Journal of Biomolecular NMR, 10 (1997) 329–336
KLUWER/ESCOM

J-Bio NMR 468

Automated ^{1}H and ^{13}C chemical shift prediction using the BioMagResBank

David S. Wishart[a,*], M. Scott Watson[a], Robert F. Boyko[b] and Brian D. Sykes[b]

[a]*Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2N8*
[b]*Department of Biochemistry, University of Alberta, Edmonton, AB, Canada T6G 2H7*

Received 14 April 1997
Accepted 9 June 1997

Keywords: Chemical shift; Homology; Prediction; BioMagResBank

Summary

A computer program has been developed to accurately and automatically predict the ^{1}H and ^{13}C chemical shifts of unassigned proteins on the basis of sequence homology. The program (called SHIFTY) uses standard sequence alignment techniques to compare the sequence of an unassigned protein against the BioMagResBank – a public database containing sequences and NMR chemical shifts of nearly 200 assigned proteins [Seavey et al. (1991) *J. Biomol. NMR*, **1**, 217–236]. From this initial sequence alignment, the program uses a simple set of rules to directly assign or transfer a complete set of ^{1}H or ^{13}C chemical shifts (from the previously assigned homologues) to the unassigned protein. This 'homologous assignment' protocol takes advantage of the simple fact that homologous proteins tend to share both structural similarity and chemical shift similarity. SHIFTY has been extensively tested on more than 25 medium-sized proteins. Under favorable circumstances, this program can predict the ^{1}H or ^{13}C chemical shifts of proteins with an accuracy far exceeding any other method published to date. With the exponential growth in the number of assigned proteins appearing in the literature (now at a rate of more than 150 per year), we believe that SHIFTY may have widespread utility in assigning individual members in families of related proteins, an endeavor that accounts for a growing portion of the protein NMR work being done today.

Introduction

Sequence databases such as the Protein Information Resource (PIR) (George et al., 1996) or SWISS-PROT (Bairoch and Apweiler, 1996) and structure databases such as the Protein Databank (PDB) (Bernstein et al., 1977) are playing an increasingly important role in all aspects of molecular and structural biology. By using powerful search engines (Altschul et al., 1990), alignment tools (Orengo et al., 1992) or threading programs (Bryant, 1996) developed especially for these databases, molecular biologists are often able to predict the structure or function of newly sequenced proteins simply on the basis of sequence homology. Similarly, many X-ray crystallographers also make use of these same databases and software tools to assist with the experimental determination of protein structures through automated homology modeling and molecular replacement (Sali et al., 1990). While X-ray crystallographers have long realized the utility of sequence and structural alignments in their work, NMR spectroscopists, in general, have not. Indeed, despite the establishment of the BioMagResBank (BMRB) in 1990 (Seavey et al., 1991) as a public repository of protein sequence and protein chemical shift data, NMR spectroscopists have yet to make a concerted effort to develop software tools necessary to exploit this valuable resource.

One example where sequential and structural alignment could have a significant impact on biomolecular NMR is in the area of chemical shift assignment. Most resonance assignment schemes depend on the spectroscopist having some knowledge of approximate chemical shifts or expec-

*To whom correspondence should be addressed.

Abbreviations: BMRB, BioMagResBank; EGF, epidermal growth factor; HPr, histidine-containing protein; PDB, Protein Databank; PIR, Protein Information Resource; PTI, pancreatic trypsin inhibitor; rms, root mean square; TGF, transforming growth factor.

ted chemical shift ranges for the residues under question. Clearly, if it were possible to accurately predict the ^{1}H, ^{13}C and ^{15}N NMR chemical shifts of a given protein prior to assignment, it would make the sequential assignment process substantially easier, significantly faster and much less dependent on NOE or scalar connectivity information.

In the early days of protein NMR, several investigators proposed using X-ray structures of the protein of interest and semiempirical chemical shift theories to directly calculate protein chemical shifts. This was done with the hope that it might facilitate the assignment process (Perkins and Wüthrich, 1979; Perkins and Dwek, 1980). However, limitations in the theory of chemical shifts led to only modest success. With recent advances in semiempirical and quantum mechanical theories of chemical shift calculation (Ösapay and Case, 1991,1994; Herranz et al., 1992; Williamson et al., 1992; de Dios et al., 1993), it is now possible, from crystal structures, to predict α-^{1}H chemical shifts with correlation coefficients of between 0.74 and 0.84 and NH shifts with correlation coefficients of between 0.57 and 0.71. However, most NMR spectroscopists have found that this level of accuracy is not yet sufficient to assist with the assignment processes. Furthermore, not all proteins of interest have readily available high-resolution X-ray structures.

An alternative and potentially far more accurate approach to chemical shift prediction is to use the fact that homologous proteins often have not only similar structures but similar chemical shifts. In other words, if one can identify a homologous protein which has already been assigned, it should be possible to use those same assignments (with suitable corrections) to predict the chemical shifts of an unassigned homologue. We call this concept 'homologous assignment' in analogy to the more familiar concept of homology modeling. This relatively simple concept has recently been described (Redfield and Robertson, 1991; Bartels et al., 1996; Gronwald et al., 1997) and put into limited practice. However, it has only been applied to situations where the homologue or homologues had been previously identified through a manual comparison or through intensive literature searches. With more than 550 different peptides and proteins already assigned by NMR and with this number expected to nearly double by the year 2000, it is increasingly unlikely that the average NMR spectroscopist will be able to rely on his/her intuition or to have the time to spend dozens of hours scanning through the literature to determine if other homologues to his/her protein of interest have already been assigned. Rather, we expect that NMR spectroscopists will eventually have to turn to the BMRB and the appropriate software tools to quickly and automatically get these answers.

In anticipation of this need, we decided to combine the searching and alignment software that has made the PDB and PIR databases so useful for X-ray crystallographers with the sequence and chemical shift information contained in the BMRB. To this end, we have developed a computer program (called SHIFTY) which automatically selects, aligns and assigns ^{1}H and ^{13}C chemical shifts of unassigned proteins using a slightly modified version of BMRB (containing nearly 200 different proteins) and a table of experimentally derived random coil chemical shifts (Wishart et al., 1995a). In this communication we describe, in detail, how SHIFTY performs the alignments and predicts the chemical shifts. We also assess the results from tests performed on nearly 30 different proteins, each of which has at least one homologue in our chemical shift database. In addition, we compare the accuracy of these 'homologous assignments' to predictions obtained by direct calculation from the corresponding X-ray crystal structures. On the basis of these and other results, we discuss the potential applications and limitations of SHIFTY and the concept of homologous assignment.

Materials and Methods

The databases

Individual data files from the BMRB (Seavey et al., 1991) were manually scanned, selected and downloaded from the BMRB server (http://www.bmrb.wisc.edu). Only those peptides and proteins containing reasonably complete ^{1}H assignments collected in aqueous conditions between pH 2.0 and 7.5 and at temperatures between 5 and 60 °C were included. A total of 147 distinct polypeptide chains were identified. In assembling the ^{1}H database, the BMRB flatfile format was converted into a more compact multicolumn format with individual columns containing (i) the residue number; (ii) the one-letter amino acid code; (iii) the secondary structure; and the chemical shifts in the following order: (iv) NH; (v) αH; (vi) βH1; (vii) βH2; (viii) γH1; (ix) γH2; (x) δH1; (xi) δH2; (xii) εH1; (xiii) εH2. Nondegenerate pairs of α, β, γ, δ and ε chemical shifts were ordered such that the largest value was always placed in the leftmost column. Aromatic resonances were excluded to simplify the presentation and to expedite the chemical shift predictions. This ^{1}H chemical shift database was supplemented with an additional 28 proteins from a previously prepared collection (Wishart et al., 1991), giving a grand total of 175 separate polypeptides. Chemical shifts in this ^{1}H database are variously referenced to TSP and DSS, but because ^{1}H chemical shift differences between these standards are so small (Wishart et al., 1995b), no further corrections were made. On the other hand, ambiguities in referencing the ^{13}C chemical shifts made many of the BMRB ^{13}C assignments unusable. Consequently, the ^{13}C database used in this paper was assembled from data originally used to develop the ^{13}C chemical shift index (Wishart and Sykes, 1994). This updated database contains ^{13}C α, β and carbonyl chemical shifts (as well as αH shifts) from 18 different

proteins all referenced to DSS. Together, the ^{13}C and ^{1}H databases in this modified version of the BMRB contain a total of 193 polypeptide chains representing 11 062 residues and nearly 56 000 chemical shift assignments. Both databases are used by the program to identify and align the query (or unassigned) protein to an assigned homologue or set of homologues. The program also uses a collection of experimentally derived random coil ^{1}H and ^{13}C NMR shifts obtained by measuring the chemical shifts of disordered hexapeptides (Wishart et al., 1995a) in 1 M urea solutions. This 'random coil' database is used to predict chemical shifts of unmatched or nonidentical residues as described below.

The algorithms

SHIFTY makes use of two algorithms. One algorithm is used for sequence comparison and alignment and the other is used for sequential assignment or chemical shift prediction. Sequential alignment and comparison are performed using the dynamic programming method of Needleman and Wunsch (1970) as implemented in the program NWALIGN (Wishart et al., 1994). In this algorithm, an input sequence (typically belonging to the unassigned protein) is systematically compared to each of the ~200 protein sequences in the chemical shift database using an amino acid scoring matrix similar to the Dayhoff PAM_{250} matrix (Dayhoff et al., 1983; Wishart et al., 1994). Gap insertion and extension penalties, in combination with the amino acid similarity scores and secondary structure information, are used to determine the alignment or parsing of the two polypeptide chains as well as the overall sequence similarity score. Upon completion of the alignment and scoring process (which typically takes a few seconds), the highest scoring sequences are selected and each of the pairwise alignments is passed on to the second algorithm.

In this second stage, chemical shifts from the sequence selected from the database are assigned to the query sequence. This assignment may be done in one of three ways:

(1) In the situation where aligned residues match exactly, the chemical shifts are directly transferred from the database protein to the query protein with no numerical adjustment. Hence, if an alanine in the query protein is aligned with an alanine in the database homologue, the complete set of alanine shifts from that database residue is written to the query residue.

```
**********(2)**********
Title....: PTI TYPE E
Id.......: 64
Score....: 377
Matches..: 25

Query Seq:   RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSA      48
Matching.:   *  ||  |   ||||| |  **||  |  || | ||||* * | |
Database.: LQHRTFCKLPAEPGPCKASIPAFYYNWAAKKCQLFHYGGCKGNANRFSTI      50
Structure: CCCHHHHHCCCCCCCBBBBBBBBBBCCCCCBBBBBBBBCCCCCCBBBBBC

Query Seq: EDCMRTCGGA                                           58
Matching.: | |   | |
Database.: EKCRHACVG                                           59
Structure: HHHHHHHHH

   I = input seq,    D = database seq, S = secondary structure
  NH = amide shift, AH = alpha proton shift
 BH1 = beta proton shift (alpha proton for GLY)
 BH2 = beta proton shift, OTH = gamma, delta, epsilon, etc

Num I D S    NH    AH   BH1   BH2   OTH
  1   L C
  2   Q C
  3 R*H C  8.34  4.29  1.84  1.75  ****  ****  ****  ****  ****  ****
  4 P R H  9.14  4.32  1.93  1.27  ****  ****  ****  ****  ****  ****
  5 D T H  8.32  4.06  2.67  2.63  ****  ****  ****  ****  ****  ****
  6 F=F H  7.16  4.49  3.13  3.13  ****  ****  ****  ****  ****  ****
  7 C=C H  7.35  4.46  3.24  2.96  ****  ****  ****  ****  ****  ****
  8 L K H  6.93  4.26  1.18  1.24  ****  ****  ****  ****  ****  ****
  9 E L C  7.41  4.46  2.63  2.40  2.73  2.71  ****  ****  ****  ****
 10 P=P C  ****  ****  ****  ****  ****  ****  ****  ****  ****  ****
 11 P A C  7.98  3.58  0.92  ****  ****  ****  ****  ****  ****  ****
 12 Y E C  7.45  5.17  2.84  2.63  ****  ****  ****  ****  ****  ****
 13 T P C  ****  4.69  5.37  1.43  ****  ****  ****  ****  ****  ****
 14 G=G C  8.62  4.22  4.05  ****  ****  ****  ****  ****  ****  ****
 15 P=P C  ****  ****  ****  ****  ****  ****  ****  ****  ****  ****
 16 C=C B  9.12  4.54  3.40  2.84  ****  ****  ****  ****  ****  ****
 17 K=K B  7.87  4.46  1.58  1.58  ****  ****  ****  ****  ****  ****
 18 A=A B  8.04  4.34  1.20  ****  ****  ****  ****  ****  ****  ****
 19 R S B  8.02  4.13  1.75  1.55  ****  ****  ****  ****  ****  ****
 20 I=I B  8.79  4.45  1.87  1.45  1.06  0.94  0.72  ****  ****  ****
 21 I P B  ****  3.92  1.63  1.38  1.29  0.87  0.95  ****  ****  ****
 22 R A B  8.24  4.42  1.08  ****  ****  ****  ****  ****  ****  ****
 23 Y*F B  8.90  5.68  2.76  2.67  ****  ****  ****  ****  ****  ****
 24 F*Y B  9.94  5.27  2.96  2.87  ****  ****  ****  ****  ****  ****
```

Fig. 1. Sample output from a SHIFTY run using bovine pancreatic trypsin inhibitor (BPTI) as the query sequence. This particular example shows the alignment and predicted ^{1}H chemical shifts derived from the homologous PTI type E protein.

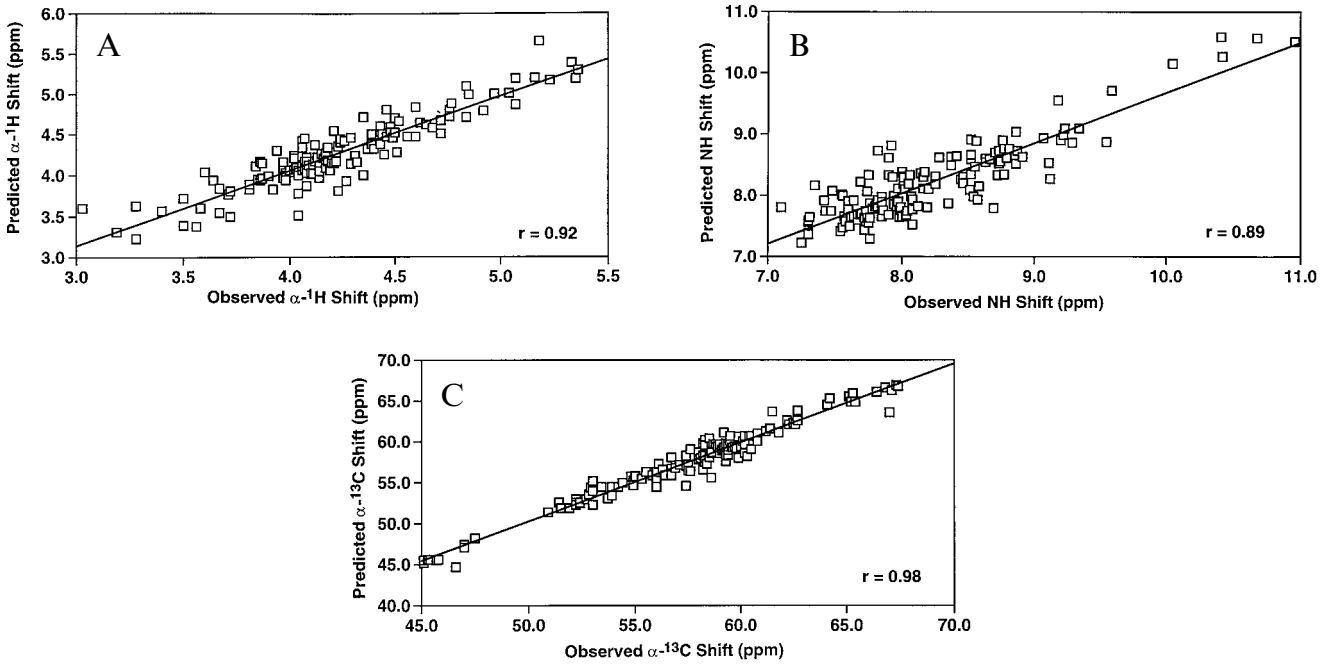

Fig. 2. Comparison of the predicted (A) α-^{1}H, (B) NH and (C) α-^{13}C chemical shifts with the observed chemical shifts of troponin C (turkey) based on the values derived from a SHIFTY alignment with calmodulin (Drosophila). The correlation coefficient (r) for each of the graphs is given in the lower right corner.

(2) In the situation where two aligned residues differ, the database protein's shifts are subtracted from the database residue's random coil shifts (Wishart et al., 1995a) and these differences are added to the random coil shifts corresponding to the query residue. Hence, if an alanine in the query protein is aligned with a proline in the database homologue, the proline chemical shifts are subtracted from random coil proline shifts and these differences are added to the random coil alanine shifts. In this particular example, only the alpha and beta ^{1}H shifts of proline could be used as predictors of the alanine shifts (proline does not have an amide ^{1}H shift).

(3) In the situation where a residue from the query protein lines up with a gap (i.e. a null residue) in the database protein, or vice versa, no chemical shift prediction is made.

It should be noted that when non-glycine α-^{1}H chemical shifts are being predicted from glycine residues, the average α-^{1}H chemical shift of the two glycine protons is used.

The program

SHIFTY is written in ANSI standard C and has been compiled, tested and run on both SUN and SGI UNIX workstations as well as PCs operating with LINUX. Using the program's parameter file, users can select the database (their own or the database packaged with the program), the gap insertion and gap extension penalties as well as the scoring matrix to be used in the alignment process. SHIFTY runs as a simple text-based program and requires only the query sequence as input. A sample of the SHIFTY output is shown in Fig. 1. This output file includes the sequences of both the query and the database protein as well as indicators of the sequence similarity (equal signs for identical residues, stars for similar residues and blanks for dissimilar residues) in the leftmost columns. The predicted chemical shifts are shown throughout the remaining columns. The order of the ^{1}H chemical shifts is (i) NH; (ii) αH; (iii) βH1; (iv) βH2; (v) γH1; (vi) γH2; (vii) δH1; (viii) δH2; (ix) εH1; and (x) εH2. The order of the ^{13}C chemical shifts is (i) αH; (ii); αC; (iii) βC; and (iv) CO. Typically, the amount of time required to perform both the alignments and the chemical shift predictions is less than 10 s on a SUN Sparcstation 5. The program is available from the authors on request.

Results and Discussion

In assessing SHIFTY we used the program to predict a complete set of ^{1}H shifts for 25 different proteins, each of which had at least one homologue in our chemical shift database. Three assessment criteria were used: (i) the overall accuracy of the predicted shifts versus the observed chemical shifts; (ii) the accuracy of the predicted shifts versus those predicted from high-resolution X-ray structures using the methods of Ösapay and Case (1991,1994) or Williamson et al. (1992); and (iii) the variation of the prediction accuracy with percent sequence identity.

Figure 2 illustrates one example of the generally high quality of predictions that can be obtained with SHIFTY.

In this case, the predicted HN, αH and α-^{13}C chemical shifts of turkey troponin C (the query protein) are plotted against the experimentally observed chemical shifts as determined by Slupsky et al. (1995). For this example, Drosophila calmodulin (Ikura et al., 1990), which had the highest level of sequence identity to troponin C (46.2%), was used as the homologous predictor protein. As can be seen by these three graphs, the agreement between predicted and observed chemical shifts is particularly strong for the αH and α-^{13}C shifts, with correlation coefficients of 0.92 and 0.98, respectively. As might be expected, the correlation is not as high for the amide chemical shifts (r = 0.89). The average (i.e. rms) error between the observed and expected αH chemical shifts is 0.18 ppm, whereas for the NH and α-^{13}C chemical shifts it is 0.32 and 0.93 ppm, respectively.

In order to compare these results to those that might have been obtained if one used the X-ray crystal structure of troponin C to predict the chemical shifts, we used the programs of Ösapay and Case (1991) and Williamson et al. (1992) to calculate ^{1}H shifts using the PDB coordinate file 5TNC (Herzberg and James, 1988). The correlations between the predicted shifts and observed shifts for the method of Ösapay and Case were determined to be 0.80 and 0.17 (for the αH and NH shifts, respectively), while for the method of Williamson et al. they were found to be 0.81 and 0.36, respectively. Despite the disadvantage of having to use only a distantly related protein (Drosophila calmodulin) for its predictions, it is clear that SHIFTY outperforms the other two methods, both of which had the distinct advantage of working with the high-resolution crystal structure of troponin C. A more appropriate comparison might have been to use the X-ray crystal structure of Drosophila calmodulin (PDB 4CLN, Taylor et al., 1991) as the template instead of troponin C. When this was done, the performance of the two coordinate-based methods, relative to SHIFTY, dropped a further 20%.

Additional comparisons between αH and NH chemical shift predictions derived from the coordinate-based methods of Ösapay and Case and Williamson et al. versus those derived from SHIFTY are shown in Table 1. For every case in which direct comparisons were possible, it is clear that SHIFTY outperformed the other two coordinate-based techniques. On average, the correlation coefficients for SHIFTY were 20–40% better for αH chemical shifts, while for NH chemical shifts they were up to 300% better. Particularly striking is SHIFTY's substantive improvement in NH chemical shift prediction accuracy relative to the other two techniques. In determining the correlation coefficients for SHIFTY, it is important to note that only those residues predicted by the program were included in these calculations. Unpredicted gaps (total = 4) and unmatched termini (total = 3) meant that 32 out of a total of 673 residues (less than 5%) did not have their αH or NH chemical shifts predicted by SHIFTY. However, excluding these 32 residues from the predictions of Ösapay and Case and Williamson et al. did not change their correlation coefficients in any measurable way.

Because SHIFTY makes its predictions on the basis of sequence and chemical shift homology, perhaps the most important question to address is how its performance varies with the sequence identity of the homologues used in the prediction process. As one might expect, the presence of a database homologue which has 99% or 100% sequence identity with the query protein will invariably allow SHIFTY to predict the query protein's shifts with 99% or 100% accuracy. Obviously as the sequence similarity falls, the quality of the predictions should fall too. In Fig. 3 we plot the quality of the αH and NH chemical shift predictions for each of the query proteins versus the percent sequence identity of the database homologues that were used to predict their chemical shifts. These data are summarized in more detail in Table 2, where we have included not only the αH and NH data but also the βH results. Insufficient data were available to plot the same curve for ^{13}C chemical shifts, but indications are that this curve would closely follow the αH distribution. From these scatter diagrams, and the superimposed hyperbolic curves, it is clear that the quality of the chemical shift

TABLE 1
COMPARISON BETWEEN THREE DIFFERENT METHODS OF CHEMICAL SHIFT PREDICTION

Protein	PDB	Resolution (Å)	Ösapay and Case		Williamson et al.		SHIFTY	
			αH	NH	αH	NH	αH	NH
Lysozyme (hen)	193L	1.33	0.85	0.45	0.88	0.57	0.99	0.99
Calbindin (bovine)	3ICB	2.30	0.73	0.40	0.85	0.61	0.99	0.99
HPr (*B. subtilis*)	1SPH	2.00	0.51	0.04	0.88	0.60	0.93	0.87
Troponin C (turkey)	5TNC	2.00	0.80	0.17	0.81	0.36	0.92	0.89
PTI (bovine)	6PTI	1.70	0.76	0.24	0.89	0.73	0.93	0.85
Bungarotoxin	2ABX	2.50	0.14	0.19	0.26	0.20	0.84	0.72
HPr (*E. coli*)	1POH	2.00	0.57	0.42	0.55	0.47	0.64	0.51
Average	–	–	0.62	0.27	0.73	0.51	0.89	0.83

The correlation coefficients for the methods of Ösapay and Case (1991) and Williamson et al. (1992) are based on the given crystal structure coordinates.

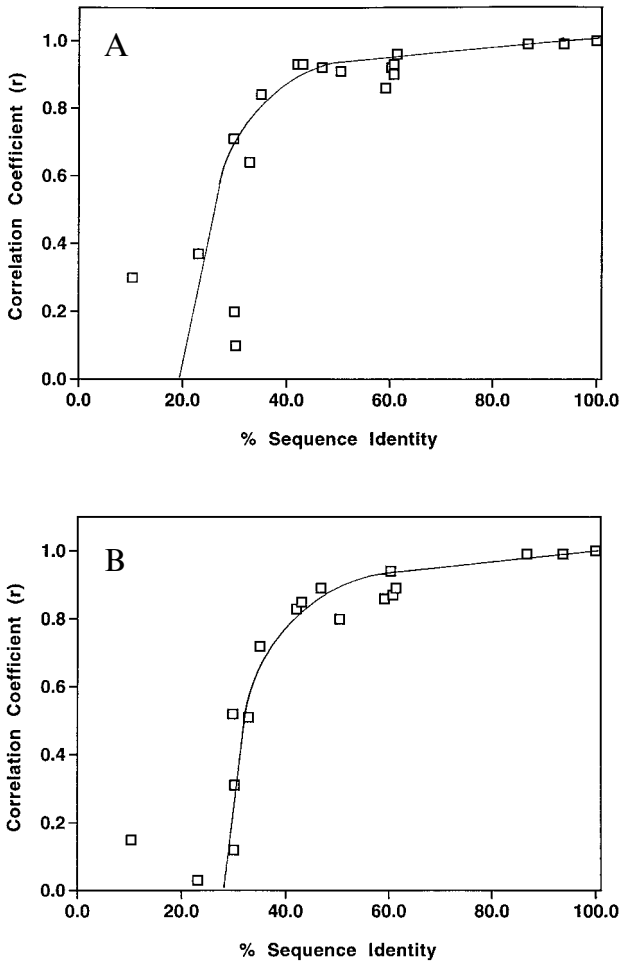

Fig. 3. Graphs illustrating how the correlation coefficients between predicted and observed (A) α-^{1}H and (B) NH chemical shifts vary with the percent sequence identity of the homologues used in the prediction process. The curves shown in both graphs are approximate and are only intended to facilitate visual comparisons.

predictions falls off rapidly once the pairwise sequence identity falls below 35%. Through extensive curve fitting, we have found that the relationship between the correlation coefficient (r) and the percent sequence identity (% ID) can be expressed as follows:

$$r_{\alpha H} = 1.16 - 16/(\%\ ID) \qquad \text{for } \%\ ID > 15 \qquad (1)$$

$$r_{NH} = 1.20 - 20/(\%\ ID) \qquad \text{for } \%\ ID > 15 \qquad (2)$$

$$r_{\beta H} = 1.03 - 3/(\%\ ID) \qquad \text{for } \%\ ID > 15 \qquad (3)$$

From Eqs. 1–3 we can conclude, at least in situations where there is more than one database homologue, that one only needs to use the predicted chemical shifts derived from the most similar homologue to obtain the best chemical shift estimates. This result is similar to the conclusions reached by Gronwald et al. (1997) using their more elaborate multiple alignment approach. It is also apparent that below the 35% level of sequence identity, SHIFTY will perform no better (and often worse) than the methods of Ösapay and Case (1991,1994) or Williamson et al. (1992) which calculate chemical shifts directly from crystal structure data.

As with any predictive process, it is important to be able to provide a quantitative estimate of the error associated with any particular prediction. We have found that four very simple equations can be used to predict the overall rms error (in ppm) for predicted αH, NH, βH and ^{13}C chemical shifts. All four equations are based on the % ID between the query protein and the database homologue. These formulae are

$$\text{rmsd (ppm)} = 0.4 - 0.004 \times (\%\ ID) \qquad \text{for } \alpha\text{-}^{1}H \qquad (4)$$

$$\text{rmsd (ppm)} = 0.8 - 0.008 \times (\%\ ID) \qquad \text{for NH} \qquad (5)$$

$$\text{rmsd (ppm)} = 0.6 - 0.006 \times (\%\ ID) \qquad \text{for } \beta H \qquad (6)$$

$$\text{rmsd (ppm)} = 2.0 - 0.02 \times (\%\ ID) \qquad \text{for } ^{13}C \qquad (7)$$

In assessing the accuracy and utility of this homologous assignment method, we also investigated whether these predicted shifts could be used directly in the assignment process. Using an approach originally developed by Redfield and Robertson (1991) based on 'minimal chemical shift distance', we wrote a simple computer program which compared an observed set of chemical shifts (from the query protein) with a predicted set (from the matching protein). Each set of predicted chemical shifts, belonging to a single spin system, was compared to each member of the observed set. The predicted/observed pair of spin systems having the smallest absolute chemical shift difference (as determined by a weighted sum of the αH, NH and side-chain proton chemical shift differences) was removed and the 'observed' spin system was assigned to that particular 'predicted' spin system. This process was repeated for each spin system in the predicted set until no more spin systems were left. In this way, the observed set of chemical shifts could be sequentially assigned purely on the basis of their similarity to the predicted chemical shifts derived from SHIFTY. While the data from this simulated sequential assignment experiment were somewhat idealized and the program was not fully optimized, it is nevertheless quite instructive to see how well this simple-minded assignment scheme worked. The results from this particular test are summarized in Table 3.

As might be expected, those pairs of proteins sharing the highest level of sequence identity (>85%) generally permitted a near-perfect assignment (>92% correct). Interestingly, when the sequence identity dropped below 85%, our assignment scheme was still able to correctly assign more than 70% (on average) of all resonances even when

TABLE 2
RELATIONSHIP BETWEEN SEQUENCE SIMILARITY AND ^{1}H CHEMICAL SHIFTS PREDICTED THROUGH HOMOLOGOUS ASSIGNMENT

Query protein	Matching protein	% identity	αH correlation	NH correlation	βH correlation
Lysozyme (hen)	Lysozyme (hen)	100	1.00	1.00	1.00
Lysozyme (hen)	Lysozyme (turkey)	93.7	0.99	0.99	0.98
Calbindin (bovine)	Calbindin (porcine)	86.8	0.99	0.99	0.99
PTI (type E)	PTI (type K)	61.4	0.96	0.89	0.98
HPr (*B. subtilis*)	HPr (*S. aureus*)	60.9	0.93	0.87	0.96
Lysozyme (hen)	Lysozyme (human)	60.8	0.90	0.87	0.96
EGF (mouse)	EGF (human)	60.4	0.92	0.94	0.95
Anthopleurin A	ATX I toxin	59.2	0.86	0.86	0.97
Plastocyanin (spinach)	Plastocyanin (algae)	50.5	0.91	0.80	0.94
Troponin C (turkey)	Calmodulin (fruit fly)	46.9	0.92	0.89	N/A
BPTI (bovine)	PTI (type E)	43.1	0.93	0.85	0.94
BPTI (bovine)	PTI (type K)	42.1	0.93	0.83	0.93
Bungarotoxin	Alpha neurotoxin	35.1	0.84	0.72	0.93
HPr (*E. coli*)	HPr (*B. subtilis*)	32.9	0.64	0.51	0.91
EGF (mouse)	TGF (human)	30.2	0.10	0.31	0.94
Cardiotoxin III	Alpha neurotoxin	30.0	0.20	0.12	0.89
Bungarotoxin	Cardiotoxin III	29.8	0.71	0.52	0.92
EGF (mouse)	Hirudin	23.1	0.37	0.03	0.86
Lac repressor	HPr (*B. subtilis*)	10.4	0.30	0.15	N/A

Percent sequence identity was determined by taking the number of matches and dividing by the number of residues in the longest of the two sequences. The values in the βH column are derived from the mean correlation coefficient of both βH1 and βH2.

the sequence identity dropped to as low as 35%. Furthermore, if the second and third choices were considered in the assignment protocol, the number of correct assignments could often climb above 80%. Table 3 also shows that an essentially random prediction is only capable of getting about 35% of the assignments correct (see EGF versus hirudin). For longer sequences with less complete assignments, this number is expected to fall well below 30%. We have no doubt that further improvements to the algorithm are possible and we believe that the results shown in Table 3 nicely illustrate the potential that high-quality chemical shift prediction methods could have in assisting NMR spectroscopists with the assignment process.

Despite the limitations imposed by both the size of the database (193 proteins in this case) and the need for moderately high sequence (35%) homology, we estimate that SHIFTY, in its present form, could assist with the assignment of approximately 30% of all new proteins. This estimate is based on the fact that approximately 55 of the 193 proteins in our current database share greater than 35% sequence homology with at least one other protein in the database. As the number of proteins in the database grows and as many NMR assignment efforts become more targeted, there is a good possibility that this proportion could grow to more than 40% or 50% of all new peptides and proteins.

In addition to SHIFTY's increased accuracy, there are a number of advantages that this program has over competing methods of chemical shift prediction. These include the fact that SHIFTY is intuitively simple and quick, it requires only the sequence of the protein of interest for input, it does not require high-resolution X-ray coordinates, it can predict ^{1}H chemical shifts as well as ^{13}C shifts with high accuracy and, most importantly, it will always get better with time. This latter point cannot be emphasized enough. With nearly 200 peptide and protein assignments already deposited in our database and the expectation that this number could grow to more than

TABLE 3
ACCURACY OF RESIDUE-SPECIFIC CHEMICAL SHIFT ASSIGNMENTS BASED ON CHEMICAL SHIFT PREDICTIONS FROM SHIFTY AND MINIMAL CHEMICAL SHIFT DISTANCE FROM OBSERVED (EXPERIMENTAL) CHEMICAL SHIFTS

Query protein	Matching protein	% identity	% correct
Lysozyme (hen)	Lysozyme (hen)	100	100
Lysozyme (hen)	Lysozyme (turkey)	93.7	92.2
Calbindin (bovine)	Calbindin (porcine)	86.8	93.4
PTI (type E)	PTI (type K)	61.4	67.8
HPr (*B. subtilis*)	HPr (*S. aureus*)	60.9	66.7
Lysozyme (hen)	Lysozyme (human)	60.8	67.4
EGF (mouse)	EGF (human)	60.4	75.5
Anthopleurin A	ATX I toxin	59.2	75.5
Plastocyanin (spinach)	Plastocyanin (algae)	50.5	64.6
BPTI (bovine)	PTI (type E)	43.1	79.3
BPTI (bovine)	PTI (type K)	42.1	74.1
Bungarotoxin	Alpha neurotoxin	35.1	70.3
HPr (*E. coli*)	HPr (*B. subtilis*)	32.9	77.4
EGF (mouse)	Hirudin	23.1	35.8
Lac repressor	HPr (*B. subtilis*)	10.4	39.2

Percent sequence identity was determined by taking the number of matches and dividing by the number of residues in the longest of the two sequences.

1000 by the year 2000, the odds that any new protein that needs to be assigned will have a close homologue already assigned should grow accordingly.

This exponential growth in NMR chemical shift assignments points to another potential advantage for SHIFTY. In particular, as the BMRB database grows in size, most NMR spectroscopists will not be able to keep up with the flood of hundreds of newly deposited protein assignments nor will they be able to readily identify which proteins might be homologous to their protein of interest. Having an automated method, such as SHIFTY, available to compare, align and predict chemical shift assignments should reduce the many hours of overhead necessary to conduct intensive literature searches or to manually scan, align and compare presumptive homologues. The fact that molecular biologists and X-ray crystallographers have been using computer programs like SHIFTY for more than a decade only underscores the fact that when a data-intensive field like biomolecular NMR matures, there eventually becomes a need to develop highly specialized computer tools to facilitate the handling, searching and interpreting of those data. We believe that biomolecular NMR has reached that point.

Conclusions

In this communication we have described a very simple concept which allows precise prediction of the ^{1}H chemical shifts of proteins and peptides. This concept, which is called homologous assignment, is based on the fact that homologous proteins tend to share very similar chemical shifts. We have implemented and tested this concept using a computer program called SHIFTY. This program combines sequence comparison and alignment algorithms with a simple chemical shift assignment algorithm. We have shown that SHIFTY can be confidently applied to the prediction of ^{1}H and ^{13}C chemical shifts whenever the query sequence shares >35% sequence identity with a previously assigned homologue. Under these circumstances, we have found that the ^{1}H chemical shift predictions are more accurate than any other method published to date. These promising results suggest that this program could have a broad range of applications extending from the assignment of individual proteins from a common family or a common fold (e.g. calcium-binding proteins, protease inhibitors, SH2 and SH3 domains, etc.) to the assignment of families of mutant proteins, and even to the assignment of individual proteins containing different ligands. It is anticipated that the utility of SHIFTY, or other programs which make use of the homologous assignment concept, will increase greatly over time. This is because homologous assignment is the only chemical shift prediction technique that makes use of the rapidly growing database of previously assigned proteins – a database which is now growing exponentially.

Acknowledgements

Financial support by Bristol-Myers Squibb (Canada), the MRC Group in Protein Structure and Function (Canada), the Protein Engineering Network of Centres of Excellence (Canada) and the Alberta Heritage Fund for Medical Research is gratefully acknowledged.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.

Bairoch, A. and Apweiler, R. (1996) *Nucleic Acids Res.*, **24**, 21–25.

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

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.V., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M.J. (1977) *J. Mol. Biol.*, **112**, 535–542.

Bryant, S.H. (1996) *Proteins Struct. Funct. Genet.*, **26**, 172–185.

Dayhoff, M.O., Barker, W.C. and Hunt, L.T. (1983) *Methods Enzymol.*, **91**, 524–545.

De Dios, A.C., Pearson, J.G. and Oldfield, E. (1993) *Science*, **260**, 1491–1495.

George, D.G., Hunt, L.R. and Barker, W.C. (1996) *Methods Enzymol.*, **266**, 41–59.

Gronwald, W., Boyko, R.F., Sönnichsen, F.D., Wishart, D.S. and Sykes, B.D. (1997) *J. Biomol. NMR*, **10**, 165–179.

Herranz, J., Gonzalez, C., Rico, M., Nieto, J.L., Santoro, J., Jimenez, M.A., Bruix, M., Neira, J.L. and Blanco, F.J. (1992) *Magn. Reson. Chem.*, **30**, 1012–1018.

Herzberg, O. and James, M.N.G. (1988) *J. Mol. Biol.*, **203**, 761–771.

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

Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.

Orengo, C.A., Brown, N.P. and Taylor, W.R. (1992) *Proteins Struct. Funct. Genet.*, **14**, 139–167.

Ösapay, K. and Case, D.A. (1991) *J. Am. Chem. Soc.*, **113**, 9436–9444.

Ösapay, K. and Case, D.A. (1994) *J. Biomol. NMR*, **4**, 215–230.

Perkins, S.J. and Wüthrich, K. (1979) *Biochim. Biophys. Acta*, **576**, 409–422.

Perkins, S.J. and Dwek, R.A. (1980) *Biochemistry*, **19**, 245–255.

Redfield, C. and Robertson, J.P. (1991) In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy* (Eds., Hoch, J.C., Poulsen, F.M. and Redfield, C.), Plenum, New York, NY, U.S.A., pp. 303–316.

Sali, A., Overington, J.P., Johnson, M.S. and Blundell, R.L. (1990) *Trends Biochem. Sci.*, **15**, 235–240.

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

Slupsky, C.M., Reinach, F.C., Smillie, L.B. and Sykes, B.D. (1995) *Protein Sci.*, **4**, 1279–1290.

Taylor, D.A., Sack, J.S., Maune, J.F. and Beckingham, K. (1991) *J. Biol. Chem.*, **266**, 21375–21384.

Williamson, M.P., Asakura, T., Nakamura, E. and Demura, M. (1992) *J. Biomol. NMR*, **2**, 83–98.

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

Wishart, D.S., Boyko, R.F. and Sykes, B.D. (1994) *Comput. Appl. Biosci.*, **10**, 121–132.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.

Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995a) *J. Biomol. NMR*, **5**, 1–22.

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