Journal of Biomolecular NMR, **25:** 1–9, 2003.
KLUWER/ESCOM

MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins

T. Kevin Hitchens, Jonathan A. Lukin, Yiping Zhan, Scott A. McCallum & Gordon S. Rule*
Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213, U.S.A.

Received 20 June 2002; Accepted 18 September 2002

Key words: automated assignment, Monte Carlo, protein NMR

Abstract

A general-purpose Monte Carlo assignment program has been developed to aid in the assignment of NMR resonances from proteins. By virtue of its flexible data requirements the program is capable of obtaining assignments of both heavily deuterated and fully protonated proteins. A wide variety of source data, such as inter-residue scalar connectivity, inter-residue dipolar (NOE) connectivity, and residue specific information, can be utilized in the assignment process. The program can also use known assignments from one form of a protein to facilitate the assignment of another form of the protein. This attribute is useful for assigning protein-ligand complexes when the assignments of the unliganded protein are known. The program can be also be used as an interactive research tool to assist in the choice of additional experimental data to facilitate completion of assignments. The assignment of a deuterated 45 kDa homodimeric Glutathione-S-transferase illustrates the principal features of the program.

Abbreviations: NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; NMR, nuclear magnetic resonance spectroscopy

Introduction

The assignment of NMR resonances is a fundamental step in using NMR to study the structure and dynamics of proteins. Traditional approaches have utilized a four step procedure for assignments: (1) The collection of inter- and intra-residue chemical shifts from the backbone and sidechains atoms into spin-systems that generally represent the spectral information associated with a single amide resonance, (2) determination of the probability that a spin-system represents a particular residue type, based on chemical shift information, (3) elucidation of sequential connections between these spin-systems by dipolar coupling or scalar coupling, (4) determination of the sequential mapping of these connected segments to the known amino acid sequence of the protein (Wüthrich, 1986).

Algorithms for assignments are straightforward and thus very amenable to automation, see Zimmerman and Montelione (1995) and Moseley and Montelione (1999) for review. One of the earlier assignment protocols was developed by Friedrichs et al. (1994). In their approach inter-residue connectivities were identified using backbone atoms and the identification of residue type was based on C_α and C_β chemical shifts. A best-first approach was then used to map the connected segments on to the primary sequence. This process was implemented using macros that were inherent to the processing software, thus allowing the user to obtain trial assignments while analyzing spectra. Zimmerman et al. (1997) expanded on the best-first approach by propagating constraints from confident initial assignments to less-confident assignments towards the end of the assignment process. A similar approach was utilized by Olson and Markley (1994) as well as by Li and Sanctuary (1997). More recently, MAPPER (Güntert et al., 2000) performs an exhaustive search to place connected fragments on to the primary sequence. Atreya et al. (2000) employ a similar approach, but simplify the assignment of

*To whom correspondence should be addressed. E-mail: rule@andrew.cmu.edu residue type by grouping like residues into eight distinct categories. Other approaches to the mapping of connected segments to the primary sequence include genetic algorithms that are similar to the Monte Carlo methods presented in this paper (GARANT, Bartels et al., 1996).

All of the above approaches rely heavily, or exclusively, on sequential connectivity that is derived from through-bond scalar coupling. A number of workers have developed assignment procedures that initially focus on the use of inter-proton distances obtained from nuclear Overhauser effects (NOE) to establish inter-residue connectivities. This method was initially proposed by Wand (Wand and Nelson, 1991) as a main-chain directed assignment method. An automated version of this approach was recently presented by Bailey-Kellogg et al. (2000). Kraulis (1994), and more recently Grishaev and Llinas (2002a, b), have developed methods that determine real-space proton densities directly from NOE data, without the need to identify sequential connectivities. Although NOE based assignment schemes show great promise for fully protonated proteins, it is unlikely that they will be widely applicable to deuterated proteins.

Available assignment programs appear to be capable of obtaining near complete assignments on smaller to mid-sized proteins. In this case, extensive proton and carbon chemical shifts provide considerable information on the most probable residue type of a spin-system as well as extensive inter-residue connectivity from scalar and dipolar coupling. In the case of larger proteins it is necessary to deuterate the protein to reduce the deleterious effect of efficient spin-spin relaxation (LeMaster, 1990). In highly deuterated proteins, the assignment process is hampered by the loss of connectivity and residue type information due to the absence of aliphatic protons. A more serious problem with the assignment of deuterated proteins is the heavy reliance on the observation of resonance signals from the amide protons. In cases where it is not possible to exchange amide deuterons for protons, the loss of observable amide resonances introduces gaps in the sequential connectivity. Such gaps can also occur if amide resonances are absent due to chemical exchange on the intermediate time scale. These gaps serve as barriers to assignment schemes that rely heavily on inter-residue connectivities. When a large number (20–30%) of the amides resonances are unobservable it can be very difficult to correctly map a series of connected spin-systems onto the primary sequence using conventional assignment schemes. In this case we have found it necessary to include additional information to assign the spin systems associated with the observable amides. This information has included residue type identification from specific labeling and inter-proton distances from dipolar coupling (McCallum et al., 1999, 2000; Hitchens et al., 2001). In a number of cases we have also used the known chemical shift assignments from one form of the protein (e.g. unliganded) to aid in obtaining assignments of other protein-ligand complexes (McCallum et al., 2000; Hitchens et al., 2001).

The large and diverse set of information that is required for the assignment of larger proteins makes manual analysis of the data impractical. Existing assignment programs cannot effectively utilize the broad type of information that is required in the assignment of large deuterated proteins. Consequently, an automated Monte Carlo approach (MONTE) has been developed that is tailored to obtaining NMR resonance assignments of proteins using a diverse set of information in the assignment process. Monte Carlo methods are particularly powerful in this particular application because they explore the landscape of possible solutions during the assignment process. Consequently, they are able to report both the most favorable set of assignments as well as an ensemble of solutions that are closely related to the best solution. This ensemble of solutions can be inspected to detect possible errors in the input data or to identify additional data that would be useful to resolve remaining difficulties associated with the assignments.

A number of workers have used Monte Carlo methods as a basis for automated assignment programs. Lukin et al. (1997) obtained resonance assignments from C_α, H_α, C_β, CO chemical shifts and inter-residue connectivity with Monte Carlo methods. A similar approach was described by Leutner et al. (1998) using a threshold accepting algorithm. Olson and Markley (1994) also explored the use of Monte Carlo methods in resonance assignment, but noted that best-first algorithms appeared to be capable of obtaining more assignments as the quantity of the available data decreased. The distinct advantage of the program described in this manuscript over existing assignment programs is that it provides a general software package for chemical shift assignments of proteins that is independent of any particular ‘required’ experimental data. For example, the program is sufficiently flexible to utilize any type of inter-residue connectivity information, from either scalar coupling or inter-proton distances. In addition, experimental data that provides informa-

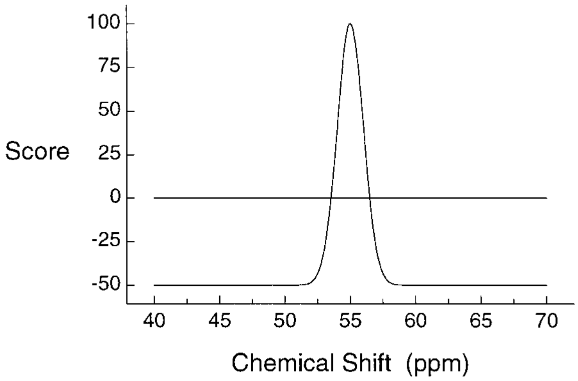

Figure 1. Gaussian scoring function. This curve shows an example of the scoring function that would be used to evaluate the match between inter- and intra-residue C_α chemical shifts. In this example, the intra-residue shift of the spin-system assigned to the (i − 1)th residue is 55 ppm. If the inter-residue shift of the spin-system assigned to the ith residue was 55, the score of the match would be 100. If the inter-residue shift was 53 ppm, the score would be zero. Larger mismatches between the shifts would result in a negative score, reducing the probability that this particular set of assignments would be retained. The height, width and offset of this function can be defined by the user.

tion on residue-type can be easily incorporated into the assignment process. Finally, known assignments from alternative forms of the protein can also be used to aid in assignment. This attribute of the program is useful when it is desirable to study a large number of protein-ligand complexes.

Methods

MONTE uses a standard Metropolis Monte Carlo algorithm with simulated annealing (Metropolis et al., 1953). The program attempts to find the best global mapping of spin-systems onto the primary sequence. In order to sample solutions that are consistent with the experimental data the program is executed a number of cycles (5–50) to generate an ensemble of possible solutions. The uniqueness of an assignment is assessed by comparing the ensemble of solutions that are obtained from the multiple independent runs. Spin-system mappings that are identical in all members of the ensemble are considered to be unique assignments. Ambiguous assignments occur when multiple spin-systems have been mapped to the same residue. Often, the nature of the ambiguity suggests additional experimental data, such as residue specific information, that may resolve the ambiguity.

The current form of the program utilizes information from the identification of residue type, NOE peaks, intra-residue chemical shifts, and inter-residue chemical shifts to the preceding or following residue. All of this information is associated with a particular amide nitrogen-proton pair and is considered to comprise a spin-system. This information is compiled from several databases in a semi-automatic fashion. At the beginning of each cycle the program generates a uniform random mapping of the observed spin-systems to either residues within the primary sequence of the protein or to positions within a cache. The size of the cache is 20% of the length of the protein and there are no restrictions on how many spin-systems can reside in the cache during the assignment process. The cache allows the temporary removal of spin-systems from the scoring process, permitting Monte Carlo moves that might otherwise be unlikely because of a poor score. In addition, spin-systems that cannot be mapped on to the primary sequence, such as those from contaminants in the sample or the existence of minor forms of the protein, will reside in the cache at the end of the assignment process. The initial mapping is scored by summing the contribution of all of the spin-systems, with the exception of those in the cache, to the overall score. The scoring function is quite flexible and *any* combination of the following terms can be employed:

– How well pairs of inter-residue scalar correlations match. Inter-residue correlations include H_α, H_β, C_α, C_β, C_γ, CO, or N_H chemical shifts. Although H_α and H_β shifts are seldom available from deuterated proteins, they have been included such that MONTE can also be used in the assignment of protonated proteins. The score of each available inter-residue connectivity is evaluated in the following fashion. Consider two spin-systems that are mapped to adjacent residues (i − 1, *i*) on the primary sequence. The program compares the inter-residue shift associated with spin system assigned to the ith residue to the intra-residue shift of the spin-system assigned to the (i − 1)th residue. The difference in these chemical shifts are used as the argument to a Gaussian function, $g(\delta^{(i)}_{\text{inter}} - \delta^{(i-1)}_{\text{intra}})$ (see Figure 1). The value of this function gives the score for a single inter-residue connectivity. Each type (e.g., C_β) of inter-residue connectivity can be given a different Gaussian function by varying both the height and width of the Gaussian function. Changing the height gives different types of connectivities different relative weights. Changing the width allows each type of connectivity to have a different response to the size of chemical shift mis-matches. Finally, it is also possible to generate a negative displacement of the

Gaussian function; this provides a repulsive term that discourages mismatched inter-residue chemical shifts.

The initial width is usually set equal to the digital resolution of the corresponding spectrum. This is likely to overestimate the distribution of chemical shift matches since the chemical shifts are more precise than the digital resolution. At the end of the assignment process the program reports the observed distribution of chemical shift matching based on the unambiguous assignments. Consequently, the user can elect to use the modified widths in future runs of the program.

– How well amide-amide NOE's match the pattern predicted from the known tertiary structure of the protein. If the tertiary structure is not known, then the program will generate an β-strand configuration for the chain in order to predict sequential NOEs. If the secondary structure is known, but the tertiary structure is unknown, then NOEs are generated from the secondary structure. Either a four dimensional amide-amide NOE data set can be utilized, or the two corresponding three-dimensional data sets (e.g., H_N-N-H_N and N-N-H_N). The program calculates the score for all predicted NOEs. To evaluate the score of any particular mapping, the contribution of each predicted NOE to the overall score is based on the difference between the chemical shift of the NOE crosspeak and the amide chemical shift of the other spin-system that is currently mapped to the residue participating in the NOE. The actual score is determined by using the difference in the amide shifts as the argument to the Gaussian scoring function. NOEs that are predicted from the input structure, but are missing from the data, do not contribute to the score.

– How well the intra-residue chemical shifts of the spin-system agree with the type of amino acid to which it is mapped. The probability of residue type is evaluated using CO, C_α, C_β, C_γ, H_α, H_β and N chemical shift distributions available from BioMagResBank (Seavey et al., 1991). If the secondary structure of the protein is known, then the expected chemical shifts for each residue are modified to agree with those observed for that residue within the particular secondary structure, as provided by the chemical shift distributions from the BioMagResBank. Secondary structural information can be provided by the user or it can be automatically extracted from the known tertiary structure of the protein. Currently, there is no provision for carbon TOCSY data (Gardner et al., 1996), however much of the residue type information in the TOCSY experiment can be supplied to MONTE via the γ-carbon. In the case of deuterated proteins, the chemical shifts of the carbon atoms are automatically adjusted for the deuterium isotope effect (see Venters et al., 1996) The contribution of residue type to the overall score is also adjustable by changing the relative weight of this contribution to the overall score.

– How well the inter-residue chemical shifts of the spin-system agree with the type of amino acid that precede the residue to which it is mapped. This contribution to the score is evaluated as described above for intra-residue chemical shifts.

– Direct determination of the residue type of a spin-system can also be utilized in the assignment process. The amino-acid type of a spin-system can be identified by either labeling the sample with specific ^{15}N-labeled amino acids (Ou et al., 2001; Lee et al., 1995; McIntosh and Dahlquist, 1990), or methods the identify methyl containing spin systems (see Gardner et al., 1996; Tugarinov et al., 2002), or by the use of residue specific pulse sequences that elucidate the residue type of a spin-system (Schubert et al., 2001; Dötsch et al., 1996). Alternatively, it is also possible to identify the amino acid type of spin-systems by uniformly labeling the protein with ^{15}N and incorporating an single type of amino acid that contains a ^{13}C at its carbonyl position. In this case the coupling between the ^{13}C on the carbonyl of the specifically labeled amino acid and the ^{15}N on the amide of the following residue can be used to edit the HSQC spectrum, leaving only those resonances from spin-systems that follow the amino acid that was labeled with ^{13}C (Griffey et al., 1986; Rule et al., 1993; McCallum et al., 1999).

Residue-specific information is incorporated into the assignment process by specifying both the residue type and the amide proton and nitrogen chemical shifts of the resonance lines that were observed in spectra from residue specific pulse sequences or in samples that have been specifically labeled ^{15}N samples. Information from samples that were specifically labeled with ^{13}C at the carbonyl position is also incorporated in the same way, however the residue-type information refers to the residue that precedes the spin-system. The amide nitrogen and proton chemical shifts of the peak positions are used to evaluate the likelihood that a spin-system is a particular residue (^{15}N labeling) or is adjacent to a particular residue (^{13}C carbonyl labeling). In this case the argument to the Gaussian function is the difference between the amide nitrogen and proton chemical shifts of the peaks in the specifically labeled sample(s) and the amide nitrogen and proton shifts of the spin-system. For example, if the amide

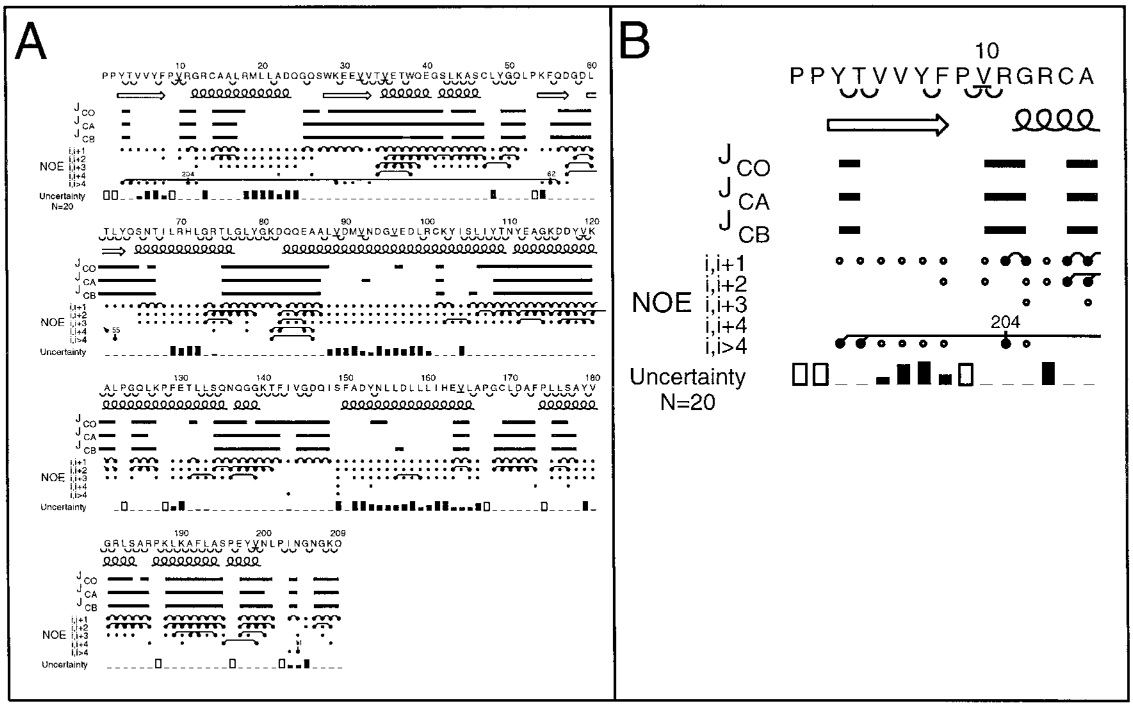

Figure 2. Summary of assignments. A complete summary of the assignments for the entire protein is shown in (A). An enlarged region of this figure is shown in (B). The amino acid sequence is given at the top of each row. The half-circle symbol under two adjacent residues indicates that this assignment agrees with information from specific ^{13}C-1 labeling. For example, the amide of Thr4 was identified as being coupled to a ^{13}C carbonyl in a sample that was uniformly labeled with ^{15}N and specifically labeled with ^{13}C-1 Tyr. An underlined residue indicates that the assignment is consistent with specific ^{15}N labeling. For example, the amide resonance for the spin-system that was assigned to Val10 was present in a sample labeled with ^{15}N Val. The following line indicates the secondary structure, either provided by the user or calculated from the three-dimensional structure of the protein. The next three lines, labeled with J_X, indicate the fidelity of matching of inter- and intra-residue chemical shifts. The thickness of the bar is related to how well the chemical shifts match. Thinner lines indicate poorer matches, such as the J_{CB} coupling between residues 37 and 38 (A). In (B), all of the indicated matches are within one σ of the Gaussian function that was used in scoring, hence all have the same thickness. The rows that display the information from NOE data show consecutive NOEs (i, i + 1), local NOEs, as well as long-range NOEs (i, i > 4). An open circle indicates that the NOE was predicted from the input structure (either primary or tertiary). A filled circle indicates that the NOE was found in the experimental data. Where possible, filled circles are connected by a solid line between the two coupled amide protons. For example, panel A shows long-range NOEs between residues 3 and 4 and 55 and 56, respectively. In the three dimensional structure these residues are across from each other on a β-sheet. If it is not possible to connect the coupled residues by a line, because they reside on separate lines in the output figure, then the NOE partner is indicated by the residue number. For example, Arg11 shows an NOE to Asn204. The last line of this figure provides information on the uncertainty of the assignments. A bar is printed under each residue, with a height that is proportional to the number of different spin systems that were assigned to this particular residue in the independent trials. A zero height bar (thin line) indicates that the same solution was found in all of the independent trials. For example Tyr3, Thr4, Val10, Arg11, Gly12, Cys14 and Ala15 appear to be uniquely assigned with this data. In contrast, residues Val5-Phe8 are not assigned for reasons discussed in the text. Open rectangles mark the location of Pro residues.

nitrogen and proton shifts of a spin-system are close to those of a peak observed in a ^{13}C-edited HSQC spectrum from a sample labeled with ^{13}C Alanine at the carbonyl position, then a high score will be obtained if this spin-system is mapped to a residue that follows Alanine in the primary sequence. The correct mapping of these peaks on to the primary sequence is indicated on the summary output figure. The user is alerted if the peaks from specifically labeled samples appear not to match the correct residue.

– Chemical shift information from one form of the protein can be used to assign another form (e.g., unliganded and liganded). In this case the similarity in the position of the crosspeak in three dimensional spectra is compared between the tentative assignment and the known assignment of the previously assigned protein variant Peaks that are in a similar position give a positive contribution to the score while mis-matched peaks do not affect the score.

The recommended scoring parameters have been optimized using several experimental and simulated data sets. However, it may be necessary for the user to modify the parameters in response to the quality of the experimental spectra. For example, it may be desirable to reduce the weighting of inter-residue Cβ chemical shift matching because of poor signal-to-noise in the HN(CA)CB spectrum.

The parameter set associated with each run is modified using a graphical user interface coded in Tcl/Tk (Tool command language). This interface allows the user to modify the width of the Gaussian matching functions, the weights applied to each of the above terms, and the annealing schedule. In addition, the user can specify whether a three dimensional structure or data from a previously assigned variant is to be used in the assignment process.

After the initial random assignments are scored, a simulated annealing-Monte Carlo approach is used to optimize the assignment of spin-systems to residues in the assignment cycle. The annealing schedule is define by the user. It can consist of one or more segments (see Table 1). Each segment is defined by a beginning temperature, a final temperature, and a temperature step. In addition, the scoring of NOEs and the repulsive term for each type of inter-residue chemical shift matching can be changed in each segment. If the scoring function is changed from segment to segment it is recommended to begin the annealing at an elevated temperature to insure that the system will reach equilibrium under the new scoring potential (see Table 1).

The program optimizes the assignment by exchanging, or swapping, one or more consecutive spin-systems from within the primary sequence with an identically sized collection of spin-systems from either the primary sequence or from the cache. These segments are selected randomly from within the primary sequence and are of random length, with the maximum size defined by the user. This new mapping is then scored. If the score improved then the new mapping is retained. If the score is lowered, then a decision is made to either keep or discard the new mapping. This decision is based on the ratio of the decrease in score to the current temperature of the system. If the decrease in the score is equal to the current temperature then, on average, 1/e of the solutions are retained. If the decrease in score is smaller than the current temperature then the probability of retaining the new mappings is larger than 1/e. A decrease in the score that exceeds the current temperature causes the probability to be less than 1/e; the larger the decrease the lower the probability. The temperature is initially set sufficiently high that most proposed swaps are accepted. The temperature is gradually lowered during the run, consequently it becomes increasingly less likely to accept swaps that decrease the overall score. To insure that the system remains in equilibrium during the annealing process it is necessary to use a large number of swaps at each temperature step. In practice, the minimum number of swaps is defined as that required for the score to convergence within each annealing segment. The suggested tolerance for convergence is a change in the score of less than 1 part in 10^3.

Results

The following example describes the assignment process of a 209 residue homodimeric protein using MONTE. Assignment of this protein was particularly difficult because a large number of amides were missing from the spectrum because the amide deuterons failed to exchange with solvent protons. Unfortunately, it is not possible to denature and then re-nature this particular protein to facilitate the amide exchange process. Data were available from the following experiments: HNCA, HN(CO)CA, HNCB, HN(CO)CB, HN(CA)CO, HNCO and a 4D amide-amide NOESY. In addition, a number of samples that were labeled with specific amino acid types at either the amide position with ^{15}N or at the carbonyl position with ^{13}C were also utilized. Although the known tertiary structure of the protein was utilized to enhance the interpretation of information from the 4D-NOESY experiment, identical assignments were obtained without the use of the tertiary structure.

The annealing schedule and the widths of the inter-residue Gaussian scoring function are given in Table 1. The resultant output from 20 independent cycles of the program is shown in Figure 2. Each cycle required approximately 5 min on a 180 MHz SGI R5000 computer, resulting in a total run time of less than two hours to calculate all 20 solutions. Figure 2 provides information on the extent of inter-residue chemical shift matching, concordance with predicted NOESY peaks, and how information from specifically labeled samples agrees with the best assignment solution. In addition, the summary output also indicates which regions are less confidently assigned based on the degree of ambiguity in the assignment of spin-systems to specific residues in the protein. This information is useful in planning further experiments to reduce these ambiguities. For example, Figure 2B indicates that the

Table 1. Annealing schedule for assignment of GSTP1-1

Segment #	T-start	T-end	T-step	N-swaps	γ	Swap size	NOE score	Inter-residue repulsive terms		
								CO	C_α	C_β
1	200	100	10	10 000	2	1	0	0	0	0
2	150	80	10	10 000	3	2	0.2	40	40	40
3	120	20	10	50 000	5	3	0.3	100	100	100
4	100	10	10	100 000	10	4	0.4	200	200	200

T-start, T-end, and T-step give the starting temperature, the final temperature, and the temperature step for each segment, respectively. N-swaps is the number of swaps per temperature step. This is modified by e^γ at lower temperatures in order to increase the number of effective swaps at low temperature. The maximum number of sequential spin systems to be swapped (Swap size) is indicated in the 6th column. The NOE score gives the contribution to the overall score of matching a single predicted NOE to an experimental NOE. The last three columns show the repulse terms for matching of inter-residue CO, C_α, and C_β shifts. The width of the Gaussian distribution used for matching N, HN, CO, C_α, and C_β shifts were 0.8, 0.2, 0.3, 0.3, 0.3 ppm, respectively.

sequence TVVY is not uniquely assigned, indicating that a number of different spin-systems were assigned to these residues during the 20 independent assignment trials. Labeling the protein with ^{13}C-^{15}N Val may be helpful in assigning this stretch of the primary sequence. In actual fact, this region of the protein is buried and exchanges slowly. Consequently, no assignments are possible in this particular case unless partially deuterated samples are used (see McCallum et al., 1999).

The assignment ambiguities are also presented in graphical form (see Figure 3). In this figure, an off-diagonal element indicates that one alternative spin system was assigned to a particular residue. For example, the off-diagonal element that is circled in Figure 3 indicates that the spin system that is assigned to residue 87 in the highest scoring solution was also assigned to residue 179 in some of the trials. A long series of off-diagonal elements that are parallel to the diagonal indicate alternative assignments for a series of sequentially connected residues. As the quality of the assignments improve, the number of off-diagonal elements decreases (see Hitchens et al., 2002).

Detailed output, in the form of an HTML file, is also generated by the program (not shown) This presentation of the data highlights three key features of the analysis. Color coded bars identify inter-residue chemical shift mismatches, facilitating the correction of recorded chemical shifts or identifying possible errors in the assignments. In addition, the NOEs that were predicted from the input structure are displayed. Those that match the input experimental data are highlighted in bold type. Finally, the lower frame of

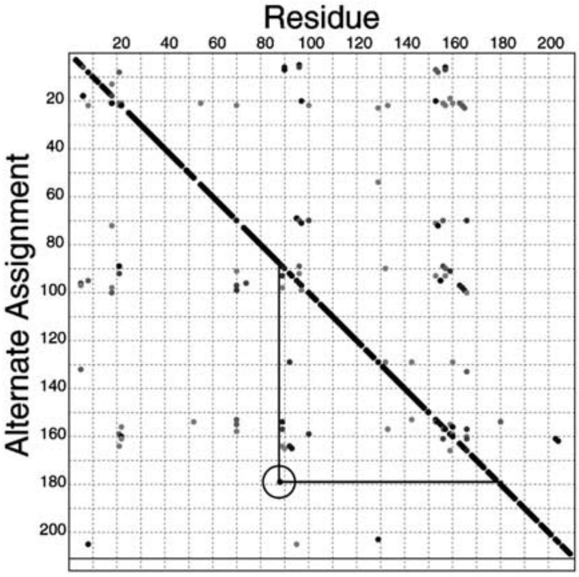

Figure 3. Correlation plot indicating uniqueness of assignments. Additional detail on the existence of alternative assignments is presented in a correlation plot. The x-axis indicates the residue number and the y-axis indicates to what other residues, besides that found in the best solution, the spin system was assigned to. For example, the circled point in the plot indicates that the spin system that was most frequently assigned to residue 88 is also assigned to 179 in one or more of the 20 solutions. The intensity of the plotted point is proportional to the frequency that a spin-system is assigned to a particular residue. The more darkly colored the point, the higher the frequency of the assignment of a single spin-system to a particular residue. For example, the dark diagonal in this plot shows that most spin-systems are uniquely assigned to a single-residue. The rectangular area below the lower horizontal line represents the cache area. Points found in this region indicate that spin-system was placed in the cache (i.e., unassigned) in some or all of the assignment solutions.

the HTML document reports the residue-type probabilities calculated for a spin-system as well as the residue-type probability of the residue that precedes that spin system. It is often useful to compare the predicted residue type to the assigned residue type to identify incorrectly recorded chemical shifts.

Conclusions

A general-purpose Monte Carlo assignment program has been developed to aid in the assignment of NMR resonances from proteins. The program is flexible and is capable of incorporating a wide variety of source data in the assignment process. Although the program is tailored to facilitate the assignment of large deuterated proteins, it can also been used to assign backbone and H_β protons in smaller protonated proteins. Planned improvements of the program include the use of Genetic algorithms to increase the efficiency of the Monte Carlo search and the use of predicted residual dipolar couplings in conjunction with secondary and tertiary structural information to facilitate assignments.

Software availability

This program is freely available to academic users. Details on how to obtain the program can be obtained from the corresponding author. Information on the most current version of the program can be found at the following URL: http://www.andrew.cmu.edu/~rule/monte/

Acknowledgements

This work was supported by an NIH grant to GSR (GM61253), the Eberly Family Professorship in Structural Biology to G.S.R and a NIH NSRA fellowship to T.K.H. J.A.L. is supported by NIH grant HL-24525. We thank Hunter Moseley for helpful discussions.

References

Atreya, H.S., Sahu, S.C., Chary, K.V.R. and Govil, G. (2000) *J. Biomol. NMR*, **17**, 125–136.

Bailey-Kellogg, C., Widge, A., Kelley, J.J., Berardi, M.J., Bushweller, J.H. and Donald, B.R. (2000) *J. Comput. Biol.*, **7**, 537–558.

Bartels, C., Güntert, P., Billeter, M. and Wüthrich, K. (1997) *J. Comp. Chem.*, **18**, 139–149.

Dötsch, V, Oswald, R.E. and Wagner, G. (1996) *J. Magn. Reson.*, **110B**, 107–111.

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

Gardner, K.H., Konrat, R., Rosen, M.K. and Kay, L.E. (1996) *J. Biomol NMR*, **8**, 351–356.

Griffey, R.H. and Redfield A.G. (1986) *J. Am. Chem. Soc.*, **108**, 6816–6817.

Grishaev, A. and Llinas, M. (2002a) *Proc. Natl. Acad. Sci. USA*, **99**, 6713–6718.

Grishaev, A. and Llinas, M. (2002b) *Proc. Natl. Acad. Sci. USA*, **99**, 6707–6712.

Güntert, P., Slazmann, M., Braun, D. and Wüthrich, K. (2000) *J. Biomol. NMR*, **18**, 129–137.

Hitchens, T.K., Mannervik, B. and Rule, G.S. (2001) *Biochemistry*, **40**, 11660–11669.

Hitchens, T.K., McCallum, S.A. and Rule, G.S. (2002) *J. Biomol. NMR*, **25**, 11–23.

Kraulis, P.J. (1994) *J. Mol. Biol.*, **243**, 696–718.

Lee, K.M., Androphy, E.J. and Baleja, J.D. (1995) *J Biomol NMR*, **5**, 93–96.

LeMaster D.M. (1990) *Quart. Rev. Biophys.*, **23**, 133–174.

Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**, 31–43.

Li, K.-B. and Sanctuary, B.C. (1997) *J. Chem. Inf. Comput. Sci.*, **37**, 467–477.

Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.

McCallum, S.A., Hitchens, T.K. and Rule G.S. (1999) *J. Mol. Biol.*, **285**, 2119–2132.

McCallum, S.A., Hitchens, T.K., Torborg, C. and Rule, G.S. (2000) *Biochemistry*, **39**, 7343–7356.

McIntosh, L.P. and Dahlquist, F.W. (1990) *Quart. Rev. Biophys.*, **23**, 1–38.

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

Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.* **9**, 635–642.

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

Ou, H.D., Lai, H.C., Serber, Z. and Dötsch, V. (2001) *J. Biomol. NMR*, **21**, 269–273.

Rule, G.S., Tjandra, N., Simplaceanu, V. and Ho, C. (1993) *J. Magn. Reson.*, **B102**, 126–128.

Schubert, M., Smalla, M., Schmieder, P. and Oschkinat, H. (2001) *J. Biomol. NMR*, **20**, 379–384.

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

Tugarinov, V., Muhandiram, R., Ayed, A. and Kay, L.E. (2002) *J. Am. Chem. Soc.* **124**, 10025–10035.

Venters, R.A., Farmer, B.T., Fierke, C.A. and Spicer, L.D. (1996) *J. Mol. Biol.*, **264**, 1101–11166.

Wand, A.J. and Nelson, S.J. (1991) *Biophys. J.*, **59**, 1101–1112.

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

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

Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C.-Y, Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.*, **269**, 592–610.