition, Table 1 has two entries for the $^{13}C^{\beta}$ chemical-shift bounds of cysteine residues. Reported ^{13}C chemical shifts of proteins containing disulfide linkages (Hansen, 1991; Constantine et al., 1992) show that cystine $^{13}C^{\beta}$ resonances are shifted downfield relative to those of reduced cysteines.

The alignment is made in two stages. In Stage 3A, only unique, strong links are used, whereas in Stage 3B all links are considered. A link between two RIDs is defined to be strong if all three sequential chemical-shift values of RID_i agree with the corresponding intraresidue chemical-shift values of RID_{i-1}. In Stage 3A a sequentially linked chain of RIDs, or fragment in our lexicon, is constructed by starting with an arbitrary RID and following the strong links in both the N- and C-terminal directions. The fragment terminates at either end when no strong link or more than one strong link is encountered. The fragments are then aligned with the protein by matching their $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift profiles (see Fig. 3A).

To find the correct alignment, a score measuring the goodness of fit of the protein's $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift profiles to those of the fragment is computed for all possible alignments

TABLE 1
UPPER AND LOWER BOUNDS FOR $^{13}C^{\beta}$ AND $^{13}C^{\alpha}$ CHEMICAL SHIFTS USED FOR THE ALIGNMENT SCORE CALCULATIONS IN STAGE 3

Residue (R)	$\sigma^{U}_{R\beta}$	$\sigma^{L}_{R\beta}$	$\sigma^{U}_{R\alpha}$	$\sigma^{L}_{R\alpha}$	Residue (R)	$\sigma^{U}_{R\beta}$	$\sigma^{L}_{R\beta}$	$\sigma^{U}_{R\alpha}$	$\sigma^{L}_{R\alpha}$
ALA	22.0	16.0	55.0	50.0	LEU	45.0	41.0	58.0	52.0
ARG	35.0	29.0	60.0	55.0	LYS	35.0	30.0	60.0	53.0
ASN	42.0	36.0	55.0	50.0	MET	37.0	32.0	57.0	53.0
ASP	43.0	38.0	57.0	51.0	PHE	42.0	38.0	59.0	54.0
CYS (reduced)	33.0	28.0	58.0	53.0	PRO	34.0	31.0	66.0	62.0
CYS (oxidized)	50.0	38.0	58.0	53.0	SER	66.0	62.0	61.0	55.0
GLN	32.0	28.0	58.0	53.0	THR	71.0	67.0	64.0	58.0
GLU	32.0	28.0	60.0	53.0	TRP	35.0	25.0	58.0	53.0
GLY			49.0	41.0	TYR	43.0	38.0	58.0	55.0
HIS	31.0	27.0	58.0	54.0	VAL	36.0	31.0	66.0	58.0
ILE	44.0	38.0	64.0	57.0					

The ranges of the $^{13}C^{\beta}$ and $^{13}C^{\alpha}$ chemical shifts used in calculating the alignment score, S, in Stage 3 are tabulated for each residue type R. U and L denote upper and lower bounds, respectively. All values are given in ppm. The values were obtained by inspection of the published $^{13}C^{\beta}$ and $^{13}C^{\alpha}$ resonance values for IL-1β (Clore et al., 1990), Factor III^{glc} (Pelton et al., 1991), and hnRNP C RBD (Wittekind et al., 1992) with consideration given to secondary ^{13}C chemical-shift effects exhibited by carbons involved in regular secondary structures (Spera and Bax, 1991). Ranges for $^{13}C^{\beta}$ resonances of cysteines involved in disulfide bonds were obtained from the NMR assignments for bovine basic pancreatic trypsin inhibitor (Hansen, 1991) and 26-10 V_L domain (Constantine et al., 1992).

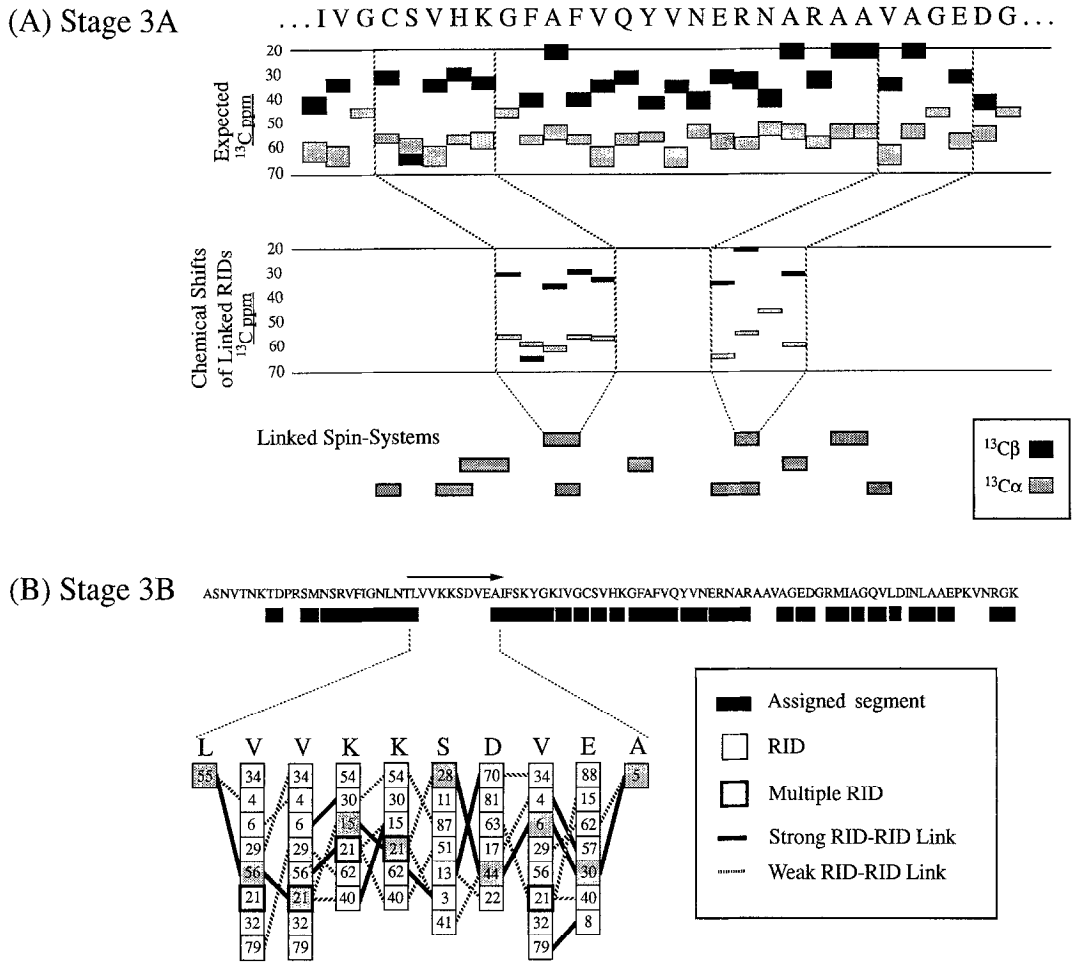

Fig. 3. Illustration of the Stage 3A and Stage 3B macros. (A) Stage 3A. The strongly linked fragments comprising RIDs connected by three-frequency ($^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$) links (strong RID–RID links) with no competing three-frequency links are arranged into ordered residue lists (represented by the grey bars at the bottom of the panel). For each fragment, the $^{13}C^{\alpha}$,$^{13}C^{\beta}$ chemical-shift profiles (probe profile, middle of the panel) are compared to the $^{13}C^{\alpha}$,$^{13}C^{\beta}$ chemical-shift profiles expected based on the primary sequence of the protein (sequence profile, top of the panel). For every possible probe/sequence profile alignment, a score comprising the contribution of each residue alignment is recorded (see text for computation of the alignment score S(i)). If the best alignment score is low and sufficiently lower than the next best score, an assignment is made. (B) Stage 3B. To assign an unassigned gap remaining after the end of the Stage 3A macro, the algorithm starts at the residue at the end of the N-terminal flanking assigned fragment and attempts to assign across the gap into the N-terminus of the C-terminal flanking assigned fragment. Only RIDs that are unassigned and are likely to be of the appropriate residue type (based on its $^{13}C^{\alpha}$, $^{13}C^{\beta}$ chemical-shift values) are considered at each unassigned residue position. The fragment that traverses the gap, has the lowest score, and has a score much lower than any other fragment traversing the gap defines the correct assignment. The assigned RIDs are then removed from further consideration for the remaining unassigned residue positions. As an illustration, a hypothetical case for the assignment of a gap in the hnRNP C1 RBD sequence is depicted. The boxes represent the RIDs that are available for assignment at residue positions. The RIDs are labeled with unique, arbitrary numbers to distinguish them from one another. Note that the same RIDs are available for each unassigned residue position of the same amino acid type. The shaded RIDs represent the correct assignments. The RID outlined with bold lines holds cross peaks from multiple residues (RID 21 in the figure). This type of RID can be assigned to more than one residue position (see text).

between the protein and fragment. The alignment score is the sum of the differences between the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift values of the RID fragment and the values expected for a particular span of the protein sequence. Specifically, for a fragment of F RIDs and a protein of P residues, P − F + 1 alignments are possible. For each alignment a score S(i) is computed, where i is the index of the first residue in the alignment (i = 1,..., P − F + 1). The scores are calculated according to the formula:

$$S(i) = \alpha'_{i-1} + \beta'_{i-1} + \sum_{j=1}^{F} \alpha_{i+j-1,j} + \beta_{i+j-1,j} \tag{1}$$

The penalty $\alpha_{i+j-1,j}$ measures the deviation of the $^{13}C^{\alpha}$ chemical shift of the *j*th RID in the fragment from the $^{13}C^{\alpha}$ chemical-shift range of residue i + j − 1 in the protein sequence; a similar definition holds for the $\beta_{i+j-1,j}$ penalty, but with the comparison based on the $^{13}C^{\beta}$ resonance. The penalties α'_{i-1} and β'_{i-1} measure the difference between the sequential information of the first RID in the fragment and the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of the residue at position i − 1 in the protein. These latter penalties are especially important for very short fragments, since they effectively increase the fragment length by one. The penalties are computed from tapered square-well functions that vary with residue type. The positions and widths of the bottoms of the wells for each residue type are given by the chemical-shift ranges in Table 1. If the $^{13}C^{\alpha}$ resonance falls within these bounds for a given residue type, the penalty is zero. On the other hand, if the $^{13}C^{\alpha}$ resonance falls outside a user-specified width, W, of the bounds, the penalty is a maximum. The penalties scale linearly over the width W, which was set to 5.0 ppm for all residue types during Stage 3A (W was set to 2.0 ppm during Stage 3B, see below). Formally, the $\alpha_{i+j-1,j}$ are given by the expression (completely analogous expressions apply for the $\beta_{i+j-1,j}$):

$$\alpha_{i+j-1,j} = \begin{cases} 0 & U_{R(i+j-1)\alpha} \geq \sigma_{j\alpha} \geq L_{R(i+j-1)\alpha} \\ (\sigma_{j\alpha} - U_{R(i+j-1)\alpha})/W & U_{R(i+j-1)\alpha} + W > \sigma_{j\alpha} > U_{R(i+j-1)\alpha} \\ (L_{R(i+j-1)\alpha} - \sigma_{j\alpha})/W & L_{R(i+j-1)\alpha} > \sigma_{j\alpha} > L_{R(i+j-1)\alpha} - W \\ 1.0 & \text{Otherwise} \end{cases} \tag{2}$$

$\sigma_{j\alpha}$ is the $^{13}C^{\alpha}$ chemical shift of the *j*th RID of the fragment. $U_{R(k)\alpha}$ and $L_{R(k)\alpha}$ are the upper and lower $^{13}C^{\alpha}$ ppm limits from Table 1 for the residue of type R (R = ALA, ARG, ...) at position k in the sequence.

Once the scores have been computed for all P − F + 1 alignments of a given fragment, the average score is calculated, and the individual alignment scores are normalized by this average. If the difference between the normalized scores for the best and next best alignment is greater than a user-specified threshold, the RIDs in the fragment are assigned to the protein residues of the best alignment. After all fragments have been examined once, the above process is repeated until no assignments are made during an iteration. Fragments that initially had two or more possible alignments may be assigned on subsequent passes if the competing alignments have been assigned to other fragments and are therefore removed from consideration.

Stage 3B. Alignment with protein sequence – all links

The final macro, Stage 3B, assigns the remaining gaps in the protein sequence. The details of

the Stage 3B assignment strategy are given in the diagram in Fig. 3B and are outlined below. The overall strategy is to start at the N-terminal residue of a gap (say, position i of the sequence) and construct a fragment that spans the gap from the list of unassigned RIDs. The macro begins by inspecting all unassigned RIDs that satisfy the following two conditions: (i) their $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift values fall within the expected chemical-shift ranges (Table 1) of the residue type at position i; and (ii) they have a sequential link (weak or strong) with the flanking RID at position i − 1. If a single RID satisfies these conditions, then it is added to the fragment, and the above process is repeated with i incremented by 1. On the other hand, if X RID candidates (X > 1) fulfill the criteria for position i, then the fragment is duplicated X − 1 times. The X candidate RIDs are then appended to one of the X − 1 clones or to the original fragment. For the next residue in the gap, each of the X fragments are considered independently of the other X − 1 fragments. This branching process can rapidly lead to many fragments for a single gap. However, most of these fragments represent incorrect alignments and will usually terminate before the entire gap is traversed.

The fragments that successfully traverse the gap are graded based upon a composite score: the fit to the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift profiles (the score S described above), and the sum of the residual ppm match deviations for each link within the fragments. The ppm match deviation D for a fragment of length F is given by the expression

$$D = \sum_{i=2}^{F} |\Delta C^{\alpha}(i-1, i)| + |\Delta C^{\beta}(i-1, i)| + 4|\Delta H^{\alpha}(i-1, i)| \quad (3)$$

$\Delta C^{\alpha}(i-1, i)$ is the chemical-shift difference between the sequential $^{13}C^{\alpha}$ assignment for RID_i and the intraresidue $^{13}C^{\alpha}$ assignment for RID_{i-1}; similar definitions hold for the $\Delta C^{\beta}(i-1, i)$ and $\Delta H^{\alpha}(i-1, i)$. For glycines, the $\Delta C^{\beta}(i-1, i)$ term is replaced by a similar one involving the $^{1}H^{\alpha 2}$ assignments. The factor of four in the last term corresponds to the ratio of the ^{1}H and ^{13}C gyromagnetic ratios; it is included to establish a uniform frequency scale. If one of the links was made with only two atom types, the term corresponding to the missing atom type is defined to be a constant Z for the ^{13}C terms or Z/4 for the ^{1}H term; Z was set to 2.0 for the applications discussed below. If the fragment's composite score, D + S, is sufficiently low and no other fragment has a score close to it, the RIDs in the fragment are assigned to the gap.

In cases where segments of the protein sequence remain unassigned after all gaps have been examined once, the process can be repeated. RIDs that were assigned on the previous pass are removed from consideration, and consequently, the number of fragments that span a gap will become progressively smaller. This process is repeated until all non-proline residues are assigned. The chemical shifts for prolines located at position i in the protein sequence can then be obtained from the sequential $^{13}C^{\alpha}_{i-1}$, $^{13}C^{\beta}_{i-1}$, and $^{1}H^{\alpha}_{i-1}$ resonances of residue i + 1. Situations in which two prolines are adjacent cannot currently be handled with the experiments employed here. Finally, the $^{1}H^{\beta}$ resonances are easily obtained from the 3D HBHA(CO)NH experiment, since the $^{15}N_i$, $^{1}HN_i$ and $^{1}H^{\alpha}_{i-1}$ chemical-shift values are now known.

RESULTS

The algorithm was applied to three different proteins: the human hnRNP C RNA-binding

TABLE 2
STAGE 3A RESULTS FOR HUMAN hnRNP C RNA-BINDING DOMAIN

Pass	Fragment	Length	Best alignment position	Best score	Next best alignment position	Next best score	Ratio	Assigned
1	1	12	15	1.85	32	89.67	0.741	X
1	2	3	57	0.00	43	13.94	0.365	X
1	3	3	83	1.81	84	12.07	0.258	X
1	4	2	90	0.00	30	0.00	0.000	
1	5	8	35	4.17	10	45.79	0.508	X
1	6	6	51	8.48	79	42.31	0.513	X
1	7	3	32	0.00	10	1.60	0.050	
1	8	2	49	0.08	29	1.36	0.049	
1	9	2	47	0.00	13	5.26	0.190	X
1	10	2	78	0.00	73	7.43	0.277	X
1	11	2	43	0.00	57	1.85	0.077	
1	12	3	73	0.00	78	7.78	0.234	X
1	13	5	60	0.00	63	23.94	0.442	X
1	14	2	76	0.00	68	8.17	0.274	X
1	15	3	92	0.00	44	14.19	0.431	X
1	16	2	9	0.00	6	3.08	0.101	X
1	17	2	45	0.00	93	3.22	0.122	X
1	18	3	27	0.00	33	0.42	0.016	
1	19	2	86	0.00	66	10.83	0.317	X
1	20	2	70	0.00	42	0.00	0.000	
1	21	2	80	0.00	31	10.00	0.342	X
1	22	2	67	0.00	64	6.72	0.187	X
1	23	2	13	0.00	47	2.32	0.090	
2	4	2	90	0.00	30	0.00	0.000	
2	7	3	32	0.00	4	10.00	0.310	X
2	8	2	49	0.08	29	1.36	0.049	
2	11	2	43	0.00	28	7.02	0.295	X
2	18	3	27	0.00	11	0.42	0.016	
2	20	2	70	0.00	90	16.86	0.659	X
2	23	2	13	0.00	49	14.55	0.563	X
3	4	2	90	0.00	30	0.00	0.000	
3	8	2	49	0.08	29	1.36	0.049	
3	18	3	27	0.00	5	10.00	0.374	X
4	4	2	90	0.00	30	0.00	0.000	
4	8	2	49	0.08	89	4.80	0.182	X
5	4	2	90	0.00	30	0.00	0.000	

Column one gives the number of applications of the algorithm to the unassigned fragments. Columns two and three are the fragment's identification number and length, respectively (the fragment number corresponds to the fragment number labels in Fig. 4A). Column four is the starting residue which aligns best with the fragment for the current pass and column five is the associated score. Columns six and seven are analogous to columns four and five, but give the next best alignment and its score. The sixth column is the ratio of (next best alignment − best score)/(average score over all alignments). The fragment was assigned to the best alignment if this ratio was greater than 0.1. An X in the final column denotes the assignment of the fragment on the current pass. After five passes all but one of the fragments were assigned (see text).

domain, calmodulin and apokedarcidin. In all three cases, the backbone resonances of all assignable residues were determined. The results for each are discussed below.

hnRNP C RNA-binding domain

A data set collected from the previously assigned human hnRNP C RNA-binding domain (Wittekind et al., 1992) was used as input to the algorithm. Upon completion of Stage 1, 102 RIDs were found, which is 12 more than expected (92 residues; two prolines). The extras were incomplete RIDs arising from side-chain ^{1}HN and ^{15}N resonances or spurious peaks. Of the remaining 90 RIDs, five were manually edited because the atom assignments either appeared inconsistent between two spectra, an assignment was missing, or extra (noise) peaks were present. One RID contained two sets of resonances due to complete overlap of the amide ^{1}HN and ^{15}N chemical shifts of Lys^{29} and Ile^{82}. Tolerances of 0.5 and 0.05 ppm were used for ^{13}C and ^{1}H, respectively, during the Stage 2 linking process.

The step-by-step progress of the Stage 3A macro can be monitored by inspection of the entries in Table 2; the corresponding fragments are given in the diagram in Fig. 4A. Twenty-three strongly linked fragments were constructed by the algorithm and all were assigned after the completion of pass 4 except for fragment 4, a dipeptide fragment made up of the RIDs corresponding to Lys^{30}-Ser^{31}. Although the sequential information was available for Ser^{31}, the intraresidue C$^{\alpha}$ and C$^{\beta}$ shifts were completely overlapped with each other and with the C$^{\alpha}$ shift of Lys^{30}. Since the entries for the intraresidue Ser^{31} C$^{\alpha}$ and C$^{\beta}$ resonances were missing, the algorithm gave no penalty for their misalignment, causing low alignment scores to be recorded at multiple positions along the protein sequence. For the other 22 assigned fragments, 73 RIDs were unambiguously assigned at the end of Stage 3A, corresponding to 82% of the assignable residues.

The Stage 3B macro was used to assign the remaining gaps. All of the backbone and ^{13}C$^{\beta}$ resonances were completely assigned, except for those of the N-terminal residue Ala^{2} and the amide nuclei of Ser^{3}, due to rapidly exchanging amide protons.

Calmodulin

A basic premise of the assignment algorithm is that unassigned gaps of residues, comprising sequences that share internal homologies, can be assigned by virtue of their linkages to flanking segments of uniquely assigned sequences. To test this idea, we generated simulated cross-peak lists using the published assignments of *Drosophila melanogaster* calmodulin (Ikura et al., 1990,1991b) and used them as input to the algorithm. Calmodulin is a challenging protein to assign, because it contains four homologous segments that form the protein's four E-F hand calcium-binding motifs.

The results are illustrated in Fig. 4B. Three RIDs contained two sets of resonances due to the overlap of the amide ^{1}HN and ^{15}N chemical shifts: Ile^{9}/Asn^{97}, Asn^{60}/Glu^{87} and Ala^{88}/Met^{144}. After running the Stage 3A macro, 114 out of the 144 assignable residues (79%) were unambiguously assigned. Although the four segments with the greatest homology (Asp^{22}-Gly^{25}, Asp^{58}-Gly^{61}, Asp^{95}-Gly^{98} and Asp^{131}-Gly^{134}; see box in Fig. 4B) were not completely assigned after the application of the Stage 3A macro, the presence of assigned RIDs flanking these segments allowed the gaps to be readily assigned during Stage 3B.

The largest unassigned gaps remaining after Stage 3A correspond to α-helical segments. The presence of these gaps was due to the large number of residues with ^{13}C$^{\beta}$ resonances in the 30 ppm

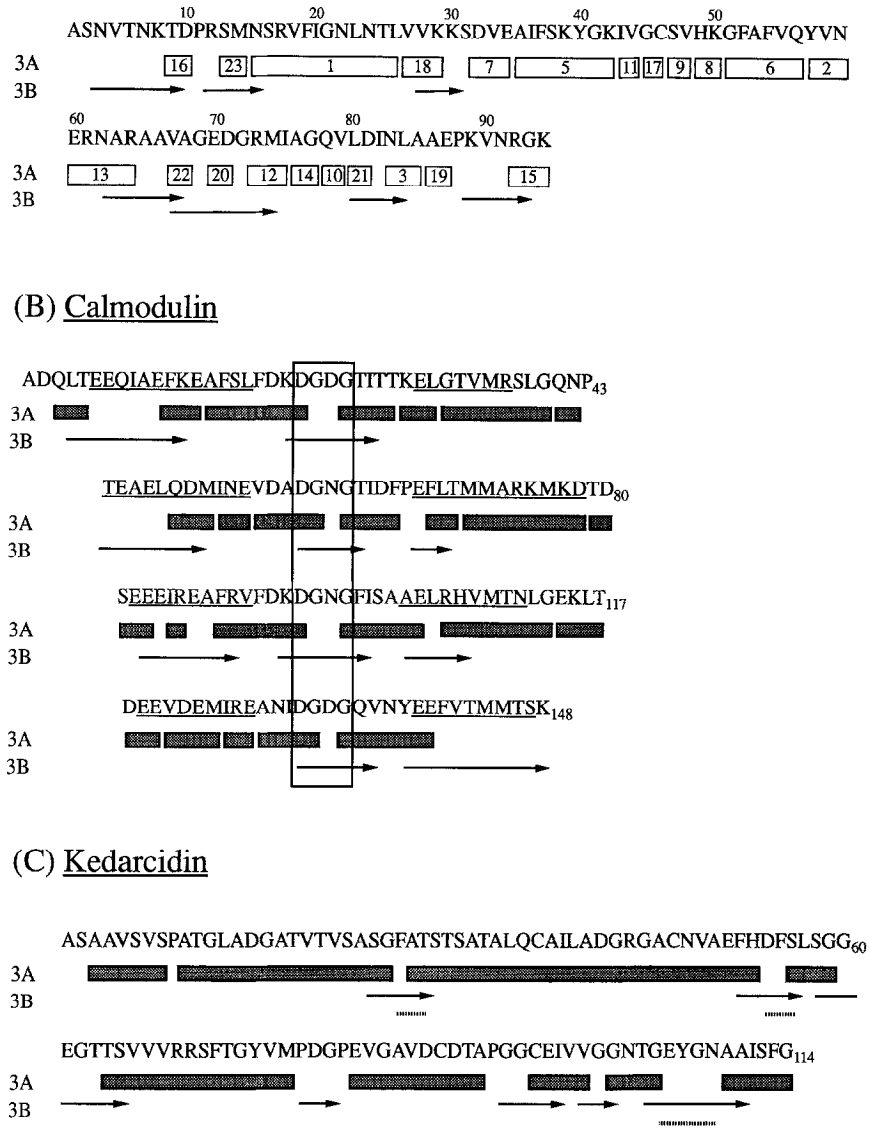

Fig. 4. Summary of the results obtained by running the Stage 3A and 3B macros with data from three different proteins. The blocks and arrows represent segments assigned with the Stage 3A and 3B macros, respectively. For Stage 3A, W was set to 5.0 ppm. During Stage 3B, W was set to 2.0 ppm and Z was set to 2.0 (see text for a description of the W and Z parameters). (A) Human hnRNP C RNA-binding domain. Real data from this previously assigned protein (Wittekind et al., 1992) was used as input to test the algorithm. The numerical labels on the open blocks correspond to the fragment numbers listed in Table 2. (B) Calmodulin. Input to the algorithm was simulated data, consisting of cross-peak lists calculated based on the published NMR assignments of *Drosophila melanogaster* calmodulin (Ikura et al., 1990, 1991a,b). The open box encloses the most highly conserved segments of the repeated E-F hand motifs. The sequences that form the α-helical regions are underlined. (C) Apokedarcidin. Real data was collected and analyzed for this previously unassigned protein (Constantine et al., manuscript in preparation). The dotted lines indicate segments that have alternate assignments due to spectral doubling (see text).

range and $^{13}C^{\alpha}$ resonances in the 57 ppm range (Grzesiek and Bax, 1993). As a result, a significant number of RIDs corresponding to these amino acids had multiple links that precluded their assignment during Stage 3A. They were, however, successfully assigned during Stage 3B.

Although factors that complicate the analysis of real data, such as missing or overlapped peaks, are not included in this test case, the results demonstrate that the Stage 3A macro can be used to assign proteins containing internally homologous segments. Even though the entries for sequential connectivities in the simulated calmodulin cross-peak lists align perfectly, the linking algorithm (Stage 2) was performed with the same tolerances used for the hnRNP C RBD case, where real data was analyzed (0.5 and 0.05 ppm were used for ^{13}C and ^{1}H, respectively). During the Stage 2 linking process, it makes no difference whether the cross peaks align perfectly or not; the same correct and incorrect RID–RID links are made. Since no penalties for resonance residuals are included at Stage 3A, the calmodulin test case is a valid test of the assignment capability of the Stage 3A macro.

In contrast, the links representing the correct linkages receive no penalty contribution during Stage 3B due to resonance matching residuals, making the calmodulin situation a less appropriate test of the robustness of the Stage 3B macro. However, the only cases where the assignment scores for competing fragments were very similar arose during the assignment of the gaps at Gly^{23} and Asn^{60}. Although the $^{13}C^{\alpha}$ resonances overlapped completely, the difference between the true Gly^{23} assignment and the next best assignment (Gly^{96}) was over 0.1 ppm for the $^{1}H^{\alpha}$ resonances. A difference this large would be discernible at the resolution used for collection of the ^{1}H dimensions of the 3D HNHA(Gly) experiment. Correct assignment of the Asn^{60} gap was complicated by the possible competing assignment of the RID corresponding to superposition of the Ile^{9} and Asn^{97} spin systems. The correct $^{1}H^{\alpha}$, $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifs of Asn^{60} (4.67, 52.7 and 37.7 ppm) are very close to the combination of the $^{1}H^{\alpha}$ and $^{13}C^{\alpha}$ resonances of Asn^{97} (4.68 and 52.8 ppm) and the $^{13}C^{\beta}$ resonance of Ile^{9} (37.8 ppm). The problem was resolved by not assigning the Asn^{60} position until after either the Ile^{9} or the Asn^{97} position was assigned.

Apokedarcidin

The algorithm was used to assign apokedarcidin, a 114-residue protein. Kedarcidin is a chromophore-containing protein that exhibits in vivo antitumor activity (Hofstead et al., 1992), and the apo-protein has recently been shown to possess proteolytic activity (Zein et al., 1993). This protein is unusual in that it has 18 glycine residues, including three occurrences of Gly-Gly pairs. The number of RIDs generated by the Stage 1 macro was 154. Ten of these contained only a few isolated peaks, representing side chains or noise artifacts and hence were discarded. Another 10 of the RIDs were manually edited to correct inconsistent or missing assignments, or to remove spurious peaks.

The remaining extra RIDs most likely arose from residues in the contaminating holo-protein, whose chemical shifts are perturbed from their apo-protein values by the presence of the chromophore. In general, the peaks comprising these RIDs have weaker intensities than the RIDs arising from apokedarcidin. However, the differences in intensities were not great enough to allow these extra RIDs to be ignored at this stage, and hence they were retained throughout the later analysis (see below). Slightly tighter tolerances (0.4 and 0.03 ppm for ^{13}C and ^{1}H, respectively) than those used for the other two examples were used during Stage 2 for apokedarcidin.

The results obtained by application of the Stage 3A and Stage 3B macros are shown in Fig. 4C.

Ninety-two of the 108 assignable residues (85%) were assigned by the Stage 3A macro. The remaining gaps were readily assigned during Stage 3B. All four cysteine residues of apokedarcidin have downfield shifted $^{13}C^{\beta}$ resonances (Cys^{37}: 50.5 ppm; Cys^{47}: 47.7 ppm; Cys^{88}: 39.9 ppm; Cys^{95}: 46.7 ppm), indicating that these residues are involved in disulfide bonds.

The algorithm successfully handled the spectral heterogeneity arising from the holo-protein contamination. The cross peaks of weaker intensity gave rise to RIDs that were linked to other RIDs containing cross peaks of similar intensities. These linked fragments were aligned with segments of the protein that were also sequentially assigned to other RIDs, comprising cross peaks of normal intensity. We interpret these results as defining two alternate assignments for these segments of the protein sequence; one for the apo-protein and the other for the holo-protein (see dotted lines in Fig. 4C). This demonstrates that the strategy to consider all possible assignments consistent with the data is a powerful approach. The resonance assignments of apokedarcidin will be published elsewhere (Constantine, K. et al., manuscript in preparation).

DISCUSSION AND CONCLUSION

The choice of NMR experiments was based on considerations regarding the sensitivity and resolution of spectra for medium- to large-sized proteins. It was important to select experiments that are relatively sensitive, since signal-to-noise ratios are often reduced for larger proteins. Therefore, we chose experiments that transfer magnetization via the relatively small intraresidue, one-bond coupling between the $^{13}C_i^{\alpha}$ and $^{15}N_i$ nuclei when the ^{15}N magnetization is transverse, capitalizing on the favorable ^{15}N relaxation properties relative to those of $^{13}C^{\alpha}$ (Clubb et al., 1992; Wittekind and Mueller, 1993). Also, correlations across the peptide bond are accomplished with high-sensitivity experiments that transfer magnetization through the carbonyl nuclei (Ikura et al., 1990; Grzesiek and Bax, 1992a).

Increases in spectral resolution will enhance the ability of automated procedures to assign spectra. The experimental data for the examples described here were recorded with acquisition parameters outlined in the published accounts of the pulse sequences. The incorporation of gradient technology (Bax and Pochapsky, 1992) to these pulse sequences (Farmer, B.T. and Mueller, L., unpublished results) allows for fewer phase cycling steps. This provides the opportunity to increase the acquisition times in some of the indirectly detected dimensions, resulting in spectra of higher resolution recorded in the same amount of time.

Data from the two 4D experiments were included so that large proteins exhibiting spectral overlap could be handled. These two experiments are primarily used to obtain the $^{1}H^{\alpha}$ resonances. If smaller proteins (molecular weight < 10 kDa) are being studied or if particularly good chemical-shift dispersion is observed in the 2D ^{1}H-^{15}N HSQC spectrum, then the $^{1}H_i^{\alpha}$ resonance can be obtained from either 3D H(CA)NNH (Kay et al., 1991), HN(CA)HA (Clubb et al., 1992) or 3D ^{15}N-edited TOCSY-HSQC-type (Fesik and Zuiderweg, 1988; Marion et al., 1989) experiments. By using the 3D instead of the 4D experiments, the total acquisition time can be substantially reduced.

The algorithm presented here for the automation of the assignments of the backbone, $^{13}C^{\beta}$ and $^{1}H^{\beta}$ resonances is fairly robust. For example, the application of the 2D inner product to gauge the alignment of peaks identifies many of the spurious peaks, which are then removed from consideration. In addition, the algorithm can accommodate multiple or missing assignments for an atom

type. The algorithm can handle a single missing atom assignment because each RID is linked to its sequential neighbor by three resonances, but usually only two are required to keep the number of competing alignments at a manageable level. For situations in which an atom type has more than one possible assignment, the alignment of the fragment's $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift profiles with those of the protein is usually sufficient to resolve the ambiguity. Since the manual resolution of these problems often requires a significant amount of human effort, the algorithm, in addition to removing tedious bookkeeping, is able to significantly reduce the time required for the backbone assignment process.

The portion of the algorithm requiring the most human input is the grouping of peaks into RIDs and their classification by atom type (the Stage 1 macro). The macro applies a simple, intuitive, rule-based approach that attempts to mimic the deductive logic employed by the user. Under the current set of rules, the macro is successful for about 90% of the RIDs. This statistic is slightly misleading, since the user should examine all aberrant RIDs (10–30% of the total number) to confirm the decisions made by the macro.

The success of the alignment process is contingent upon the lengths of the RID fragments and the degree of degeneracy in the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical-shift profiles of the protein. The length of the fragment determines the number of constraints available for the alignment – longer fragments are better. However, as recently noted (Grzesiek and Bax, 1993a,b), fragment lengths of three and even two are often sufficient to allow alignment with the sequence. In large part, this can be attributed to the relatively large dispersion in the $^{13}C^{\beta}$ chemical-shift values. In addition, the algorithm outlined here utilizes the sequential information of the RID at the N-terminus of the linked fragment, lengthening the probe length by one, resulting in increased assignment efficiency for small fragments.

Overlap of the carbon chemical-shift profiles of segments within the protein sequence makes the alignment process more difficult. This becomes an especially important consideration when dealing with proteins exhibiting significant internal sequence homology. However, as demonstrated with the simulated data set from calmodulin, the algorithm can be used to successfully assign a protein of this type.

Finally, the $^{13}C^{\alpha}$ or $^{13}C^{\beta}$ chemical shifts of a residue may lie outside the ranges shown in Table 1 for a variety of reasons (e.g., proximity of an aromatic ring, or secondary structure conformation). For fragments of only two or three RIDs, the resulting alignment penalty may prevent the automated assignment from being made. For larger fragments, however, the contribution of each residue to the alignment can be monitored. Therefore, if only one residue is making a significant contribution to the total score, the alignment can be checked by the user. In these problem cases, the user can widen the tolerances for the ^{13}C chemical-shift ranges and run the macro again. Also, the assignment of the segment can be delayed until most RIDs have been assigned, so that fewer unassigned RIDs are available as choices.

The current version of the algorithm has several limitations. Because the six experiments involve recording the amide proton chemical shifts, proline residues are not directly represented. The $^{1}H_i^{\alpha}$, $^{13}C_i^{\alpha}$, $^{1}H_i^{\beta}$ and $^{13}C_i^{\beta}$ chemical shifts of proline residue i can be obtained only after assignment of the non-proline i + 1 residue, via sequential cross-peak information. Other experiments will be required to assign segments containing proline-proline pairs.

Improvements also can be made in the pattern-recognition capabilities of the Stage 1 and Stage 3 macros. For example, the Stage 1 macro could be modified to recognize the overlap of $^{13}C^{\alpha}$ and

$^{13}C^{\beta}$ chemical shifts in RIDs containing serines. In addition, situations in which two peaks overlap and the peak maxima are shifted could be explicitly recognized and handled by the macro. Thus far, however, these cases have proven to be relatively rare. As the size of proteins that can be structurally determined by NMR increases, the programming effort required to reliably handle situations like these may become warranted.

The pattern-recognition capabilities of the Stage 3 macros could also be enhanced. The $^{13}C^{\alpha}$ chemical shifts of α-helical segments are shifted downfield relative to their random-coil resonances (Spera and Bax, 1991). In the current version of the program, the penalty is weighted lightly, so that fragments encoding α-helical segments are not overpenalized. In future versions, a provision could be made for recognizing and exempting systematic carbon chemical-shift deviations along a fragment. Alternatively, the allowed chemical-shift ranges could be locally adjusted on the basis of results obtained from NOE and J-coupling data or from secondary structure prediction algorithms. In addition, it should be straightforward to automatically assign the aliphatic portions of side chains by aligning the $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{1}H^{\alpha}$ and $^{1}H^{\beta}$ resonances with cross peaks in the HCCH-COSY and HCCH-TOCSY spectra (Bax et al., 1990; Olejniczak et al., 1992b) or from amide-correlated carbonyl-assisted ^{13}C TOCSY experiments (Logan et al., 1992; Montelione et al., 1992; Clowes et al., 1993; Grzesiek et al., 1993). This would confirm the established backbone assignments, and aid the assignment of gaps during the alignment process.

In conclusion, although we cannot claim to have rigorously completely automated the assignment process of the protein's backbone nuclei, the required time has been dramatically reduced. Instead of taking weeks, the backbone assignments can be made in one or two days following data acquisition and processing; almost all of this time is spent checking the peak picks and verifying the Stage 1 results. As an added benefit, implementation of the algorithm within a relational database simplifies the bookkeeping chores and expedites the transition to the assignment of the side-chain atoms and NOE spectra.

ACKNOWLEDGEMENTS

We thank Mitsu Ikura and Beverly Seavy for providing the files containing the calmodulin NMR assignments. We thank Kim Colson for helping with the assignment of apokedarcidin. We also thank Keith L. Constantine, Bennett T. Farmer II, and William J. Metzler for their critical reading of the manuscript.

NOTE ADDED IN PROOF

While this manuscript was being reviewed, Meadows et al. (*J. Biomol. NMR*, **4**, 79–96, 1994), Hare and Prestegard (*J. Biomol. NMR*, **4**, 35–46, 1994) and Olson and Markley (*J. Biomol. NMR*, **4**, 385–410, 1994) published papers on automated protein NMR assignment methods.

REFERENCES

Bax, A., Clore, G.M. and Gronenborn, A.M. (1990) *J. Magn. Reson.*, **88**, 425–431.
Bax, A. and Pochapsky, S.S. (1992) *J. Magn. Reson.*, **99**, 638–643.
Bernstein, R., Cieslar, C., Ross, A., Oschkinat, H., Freund, J. and Holak, T.A. (1993) *J. Biomol. NMR*, **3**, 245–251.

Boucher, W., Laue, E.D., Campbell-Burk, S.L. and Domaille, P.J. (1992) *J. Am. Chem. Soc.*, **114**, 2262–2264.
Campbell-Burk, S.L., Domaille, P.J., Starovasnik, M.A., Boucher, W. and Laue, E.D. (1992) *J. Biomol. NMR*, **2**, 639–646.
Catasti, P., Carrara, E. and Nicolini, C. (1990) *J. Comput. Chem.*, **11**, 805–818.
Cieslar, C., Clore, G.M. and Gronenborn, A.M. (1988) *J. Magn. Reson.*, **80**, 119–127.
Cieslar, C., Holak, T.A. and Oschkinat, H. (1990) *J. Magn. Reson.*, **87**, 400–407.
Clore, G.M., Bax, A., Driscoll, P.C., Wingfield, P.T. and Gronenborn, A.M. (1990) *Biochemistry*, **29**, 8172–8184.
Clowes, R.T., Boucher, W., Hardman, C.H., Domaille, P.J. and Laue, E.D. (1993) *J. Biomol. NMR*, **3**, 349–354.
Clubb, R.T., Thanabal, V. and Wagner, G. (1992) *J. Biomol. NMR*, **2**, 203–210.
Constantine, K.L., Goldfarb, V., Wittekind, M., Friedrichs, M.S., Anthony, J., Ng, S.-C. and Mueller, L. (1992) *J. Biomol. NMR*, **3**, 41–45.
Domaille, P. (1991) Presentation at the Eastern Analytical Symposium, Sommerset, NJ.
Eads, C.D. and Kuntz, I.D. (1989) *J. Magn. Reson.*, **82**, 467–482.
Eccles, C., Güntert, P., Billeter, M. and Wüthrich, K. (1991) *J. Biomol. NMR*, **1**, 111–130.
Fesik, S.W. and Zuiderweg, E.R.P. (1988) *J. Magn. Reson.*, **78**, 588–593.
Fesik, S.W. and Zuiderweg, E.R.P. (1990) *Q. Rev. Biophys.*, **23**, 97–131.
Gao, X. and Burkhart, W. (1991) *Biochemistry*, **30**, 7730–7739.
Griesinger, C., Sørensen, O.W. and Ernst, R.R. (1987) *J. Magn. Reson.*, **73**, 574–579.
Grzesiek, S. and Bax, A. (1992a) *J. Am. Chem. Soc.*, **114**, 6291–6293.
Grzesiek, S. and Bax, A. (1992b) *J. Magn. Reson.*, **99**, 201–207.
Grzesiek, S., Döbeli, H., Gentz, R., Garotta, G., Labhardt, A.M. and Bax, A. (1992) *Biochemistry*, **31**, 8180–8190.
Grzesiek, S., Anglister, J. and Bax, A. (1993) *J. Magn. Reson.*, **101**, 114–119.
Grzesiek, S. and Bax, A. (1993a) *J. Biomol. NMR*, **3**, 185–204.
Grzesiek, S. and Bax, A. (1993b) *Acc. Chem. Res.*, **26**, 131–138.
Hansen, P.E. (1991) *Biochemistry*, **30**, 10457–10466.
Hofstead, S.J., Matson, L.A., Malecko, A.R. and Marquardt, H. (1992) *J. Antibiot.*, **45**, 1250–1254.
Ikura, M., Kay, L.E. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.
Ikura, M., Kay, L.E., Krinks, M. and Bax, A. (1991a) *Biochemistry*, **30**, 5498–5504.
Ikura, M., Spera, S., Barbato, G., Kay, L.E., Krinks, M. and Bax, A. (1991b) *Biochemistry*, **30**, 9216–9228.
Kay, L.E., Ikura, M., Tschudin, R. and Bax, A. (1990a) *J. Magn. Reson.*, **89**, 496–514.
Kay, L.E., Clore, G.M., Bax, A. and Gronenborn, G.M. (1990b) *Science*, **249**, 411–414.
Kay, L.E., Ikura, M. and Bax, A. (1991) *J. Magn. Reson.*, **91**, 84–92.
Kay, L.E., Wittekind, M., McCoy, M.A., Friedrichs, M.S. and Mueller, L. (1992) *J. Magn. Reson.*, **98**, 443–450.
Kleywegt, G.J., Boelens, R., Cox, M., Llinás, M. and Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 23–47.
Kleywegt, G.J., Vuister, G.W., Padilla, A., Knegtel, R.M., Boelens, R. and Kaptein, R. (1993) *J. Magn. Reson.*, **102**, 166–176.
Logan, T.M., Olejniczak, E.T., Xu, R.X. and Fesik, S.W. (1992) *FEBS Lett.*, **314**, 413–418.
Logan, T.M., Olejniczak, E.T., Xu, R.X. and Fesik, S.W. (1993) *J. Biomol. NMR*, **3**, 225–231.
Lyons, B.A., Tashiro, M., Cedargren, L., Nilsson, B. and Montelione, G.T. (1993) *Biochemistry*, **32**, 7839–7845.
Marion, D., Kay, L.E., Sparks, S.W., Torchia, D.A. and Bax, A. (1989) *J. Am. Chem. Soc.*, **111**, 1515–1517.
Montelione, G.T., Lyons, B.A., Emerson, S.D. and Tashiro, M. (1992) *J. Am. Chem. Soc.*, **114**, 10974–10975.
Nelson, S.J., Schneider, D.M. and Wand, A.J. (1991) *Biophys. J.*, **59**, 1113–1122.
Oh, B.H., Westler, W.M., Darba, P. and Markley, J.L. (1988) *Science*, **240**, 908–911.
Olejniczak, E.T., Xu, R.X., Petros, A.M. and Fesik, S.W. (1992a) *J. Magn. Reson.*, **100**, 444–450.
Olejniczak, E.T., Xu, R.X. and Fesik, S.W. (1992b) *J. Biomol. NMR*, **2**, 655–659.
Pelton, J.G., Torchia, D.A., Meadow, N.D., Wong, C.-Y. and Roseman, S. (1991) *Biochemistry*, **30**, 10043–10057.
Richarz, R. and Wüthrich, K. (1978) *Biopolymers*, **17**, 2133–2141.
Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.
Van de Ven, F.J.M. (1990) *J. Magn. Reson.*, **86**, 633–644.
Vuister, G.W. and Boelens, R. (1987) *J. Magn. Reson.*, **73**, 328–333.
Weber, P.L., Malikayil, J.A. and Mueller, L. (1988) *J. Magn. Reson.*, **82**, 419–426.

Wehrens, R., Lucasius, C., Buydens, L. and Kateman, G. (1993) *J. Chem. Inf. Comput. Sci.*, **33**, 245–251.
Wittekind, M., Görlach, M., Friedrichs, M.S., Dreyfuss, G. and Mueller, L. (1992) *Biochemistry*, **31**, 6254–6265.
Wittekind, M., Metzler, W.J. and Mueller, L. (1993) *J. Magn. Reson.*, **101**, 214–217.
Wittekind, M. and Mueller, L. (1993) *J. Magn. Reson.*, **101**, 201–205.
Xu, J., Sanctuary, B.C. and Gray, B.N. (1993) *J. Chem. Inf. Comput. Sci.*, **33**, 475–489.
Zein, N., Casazza, A.M., Doyle, T.W., Leet, J.E., Schroeder, D.R., Solomon, W. and Nadler, S.G. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 8009–8012.
Zimmerman, D.E., Kulikowski, C.A., Wang L., Lyons, B. and Montelione, G.T. (1994) *J. Biomol. NMR*, **4**, 241–256.

APPENDIX

Stage 1

The details of the Stage 1 macro are presented here. The macro begins by searching through the list of picked peaks for an unclassified peak p. It then collects all peaks that have approximately the same ^{1}HN and ^{15}N chemical-shift values as p from the first four spectra depicted in Fig. 1. To allow for the different resolutions among the spectra, each spectrum has two sets of tolerance bounds assigned to it. One set is a narrow tolerance range used for intraspectra searches, and the other set is a wider tolerance range used for interspectra searches.

The ^{1}HN and ^{15}N tolerances are initially set to the appropriate lower bound. The macro then gradually increases the tolerances until the upper bound is reached or the expected number of peaks is found (e.g., four peaks in the HNCACB spectrum for non-glycine RIDs). The major advantage of this strategy is that peaks belonging to a common RID are closely aligned along the ^{1}HN and ^{15}N dimensions, and the macro collects these peaks using a minimal tolerance. As a consequence, nearby but extraneous peaks are ignored. A disadvantage to this approach is that not all valid RIDs have the same number of peaks from a particular spectrum. For example, RIDs in which at least one pair of peaks overlap will have fewer than the expected number of peaks, so that the widest tolerances will be used in the search.

An inner product measure of the alignment in the ^{15}N and ^{1}HN dimensions between peak p and each of the other peaks, q, from the same spectrum as p is calculated next. The two-dimensional inner product formula is

$$IP = \frac{\sum_{i=1}^{N} p_i q_i}{\sqrt{\sum_{i=1}^{N} p_i^2}\ \sqrt{\sum_{i=1}^{N} q_i^2}} \qquad \text{(A1)}$$

where p_i is the intensity of p at grid point i of a rectangle centered at the peak's maximum and spanning the peak's footprint in the ^{15}N and ^{1}HN dimensions. The footprint is the rectangle enclosing the baseline contour of p in the two dimensions. N is the total number of data points comprising the rectangle; thus, $N = (n \times m)$, where n and m are the number of grid points spanning the ^{15}N and ^{1}HN dimensions of the footprint. Similarly, q_i is the intensity of q at grid point i of the rectangle centered at the maxima of q in the non-^{15}N and ^{1}HN dimensions and spanning the same data-point ranges in the ^{15}N and ^{1}HN dimensions as the rectangle enclosing p; this rectangle, therefore, does not necessarily coincide with the footprint of q. The two main advantages of applying a 2D inner product instead of two 1D products are the following: the 2D

product is more restrictive than two 1D inner products, and the number of points used in computing the 2D product is larger, resulting in a more reliable metric.

For well-aligned peaks, the absolute value of the inner product is close to 1.0, whereas for poorly aligned peaks this value is much less. Empirically, we have found that peaks are well aligned if the inner product score is 0.85 or higher. Peaks with an inner product value with p below 0.85 are removed from consideration for the current RID and placed back into the pool of unclassified peaks.

Next, to determine the $^{13}C_i^{\alpha}$, $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$, $^{1}H_i^{\alpha}$ and $^{1}H_{i-1}^{\alpha}$ frequencies, the peak positions among the different spectra are compared with each other. For example, if the ^{13}C chemical shift of a negative peak from the HNCACB spectrum matches the chemical shift of a peak in the CBCA(CO)HN spectrum, then the resonance must be the $^{13}C_{i-1}^{\beta}$ assignment. As in the above searches for peaks with common ^{1}HN and ^{15}N chemical shifts, the program initially uses a user-specified lower bound for the ^{13}C and ^{1}H (in the case of the 4D spectra) chemical-shift tolerances. The tolerances are then gradually increased until either a match is found or the user-specified upper bound is reached. The upper and lower bound tolerances were empirically determined; the algorithm is relatively insensitive to the exact values employed and therefore crude estimates are sufficient. If the atom types for each of the spectra are uniquely identified and all peaks have been classified, then the algorithm moves on to the next RID by finding a peak that has not been previously classified. The above steps are then repeated using the ^{15}N and ^{1}HN chemical shifts of this peak.

If the number of peaks with common ^{15}N and ^{1}HN chemical shifts is larger than expected for a particular spectrum, the algorithm will try to identify the extraneous peaks. For example, in the case of the CBCA(CO)HN spectrum, if there is at least one assignment for both the $^{13}C_{i-1}^{\alpha}$ and $^{13}C_{i-1}^{\beta}$, then any unclassified peaks can be placed back in the pool of unclassified peaks; these peaks either belong to another RID or are artifacts.

This approach, however, cannot be applied when unclassified peaks with negative intensities are present in the HNCACB spectrum. Only this spectrum contains the $^{13}C_i^{\beta}$ peak, which therefore cannot be classified by comparing its position to those of peaks in the other spectra. In this situation, the macro will attempt to identify the $^{13}C_i^{\beta}$ resonance by comparing the volumes of the peaks in the spectrum. If the volume of one of the negative peaks is much larger in absolute value than those of the other negative peaks and it is comparable in magnitude to the largest positive peak in the group of peaks under consideration from that spectrum, then it is classified as the $^{13}C_i^{\beta}$ resonance. Otherwise, the macro cannot reliably assign the resonance and it is omitted.

After the above filters have been applied, if more peaks than expected are still present from a particular spectrum and the chemical-shift assignment for any atom type is unique, then the inner product measure is used to identify extraneous peaks. This is useful at this point in the macro only if the original peak p used to collect the peaks in the current RID belongs to a different spectrum; the inner product calculation performed earlier would have eliminated spurious peaks in the RID from the same spectrum as p. The restriction that the assignment be unique is made to ensure that the assignments for the resulting RID are internally consistent. For nonunique assignments (e.g., two peaks assigned to the $^{13}C_i^{\alpha}$ resonance), only one peak is correct for the RID; however, the macro cannot identify this peak without additional information. Consequently, none of these peaks can be used as a 'basis' peak for the RID. For RIDs with more than one atom type with a unique peak assignment, any one of these peaks should be equally good in gauging the alignment.

The 2D inner product is calculated between the peak from which the unique assignment was obtained and the other peaks in the RID from the same spectrum. As in the inner product calculations discussed above, the alignment is measured in the ^{15}N and ^{1}HN dimensions. Low inner-product values identify peaks that are outliers to the current RID, and therefore these peaks are deleted from the list of peaks belonging to the RID. The inner product metric has proven to be effective in identifying peaks that are artifacts, that are members of other RIDs, or that are present due to noise.

The above discussion focused on cases in which an RID has more peaks from a spectrum than expected. We now turn to the situations in which an RID has fewer peaks from a spectrum than expected. This case is often due to a degeneracy or near-degeneracy between the intraresidue and sequential $^{13}C^{\alpha}$ or $^{13}C^{\beta}$ resonances. In these cases, the macro applies a simple set of rules in an attempt to isolate the degeneracy. The rules are slightly different for each spectrum, but similar in spirit. Below an example from the HNCACB spectrum is described.

Suppose three peaks in the HNCACB spectrum are aligned in the ^{15}N and ^{1}HN dimensions: two with positive intensities and one with a negative intensity. Three scenarios are possible: (i) the RID represents a glycine residue or a residue that is sequential to a glycine residue; (ii) the $^{13}C^{\alpha}$ chemical shifts of both positive peaks are in the range 51–60 ppm; or (iii) at least one $^{13}C^{\alpha}$ chemical shift is greater than 60 ppm. Glycine residues are easily recognized by their characteristic $^{13}C^{\alpha}$ chemical shift; therefore, the absence of a $^{13}C^{\beta}$ peak is readily handled. In the second case, the macro can safely surmise that the intraresidue and sequential $^{13}C^{\beta}$ chemical shifts are degenerate. This conclusion follows from the observation that $^{13}C^{\beta}$ resonances rarely occur in the interval 51–60 ppm, and consequently the missing $^{13}C^{\beta}$ peak is unlikely to be overlapping one of the $^{13}C^{\alpha}$ peaks. Therefore, the $^{13}C_i^{\beta}$ and $^{13}C_{i-1}^{\beta}$ resonances must be degenerate by default. This conclusion is verified by inspecting the data from the CBCA(CO)NNH experiment. In the third scenario, one has to consider the possibility that the RID may contain a serine residue; the $^{13}C^{\beta}$ resonance of serine typically falls within the interval 60–68 ppm and hence it may overlap the $^{13}C^{\alpha}$ resonances. Similar considerations apply for threonine residues. As a result, reliably determining which peak is actually the superposition of two resonances is difficult. If no conclusion can be made after comparing cross-peak positions among the various spectra, the $^{13}C_i^{\beta}$ assignment is allowed to 'float' between the multiple ppm values, as outlined elsewhere in this paper.