Using Neural Network Predicted Secondary Structure Information in Automatic Protein NMR Assignment

W. Y. Choy and B. C. Sanctuary*

Department of Chemistry, McGill University, 801 Sherbrooke Street West, Montreal, Quebec, Canada H3A 2K6

Guang Zhu

Department of Biochemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Received February 24, 1997[⊗]

In CAPRI, an automated NMR assignment software package that was developed in our laboratory, both chemical shift values and coupling topologies of spin patterns are used in a procedure for amino acids recognition. By using a knowledge base of chemical shift distributions of the 20 amino acid types, fuzzy mathematics, and pattern recognition theory, the spin coupling topological graphs are mapped onto specific amino acid residues. In this work, we investigated the feasibility of using secondary structure information of proteins as predicted by neural networks in the automated NMR assignment. As the ^{1}H and ^{13}C chemical shifts of proteins are known to correlate to their secondary structures, secondary structure information is useful in improving the amino acid recognition. In this study, the secondary structures of proteins predicted by the PHD protein server and our own trained neural networks are used in the amino acid type recognition. The results show that the predicted secondary structure information can help to improve the accuracy of the amino acid recognition.

INTRODUCTION

The advancement of technologies in recent years makes NMR an important tool in protein structure determination. In the structure determination procedure, the first step is to assign the protein spectra so that distance constraints can be extracted and used in the 3D structure calculations. It is generally accepted that the NMR resonance assignment procedure is tedious and time-consuming. In order to save human labor, our group has developed a computer assisted protein NMR assignment software package.[1−5] At the present time, our program can be used successfully to process both 2D and 3D NMR spectral data and automatically assign backbone and side chain protons.

Following are the logical steps involved in the computer assisted NMR assignment procedure: (1) extracting all possible spin coupling fragments based upon the supplied spectral data; (2) identifying spin coupling topological fragments by a pattern recognition algorithm; (3) sequentially mapping the identified spin coupling patterns to the protein sequence. The detail of the software algorithms can be found in refs 1−3.

In amino acid pattern recognition (step 2), fuzzy mathematics, pattern recognition theory, and a knowledge base of chemical shift distributions of the 20 amino acid types are used to map onto specific amino acid residues all the spin coupling topological graphs created in step 1. Numerical similarity scores of spin coupling patterns are calculated from chemical shift averages and standard deviations of the 20 amino acid types. The chemical shift knowledge base that we are now using does not include secondary structure information, and the expected chemical shift for each amino acid type is the average of all observed data of the same type. Since the ^{1}H and ^{13}C chemical shifts of proteins correlate to the secondary structures,[6−8] including secondary structure information into the chemical shift knowledge base may help to improve the performance of amino acid recognition. In this work, we investigated the feasibility of using neural network predicted secondary structure information of proteins to improve the amino acid type recognition in automated NMR assignment. The predicted protein secondary structures are obtained by using both the neural network systems developed by Rost and Sander[9] and the one developed by us.

THEORY

Background. In CAPRI, the spin coupling fragments are first extracted from the peak lists of input multidimensional NMR spectra. An efficient pattern recognition algorithm is then used to map each fragment to one of the 20 possible amino acid types. The algorithm that is used by CAPRI is the so called fuzzy pattern recognition algorithm (FPRA). FPRA uses HBA (heuristic-back tracking algorithm) to produce a fuzzy mapping set from a query topological space, QG (which represent experimental *J*-coupling topologies), to a fuzzy topological pattern space, SG (which is a set of cluster centers containing the chemical shift and spin coupling topological properties of all 20 amino acid types). The detail of this fuzzy pattern recognition procedure can be found in ref 1.

In fuzzy pattern recognition, every spin coupling network can be represented as a fuzzy graph (FG), A FG is defined as

$$\mathrm{FG} = \{V,\Delta_v,\mu_v,E\} \quad (1)$$

* To whom correspondence should be addressed.
[⊗] Abstract published in *Advance ACS Abstracts,* November 1, 1997.

S0095-2338(97)00012-7 CCC: $14.00

where V is a cluster center which represents a group of chemical shifts, $E \in \{V \times V\}$, and each edge in E represents a spin coupling between two spins i and j. Δ_v is the set of distributions associated with every element in V and μ_v is the membership function set. With consideration of spin i in the coupling network, μ_{vi} can be calculated as

$$\mu_{vi} = 1 - \exp\{-[(V_i - V_{xi})/\Delta_{vi}]^2\} \qquad (2)$$

where V_i is the expected chemical shift value for spin i, V_{xi} is the experimental chemical shift to be assigned, and Δ_{vi} is the deviation of V_i. μ_{vi} is a quality factor which is used to determine whether an experimental frequency value belongs to a given spin system and to determine the quality of this assignment.

In eq 2, we assume that the expected chemical shift values obey Gaussian distributions. However, recent studies show that the chemical shifts of proteins are related to their secondary structures. For example, upfield secondary structure shifts of Hα ranging from −0.15 to −0.60 ppm with a mean of −0.39 ppm had been observed for residues in helical environments and downfield conformational shifts of an average of +0.37 ppm occur for residue in β-sheet conformations. This means that the expected chemical shift values of those spins which are sensitive to the conformation environments cannot be perfectly described by Gaussian distributions as assumed in eq 2. To resolve this problem, we need to know the structural information of the protein being studied. However, the protein structures are usually unknown before the NMR experiments are carried out.

In this work, neural network predicted secondary structure information is included to calculate a membership function. As the protein primary sequence is the only input for the secondary structure prediction, the method can be applied to proteins with unknown structures. However, as the secondary structure prediction is not 100% correct, we need to answer the following questions. Can the predicted secondary structure information improve the amino acid classification? How significant can the improvement be compared to the previous Gaussian distribution assumption?

Secondary Structure Prediction by Neural Networks. Artificial neural networks are mathematical models of biological neural systems. A feed-forward neural network receives a set of input facts. These facts are processed by the network, and a set of output values are given as the response. Before a neural network can be used as a prediction or classification tool, it has to be trained. The most widely used training algorithm is so called back-propagation (BP) proposed by Rumelhart et al.[10] However, the BP training algorithm has been criticized for its slow convergence speed and the possibility of sticking in local minima. In this work, the enhancement technique called RPROP[11,12] is used to speed up the BP learning process. The algorithm uses the adaptive local learning rate in the training process. As local adaptation uses independent learning rates for every adjustable parameter; usually, it is able to find optimal learning rates for every weight, and this can speed up the back-propagation learning speed. Detailed descriptions of the RPROP algorithm can be found in refs 11 and 12. In protein secondary structure prediction, the inputs to the neural networks are usually the amino acid sequences, while the outputs are the predicted secondary structures. For the network inputs, the 20 different amino acid types are each represented by a 21-bit binary number with 20 zeros and a single 1. The extra bit is used to represent the "virtual" amino acids at both ends of the protein sequences. From the output, the secondary structure assignments based on the atomic coordinates can be assigned by the method DSSP of Kabsch and Sander.[13] As the local structure of an amino acid can be greatly influenced by its neighboring residues, in the prediction process, it is important to consider the local neighboring residues that occur close in the sequence. As a consequence, segments of sequences are used as inputs to the neural networks rather than treating isolated amino acids alone. The number of residues in a segment which is presented to the neural networks in each time is known as the window size. The input layer encodes a moving window in the amino acid sequence, and prediction is made for the residues of a specific position in the windows, e.g. the central residues. Instead of using single sequences, Rost and Sander used multiple sequence alignments rather than single sequences as input to the networks.[9] At the training stage, a database of protein families aligned to proteins of known structure was used. At the prediction stage, the database of sequences was scanned for all homologues of the protein to be predicted, and the family profile of amino acid frequencies at each alignment position was fed into the networks. On average, the sustained prediction accuracy increased by 6 percentage points compared to single sequence training. The additional information comes from the fact that the pattern of residue substitutions reflects the folding of the family's protein.

If we consider a three-layer feed-forward neural network with 3 hidden neurons, a window size of 13 and 3 units in the output layer (each representing one of the possible secondary structures), then there will be a total 834 weights or parameters that need to be optimized during the network training. The large number of weights causes the network training to be slow, and there is a high possibility of getting trapped in local minima. In this work, we attempted to reduce the network training time by using reduced representations of amino acids as inputs. Instead of using 21 bits to represent one amino acid residue, we use the physical characteristics of each amino acid as the network inputs. For example, if 3 physical characteristics are used to represent one amino acid, the number of weights needed to optimize can be reduced to 132. However, this may have an adverse effect on the prediction accuracy of the neural networks since some structural information may be missed in the reduced representation.

EXPERIMENTATION

A nonredundant database of 123 protein chains were used to train the feed-forward neural networks.[9] The correspondence protein sequence alignment profiles and secondary structure information were taken from the HSSP protein databank[14] (the HSSP database can be accessed by ftp://ftp.ebi.ac.uk/database/). We classified *H*, *G*, and *I* as helix and E as sheet; residues that are neither helices nor sheets are classified as "coils". The output layer had 3 units, each representing one of the possible secondary structures of the central residue in the specified window size. It was shown by Rahman and Rackovsky that the most important properties for the correlation between sequences and structure spaces are related to helix, β-structure preference, side chain bulk,

Table 1. Prediction Accuracies of the Neural Networks Trained by Different Styles

	Q_{helix}	C_{helix}	Q_{sheet}	C_{sheet}	Q_{coil}	C_{coil}	Q_{total}
nonbalanced	61.73%	0.4714	46.85%	0.4231	77.76%	0.4627	65.99%
partially balanced	58.82%	0.5054	59.91%	0.4492	74.67%	0.4646	66.47%
balanced	64.69%	0.5106	68.29%	0.4499	64.32%	0.4579	65.30%

Table 2. Average $^{13}C\alpha$ Chemical Shift Values and Standard Deviations (ppm) Categorized According to Secondary Structural Assignment[a]

amino acid	overall	helix	sheet	coil
ALA	53.206 (112)	54.430 (67)	50.293 (13)	51.827 (32)
	1.995	1.072	1.334	1.534
CYS	57.286 (32)	58.756 (10)	56.834 (9)	56.467 (13)
	2.930	3.316	1.894	2.977
ASP	54.468 (75)	56.616 (25)	53.190 (7)	53.427 (43)
	2.043	1.206	0.987	1.501
GLU	57.405 (118)	58.944 (65)	53.678 (19)	56.547 (34)
	2.497	1.264	1.070	2.229
PHE	57.693 (67)	60.671 (27)	55.729 (26)	55.598 (14)
	2.867	1.572	1.385	1.491
GLY	45.105 (121)	46.934 (16)	44.129 (16)	44.952 (89)
	1.400	1.197	1.239	1.167
HIS	56.638 (24)	59.085 (8)	56.405 (4)	55.084 (12)
	2.405	1.098	1.392	1.947
ILE	61.863 (67)	64.307 (26)	60.041 (26)	60.784 (15)
	2.861	1.886	2.176	2.297
LYS	56.744 (127)	59.223 (45)	54.429 (19)	55.670 (63)
	2.431	1.241	1.206	1.794
LEU	55.701 (104)	57.339 (51)	53.836 (29)	54.476 (24)
	2.027	1.220	1.289	1.112
MET	56.111 (48)	57.710 (26)	53.865 (11)	54.578 (11)
	2.202	1.562	0.956	1.001
ASN	53.784 (63)	55.668 (24)	50.720 (3)	52.783 (36)
	2.029	0.994	1.526	1.496
PRO	63.162 (50)	65.288 (10)	62.230 (7)	62.715 (33)
	1.776	0.849	1.133	1.605
GLN	56.287 (72)	58.447 (27)	53.681 (15)	55.646 (30)
	2.321	1.520	1.124	1.493
ARG	57.419 (87)	59.103 (47)	53.555 (11)	56.154 (29)
	2.629	1.374	1.720	2.187
SER	58.185 (68)	61.424 (14)	56.309 (20)	57.954 (34)
	2.413	0.986	1.193	1.995
THR	62.092 (112)	65.730 (29)	60.883 (29)	60.788 (54)
	3.112	1.817	2.022	2.571
VAL	62.333 (71)	65.756 (25)	60.259 (28)	60.806 (18)
	3.218	1.771	1.881	2.411
TRP	57.971 (18)	59.400 (12)	53.050 (2)	56.145 (4)
	3.242	2.859	1.485	0.295
TYR	57.523 (44)	60.954 (13)	55.803 (15)	56.349 (16)
	2.784	1.705	1.510	1.771

[a] Data in the first row of each amino acid category are the average chemical shift values in different conformations; the number in the brackets is the number of residues observed. Data in the second row are the standard deviations. Proteins used to compile this database: the salmonella phage P22 c2 repressor (1ADR),[18] the zinc finger Xfin31 from *Xenopus laevis* (1AHD),[19] the pheromone Er-2 from *Euplotes raikovi* (1ERD),[20] FK506-binding protein (1FKB),[21] Grx-SSG- mixed disulfide between *E.coli* glutaredoxin and glutathione (1GRX),[22] Human Interleukin-4 (1ITI),[23] HPr from *E. coli* (1POH),[24] Snase-pdTp-Ca2+, ternary complex among a nuclease from Staphylococcus aureus, 3′,5′-deoxythymidinediphosphate, and calcium (1SNC),[25] Interleukin-1β (2I1B),[26] bacteriophage T4 lysozyme (2LZM),[27] Cyp−CsA complex between human cyclophilin A and cyclosporin A (2RMA),[28] Ribonuclease H from *E. coli* (2RN2),[29] *Drosophila* Calmodulin (4CLN),[30] and BPTI from Bovine (5PTI).[31] The chemical shifts are used as reported in the literature.

and hydrophobicity.[15] We decided to use these four as the physical characteristics to represent each amino acid residue as network inputs. The residue bulk and hydrophobicity data were taken from ref 16. The helix and sheet preferences that we used were simply calculated as follows:

$$H_i = \frac{\alpha_i}{\mathrm{sum}_i} \frac{\sum_i \mathrm{sum}_i}{\sum_i \alpha_i} \tag{3}$$

$$E_i = \frac{\beta_i}{\mathrm{sum}_i} \frac{\sum_i \mathrm{sum}_i}{\sum_i \beta_i} \tag{4}$$

where H_i and E_i are the calculated helix and sheet preferences, α_i and β_i are the number of residues of the specific amino acid type in helix and sheet states, and sum_i is the total number of the specific amino acid types. For every position in a protein sequence, the values of the physical characteristics used are the weighted average of all amino acids in the aligned sequence profiles in that position. As stated by others, the number of hidden neurons used did not have a significant effect on the predictive power of the networks. We arbitrarily chose to use 4 hidden neurons and a typical window size of 13 in our training.

The parameters that were used in RPROP training were $\Delta_0 = 0.1$, $\Delta_{max} = 50.0$, and $\Delta_{min} = 10^{-6}$. The increase and decrease factors are fixed to $\eta^+ = 1.2$ and $\eta^- = 0.5$. The physical meaning of these parameters and the details of the RPROP algorithm can be found in refs 11 and 12. To circumvent the overtraining problem, a subset of these proteins was taken out for the testing and the remaining proteins were used for the training set. As different accuracies may obtained by choosing different testing sets, the whole protein database was divided into seven subsets and we worked with 7-fold cross-validation. Each time, one subset was used for testing while the remaining were used for training. To assess the performance, Q_3, which is the percentage of correctly predicted residues on all 3 types of secondary structure, was used:

$$Q_3 = \frac{P_\alpha + P_\beta + P_{coil}}{N} \tag{5}$$

where N is the total number of predicted residues and P_α, P_β, and P_{coil} are the number of correctly predicted secondary structures of type α-helix, β-sheet, and coil, respectively. In addition to Q_3, Matthews' correlation coefficient was also used to assess the performances.[17] This makes allowance for the possible preponderance of certain structures in the database by punishing overprediction and underprediction of residues. The coefficient for a structure μ is given by

$$\mathrm{Cm} = \frac{p_\mu n_\mu - u_\mu o_\mu}{\sqrt{(n_\mu + u_\mu)(n_\mu + o_\mu)(p_\mu + u_\mu)(p_\mu + o_\mu)}} \tag{6}$$

where p_μ is the number of residues with structure μ predicted correctly, n_μ is the number that do not have structure μ and

Table 3. Results of Similarity Scores Calculations of $^{13}C\alpha$ Chemical Shift According to Equations 7−10

(a)

	no. of residues				XU		DSSP		PHD		
protein	total	helix	sheet	coil	rank[a]	SC[b]	rank[a]	SC[b]	rank[a]	SC[b]	Q_{total}[c] (%)
1ADR	75	43	0	32	6.64	0.678	4.20	0.758	5.43	0.544	68.0
1AHD	68	40	0	28	7.40	0.600	6.03	0.603	6.10	0.581	88.2
1ERD	40	24	0	16	5.68	0.654	3.50	0.717	5.10	0.431	47.5
1FKB	107	11	43	53	5.25	0.749	4.42	0.654	4.39	0.623	86.9
1GRX	85	39	18	28	6.32	0.689	3.92	0.763	4.55	0.674	71.8
1ITI	133	79	4	50	6.80	0.683	4.26	0.795	4.64	0.720	85.7
1POH	85	29	23	33	5.95	0.722	3.93	0.743	3.99	0.688	74.1
1SNC	88	21	38	29	6.56	0.583	4.85	0.618	5.85	0.485	64.8
2I1B	152	7	72	73	6.27	0.699	4.70	0.715	4.77	0.666	79.6
2LZM	164	109	15	40	7.05	0.626	4.13	0.707	4.98	0.595	74.4
2RN2	122	41	38	43	6.63	0.544	5.75	0.446	6.03	0.391	77.9
4CLN	145	90	4	51	6.29	0.629	4.32	0.733	4.28	0.715	91.0
5PTI	54	11	13	30	5.69	0.656	4.35	0.639	3.96	0.641	81.5
overall	1318	544	257	517	6.41	0.654	4.52	0.686	4.91	0.609	78.3

(b)

	balanced			partially balanced			nonbalanced		
protein	rank	SC	Q_{total} (%)	rank	SC	Q_{total} (%)	rank	SC	Q_{total} (%)
1ADR	5.76	0.508	60.0	5.69	0.524	64.0	5.77	0.519	66.7
1AHD	6.91	0.477	64.7	6.68	0.495	77.9	6.90	0.479	67.6
1ERD	4.80	0.453	47.5	4.80	0.461	52.5	4.88	0.475	52.5
1FKB	4.40	0.538	79.4	4.53	0.547	72.9	4.42	0.575	73.8
1GRX	4.58	0.607	74.1	4.71	0.593	76.5	4.60	0.608	76.5
1ITI	5.32	0.570	62.4	5.26	0.569	70.7	5.27	0.583	68.4
1POH	4.56	0.588	68.2	4.53	0.593	70.6	4.60	0.590	68.2
1SNC	5.75	0.450	59.6	5.60	0.469	61.8	5.88	0.442	59.6
2I1B	5.36	0.536	64.5	5.20	0.547	66.4	5.38	0.535	58.6
2LZM	5.23	0.529	68.3	5.18	0.529	65.9	5.22	0.537	70.1
2RN2	6.15	0.352	49.2	6.26	0.362	50.0	6.14	0.363	56.6
4CLN	4.80	0.606	82.1	4.68	0.598	84.8	4.74	0.608	83.4
5PTI	4.39	0.550	75.9	4.43	0.571	72.2	4.50	0.357	74.1
overall	5.24	0.525	66.7	5.20	0.531	68.7	5.25	0.525	68.0

[a] The average ranking score of the correct amino acid types in the candidate lists. For example, the candidate list of a Ile residue in a coil conformation is found to be Asn, His, Ile, Thr, ... in descending order according to the magnitude of the similarity scores. Hence, the ranking score for this residue is 3. [b] The average similarity scores of the correct amino acid types in the candidate lists. [c] Q_{total} is the overall secondary structure prediction accuracy of the protein by the PHD method.

were correctly rejected, u_μ is the number of underpredicted cases, and o_μ is the number of overpredicted cases. The coefficient ranges are between −1.0 and +1.0, with +1.0 representing prefect correlation.

The neural network system consists of two levels. The first level is a three layered neural network with 4 hidden neurons. To train this network, protein sequence segments encoded with physical characteristics were used as inputs, while the outputs are the secondary structures of central residues in the segments. As 4 characteristics were used and the window size was 13, the resulting network was 52−4−3 in size.

A second neural network which took the correlation between the consecutive patterns of secondary structure in protein sequences was used. The predicted outputs of the first network were used as inputs to the second network, while the outputs still represent secondary structures of the central residues in the segments. The window size used in the second network was the same as that used in the first network so that the second network was 39−4−3 in size.

Although an extensive compilation of chemical shift data in proteins was published by Wishart et al.,[6] some statistical information, such as variance, was missed. In this work, we used our own compiled $^{13}C\alpha$, $^{1}H\alpha$, and ^{1}HN chemical shift databases in the calculation of the similarity scores.

Some proteins with known NMR chemical shifts and structures are used as tests. The secondary structures of the proteins were defined by using the DSSP method, and the homology alignment profiles were retrieved from the HSSP

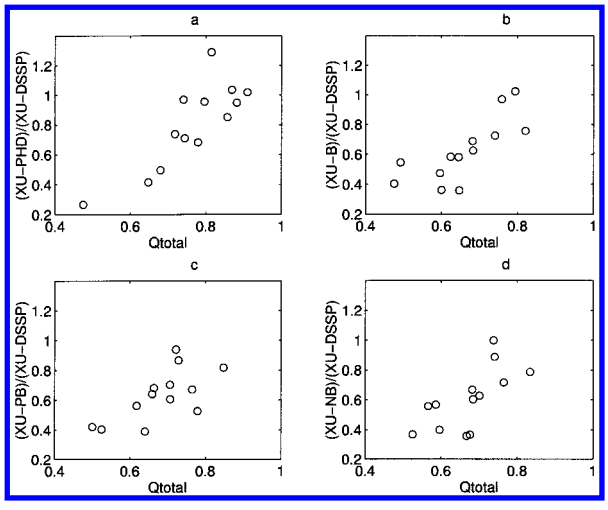

Figure 1. (a) Comparison of XU, DSSP, and PHD, the average ranks of the amino acid classifications, by using no secondary structure information, secondary structure information from the DSSP mehtod, and information from the PHD prediction, respectively, with Q_{total}, the overall prediction accuracy. (b) Balanced (B), using secondary structure information predicted by neural networks using balanced training. (c) PB, using secondary structure information predicted by neural networks using partially balanced training. (d) NB, using secondary structure information predicted by neural networks using nonbalanced training.

database. The protein primary sequences were then submitted to the PredictProtein Server (http://www.embl-heidelberg.de/predictprotein/) by electronic mail in order to receive the secondary structure prediction results by the PHD method. We also used our trained neural networks to make the prediction. For each chemical shift data, the similarity scores defined as follows were calculated.

$$(xu)_{ki} = \exp(-2(x_k - \bar{x}_i)^2/2\sigma_i^2) \tag{7}$$

$$(\mathrm{dssp})_{ki} = \exp(-2(x_k - \bar{x}_{is})^2/2\sigma_{is}^2) \tag{8}$$

$$(\mathrm{phd})_{ki} = \frac{\sum_{s=1}^{3} \mathrm{pr}_s \exp(-2(x_k - \bar{x}_{is})^2/2\sigma_{is}^2)}{\sum_{s=1}^{3} \mathrm{pr}_s} \tag{9}$$

$$(\mathrm{psnn})_{ki} = \frac{\sum_{s=1}^{3} o_s \exp(-2(x_k - \bar{x}_{is})^2/2\sigma_{is}^2)}{\sum_{s=1}^{3} o_s} \tag{10}$$

where $(xu)_{ki}$ is the similarity score for classifying the chemical shift x_k to the amino acid type i, where $i \in$ (Ala, Cys, Asp, ..., Trp), with the average chemical shift $\bar{x}_i$ over all conformations and σ_i is the standard deviation of the chemical shifts of all amino acids i.

This equation is used in the similarity calculation of the FPRA algorithm. The quantities $(\mathrm{dssp})_{ki}$, $(\mathrm{psnn})_{ki}$, and $(\mathrm{phd})_{ki}$ are the similarity scores for classifying the chemical shift x_k to the amino acid type i with secondary s. S is the secondary structure predicted by DSSP, by our trained neural networks and PHD, respectively. $\bar{x}_{is}$ is the average chemical shift of amino acid type i in conformation s, $s \in$ (helix, sheet, coil); σ_{is} is the standard deviation of chemical shift values of amino acid i in conformation s; o_s is the value of the output unit of conformation s; and pr_s in eq 9 is the probability of assigning the residue to conformation s reported by the PHD server. For example, $\mathrm{pr}_{\mathrm{helix}} = 0.5$ means the output node for the helix is about 0.5−0.6. The amino acid candidates were then ranked according to the similarity scores. Due to the overlapping of the chemical shift ranges of different amino acid types, the correct one may not come up to be the one with the highest similarity score. In order to evaluate the performance of the classifications, the average ranking of the correct amino acid types of the chemical shift data were calculated.

RESULTS AND DISCUSSIONS

RPROP was used to train the neural networks in three different ways: nonbalanced training, partially balanced training, and balanced training. The results are shown in Table 1. Our protein database is composed of 31.43% helix, 21.78% sheet, and 46.79% coil. Since the percentages of helix and sheet are lower than that for coil, the coil has a higher prediction accuracy. In the balanced training, for each training pattern, the weight correction terms are multiplied by a balancing factor, which is equal to the total number of coil residues in the training set, divided by the total number of residues having the same conformation as the output of the training pattern. In partially balanced training, the balancing factors for helix and coil are set equal to 1, while, for β-sheets, the balancing factor is the number of helix residues divided by the number of β-sheet residues. The prediction accuracies for three conformations are not the same in each style of training, as shown in Table 1. We compared the prediction accuracy of the our neural networks using a reduced representation with the PHD networks as reported by Rost and Sander. It was found that the prediction accuracy is lowered by 1.5% compared to the PHD networks if the jury decision system is not considered. On the other hand, if the reduced representation is not used, the size of the networks are reduced by a factor of 5.

The statistical results of the compiled ^{13}Cα chemical shifts database are shown in Table 2. As the chemical shift data available in literature are limited, rather than using another

Table 4. Average ^{1}Hα Chemical Shift Values and Standard Deviations (ppm) Categorized According to Secondary Structural Assignment[a]

amino acid	overall	helix	sheet	coil
ALA	4.226 (191)	4.011 (91)	4.853 (25)	4.279 (75)
	0.479	0.442	0.442	0.308
CYS	4.771 (80)	4.370 (9)	4.942 (24)	4.760 (47)
	0.521	0.569	0.700	0.344
ASP	4.610 (141)	4.452 (49)	5.034 (16)	4.623 (76)
	0.298	0.182	0.255	0.277
GLU	4.311 (159)	4.084 (69)	4.910 (26)	4.312 (64)
	0.470	0.322	0.464	0.390
PHE	4.722 (92)	4.185 (30)	5.309 (21)	4.815 (41)
	0.594	0.382	0.474	0.424
GLY	3.971 (183)	3.824 (20)	4.034 (22)	3.982(141)
	0.338	0.212	0.569	0.301
HIS	4.665 (43)	4.442 (8)	4.784 (7)	4.698 (28)
	0.650	0.369	0.844	0.666
ILE	4.201 (111)	3.642 (36)	4.671 (35)	4.293 (40)
	0.707	0.277	0.793	0.546
LYS	4.248 (171)	3.989 (63)	4.666 (27)	4.310 (81)
	0.497	0.296	0.583	0.480
LEU	4.262 (164)	4.026 (83)	4.654 (25)	4.436 (56)
	0.387	0.247	0.330	0.348
MET	4.371 (41)	4.210 (24)	4.923 (4)	4.497 (13)
	0.417	0.344	0.496	0.347
ASN	4.647 (95)	4.370 (27)	5.010 (10)	4.713 (58)
	0.387	0.283	0.591	0.302
PRO	4.418 (75)	4.176 (9)	4.732 (4)	4.433 (62)
	0.398	0.294	0.189	0.405
GLN	4.324 (80)	3.981 (32)	5.021 (19)	4.245 (29)
	0.523	0.303	0.438	0.274
ARG	4.280 (87)	3.926 (33)	4.919 (16)	4.319 (38)
	0.551	0.363	0.471	0.460
SER	4.597 (130)	4.192 (19)	5.127 (24)	4.539 (87)
	0.450	0.338	0.470	0.319
THR	4.468 (146)	3.914 (28)	4.958 (41)	4.408 (77)
	0.551	0.336	0.466	0.424
VAL	4.099 (170)	3.552 (55)	4.569 (60)	4.133 (55)
	0.576	0.316	0.425	0.436
TRP	4.668 (36)	4.392 (13)	5.282 (12)	4.324 (11)
	0.665	0.248	0.521	0.703
TYR	4.656 (87)	4.242 (21)	5.055 (34)	4.502 (32)
	0.538	0.365	0.461	0.418

[a] Data in the first row of each amino acid category are the average chemical shift values in different conformations; the number in the brackets are the number of residues observed. Data in the second row are the standard deviations. Only the PDB code and the references from which the chemical shifts data were retrieved are listed (the chemical shifts are used as reported in the literature): 1APC,[32] 1PUT,[33] 1TPK,[34] 2RN2,[29] 2ETI,[35] 1ZRP,[36] 1GCN,[37] 1LAC,[38] 2ABD,[39] 1TTF,[40] 1EGR,[41] 3B5C,[42] 2INS,[43] 1HIU,[44] 1HCC,[45] 1GF1,[46] 4TGF,[47] 3RN3,[48] 1APS,[49] 1SH1,[50] 1ATX,[51] 2LZM,[27] 1ACP,[52] 2PAS,[53] 1EGL,[54] 1POH,[24] 2MLT,[55] 1BHB,[56] 1BCT,[57] 2AIT,[58] 2RMA,[28] 1PDC,[59] and 2I1B.[26]

Table 5. Results of Similarity Scores Calculations of the ^{1}Hα Chemical Shift According to Equations 7−9

	no. of residues				XU		DSSP		PHD		
protein	total	helix	sheet	coil	rank[a]	SC[b]	rank[a]	SC[b]	rank[a]	SC[b]	Q_{total}[c] (%)
1ACP	76	39	0	37	7.30	0.816	7.14	0.810	7.86	0.734	76.3
1APC	104	72	0	32	7.15	0.846	7.09	0.805	7.02	0.780	56.7
2I1B	151	7	69	75	9.58	0.664	8.88	0.725	9.11	0.667	78.8
1EGL	53	9	13	31	8.28	0.761	8.02	0.740	8.45	0.678	69.8
1EGR	84	31	20	33	8.54	0.745	7.93	0.748	8.01	0.726	82.1
1GCN	29	14	0	15	7.93	0.864	7.66	0.800	7.28	0.830	69.0
1POH	85	29	23	33	8.24	0.726	7.86	0.769	8.21	0.710	74.1
2ABD	85	49	0	36	8.25	0.777	7.93	0.766	7.61	0.743	88.2
2AIT	74	0	30	44	8.99	0.728	8.77	0.760	9.65	0.671	66.2
2LZM	157	105	13	39	7.66	0.752	6.65	0.782	6.95	0.725	74.5
2RN2	155	55	44	56	8.37	0.674	7.57	0.725	8.11	0.662	81.9
overall	1053	410	212	431	8.26	0.743	7.73	0.762	8.02	0.711	75.3

Table 6. Average ^{1}HN Chemical Shift Values and Standard Deviations (ppm) Categorized According to Secondary Structural Assignment[a]

amino acid	overall	helix	sheet	coil
ALA	8.162 (197)	8.048 (92)	8.530 (27)	8.169 (78)
	0.674	0.646	0.746	0.643
CYS	8.466 (85)	7.929 (11)	8.757 (27)	8.425 (47)
	0.736	0.792	0.737	0.653
ASP	8.416 (153)	8.360 (53)	8.701 (17)	8.393 (83)
	0.627	0.659	0.378	0.636
GLU	8.361 (171)	8.300 (70)	8.590 (30)	8.324 (71)
	0.628	0.647	0.456	0.657
PHE	8.404 (106)	8.162 (33)	9.099 (30)	8.105 (43)
	0.762	0.564	0.540	0.725
GLY	8.330 (192)	8.502 (21)	8.078 (25)	8.349 (146)
	0.802	0.948	0.876	0.762
HIS	8.366 (44)	8.005 (8)	8.529 (7)	8.426 (29)
	0.895	0.419	0.568	1.034
ILE	8.345 (119)	8.186 (36)	8.517 (41)	8.314 (42)
	0.665	0.537	0.676	0.727
LYS	8.113 (191)	7.996 (64)	8.510 (30)	8.067 (97)
	0.690	0.653	0.780	0.647
LEU	8.188 (182)	8.045 (84)	8.644 (40)	8.080 (58)
	0.771	0.845	0.651	0.608
MET	8.238 (46)	8.262 (26)	8.594 (8)	7.928 (12)
	0.516	0.430	0.518	0.558
ASN	8.245 (108)	8.209 (27)	8.675 (14)	8.170 (67)
	0.761	0.597	0.744	0.803
GLN	8.303 (92)	8.089 (33)	8.665 (22)	8.280 (37)
	0.647	0.520	0.811	0.559
ARG	8.246 (98)	8.160 (37)	8.489 (19)	8.212 (42)
	0.658	0.581	0.807	0.639
SER	8.361 (140)	8.048 (19)	8.574 (32)	8.351 (89)
	0.641	0.397	0.612	0.668
THR	8.286 (154)	8.163 (29)	8.698 (46)	8.091 (79)
	0.753	0.587	0.754	0.719
VAL	8.238 (181)	7.969 (55)	8.597 (68)	8.071 (58)
	0.737	0.570	0.728	0.731
TRP	8.296 (37)	7.941 (13)	8.771 (12)	8.207 (12)
	0.611	0.477	0.505	0.560
TYR	8.482 (97)	8.320 (22)	8.747 (40)	8.282 (35)
	0.706	0.515	0.739	0.690

[a] Data in the first row of each amino acid category are the average chemical shift values in different conformations; the number in the brackets are the number of residues observed. Data in the second row are the standard deviations. Only the PDB code and the references from which the chemical shifts data were retrieved are listed (the chemical shifts are used as reported in the literature): 1APC,[32] 1PUT,[33] 1TPK,[34] 2RN2,[29] 2ETI,[35] 5PTI,[31] 1ZRP,[36] 1GCN,[37] 1LAC,[38] 2ABD,[39] 1TTF,[40] 1EGR,[41] 3B5C,[42] 2INS,[43] 1HIU,[44] 1HCC,[45] 1GF1,[46] 4TGF,[47] 3RN3,[48] 1APS,[49] 1SH1,[50] 1ATX,[51] 2LZM,[27] 1ACP,[52] 2PAS,[53] 1EGL,[54] 1POH,[24] 2MLT,[55] 1BHB,[56] 1BCT,[57] 2AIT,[58] 2RMA,[28] 1PDC,[59] and 2I1B.[26]

set of chemical shift data that were not used in compiling the database, we used a jackknife method for the similarity score calculation. The chemical shift data of a particular protein were not used in the database compilation whenever the similarity scores of the chemical shifts of that protein were being calculated. The similarity scores were calculated according to eqs 7−10. Table 3 shows the results of such calculations.

The results clearly demonstrate that amino acid type classification can be significantly improved if the secondary structure information is taken into consideration. Due to large variations of chemical shifts in different environments and overlapping of chemical shift regions of different amino acids, it is impossible to get an exact classification based solely on chemical shift data. However, the average rank acts as an indicator that secondary structure information must be included in the amino type determination. Using the average chemical shifts and standard deviations of amino acids over all conformations, the average rank of the correct amino acid types in the candidate lists is about 6.4. This value shows that the ^{13}C chemical shift ranges of the 20 amino acid types are partially overlapped. If the "correct" secondary structure (DSSP) of the proteins is considered, the average rank of the correct amino acid type improved to about 4.5, an improvement in average rank of about 30%. On the other hand, if only the predicted secondary structures are used in the calculation, the improvement is smaller. This is expected because the neural network prediction accuracy is not 100%, but this still gives a significant improvement in the average ranks. From Figure 1a−d, we found that the improvement in the average ranks is directly proportional to the overall prediction accuracy (Q_{total}). Although different styles of training give different prediction accuracies for the three conformations, the average ranks for the nonbalanced, partially balanced and the balanced training are almost the same, being around 5.2. The improvement in average rank is about 19%. The insignificant difference between the three training methods implies that the style of training has little effect on the final amino acid classification results as long as there is not a great different between the overall prediction accuracies. The recently updated PHD protein prediction server uses an even larger protein database for the network training; this leads to a higher prediction accuracy. The results in Table 3 show that the average rank is about 4.9 if the secondary structure prediction results from the PHD protein prediction server are used. This is about 0.3 higher than using our own network results.

There are several points that warrant discussion. First, the ^{13}Cα database that we compiled can be further improved. Due to the limited data available in the literature, the composition of the helix in the database is relatively higher than normal, while the composition of the β-strand is

Table 7. Results of Similarity Scores Calculations of the ^{1}HN Chemical Shift According to Equations 7−9

	no. of residues				XU		DSSP		PHD		
protein	total	helix	sheet	coil	rank[a]	SC[b]	rank[a]	SC[b]	rank[a]	SC[b]	Q_{total}[c] (%)
1ACP	73	39	0	34	9.93	0.754	8.68	0.748	9.08	0.728	75.3
1EGL	53	9	13	31	9.89	0.744	8.92	0.761	9.40	0.736	69.8
1EGR	79	31	20	28	7.66	0.764	8.32	0.768	7.72	0.758	82.3
1GCN	27	14	0	13	9.33	0.861	7.75	0.875	8.70	0.859	66.7
1POH	82	28	23	31	9.11	0.693	8.43	0.710	8.59	0.701	73.2
2ABD	82	49	0	33	8.65	0.740	8.41	0.721	7.90	0.722	87.8
2AIT	70	0	30	40	10.20	0.724	8.66	0.748	8.74	0.704	64.3
2LZM	160	106	14	40	9.29	0.738	8.78	0.720	8.80	0.744	74.4
2RN2	148	54	43	51	8.61	0.725	9.29	0.725	8.35	0.716	82.4
2I1B	144	7	67	70	9.56	0.699	9.17	0.752	9.00	0.705	79.2
overall	918	337	210	371	9.15	0.732	8.79	0.740	8.63	0.721	77.0

relatively low. This leads to the average chemical shift values for the coil conformation being systematically smaller than the average values over all conformation. We investigated the effect of this to the XU similarity scores calculations. Without use of the jackknife method and simply compilation of the database using all proteins, the average rank for XU is 6.14, which is only about 5% smaller than that obtained by the jackknife method. The average rank becomes 6.25 if the average "coil" shifts and standard deviations are used. We also used the average coil shifts with the "overall" standard deviations, and this approach gave the average rank of 6.48. The differences between these three results are not significant. From these results we can conclude that imperfections in the database do not significantly affect the final results and the secondary structure is the factor which leads to improvements in the average rank. In brief, the better the prediction accuracy of the method used in secondary structure prediction, the larger will be the improvement. We conclude that the average ranks are always higher compared to XU whenever neural network predicted secondary structure information is used, and the extent of improvement is proportional to the prediction accuracy (Figure 1). The database can be continually improved as more data become available.

Second, when we compared the average similarity scores of each protein obtained by the different methods, all appear in a narrow range and DSSP has larger average similarity scores than XU. From Table 2, we find all amino acid types have larger chemical shift standard deviations in the overall category than in various secondary structure conformations. These imply that the absolute differences between the observed chemical shift values and the expected chemical shift values are actually smaller when the secondary structures are taken into account.

We compiled the ^{1}Hα chemical shift database and did a similar analysis on the chemical shift data. For Gly, which has two α protons in each residue, we simply used the average chemical shift values of the two. The statistical results of the database compilation are shown in Table 4. Since the ^{1}Hα chemical shift database is larger in size than the ^{13}Cα database, we decided not to use the jackknife method. This means that the data in Table 4 was used throughout the similarity calculation for the ^{1}Hα data. Some proteins were chosen, and the results of similarity calculations are shown in Table 5. For ^{1}Hα, the average ranks calculated by the three methods are far lower as compared to those for ^{13}Cα. This clearly demonstrates that ^{1}Hα chemical shifts are not as good for amino acid classification as ^{13}Cα. The average rank of about 8.2 for the XU method shows that the chemical shift ranges for the 20 amino acid types are actually heavily overlapped, compared to the average rank equal to 10 for those totally overlapped. If the similarity scores are calculated by using eqs 8 and 9, the average ranks improve by only 0.5 and 0.25, respectively. Due to the serious overlapping of proton chemical shift ranges of the 20 amino acids, the use of secondary structure information does not improve the amino acid classification. Similar results are observed for ^{1}HN (Tables 6 and 7). Besides the secondary structure conformation, the chemical shifts of ^{1}HN in proteins are also sensitive to hydrogen bonding. This causes the ^{1}HN chemical shifts to have larger standard deviations. It is not surprising that the average rank of ^{1}HN chemical shifts is closer to 10 than ^{1}Hα, and the inclusion of secondary structure information cannot further improve the amino acid classification.

In this work, we only focused on the ^{13}Cα, ^{1}Hα, and ^{1}HN chemical shifts which are shown to have high correlations to the protein's secondary structures. For those spins such as ^{1}Hβ and ^{13}Cβ which have a small chemical shift difference between various amino acids and conformations, secondary structure information is not expected to lead to an improved amino acid classification. For ^{15}N, which are sensitive to hydrogen bonding in proteins, they are also greatly affected by neighboring effects. These factors blur the secondary structure dependency of ^{15}N chemical shifts. In conclusion, ^{13}Cα chemical shifts are more suitable for the amino acid classification than ^{1}Hα or ^{1}HN's. However, the topologies of the ^{1}H spin coupling network is still crucial in the amino acid classification. The classification process can be further improved by using the secondary structure information obtained from known crystal structures or predicted by neural networks. As the input for the neural network prediction is simply the primary sequence, this method can be implemented to a computer-assisted NMR assignment program. Investigations into how to combine the chemical shift data, secondary structure information, and spin topologies for amino acid pattern recognition are in progress.

ACKNOWLEDGMENT

This work is supported by NSERC, Natural Science and Engineering Research Council of Canada, operating and collaborating grants.

REFERENCES AND NOTES

(1) Xu, J.; Straus, S. K.; Sanctuary, B. C. Automation of Protein 2D Proton NMR Assignment by Means of Fuzzy Mathematics and Graph Theory. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 668−682.

(2) Xu, J.; Sanctuary, B. C. CPA: Constrained Partitioning Algorithm for Initial Assignment of Protein ^{1}H Resonances from MQF-COSY. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 490−500.
(3) Xu, J.; Sanctuary, B. C.; Gray, B. N. Automated Extraction of Spin Coupling Topologies from 2D NMR Correlation Spectra for Protein ^{1}H Resonance Assignment. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 475−489.
(4) Xu, J.; Straus, S. K.; Sanctuary, B. C.; Trimble, L. Use of Fuzzy Mathematics for Complete Automated Assignment of Peptide 1H 2D NMR Spectra. *J. Magn. Reson. B* **1994**, *103*, 53−58.
(5) Li, K. B.; Sancutary, B. C. Automated Extracting of Amino Acid Spin Systems in Proteins Using 3D HCCH-COSY/TOCSY Spectroscopy and Constrained Partitioning Algorithm (CPA). *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 585−593.
(6) Wishart, D. S.; Sykes, B. D.; Richards, F. M. Relationship between Nuclear Magnetic Resonance Chemical Shift and Protein Secondary Structure. *J. Mol. Biol*. **1991**, *222*, 311−333.
(7) Spera, S.; Bax, Ad. Empirical Correlation between Protein Backbone Conformation and Cα and Cβ ^{13}C Nuclear Magnetic Resonance Chemical Shifts. *J. Am. Chem. Soc*. **1991**, *113*, 5490−5492.
(8) Szilagyi, L. Chemical shifts in proteins come of age. *Prog. Nucl. Magn. Reson. Spectrosc*. **1995**, *27*, 325−443.
(9) Rost, B.; Sander, S. Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J. Mol. Biol*. **1993**, *232*, 584−599.
(10) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533−536.
(11) Fahlman, S. E. An Empirical Study of Learning Speed in Back-Propagation Networks. Technical Report CMU-CS-88-162.
(12) Riedmiller, M.; Braun, H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. *Proc. IEEE Int. Conf. Neural Networks* **1993**, 586−591.
(13) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structures: Pattern of Hydrogen-bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577−2637.
(14) Schneider, R.; Sander, C. Database of Homology Derived Protein Structures and the Structural Meaning of Sequence Alignment. *Proteins* **1991**, *9*, 56−68.
(15) Rahman, R. S.; Rackovsky, S. Protein Sequence Randomness and Sequence/Structure Correlations. *Biophys. J*. **1995**, *68*, 1531−1539.
(16) Argos, P. A Sensitive Procedure to Compare Amino Acid Sequences. *J. Mol. Biol*. **1987**, *193*, 385−396.
(17) Matthews, B. W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.
(18) Sevilla-Sierra, P.; Otting, G.; Wuthrich, K. Determination of the Nuclear Magnetic Resonance Structure of the DNA-Binding Domain of the P22 C2 Repressor (1−76) in Solution and Comparison with the DNA-Binding Domain of the 434 Repressor. *J. Mol. Biol*. **1994**, *235*, 1003−1020.
(19) Qian, Y. Q.; Otting, G.; Billeter, M.; Müller, M.; Gehring, W.; Wüthrich, K. Nuclear Magnetic Resonance Spectroscopy of a DNA Complex with the Uniformly 13C-labeled Antennapedia Homeodomain and Structure Determination of the DNA-bound Homeodomain. *J. Mol. Biol*. **1993**, *234*, 1070−83.
(20) Ottiger, M.; Szyperski, T.; Luginbuhl, P.; Ortenzi, C.; Luporini, P.; Bradshaw, R. A.; Wüthrich, K. The NMR Solution Structure of the Pheromone Er-2 from the Ciliated Protozoan Euplotes Raikovi. *Protein Sci*. **1994**, *3*, 1515−1526.
(21) Xu, R. X.; Nettesheim, D.; Olejniczak, E. T.; Meadows, R.; Gemmecker, G.; Fesik, S. W. ^{1}H, ^{13}C and ^{15}N Assignments and Secondary Structure of the FK506 Binding Protein When Bound to Ascomycin. *Biopolymers* **1993**, *33*, 535−550.
(22) Büshweller, J. H.; Holmgren, A.; Wüthrich, K. Biosynthetic ^{15}N and ^{13}C Isotope Labelling of Glutathione in the Mixed Disulfide with *Escherichia Coli* Glutaredoxin Documented by Sequence-Specific NMR Assignments. *Eur. J. Biochem*. **1993**, *218*, 327−334.
(23) Powers, R.; Garrett, D. S.; March, C. J.; Frieden, E. A.; Gronenborn, A. M. Clore, G. M. ^{1}H, ^{15}N, ^{13}C, and ^{13}CO Assignments of Human Interleukin-4 Using Three-Dimensional Double- and Triple-Resonance Heteronuclear Magnetic Resonance Spectroscopy. *Biochemistry* **1992**, *31*, 4334−4346.
(24) van Nuland, N. A.; van Dijk, A. A.; Dijkstra, K.; van Hoesel, F. H. J.; Scheek, R. M.; Robillard, G. T. Three-dimensional ^{15}N-^{1}H-^{1}H and ^{15}N-^{13}C-^{1}H nuclear-magnetic resonance studies of HPr a central component of the phospho*enol*pyruvate-dependent phosphotransferase system from *Escherichia coli*. *Eur. J. Biochem*. **1992**, *203*, 483−491.
(25) Wang, J. F.; Hinck, A. P.; Loh, S. N.; Markley, J. L. Two-Dimensional NMR Studies of Staphylococcal Nuclease. 2. Sequence-Specific Assignments of Carbon-13 and Nitrogen-15 Signals from the Nuclease H124L−Thymidine 3′,5′-bisphosphate−Ca^{2+} Ternary Complex. *Biochemistry* **1990**, *29*, 102−113.
(26) Clore, G. M.; Bax, Ad; Drisccoll, P. C.; Wingfield, P. T.; Gronenborn, A. M. Assignment of the Side-Chain ^{1}H and ^{13}C Resonances of Interleukin-1β Using Double- and Triple-Resonance Heteronuclear Three-Dimensional NMR Spectroscopy. *Biochemistry* **1990**, *29*, 8172−8184.
(27) Fischer, M. W. F.; Majumdar, A.; Dahlquist, F. W.; Zuiderweg, E. R. P. ^{15}N, ^{13}C, ^{1}H NMR Assignments and Secondary Structure for T4-Lysozyme. *J. Magn. Reson. B* **1995**, *108*, 143−154.
(28) Neri, P.; Meadows, R.; Gemmecker, G.; Olejniczak, E.; Nettesheim, D.; Logan, T.; Simmer, R.; Helfrich, R.; Holzman, T.; Severin, J.; Fesik, S. ^{1}H, ^{13}C and ^{15}N Backbone Assignments of Cyclophilin When Bound to Cyclosporin A (CsA) and Preliminary Structural Characterization of the CsA Binding Site. *FEBS Lett*. **1991**, *294*, 81−88.
(29) Yamazaki, T.; Yoshida, M.; Kanaya, S.; Nakamura, H.; Nagayama, K. Assignments of Backbone ^{1}H, ^{13}C, and ^{15}N Resonances and Secondary Structure of Ribonuclease H from *Escherichia coli* by Heteronuclear Three-Dimensional NMR Spectroscopy. *Biochemistry* **1991**, *30*, 6036−6047.
(30) Ikura, M.; Kay, L. E.; Bax, Ad. A Novel Approach for Sequential Assignment of ^{1}H, ^{13}C, and ^{15}N Spectra of Larger Proteins: Heteronuclear Triple-Resonance Three-Dimensional NMR Spectroscopy. Application to Calmodulin. *Biochemistry* **1990**, *29*, 4659−4667.
(31) Wagner, G.; Brühwiler, D. Toward the Complete Assignment of the Carbon Nuclear Magnetic Resonance Spectrum of the Basic Pancreatic Trypsin Inhibitor. *Biochemistry* **1986**, *25*, 5838−5843.
(32) Feng, Y. Q.; Wand, A. J.; Sligar, S. G. ^{1}H and ^{15}N NMR Resonance Assignments and Preliminary Structural Characterization of *Escherichia coli* Apocytochrome b562. *Biochemistry* **1991**, *31*, 7711−7717.
(33) Pochapsky, T. C.; Ye, X. M. ^{1}H NMR Identification of a β-Sheet Structure and Description of Folding Topology in Putodaredoxin. *Biochemistry* **1991**, *30*, 3850−3856.
(34) Byeon, I. J.; Kelley, R. F.; Llinas, M. Kringle-2 Domain of the Tissue-type Plasminogen Activator. *Eur. J. Biochem*. **1991**, *197*, 155−165.
(35) Heitz, A.; Chiche, L.; Le-Nguyen, D.; Castro, B. 1H 2D NMR and Distance Geometry Study of the Folding of *Ecballium elaterium* Trypsin Inhibitor, a Member of the Squash Inhibitors Family. *Biochemistry* **1989**, *28*, 2392−2398.
(36) Blake, P. R.; Park, J. B.; Bryant, F. O.; Aono, S.; Magnuson, J.; Eccleston, E.; Howard, J. B.; Summers, M. F.; Adams, M. W. Determinants of Protein Hyperthermostability: Purification and Amino Acid Sequence of Rubredoxin from the Hyperthermophilic Archaebacterium *Pyrococcus furiosus* and Secondary Structure of the Zinc Adduct by NMR. *Biochemistry* **1991**, *30*, 10885−10895.
(37) Wider, G.; Lee, K. H.; Wüthrich, K. Sequential Resonance Assignments in Protein ^{1}H Nuclear Magnetic Resonance Spectra (Glucagon Bound to Perdeuterated Dodecylphosphocholine Micelles). *J. Mol. Biol*. **1982**, *155*, 367−388.
(38) Dardel, F.; Laue, E. D.; Perham, R. N. Sequence-specific ^{1}H-NMR Assignments and Secondary Structure of the Lipoyl Domain of the Bacillus Sterarothermophilus Pyruvate Dehydrogenase Multienzyme Complex. *Eur. J. Biochem*. **1991**, *201*, 203−209.
(39) Andersen, K. V.; Ludvigsen, S.; Mandrup, S.; Knudsen, J.; Poulsen, F. U. The Secondary Structure in Solution of Acyl−Coenzyme A Binding Protein from Bovine Liver Using ^{1}H Nuclear Magnetic Resonance Spectroscopy. *Biochemistry* **1991**, *30*, 10654−10663.
(40) Baron, M.; Main, A. L.; Driscoll, R. C.; Mardon, H. J.; Boyd, J.; Campbell, I. D. ^{1}H NMR Assignment and Secondary Structure of the Cell Adhesion Type III Module of Fibronectin. *Biochemistry* **1992**, *31*, 2068−2073.
(41) Sodano, P.; Xia, T. H.; Büshweller, J. H.; Bjornberg, O.; Holmgren, A.; Billeter, M.; Wüthrich, K. Sequence-specific ^{1}H NMR Assignments and Determination of the Three-dimensional Structure of Reduced *Escherichia coli* Glutaredoxin. *J. Mol. Biol*. **1991**, *221*, 1311−1324.
(42) Veitch, N. C.; Whitford, D.; Williams, R. J. An analysis of pseudocontact shifts and their relationship to structural features of the redox states of cytochrome b5. *FEBS Lett*. **1990**, *269*, 297−304.
(43) Higgins, K. A.; Craik, D. J.; Hall, J. G. 2D ^{1}H NMR Studies of Monomeric Insulin. *Biochem. Int*. **1990**, *22*, 627−637.
(44) Kline, A. D.; Justice, R. M. J. Complete Sequence-Specific ^{1}H NMR Assignments for Human Insulin. *Biochemistry* **1990**, *29*, 2906−2913.
(45) Barlow, P. N.; Baron, M.; Norman, D. G.; Day, A. J.; Willis, A. C.; Sim, R. S.; Campbell, I. D. Secondary Structure of a Complement Control Protein Module by Two-Dimensional ^{1}H NMR. *Biochemistry* **1991**, *30*, 997−1004.
(46) Sato, A.; Nishimura, S.; Ohkubo, T.; Kyogoku, Y.; Koyama, S.; Kobayashi, M.; Yasuda, T.; Kobayashi, Y. ^{1}H -NMR Assignment and Secondary Structure of Human Insulin-Like Growth Factor-I (IGF-I) in Solution. *Biochemistry* **1992**, *111*, 529−536.
(47) Brown, S. C.; Mueller, L.; Jeffs, P. W. ^{1}H NMR Assignment and Secondary Structural Elements of Human Transforming Growth Factor alpha. *Biochemistry* **1989**, *28*, 593−599.
(48) Rico, M.; Bruix, M.; Santoro, J.; Gonzalez, C.; Neira, J. L.; Nieto, J.; Herranz, J. Sequential ^{1}H-NMR assignment and solution structure of bovine pancreatic ribonuclease A. *Eur. J. Biochem*. **1989**, *183*, 623−638.
(49) Saudek, V.; Wormald, M. R.; Williams, R. J.; Boyd, J.; Stefani, M.; Ramponi, G. Identification and Description of beta-Structure in Horse

Muscle Acylphosphatase by Nuclear Magnetic Resonance Spectroscopy. *J. Mol. Biol.* **1989**, *207*, 405−415.

(50) Fogh, R. H.; Mabbutt, B. C.; Kem, W. R.; Norton, R. S. Sequence-Specific ^{1}H NMR Assignments and Secondary Structure in the Sea Anemone Stichodactyla helianthus Neurotoxin I. *Biochemistry* **1989**, *28*, 1826−1834.

(51) Widmer, H.; Wagner, G.; Schweitz, H.; Lazdunski, M.; Wüthrich, K. The Secondary Structure of the Toxin ATX Ia from Anemonia sulcata in Aqueous Solution Determined on the Basis of Complete Sequence-specific ^{1}H NMR Assignments. *Eur. J. Biochem.* **1988**, *171*, 177−192.

(52) Holak, T. A.; Prestegard, J. H. Secondary Structure of Acyl Carrier Protein As Derived from Two-Dimensional ^{1}H NMR Spectroscopy. *Biochemistry* **1986**, *25*, 5766−5774.

(53) Padilla, A.; Cave, A.; Parello, J. Two-dimensional ^{1}H Nuclear Magnetic Resonance Study of Pike pI 5.0 Parvalbumin (Esox lucius) Sequential Resonance Assignments and Folding of the Polypeptide Chain. *J. Mol. Biol.* **1988**, *204*, 995−1017.

(54) Hyberts, S. G.; Marki, W.; Wagner, G. Stereospecific assignments of side-chain protons and characterization of torsion angles in Eglin c. *Eur. J. Biochem.* **1987**, *164*, 625−635.

(55) Inagaki, F.; Shimada, I.; Kawaguchi, K.; Hirano, M.; Terasawa, I.; Ikura, T.; Go, N. Structure of Melittin Bound to Perdeuterated Dodecylphosphocholine Micelles As Studied by Two-Dimensional NMR and Distance Geometry Calculations. *Biochemistry* **1989**, *28*, 5985−5991.

(56) Sobol, A. G.; Arseniev, A. S.; Abdulaeva, G. V.; Musina, L.; Bystrov, V. F. Sequence-specific resonance assignment and secondary structure of (1−71) bacterioopsin. *J. Biomol. NMR* **1992**, *2*, 161−171.

(57) Barsukov, I. L.; Abdulaeva, G. V.; Arseniev, A. S.; Bystrov, V. F. Sequence-specific ^{1}H-NMR assignment and conformation of proteolytic fragment 163−231 of bacterioopsin. *Eur. J. Biochem.* **1990**, *192*, 321−327.

(58) Kline, A. D.; Wüthrich, K. Complete Sequence-specific ^{1}H Nuclear Magnetic Resonance Assignments for the alpha-Amylase Polypeptide Inhibitor Tendamistat from Streptomyces tendae. *J. Mol. Biol.* **1986**, *192*, 869−890.

(59) Constantine, K. L.; Ramesh, V.; Banyai, L.; Trexler, M.; Patthy, L.; Llinas, M. Sequence-specific ^{1}H NMR Assignments and Structural Characterization of Bovine Seminal Fluid Protein PDC-109 Domain b. *Biochemistry* **1991**, *20*, 1663−1672.

CI970012C