Journal of Biomolecular NMR, 10 (1997) 165–179
KLUWER/ESCOM

J-Bio NMR 453

ORB, a homology-based program for the prediction of protein NMR chemical shifts

Wolfram Gronwald[a,b], Robert F. Boyko[b], Frank D. Sönnichsen[c], David S. Wishart[a,d] and Brian D. Sykes[a,b,*]

[a]*Protein Engineering Network of Centres of Excellence, 713 Heritage Medical Research Centre, University of Alberta, Edmonton, AB, Canada T6G 2S2*
[b]*Department of Biochemistry, University of Alberta, Edmonton, AB, Canada T6G 2H7*
[c]*Department of Physiology and Biophysics, Case Western Reserve University, Cleveland, OH 44106-4970, U.S.A.*
[d]*Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB, Canada T6G 2N8*

Received 31 January 1997
Accepted 1 April 1997

Keywords: Automated sequential assignment; Homology-based protein NMR chemical shift prediction; Protein NMR chemical shift databases; Multiple alignment

Summary

A computer program (ORB) has been developed to predict ^{1}H, ^{13}C and ^{15}N NMR chemical shifts of previously unassigned proteins. The program makes use of the information contained in a chemical shift database of previously assigned proteins supplemented by a statistically derived averaged chemical shift database in which the shifts are categorized according to their residue, atom and secondary structure type [Wishart et al. (1991) *J. Mol. Biol.*, **222**, 311–333]. The prediction process starts with a multiple alignment of all previously assigned proteins with the unassigned query protein. ORB uses the sequence and secondary structure alignment program XALIGN for this task [Wishart et al. (1994) *CABIOS*, **10**, 121–132; 687–688]. The prediction algorithm in ORB is based on a scoring of the known shifts for each sequence. The scores depend on global sequence similarity, local sequence similarity, structural similarity and residue similarity and determine how much weight one particular shift is given in the prediction process. In situations where no applicable previously assigned chemical shifts are available, the shifts derived from the averaged database are used. In addition to supplying the user with predicted chemical shifts, ORB calculates a confidence value for every prediction. These confidence values enable the user to judge which predictions are the most accurate and they are particularly useful when ORB is incorporated into a complete autoassignment package. The usefulness of ORB was tested on three medium-sized proteins: an interleukin-8 analog, a troponin C synthetic peptide heterodimer and cardiac troponin C. Excellent results are obtained if ORB is able to use the chemical shifts of at least one highly homologous sequence. ORB performs well as long as the sequence identity between proteins with known chemical shifts and the new sequence is not less than 30%.

Introduction

NMR spectroscopy is now used routinely for the determination of protein solution structures. The process of generating a new NMR structure is, however, quite time-consuming, with the sequential assignment of the NMR spectra being the time-limiting step. For small or medium-sized proteins and peptides, the sequential assignment process relies primarily on conventional two-dimensional homonuclear methods (Wüthrich, 1986). For larger molecules, several strategies based on various heteronuclear experiments have been proposed for sequence specific

[*]To whom correspondence should be addressed at: Protein Engineering Network of Centres of Excellence, 713 Heritage Medical Research Centre, University of Alberta, Edmonton, AB, Canada T6G 2S2.

Abbreviations: CalB, calcineurin B; CaM, calmodulin; cNTnC, N-domain cardiac muscle troponin C; CTnC, C-domain chicken cloned TnC; IL-8, interleukin-8; NTnC0c, N-domain chicken cloned TnC with no Ca^{2+} bound; NTnC2c, N-domain chicken cloned TnC with two Ca^{2+} bound; ppm, parts per million; TF, troponin C heterodimer; TnC, troponin C; TT, troponin C homodimer.

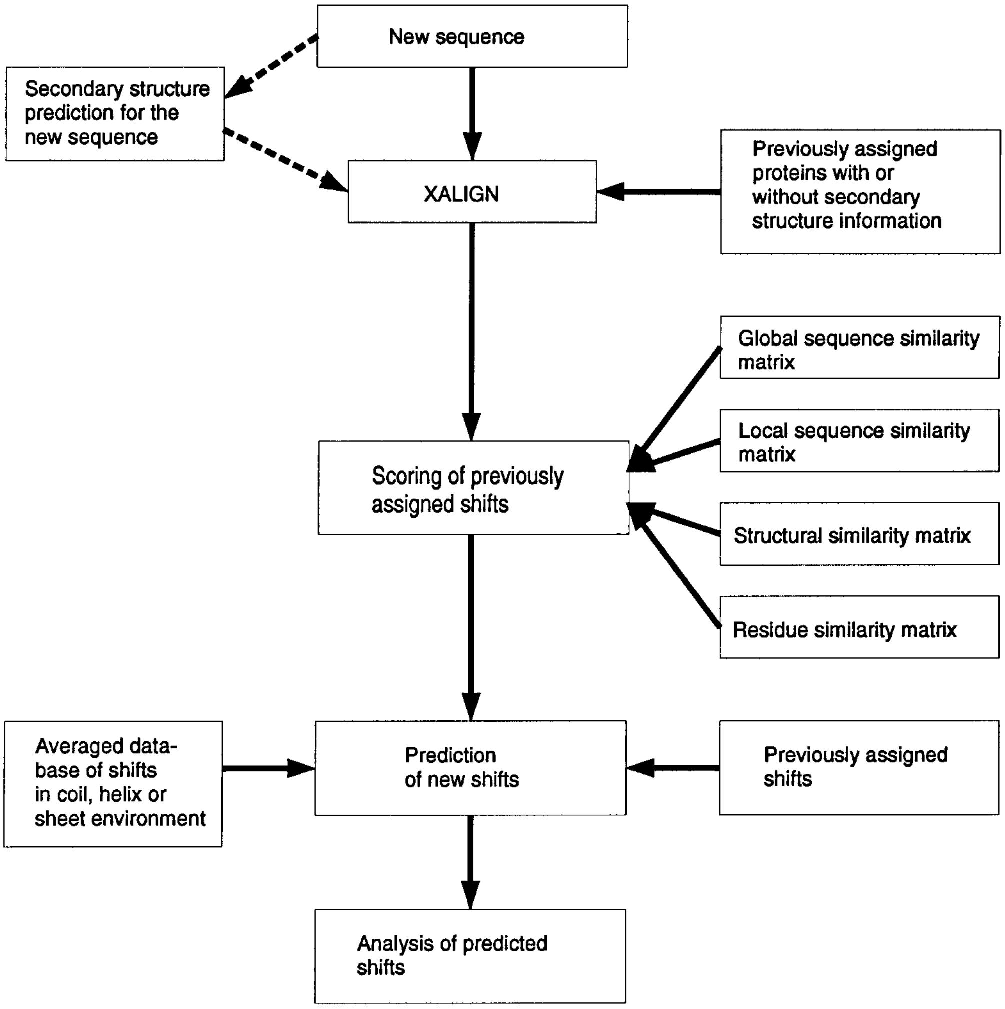

New sequence
Secondary structure prediction for the new sequence
XALIGN
Previously assigned proteins with or without secondary structure information
Global sequence similarity matrix
Local sequence similarity matrix
Scoring of previously assigned shifts
Structural similarity matrix
Residue similarity matrix
Averaged database of shifts in coil, helix or sheet environment
Prediction of new shifts
Previously assigned shifts
Analysis of predicted shifts

Fig. 1. Flow chart of the computer program ORB, presenting an outline of the chemical shift prediction process using ORB. All steps are further described in the Algorithm section.

assignments (Bax and Grzesiek, 1993). To accelerate the assignment process, some of these methods have been automated over the last several years (Kleywegt et al., 1991; Hare and Prestegard, 1994; Kjaer et al., 1994; Olson and Markley, 1994; Zimmermann et al., 1994; Morelle et al., 1995; Bartels et al., 1996). With the exception of Bartels et al. (1996) all these methods use only the information contained in the NMR spectra of interest. The latter method is able to incorporate additional structural and/or chemical shift information obtained from one highly homologous protein.

It is obvious that it would be of considerable advantage for the assignment process, whether manual or automatic, to have a precise prediction of the expected ^{1}H, ^{13}C and ^{15}N NMR chemical shifts. During the last years, several attempts have been made to theoretically calculate NMR chemical shifts from known secondary and tertiary structure elements (de Dios et al., 1993; Ösapay and Case, 1994; Williamson et al., 1995). In this paper, we use a different approach. With complete or nearly complete chemical shift assignments for more than 200 peptides and proteins now deposited in the BioMagResBank (Seavey et al., 1991), it seems logical to use all this prior knowledge in the process of predicting chemical shifts for a new protein. With a knowledge of the expected chemical shifts, the assignment process becomes less dependent on NOE and/or other connectivity information. This could substantially reduce the time required for NMR data collection and interpretation. The purpose of ORB is to appropriately combine the information of the database of previously assigned homologous sequences with averaged NMR chemical shift information to obtain a chemical shift prediction for a new sequence. In the process of testing ORB, we have decided to investigate the advantages of a database containing multiple homologous assignments. This is important if the previously assigned proteins have varying homologies to different regions of the query sequence. ORB is able to calculate a separate weight for every known shift to take these different sequence similarities into account. Using a database of previously assigned sequences allows ORB to estimate error ranges for every predicted shift, therefore providing the user with a confidence measurement for any given shift.

In the last few years, there have been systematic investigations to create averaged chemical shift databases based on previously assigned proteins or model peptides (Wishart et al., 1991,1995; Wishart and Sykes, 1994; Merutka et al., 1995). ORB incorporates the database developed by Wishart et al. (1991) in which the chemical shifts are categorized with respect to their residue, atom and secondary structure type. This database allows ORB to perform predictions in regions for which no homology-based assignments are possible. We describe in detail how ORB evaluates the applicability of previously assigned proteins and how it uses this information supplemented with averaged database values to predict chemical shifts for the unassigned homolog. ORB was tested on three different medium-sized proteins: an interleukin-8 analog, a synthetic troponin C heterodimer and cardiac troponin C. To predict chemical shifts for these proteins, databases compiled from six to 11 previously assigned proteins were used. The interleukin-8 analog was selected because a database of very highly homologous previously assigned proteins was available for this protein. For the second and third proteins, the available databases contained less homologous previously assigned proteins. This enables us to investigate how ORB performs under increasingly difficult conditions. In addition, we explore the possibility of predicting the shifts of an intact protein from the shifts of several smaller fragments of similar proteins. We also investigate using only one homologous previously assigned protein as opposed to a complete database of assigned proteins. On the basis of these results, we discuss the strengths and limitations of homology-based chemical shift predictions as used in ORB.

Algorithm

In essence, the ORB prediction algorithm is based on the use of previously assigned homologous proteins supplemented by an averaged NMR chemical shift database. Figure 1 displays the ORB flow chart. As seen in this chart, it is necessary to begin by obtaining a multiple sequence alignment of the unassigned protein of interest with a set of assigned homologous proteins. The alignment enables the program to find all the homologous shift information which pertains to any particular shift of the query protein. It is optional at this point to incorporate predicted secondary structure information for the new sequence prior to the sequential alignment.

Making a single chemical shift prediction can simply be a matter of taking a weighted average of the corresponding chemical shift data of the homologous proteins. In ORB the weight assigned to each homologous shift is determined by the following factors, which will be described in more detail below: (i) global sequence similarity; (ii) local sequence similarity; (iii) secondary structure similarity; and (iv) residue similarity. The weighting function becomes somewhat more complex because averaged NMR chemical shift database values are considered in the prediction process when the homologous chemical shift data are deemed to be poor or unavailable. In the final step of the assignment process, the program calculates a confidence interval for each chemical shift prediction based on the quality of the homologous shift data.

Averaged chemical shift database

ORB contains three averaged NMR chemical shift databases for ^{1}H, ^{13}C and ^{15}N chemical shifts. The chemical shift databases were derived from a statistical analysis of the chemical shifts, sequences and secondary structures of 78 different proteins for the ^{1}H chemical shifts (Wishart et al., 1991) and of 12 different proteins for the ^{13}C chemical shifts (Wishart and Sykes, 1994). From these data, it is possible to calculate average ^{1}H and ^{13}C chemical shifts for all 20 amino acids in each of three secondary structure categories: α-helices, β-sheets and 'coil' regions. Because ^{15}N chemical shifts are almost invariant to secondary structure, it is sufficient to use random coil values for the averaged ^{15}N chemical shift database. The ^{15}N chemical shifts were derived from an analysis of Gly-Gly-X-Ala-Gly-Gly hexapeptides (Wishart et al., 1994). These three databases enable ORB to predict chemical shifts according to the presumed or known secondary structure of the query sequence. They are particularly useful if no applicable homologous chemical shift information is available.

Alignment of previously assigned homologous sequences with the query sequence

A correct alignment between the query sequence and its homologs is critical to comparing, analyzing and predicting chemical shifts. At this point, it is the responsibility of the user to provide ORB with a set of previously assigned homologous proteins. Because the multiple sequence alignment problem is intrinsically difficult to solve, it was decided, for the sake of functionality, that an established program should be used to address this issue. ORB makes use of the XALIGN program (Wishart et al., 1994) that is included in the ORB package to accomplish this task. Consequently, ORB was designed to read the XALIGN output format. A user can choose, by some other method, to create his/her own alignment file, provided it conforms to the XALIGN output format.

Weighting homologous shift data

There are many factors which can be considered in determining the applicability of homologous shift data to predict the chemical shifts of a query protein. Each homologous protein, more accurately, each shift/atom, is com-

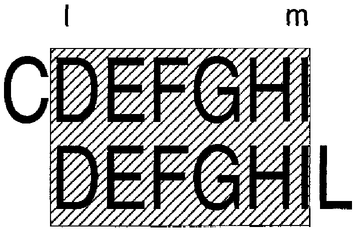


Fig. 2. Example of an alignment window where l is the residue number of the first two aligned amino acids and m is the residue number of the last two aligned amino acids. The reference residue numbers are obtained from the query sequence.

pared to the query protein and weighted according to the following criteria. The specified variables are set in an ORB parameter file. The program model assumes hereby that previously assigned shifts with higher similarity scores are considered more reliable for chemical shift prediction than those with lower similarity scores.

Global sequence similarity

Each homologous protein is compared to the query protein and the overall level of primary sequence similarity is determined, using the following four steps:

(1) Define the alignment window W(l,m), where l is the residue number of the first pair of amino acids which align between the two sequences and m is the last pair of amino acids in the alignment. This is clarified in Fig. 2.

(2) Using an amino acid similarity weighting matrix, find y_o, the sum of all amino acid pair scores in W. The chosen similarity weighting matrix was originally developed for multiple sequence alignments in the program SEQSEE (Wishart et al., 1994).

(3) Determine the perfect alignment score y_p, by summing all amino acid pair scores of the query protein with itself in W.

```
#
# B = beta Strand
# C = random coil
# H = helical
# T = beta turn
# X = unassigned
#
#   B   C   H   T   X
B   5
C   1   3
H   0   1   5
T   2   1   1   5
X   3   3   3   3   3
```

Fig. 3. The secondary structure similarity matrix used within ORB. This matrix is used by ORB to determine regions in which the secondary structure is conserved between sequences. The highest values are assigned to β-sheet, α-helix and turn regions.

(4) Calculate the global sequence homology gss(i), with i being the sequence number of the particular homologous sequence in the multiple alignment:

$$\mathrm{gss(i)} = \frac{100y_o}{y_p} \qquad (1)$$

Local sequence similarity

Local sequence similarity lss(i,r) uses the same algorithm as global sequence similarity except that the alignment window W is defined as W(r–s, r+s) and 2s+1 defines the window size. In calculating the local sequence similarity, the amino acids in all previously assigned sequences are renumbered according to their alignment with the query sequence to ensure that all residues that match in the multiple alignment have the same residue number r. In our program model, we typically set $s = 3$.

Structural similarity

Structural similarity considerations within ORB are limited to an assessment of secondary but not tertiary structure similarity. If the secondary structure of the query sequence is unknown, it can be obtained from a homologous X-ray or NMR structure. Alternatively, this information can be obtained from a variety of secondary structure prediction programs (Chou and Fasman, 1974, 1978; Garnier et al., 1978; Eisenberg et al., 1984; Levin et al., 1986; Gibrat et al., 1987; Williams et al., 1987; Levin and Garnier, 1988; Rooman and Wodak, 1988,1990,1991).

The calculation of the structural homology score sss(i,r) is identical to the local sequence similarity score, except that a secondary structure similarity weighting matrix (Fig. 3) is used instead of an amino acid similarity matrix in the calculation of y_o and y_p.

Note that, in this example, the sum for the perfect score y_p is calculated according to the following formula:

$$y_p = (2s + 1)\,\mathrm{max(secondary\_structure\_matrix\_value)} \qquad (2)$$

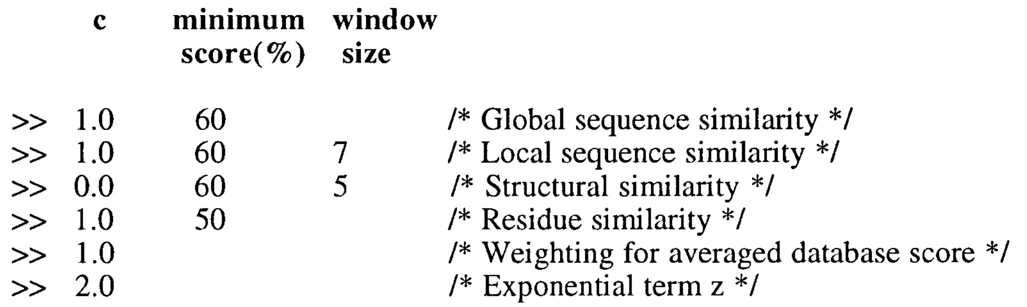
This ensures that perfect matches of 'coil' with 'coil' are scored lower than perfect matches between defined secondary structure elements. In our programming model, we typically set $s = 2$.

Residue similarity

Residue similarity describes the similarity between shifts arising from residues that differ in type, but have the same sequence position. For example, one could use an assigned leucine H^α to predict a corresponding alanine H^α in the query sequence via the following formula:

$$A(H^\alpha) = (L(H^\alpha) - L(dbH^\alpha)) + A(dbH^\alpha) \qquad (3)$$

where $A(dbH^\alpha)$ and $L(dbH^\alpha)$ are statistical chemical shift values derived from the averaged chemical shift database.



```
        c     minimum   window
              score(%)   size

>>   1.0       60                    /* Global sequence similarity */
>>   1.0       60          7         /* Local sequence similarity */
>>   0.0       60          5         /* Structural similarity */
>>   1.0       50                    /* Residue similarity */
>>   1.0                             /* Weighting for averaged database score */
>>   2.0                             /* Exponential term z */
```

Fig. 4. Example of an ORB parameter file. It is possible for users to adapt this file to their specific needs. All the parameters are described in detail in the Algorithm section.

The applicability of this converted shift information is determined by the residue similarity matrix, and a residue similarity score rss(i,r,a) is assigned to each known shift, where a is the specified atom in a particular residue.

The residue similarity matrix was generated by comparing the random coil chemical shifts of the same atoms in different amino acids, for instance a leucine H^{α} with a threonine H^{α}. The scores in this residue similarity matrix range from zero to 10, with the highest scores assigned to the pairs with the lowest random coil chemical shift differences. To generate the matrix, individual step sizes were used depending on the type of atom. In particular, 0.2 ppm steps were used for ^{1}H chemical shifts, 1.5 ppm steps for ^{15}N chemical shifts, 5.0 ppm steps for ^{13}C C^{β} chemical shifts and 1.8 ppm steps for all other ^{13}C chemical shifts. For example, the score for a leucine H^{α}/alanine H^{α} pair would be 10 because the difference in their random coil chemical shifts is less than 0.2 ppm.

Combining the various similarity factors

ORB uses a special weighting scheme to combine the above factors into a single chemical shift applicability score. The ORB programming model uses the equation below:

$$\begin{aligned} x(i,r,a) &= c_{gss}\,(gss(i) - gss_0) + c_{lss}\,(lss(i,r) - lss_0) \\ &\quad + c_{sss}(sss(i,r) - sss_0) + c_{rss}\,(rss(i,r,a) - rss_0) \end{aligned} \tag{4}$$

where x(i,r,a) is the applicability score for homologous shift i,r,a, and c is the coefficient for relative factor weighting (for example, one could choose to weight local homology more heavily than global homology). gss,lss,sss,rss represent the score for a particular factor. gss_0,lss_0,sss_0,rss_0 represent the minimum score for a particular factor. The advantage of including this term enables x(i,r,a) to be <0. This identifies the shifts which do not meet a minimum criterion. Any $x(i,r,a) < 0$ is set to 0 for programming convenience, which means this particular previously assigned shift is not used in the prediction process. The current values were determined in an extensive testing process. However, they can be changed individually in the ORB parameter file.

Calculating the predicted shift

The following equation is used by ORB to calculate a final predicted shift s(i_0,r,a), with i_0 being the sequence number of the query protein:

$$s(i_0,r,a) = \frac{x_0\,shift_0(i_0,r,a) + \sum_{i=1}^{n,i\neq i_0} x(i,r,a)new\_shift(i,r,a)}{x_0 + \sum_{i=1}^{n,i\neq i_0} x(i,r,a)} \tag{5}$$

where x_0 is the weight assigned to database shift, $shift_0(i_0$, r,a) is the averaged database shift value, x(i,r,a) is the homologous shift applicability score, new_shift(i,r,a) is the new_shift calculated according to Eq. 3 and n is the number of sequences in multiple alignment. Typically, x_0 is set to a small number in the ORB parameter file in order to emphasize homologous shifts which exceed the minimum applicability standards.

Equations 1–5 describe the essence of the chemical shift prediction protocol in ORB. We have further experimented with an exponential transformation function on the x(i,r,a) values. This serves to weight the most similar shifts to an even higher degree. The following equation will accomplish this:

$$a(i,r,a) = x(i,r,a)^{z} \tag{6}$$

where z is an exponential factor >1.

In summary, the chemical shift prediction is a weighted average of known homologous chemical shifts which is supplemented in cases of low homology by averaged database information. In the shown example parameter file (Fig. 4), *global sequence similarity*, *local sequence similarity* and *residue similarity* are all assigned the same weight, while *structural similarity* is not used. However, all parameters are user adjustable for increased flexibility. Figure 5 shows an example of the extended ORB prediction file, illustrating how the predicted shifts are derived.

Analysis of predicted shifts

The final prediction file (Fig. 6) contains all predicted ^{1}H, ^{13}C and ^{15}N chemical shifts for the query sequence. It

A

	Sequence(i)	*Residue(r)*	*gss(i,r)*	*lss(i,r)*	*sss(i,r)*
	IL6-72	I-10	98	100	60
	IL5-72	I-10	98	100	60
	ILL25N	I-10	99	100	60
	IL8.1	I-10	99	100	60
Predict	IL4-72	I-10			
	IL8.2	I-10	99	100	60
	IL1-72	I-10	99	100	60
	ILR6K	I-10	99	100	60
	ILH33A	I-10	97	100	60
	ILE38A	I-10	98	100	60
	ILI10A	A-10	98	88	60
	MGSA	L-12	64	71	60

B

Atom(i_0,r,a)		*Sequence(i)*		*rss(i,r,a)*		*new_shift(i,r,a)*	
	s(i_0,r,a)		*PPMVal(i,r,a)*		*Offset(i,r,a)*		*PctWt(i,r,a)*
HA	4.19						
		Random Coil				4.09	0.0%
		IL6-72	4.20	100	0.11	4.20	11.1%
		IL5-72	4.19	100	0.10	4.19	11.1%
		IL8.1	4.18	100	0.09	4.18	11.3%
		IL8.2	4.20	100	0.11	4.20	11.3%
		IL1-72	4.20	100	0.11	4.20	11.3%
		ILR6K	4.19	100	0.10	4.19	11.3%
		ILH33A	4.19	100	0.10	4.19	10.9%
		ILE38A	4.14	100	0.05	4.14	11.1%
		ILI10A	4.38	100	0.19	4.28	9.0%
		MGSA	4.38	90	0.03	4.12	1.6%

Fig. 5. Example of an extended ORB prediction file. All the information necessary to predict a single chemical shift is shown. Figure 5A displays, for a single specified residue, all previous assigned sequences Sequence(i) that were used in the prediction process together with the corresponding global, local and structural similarity values gss(i), lss(i,r) and sss(i,r). The predicted sequence Predict is the sequence(i_0). Figure 5B shows the predicted Atom(i_0,r,a) in the first column and the predicted shift s(i_0,r,a) in the next column. The third column contains all previously assigned sequences Sequence(i) together with their corresponding chemical shift values PPMVal(i,r,a) in the next column. Column five contains the residue similarity scores rss(i,r,a) for the previous assigned atoms. The offsets Offset(i,r,a) between measured ppm values and their corresponding random coil values are displayed in column six. In the next column, these offsets are added to the corresponding random coil value of the predicted atom to get the so-called new_shift(i,r,a). The last column contains the percentage to which one particular previous assigned shift is used in the prediction process PctWt(i,r,a).

also includes the expected standard deviations for all predicted shifts. These standard deviations provide an estimate of the accuracy (and a level of confidence) for the predicted shifts. Currently, two methods are used to calculate these standard deviations. Method 1 uses Eq. 7:

$$\text{stdev\_1}(i_0,r,a) = \frac{sd_{rdmc}(i_0,r,a)}{ns} \tag{7}$$

This method uses the random coil standard deviation $sd_{rdmc}(i_0,r,a)$ (Wishart et al., 1991) divided by the number of sequences in the multiple alignment for which shifts are known (ns). If one only has a single previously assigned sequence, one would always obtain half of the random coil standard deviation by this method. Method 2 uses Eq. 8:

$$\text{stdev\_2}(i_0,r,a) = \frac{x_0\, sd_{rdmc}(i_0,r,a) + \sum_{i=1}^{n,i\neq i_0} x(i,r,a)\, sd(i_0,r,a)}{x_0 + \sum_{i=1}^{n,i\neq i_0} x(i,r,a)} \tag{8}$$

Method 2 (which is more accurate) is only used if more than one previously assigned sequence is available for the prediction process. In this method, the standard deviation sd(i_0,r,a) is first calculated from all new_shifts that correspond to one particular predicted shift s(i_0,r,a). Second, a weighted average of sd(i_0,r,a) and $sd_{rdmc}(i_0,r,a)$ is calculated to obtain stdev_2(i_0,r,a).

```
! Atom           s(i0,r,a)  stdev_1/2  shift0(i0,r,a)  sdrdmc    confidence   CSI
!
1:LEU_5:N          122.10     1.50       121.80         3.00        *
1:LEU_5:CA          54.80     0.69        55.42         1.38        *          C
1:LEU_5:HN           8.36     0.26         7.99         0.61        *
1:LEU_5:HA           4.43     0.04         4.35         0.28        ***        C
1:LEU_5:C          177.28     1.32       177.28         1.32        -
1:LEU_5:CB          41.90     0.84        42.29         1.68        *
1:LEU_5:HB1          1.61     0.06         1.73         0.24        **
1:LEU_5:HB2          1.52     0.09         1.57         0.35        **
1:LEU_5:HG           1.60     0.07         1.55         0.29        **
1:LEU_5:HD1#         0.82     0.07         0.85         0.28        **
1:LEU_5:HD2#         0.85     0.11         0.71         0.44        **
```

Fig. 6. Example of a final prediction file. The first two columns contain the atom names together with the corresponding predicted chemical shifts. The predicted standard deviations are displayed in the third column. The next two columns show the corresponding random coil shifts and standard deviations. The last two columns contain the confidence values, ranging from – for random coil shifts only to **** for highest confidence, and a prediction of the expected secondary structure.

In order not to overestimate the quality of the predictions, the method that gives the larger standard deviations in a particular case is generally used. To provide a simple measure of the quality of the results, a qualitative assessment called Confidence is incorporated in the final prediction file. The Confidence level is based on the standard deviation assigned to the predicted shifts. A small standard deviation suggests a high level of confidence and vice versa. The level of confidence is qualitatively indicated by a symbolic range from '–' to four '*'. The last column of the prediction file contains a prediction of the expected secondary structure derived from the predicted H^{α} and C^{α} chemical shifts. This secondary structure prediction is based on the chemical shift index method (Wishart et al., 1992; Wishart and Sykes, 1994). If a set of reference shifts is available (these could be, for example, shifts from a previous NMR assignment at a different pH or temperature), it is possible to graphically analyze the predicted shifts. This can be done individually for each type of atom (H^{α}, H^{N}, etc.). Examples for this procedure are shown in the Results section.

Results

Three systems were investigated to evaluate the extent to which ORB is able to predict ^{1}H, ^{13}C and ^{15}N chemical shifts for an unassigned query sequence. To evaluate the performance of ORB under various conditions, we used the following four testing schemes on the three test systems: (i) predictions using random coil shifts alone; (ii) predictions using only the shifts of the least homologous sequence; (iii) predictions using only the shifts of the most homologous sequence; and (iv) predictions using the shift information from all available homologs. The ORB parameters shown in Fig. 4 were used throughout the prediction process, if not otherwise stated. Structural homology considerations were not included in these tests because the secondary structures of all query proteins were assumed to be unknown.

The first protein analyzed was IL-8 (4–72), a 69-residue homodimeric protein of the CXC chemokine family which is missing the first three N-terminal residues. IL-8 (4–72) contains extensive β-sheet structures and one α-helix

TABLE 1
DATABASE OF PREVIOUSLY ASSIGNED PROTEINS OF THE CXC CHEMOKINE FAMILY

Sequence	Percent identity to IL-8 (4–72)	Reference
IL-8 analog (6–72)	97	Rajarathnam et al. (1994)
IL-8 analog (5–72)	99	Rajarathnam et al. (1994)
IL-8 L25NMe (4–72) monomer	100	Rajarathnam et al. (1995)
IL-8 analog (4–72)	100	Rajarathnam et al. (1994)
IL-8 native (1–72) (A)	100	Rajarathnam et al. (1994)
IL-8 native (1–72) (B)	100	Clore et al. (1989)
IL-8 R6K analog (4–72)	99	Rajarathnam et al. (1994)
IL-8 H33A analog (4–72)	97	Rajarathnam et al. (1994)
IL-8 E38A analog (4–72)	99	K. Rajarathnam (1996) personal communication
IL-8 I10A analog (4–72)	99	Rajarathnam et al. (1994)
PF4-M2 (1–67)	38	Mayo et al. (1995)
MGSA (1–72)	44	Kim et al. (1994)

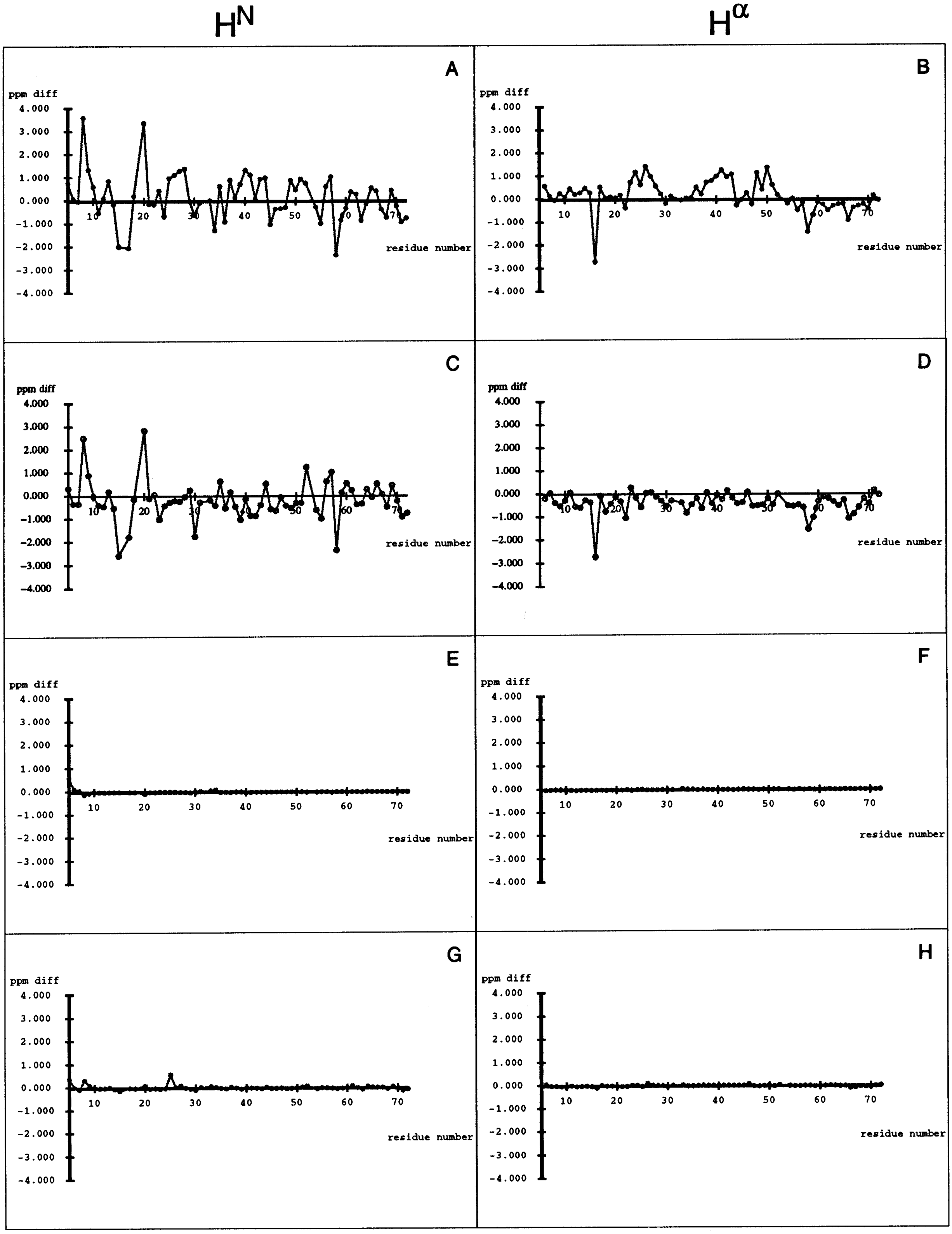


Fig. 7. Results from the chemical shift predictions of the backbone H^N and H^α nuclei of IL-8 (4–72) plotted as a function of residue number. The diagrams show observed shifts minus predicted shifts. The leftmost column contains the results for H^N shifts while the rightmost column contains the corresponding results for H^α shifts. (A, B) Results obtained using only random coil data. (C, D) Results obtained using the least homologous sequence supplemented with random coil values. (E, F) Results obtained using the most homologous sequence supplemented with random coil values. (G, H) Results where the prediction process uses the whole database of previously assigned sequences supplemented with random coil values. Note that no points are shown for proline residues and that the comparisons for the H^N shifts start with residue 5 while the comparisons for the H^α shifts start with residue 6.

TABLE 2
STATISTICS FOR CHEMICAL SHIFT PREDICTION FOR IL-8 (4–72)

	Average error H^N (ppm)	Correlation coefficient H^N	Average error H^α (ppm)	Correlation coefficient H^α
Random coil shifts	0.77	–0.14	0.49	0.18
Least homologous shifts	0.61	0.53	0.41	0.75
Most homologous shifts	0.02	1.00	0.01	1.00
All homologous shifts	0.06	0.99	0.03	1.00

packed into a well-defined tertiary structure. To predict the chemical shifts for this protein, a chemical shift database consisting of the ^1H chemical shifts of 11 homologous proteins was used (Table 1). Note that the shifts of IL-8 (4–72) were not used for the prediction process; they were only used for comparing predicted shifts with the observed shifts to assess the quality of the predictions. The prediction process began with a multiple sequence alignment of all 12 sequences including IL-8 (4–72) (data not shown) using the program XALIGN (Wishart et al., 1994). The sequence identities for all the sequence pairs compared to IL-8 (4–72) are shown in Table 1. Note that the sequence identities were obtained using the multiple sequence alignment of all sequences in Table 1.

For the first test with this system, only random coil chemical shifts were used to predict shifts for IL-8 (4–72). For the second test, only the shifts of PF4-M2, which has the lowest pairwise sequence identity of 38% with IL-8 (4–72), were used supplemented with random coil shifts to predict IL-8 (4–72). For the third test case, the shifts of the most homologous sequence IL-8 (1–72) (A), which has 100% sequence identity to IL-8 (4–72) and contains three additional N-terminal residues, were used supplemented with random coil data for predictive calculations using ORB. For the fourth test case, the shifts of all assigned homologous sequences were used supplemented with random coil values for the predictions. For the entire testing, H^N and H^α atoms were chosen as representative atoms for all residues. In the following, the predicted H^N and H^α shifts were compared with the experimentally determined shifts of IL-8 (4–72) (Fig. 7 and Table 2). The average error shown in Table 2 is the sum of the absolute values of all the chemical shift differences in parts per million (ppm) between the observed and predicted shifts for the selected nuclei divided by the number of residues. In addition to the average errors, Table 2 contains the corresponding Pearson correlation coefficients (Larsen and Marx, 1981). Both methods are well known and widely used for NMR chemical shift comparisons and both are included for the sake of comparison. The results from the first case show that using only random coil values for the prediction can lead to large errors of up to 4 ppm for the H^N shifts and up to 3 ppm for the H^α shifts. As might be expected, there is no regular pattern observed for the predicted H^N shifts and they vary widely throughout the sequence. This indicates, at least in this example, that random coil chemical shifts alone are not suitable for accurately predicting H^N shifts. Similar conclusions can be made about the H^α shifts. If the predicted H^α shifts are compared with the secondary structure of other IL-8 analogs (data not shown), it is apparent that the observed shifts in regions that are in β-sheets (Lys^{23} to Glu^{29}, Ala^{35} to Leu^{43}, Gly^{46} to Cys^{50}) are shifted downfield compared to the random coil shifts, while the observed shifts in α-helical regions (Asn^{56} to Ser^{72}) are shifted upfield compared to the random coil shifts (Fig. 7). Using secondary structure information would certainly help the predictions if information about the secondary structure was available. A large error of almost 3 ppm is observed for Pro^{16} (Fig. 7) where the real observed shift is at 1.6 ppm due to a ring current effect caused by Trp^{57} (actual sequence not shown). In the second test case, large shift differences of up to 3 ppm were still observed between observed and predicted shifts for the H^N protons, especially in the N-terminal region. The average error decreased for the H^N shifts from 0.77 to 0.61 ppm. The same was true for the H^α shifts, where the average error decreased from 0.49 to 0.41 ppm. The prediction for the H^α of Pro^{16} is still off by about 3 ppm.

Dramatic improvements occurred in the third test. The predictions were almost exact for the H^N and H^α shifts, with average errors of 0.02 and 0.01 ppm, respectively. These results are close to the precision at which these shifts were experimentally determined. In the last test case, the predictions are very close to the observed shifts for both the H^N and H^α shifts, with average errors of 0.06 and 0.03 ppm for the H^N and H^α shifts, respectively. Compared to the previous test case, the average error increased by a small amount for both the H^N and H^α shifts. The likely cause for this increase is the inclusion of sequences with low homology (38% sequence identity) in the prediction. A possible solution would be to increase the exponential factor z in the ORB parameter file, thereby putting more weight on the shifts with the highest sequence identity to the query sequence.

The second protein studied was the troponin C III.IV heterodimer ((93–126) and (129–162)) (TF). This protein represents the calcium binding sites III and IV of the muscle protein troponin C. It consists mainly of α-helical sections and some small β-sheet regions. A chemical shift database consisting of the chemical shifts of eight additional homologous proteins was used in the prediction

TABLE 3
DATABASE OF PREVIOUSLY ASSIGNED CALCIUM BINDING PROTEINS

Sequence	Percent identity to TF	Percent identity to cNTnC	Reference
TnC	96	53	Slupsky et al. (1995) and Slupsky and Sykes (1995)
NTnC0c	–	53	Gagne et al. (1994)
NTnC2c	–	53	Gagne et al. (1994)
CTnC	96	–	Calhoun and Sykes (1996)
TF	100	–	Shaw et al. (1992)
TT	65	–	Shaw et al. (1990)
CaM	34	47	Ikura et al. (1990)
cNTnC	–	100	M.X. Li (1996) personal communication
CalB	36	19	Anglister et al. (1994)

process. The percent pairwise sequence identities of all sequences versus TF are shown in Table 3. Table entries marked with a dash indicate, for all investigated proteins, sequences which do not overlap with the query sequence in the multiple alignment and, therefore, are excluded from the prediction process.

The results for TF for all four test cases are displayed in Fig. 8 and Table 4. Again, random coil shifts alone are not sufficient for obtaining accurate predictions with average errors of 0.50 and 0.27 ppm for the H^N and H^α shifts, respectively. On the other hand, a comparison with the corresponding results for IL-8 (4–72) shows that the random coil shifts give better predictions in this case, especially for the H^α shifts. The improvement over the IL-8 (4–72) random coil only predictions is due to the absence of extended β-sheet structure in TF. In the next test case, the shifts of calmodulin (CaM), the database member with the lowest pairwise sequence identity to TF (34%) of all sequences that align with TF, were used supplemented with random coil values. Compared to the previous test case, the quality of the predictions improved for the H^N and the H^α shifts, with average errors of 0.25 and 0.13 ppm, respectively. The use of sequences with such a low pairwise sequence identity is still sufficient to substantially improve the predictions made solely using random coil values. Due to the low sequence homology, the ORB parameters were changed for this test. Specifically, the gss_0, lss_0, sss_0 and rss_0 values were divided by two and the c_0 value for the averaged database weighting was divided by 10 to ensure a proper weighting of the CaM shifts. Next, the shifts of the most homologous sequence, C-domain cloned chicken troponin C (88–162) (CTnC) with a pairwise sequence identity of 96% to TF, were used supplemented with random coil values for the prediction. CTnC was used in this case and not troponin C 1–162 (TnC), which has the same pairwise sequence identity to TF of 96%, because CTnC is more similar in length to TF than TnC. The results improved dramatically, with average errors of 0.09 and 0.04 ppm for the H^N and H^α shifts, respectively. The only regions where small errors occur in the predictions for both the H^N and H^α shifts are in the region around the gap in TF between Gly^{126} and Val^{129}, and for the last three residues in the C-terminal portion of the molecule. Using the whole database of all eight assigned sequences, predictions with average errors of 0.09 and 0.04 ppm were obtained for the H^N and H^α shifts, respectively. This result is identical to the previous test case. On the other hand, a graphical inspection of the results (Fig. 8) shows that, compared to the previous case (CTnC), the errors in the gap region are getting smaller, especially for the H^α shifts. Overall, the best results for TF were obtained if all eight assigned proteins were used in the prediction process.

To simulate a case where only a smaller and less homologous database of assigned sequences is available, TnC and CTnC were eliminated from the database. In this smaller database, the troponin C III.III homodimer ((93–126) and (129–162)) (TT) has the highest pairwise sequen-

TABLE 4
STATISTICS FOR CHEMICAL SHIFT PREDICTION FOR TF USING THE WHOLE AND SMALL DATABASES OF PREVIOUSLY ASSIGNED PROTEINS

	Average error H^N (ppm)	Correlation coefficient H^N	Average error H^α (ppm)	Correlation coefficient H^α
Whole database				
Random coil shifts	0.50	–0.07	0.27	0.53
Least homologous shifts	0.25	0.88	0.13	0.92
Most homologous shifts	0.09	0.98	0.04	0.99
All homologous shifts	0.09	0.98	0.04	0.99
Small database				
Most homologous shifts	0.22	0.90	0.15	0.87
All homologous shifts	0.18	0.93	0.10	0.95

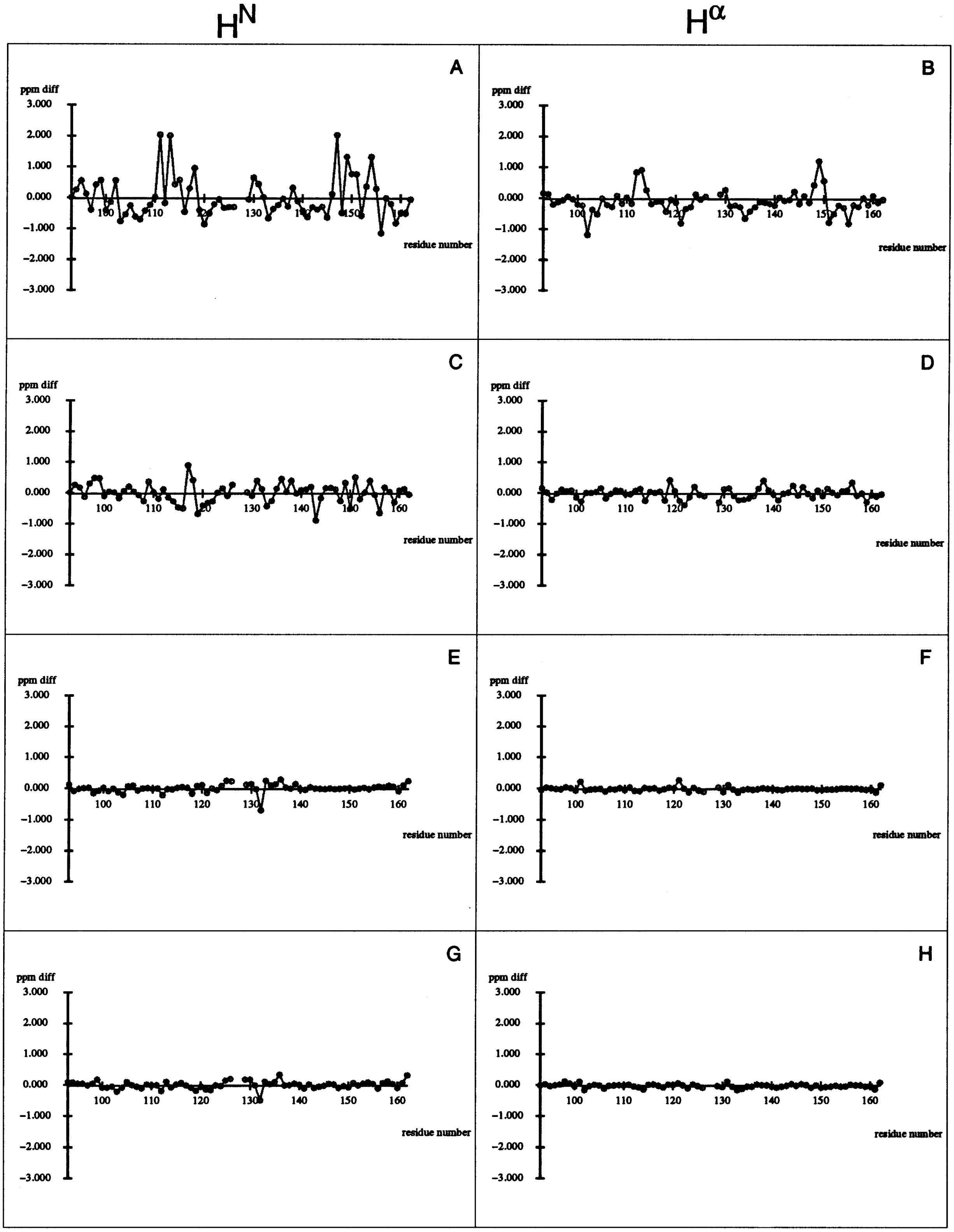


Fig. 8. Results from the chemical shift predictions of the backbone H^N and H^α nuclei of TF plotted as a function of residue number. Other features of the figure are as described in the caption to Fig. 7.

ce identity of 65% to TF. The pairwise sequence identity of TT is at approximately the midpoint between CalB and CTnC in pairwise sequence homology to TF. First, the shifts of TT supplemented with random coil values were used to predict shifts for TF. The results displayed in Fig. 9 and Table 4 show an average error of 0.22 and 0.15

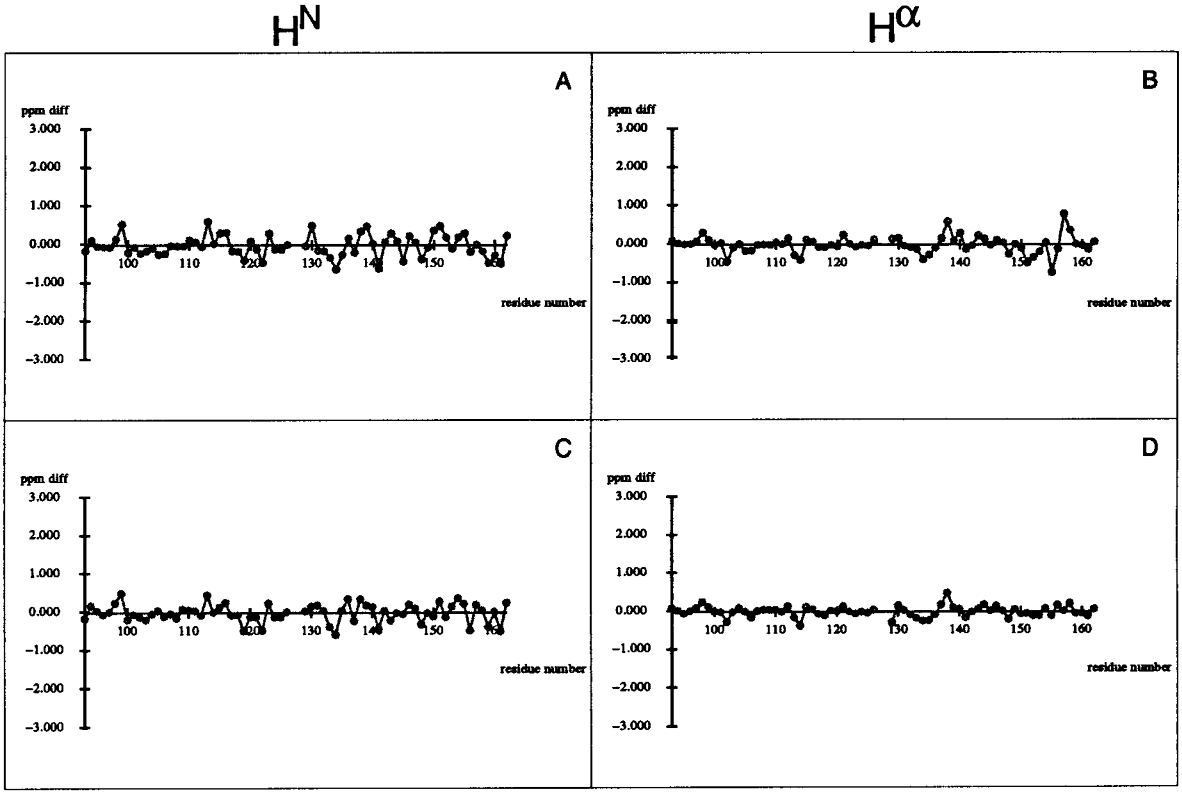


Fig. 9. Results from the chemical shift predictions of the backbone H^N and H^α nuclei of TF plotted as a function of residue number. In this example, a reduced database of previously assigned proteins was used in the prediction process. The first two diagrams (A, B) contain the results obtained when the shifts of the most homologous sequence supplemented with random coil values were used. The next two diagrams (C, D) display the results obtained when the entire reduced database was used supplemented with random coil values.

ppm for the H^N and H^α shifts, respectively. By using TT instead of CTnC as the protein with the highest sequence identity, the average error increased for the H^N shifts by around 2 times and for the H^α shifts by around 4 times. Next, all six assigned sequences of the small database were used supplemented with random coil values to predict shifts for TF (Fig. 9 and Table 4). The average errors for the H^N and H^α shifts of 0.18 and 0.10 ppm, respectively, are 18% and 38% smaller than the errors obtained in the previous case where only the TT shifts supplemented with random coil shifts were used for the prediction. Here it is very clear that the use of the entire database really improves the results compared to the case where only the most homologous sequence is used.

The third protein investigated was the N-domain of cardiac troponin C (1–89) (cNTnC). This protein is an 89-residue monomeric protein. It consists mainly of α-helical sections and some small β-sheet regions. The same database of assigned proteins that was used for the first four TF test cases, now including the shifts for TF and excluding the shifts for cNTnC, was used for the predictions of cNTnC. The four test cases were investigated as described previously and the results for cNTnC are displayed in Fig. 10 and Table 5.

Using random coil values for the predictions gave to the corresponding TF test case comparable average errors of 0.45 and 0.27 ppm for the H^N and H^α shifts, respectively.

In the next case, the shifts of CalB were used supplemented with random coil data for the predictions. The sequence identity of CalB to cNTnC is only 19%. To ensure that CalB shifts were preferentially used in the prediction process, the ORB parameters were changed as described above for TF. The average errors of 0.45 and 0.21 ppm for the H^N and H^α shifts, respectively, are simi-

TABLE 5
STATISTICS FOR CHEMICAL SHIFT PREDICTION FOR cNTnC

	Average error H^N (ppm)	Correlation coefficient H^N	Average error H^α (ppm)	Correlation coefficient H^α
Random coil shifts	0.45	0.23	0.27	0.48
Least homologous shifts	0.45	0.52	0.21	0.68
Most homologous shifts	0.34	0.59	0.17	0.84
All homologous shifts	0.33	0.61	0.15	0.83

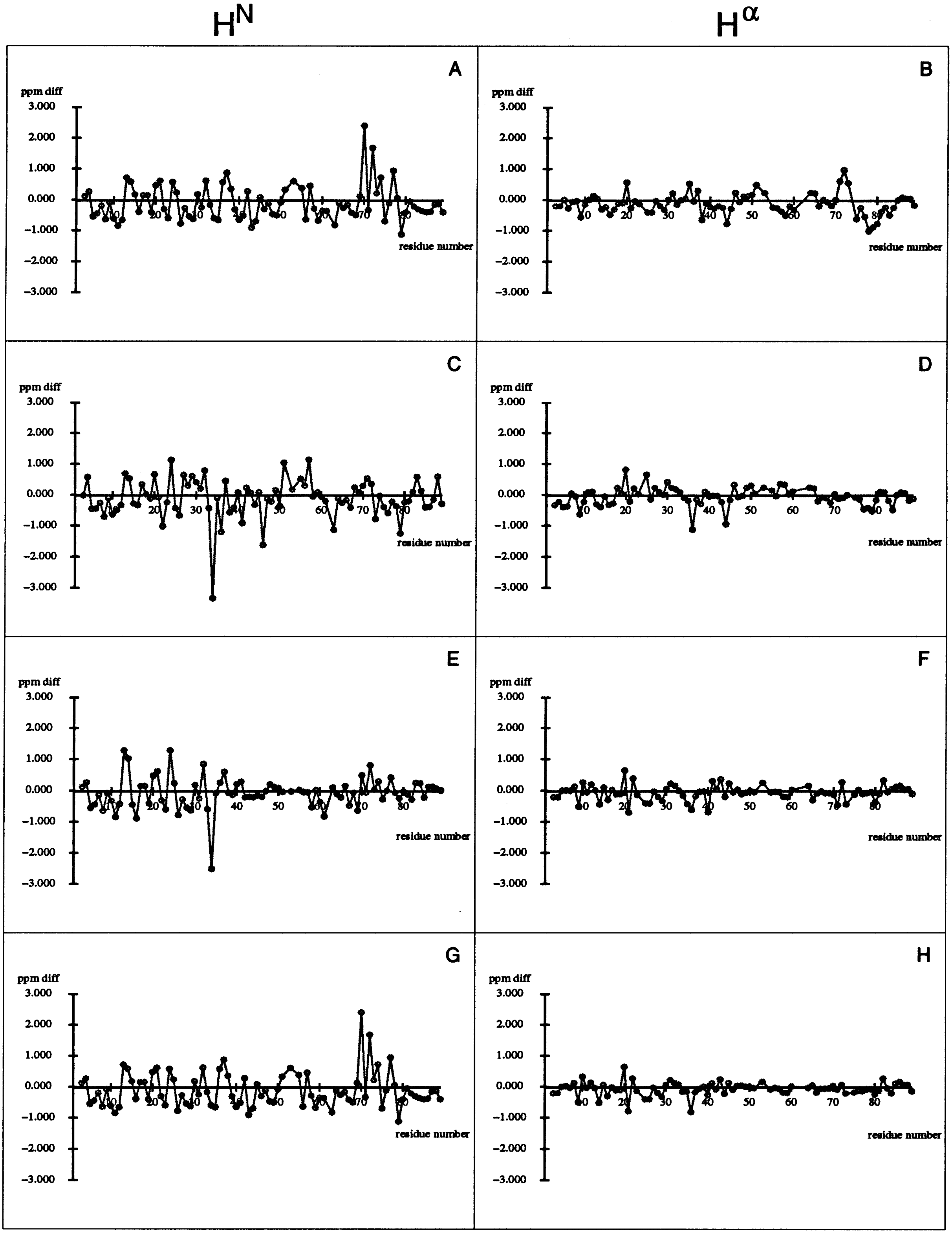


Fig. 10. Results from the chemical shift predictions of the backbone H^N and H^α nuclei of cNTnC plotted as a function of residue number. Other features of the figure are as described in the caption to Fig. 7.

lar to the errors obtained using only random coil data. Apparently, proteins with sequence identities to the query protein of less than 20% are not useful in the prediction process. The results improved significantly when the shifts of the most homologous sequence of apo N-domain troponin C (1–90) (NTnC0c) with a sequence identity of

53% to cNTnC supplemented with random coil shifts were used for the prediction. Average errors of 0.34 and 0.17 ppm were obtained for the H^N and H^α shifts, respectively. Compared to the corresponding test cases for IL-8 (4–72) and the first series for TF, the precision of the predictions has dropped, which is a result of the lower sequence identity of the most homologous sequence. Using the whole database of all assigned sequences improves the results further. Average errors of 0.33 and 0.15 ppm were obtained for the H^N and H^α shifts, respectively. In this series of tests, the use of all assigned sequences supplemented with random coil values is again superior to the use of only the shifts of the most homologous sequence supplemented with random coil shifts.

A comparison of the average errors with the correlation coefficients (Tables 2, 4 and 5) shows that, in general, both go hand in hand. A small average error corresponds to a large correlation coefficient and vice versa. One exception occurs for the H^α chemical shifts in the second test case for IL-8 (4–72). The average error is unusually high for the corresponding correlation coefficient. A visual inspection of Fig. 7 shows that a small systematic shift occurs for the predicted H^α shifts. This is probably caused by a referencing error for the PF4-M2 shifts, which were used in this test case. This points to the importance of proper referencing of chemical shifts (Wishart and Sykes, 1994).

Throughout this section, results were only discussed for the H^N and H^α atoms. However, tests have shown that the results obtained for the side-chain atoms (e.g. H^β, H^γ, etc.) are comparable to the results obtained for the H^α atoms (data not shown).

Discussion and Conclusions

From the investigation of all three test systems, it is clear that ORB obtains excellent results when at least one highly homologous, fully assigned protein is available for the prediction process. This is very clear to see for IL-8 (4–72) and for the first series of the TF tests. For all three proteins, it is possible to obtain accurate predictions as long as a whole database of previously assigned proteins is used. With decreasing sequence identity between the query protein and the assigned proteins, the use of more than one sequence becomes increasingly important for the prediction process. This can easily be seen for the cNTnC and the second series of the TF tests. Using a whole database of assigned shifts, ORB gives reasonably good results as long as one or more of the assigned sequences has at least 30% sequence identity to the query sequence. In the test cases where only the least homologous sequence was used, the results were sometimes no better than if random coil values were used. This shows that homologous assignment techniques using proteins with sequence identities of less than 30% to the query sequence do not seem to be useful for the prediction process. Tests involving cNTnC and the second series of TF (Figs. 10 and 9) clearly show that some predictions can be very precise while others can be off by a considerable margin. To allow the selection of the predictions that are most probably correct, a Confidence value is calculated for every predicted shift, as described in the Algorithm section. On the basis of these confidence values, a user can select dependable starting points to assist with the manual assignment process. This procedure was used with great success during the initial assignment of cNTnC (M.X. Li (1996) personal communication).

ORB can be used as a stand-alone program to assist the user during the manual assignment process or it is possible to combine ORB with other computer assignment procedures to allow a fast automatic assignment. With the number of published protein and peptide assignments increasing on a day-to-day basis, a program like ORB that makes use of this information will facilitate rapid assignment of a new protein. The program can be accessed from the following URL: http://www/pence.ualberta.ca/export/docs.

Acknowledgements

We thank T. Jellard for programming the graphical user interface for ORB, L. Willard for helpful discussions regarding the programming of ORB and K. Rajarathnam for providing most of the shifts for the interleukin family and for a critical reading of the manuscript. In addition, we would like to thank M. Li for providing the unpublished shifts of cNTnC and B. Lix and M. Crump for many helpful discussions. This work is supported by the Protein Engineering Network of Centres of Excellence of Canada.

References

Anglister, J., Grzesiek, S., Wang, A.C., Ren, H., Klee, C.B. and Bax, A. (1994) *Biochemistry*, **33**, 3540–3547.

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

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

Chou, P.Y. and Fasman, G.D. (1974) *Biochemistry*, **13**, 222–245.

Chou, P.Y. and Fasman, G.D. (1978) *Annu. Rev. Biochem.*, **47**, 251–276.

Clore, G.M., Apella, E., Yamada, M., Matsushima, K. and Gronenborn, A.M. (1989) *J. Biol. Chem.*, **264**, 18907–18911.

de Dios, A.C., Pearson, J.G. and Oldfield, E. (1993) *Science*, **260**, 1491–1496.

Eisenberg, D., Weiss, R.M. and Terwilliger, R.C. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 140–144.

Gagne, S.M., Tsuda, S., Li, M.X., Chandra, M., Smillie, L.B. and Sykes, B.D. (1994) *Protein Sci.*, **3**, 1961–1974.

Garnier, J., Ogusthorpe, D.J. and Robson, B. (1978) *J. Mol. Biol.*, **120**, 97–120.

Garrett, D.S., Powers, R., Gronenborn, A.M. and Clore, M. (1991) *J. Magn. Reson.*, **95**, 214–220.

Gibrat, J.F., Garnier, J. and Robson, B. (1987) *J. Mol. Biol.*, **198**, 425–443.

Hare, B.J. and Prestegard, J.H. (1994) *J. Biomol. NMR*, **4**, 35–46.

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

Kim, K.-S., Clark-Lewis, I. and Sykes, B.D. (1994) *J. Biol. Chem.*, **269**, 32909–32915.

Kjaer, M., Andersen, K.V. and Poulsen, F.M. (1994) *Methods Enzymol.*, **239**, 288–318.

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

Larsen, R.J. and Marx, M.L. (1981) *An Introduction to Mathematical Statistics and its Applications*, Prentice-Hall, Englewood Cliffs, NJ, U.S.A.

Levin, J.M., Robson, B. and Garnier, J. (1986) *FEBS Lett.*, **205**, 303–308.

Levin, J.M. and Garnier, J. (1988) *Biochim. Biophys. Acta*, **955**, 283–295.

Mayo, K.H., Roongta, V., Ilyina, E., Milius, R., Barker, S., Quinlan, C., La Rosa, G. and Daly, T.J. (1995) *Biochemistry*, **34**, 11399–11409.

Merutka, G., Dyson, H.J. and Wright, P.E. (1995) *J. Biomol. NMR*, **5**, 14–24.

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

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

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

Rajarathnam, K., Clark-Lewis, I. and Sykes, B.D. (1994) *Biochemistry*, **33**, 6623–6630.

Rajarathnam, K., Clark-Lewis, I. and Sykes, B.D. (1995) *Biochemistry*, **34**, 12983–12990.

Rooman, M.J. and Wodak, S.F. (1988) *Nature*, **335**, 45–49.

Rooman, M.J. and Wodak, S.F. (1990) *J. Mol. Biol.*, **213**, 337–350.

Rooman, M.J. and Wodak, S.F. (1991) *Proteins Struct. Funct. Genet.*, **9**, 68–78.

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

Shaw, G.S., Hodges, R.S. and Sykes, B.D. (1990) *Science*, **249**, 280–283.

Shaw, G.S., Findlay, W.A., Semchuk, P.D., Hodges, R.S. and Sykes, B.D. (1992) *J. Am. Chem. Soc.*, **114**, 6258–6259.

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

Slupsky, C.M. and Sykes, B.D. (1995) *Biochemistry*, **34**, 15953–15964.

Williams, R.W., Chang, A., Juretic, D. and Loughram, S. (1987) *Biochim. Biophys. Acta*, **916**, 200–204.

Williamson, M.P., Kikuchi, J. and Asakura, T. (1995) *J. Mol. Biol.*, **247**, 541–546.

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

Wishart, D.S., Richards, F.M. and Sykes, B.D. (1992) *Biochemistry*, **31**, 1647–1651.

Wishart, D.S., Boyko, R.F. and Sykes, B.D. (1994) *CABIOS*, **10**, 687–688.

Wishart, D.S., Boyko, R.F., Willard, L., Richards, F.M. and Sykes, B.D. (1994) *CABIOS*, **10**, 121–132.

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

Wishart, D.S. and Sykes, B.D. (1994) *Methods Enzymol.*, **239**, 363–391.

Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995) *J. Biomol. NMR*, **5**, 67–81.

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

Zimmermann, D., Kulikowski, C., Wang, L., Lyons, B. and Montelione, G.T. (1994) *J. Biomol. NMR*, **4**, 241–256.