Journal of Biomolecular NMR, 9 (1997) 151–166
ESCOM

J-Bio NMR 416

Automated probabilistic method for assigning backbone resonances of (^{13}C,^{15}N)-labeled proteins*

Jonathan A. Lukin[a], Andrew P. Gove[b], Sarosh N. Talukdar[b] and Chien Ho[a,]**

[a]*Department of Biological Sciences and* [b]*Robotics Institute, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213-2683, U.S.A.*

Received 15 August 1996
Accepted 26 November 1996

Keywords: Three-dimensional NMR; Automated polypeptide resonance assignment

Summary

We present a computer algorithm for the automated assignment of polypeptide backbone and $^{13}C^{\beta}$ resonances of a protein of known primary sequence. Input to the algorithm consists of cross peaks from several 3D NMR experiments: HNCA, HN(CA)CO, HN(CA)HA, HNCACB, COCAH, HCA(CO)N, HNCO, HN(CO)CA, HN(COCA)HA, and CBCA(CO)NH. Data from these experiments performed on glutamine-binding protein are analyzed statistically using Bayes' theorem to yield objective probability scoring functions for matching chemical shifts. Such scoring is used in the first stage of the algorithm to combine cross peaks from the first five experiments to form intraresidue segments of chemical shifts $\{N_i,H_i^N,C_i^{\alpha},C_i^{\beta},C_i'\}$, while the latter five are combined into interresidue segments $\{C_i^{\alpha},C_i^{\beta},C_i',N_{i+1},H_{i+1}^N\}$. Given a tentative assignment of segments, the second stage of the procedure calculates probability scores based on the likelihood of matching the chemical shifts of each segment with (i) overlapping segments; and (ii) chemical shift distributions of the underlying amino acid type (and secondary structure, if known). This joint probability is maximized by rearranging segments using a simulated annealing program, optimized for efficiency. The automated assignment program was tested using CBCANH and CBCA(CO)NH cross peaks of the two previously assigned proteins, calmodulin and CheA. The agreement between the results of our method and the published assignments was excellent. Our algorithm was also applied to the observed cross peaks of glutamine-binding protein of *Escherichia coli*, yielding an assignment in excellent agreement with that obtained by time-consuming, manual methods. The chemical shift assignment procedure described here should be most useful for NMR studies of large proteins, which are now feasible with the use of pulsed-field gradients and random partial deuteration of samples.

Introduction

A necessary stage of any NMR investigation of protein structure or dynamics is the assignment of resonances to their originating nuclei. Resonance assignment, in contrast to the automated processing of raw data and modeling of solution protein structures, is often performed manually, making it perhaps the most tedious, time-consuming stage of NMR studies. In order to accelerate this process, several workers have proposed automated computer procedures for the assignment of 2D ^{1}H-NMR spectra (Cieslar et al., 1988; Weber et al., 1988; Eads and Kuntz, 1989; Van de Ven, 1990; Eccles et al., 1991; Kleywegt et al., 1991; Xu et al., 1994). Such methods generally use homonuclear NOESY and COSY spectra to identify and link amino acid spin systems with reference to the primary sequence of a protein. This technique fails for proteins larger than 10 kDa, as line broadening due to slower molecular tumbling contributes to spectral overlap while decreasing the sensitivity of the experiments. Even for proteins below this size limit, the poor resolution of proton chemical shifts can present problems for the automated pattern recognition of spin systems (Leopold et al., 1994).

*Preliminary accounts of the research presented in this paper were given at the 16th International Conference on Magnetic Resonance in Biological Systems, Veldhoven, The Netherlands, 1994, and at the 39th and 40th meetings of the Biophysical Society at San Francisco, CA, U.S.A., 1995, and Baltimore, MD, U.S.A., 1996. Our Fortran programs for automated assignment of NMR cross peaks are available on the World Wide Web page of the Pittsburgh NMR Center for Biomedical Research at http://info.bio.cmu.edu/NMR-Center/NMR.html.

**To whom correspondence should be addressed.

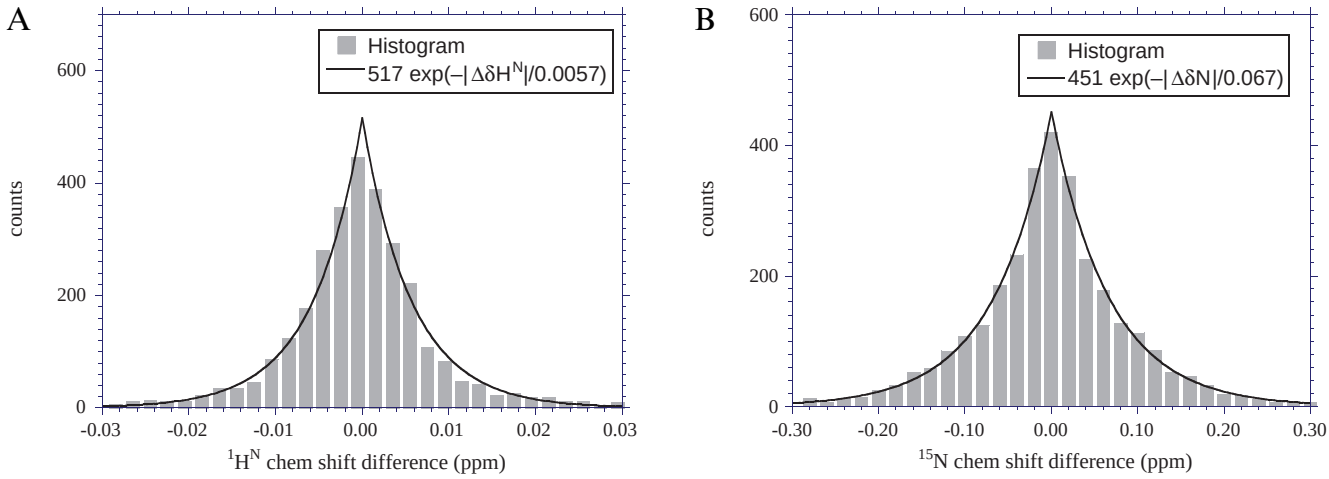

Fig. 1. Histograms of $^{1}H^{N}$ (A) and ^{15}N (B) chemical shift differences between closest-matching peaks in different 3D NMR experiments performed on GlnBP. The fitted curves are used to evaluate the probability that two peaks originate from the same nuclei.

The assignment of larger proteins has been made possible in recent years by the development of triple-resonance pulse sequences applicable to proteins labeled with ^{13}C and ^{15}N (for reviews, see Clore and Gronenborn (1991) and Bax and Grzesiek (1993)). The resonances of these nuclei are used to resolve spectra along a third and fourth frequency dimension, relieving the crowding seen in 2D spectra. At the same time, sensitivity is improved due to large one-bond heteronuclear J-couplings which allow for the efficient transfer of magnetization. These couplings have distinct values which enable the design of experiments to transfer magnetization along specific pathways in the polypeptide, correlating particular sets of ^{1}H, ^{13}C, and ^{15}N chemical shifts. Cross peaks from several 3D or 4D NMR experiments provide chemical shift coordinates which collectively cover the protein backbone. By matching cross peaks between experiments which detect overlapping sets of nuclei, an assignment can be extended in both directions along the main chain (Ikura et al., 1990). This method, unlike the assignment of homonuclear ^{1}H-NMR spectra, does not require the recognition of multiple-peak spin systems, so it is more amenable to automation.

In recent years, a number of strategies for the assignment of multinuclear NMR data have been proposed (some of these are discussed below; see Zimmerman and Montelione (1995) for a review). We have developed a suite of Fortran programs which automate the assignment of protein backbone and $^{13}C^{\beta}$ chemical shifts. Input to the programs consists of peak lists obtained from several 3D NMR experiments: HNCA, HNCO, HN(CO)CA (Grzesiek and Bax, 1992a), HN(CA)HA (Clubb et al., 1992a), HN(CA)CO (Clubb et al., 1992b), HNCACB (Wittekind and Mueller, 1993), COCAH (Dijkstra et al., 1994), HCA(CO)N (Powers et al., 1991), HN(COCA)HA (Clubb and Wagner, 1992), and CBCA(CO)NH (Grzesiek and Bax, 1992b). The likelihood that peaks from overlapping experiments originate from the same nuclei is evaluated using Bayes' theorem (Press, 1989; Press et al., 1992). This theorem is also applied to statistics relating chemical shift values and amino acid types, to yield the probability of assigning peaks to the protein's primary sequence. These two probability factors are combined to provide an objective overall score for a tentative peak assignment. This score is maximized by rearranging the assignment according to a simulated annealing protocol. The algorithm is tested using observed NMR cross peaks of the previously assigned proteins, calmodulin (Tjandra et al., 1995) and CheA (McEvoy et al., 1995), and applied to glutamine-binding protein (GlnBP) of *Escherichia coli* (Yu et al., 1997).

Materials and Methods

Experimental data: Glutamine-binding protein

The development of an automated assignment method was motivated by our ongoing NMR studies of GlnBP (Tjandra et al., 1992; Tjandra, 1993; Hing et al., 1994; Yu et al., 1995,1997). This 226-residue (25 kDa) monomeric protein is among the largest to be investigated by NMR. The 10 3D NMR experiments listed above were carried out on liganded, (^{13}C,^{15}N)-labeled GlnBP. Each of these experiments provides a characteristic set of three chemical shifts for approximately 226 cross peaks (one for each residue). By correlating pairs of cross peaks from different 3D experiments having two chemical shift dimensions in common, it is possible to 'walk along' the protein backbone (Ikura et al., 1990), connecting peaks from one residue to the next, towards either the amino or carboxyl end of the polypeptide. Definite starting points for the assignment are provided by experiments performed on specifically labeled protein, obtained by site-directed mutagenesis (Shen et al., 1989a,b; Tjandra et al., 1992).

Unfortunately, this straightforward assignment strategy is complicated by two factors arising from the large size of the protein (Leopold et al., 1994). First, the sheer number of cross peaks (broadened by the rapid relaxation of spins in a slowly tumbling macromolecule) results in significant overlap of the relevant 2D projections of the 3D spectra. Thus, for example, more than one HNCA peak is often found to match the (^{1}H,^{15}N) chemical shifts of a given HNCO peak to within reasonable tolerance. Such chemical shift degeneracy causes ambiguity in extending an assignment, resulting in an unmanageable number of branching possibilities. A second problem is related to the imperfect sensitivity of some 3D NMR experiments when applied to a large protein. Experiments such as HNCACB and HN(CA)CO, which rely on the relatively weak ^{1}J(^{15}N,^{13}C$^\alpha$) coupling, may yield fewer than the expected number of cross peaks. Glycine residues may not be detected by HN(CA)HA or HN(COCA)HA experiments (Clubb et al., 1992a; Clubb and Wagner, 1992), which are tuned for ^{13}C$^\alpha$ coupled to a single proton. None of the experiments which rely on amide ^{1}H-^{15}N coupling will detect proline residues. Amide protons in rapid exchange with the solvent may also go unobserved. Thus, there are several factors which can lead to 'missing peaks', causing the assignment process to come to a halt after only a short stretch of amino acid residues have been connected.

We attempt to overcome the problems of chemical shift degeneracy and missing peaks by performing a large number of 3D NMR experiments, which provide redundant coverage of the backbone and ^{13}C$^\beta$ nuclei. A deterministic, 'best-first' procedure is used to combine the cross peaks into segments of six chemical shifts, which then become the units for a probabilistic, Monte Carlo assignment algorithm. The first step in the treatment of data is to apply slight adjustments to the chemical shifts, in order to compensate for systematic differences due to varying temperature, pH, and isotope effects. The amide proton and nitrogen chemical shifts of 3D NMR experiments which detect these nuclei are aligned with 2D (^{1}H,^{15}N)-HSQC cross peaks. An initial, rough alignment is carried out 'by eye', using interactive plotting software. The alignment is then refined using programs (described below) which tabulate chemical shift differences between matching cross peaks. For example, the mean chemical shift differences between HNCA peaks and the closest-matching 2D HSQC peaks in the (^{1}H,^{15}N) plane indicate the amount by which the former spectrum must be shifted to match the latter.

Similarly, the ^{13}C$^\alpha$ chemical shifts of HNCA and HN(CO)CA peaks are aligned, using the fact that the former experiment often detects interresidue correlations which are duplicated in the latter (Ikura et al., 1990). After being used for alignment, the 'back-connecting' HNCA peaks are discarded. Similar treatment is given to the pairs of experiments [HN(CA)HA, HN(COCA)HA] and [HN(CA)CO, HNCO]. Then, the chemical shifts of the HCA(CO)N experiment (performed in D_2O solution) are adjusted to agree with cross peaks from the other experiments. Finally, HNCACB peaks are classified as C$^\alpha$ or C$^\beta$ peaks, and combined with HNCA, which usually has higher sensitivity, and better resolution in the ^{13}C dimension; similar treatment is accorded to CBCA(CO)NH and HN(CO)CA.

Chemical shift difference distributions

Even after an overall adjustment has been applied to chemical shifts of different experiments so that they align in an 'average' sense, the matches between individual cross peaks are rarely exact. In order to evaluate the probability that two 3D cross peaks originate from the same nuclei, it is useful to tabulate chemical shift differences between closest-matching pairs of peaks. A Fortran program has been written which finds, for each HNCA peak, the closest HNCO peak (matching on ^{1}H^N and ^{15}N chemical shifts) and records the differences between the corresponding chemical shifts. The procedure is repeated for all distinct pairs of experiments in the set [HNCA, HN(CA)HA, HN(CA)CO, HNCO, HN(CO)CA, HN(COCA)HA]. This yields a large number of chemical shift differences ($\Delta\delta$H^N, $\Delta\delta$N) which can be sorted into bins, and plotted as histograms (see Fig. 1).

The histograms were empirically fitted to curves of the form $P(\Delta x) = A \exp(-|\Delta x|/D)$, where D, the characteristic width of the distribution, is 0.0057 ppm for ^{1}H^N and 0.067 ppm for ^{15}N. The value of the fitted function $P(\Delta x)$ is proportional to the probability of observing a chemical shift difference Δx between two cross peaks which originate from the same nucleus. According to Bayes' theorem (Press, 1989), it follows that $P(\Delta x)$ is proportional to the likelihood that two peaks originate from the same nucleus, given that the observed chemical shift difference is Δx. Thus, any pair of peaks from different HN experiments can be given a probability score for matching, proportional to

$$P = A \exp\left(-\frac{|\Delta\delta H|}{0.0057}\right) \exp\left(-\frac{|\Delta\delta N|}{0.067}\right) \qquad (1)$$

based on their ^{1}H^N and ^{15}N chemical shift differences, with an arbitrary normalization A. Similar programs tabulate chemical shift differences between other pairs of experiments which detect two nuclei in common, including [COCAH, HCA(CO)N], [HCA(CO)N, HN(CO)CA], [HCA(CO)N, HN(COCA)HA], and [HNCACB, CBCA(CO)NH]. Small systematic shifts are applied to cross peaks from one member of each pair of experiments so that the final chemical shift difference distributions are centered at zero. Curves fitted empirically to these distributions allow for the objective scoring of matches be-

TABLE 1
WIDTHS (ppm) OF EXPONENTIAL DISTRIBUTIONS FITTED TO GlnBP CHEMICAL SHIFT DIFFERENCES

	N	H^N	H^α	C^α	C'	C^β
$H_2O \leftrightarrow H_2O$	0.067	0.0057	0.010	0.068	0.048	0.13
$H_2O \leftrightarrow D_2O$	0.12		0.014	0.092		

The first row refers to matches among experiments performed on protein in H_2O solution. Matches between such experiments and HCA(CO)N, which was performed on a D_2O solution, are of poorer quality; the corresponding distribution widths are listed in the second row.

tween cross peaks. The characteristic widths of these exponential fits are summarized in Table 1.

Connecting peaks to form segments

Given a set of 3D NMR cross peaks from the 10 experiments listed above, the first step in the assignment process is to combine them into longer segments of six chemical shifts each. First, peaks from experiments having three chemical shifts in common [HNCA and HNCACB; CBCA(CO)NH and HN(CO)CA] are combined. Then, as shown schematically in Fig. 2, cross peaks are matched in groups of four to yield intra- and interresidue segments, which then become the units of assignment. As a preliminary step, fitted distributions such as Eq. 1 are used to tabulate likely connections between cross peaks. For example, the five best HNCACB matches to each HN(CA)HA peak are recorded, together with their scores. Connections can then be built up among three experiments at a time. A program searches for the matched pair of [HNCACB, HN(CA)HA] peaks which most closely match the ($^{13}C^\alpha$,$^1H^\alpha$) chemical shifts of each COCAH peak. The chemical shift differences ($\Delta\delta C^\alpha$ and $\Delta\delta H^\alpha$) are tabulated, and fitted by curves of the form shown in Fig. 1, which can then be used (together with $\Delta\delta H^N$ and $\Delta\delta N$) to score the overall probability of the three-way match:

$$P_3 = A\exp\left(-\frac{|\Delta\delta H^N|}{0.0057}\right)\exp\left(-\frac{|\Delta\delta N|}{0.067}\right) \times \exp\left(-\frac{|\Delta\delta H^\alpha|}{0.010}\right)\exp\left(-\frac{|\Delta\delta C^\alpha|}{0.068}\right) \quad (2)$$

Using a similar strategy, matching probabilities are calculated for other combinations of three experiments, and finally for all four intra- and interresidue experiments.

After possible connections between sets of four cross peaks have been tabulated, a program uses a 'best-first' approach to find a set of matches satisfying uniqueness, i.e., each peak can be a member of only one segment. The uniquely matched segments are written in descending order of scores to a file; the chemical shifts are simply the averages of the contributing cross peaks. The user specifies a cutoff score such that segments with matching scores higher than this value are accepted. Cross peaks which are not members of these segments are classified as 'unmatched'. Next, within the pool of unmatched peaks, possible two- and three-way matches are tabulated. A program finds the best set of three-way matches satisfying uniqueness, and writes the combined segments in descend-

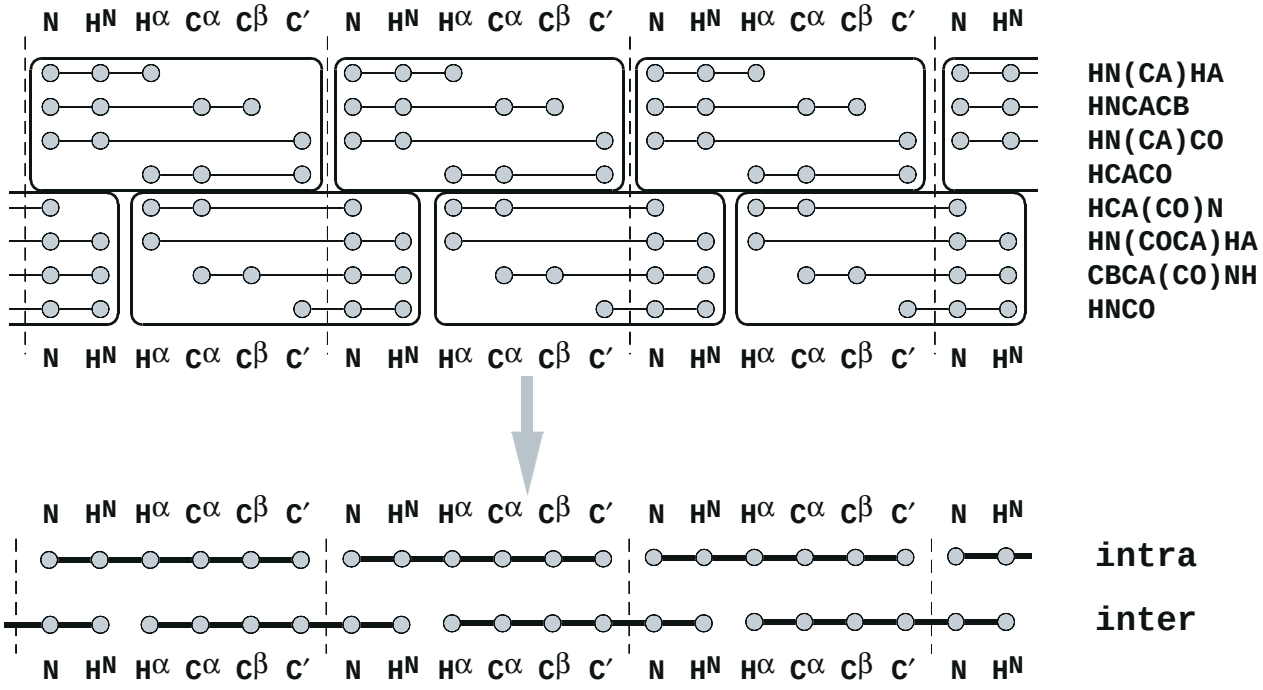

Fig. 2. Schematic diagram of 3D NMR experiments. Filled circles indicate the resonances detected by each experiment. Cross peaks from several experiments are combined as shown to yield intraresidue and interresidue 'segments' of six chemical shifts each.

TABLE 2
MEAN CHEMICAL SHIFTS USED TO EVALUATE ASSIGNMENT PROBABILITY

Amino Acid	Helix				Strand				Coil			
	$^{13}C'$	$^{13}C^{\alpha}$	$^{13}C^{\beta}$	^{15}N	$^{13}C'$	$^{13}C^{\alpha}$	$^{13}C^{\beta}$	^{15}N	$^{13}C'$	$^{13}C^{\alpha}$	$^{13}C^{\beta}$	^{15}N
Gly	175.35	46.71	0.00	108.50	171.87	44.97	0.00	110.24	173.96	45.41	0.00	110.03
Ala	179.37	54.77	18.40	122.02	175.90	51.15	21.02	126.11	177.30	52.42	19.03	125.10
Ser	175.69	61.31	62.98	116.19	173.60	57.11	65.08	118.02	174.41	58.27	64.14	116.43
Cys[a]	176.68	61.62	26.75	118.38	174.18	56.08	29.02	121.64	174.84	58.01	28.20	119.19
Met	178.04	58.34	32.87	118.56	175.53	54.53	35.13	122.41	175.45	55.34	33.00	120.39
Lys	178.51	59.03	32.19	120.18	175.07	55.23	34.75	123.79	176.39	56.59	32.62	121.82
Val	177.73	66.16	31.41	119.74	174.50	61.01	34.19	123.91	175.79	62.13	32.65	120.93
Thr	176.35	65.85	68.29	115.75	173.71	61.18	70.70	116.47	174.75	61.62	69.83	114.17
Ile	177.52	64.61	37.75	120.65	174.79	60.09	40.47	123.78	175.69	60.98	38.87	121.19
Leu	178.64	57.52	41.41	120.56	175.76	54.01	43.88	125.70	177.15	54.82	42.82	122.22
Asp	178.44	57.11	40.25	120.12	175.40	53.50	42.38	123.63	176.45	54.12	40.83	120.51
Asn	176.73	55.62	38.46	118.76	174.55	52.47	39.78	122.01	174.65	53.22	38.74	119.17
Glu	178.73	59.06	29.30	119.71	175.28	54.96	31.93	123.37	176.27	56.66	30.13	121.44
Gln	178.35	58.87	28.46	119.63	174.54	54.77	31.57	123.22	175.54	55.78	29.34	120.34
Arg	178.68	59.18	30.02	120.01	174.85	54.73	32.54	123.72	176.05	56.25	30.56	122.42
His	176.88	58.71	29.63	119.57	174.34	54.46	32.13	120.79	174.54	55.78	29.78	120.21
Phe	177.04	60.75	38.95	120.07	174.43	56.34	41.45	121.24	174.79	57.91	39.34	119.84
Tyr	177.09	60.82	38.57	119.69	174.22	56.57	41.19	122.01	175.80	57.77	38.88	120.41
Trp	178.14	59.51	29.26	120.52	174.78	56.49	30.98	124.74	175.85	57.50	29.09	120.19
Pro	179.45	65.28	30.90	136.12	175.56	62.71	32.03	135.54	176.60	63.27	32.09	136.73

See the text for a description of the database used.
[a] Reduced Cys.

ing order of scores. Again, segments with connection scores higher than a specified cutoff are accepted, while peaks which do not participate in them are classified as unmatched. These peaks are next considered for two-way matches. Finally, any cross peaks which do not comprise four-, three-, or two-way matches are kept individually.

This algorithm, by accepting multiple-peak matches first, gives priority to peak connections which are verified by the redundant coverage present in the large number of 3D NMR experiments. After four-way matches are set aside, fewer peaks remain available for three-way matches (and so on), thereby minimizing the ambiguity of connections. Two-way matches and individual peaks yield segments with 'holes', or missing chemical shifts. These are taken into account in the final connection step, where matches are tabulated between intra- and interresidue segments. Each segment is given an arbitrary, unique ID number. The fitted chemical shift difference distributions are used to tabulate, for each intraresidue segment, the six closest-matching interresidue segments, based on $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and $^{13}C'$ shifts. Similarly, each interresidue segment is matched to the six closest intraresidue segments, based on their common $^{1}H^{N}$ and ^{15}N chemical shifts. Possible matches and their scores are written to tables, which are used as input to the main assignment program.

Amino acid statistics

The procedure outlined above allows for the connection of overlapping cross peaks using an objective probability score, derived from chemical shift matches. Maximizing the overall connection probability would be sufficient to assign the peaks, given a complete data set with good chemical shift dispersion, and little variation between experiments. However, the missing peaks and chemical shift degeneracy seen in a large protein may lead to gaps or ambiguities when connecting peaks. Therefore, it is advantageous to use the information of the protein's primary sequence, together with statistics correlating chemical shifts with amino acid types, to establish an additional type of assignment probability. In particular, the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts have been shown to be useful in assigning residues by amino acid type (Grzesiek and Bax, 1993). These chemical shifts are also sensitive to local secondary structure (Spera and Bax, 1991; Wishart et al., 1991), so that a knowledge of α-helices and β-strands in the protein can provide additional information for their assignment.

To quantify the relationship between chemical shifts and amino acid and secondary structure types, we have analyzed a large database of protein NMR. We used the BioMagResBank database (Seavey et al., 1991) supplemented by data on several proteins scanned in from the recent published literature. The data consist of backbone and $^{13}C^{\beta}$ chemical shifts for over 40 distinct, non-paramagnetic proteins of known secondary structure, including about 4000 amino acid residues. Proton and carbon chemical shifts were adjusted where necessary for referencing to DSS at 0 ppm (Wishart and Sykes, 1994; Wishart et al., 1995). Relational database software (Microsoft FoxPro) was used to average the chemical shifts within

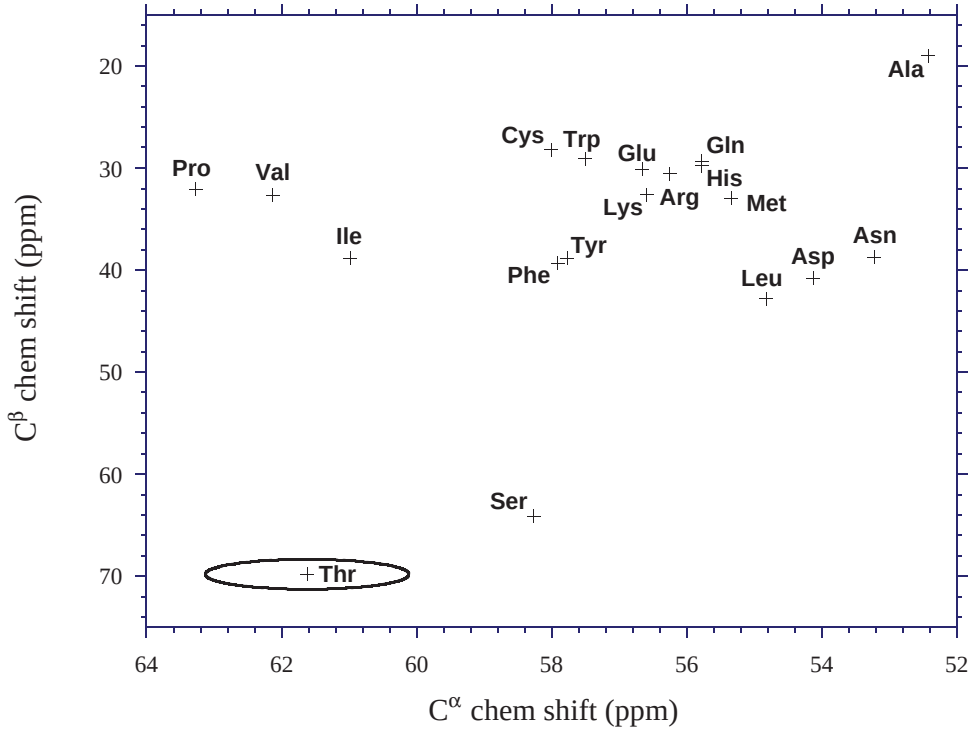

Fig. 3. Average $^{13}C^\alpha$ and $^{13}C^\beta$ chemical shifts for assigned residues in the coil conformation. The secondary chemical shifts (deviations from the mean value for the respective amino acid type) were fitted by a Gaussian distribution (see the text). One standard deviation of this distribution is indicated by the curve centered at threonine.

each of the 20 amino acids and three secondary structure types. These averages are shown in Table 2. Figure 3 illustrates the mean ($^{13}C^\alpha$,$^{13}C^\beta$) chemical shifts for the 19 amino acids (excluding glycine) in the coil conformation, defined as all residues which are not part of α-helices or β-strands.

The variation about the mean is characterized by computing the secondary chemical shift, defined as the difference between each individual chemical shift and the average value for the originating amino acid type, e.g.,

$$\delta_2^{i,j}C^\alpha \equiv \delta C^\alpha - \langle \delta C^{\alpha\ i,j} \rangle \tag{3}$$

where $\langle \delta C^{\alpha\ i,j} \rangle$ is the average $^{13}C^\alpha$ chemical shift for amino acid i and secondary structure j. The secondary chemical shifts for $^{13}C^\alpha$, $^{13}C^\beta$, $^{13}C'$, and ^{15}N were pooled together for all values of i and j, and fitted by a four-dimensional Gaussian distribution

$$G(\delta_2 C^\alpha, \delta_2 C^\beta, \delta_2 C', \delta_2 N) = \exp\left\{-\frac{1}{2}\left[\left(\frac{\delta_2 C^\alpha}{\sigma_{C^\alpha}}\right)^2 + \left(\frac{\delta_2 C^\beta}{\sigma_{C^\beta}}\right)^2 + \left(\frac{\delta_2 C'}{\sigma_{C'}}\right)^2 + \left(\frac{\delta_2 N}{\sigma_N}\right)^2\right]\right\} \tag{4}$$

with the fitted standard deviations $\sigma_{C^\alpha} = 1.42$ ppm, $\sigma_{C^\beta} = 1.31$ ppm, $\sigma_{C'} = 1.32$ ppm, and $\sigma_N = 4.07$ ppm. The secondary chemical shift distributions and fitted curves are shown in Fig. 4.

The ellipse in Fig. 3 indicates one standard deviation of the fitted Gaussian distribution in the ($^{13}C^\alpha$,$^{13}C^\beta$) plane about the mean chemical shifts of threonine, in the coil conformation. Approximately 39% of the chemical shifts originating at threonine residues are expected to fall within this curve; an identical distribution about the mean is assumed for each of the other amino acid types. The use of $^{13}C'$ and ^{15}N statistics (not shown) adds somewhat to the 'discriminating power' of the chemical shifts. The fitted Gaussian distribution can be interpreted as the probability of observing a set of chemical shift values from a given residue type (we use the term 'residue type (i,j)' as shorthand for amino acid type i and secondary structure type j). Using Bayes' theorem, this can be inverted to give the relative probability $P_{i,j}$ that an observed set of chemical shifts originated from a residue of type (i,j):

$$P_{i,j} \propto G(\delta_2^{i,j}C^\alpha, \delta_2^{i,j}C^\beta, \delta_2^{i,j}C', \delta_2^{i,j}N) \tag{5}$$

where each of the secondary chemical shifts $\delta_2^{i,j}C^\alpha$, etc. is the difference between the observed chemical shift and the mean value for the residue type at the given sequence position.

The probability of assigning a cross peak to any of the positions of a protein's sequence can be evaluated using Eq. 5, provided the detailed secondary structure is known. This information is sometimes available from previous

crystallographic studies of the protein now under NMR investigation, or of homologous proteins. On the other hand, the secondary structure may not be available at the stage of chemical shift assignment. In that case, it is often possible to estimate the fractional secondary structure content of the protein by circular dichroism (CD) spectroscopy or by counting the number of NMR cross peaks within certain ranges of chemical shifts (Wishart and Sykes, 1994). Suppose that, of all the residues of a protein, a fraction f_1 are found in helices, f_2 in β-strands, and f_3 in coils, where $f_3 = 1 - f_1 - f_2$. Then, the probability of observing a given $^{13}C^{\alpha}$ chemical shift from an amino acid of type i is proportional to

$$P_i = \sum_{j=1}^{3} f_j G_{i,j}\left(\delta_2^{i,j} C^{\alpha}\right) \qquad (6)$$

with similar expressions for other nuclei. As above, this equation is interpreted as the relative probability that the given chemical shift originates from an amino acid type i.

Assignment by simulated annealing

Using the procedure described above, we can combine 3D NMR cross peaks into longer intra- and interresidue segments, and evaluate the probability of (i) linking overlapping segments; and (ii) assigning a segment to a given position along the protein backbone. By arranging segments so as to maximize this overall probability, the optimal assignment of chemical shifts can be obtained. The best arrangement of segments is determined using simulated annealing (Kirkpatrick et al., 1983; Press et al., 1992), a versatile technique suitable for large-scale optimization problems where a desired global extremum may be hidden among many poorer, local extrema. The problem to be solved is formulated so that its solution corresponds to the minimum of a target function, the 'energy', calculable from the state of the system.

By an analogy with thermodynamics, the simulated annealing (or Monte Carlo) method allows the system to sample all configurations consistent with a defined value of 'temperature' T. Under the usual Metropolis algorithm (Metropolis et al., 1953), a random change (Monte

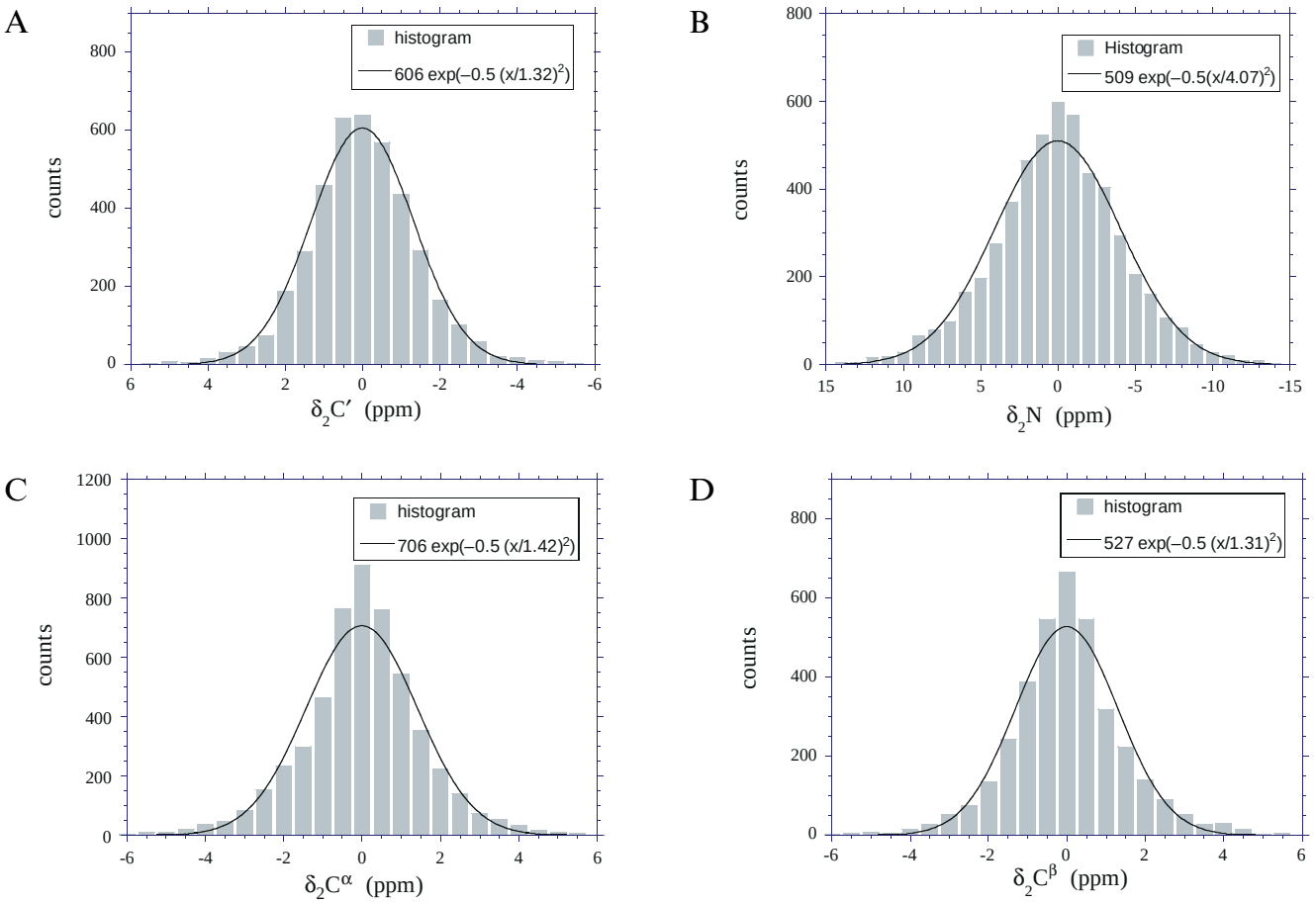

Fig. 4. Histograms of $^{13}C'$ (A), ^{15}N (B), $^{13}C^{\alpha}$ (C), and $^{13}C^{\beta}$ (D) secondary chemical shifts, with fitted Gaussian curves. For each chemical shift in the database described in the text, the secondary shift was calculated by subtracting the mean value for the corresponding amino acid and secondary structure type.

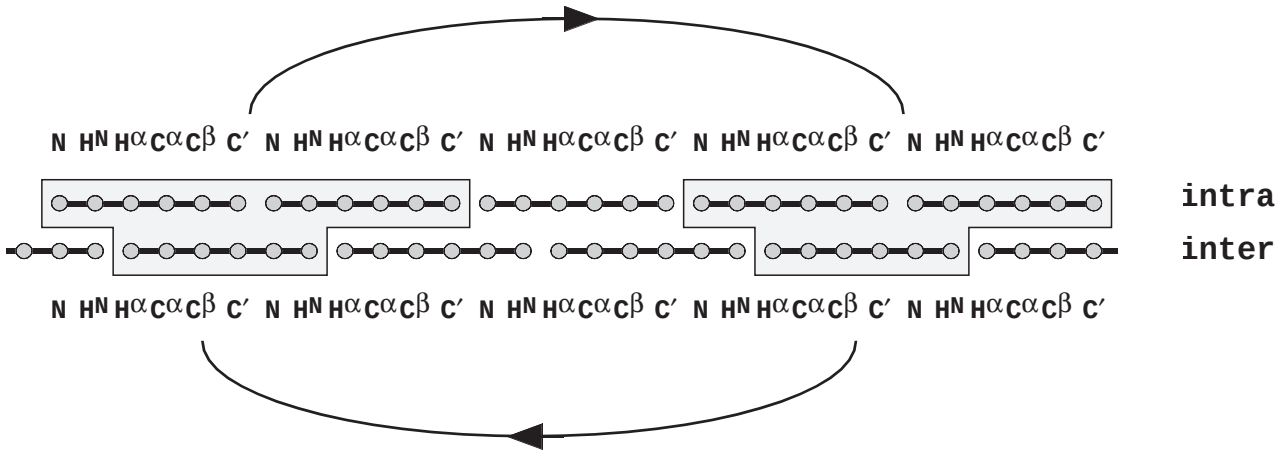

Fig. 5. The assignment of intra- and interresidue segments is optimized by simulated annealing. A continuous block of segments (from a single segment up to a specified maximum length) is chosen at random. A nonoverlapping, identically shaped block is chosen elsewhere along the protein sequence. A proposed Monte Carlo move consists of swapping the chemical shift assignments of the two blocks; if the swap violates any constraints, it is rejected.

Carlo move) is proposed in the configuration of the system, and the resulting change in energy ΔE is calculated. If $\Delta E > 0$, the proposed change is accepted with probability $\exp(-\Delta E/kT)$; if $\Delta E \leq 0$, it is always accepted. As independent Monte Carlo moves are generated at a given temperature, a Boltzmann distribution is approached, where the probability of the system having an energy E is given by $P_B(E) \propto \exp(-E/kT)$. Thus, at infinite temperature (effectively, a temperature much higher than the characteristic energies of the system), all states of the system are equally likely since all proposed Monte Carlo moves are accepted. As T is slowly reduced, the system tends to occupy lower energy states, while retaining a finite probability of escaping from a local energy minimum in favor of finding a better, more global one.

To apply simulated annealing to the NMR assignment problem, we begin with a tentative assignment of intra- and interresidue chemical shift segments, and calculate all connection and assignment probability scores P_i as described above. For each such score, we compute an energy $E_i = -\log(P_i)$ so that a high probability corresponds to a low energy and vice versa. The total assignment energy is then minimized using simulated annealing. A Monte Carlo move consists of choosing a continuous block of connected segments (see Fig. 5), choosing a similar block elsewhere along the protein chain, and swapping the corresponding segment assignments. The net energy change is calculated from the probability scores affected by the proposed swap, including amino acid type assignment scores for the swapped stretches of chemical shifts, and connection scores at their borders. The move is accepted or rejected according to the Metropolis criterion at the given temperature, with the Boltzmann factor $k \equiv 1$. At each value of T, many Monte Carlo steps (typically $\approx 10^5$) are attempted, allowing the system to come to equilibrium. The annealing temperature is reduced following a geometric sequence from initial to final values specified in an input file.

The Monte Carlo moves take into account assignment constraints provided by isotopic labeling of amino acids. For example, a sample of GlnBP was prepared with ^{13}C at phenylalanine only, and ^{15}N at all the other amino acid types (Yu et al., 1997). An HNCO experiment performed on this sample yielded nine cross peaks which determined the C' chemical shifts of phenylalanine and the amide ^{15}N and $^{1}H^{N}$ shifts of the amino acid residues which directly follow them in the sequence (2 of the 11 phenylalanine residues in GlnBP precede prolines, and were not detected by this experiment). The HNCO cross peaks are incorporated into nine interresidue segments, which are identified as a 'constrained set' in the simulated annealing program. Monte Carlo moves which swap members of this set with segments which are not in the same set are disallowed. However, swaps may take place within the set of phenylalanine-identified segments. Similarly, proline residues are identified with a constrained set of segments which lack amide proton chemical shifts; these segments are simply individual COCAH and HCA(CO)N cross peaks which have not been found to match other intra- and interresidue peaks. The annealing program must be supplied with an initial assignment which satisfies all constraints. During annealing, the program keeps track of the best (lowest energy) assignment achieved, as well as the most recent one. At the end of a run, the best assignment is written to an output file in a form which may be used as input to a later run.

Results

Calmodulin

The assignment algorithm has been tested on the previously assigned 148-residue protein calmodulin (Tjan-

dra et al., 1995), in its unliganded form. Lists of cross peaks for the CBCANH (Grzesiek and Bax, 1992c) and CBCA(CO)NH (Grzesiek and Bax, 1992b) experiments were provided by Dr. Nico Tjandra and Dr. Ad Bax. These researchers did not find it necessary to perform most of the other triple-resonance experiments listed above. Therefore, the available data on this protein do not provide a test of our best-first approach to combining data from several 3D NMR experiments. However, similar ideas were used to match the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ peaks of each amide (H,N) pair. Recall that the CBCA(CO)NH experiment correlates (H_i^N,N_i,C_{i-1}^{α}) and (H_i^N,N_i,C_{i-1}^{β}) chemical shifts; all cross peaks appear with the same (positive) sign. In a CBCANH spectrum, each amide pair may appear in four cross peaks: (H_i^N,N_i,C_i^{α}), (H_i^N,N_i,C_i^{β}), (H_i^N,N_i,C_{i-1}^{α}), and (H_i^N,N_i,C_{i-1}^{β}). The C^{α} peaks of glycine and the C^{β} peaks of all the other residues have negative amplitude. The first step in our treatment of these data is to align the CBCA(CO)NH and CBCANH spectra, and tabulate chemical shift difference distributions in the (H^N,N) plane. These histograms were fitted to curves similar to Eq. 1. The resulting characteristic widths $D_{H^N} = 0.0030$ ppm and $D_N = 0.025$ ppm indicate that the spectra align very well.

The next task is to identify glycine cross peaks, which appear in the CBCA(CO)NH spectrum with non-degenerate (H^N,N) chemical shifts. Some of the same interresidue peaks are seen in CBCANH. Intraresidue glycine cross peaks in CBCANH can usually be identified by their characteristically low ^{15}N and $^{13}C^{\alpha}$ chemical shifts. Once the glycine peaks have been set aside, all the remaining positive and negative peaks in CBCANH are associated with $^{13}C^{\alpha}$ and $^{13}C^{\beta}$, respectively. We search for matches to these peaks in CBCA(CO)NH, in order to classify the CBCANH peaks as inter- or intraresidue, and the CBCA(CO)NH peaks as $^{13}C^{\alpha}$ or $^{13}C^{\beta}$. Next, we search the CBCANH spectrum for intraresidue $^{13}C^{\alpha}$ or $^{13}C^{\beta}$ peaks with matching (H^N,N) chemical shifts. Uniquely matched pairs provide the equivalent of four-dimensional cross peaks giving (H_i^N,N_i,C_i^{α},C_i^{β}). Similarly, the $^{13}C^{\alpha}$ peaks of CBCA(CO)NH peaks are matched to the $^{13}C^{\beta}$ and unclassified peaks of the same spectrum. We can resolve some of the ambiguous matches and unclassified peaks by referring to matched (C_i^{α},C_i^{β}) pairs already found in CBCANH. Next, we try to resolve ambiguous matches remaining in CBCANH using the (C_i^{α},C_i^{β}) pairs seen in CBCA(CO)NH, and repeat the cycle until no more peaks can be confidently matched.

After applying this procedure to the calmodulin data, we have 147 pairs of CBCA(CO)NH peaks with matched (C^{α},C^{β}) chemical shifts, 28 C^{α} peaks lacking a matching C^{β}, and 17 C^{β} with no matching C^{α}. Of the CBCANH

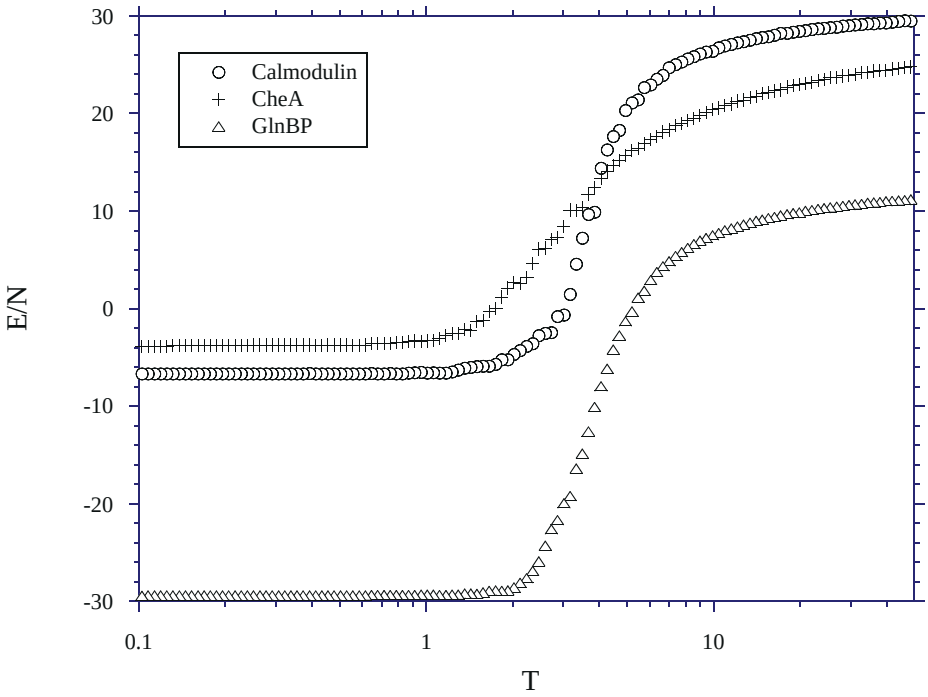

Fig. 6. Assignment energy (E) per residue (N) versus annealing temperature for runs of the simulated annealing assignment program applied to calmodulin, CheA, and GlnBP data. In all three cases, the assignment was initially randomized, then annealed using Monte Carlo moves as described in the text, as the annealing temperature T was reduced by a constant factor in 500 steps from 50.0 to 0.10. For the calmodulin and CheA runs, one hundred thousand such moves were attempted at each value of T. These runs each lasted ≈106 min on a Silicon Graphics Indy computer, and resulted in assignments which were essentially correct for calmodulin and the structured region of CheA. During the GlnBP run (which lasted 14 h), one million moves were proposed at each temperature. The lower energies achieved during this run are the result of additional, negative energy scores associated with C' and H^{α} chemical shifts, which were not available for calmodulin and CheA.

peaks, there are 135 matched (C^{α},C^{β}) pairs, 29 unmatched C^{α} peaks, and 27 unmatched C^{β} peaks. Chemical shift difference distributions are tabulated between the matched (C^{α},C^{β}) peaks of the two spectra. The histograms are fitted to expressions similar to Eq. 1 with characteristic widths $D_{C^{\alpha}} = D_{C^{\beta}} = 0.060$ ppm. As described above, curves fitted to the H^N, N, C^{α}, and C^{β} chemical shift differences are used to evaluate the likelihood that any two CBCA(CO)NH and CBCANH cross peaks originate from the same nuclei of the protein. Using these fitted curves, and the probability scores for assigning chemical shifts to known amino acid residue types, we can proceed to apply the simulated annealing program, where the intra- and interresidue segments are simply the CBCANH and CBCA(CO)NH peaks, respectively.

The procedure for pairing (C^{α},C^{β}) peaks generally provides more segments than the number of available sequence positions. Here, we have 191 intraresidue and 192 interresidue segments, to assign to the 146 non-proline residues of calmodulin. Excess segments are placed in a pool of 'unassigned' segments kept in 'null' positions following the actual protein sequence. During simulated annealing, unassigned segments may be swapped with assigned ones; however, only the latter contribute to the total energy. The assignment of calmodulin segments was initially randomized by performing 50 000 random swaps. The energy versus temperature curve of the subsequent simulated annealing run is shown in Fig. 6. Swaps were constrained only by the requirements that proline and glycine lack $^{1}H^{N}$ and $^{13}C^{\beta}$ chemical shifts, respectively.

Swapped blocks of chemical shifts were chosen at random from single segments to six segments (three residues) in length. The annealing temperature was reduced by a constant factor in 500 steps from 50 to 0.1. At each value of T, 1×10^5 Monte Carlo steps were attempted. By the end of the run (which took 106 min on a Silicon Graphics Indy workstation), the energy appeared to reach a minimum. The program's output shows excellent agreement with the previous, manual assignment of calmodulin (Tjandra et al., 1995, and personal communication). The $^{1}H^{N}$ and ^{15}N chemical shifts were correctly assigned for all 147 residues excluding the amino terminus. The $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ assignments were correct for all residues except Phe^{65}, for which the manually assigned chemical shifts were very different from the expected values for phenylalanine in a β-strand. The automated procedure yielded zero (i.e., missing) ^{13}C chemical shifts for this residue. When the program was modified so as to reduce the strict energy penalties for assigning outlying chemical shifts, a subsequent run yielded results in perfect agreement with the manual assignment.

CheY-binding domain of CheA

As a second test, the assignment procedure was applied to a polypeptide consisting of residues 124–257 of CheA with five additional residues at the carboxy-terminal end. This protein fragment comprises a structured segment of 68 amino acids (residues 160–227) flanked by two unstructured regions at either end (McEvoy et al., 1995, 1996). HNCACB and CBCA(CO)NH data on this protein were kindly provided by Dr. Frederick F. Dahlquist. The first step in our treatment of these data was to combine individual peaks within each spectrum by matching ($^{1}H^{N}$,^{15}N) chemical shifts, using an algorithm similar to that described above for calmodulin. This procedure resulted in 135 matched pairs of HNCACB peaks including both C^{α} and C^{β} chemical shifts, 40 C^{α} peaks without a matching C^{β}, and 39 C^{β} peaks lacking a matching C^{α}. Among the CBCA(CO)NH peaks, 133 matching pairs were found, leaving 26 unmatched C^{α} and 40 unmatched C^{β} peaks. Histograms of chemical shift differences were fitted by functions of the form of Eq. 1, with characteristic widths $D_{H^N} = 0.0061$ ppm, $D_N = 0.030$ ppm, and $D_{C^{\alpha}} = D_{C^{\beta}} = 0.085$ ppm. These values indicate somewhat poorer matches of chemical shifts than those of the calmodulin spectra.

Using the fitted curves to evaluate probability scores for matching HNCACB and CBCA(CO)NH peaks, we have applied the simulated annealing assignment program to the CheA data. The input parameters of the program (initial and final temperatures, number of Monte Carlo steps, etc.) were the same as those used for calmodulin. The annealing curve is shown in Fig. 6. The output of the run was compared to the published assignment (McEvoy et al., 1995) of the CheA fragment. Little agreement was found for residues in the unstructured regions at either end of the fragment, due to the poor chemical shift dispersion of such flexible domains. However, for the 68-residue structured region of the CheA fragment, only two $^{13}C^{\alpha}$ chemical shifts in the program's output (those of Leu^{199} and Pro^{200}) differed from the published assignment.

The $^{13}C^{\alpha}$ shift of Leu^{199} was left unassigned, because our programs failed to combine HNCACB cross peaks with ($^{1}H^{N}$,^{15}N,^{13}C) chemical shifts (9.16,128.06,51.33) and (9.16,128.08,42.56). These peaks give $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts, respectively, but a match between them was considered ambiguous because another $^{13}C^{\beta}$ peak (9.16,128.07,37.02) was also present. The last is actually an interresidue peak, but was not identified as such because the closest CBCA(CO)NH peak (9.15,127.91,37.52) lies slightly outside the tolerance limits used for matching. As a result, the three HNCACB cross peaks were treated as individual intraresidue segments, and only the second one was assigned. Note that, since the next residue is proline, the missing $^{13}C^{\alpha}$ chemical shift could not be provided by an overlapping CBCA(CO)NH peak. Similarly, the ($^{13}C^{\alpha}$, $^{13}C^{\beta}$) shifts of Pro^{200} were only available from an interresidue segment. Our program assigned $^{13}C^{\alpha} = 62.49$ ppm to this residue, but the published assignment lists $^{13}C^{\alpha} = 63.43$ ppm; both assignments agree on the amide ($^{1}H^{N}$,^{15}N) =

(8.51,106.04) of the next residue, Gly^{201}. However, the correlated chemical shifts (8.51,106.03,63.43) have no matching peaks in the CBCA(CO)NH or C(CO)NH spectra (the latter was provided by Dr. Dahlquist, but was not used by our programs). On the other hand, our assignment is supported by a C(CO)NH peak (8.51,106.01, 62.54) and an interresidue HNCACB peak (8.52,106.03, 62.43). This suggests that the published assignment of CheA contains a misprint, so that $^{13}C^{\alpha}$ of Pro^{200} should be 62.43 ppm rather than 63.43. In any case, the assignment of (H_i^N,N_i,C_i^{α},C_i^{β}) shifts provided by our algorithm agrees with that of McEvoy et al. (1995) for over 99% of the structured region of the CheA fragment.

Glutamine-binding protein

Having tested our method on the previously assigned proteins calmodulin and CheA, we proceeded to apply it to liganded GlnBP, using the full list of experiments: HNCA, HNCACB, HN(CA)CO, HN(CA)HA, COCAH, HN(CO)CA, CBCA(CO)NH, HNCO, HN(COCA)HA, and HCA(CO)N. Peaks were picked from these spectra at rather low contour levels, in order to reduce the chance of missing any real but weak cross peaks, at the risk of including some noise in the data. Noise in several of the spectra was reduced by filtering the peak lists against a 2D (^1H,^{15}N)-HSQC spectrum. Next, the HNCA peaks were combined with HNCACB, and HN(CO)CA with CBCA(CO)NH in a straightforward way; these combined spectra will be referred to as HNCACB and CBCA(CO)-NH below.

One possible stumbling block for the assignment procedure is the presence of 18 glycine residues in GlnBP. Because C^{α} of glycine is coupled to two protons, this amino acid is often not detected by HN(CA)HA, HN-(COCA)HA, or HN(CA)CO experiments. However, each glycine residue should yield two cross peaks in the COCAH and HCA(CO)N spectra. Applying the best-first method for connecting cross peaks, as described above, would result in several isolated peaks. Thus, the assignment program could not reliably connect segments across glycine residues. For this reason, we searched for glycine cross peaks separately, before applying the automated connection algorithm to the rest of the peaks. The COCAH and HCA(CO)N spectra were displayed as 2D contour plots on planes of constant $^{13}C^{\alpha}$ chemical shift. Corresponding planes of the two spectra were searched visually for pairs of cross peaks of the form (H_1^{α},C^{α},C'), (H_2^{α},C^{α},C') and (H_1^{α},C^{α},N), (H_2^{α},C^{α},N). The C^{α} chemical shifts of such peaks were matched to HNCACB peaks lacking C^{β} shifts, which exhibited the upfield-shifted ^{15}N values characteristic of glycine.

Connections were extended to the following amino acid residue by matching the $^{13}C^{\alpha}$, $^{13}C'$, and ^{15}N chemical shifts of the [COCAH,HCA(CO)N] clusters to matched pairs of HNCO and CBCA(CO)NH cross peaks. A few additional glycine candidates were identified by matching the $^{13}C^{\alpha}$ chemical shifts of HNCACB and CBCA(CO)NH peaks with $^{13}C^{\alpha} < 50$ ppm and no $^{13}C^{\beta}$. The HNCACB and COCAH peaks identified with glycine were grouped together as intraresidue segments, while HCA(CO)N, CBCA(CO)NH, and HNCO were combined to form interresidue segments. Altogether, 21 intraresidue and 23 interresidue segments were identified as candidates for the 18 glycine positions. Cross peaks participating in these segments were set aside, leaving 223 non-glycine HNCACB peaks, 188 HN(CA)CO, 218 HN(CA)HA, 231 COCAH, 219 CBCA(CO)NH, 223 HNCO, 200 HN(COCA)HA, and 242 HCA(CO)N.

Our chemical shift connection algorithm, applied to the intraresidue spectra HNCACB, HN(CA)CO, HN(CA)HA, and COCAH, yielded 132 four-way, 55 three-way, and 22 two-way matches, leaving 123 unmatched peaks. Similar treatment of the interresidue peak lists provided 135 four-way, 55 three-way, and 34 two-way matches, with 111 individual peaks left unmatched. At this stage, the segments previously identified as glycine candidates were added to the 332 intraresidue and 335 interresidue segments. In assigning these chemical shift segments to GlnBP, we have made use of NMR experiments performed on specifically labeled samples of this protein (Tjandra et al., 1992; Tjandra, 1993; Yu et al., 1995, 1997). Such experiments provided the backbone amide $^1H^N$ and ^{15}N chemical shifts of the two tryptophan and three methionine residues. Measurements on a sample labeled with ^{13}C at the carbonyl of phenylalanine and ^{15}N at all the other amino acid types provided a set of nine HNCO peaks identified with the phenylalanine residues not followed by proline in the primary sequence. Another sample, labeled similarly with $^{13}C'$-Tyr, yielded cross peaks originating from tyrosine and the amide of the next residues. A 1D ^{13}C spectrum of a $^{13}C'$-Pro-labeled sample allowed us to identify possible COCAH peaks to assign to the seven prolines of GlnBP. These constraints, as well as the visual identification of glycine peaks, have been used to limit the number of possible assignments.

An initial assignment consistent with the specific labeling constraints, but otherwise random, was provided as input to the simulated annealing assignment program. Residue-type assignment scores were evaluated under the assumption that the secondary structure of liganded GlnBP is identical to that of the unliganded protein, as determined by X-ray crystallography (Hsiao, 1993). In order to obtain the best assignment of the large number of segments available, a long run was performed. As before, the Monte Carlo moves consisted of swapping the assignments of two blocks of chemical shifts, chosen at random from one to six segments in length. One million such moves were proposed at each value of the annealing temperature, which was reduced in 500 steps from 50 to 0.1. The run lasted 14 h on an SGI Indy workstation,

1	2	3	4	5	6	7	8	9
Glu	38	112	111	N	125.09	125.12	-6.02	
Glu	38	112	111	Hn	7.486	7.493		-7.52
Glu	38	112	162	Ha	0.000	4.023		-11.56
Glu	38	112	162	Ca	58.80	58.75	-6.89	
Glu	38	112	162	Co	0.00	178.05	-3.40	
Glu	38	112	162	Cb	29.36	29.59	-6.89	
Leu	39	164	162	N	118.99	118.98	-6.83	
Leu	39	164	162	Hn	7.838	7.854		-6.29
Leu	39	164	146	Ha	4.213	4.214		-14.92
Leu	39	164	146	Ca	54.52	54.48	-4.65	
Leu	39	164	146	Co	175.15	175.15	-3.33	
Leu	39	164	146	Cb	0.00	42.86	-3.14	
Lys	40	146	146	N	122.40	122.43	-6.90	
Lys	40	146	146	Hn	7.722	7.723		-8.60
Lys	40	146	201	Ha	3.874	3.860		-17.13
Lys	40	146	201	Ca	57.35	57.37	-6.76	
Lys	40	146	201	Co	175.64	175.64	-6.75	
Lys	40	146	201	Cb	29.20	29.21	-3.50	
Leu	41	200	201	N	122.21	122.23	-6.55	
Leu	41	200	201	Hn	8.134	8.136		-8.49
Leu	41	200	285	Ha	4.689	4.693		-14.84
Leu	41	200	285	Ca	53.77	53.77	-6.89	
Leu	41	200	285	Co	176.01	176.02	-6.89	
Leu	41	200	285	Cb	0.00	44.35	-3.43	
Asp	42	275	285	N	126.76	126.72	-6.62	
Asp	42	275	285	Hn	8.645	8.639		-7.42
Asp	42	275	210	Ha	5.025	5.021		-16.74
Asp	42	275	210	Ca	53.36	53.27	-6.90	
Asp	42	275	210	Co	175.98	175.99	-6.81	
Asp	42	275	210	Cb	43.03	43.06	-6.78	
Tyr	43	208	210	N	121.43	121.45	-6.90	
Tyr	43	208	210	Hn	8.173	8.174		-8.70
Tyr	43	208	318	Ha	5.850	5.865		-15.77
Tyr	43	208	318	Ca	56.36	56.36	-6.90	
Tyr	43	208	318	Co	172.61	172.64	-6.16	
Tyr	43	208	318	Cb	41.51	41.37	-6.89	

Fig. 7. A portion of the output file written by the simulated annealing assignment program. The first two columns list the amino acid and sequence positions. The next two columns list the arbitrary ID numbers used to identify the assigned intra- and interresidue segments, whose chemical shifts are printed after a column which identifies the type of nucleus. Chemical shifts of the intra- and interresidue segments are enclosed in boxes, for clarity. The eighth column lists individual amino acid type assignment energies, including ^{15}N, $^{13}C^{\alpha}$, $^{13}C'$, and $^{13}C^{\beta}$ scores. The final column lists energies for matching overlapping segments on (^{15}N,$^{1}H^{N}$) and ($^{1}H^{\alpha}$,$^{13}C^{\alpha}$,$^{13}C'$,$^{13}C^{\beta}$) shifts. These scores are printed on the same lines as $^{1}H^{N}$ and $^{1}H^{\alpha}$, respectively. The 'energies', calculated from integer probability scores scaled as described in the text, are dimensionless.

giving the annealing curve shown in Fig. 6. The quality of the assignment produced by the program was checked by inspection of an output file listing the chemical shifts and energy scores, as shown in Fig. 7 (see also the discussion below). On the basis of this output, we have confidently assigned over 95% of the backbone resonances of GlnBP.

For the residues where our assignment was judged reliable, over 97% of the chemical shifts agreed with those of a manual assignment (Yu et al., 1995,1997), which has been carried out independently, during the development of our programs. Furthermore, the automated procedure was able to assign a few chemical shifts which were left unassigned by the manual method, because of connections which appeared ambiguous when peaks were matched one at a time. On the other hand, the manual method yielded some $^{13}C^{\beta}$ assignments which were not provided by our automated algorithm. Our treatment of the HNCACB and CBCA(CO)NH spectra connected individual $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ peaks only when their ($^{1}H^{N}$,^{15}N) shifts matched strongly and uniquely. This resulted in several residues where $^{13}C^{\beta}$ was missing in the assignment. However, we can refer to the random coil chemical shifts for those residues to resolve ambiguous connections and fill in several of the missing $^{13}C^{\beta}$ shifts, from those HNCACB and CBCA(CO)NH peaks left unassigned. We are currently using NOESY data to decide between the manual and automated assignment results where they disagree. The final assignment of GlnBP is presented, along with experimental details, in Yu et al. (1997).

Discussion and Conclusions

Our algorithm has some features in common with previously published strategies for chemical shift assignment. The CONnectivity TRacing ASsignment Tools (CONTRAST) method of Olson and Markley (1994) combines cross peaks from 3D NMR experiments HNCO, HN(CO)CA, HNCA, TOCSY-HMQC, HCACO, and HCA(CO)N to form 'fragments' spanning chemical

shifts C^{α}_{i-1},...N_{i+1} (in terms of the order of chemical shifts given in Fig. 2). These fragments are joined on the basis of overlapping chemical shifts (C^{α},C',N); the amino acid sequence of the protein is not used. The output of the program includes information on the ambiguity of the assignment. For imperfect data, a 'best-first' procedure for linking fragments was found to be more reliable than simulated annealing. A somewhat similar assignment procedure was used by Morelle et al. (1995). These researchers combined peaks from 'reduced dimensionality' 2D versions of HNCO, HN(CO)CA, H(N)COCA, HNCA, HN(CA)HA, HN(CA)CO, and HN(COCA)HA to form 'pseudoresidues' spanning H^{α}_{i-1},...C'_{i+1}. The primary sequence of the protein was used to rule out some assignments based on loose ranges of $^{13}C^{\alpha}$ chemical shifts. The pseudoresidues were assigned by two alternative methods: (i) a systematic search which proceeds from the N- to the C-terminus, storing all possible solutions at each stage; and (ii) simulated annealing. The latter method was more robust with respect to data with missing peaks.

Greater use of the amino acid sequence information is made by Meadows et al. (1994). Here, data from 4D HNCAHA, HN(CO)CAHA, and HC(CO)NH-TOCSY are correlated to obtain spin systems which include the ($N,H^N,H^{\alpha},C^{\alpha}$) resonances of residue i and all aliphatic ^1H and ^{13}C resonances of residue i–1. These spin systems are classified by an empirical scoring system which uses simple 'rewards' and 'penalties' to compare the observed aliphatic chemical shifts with their characteristic ranges for different amino acid types. The spin systems are extended into longer fragments by matching (N,H^N) and (H^{α},C^{α}) chemical shifts. As each fragment is extended, it is checked for consistency with the protein sequence. These steps are iterated until all fragments spanning up to six peptides have been obtained. Then, the largest fragments are assigned to the sequence first; any gaps that remain are filled with smaller peptide fragments. A comparable approach (Friedrichs et al., 1994) uses data from 3D HNCACB, 3D CBCA(CO)NH, 4D HNCAHA, and 4D HN(CO)CAHA experiments. Peaks from these spectra which share amide (N,H^N) chemical shifts are grouped into 'residue information data structures' (RIDs) and classified as $^{13}C^{\alpha}_i$, $^{13}C^{\alpha}_{i-1}$, $^{13}C^{\beta}_i$, or $^{13}C^{\beta}_{i-1}$. The RIDs are then linked based on matches between overlapping ($H^{\alpha},C^{\alpha},C^{\beta}$) shifts. Glycine RIDs are treated separately, using HNHA-Gly and HBHA(CO)NH spectra. Stretches of strongly linked RIDs are assigned to the sequence, based on (C^{α},C^{β}) chemical shift profiles of the underlying amino acids. The remaining RIDs are used to consistently fill the gaps between these assigned fragments.

All of the automated assignment methods outlined above begin by combining individual 3D or 4D NMR cross peaks into longer fragments of chemical shifts. Our own procedure accomplishes this in a rigorous way, using connection scores based on Bayes' theorem to establish the strongest connections first. Matches among amide-detected experiments are confirmed by COCAH and HCA(CO)N data. While most of the above procedures continue to use 'best-first' methods to assign the chemical shift segments, ours uses simulated annealing. Olson and Markley (1994) concluded that, while simulated annealing finds the global optimal solution to the problem of joining fragments, it may do so at the expense of failing to form some of the locally strongest, least ambiguous connections. On the other hand, the chemical shift degeneracy of a large protein may result in 'accidental' matches of cross peaks when a best-first connection procedure is given priority over amino acid type assignment. Any locally incorrect connections formed by such a deterministic, irreversible procedure may force assignment errors to propagate over long stretches of the sequence. We believe that the application of simulated annealing to short, deterministically formed chemical shift segments represents a good compromise between these pitfalls.

Our probabilistic assignment method has proven to be effective when applied to data on calmodulin, CheA, and GlnBP. For the algorithm to succeed, two general conditions must be met: (i) the simulated annealing program must converge on the lowest energy arrangement of segments; and (ii) this solution should correspond to the correct assignment of chemical shifts. In order to satisfy the first condition, the annealing temperature must be reduced slowly while a sufficient number of Monte Carlo moves are performed to keep the system in equilibrium. This can be checked by performing successive runs with overlapping temperature ranges. If annealing has been carried out slowly enough, the E versus T plots (as in Fig. 6) of the overlapping runs should coincide over the shared temperature range.

Under conditions of poor dispersion of chemical shifts and low signal-to-noise, the data may not be good enough to determine a unique assignment. Even if the data are of high quality, effects such as chemical exchange and conformational heterogeneity may allow certain regions of the protein to exhibit two or more different sets of chemical shifts. However, our simulated annealing program will yield only a single assignment, ideally the one which minimizes the total energy score, even if other possible assignments exist which have energies nearly as low. Taking this possibility into account, we have modified the program to keep track of the 10 best assignments achieved so far, at any time during an annealing run. At the end of the run, these assignments and their energy scores are written to a file, which can be inspected by the user. Extra spin systems and poorly determined chemical shifts will appear as alternative assignments with nearly equal energies. However, a single run may not reveal all such alternative assignments; in general, simulated annealing may not sample every possible low-energy state of a system. Therefore, it is often desirable to check the out-

put of several annealing runs, which include temperatures high enough to allow the system to escape from local energy minima.

The efficiency of the annealing program has been optimized so that multiple runs may be performed in a reasonable amount of time. After a Monte Carlo move is proposed, the affected probability scores for connecting segments are rapidly reevaluated using two look-up tables prepared in advance. A file *intra_inter.dat* tabulates, for each intraresidue segment, the six likeliest connecting interresidue segments, matching on $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C'$, and $^{13}C^{\beta}$ chemical shifts. A second file *inter_intra.dat* records the six most probable intraresidue segments matching the $^{1}H^{N}$ and ^{15}N chemical shifts of each interresidue segment. These files tabulate the probability scores for each match as integers, with maximum values of 10^8 for *intra_inter.dat* and 10^4 for *inter_intra.dat*. Amino acid assignment scores (normalized to integers with a maximum of 1000) are evaluated without performing floating-point arithmetic, through the use of a discrete look-up table representing the Gaussian distribution. These time-saving measures allow Monte Carlo moves to be performed at rates of several thousand per second on a modern computer workstation.

An earlier version of the program which used individual 3D NMR cross peaks as the moveable units of assignment failed to converge on the global energy minimum. Apparently, accidental peak connections due to chemical shift degeneracy led to an excess of local minima separated by barriers which were difficult to overcome by moving one peak at a time. The current approach deterministically combines peaks into longer segments, verified using the redundant coverage provided by a large set of 3D NMR experiments. This has the effect of smoothing the energy 'landscape' so that simulated annealing can find the overall minimum. If the data are sufficiently complete, it may be advantageous to combine the two types of segments to form a stretch of chemical shifts:

$$\{N_i, H_i^N, C_i^{\alpha}, C_i^{\beta}, C_i', N_{i+1}, H_{i+1}^N\}$$

With this single type of segment, matched at both ends using the amide proton and nitrogen chemical shifts, simulated annealing might proceed more efficiently.

Independently of the effectiveness of the Monte Carlo program in finding the minimum-energy arrangement of segments, the success of the method requires accurate scoring so that this solution indeed corresponds to the correct chemical shift assignment. This is achieved by applying Bayes' theorem to objectively evaluate the probabilities of matching overlapping segments and assigning them to given sequence positions. Each probability score is translated into an energy value using the relation $E_i = -\log(P_i)$. Probabilities below a given threshold are assigned a constant, positive energy in order to discourage poor connections. At the end of a run, the simulated annealing program writes an output file which lists the assigned chemical shifts and individual energy scores (see Fig. 7). The user can check this output for energies with small magnitudes, which may indicate a locally incorrect assignment. One possible reason for an incorrect assignment may be that some chemical shifts lie outside the normal range for the originating amino acid, due to the local protein conformation (such as the proximity of an aromatic ring, i.e., the ring-current effect). In order not to overpenalize such a case, the program includes the option of giving different relative weights to connection and assignment energies.

An erroneous assignment may also result from incorrect segments formed by cross peaks whose chemical shifts overlap accidentally. Such incorrect segments are usually avoided by the 'best-first' connection procedure which uses the redundant information provided by several 3D NMR experiments. For example, HN(CA)HA and HN(CA)CO cross peaks with nearby $^{1}H^{N}$ and ^{15}N chemical shifts are matched with high priority if the connection is verified by the H^{α} and C' shifts of a COCAH peak. Some intervention is necessary to decide on the threshold of connection scores below which a match is not accepted. If this threshold is set too high, the program will fail to use all of the consistent connections and there will be an excess of segments with missing chemical shifts. If it is too low, incorrect segments will be formed, which may force assignment errors to propagate over long stretches of the sequence. It may also be necessary to adjust the energy penalties for connecting and assigning segments with missing chemical shifts, which are somewhat arbitrary.

The accuracy of the assignment method is less sensitive to the fine-tuning of energy scores, the larger the number of assignment constraints provided by specific labeling. The program can be run several times with different partial sets of constraints as a check of consistency. To ensure the correctness of the lowest energy assignment, the coordinates of cross peaks from different experiments must be carefully aligned, to eliminate systematic differences. The quality of chemical shift matches may be improved by enhancing spectral resolution using numerical techniques such as zero filling and linear prediction.

The programs for combining cross peaks can be adapted for different sets of experiments. The results presented above for calmodulin and CheA indicate that, for proteins under 150 amino acids, HNCACB and CBCA(CO)NH may be sufficient. For larger proteins with good chemical shift dispersion and few prolines, HCA(CO)N could be omitted. The other experiments may all be performed on a single protein sample dissolved in H_2O, leading to better matches among chemical shifts. If there are too many ambiguous connections, 4D HNCAHA and

HN(CO)CAHA (Olejniczak et al., 1992) could be substituted for HNCA, HN(CA)HA, HN(CO)CA, and HN(COCA)HA. Further restrictions on amino acid types obtained from HCCH-TOCSY spectra (Bax et al., 1990) could be imposed in the same way as information provided by specific amino acid labeling. In any case, once the segments of chemical shifts are formed, the same simulated annealing program (with slight changes in scoring) can be used to find the optimal resonance assignment.

The quality of matches between cross peaks from different spectra seems to depend on the specific conditions under which experiments were performed. Therefore, in order to evaluate matching probabilities rigorously using Bayes' theorem, it is necessary to tabulate (and fit curves to) chemical shift differences for each new set of experiments. Then, the best-first connection procedure can be applied to link the cross peaks into longer segments of chemical shifts, which become the units of the simulated annealing assignment program. Taking into account the effort involved in these preliminary steps, our procedure may not be more efficient than manual assignment when applied to a very small protein. For larger proteins, however, the numerous cross peaks and consequent chemical shift overlap confer a definite advantage to an automated assignment method. This was evident in the case of GlnBP, where the manual assignment of this 25 kDa polypeptide took many days of a researcher's time. Even larger proteins can now be studied by NMR, thanks to recent technical advances such as high-field magnets, pulsed-field gradients, and ^{2}H decoupling applied to partially deuterated samples. By improving the efficiency of chemical shift assignment, the automated procedure described here should help make such studies practical.

Acknowledgements

We wish to thank Ms. Patricia F. Cottam for preparing labeled samples of GlnBP and Ms. Jinghua Yu and Mr. Virgil Simplaceanu for performing many 3D NMR experiments which provided data used in this work. We also thank Dr. Gordon S. Rule (formerly at the University of Virginia, now at Carnegie Mellon University) and Dr. Clemens Anklin (Bruker Instruments) for the use of their spectrometers for the CBCA(CO)NH and HNCACB experiments, respectively. Dr. John L. Markley (University of Wisconsin) kindly provided an early release of the BioMagResBank database. Dr. Ad Bax and Dr. Nico Tjandra (National Institute of Diabetes, Digestive, and Kidney Disease, National Institutes of Health) provided the complete chemical shift assignment of calcium-free calmodulin in advance of publication, and the CBCANH and CBCA(CO)NH peak coordinates. Peak lists for the CheA protein fragment were provided by Dr. Frederick F. Dahlquist and Ms. Megan McEvoy (University of Oregon). We also wish to thank Drs. Michael Widom and Robert H. Swendsen of the Department of Physics, Carnegie Mellon University, for helpful discussions on simulated annealing. J.A.L. is the recipient of an NIH NRSA award (F32GM17034). This work is also supported by research grants from the National Institutes of Health (HL-24525 and GM-26874).

References

Bax, A., Ikura, M., Kay, L.E., Torchia, D.A. and Tschudin, T. (1990) *J. Magn. Reson.*, **86**, 304–318.

Bax, A. and Grzesiek, S. (1993) *Acc. Chem. Res.*, **26**, 131–138.

Cieslar, C., Clore, G.M. and Gronenborn, A.M. (1988) *J. Magn. Reson.*, **80**, 119–127.

Clore, G.M. and Gronenborn, A.M. (1991) *Science*, **252**, 1390–1399.

Clubb, R.T., Thanabal, V. and Wagner, G. (1992a) *J. Biomol. NMR*, **2**, 203–210.

Clubb, R.T., Thanabal, V. and Wagner, G. (1992b) *J. Magn. Reson.*, **97**, 213–217.

Clubb, R.T. and Wagner, G. (1992) *J. Biomol. NMR*, **2**, 389–394.

Dijkstra, K., Kroon, G.J.A., Van Nuland, N.A.J. and Scheek, R.M. (1994) *J. Magn. Reson.*, **A107**, 102–105.

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*, **97**, 111–130.

Friedrichs, M.S., Mueller, L. and Wittekind, M. (1994) *J. Biomol. NMR*, **4**, 703–726.

Grzesiek, S. and Bax, A. (1992a) *J. Magn. Reson.*, **96**, 432–440.

Grzesiek, S. and Bax, A. (1992b) *J. Am. Chem. Soc.*, **114**, 6291–6293.

Grzesiek, S. and Bax, A. (1992c) *J. Magn. Reson.*, **99**, 201–207.

Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.

Hing, A.W., Tjandra, N., Cottam, P.F., Schaeffer, J. and Ho, C. (1994) *Biochemistry*, **33**, 8651–8661.

Hsiao, C.-D. (1993) Ph.D. Thesis, University of Pittsburgh, Pittsburgh, PA, U.S.A.

Ikura, M., Kay, L.E. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.

Kirkpatrick, S., Gelatt Jr., C.D. and Vecchi, M.P. (1983) *Science*, **220**, 671–680.

Kleywegt, G.J., Boelens, R., Cox, M., Llinás, M. and Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 23–47.

Leopold, M.F., Urbauer, J.L. and Wand, A.J. (1994) *Mol. Biotechnol.*, **2**, 61–93.

McEvoy, M.M., Zhou, H., Roth, A.F., Lowry, D.F., Morrison, T.B., Kay, L.E. and Dahlquist, F.W. (1995) *Biochemistry*, **34**, 13871–13880.

McEvoy, M.M., Muhandiram, D.R., Kay, L.E. and Dahlquist, F.W. (1996) *Biochemistry*, **35**, 5633–5640.

Meadows, R.P., Olejniczak, E.J. and Fesik, S.W. (1994) *J. Biomol. NMR*, **4**, 79–96.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) *J. Chem. Phys.*, **21**, 1087–1092.

Morelle, M., Brutscher, B., Simorre, J.-P. and Marion, D. (1995) *J. Biomol. NMR*, **5**, 154–160.

Olejniczak, E.T., Xu, R.X., Petros, A.M. and Fesik, S.W. (1992) *J. Magn. Reson.*, **100**, 444–450.

Olson, J.B. and Markley, J.L. (1994) *J. Biomol. NMR*, **4**, 385–410.

Powers, R., Gronenborn, A.M., Clore, G.M. and Bax, A. (1991) *J. Magn. Reson.*, **94**, 209–213.

Press, S.J. (1989) *Bayesian Statistics: Principles, Models, and Applications*, Wiley, New York, NY, U.S.A.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992) *Numerical Recipes in Fortran*, 2nd ed., Cambridge University Press, Cambridge, U.K., pp. 436–438.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.

Shen, Q., Simplaceanu, V., Cottam, P.F. and Ho, C. (1989a) *J. Mol. Biol.*, **210**, 849–857.

Shen, Q., Simplaceanu, V., Cottam, P.F., Wu, J.-L., Hong, J.-S. and Ho, C. (1989b) *J. Mol. Biol.*, **210**, 859–867.

Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.

Tjandra, N., Simplaceanu, V., Cottam, P.F. and Ho, C. (1992) *J. Biomol. NMR*, **2**, 149–160.

Tjandra, N. (1993) Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, U.S.A.

Tjandra, N., Kuboniwa, H., Ren, H. and Bax, A. (1995) *Eur. J. Biochem.*, **230**, 1014–1024.

Van de Ven, F.J.M. (1990) *J. Magn. Reson.*, **86**, 633–644.

Weber, P.L., Malikayil, J.A. and Mueller, L. (1988) *J. Magn. Reson.*, **82**, 419–426.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311–333.

Wishart, D.S. and Sykes, B.D. (1994) *Methods Enzymol.*, **239**, 363–392.

Wishart, D.S., Bigham, C.G., Yao, J., Abildgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995) *J. Biomol. NMR*, **6**, 135–140.

Wittekind, M. and Mueller, L. (1993) *J. Magn. Reson.*, **B101**, 201–205.

Xu, J., Straus, S.K., Sanctuary, B.C. and Trimble, L. (1994) *J. Magn. Reson.*, **B103**, 53–58.

Yu, J., Tjandra, N., Simplaceanu, V., Cottam, P.F., Lukin, J.A. and Ho, C. (1995) *Biophys. J.*, **68**, 420.

Yu, J., Simplaceanu, V., Tjandra, N.L., Cottam, P.F., Lukin, J.A. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 167–180.

Zimmerman, D.E. and Montelione, G.T. (1995) *Curr. Opin. Struct. Biol.*, **5**, 664–673.