Journal of Biomolecular NMR **29:** 1–9, 2004.

Assignment of protein backbone resonances using connectivity, torsion angles and $^{13}C^{\alpha}$ chemical shifts

Laura C. Morris, Homayoun Valafar & James H. Prestegard*
Complex Carbohydrate Research Center, University of Georgia, 220 Riverbend Road, Athens, GA 30602-4712, U.S.A.

Received 29 July 2003; Accepted 21 November 2003

Key words: assignment, chemical shift, probability density function, protein backbone, structural genomics, torsion angle

Abstract

A program is presented which will return the most probable sequence location for a short connected set of residues in a protein given just $^{13}C^{\alpha}$ chemical shifts ($\delta(^{13}C^{\alpha})$) and data restricting the φ and ψ backbone angles. Data taken from both the BioMagResBank and the Protein Data Bank were used to create a probability density function (PDF) using a multivariate normal distribution in $\delta(^{13}C^{\alpha})$, φ, and ψ space for each amino acid residue. Extracting and combining probabilities for particular amino acid residues in a short proposed sequence yields a score indicative of the correctness of the proposed assignment. The program is illustrated using several proteins for which structure and $^{13}C^{\alpha}$ chemical shift data are available.

Introduction

The first step in a traditional approach to NMR structure determination of a protein is sequential assignment of backbone resonances. The most reliable route to assignments relies heavily on a suite of four or more triple resonance experiments collected on proteins uniformly enriched to high levels in both ^{15}N and ^{13}C (Kanelis et al., 2001; Lian and Middleton, 2001). This route to assignment has been automated to a great extent (Atreya et al., 2002; Bartels et al., 1996; Coggins and Zhou, 2003; Moseley et al., 2001), but the amount of data, time required for data acquisition, and the effort required to label protein is still significant. There are now a number of activities that could benefit from a reduction of time and effort devoted to assignment. For example, structural genomics, where rapid structure determination of a large number of proteins is a goal, would clearly benefit, and drug design, where assignment of protein resonances is often a prerequisite to NMR based ligand screens, may also benefit. Both activities present opportunities for use of prior structural knowledge in lieu of exhaustive data acquisition for assignment purposes. Here we present a program that capitalizes on these opportunities; it relies largely on connectivity and $^{13}C^{\alpha}$ shift ($\delta(^{13}C^{\alpha})$) data that can be collected in a single experiment on partially labeled samples.

As a part of a structural genomics effort a procedure in which three NMR experiments yield all the data required for resonance assignment and structure determination has been recently devised (Tian et al., 2001). Data from these three experiments (phase-modulated HSQC (Tolmann et al., 1996), soft HNCA-E.COSY (Weisemann et al., 1994) and 2D IP-HSQC (Wang et al., 1998)), when combined, provide $\delta(^{13}C^{\alpha})$, $^{13}C^{\alpha}_{i} - {}^{13}C^{\alpha}_{i-1}$ connectivities and residual dipolar couplings (RDCs). In this particular case the $\delta(^{13}C^{\alpha})$ and $^{13}C^{\alpha}_{i} - {}^{13}C^{\alpha}_{i-1}$ connectivities are provided by a soft HNCA-E.COSY, but any HNCA type experiment (Kay et al., 1990) could provide this information. The additional information provided by the soft HNCA-E.COSY comes in the form of RDCs. The combined RDCs from all three experiments permit calculation of the relative orientation of the peptide planes, thereby providing structural information in the form of back-

*To whom correspondence should be addressed. E-mail: jpresteg@ccrc.uga.edu

bone torsion angles (φ, ψ). Thus, the Tian procedure provides connectivity of small fragments and defines backbone torsion angles (φ, ψ) before assignment of amino acid residue types or sequential location. This information will be exploited along with backbone atom chemical shift data in the assignment tool presented. The same tool may be applicable to cases where φ restrictions are derived from measurements of scalar coupling and ψ restrictions from cross-correlation effects (Schwalbe et al., 2001), rather than structure determination of a connected fragment.

In the drug discovery area the study of protein ligand interaction by NMR has also blossomed (Stockman and Dalvit, 2002). In some applications, perturbation of protein NMR resonances (HSQC cross-peaks) by ligand addition is a primary source of information. Often x-ray structures of the target proteins exist and resonance assignment becomes the primary obstacle. Here, φ and ψ information from deposited structure files can be used, thereby reducing the experimental requirements to one HNCA experiment (Kay et al., 1990), or a combination of HNCA and HN(CO)CA (Bax and Ikura, 1991) experiments, to obtain the necessary experimental $^{13}C^{\alpha}$ shifts and connectivities. An expeditious determination of the sequential position of peptide fragments using these easily obtained data would clearly expedite research.

Methods

It is well known that $^{13}C^{\alpha}$ chemical shifts are sensitive to both amino acid residue type and local backbone (secondary) structure. Normally $\delta(^{13}C^{\alpha})$ resonances are assigned to amino acid residue types and the deviation of chemical shifts from random coil values for the amino acid residue is used to deduce secondary structure (Spera and Bax, 1991; Wishart and Case, 2001). The situations described above are different in that local structures are known first. Since the BioMagResBank's (BMRB URL: http://www.bmrb.wisc.edu; (Seavey, 1991; Ulrich et al., 1998)) database of chemical shifts and other NMR data contains thousands of entries, it seemed possible to combine shift data from the BMRB with structural data available from the Protein Data Bank (PDB URL: http://www.rcsb.org/pdb; (Berman et al., 2000)) to provide a statistic that can predict amino acid residue type from $^{13}C^{\alpha}$ shift and local structure characteristics.

Chemical shifts from proteins deposited in the BMRB were associated with protein structures in the PDB. For the associated proteins, two different data sets were obtained. The first contains ($\delta(^{13}C^{\alpha})$, φ, ψ) data for each amino acid residue. To construct this data set dihedral angles were determined from PDB files using the program **dang** (URL: http://kinemage.biochem.duke.edu; (Word, 2000)). The resulting number of data points for each amino acid residue is given in Table 1. These correspond to roughly the same amino acid residue distribution percentages as are available in the BMRB. Approximately half of the $^{13}C^{\alpha}$ chemical shifts available in the BMRB were used. The second data set was derived from the first and contains ($\delta(^{13}C^{\alpha})$, $^{3}J_{HNHA}$) data points. Couplings were calculated from the previously determined φ angles using a standard Karplus equation (1)

$$^{3}J_{HNHA} = A\cos^{2}(\varphi - 60) + B\cos(\varphi - 60) + C \quad (1)$$

with two different parameter sets, A = 6.4, B = −1.4, C = 1.9 (Pardi et al., 1984) and A = 6.51, B = −1.76, C = 1.6 (Vűister and Bax, 1993). Results obtained from both parameter sets were indistinguishable.

The plot of a probability density function (PDF) could be represented as a three dimensional ($\delta(^{13}C^{\alpha})$, φ, ψ, frequency) or two dimensional ($\delta(^{13}C^{\alpha})$, $^{3}J_{HNHA}$, frequency) histogram in which the amplitude of a point is simply the number of occurrences within appropriate φ, ψ and $\delta(^{13}C^{\alpha})$ or $^{3}J_{HNHA}$ and $\delta(^{13}C^{\alpha})$ intervals. However, kernel density estimation (Silverman, 1986) is more flexible than histograms because it avoids the problems associated with bins (size and placement) and allows the use of kernels, or error functions, which better describe the distribution of data. As an example, Figure 1 shows the distribution of the $^{13}C^{\alpha}$ chemical shifts obtained for glutamine. If we were to use a histogram to describe this distribution we would have to choose a bin size. For purposes of illustration, let us suppose that the error in measurement of the chemical shift is ± 0.1 ppm. The top plot (Figure 1a) is an example of what would happen if a histogram with a bin size of 0.2 were used to describe the distribution. The highest point (at 59.0 ppm) has a normalized value of 0.0575 while the bin next to it (58.8 ppm) has a value of 0.0343. In addition there are many areas for which the probability density is zero. These discontinuities and zero points are avoided when using PDFs. Figure 1b illustrates the distribution of data using a normal distribution for the kernel

Table 1. Number of data points used to determine the probability distribution over ($\delta(^{13}C^{\alpha})$, φ, ψ) for each amino acid residue

Amino acid	# Data points	%[a] occurrence	Amino acid	# Data points	%[a] occurrence
ala	1521	7.5	leu	1779	8.8
arg	919	4.5	lys	1590	7.8
asn	793	3.9	met	378	1.9
asp	1312	6.5	phe	757	3.7
cys	313	1.5	pro	870	4.3
gln	817	4.0	ser	1206	5.9
glu	1591	7.8	thr	1138	5.6
gly	1446	7.1	trp	217	1.1
his	467	2.3	tyr	601	3.0
ile	1137	5.6	val	1426	7.0

[a]The percent occurrence of each amino acid residue.

to estimate the distribution. In this case the standard deviation was chosen to be 0.2 ppm. This oversimplified one dimension example is meant to give a general feel for the issues regarding histograms and not to actually provide information about the distribution of a particular amino acid residue's $^{13}C^{\alpha}$ chemical shifts. Kernel density estimation allows one to determine the PDF with fewer assumptions thereby reducing the propagation of error over multiple dimensions.

For our kernel we have chosen a multidimensional normal distribution function (2) that allows us to compensate for regions of $\delta(^{13}C^{\alpha})$, φ, ψ space or $\delta(^{13}C^{\alpha})$, $^{3}J_{HNHA}$ space containing sparse data. In the kernel equation below, $\vec{x}$ is the point at which the PDF is being calculated ($\delta(^{13}C^{\alpha})$, φ, ψ) or ($\delta(^{13}C^{\alpha})$, $^{3}J_{HNHA}$), p is the dimensionality of the problem (2 or 3 in this study), Σ is the covariance matrix, and $\overrightarrow{x_i}$ represents a point in the experimental data set. This kernel was then used to generate a continuous PDF from the discretely sampled data for each amino acid residue.

$$PDF(\overrightarrow{x}) = (2\pi)^{-p/2} \left|\sum\right|^{-1/2} \exp\{-1/2(\overrightarrow{x} - \overrightarrow{x_i}) \cdot \sum^{-1}(\overrightarrow{x} - \overrightarrow{x_i})\}. \quad (2)$$

The resulting function for each amino acid residue covers $^{13}C^{\alpha}$ chemical shifts ranging from 40 to 70 ppm and all of φ, ψ space for the first data set and the same chemical shift range and $^{3}J_{HNHA}$ values from 1.5 to 10 Hz for the second data set. All of the resulting PDFs were normalized over all space for each amino acid residue. The resulting PDFs were then normalized over all amino acid residues such that the sum of the probabilities for a specific $\vec{x}$ is one.

The utility of the PDFs in discriminating amino acid residue types, given a $^{13}C^{\alpha}$ chemical shift, a φ, and a ψ, is a complex function of differences in the shapes of the multidimensional PDFs. If values in the PDFs were narrowly distributed, displacements of the means for various amino acid residues would provide a good measure of discrimination power. Our means are represented in the form of the vector, $\vec{x}$, which indicates the location of the center of mass for the collected data on each amino acid residue. However, our distributions are broad leading to extensive overlaps of PDFs for different amino acid residues. Hence, the covariance matrix Σ, also plays an important role in the study of multidimensional discrimination (Fukunaga, 1990). Analogous to the standard deviation in the one-dimensional studies, this matrix provides detailed information regarding the scattering pattern of the data. When diagonalized it can be visualized as hyper-ellipsoids of n^{th} dimension. The eigenvectors of this matrix will represent the direction of the principle axes while the eigenvalues represent the scatter of the ellipsoid (data) in the direction of each principle axis. Discrimination ultimately arises from overlap of differently oriented hyper-ellipsoids that represent a level curve of the distribution functions. We have embodied the extent of these overlaps in our probabilities by normalizing each PDF point by point over all amino acid residues.

Since, for a given $\vec{x}$, the probability for assignment to the best amino acid residue may be only marginally higher than the probability for assignment to the

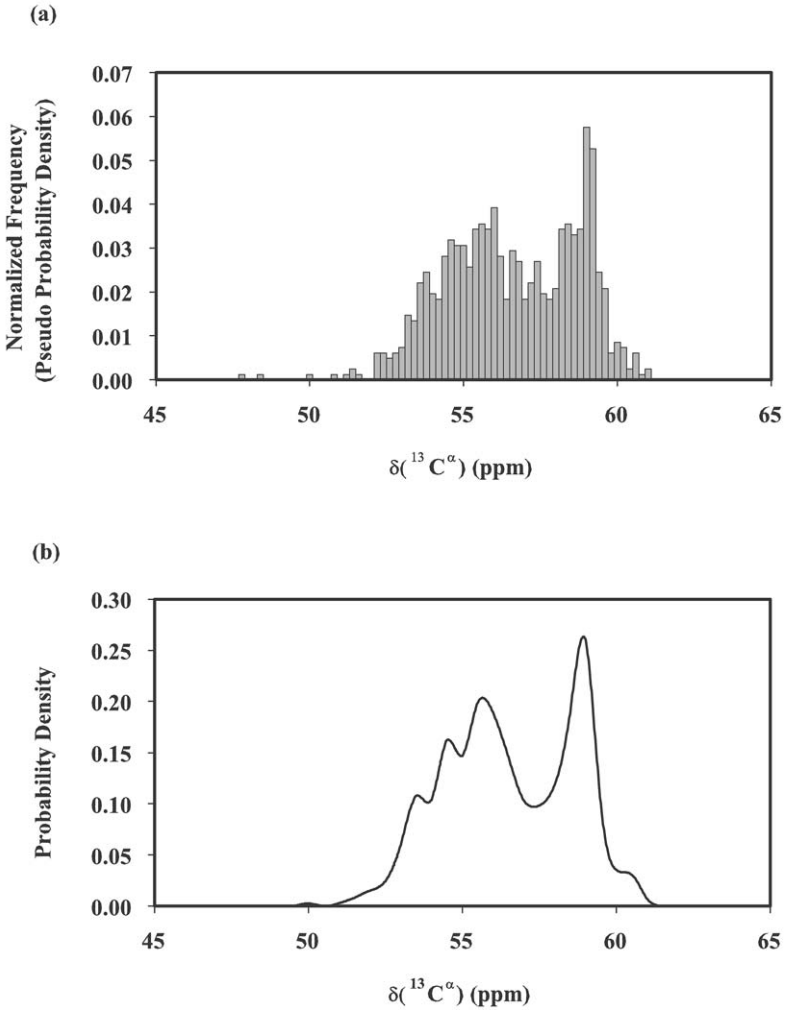

Figure 1. Example of a histogram and a probability density function calculated for the $^{13}C^{\alpha}$ chemical shifts obtained from the BMRB for glutamine. The top plot (a) is a histogram with normalized frequencies (pseudo probability densities). The bottom plot (b) is the plot of a probability density function created using kernel density estimation.

Table 2. Comparison of assignment results for 1M2Y

# Residues in fragment	% Correctly assigned		
	$\delta(^{13}C^{\alpha})$, φ, ψ[a]	$\delta(^{13}C^{\alpha})$, $^{3}J_{HNHA}$[b]	$\delta(^{13}C^{\alpha})$ (+structure)[c]
1	8	2	12
2	36	14	28
3	61	28	43
4	84	45	62
5	95	53	77
6	100	62	86

[a]Using φ, ψ data calculated from dipolar couplings.
[b]Using experimental $^{3}J_{HNHA}$ values.
[c]Using a modification in which the structure is already known and torsion angles are paired with the residues.

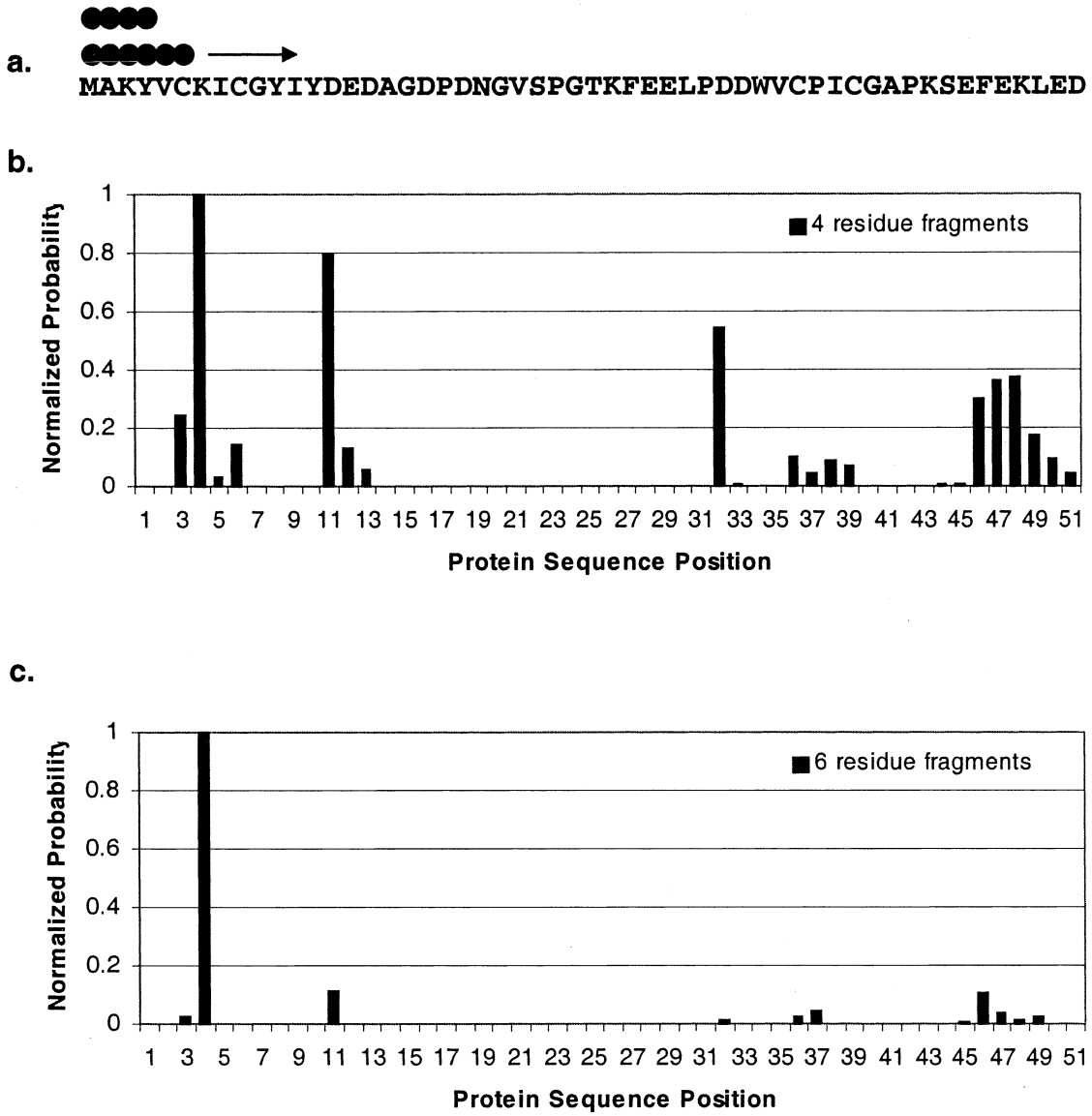

Figure 2. Example of assignment of a fragment to a position in a protein sequence. (a) Sequence of a rubredoxin mutant (1M2Y) from *Pyrococcus furiosus.* Four and six residue fragments (circles) are moved along the sequence as probabilities are calculated for each possible position. Graphs of the normalized probabilities calculated for the 4 residue (b) and 6 residue (c) fragments, ICGY and ICGYIY respectively. The ($\delta(^{13}C^{\alpha})$, φ, ψ) data used for the analysis are 4Y = (56.63, −125.1, 135.1), 5V = (58.28, −110.2, 140.1), 6C = (59.79, −75.1, 64.8), 7K = (59.08, −14.7, −9.9), 8I = (62.43, −115.1, 15.2), and 9C = (59.1, −140, −30.2).

next best amino acid residue, it is necessary to improve discrimination by using the fact that data can be connected for several residues. This allows combined probabilities to be calculated for sequentially connected amino acid residue types appearing in the protein primary sequence. A program that we will call SEASCAPE (SEquential Assignment by Structure and Chemical shift Assisted Probability Estimation) was therefore written to make assignments based on known connectivities between any number of resonances, for which $\delta(^{13}C^{\alpha})$, φ and ψ or $\delta(^{13}C^{\alpha})$ and $^{3}J_{HNHA}$ are available. Given data from the fragment to be assigned, the program takes the protein sequence and PDFs for each amino acid residue and calculates the combined probability for the fragment to be placed at each possible position in the sequence. A section of contiguous residues for which the connectivities and each residue's $\delta(^{13}C^{\alpha})$, φ and ψ, or

$\delta(^{13}C^{\alpha})$, $^{3}J_{HNHA}$ are known (hereafter called a fragment) is aligned along the beginning of the sequence, and the probability that the fragment is correctly positioned is calculated. The fragment is then repositioned by sliding it over one residue and the probability at that position is calculated. This procedure is repeated until all of the positional probabilities have been calculated. The most probable alignment is the position with the highest calculated probability. The program is written in C++ and is available from our web site (http://tesla.ccrc.uga.edu).

Results and discussion

Figure 2 illustrates the operation of the program on a mutant of rubredoxin from *Pyrococcus furiosus* (1M2Y). The example fragments used include a four residue segment from 4Y to 7K and a six residue segment from 4Y to 9C. For the four residue fragments the highest probability is generally less than a factor of two greater than the second highest probability. But due to variations in this difference, the confidence is not particularly high even though the position with the highest probability is correct. For the six residue fragments the highest probability is, on the average, slightly more than an order of magnitude greater than the next highest probability. When probed with known fragments, the robustness of the method was found to be steeply dependent upon the length of the fragment. Table 2 shows a comparison of how often the correct assignment resulted in the highest probability for fragments of different lengths given either experimental $\delta(^{13}C^{\alpha})$, φ, ψ data (column 1) or experimental $\delta(^{13}C^{\alpha})$, $^{3}J_{HNHA}$ data (column 2) averaged over all possible assignments for 1M2Y. In these cases, φ and ψ were directly determined from NMR data and are associated with positions in the connected amino acid residues of the fragment. The utility of assignment using just $^{13}C^{\alpha}$ shifts and $^{3}J_{HNHA}$ data seems marginal unless very long stretches of connectivities can be established, but with $^{13}C^{\alpha}$, φ, and ψ data, sequential stretches of five or more prove to give reliable assignments. Also included in Table 2 is a case in which φ and ψ angles are taken from the structure (column 3). Here φ and ψ values associated with positions in the fragment vary as the fragment is moved down the sequence. The percent of correct assignments is slightly less, but largely parallels that for experimentally determined φ, ψ angles given in column one.

The program was further tested using 10 proteins chosen from a set having a significant percentage of their $^{13}C^{\alpha}$ chemical shifts deposited in the BMRB and a structure available from the PDB. Here, angular constraints from either X-ray or NMR derived structures and $^{13}C^{\alpha}$ data from the BMRB were used (Table 3). These cases represent those in which assignments may be sought for the purpose of ligand screening using HSQC data, and just an HNCA data set may have been collected to obtain $^{13}C^{\alpha}$ shifts and connectivity information. Table 3 clearly indicates some variability in the level of successful assignments, but in all proteins (with one notable exception), a six residue connected fragment can be placed in the sequence with greater than 70% certainty using just $^{13}C^{\alpha}$ shifts, φ and ψ data. The one exception is a DNA binding protein (1IRF) whose binding domain has been categorized as a novel subgroup of the winged helix-turn-helix family (Furui et al., 1998).

In order to understand why assignment of the 1IRF protein proved difficult we repeated the analysis with structural data from a corresponding crystal structure (2IRF). The crystal structure is from a DNA bound form of the protein; the structural information was nevertheless combined with $\delta(^{13}C^{\alpha})$ data from solution in the absence of DNA. Four and six residue fragments were again examined and the results are reported in Table 3. When using the crystal structure we obtained double the number of correct identifications of fragment position. It is possible that the dynamic nature of the protein in solution when not bound to DNA contributed to a set of averaged torsion angles that do not correlate well with chemical shift. Indeed, the torsion angles for half of the residues in the protein differ substantially (>30°) between the unbound solution structure and the bound crystal structure.

Why certain of the other proteins show more successful assignments than others might be expected to depend on issues such as amino acid residue composition or secondary structure population. The PDFs of the individual amino acid residues give an indication of how distinctive the distributions are and how well the program might be expect to perform on a fragment of a given composition. In order to simplify representation of data in the 3D PDFs and produce a more user friendly form, the probability densities were summed across the entire $\delta(^{13}C^{\alpha})$, φ, ψ data set for each amino acid residue. Since the densities had previously been normalized over all amino acid residues at each φ, ψ and $\delta(^{13}C^{\alpha})$ point, the resulting sums provide an indication of how well separated

Table 3. Assignment results for a set of 10 proteins

PDB id (#residues)	BMRB accession number	($\delta(^{13}C^{\alpha})$, φ, ψ) % Correct assignments		Experimental structure data type	% Secondary structure[c]	
		4 Residues	6 Residues		α-Helix	β-Strand
1M2Y (54)	5601	84	100	NMR[a]	0	4
1C0V (79)	4146	85	96	NMR	82	0
1QJT (99)	4140	44	70	NMR	42	0
1IRF (112)	4161	23	43	NMR	29	13
2IRF[b] (113)	4161[b]	52	81	X-ray	33	14
1DMO (148)	4056	39	71	NMR	55	5
1SYM (184)	4001	62	84	NMR	55	4
1CDC (198)	4109	78	98	X-ray	0	35
1EZA (259)	4264	39	76	NMR	50	10
1AZM (260)	4022	47	80	X-ray	8	28
1L6N (289)	5316	37	74	NMR	56	0
Average $\pm$ sd[d]		54 ± 21	79 ± 16		37 ± 26	10 ± 12

[a] φ, ψ Values determined using dipolar couplings, see Tian et al. (2001).
[b] Chemical shift data from the unbound solution structure (1IRF) was used in this analysis.
[c] Secondary structure content as listed in the PDB in the Sequence Details section.
[d] Averages and standard deviations for calculated values over all proteins used.

the distributions are. In Figure 3 the sums have been divided by the sum for leucine, which has the smallest sum, to produce a single number related to the value of $\delta(^{13}C^{\alpha})$, φ and ψ information in identifying each amino acid residue. Not surprisingly, glycine has the highest identification value by far (almost 16 $\times$ leucine). Although this could easily be predicted by glycine's unique $^{13}C^{\alpha}$ chemical shift, the reason that threonine has the second highest value is less obvious. Threonine valine, proline, and isoleucine, all have, on average, similar $^{13}C^{\alpha}$ shifts. This suggests that the basis for distinction is more complex.

Surprisingly, regions with regular secondary structure (α-helices and β-strands) do not result in more accurate assignments than regions that lack regular secondary structure (everything not defined as an α-helix or β-strand). Of the proteins tested, about half of the data are for regions lacking regular secondary structure and the accuracy is no worse or better than structured regions. Among the fragments lacking regular secondary structure, 46% of the four residue fragments are assigned correctly while 76% of the six residue fragments are assigned correctly. In the fragments having regular secondary structure, 47% of the α helices and 48% of the β-strands are assigned correctly in the four residue fragments whereas 77% of the α-helices and 68% of the β-strands are assigned correctly. We still do not expect the program to perform well on highly flexible regions of proteins where chemical shifts may not correlate with average structural parameters or in cases where there are redundancies in sequence. This may have in fact been a factor in our second lowest score (1DMO) at 71%.

An indication of the confidence one should have in a given sequential assignment is also an important issue. The raw probability score obtained for a fragment can be used to give this indication. Figure 4 shows the probability of correct assignment for the top score given a fragment of four or six residues. Since individual probability densities in the data sets are always between zero and one, it is not unusual to obtain overall raw scores on the order of 10^{-4} for a four residue fragment and 10^{-6} for a six residue fragment. The highest scores obtained so far are 3×10^{-2} and 2×10^{-3}, for four and six residue fragments respectively. Respective scores for these fragments can be as low as 10^{-5} and 10^{-7}. In cases where the raw score is equal to or greater than 2×10^{-3} for a four residue fragment the probability of a correct assignment is greater than 80%. For a six residue fragment a score equal to or greater than 3×10^{-6} results in a more than 90% probability of correct assignment. In addition, scores at or above 10^{-4} for a six residue fragment lead to assignment with near certainty.

A problem that requires a program to go a step beyond the proper sequential placement of correctly connected fragments is one in which there is some uncertainty in the connection, possibly due to degeneracy

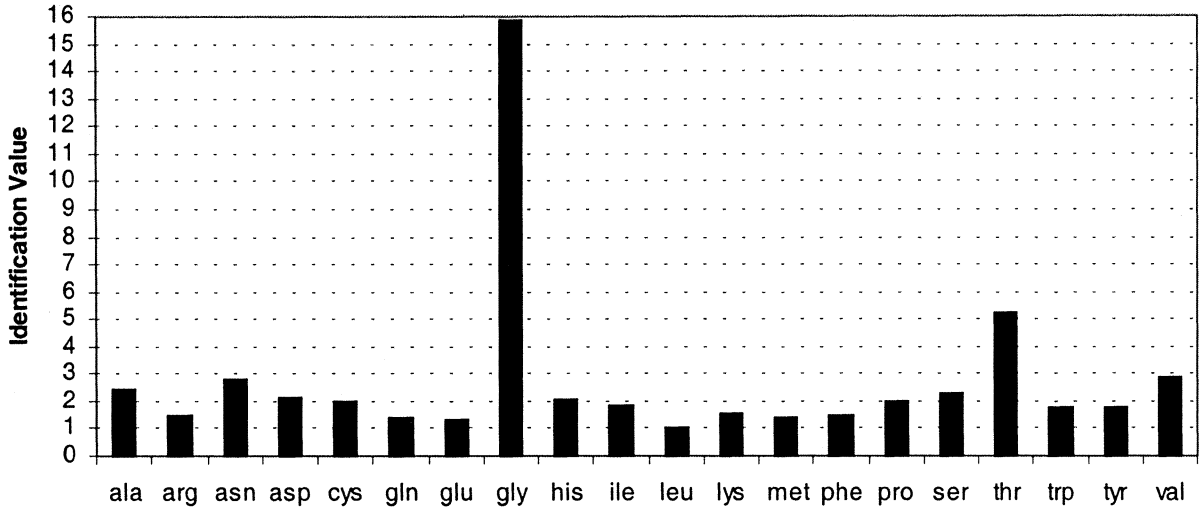

Figure 3. Value of $\delta(^{13}C^{\alpha})$, φ, ψ in identifying an isolated single amino acid residue. All values are relative to the lowest value (leucine) and are multiples of that value.

in $^{13}C^{\alpha}$ shifts used to establish connections between the residues. Connectivities correctly representing the fragment may be determined by comparison of the probabilities of each of the proposed connections. Each of the possible fragments is threaded through the sequence taking the torsion angles, ϕ and ψ, from a proposed structure; the probabilities are calculated for each possible assembly and the highest score, or probability, is used to identify the correct assembly in addition to the proper sequential position. The rightmost column in Table 2 compared the results of this adaptation, using correctly assembled fragments, with the original program in which the $^{13}C^{\alpha}$ chemical shifts were paired with experimentally determined φ and ψ To test the method's ability to identify fragments that are not correctly connected several pairs of sequences 6 or 7 residues in length were threaded through the sequence with one member of the pair being correctly connected and the other incorrectly connected. In all cases the correctly connected sequence gave the highest score and was properly placed in the sequence. The same criteria given above for confidence in assignment, 3×10^{-6} for a six residue fragment and 2×10^{-3} for a four residue fragment, resulted in a confirmation of correct assignment for the six residue fragments but scores for the four residue fragment were below the 80% standard in all cases.

Conclusion

Through this work we have demonstrated a new assignment strategy based on the availability of local backbone geometry and limited chemical shift data. Although results for only one protein using the Tian protocol (Tian et al., 2001) have been shown, examples using larger proteins will follow in the near future. These proteins will enable us to better determine the practical limits of the program with such data. We anticipate applications where high throughput structure determination is an issue and where assignment of resonances for proteins of known structure is a research objective. As expected, the likelihood of correct assignment increases with increased connectivity and success varies somewhat depending on amino acid residue type. A detailed examination of the results indicates that fragments containing glycine are most often identified correctly. This is not surprising given that the $\delta(^{13}C^{\alpha})$ for glycine is the furthest upfield of the amino acid residues and is narrowly distributed. Histidine and tryptophan have so far been the two amino acid residues least likely to be correctly identified, but as with other amino acid residues, their probabilities for correct assignment rise with an increase in the fragment length in which they are found.

The data in Table 2 illustrate the reduction of performance when using $^{3}J_{HNHA}$ in the place of φ, ψ. Although this result is expected, the performance of the algorithm with the use of $^{3}J_{HNHA}$ is still useful. The program presented has relied heavily on $^{13}C^{\alpha}$ shifts

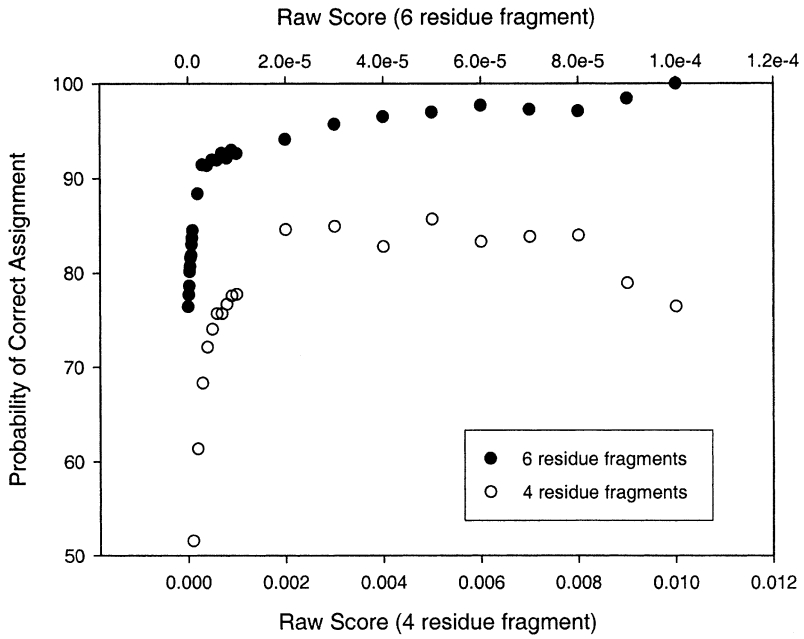

Figure 4. Correct assignment probability as a function of raw score for four residue fragments (bottom scale) and six residue fragments (top scale).

because of their known sensitivity and ready availability from high sensitivity triple resonance experiments. Other carbon, nitrogen, and proton chemical shifts may also be incorporated in future versions of the program.

Acknowledgements

The work presented was supported by grants from the National Science Foundation, MCB0092661 and the National Institutes of Health, GM62407 and RR05351.

References

Atreya, H.S., Chary, K.V.R. and Govil, G. (2002) *Curr. Sci.*, **83**, 1372–1376.

Bartels, C., Billeter, M., Güntert, P. and Wüthrich, K. (1996) *J. Biomol. NMR*, **7**, 207–213.

Bax, A. and Ikura, M. (1991) *J. Biomol. NMR*, **1**, 99–104.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucl. Acids Res.*, **28**, 235–242.

Coggins, B.E. and Zhou, P. (2003) *J. Biomol. NMR*, **26**, 93–111.

Fukunaga, K. (1990) *Introduction to Statistical Pattern Recognition*, Academic Press, Boston, MA.

Furui, J., Uegaki, K., Yamazaki, T., Shirakawa, M., Swindells, M.B., Harada, H., Taniguchi, T. and Kyogoku, Y. (1998) *Struct. Fold. Des.*, **6**, 491–500.

Kanelis, V., Forman-Kay, J.D. and Kay, L.E. (2001) *Iubmb Life*, **52**, 291–302.

Kay, L.E., Ikura, M., Tschudin, R. and Bax, A. (1990) *J. Magn. Reson.*, **89**, 496–514.

Lian, L.Y. and Middleton, D.A. (2001) *Prog. Nucl. Magn. Reson. Spectrosc.*, **39**, 171–190.

Moseley, H.N.B., Monleon, D. and Montelione, G.T. (2001) *Meth. Enzymol.*, **339**, 91–108.

Pardi, A., Billeter, M. and Wüthrich, K. (1984) *J. Mol. Biol.*, **180**, 741–751.

Schwalbe, H., Carlomagno, T., Hennig, M., Junker, J., Reif, B., Richter, C. and Griesinger, C. (2001) *Meth. Enzymol.*, **338**, 35–81.

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

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, NY.

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

Stockman, B.J. and Dalvit, C. (2002) *Prog. Nucl. Magn. Reson. Spectrosc.*, **41**, 187–231.

Tian, F., Valafar, H. and Prestegard, J.H. (2001) *J. Am. Chem. Soc.*, **123**, 11791–11796.

Tolman, J.R. and Prestegard, J.H. (1996) *J. Magn. Reson.*, **112**, 245–252.

Vűister, G.W. and Bax, A. (1993) *J. Am. Chem. Soc.*, **115**, 7772–7777.

Wang, Y.X., Marquardt, J.L., Wingfield, P., Stahl, S.J., Huang, S., Torchia, D. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 7385–7386.

Weisemann, R., Ruterjans, H., Schwalbe, H., Schleucher J., Bermel, W. and Griesinger, C. (1994) *J. Biomol. NMR*, **4**, 231–240.

Wishart, D.S. and Case, D.A. (2001) *Meth. Enzymol.*, **338**, 3–34.

Word, J.M. (2000) Ph.D. Thesis, Duke University, Durham, NC.