J-Bio NMR 155

Application of neural networks to automated assignment of NMR spectra of proteins

Brian J. Hare[a] and James H. Prestegard[b,*]

[a]Department of Molecular Biophysics and Biochemistry and [b]Department of Chemistry, Yale University, New Haven, CT 06511, U.S.A.

Received 22 June 1993
Accepted 31 August 1993

Keywords: Feed forward; Back propagation; Acyl carrier protein; Boltzmann machine

SUMMARY

Simulated neural networks are described which aid the assignment of protein NMR spectra. A network trained to recognize amino acid type from TOCSY data was trained on 148 assigned spin systems from *E. coli* acyl carrier proteins (ACPs) and tested on spin systems from spinach ACP, which has a 37% sequence homology with *E. coli* ACP and a similar secondary structure. The output unit corresponding to the correct amino acid is one of the four most activated units in 83% of the spin systems tested. The utility of this information is illustrated by a second network which uses a constraint satisfaction algorithm to find the best fit of the spin systems to the amino acid sequence. Application to a stretch of 20 amino acids in spinach ACP results in 75% correct sequential assignment. Since the output of the amino acid type identification network can be coupled with a variety of sequential assignment strategies, the approach offers substantial potential for expediting assignment of protein NMR spectra.

INTRODUCTION

Since manual assignment of protein NMR spectra is tedious and prone to error, the development of automated assignment strategies has attracted a great deal of interest. Recently, spin–spin couplings between adjacent backbone atoms have been used to make sequential assignments in samples uniformly enriched in ^{13}C and ^{15}N (Clore and Gronenborn, 1991; Bax and Grzesiek, 1993). However, most work is still based on the strategy originally proposed by Wüthrich and coworkers for application to unlabeled proteins or proteins labeled only with ^{15}N (Wüthrich, 1986). Wüthrich's sequential assignment strategy relies on detecting and classifying spin systems from COSY or TOCSY data, detecting interspin system connectivities from NOESY data and then matching connected spin systems to specific segments of the protein sequence.

Both methods are amenable to automation and a number of software packages have been

*To whom correspondence should be addressed.

reported which automate one or more of the assignment steps (Eads and Kuntz, 1989; Catasti et al., 1990; Van de Ven, 1990; Eccles et al., 1991; Kleywegt et al., 1991; Bernstein et al., 1993; Grzesiek and Bax, 1993; Wittekind et al., 1993). An important component of most of these packages, especially those based on Wüthrich's strategy, is the use of side chain resonances to classify spin systems by amino acid type. The classification is based either on amino acid connectivity patterns or rules for allowed deviation of side chain chemical shifts from random coil values. While the approach is reasonable, neither the rules nor the patterns are simple and programming is a major undertaking.

We demonstrate here a new method for amino acid type determination employing a neural network algorithm. The use of neural networks offers several potential advantages over other methods for amino acid type determination. Since the neural network uses a distributed representation of the spin systems, it generalizes automatically to novel situations, such as missing or shifted resonances. The network is trained on examples of the type of data which is to be recognized, so programming rules to identify amino acids or connectivity patterns are not necessary. Finally, a neural network offers a flexible environment for pattern recognition. A variety of different types of data, such as ^{1}H and heteronuclear chemical shifts, may be easily input into a single network. Since the chemical shift of side chain resonances is correlated with secondary structure as well as amino acid type (Wishart et al., 1991), a larger training set with more diverse proteins should allow identification of elements of secondary structure, which would be very useful in the sequential assignment of the protein. Once trained, the weight matrix is a compact representation of the rules extracted from the training set.

Neural networks have been applied to a number of biological problems, including prediction of protein structure and identification of functional nucleic acid sequences (Hirst and Sternberg, 1992). NMR applications have included automated peak-picking (Kjaer and Poulsen, 1991; Corne and Johnson, 1992) and identification of ^{1}H NMR spectra of complex oligosaccharides (Meyer et al., 1991).

We chose TOCSY experiments (2D or isotope-edited 3D) to use as input to our amino acid type assignment network. The information content of the TOCSY experiment is high since it provides correlations throughout much of a spin system in a single column of a 2D or 3D plot. Also, the in-phase cross peaks show minimal multiplicity, greatly simplifying automated peak-picking of the spectrum. We used a feed-forward network, trained using the back propagation of errors algorithm (McClelland and Rumelhart, 1988). The network was trained with 148 spin systems from three assigned 2D and 3D ^{1}H TOCSY data sets from mutant and wild-type *E. coli* acyl carrier proteins (Horvath and Prestegard, unpublished results; Hill et al., unpublished results). The trained network was tested on 53 spin systems from spinach ACP, which has 37% sequence homology with the *E. coli* acyl carrier proteins (Kim et al., unpublished results). By addition of appropriate bias terms to the output units before their activation was calculated, a distribution of possible amino acids was found for each input spin system.

To illustrate that the output of the amino acid type determination network is useful in sequential assignment, we employed simulated annealing in a simple constraint-satisfaction algorithm to optimize the sequential assignment of the input spin systems. We used constraints from NOESY data (2D or 3D) in addition to the constraints on amino acid type. Once intraresidue cross peaks are eliminated, NH-NH and NH-C$^\alpha$H cross peaks provide useful information for sequential assignment. The second step was tested on a 20-residue subset of spinach ACP data.

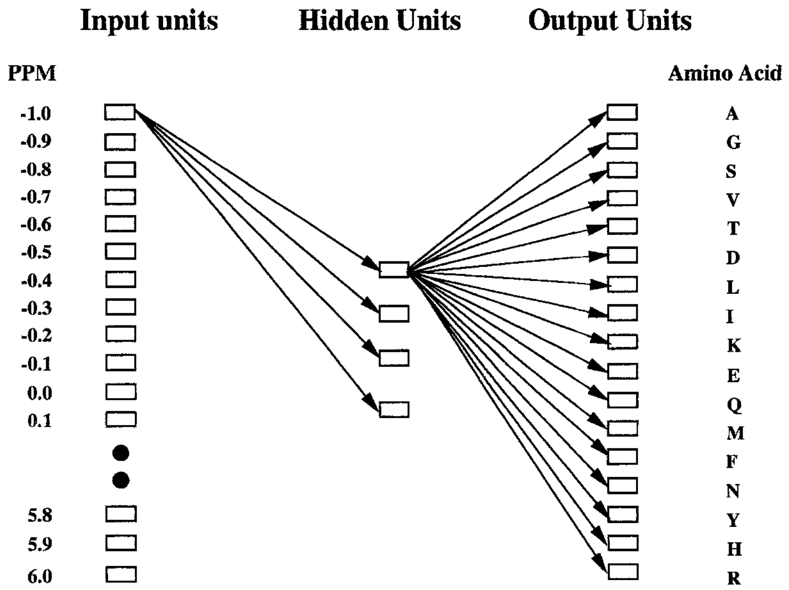

Fig. 1. The network used to identify the amino acid type of a spin system. Each unit in the network is connected to all units in the following layer by an adjustable weight. The arrows show the connections of the first unit of the input layer and the first unit of the hidden layer. The values of the input units are the volumes of cross peaks in a spin system coded in the 71-unit grid described in Experimental Methods. The number of hidden units is variable. The value of each output unit calculated by the network is interpreted as the probability that the input spin system is a particular type of amino acid.

EXPERIMENTAL METHODS

Network architecture

Spin system identification. The neural networks are part of a commercial package (McClelland and Rumelhart, 1988). The first network used is a pattern associator and is shown in Fig. 1. It contains an input layer, a hidden layer, and an output layer. Each unit is connected to all units in the following layer by an adjustable weight. These weights are varied during training to allow a pattern presented at the input layer to activate appropriate output units.

The use of hidden units is required for the network to extract rules based on second-order features of the training set. In the context of amino acid type determination of spin systems, first-order features are the part of the mapping which can be predicted by each individual cross peak in the input grid and second-order features are the part determined by pairs of cross peaks. The number of cross peaks in the methyl region of a spin system, for example, is a second-order feature which is crucial to distinguishing alanine from valine, so the use of at least one hidden layer is essential. The network implementation is flexible, allowing the number of hidden units and layers to be varied.

The data used to train and test the network were presorted as described below so that sets of cross peaks belonging to a single spin system were presented one set at a time. NH and $C^{\alpha}H$ peak positions were ignored, since they are more likely to reflect secondary structure than amino acid type (Wishart et al., 1991). Removing the most downfield resonance of each spin system between 6 ppm and 3 ppm eliminates the $C^{\alpha}H$ in most cases. Without an output unit that explicitly identifies secondary structure, inclusion of NH and C$^{\alpha}$H peaks would unnecessarily add noise to the training sets. When a larger network is trained to recognize both amino acid type and secondary structure, the NH and C$^{\alpha}$H peaks will be included.

The volumes of remaining peaks in the −1.0 to 6.0 ppm region were encoded in a grid with an increment of 0.1 ppm. The activation of each of the 71 input units in that grid is the sum of the volumes of any peaks that fall in the corresponding 0.1 ppm region. The choice of 71 units is a compromise between enhancing resolution of cross peaks within the neural network and limiting computational time and the number of patterns which must be presented in the process of training. The output layer contains 17 units, corresponding to the 17 amino acids which occur in either *E. coli* or spinach ACP sequences.

The activations of the input units are clamped by the externally supplied patterns. The activation of each hidden and output unit is between 0 and 1 and is given by Eq. 1:

$$\text{activation} = \frac{1}{1 + e^{-\text{net}_i}} \qquad (1)$$

where

$$\text{net}_i = \sum_j a_j w_{ij} + (\text{bias}_i) \qquad (2)$$

In the above equation, a_j is the activation of unit j, w_{ij} is the weight connecting unit i to unit j of the previous layer and bias_i is a term which is either unmodifiable and supplied externally or modified by learning. Since the weights which connect the units can be either excitatory or inhibitory, net_i can be either positive or negative.

Sequential assignment. The network used for sequential assignment of the protein is an implementation of a Boltzmann machine and a portion of its weight matrix is shown in Table 1. Each unit represents the hypothesis that a particular spin system should be assigned to a position in the amino acid sequence. The units are duplicated along the horizontal and vertical axes in Table 1 and the letters represent weights connecting the units. Only those units associated with positions 21 and 22 of the primary sequence of spinach ACP are shown because of space limitations. The spin system to be hypothetically assigned to each residue is denoted by either the one-letter codes for the three most probable amino acids identified by the first neural network or 'not determined' for those spin systems which contained too little information for the first network to assign. As an example, the 21st position of the primary sequence of spinach ACP is alanine so every spin system for which alanine is among the top three outputs of the first network as well as every spin system which could not be assigned an amino acid type will have a unit in the second network corresponding to residue 21. In our case, there were seven units for this residue.

Limiting choices to the top three produces a workable number of input units. In the limit of no information on the amino acid type of any of the spin systems, the network would contain N^2 units, where N is the number of amino acids in the primary sequence. The network becomes quite large under these circumstances. Using the output from the neural network shown in Fig. 1, however, the possible amino acid identity of each spin system can be restricted, and the number of output units reduced.

The units are binary and stochastic and their activation is calculated from the following equation:

$$\text{probability(activation)} = \frac{1}{1 + e^{-\text{net}_i/T}} \tag{3}$$

here, net_i is defined as it was in Eq. 2 and T is a modifiable 'temperature' factor used in simulated annealing. The bias term for a given unit is the output from the amino acid type determination network. In cases where a spin system contains too few cross peaks to determine its amino acid type, a bias of 0.05 was entered, corresponding to the spin system being assigned with equal probability to each amino acid in the sequence.

As shown in Table 1, all of the units are connected to each other by elements of a weight matrix. These weights were not determined by training as in the previous network, but were set according to a predetermined protocol. All weights were initially set to 0 and a computer program in C++ was written to make a symmetric weight matrix based on the following considerations:

1. A negative weight between units representing the same position in the amino acid sequence was added to discourage the assignment of more than one spin system to the same sequential position (a in Table 1).
2. A negative weight between units which correspond to the same spin system was added to discourage the assignment of the same spin system to more than one sequential position (b in Table 1).

TABLE 1
A PORTION OF THE WEIGHT MATRIX USED TO SEQUENTIALLY ASSIGN A 20-RESIDUE SEGMENT OF SPINACH ACP[a]

	Residue 21								Residue 22						
	Spin system	ATS	N.D.	N.D.	N.D.	ALK	ALK'	ALT	N.D.	N.D.	DKE	EDK	N.D.	EDQ	DEQ
Resi-due 21	ATS		a	a	a	a	a	a							
	N.D.	a		a	a	a	a	a	b						
	N.D.	a	a		a	a	a	a		b	2c+d				
	N.D.	a	a	a		a	a	a					b		
	ALK	a	a	a	a		a	a							
	ALK'	a	a	a	a	a		a							
	ALT	a	a	a	a	a	a								c+d
Resi-due 22	N.D.		b							a	a	a	a	a	a
	N.D.			b					a		a	a	a	a	a
	DKE			2c+d					a	a		a	a	a	a
	EDK								a	a	a		a	a	a
	N.D.				b				a	a	a	a		a	a
	EDQ								a	a	a	a	a		a
	DEQ							c+d	a	a	a	a	a	a	

[a] The spin systems which may be assigned to residues 21 and 22 in the primary sequence are shown along the horizontal and vertical axes. The three-letter names for each spin system are the one-letter codes for the three most probable amino acid types of that spin system based on the output of the network in Fig. 1. The units denoted 'N.D.' represent spin systems whose identity could not be determined by the first network. The weight matrix contains inhibitory connections, a and b, and excitatory connections, c and d, described in the Experimental Methods section. All blank entries indicate weights of 0.

3. NOE data identifying two spin systems as potential sequential neighbors was encoded in the weight matrix by reinforcing weights between units whose mutual activation is consistent with the observed NOEs (c and d in Table 1).

A weight of −1.1 was added between units based on 1 and 2 above. A weight of 1.0 was added between units assigned to sequential residues whose spin systems are connected by NH-NH or NH-C$^\alpha$H NOEs (3, above). The simple assumption was made that NH-NH connectivities could be in either direction, but that NH-C$^\alpha$H connectivities connected the NH of one residue to the C$^\alpha$H of its N-terminal neighbor. The weights c and d in Table 1 show spin systems connected by NH-NH and NH-C$^\alpha$H NOEs, respectively. While optimal for a highly α-helical protein, this protocol may have to be modified for other secondary structure types.

Implementation of the spin system identification network

Spin systems extracted from three *E. coli* ACP data sets were used to train the first network. *E. coli* ACP is a 77-residue protein with high α-helical content. A 2D TOCSY of wild-type *E. coli* ACP at 25 °C and pH 6.6 in 50 mM sodium acetate and 10 mM calcium chloride was collected on a Bruker AM500 spectrometer with a mixing time of 42 ms. A 2D TOCSY of *E. coli* ACP with valine at position 54 in place of an isoleucine was collected on a GE Omega 500 spectrometer at 30 °C and pH 7.0 in 250 mM potassium phosphate with a mixing time of 50 ms. A 3D TOCSY-HMQC spectrum of ^{15}N-labeled *E. coli* ACP at 30 °C and pH 6.6 in 100 mM sodium acetate and 7 mM calcium chloride was collected on a GE Omega 500 spectrometer with a mixing time of 60 ms.

The trained network was tested on spin systems from spectra of spinach ACP. Spinach ACP has 82 residues and a 37% sequence homology to *E. coli* ACP. Both a 2D TOCSY and a 2D NOESY dataset were collected on spinach ACP in 50 mM phosphate buffer at pH 5.9. The TOCSY dataset was collected at 25 °C on a Bruker AM500 spectrometer with a mixing time of 58 ms. The NOESY dataset was collected at 30 °C on a home-built 490 MHz spectrometer with a mixing time of 180 ms.

The spectra were processed by using Felix V2.05 (Hare Research Inc., Bothell, WA) and peaks were picked by using the Felix peak-picking routine. Spectra were assigned manually, using published chemical shift values and standard sequential assignment strategies (Holak and Prestegard, 1986). Prior to assignment or presentation to the networks, the 2D TOCSY spectra were sorted into spin systems with the help of an automated sorting algorithm written as a macro in the commercial database software package DBASE (Borland, Scotts Valley, CA) running on IBM PCs and a Sun Sparcstation I. The algorithm verifies NH to side chain proton connectivities along a column in the NH region of the directly detected dimension by finding a corresponding C$^\alpha$H to side chain connectivity in a column corresponding to the C$^\alpha$H chemical shift in the directly detected dimension. Other automated algorithms have been reported which further exploit the redundancy of 2D TOCSY spectra by continuing the search for side chain connectivities in the aliphatic region of the spectrum (Kleywegt et al., 1991). A sorting algorithm based simply on assumed resolution of columns emanating from ^{15}N/NH position in a 3D spectrum was used for the 3D TOCSY data.

Using the automated sorting algorithm with manual error checking and eliminating spin systems which contained only one cross peak, 25 sequentially assigned spin systems with a total of 65 cross peaks were extracted from the wild-type ACP data. A total of 65 sequentially assigned spin systems with 212 cross peaks were extracted from the mutant *E. coli* ACP data set. The 3D TOCSY-HMQC data set produced 58 sequentially assigned spin systems with 176 cross peaks. The spinach ACP test produced 53 spin systems with a total of 111 cross peaks.

The spin systems were prepared for input to the network by using DBASE macros to enter the volume of each cross peak into the 71-unit grid described above. NH and $C^{\alpha}H$ cross peaks were eliminated as described and the remaining cross peaks were normalized so that the sum of the volumes in each pattern was one. In order to be confident that the network had been presented with enough examples to extract the necessary rules for matching spin systems to amino acid type, we created four new patterns from each cross peak in each training pattern by moving the cross peak ± 0.2 and ± 0.1 ppm. The approach is reasonable since distributions of ^{1}H chemical shifts of aliphatic resonances in assigned proteins have standard deviations which are typically 0.5–1.0 ppm (Wishart et al., 1991). This expanded our training set to 1340 patterns.

During training, the network was presented with patterns representing spin systems of different amino acid types. The target output activation was set to 1 for the output unit corresponding to the correct amino acid type and 0 for all others. The weights were adjusted to minimize the difference between target and output activations for all patterns in the training set using the back propagation of errors algorithm (Rumelhart et al., 1986). Initial biases were 0 and were not modifiable during training. To avoid settling into a local minimum, a small learning rate of 0.05 was used and the patterns were presented randomly during each epoch of training. The weights were adjusted after the presentation of each pattern. Using a momentum value of 0.9 to filter out high-frequency variations in the error surface, a well-trained network was obtained after presenting the training set to the network 2000 times, requiring approximately 5 h of CPU time on a Sun Sparcstation I for the training set of 1340 patterns. The training procedure was repeated with 3, 4 and 5 hidden units.

Implementation of the sequential assignment network

A weight matrix analogous to that shown in Table 1 was constructed by using the C++ program described and interresidue cross peaks from the 9 to 3 ppm region in a 2D NOE spectrum of spinach ACP. The activations from the first network were entered as biases and the constraint satisfaction algorithm was initiated. This algorithm proceeds by randomly selecting units and calculating their activation by using Eq. 3. Simulated annealing was used to avoid settling into a local minimum near the starting configuration by adjusting T in Eq. 3 according to the protocol shown in Table 2. On average, the activation of each unit is updated once during a minimization cycle. The 2000 cycles of minimization shown in Table 2 required about 5 min of CPU time on a Sun Sparcstation I with 16 MB of RAM.

A combination of commercial and custom software was used in this work. The software written by us, including the sorting macros for DBASE and the program to construct the weight matrix for the second network, are available by anonymous FTP from the Internet node psun.chem.yale.edu.

RESULTS AND DISCUSSION

Amino acid type determination

The results of testing the 53 spin systems from spinach ACP using the network in Fig. 1 with

TABLE 2
SIMULATED ANNEALING PROTOCOL[a]

Cycle	T	Cycle	T
0	2	1300	0.3
100	0.9	1700	0.15
500	0.7	2000	0
900	0.5		

[a] The simulated annealing protocol used in the network in Table 1 used 2000 cycles of minimization along with a value for T which was varied continuously between the values shown.

between three and five hidden units are shown in Table 3. The correct responses are classified according to their ambiguity, defined as the total number of output units whose activation is greater than or equal to the activation of the correct unit. A response with ambiguity greater than eight is classified as incorrect. A response is defined as undetermined if the largest activation among the output units is less than 0.05.

In order to use primary sequence information in the determination of amino acid type, a positive bias was added to each output unit during testing which was proportional to the occurrence of that amino acid in the primary sequence of the protein. The added bias was the number of times the amino acid occurred in the primary sequence of spinach ACP multiplied by the bias

TABLE 3
TEST SET RESULTS FOR THE NETWORK USED TO IDENTIFY AMINO ACID TYPE[a]

Hidden units	Bias weight factor	Ambiguity								Incorrect	Undetermined
		1	2	3	4	5	6	7	8		
3	0	21	4	3	2	1	0	1	1	16	4
3	0.5	19	17	4	2	1	0	1	4	14	1
3	1.0	18	6	6	4	1	0	3	2	16	1
4	0	24	9	2	0	4	0	0	0	14	0
4	0.5	25	12	4	1	0	0	0	0	11	0
4	1.0	21	15	5	1	0	0	0	2	9	0
5	0	21	8	1	2	2	1	0	0	18	0
5	0.5	25	5	2	3	2	1	0	0	15	0
5	1.0	22	7	6	2	3	2	1	0	10	0

[a] The entries are the number of spin systems which fall into each category out of a total of 53. The results are classified according to their ambiguity (defined as the total number of units whose activation is greater than or equal to the activation of the correct unit). A response is undetermined if the largest activation among the output units is less than 0.05. A positive bias is added to each output unit during testing which is the number of times the amino acid appears in the primary sequence of spinach ACP multiplied by the bias weight factor. For nonzero bias weight factor, a bias of −60 was added to those units representing amino acids which are not in spinach ACP.

weight factor. A weight of −60 was added to units representing tyrosine, histidine, and arginine, since these amino acids do not appear in the primary sequence of spinach ACP. As shown in Eq. 1, the effect of a positive bias is to increase the activity of the unit, with the effect most pronounced for those units whose net input is closest to zero. For these units, the input data is insufficient to uniquely determine an amino acid type.

Table 3 shows that when no bias term is used, the correct amino acid type is identified as being among the top eight choices for 66 to 74% of the spin systems tested. About 60% of these responses are unambiguous. Addition of bias terms reduces the number of incorrect responses significantly, but at the expense of greater ambiguity in the responses. The correct amino acid type is among the top eight choices for 83% of the spin systems when the network with four hidden units and a bias weight factor of 1.0 is used. The correct amino acid is among the top three choices for 77% of the spin systems tested under these conditions.

When the training set was again presented to the trained network, the correct unit was the most activated for 65, 72 and 79% of the patterns using three, four, or five hidden units, respectively. Table 3 shows that when the test set was presented, the number of times the correct unit was most activated initially increased, but then decreased, as the number of hidden units was increased. These results demonstrate the well-documented observation that while the ability of a network to classify training patterns improves with the addition of hidden units, its ability to recognize test patterns eventually decreases because the network loses the ability to generalize (Hertz et al., 1991). The number of correct responses and the ambiguity of those responses suggests that the use of four hidden units is optimal for our problem. The expansion of the network to include all 20 amino acids or the addition of more input or output units to recognize other features of the spin systems would probably increase the optimal number of hidden units, but would also require additional training.

As the number of patterns presented to the network during training increases, the difference between the performance of the network on the training set and the test set would be expected to decrease. Defining $g_p(f)$ to be the fraction of the training set correctly classified by the trained network and $g(f)$ to be the fraction of all patterns correctly classified by the trained network, it has been shown that approximately W/ε training examples are required to expect a generalization error $(g_p(f) - g(f))$ of less than ε for a network with W weights and a threshold activation function (Baum and Haussler, 1989). For the network pictured in Fig. 1 with four hidden units, 3520 patterns are required to expect a generalization error of less than 0.1. Although the activation function for the hidden and output units in our network is a continuous function instead of a threshold function, it is quite steep and may be approximated by a binary function with a threshold value of 0.5.

By defining a response as correct only if the activation of the correct unit is the largest among the output units, the generalization error for the network with four hidden units was 0.27. A value greater than 0.1 was expected since the network was trained with only 1320 patterns. Reduction of the generalization error can be achieved by adding more patterns to the training set, which may result in more accurate and unambiguous determination of amino acid type.

Sequential assignment

Information about probable amino acid type can be used in optimization algorithms which incorporate both information on amino acid type and interresidue connectivities to sequentially assign NMR spectra of proteins (Bernstein et al., 1993; Grzesiek and Bax, 1993; Wittekind et al., 1993) or as an aid to manual assignment. One sequential assignment approach is illustrated here which is at present limited to rather short protein segments, but provides a clear illustration of how the probabilities from the first network can be used. The network shown in Table 1 was constructed to sequentially assign an α-helical stretch of 20 amino acids in spinach ACP. The region encompasses residues 5 to 24, where at least one sequential NOE is observed for each residue. We were not able to use a larger segment of the protein at present because of memory limitations on our Sun workstation.

For each spin system, the activations of the three most activated output units from the network in Fig. 1 with four hidden units and a bias weight factor of 0.5 were input as biases to the units corresponding to that spin system in the network shown in Table 1. The weight matrix for this network is defined by the user. In addition to negative weights discouraging the assignment of more than one spin system to the same residue or the same spin system to more than one residue, positive weights connect units whose mutual activation is consistent with observed NOEs, as described in Experimental Methods. The assumption that all NH-NH and NH-$C^{\alpha}H$ NOEs are between sequential residues obviously introduces some errors into the weight matrix. In this case, 39 of the NOEs within the 20-amino acid segment were in fact sequential, while 8 were between nonsequential residues within the segment.

Table 4 shows the results of the sequential assignment of the stretch of 20 amino acids in spinach ACP. For convenience, the spin systems are numbered the same as the residues to which they should be assigned. Using 2000 cycles of optimization and the simulated annealing protocol shown in Table 2, the network assigned 15 of the 20 spin systems to the correct amino acid. The results in Table 4 were achieved consistently from a variety of randomly generated starting configurations, suggesting that it is the global minimum configuration. The incorrectly assigned residues were grouped around i12, k14 and k16, which were identified incorrectly by the type assignment network. The i12 and k16 spin systems presented to the network contained only a single $C^{\beta}H$. The information in the β-proton chemical shift is probably not sufficient in these cases to identify amino acid type with an ambiguity of just three or four. Additional experiments, such as COSY, or analysis of complementary TOCSY columns in the aliphatic region may help to increase the number of cross peaks in each spin system and avoid this problem. The spin system which should have been assigned to K14 contained an ε proton shifted significantly downfield from its usual value. Addition of more patterns to the training set could reduce the number of errors of this type by including more examples of unusual chemical shifts.

Among the correct assignments, only the assignments of leucines 17 and 19 were ambiguous. This ambiguity is not surprising since they both exhibited a sequential NH-NH NOE to residue 18. In cases of an ambiguous assignment, the output from the network may serve as a good starting point for manual error checking.

The advantage of using the network shown in Table 1 to sequentially assign protein NMR spectra is that constraints on both amino acid type and sequential connectivities can be easily incorporated into a single weight matrix. The most serious limitation is the size of the weight matrix required for larger proteins. The size of the weight matrix for a protein with N residues is in the order of $N^4 * A^2/20^2$, where A is the average ambiguity of amino acid type associated with each spin system. While large, the weight matrix for a moderate-sized protein could be accommodated by a workstation with substantial memory.

CONCLUSIONS

The application of neural networks to the assignment of NMR spectra of proteins is clearly promising. After the application of the two networks, 75% correct sequential assignment was achieved by using just a TOCSY and NOESY dataset. The network used to determine the amino acid type of spin systems can easily incorporate a variety of different kinds of information, including proton and heteronuclear chemical shifts, coupling constants, and elements of secondary structure. Also, the assignment of amino acid types to the spin systems is probabilistic, so that a final decision on amino acid type need not be made until the spin systems are assigned sequentially.

Although we used a neural network to make sequential assignments, the two networks are completely separable, so the amino acid type determination from the first step may be useful by itself in the manual assignment of the protein or as input for another automated sequential assignment procedure. The architecture of the amino acid type determination network illustrated here is particularly suited to the assignment of homologous proteins, but we are optimistic that the method can be applied to a more diverse set of proteins by using a larger training set and training the network to recognize elements of secondary structure as well as amino acid type.

TABLE 4
THE RESULTS OF THE NETWORK SHOWN IN TABLE 1 APPLIED TO A 20-AMINO ACID STRETCH OF SPINACH ACP[a]

Residue number	Amino acid type	Spin system number
5	T	5
6	I	6
7	D	7
8	K	8
9	V	9
10	S	10
11	D	11
12	I	8, 12
13	V	7, 9
14	K	8, 11
15	E	12, 22
16	K	11, 18
17	L	17, 19
18	A	18
19	L	17, 19
20	G	20
21	A	21
22	D	22
23	V	23
24	V	24

[a] The network was tested on residues 5 to 24 of spinach ACP. For convenience, the spin systems are numbered the same as the amino acid residues to which they should be assigned. Multiple spin systems assigned to the same residue by the network are separated by a comma.

ACKNOWLEDGEMENTS

We thank Anne Horvath, Blake Hill, Mike Andrec, and Mark Oswood for supplying assigned spectra for the training and test sets, and Kai Huang, Anne Dallas, and Ranijet Ghosh for doing much of the work with DBASE. We also thank Dr. B. Meyer for introducing us to the potential application of neural networks to NMR and Erich Rescorla for computational assistance. Financial support was provided by NIH Grant No. GM32243 and acquisition of data was facilitated by NSF Instrumentation Grant DIR9015967.

REFERENCES

Baum, E.B. and Haussler, D. (1989) *Neur. Comp.*, **1**, 151–160.

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

Bernstein, R., Cieslar, C., Ross, A., Oschkinat, H., Freund, J. and Holak, T.A. (1993) *J. Biomol. NMR*, **3**, 245–251.

Catasti, P., Carrara, E. and Nicolini, C. (1990) *J. Comput. Chem.*, **11**, 805–818.

Clore, G.M. and Gronenborn, A.M. (1991) *Annu. Rev. Biophys. Biophys. Chem.*, **20**, 29–63.

Corne, S.A. and Johnson, A.P. (1992) *J. Magn. Reson.*, **100**, 256–266.

Eads, C.D. and Kuntz, I.D. (1989) *J. Magn. Reson.*, **82**, 467–482.

Eccles, C., Güntert, P., Billeter, M. and Wüthrich, K. (1991) *J. Biomol. NMR*, **1**, 111–130.

Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.

Hertz, J., Krogh, A. and Palmer, R.G. (1991) *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, CA.

Hirst, J.D. and Sternberg, M.J.E. (1992) *Biochemistry*, **31**, 7211–7218.

Holak, T.A. and Prestegard, J.H. (1986) *Biochemistry*, **25**, 5766–5774.

Kjaer, M. and Poulsen, F.M. (1991) *J. Magn. Reson.*, **94**, 659–663.

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

McClelland, J.L. and Rumelhart, D.E. (1988) *Explorations in Parallel Distributed Processing*, MIT Press, Cambridge, MA.

Meyer, B., Hansen, T., Nute, D., Albersheim, P., Darvill, A., York, W. and Sellers, J. (1991) *Science*, **251**, 542–544.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) *Nature*, **323**, 533–536.

Van de Ven, F.J.M. (1990) *J. Magn. Reson.*, **86**, 633–644.

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

Wittekind, M., Friedrichs, M.S., Constantine, K.L., Metzler, W.J., Bassolino, D. and Mueller, L. (1993) *J. Cell. Biochem.*, **17C**, 258.

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