

## Prediction of protein supersecondary structures based on the artificial neural network method

Zhirong Sun, Xiaoqian Rao, Liwei Peng and Dong Xu<sup>1,2</sup>

The State Key Laboratory of Biomembrane and Membrane Engineering, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, P. R. China, <sup>1</sup>Laboratory of Mathematical Biology, SAIC Frederick, NCI-FCRDC, Frederick, MD 21702-1201, USA

<sup>2</sup>To whom correspondence should be addressed

**The sequence patterns of 11 types of frequently occurring connecting peptides, which lead to a classification of supersecondary motifs, were studied. A database of protein supersecondary motifs was set up. An artificial neural network method, i.e. the back propagation neural network, was applied to the predictions of the supersecondary motifs from protein sequences. The prediction correctness ratios are higher than 70%, and many of them vary from 75 to 82%. These results are useful for the further study of the relationship between the structure and function of proteins. It may also provide some important information about protein design and the prediction of protein tertiary structure.**

**Keywords:** supersecondary structure/protein structure prediction/artificial neural network

### Introduction

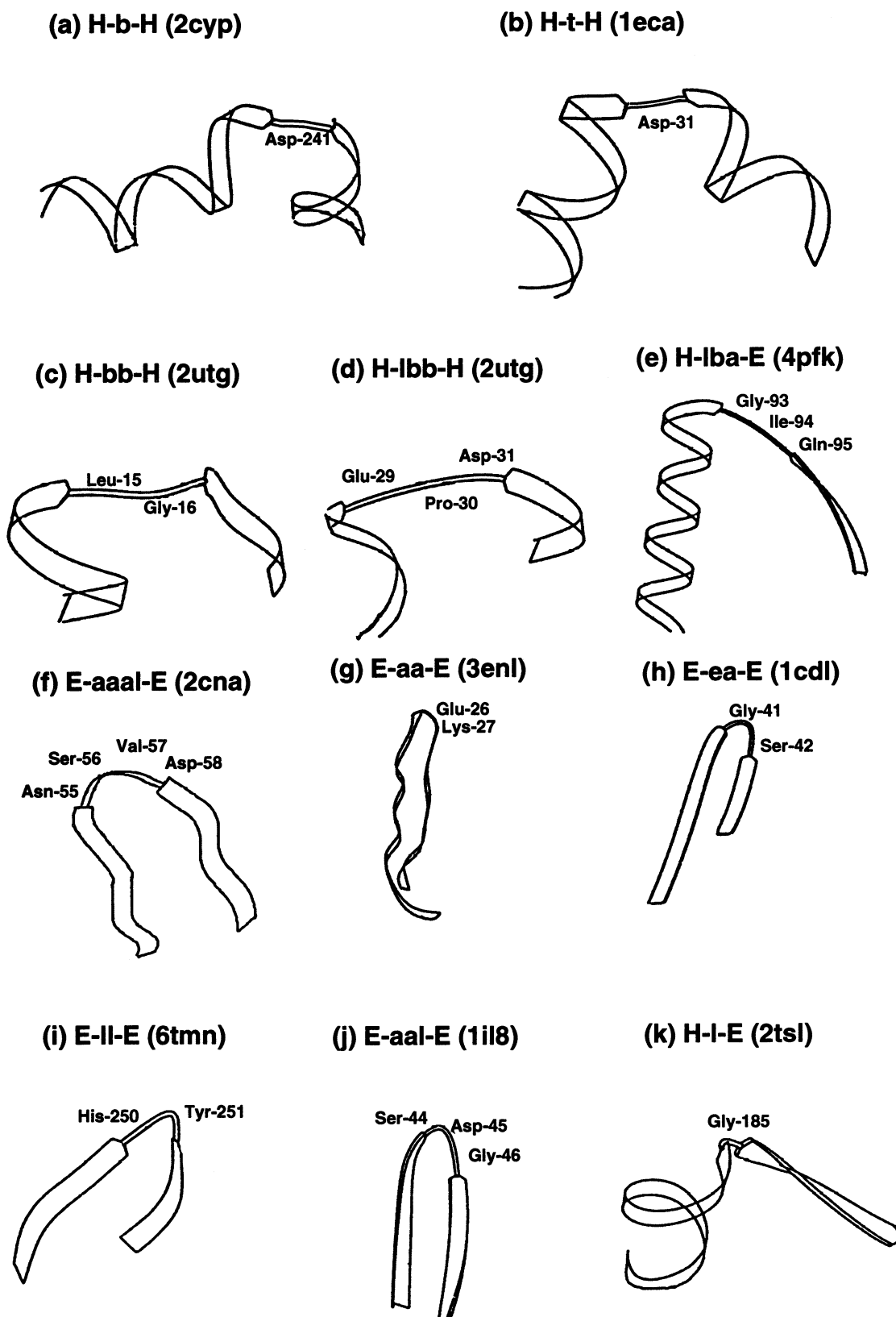
Although tremendous effort has been made, the protein folding problem, namely, prediction of the structure of a protein from its primary amino acid sequence, has yet to be solved. Many methods, such as Chou–Fasman method (Chou and Fasman, 1974), GOR method (Garner *et al.*, 1978), the pattern matching approach (Cohen *et al.*, 1986) and artificial neural network (AAN) method (Qian and Sejnowski, 1988; Holley and Karplus, 1989) have been developed and improved for the secondary structure prediction, which is an important element of the protein folding problem. Overall accuracy for predicting the three-state secondary structures (helix, strand and coil) has reached more than 70% (Salamov and Solovyev, 1995; Rost and Sander, 1995; Chandonia and Karplus, 1996). However there is still a long way to go for the tertiary structure prediction from the secondary structure assignment. Despite some success, recent trials to resolve atomic coordinates from secondary structures typically result in low resolution structures (Gunn *et al.*, 1994; Hu *et al.*, 1995).

One important step towards building a tertiary structure from the specified secondary structures is to identify how secondary structures as building blocks arrange themselves in space. High resolution X-ray analysis of protein structures shows that the conformational categories of the connecting peptides which link the  $\alpha$ -helices and  $\beta$ -sheets are limited (Thornton *et al.*, 1988; Efimov, 1993). These conformations are characteristically categorized by the angles between the secondary structures of  $\alpha$ -helices and  $\beta$ -sheets which are linked by the connecting peptides. Such well-defined types of folding units or structural motifs, e.g.  $\alpha\alpha$ - and  $\beta\beta$ -hairpins,

$\alpha\beta$ - and  $\beta\alpha$ -arches, and  $\alpha\alpha$ - and  $\beta\beta$ -corners, are referred to as supersecondary structures. A supersecondary structure is not only an important building block of the tertiary structure, but also can play an important role in the energetics of protein folding, e.g. to enhance the helix stabilization (Gurunath *et al.*, 1995).

The conformation of the coils is the key issue in identifying a supersecondary motif. The conformations of backbones on  $\alpha$ -helices or  $\beta$ -sheets are well defined, although some variation may exist. However, a coil can have a large number of conformations which play an important role in defining protein structures. Connecting peptides usually change the trend of the protein backbones so as to form an antiparallel turn, a vertical corner, a twist or just a slight bend in peptide chains. Hence, the coil in the three-state secondary structures needs to be described in detail. We found that there are five major clusters, namely *a*, *b*, *e*, *l* and *t* (Sun and Jiang, 1996) in the Ramachandran plot, for amino acids in the coil conformation. By such a clustering, we further discovered that there are 34 types of supersecondary motifs which occur more than five times in the selected 240 proteins (Sun and Jiang, 1996). Of these 34 types there are 11 types of supersecondary motifs which occur more than 25 times. We called them frequently occurring supersecondary motifs. Each motif corresponds a well defined 3-dimensional pattern, as seen in Figure 1. If the category of the supersecondary structure can be predicted from a peptide sequence, it would be extremely helpful to identify the tertiary architecture of the peptide backbones. One can also use the information in the *de novo* design of particular supersecondary structures.

In this paper, we employed an artificial neural network (ANN) method to predict super-secondary motifs from protein sequences. For this purpose, the back propagation (BP) neural network (Bryson and Ho, 1969) was applied. The BP algorithm is a classical paralleled calculation. Compared with other algorithms, it is advantageous in associating the sequence patterns directly to their 3-dimensional conformations without setting up a special theoretical model for each conformation. This feature is of particular value in the structure prediction of supersecondary motifs, which are far more complex to set up any model than the secondary structures. Another advantage of ANN method in general is that it includes the effect due to correlation of neighboring residues, while some statistical analyses, such as the Chou–Fasman method (Chou and Fasman, 1974), often derive 3-dimensional information from the propensity of a single residue. ANN has been applied to predict protein folding classes, such as all- $\alpha$ -helical proteins (Dubchak *et al.*, 1993; Reczko and Bohr, 1994; Chandonia and Karplus, 1995). Nevertheless, the conformation on how  $\alpha$ -helices and  $\beta$ -sheets are connected was not provided by these studies. To our knowledge, the work described in this paper is the first attempt to use ANN in the prediction of the supersecondary motifs.



**Fig. 1.** The topology of 11 commonly occurring supersecondary structures: (a) H-b-H (2cyp); (b) H-t-H (1eca); (c) H-bb-H (2utg); (d) H-lbb-H (2utg); (e) H-lba-E (4pfk); (f) E-aaal-E (2cna); (g) E-aa-E (3enl); (h) E-ea-E (1cdl); (i) E-ll-E (6tmn); (j) E-aal-E (1il8); (k) H-l-E (2tsl). The four letters in brackets are the PDB codes. H and E represent  $\alpha$ -helix and  $\beta$ -strand, respectively; a, b, l, e and t represent the special conformational locations on the Ramachandran plot.

**Table I.** Protein families in the 240 proteins

PDB code	Number of residues	Resolution (Å)	Family	PDB code	Number of residues	Resolution (Å)	Family
2hsc	381	2.2	01	2mhu	30	(nmr)	29
1ak3	225*2	1.9	02	1pal	108	1.65	30
2lbp	346	2.4	03	1pfk	320*2	2.4	31
1mpp	336	2.1	04	1bp2	123	1.7	32
2aza	129*2	1.8	05	2cro	71	2.35	33
7rsa	124	1.26	06	7rxn	52	1.5	34
2cab	261	2.0	07	2rsp	124*2	2.0	35
3cln	148	2.2	08	2alp	198	1.7	36
1ger	174	1.6	09	1sgt	223	1.7	37
2act	220	1.7	10	1sbt	275	2.5	38
1cy3	118	2.5	11	2gbp	309	1.9	39
256b	106*2	1.4	12	3trx	105	(nmr)	40
3dfr	162	1.7	13	1tim	247*2	2.5	41
3fxc	98	2.5	14	3tms	264	2.1	42
4fd1	106	1.9	15	4xia	393	2.3	43
1fx1	148	2.0	16	2gd1	334*4	2.5	44
4mbn	153	2.0	17	1mbd	153	1.4	45
5p21	166	1.35	18	3ebx	82	1.4	46
1hip	85	2.0	19	1fd2	106	1.9	47
2fb4	216+229	1.9	20	1tpa	223+58	1.9	48
1rei	107*2	2.0	21	4bp2	130	1.6	49
3ins	(21+30)*2	1.5	22	3tlh	316	1.6	50
1ovo	56*4	1.9	23	4ptp	223	1.34	51
5pti	58	1.8	24	1cho	245+56	1.8	52
6ldh	329	2.0	25	2cpp	414	1.63	53
1rbp	182	2.0	26	3adk	195	2.1	54
1lzt	129	1.97	27	1il8	72*2	(nmr)	55
1mrb	31	(nmr)	28	1tec	279+70	2.2	56

## Methods

### Protein structure data set

The coordinate data of 326 proteins derived from Brookhaven Protein Data Bank (PDB) (Bernstein *et al.*, 1977) were chosen by resolution. We selected the proteins with resolutions of 2.5 Å or less and used the program PROCHECK to delete those of poor quality in the X-ray diffraction analysis (MacArthur *et al.*, 1993). 240 high quality proteins were then selected. We further employed the structural comparison program COMPARER (Sali and Blundell, 1990) to analyze the homologous families among the 240 proteins in the database. Finally 56 non-redundant proteins were determined to represent all the families in the 240 proteins, as shown in Table I. In the 240 proteins, the number of sequence segments of each frequently occurring supersecondary motif is in a range of 25–87 (Sun and Jiang, 1996). Eighty percent of them are used in training the neural networks, and 20% in the test sample for predictions. To ensure a reliable test, we excluded any significant homology between the training and test proteins when we grouped the two sets. The threshold of the maximum percentage sequence identity between any protein from the training set and any protein from the test set was 30%.

### Supersecondary structure motifs

According to the 240 high quality protein structures, we set up a supersecondary structure motif database. A supersecondary structure motif in this database consists of two regular secondary structures ( $\alpha$ -helix or  $\beta$ -sheet) and the connecting peptides that link them together. A linking residue is in one of the five clustered regions (a, b, l, e and t) on the Ramachandran plot for the residues in the coil conformation (Sun and Blundell, 1995). The secondary structure unit of  $\alpha$ -helix or  $\beta$ -sheet is composed of at least three contiguous residues. For instance, the sequence HHHlbbHHH means that two  $\alpha$ -helices are

**Table II.** Examples of sequences pattern in supersecondary motif H-I-E

Conformation	Sequence pattern	PDB code	Loop range	Family
HHHH-I-EEEE .....	QIEA-G-YVLT	1fxa	73–73	14
HHHH-I-EEEX	LSAY-G-ATVL	2sga	176–176	36
bHHH-I-EEEE	TPAD-H-FTFG	4xia	7–7	43
bHHH-I-EEEE	TPED-R-FTFG	6xia	8–8	43
bHHH-I-EEEE	TPED-R-FTFG	7xia	9–9	43
HHHH-I-EEEE	HEQF-G-IVRG	2gd1	167–167	44
HHHH-I-EEEH	LRPQ-G-QCNF	2cpp	136–136	53
HHHH-I-EEEE	YETE-G-CRLQ	2ts1	184–184	
HHHH-I-EEEE	LGPR-G-LVVL	1gp1	56–56	

H and E represent  $\alpha$ -helix and  $\beta$ -strand, respectively.

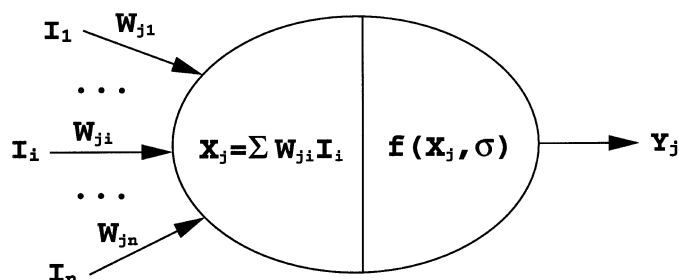
linked by three residues whose conformations are l, b and b, respectively; EEEeaEEE means that two  $\beta$ -sheets are linked by two residues in e and a conformations. H and E represent  $\alpha$ -helix and  $\beta$ -strand, respectively.

We searched the sequence patterns of the supersecondary structure motifs with a program written by ourselves in the FORTRAN language. There were 34 types of supersecondary structure motifs with the occurrence of five times or higher. As an example, Table II shows the sequence pattern of supersecondary motif H-I-E, which occurs 74 times in the 240 proteins. Among these 34 types of supersecondary structure motifs, there were 11 types whose occurrence was higher than 25 times: H-b-H, H-t-H ( $\alpha$  corner), H-bb-H, H-lbb-H ( $\alpha$  hairpin), H-lba-E, E-aaal-E, E-aa-E, E-ea-E, E-ll-E, E-aal-E ( $\beta$  hairpin) and H-l-E (arch). They can be classified into four classes ( $\alpha$ -loop- $\alpha$ ,  $\beta$ -loop- $\beta$ ,  $\alpha$ -loop- $\beta$  and  $\beta$ -loop- $\alpha$ ) (Sun and Jiang, 1996). The probabilities of 20 amino acids at every conformation position of 11 supersecondary motifs can be

**Table III.** Residue occurrence in the supersecondary motif H-1-E

Conformation position	H -3	H -2	H -1	LOOP 0	E +1	E +2	E +3
Ala	0	1.5	1	0	1	0.5	0
Arg	1	0	1	1	0	1	1
Asn	0	0	0	0.5	0	0	1
Asp	0	0	1.5	0	0	0	0
Cys	0	0	0	0	1	1	0
Gln	0	1	1	0	1	0	0
Glu	2	2	1	0	0	0	0
Gly	1	0	0	6	0	0	0
His	0	0	0	0.5	0	0	0
Ile	1	0	0	0	1	0	0
Leu	0	0	0	0	1	0	2
Lys	0.5	0	0	0	0	0.5	0
Met	0	0	0	0	0	0	0
Phe	0	0	1	0	1.5	0	1.5
Pro	1.5	2.5	0	0	0	0	0
Ser	1	0.5	1	0.5	0	0	0
Thr	0.5	1	0	0	1	2.5	0
Trp	0	0	0	0	0	0	0
Tyr	0	0	1	0	1	0	0.5
Val	0	0	0	0	0	3	2.5

The numbers are the sum of the corresponding residue's weighted occurrence in a protein. If a certain supersecondary motif occurs in the same protein family for  $n$  times, then the occurrence of the residue at a certain position of the sequence pattern is weighted by  $1/n$ .



**Fig. 2.** A model of an artificial neuron.

calculated. Table III shows an example of the statistical results for the supersecondary motif H-1-E.

*Training procedure of artificial neural network*

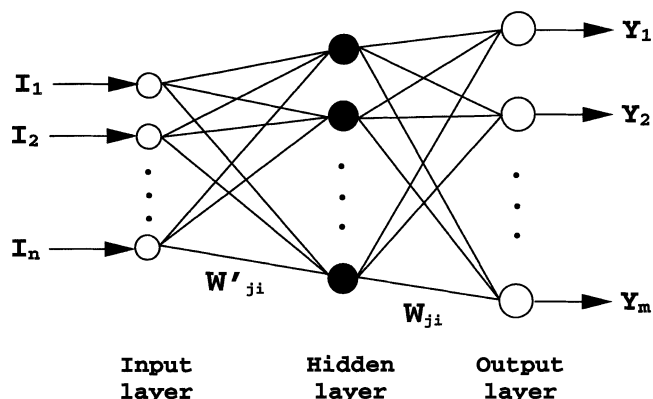
The basic unit of an ANN is the artificial neuron which is a simulation of a physiological neuron. A classical artificial neuron has the following features: (a) all or none output; (b) integration between the input messages; (c) non-linear relationship between the inputs and the outputs. Figure 2 shows a model of an artificial neuron. We assume  $I_i$  represents the  $i$ -th input, and  $X_j$  represents the overall weighted sum of the inputs to the  $j$ -th neuron, i.e.

$$X_j = \sum_i W_{ji} I_i, \tag{1}$$

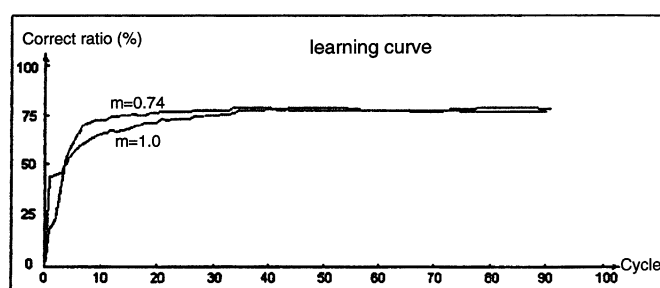
where  $W_{ji}$  is the weight of  $I_i$  for the contribution to  $X_j$ . The output of the  $j$ -th neuron is

$$Y_j = \frac{1}{1 + e^{-(X_j - \sigma)}} \tag{2}$$

where  $\sigma$  is the threshold of the neuron, and is chosen to be 0.5 (Qian and Sejnowski, 1988).



**Fig. 3.** The BP neural network that was applied to predict supersecondary structures.



**Fig. 4.** The learning curves of the training procedure for the supersecondary motif H-1-E with momentum  $m = 1$  and  $m = 0.74$ .

Figure 3 shows the BP neural network used to predict the supersecondary structures. The first layer is the input layer; the secondary layer is the hidden layer; the third layer is the output layer. The output from the input layer to the hidden layer is the input from the hidden layer to the output layer. It has been demonstrated that the neural networks with more than four layers have no remarkable improvement in prediction (Dubchak *et al.*, 1993) while the computational time is drastically increased. Therefore we used a 3-layer network in this research.

The training procedure was carried out through iterations. First we randomly assigned  $W_{ji}$  and compared the output  $Y_j$  calculated from Equation 2 with its desired output  $D_j$ , i.e. we computed the error

$$\delta_j = Y_j - D_j. \tag{3}$$

Then the weight matrix  $W$  between the hidden layer and the output layer was modified by the BP algorithm (Hertz *et al.*, 1991), i.e.

$$W_{ji}(t + 1) = W_{ji}(t) + \eta \delta_j X_i \tag{4}$$

where  $t$  represents the calculation step.  $\eta$  is called the learning rate. It is a coefficient ranged from 0 to 1. After many repeated tests, we chose an optimal value 0.3. Then the error in the hidden layer can be calculated by propagating  $\delta_j$  (the error in the output layer) backwards, i.e.

$$\delta'_i = \sum_j W_{ji} \delta_j. \tag{5}$$

The weight matrix between the input layer and the hidden layer can be modified similarly as the process (1–4). The training procedure stops until the convergence is reached.

The sequences of test samples are encoded to form the input

**Table IV.** The weight matrix for H-lba-E (input and hidden layer)

Input layer		H	H	H	H	l	b	a	E	E	E	E
Hidden layer	Unit 1	-0.038	-0.224	0.065	0.105	-0.227	-0.414	0.273	0.080	0.167	-0.085	-0.224
	Unit 2	0.078	-0.030	0.052	-0.056	0.129	-0.287	-0.225	-0.032	-0.121	-0.215	0.099
	Unit 3	0.247	0.230	0.176	0.014	0.569	0.282	0.385	-0.144	-0.235	-0.158	0.023
	Unit 4	-0.176	-0.213	-0.387	-0.251	0.142	-0.232	0.087	-0.294	0.104	-0.274	-0.131

**Table V.** The weight matrix for H-lba-E (hidden and output layer)

Hidden layer	Unit 1	Unit 2	Unit 3	Unit 4
Output layer	0.129	0.105	0.518	0.354

$I_i$  in Equation 1. Each residue is arbitrarily encoded to a number, from 1, 2, 3 to 20. For example, alanin was encoded to 6 and asparagine was encoded to 4. Therefore  $I_i$  was chosen to be 6 as input for an alanin along a sequence. The specific correspondence between an amino acid and a number is unimportant for the prediction, since the weight matrices can be adjusted accordingly during the training. For each supersecondary motif, we setup an individual neural network. The window size of the input is the number of amino acids in a supersecondary motif, e.g. 11 units for the motif H-lba-E (three residues in the loop region, and four residues at each side of the loop). Compared with the decoding method of Qian and Sejnowsky (1988), our method is equivalent to degenerate the 21 units for 20 amino acid types to the one dimension. The number of weight coefficients can be reduced substantially while the sequence pattern of supersecondary motifs is still well established in the weight matrices by the training, as shown by the good prediction results in the following.

The number of units in hidden layers equals about half of the number of units in input layers. It was proposed that the number of units in hidden layers is important to the prediction of neural network (Rumelhart and McClelland, 1987). More units in hidden layers results in a higher prediction correctness ratio. However, when the number of units in hidden layers exceeds a certain value, its importance is insignificant. After testing, we found that the ratio between prediction effect and CPU time is optimal when the number of units in hidden layers is about half of the number of units in input layers.

We define the output vector to represent the different conformation of the supersecondary structure as has been done in the prediction of the secondary structure. There are 11 elements in the output vector, corresponding to the 11 supersecondary motifs. During the training, we set the desired output  $D_j$  in Equation 3 to be 1 for the element of the actual motif, and to be 0 for the others. The software used was PREDICTOR written by ourselves.

#### Prediction by the trained neural network

During the prediction, we scanned the sequence by the 11 trained neural networks. A network has an output vector with 11 elements. We only calculated the element which corresponds to this particular neural network, i.e. whose desired value  $D_j$  in Equation 3 is 1. Hence, for a given sequence centered at a specific residue, 11 networks yield 11 such output values. The motif whose neural network gives the largest output value is picked to assign the supersecondary structure for the sequence segment.

Our predictions are judged by the correctness ratios and the Matthews correlation coefficients. Assume there are  $n$

occurrence of motif **A** in our test proteins. If the neural network predicted those region as motif **A** for  $m$  times, then  $m/n$  is defined as the correctness ratio. We also calculated the Matthews correlation coefficients  $C_j$  for all the sequence segments with 11 commonly occurring supersecondary structures (sequence segments which have structures other than the 11 commonly occurring supersecondary motifs were ignored).  $C_j$  which corresponds to the motif  $j$  is defined as (Matthews, 1975)

$$C_j = \frac{p_j n_j - u_j o_j}{\sqrt{(n_j + u_j)(n_j + o_j)(p_j + u_j)(p_j + o_j)}}, \quad (6)$$

where  $p_j$  is the number of correctly predicted sequence segments with motif  $j$ ,  $n_j$  is number of segments that are correctly identified as something other than motif  $j$ ,  $o_j$  is the number of segments which do not have motif  $j$  but are predicted as motif  $j$ , and  $u_j$  is the number of segments of motif  $j$  that are missed by the prediction.

## Results

### Training of the artificial neural network

For a network of a particular commonly occurring supersecondary structure, we trained the weight matrices with the regions known to correspond to this motif against regions of other motifs, i.e. all the regions of commonly occurring supersecondary structures were used. We trained the neural network iteratively by using the procedure described above. The results show that the error  $\delta_j$  in Equation 3 usually meet convergence after 15–20 cycles. The learning curves of each training are very similar. Tables IV and V show an example of the weight matrices derived from the training procedure with the data of H-lba-E. In this case, there are four units in the hidden layers.

In order to reduce the computing time, we introduced a coefficient  $m$ , called momentum to the algorithm, as Alum Blum (1992) advised. Then the equation can be changed to

$$W_{ji}(t+1) = mW_{ji}(t) + \eta\delta_j X_i, \quad (7)$$

where  $m$  varies from 0 to 1. The training procedure will meet convergence faster if  $m$  is properly chosen. By testing, we chose  $m = 0.74$  for the calculations. Figure 4 shows that the curves of two training procedures for the supersecondary motif H-1-E with the same data but different  $m$  values. It can be seen that convergence is faster when  $m$  equals 0.74 than that when  $m$  equals 1.0.

### Results of structure prediction

The trained neural networks (in fact, the weight matrices) were applied to predict the data which was not included in the training procedure. We obtained the correctness ratio of 75.23% to the H-1-E motif and 80.26% to the H-1bb-H motif. Furthermore, instead of a yes-or-no prediction, we put three types of motifs with 3-residue connecting peptides (H-1bb-H, H-1ba-E and E-aal-E) together and obtained a prediction correctness

**Table VI.** Correctness ratios and Matthews correlation coefficients

Motif	E-aa-E	E-aal-E	E-aaal-E	E-ea-E	E-ll-E	H-b-H
Correctness ratio (%)	75.4	74.8	78.6	80.4	78.0	72.8
Correlation coefficient	0.40	0.42	0.45	0.50	0.48	0.42
Motif	H-bb-H	H-lbb-H	H-t-H	H-l-E	H-lba-E	
Correctness ratio (%)	76.7	80.6	68.0	72.1	80.6	
Correlation coefficient	0.41	0.48	0.39	0.42	0.57	

ratio of 67.4%. The results of the prediction ratio and the Matthews correlation coefficients of 11 types of common supersecondary structures are shown in Table VI.

#### *Frequencies of the residues in the connecting peptides*

As shown above, the BP algorithm is a classical parallel calculation. The rules for further prediction obtained from the neural network training are represented in the weight matrices. Apparently, if a weight value is 0, the corresponding input value has no effect on the final output. Therefore the weight value represents the importance of the corresponding input value. We find that the weights of the amino acids in connecting peptides are significantly higher than those of the residues in the corresponding secondary elements of  $\alpha$ -helices and  $\beta$ -sheets, as demonstrated in an example in Table V. This means that the sequence patterns of the connecting peptides are the dominant factor to form a supersecondary structure.

Frequencies of the residues in the connecting peptides can provide information about the sequence pattern in each supersecondary structure motif (Sun and Jiang, 1996; Sun *et al.*, 1996). It is found that in some motifs glycine is a frequently occurring residue in connecting peptides. For example, statistics have shown that glycine occupies 73.8% of positions in motif H-l-E. This phenomenon can be explained by the fact that the side chain of glycine is the smallest one among the 20 amino acids. Therefore the dihedral angle of glycine can vary over a larger area in the Ramachandran plot than other residues. This feature facilitates the construction of a connecting peptide which usually changes the trend of a peptide chain. On the other hand, we also find that some charged/polar amino acid residues, such as aspartate, glutamate and glutamine, have higher frequencies in connecting peptides. This implies that there are factors other than side chain volumes that influence the construction of connecting peptides and the corresponding supersecondary motifs.

#### Discussions

It is very hard to set up a model for an *a priori* calculation of the supersecondary structure motifs due to lack of detailed knowledge of protein folding. Therefore, it is particularly valuable to introduce a statistical method. Statistical methods, in particular ANN, have been successful in the secondary structure prediction. Given the complexity of supersecondary structure, the prediction correctness ratios of 11 types of frequently occurring supersecondary structures are still higher than 70%. It implies that the ANN method is feasible to predict higher level conformations of a protein than the secondary structures. Our work may provide some important information for protein engineering and for the research of high level protein structures.

The high prediction rates of supersecondary structures reveal that each of the super-secondary structure motifs in our database has a particular sequence pattern whether the sequences are

from homologous protein families or not. From the ANN weight matrices, we found that the patterns are determined mostly by the sequences of the connecting peptides, and to a lesser degree by sequences of  $\alpha$ -helices and  $\beta$ -sheets around the peptides. Such underlying sequence patterns are very important to the conformations of the supersecondary peptides. We have statistically analyzed the sequence patterns of 34 types of connecting peptides and their corresponding secondary structure units, in particular, the probability of a certain amino acid occurring at a given position (Sun and Jiang, 1996; Sun *et al.*, 1996). However, more work needs to be done for identifying the statistical patterns of supersecondary structure motifs in the future.

Further improvement of our work is ongoing. We found that if we include the secondary structures at each end of the connecting peptide as an entire block and then predict the supersecondary structure motif, the prediction correctness ratio was improved considerably. This suggests that one may improve the prediction of supersecondary motifs by integrating secondary structure predictions.

The supersecondary structure prediction, as the secondary structure prediction, has its imitations, namely the conformations of connecting peptides are not only determined by their local sequence patterns, but also associated with their protein environments. But on the other hand, supersecondary motifs, as more energetically stable units in proteins than secondary structures, have a potential to be predicted more accurately from sequences. Our current investigation may provide some hints for further investigation along this line.

#### Acknowledgements

This work was supported in part by the National Natural Science Grant of China. D.X. has been supported by the National Cancer Institute, DHHS, USA. We thank Prof. Tom Blundell for helpful suggestions. We also thank Drs Ruth Nussinov and Jacob Maizel for encouragement. We are grateful to the anonymous referee for the constructive suggestions. The content of this publication does not necessarily reflect the views or policies of the Department of Human Service, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

#### References

- Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542.
- Blum, A. (1992) *Neuro Networks*. Cambridge, MIT Press.
- Bryson, A.E. and Ho, Y.C. (1969) *Applied Optimal Control*. New York, Blaisdell.
- Chandonia, J.M. and Karplus, M. (1995) *Protein Sci.*, **4**, 275–285.
- Chandonia, J.M. and Karplus, M. (1996) *Protein Sci.*, **5**, 768–774.
- Chou, P.Y. and Fasman, G.D. (1974) *Biochemistry*, **13**, 222–245.
- Cohen, F.E., Abarbanel, R.A., Kuntz, I.D. and Fletterick, R.J. (1986) *Biochemistry*, **25**, 266–275.
- Dubchak, I., Holbrook, S.R. and Kim, S.H. (1993) *Proteins: Struct. Funct. Genet.*, **16**, 79–91.
- Efimov, A.V. (1993) *Curr. Opin. Struct. Biol.*, **3**, 379–384.

- Garner, J., Osguthorpe, D.J. and Robson, B. (1978) *J. Mol. Biol.*, **120**, 97–118.
- Gunn, J.R., Monge, A., Friesner, R.A. and Marshall, C. (1994) *J. Phys. Chem.*, **98**, 702–711.
- Gurunath, R., Beena, T.K., Adiga, P.R. and Balaram, P. (1995) *FEBS Lett.*, **361**, 176–178.
- Hertz, J., Krogh, A. and Palmer, R.G. (1991) *Introduction to the Theory of Neural Computation*. New York, Addison-Wesley.
- Holley, L.H. and Karplus, M. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 152–156.
- Hu, X., Xu, D., Hamer, K., Schulten, K., Köpke, J. and Michel, H. (1995) *Protein Sci.*, **4**, 1670–1682.
- MacArthur, M.W., Laskowski, R.A., Moss, D.S. and Thornton, J.M. (1993) *J. Appl. Crystallogr.*, **26**, 283–291.
- Matthews, B.W. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.
- Qian, N. and Sejnowski, T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.
- Reczko, M. and Bohr, H. (1994) *Nucleic Acids Res.*, **22**, 3616–3619.
- Rost, B. and Sander, C. (1995) *Proteins: Struct. Funct. Genet.*, **23**, 295–300.
- Rumelhart, D.E. and McClelland, J.L. (1987) *Parallel Distributed Processing*, Vol. 1. Cambridge, MIT Press.
- Salamov, A.A. and Solovyev, V.V. (1995) *J. Mol. Biol.*, **247**, 11–15.
- Sali, A. and Blundell, T.L. (1990) *J. Mol. Biol.*, **212**, 403–428.
- Sun, Z. and Blundell, T. (1995) In Hunter, L. and Shriver, B.D., (eds), *28th Hawaii International Conference Proceeding on Systems Sciences*. Vol. 5, IEEE Computer Society Press, pp. 312–318.
- Sun, Z., Zhang, C.-T., Wu, F.-H. and Peng, L.-W. (1996) *Protein Chem.*, **15**, 721–729.
- Sun, Z., and Jiang, B. (1996) *J. Protein Chem.*, **15**, 675–690.
- Thornton, J.M., Sibanda, B.L., Edwanlo, M.S. and Barlow, D.J. (1988) *Bioessays*, **8**, 63–70.

Received on November 14, 1996; revised on March 23, 1997; accepted on March 26, 1997