Prediction of Protein Secondary Structure Based on a Multi-modal Neural Network: with Modified Profiles of MSA and PSSM

Hanxi ZHU [1], Ikuo YOSHIHARA [2], Kunihito YAMAMORI [3], Moritoshi YASUNAGA [4]

Abstract:

Prediction of protein secondary structure is considered as an important step towards elucidating its three-dimensional structure, as well as its function. We have developed a multi-modal neural network for predicting protein secondary structure. The prediction is based on the frequency profile of multiple sequences alignment and the position specific scoring matrices (PSSM) generated by BLOCK. The multi-modal neural network is composed of two steps: The first step is to develop three neural networks to predict the secondary structure states of proteins: α -helix, β -sheet and non-regular structure respectively. The single-state prediction neural networks use a local input window of consecutive amino acids to predict the secondary structure state of the amino acid located at the center of the input window; The second step is to develop a decision neural network to combine all of the single-state predictions to obtain an overall prediction on three states. This method gives an overall accuracy of 67.8% when using seven-fold cross-validation on a database of 126 non-homologous proteins. To improve the accuracy further, majority decision is introduced to each network for single-state prediction in the first step. By using majority decision, the overall accuracy is improved to 70.2% with corresponding Matthews' correlation coefficients C_α=0.61, C_β=0.48.

Key words:
Multi-model neural network, Protein secondary structure, Multiple sequence alignment, Position specific scoring matrix, Majority decision

1. Introduction

Proteins are polypeptide chains carrying out most of the basic function of life at the cell molecular level. They are large, complex molecules made up of long chains of subunits called amino acids that are attached one by one in a linear string. The amino acid sequence is called its primary structure. The secondary structure is defined as a part of three-dimensional structure that is fold in regular form: such as α -helices and β -strands. Since the increasing of the number of sequences in public database is much faster than our ability to solve their structures experimentally, the prediction of protein secondary structure from the primary amino acids sequence is considered as a very challenging task. Over the last 10 years, the prediction methods have gradually improved in accuracy. The improvement is partly due to the increased number of reliable protein structures and partly due to the improvement of prediction methods. Among these methods, computational predictive tools have become more and more refined. The problem has been approached from several angles. Many different neural networks have been applied to this task, such as NNPREDICT [1], PHD [2] and PSIPRED [3]. The implicit

1) Graduate student, Graduate School of Computer Science and Systems Engineering, Miyazaki University
2) Professor, Dept. of Computer Science and Systems Engineering, Miyazaki University
3) Associate professor, Dept. of Computer Science and Systems Engineering, Miyazaki University
4) Associate professor, Institute of Information Sciences and Electronics, University of Tsukuba

hypothesis of the prediction is that the secondary structure of a protein is uniquely determined by its sequence of amino acids. The traditional prediction method is based on a local input window of consecutive amino acids. The window slides along the sequence and the corresponding output is defined to be the secondary structure of the center residue of the input window.

Alignment has been improved to be very important and widely used means for sequence analysis. We make profile matrices to represent amino acids by using multiple sequence alignment and PSSM generated by BLOCK.

One of the characteristics of our work is to predict the three secondary structure states independently. The results of them are input into a decision feed-forward neural network to get an overall prediction. In order to improve the prediction accuracy, majority decision is introduced to the single state prediction, namely, several neural networks are developed to predict the single state in parallel and the final decisions are made by majority rule. By using majority decision, the overall accuracy of the multi-modal neural network is improved to 70.2%.

2. Genome data compiling for prediction

2.1 Data set

The information of the protein sequences, including amino acid chains and reference secondary structures, are taken from PDB (Protein Data Bank) [4)]. Usually, amino acids are classified into about 20 kinds. Here, we adopt the classification of 23 kinds, which are represented by 23 capital letters. Secondary structure is most often assigned-based on the hydrogen bond pattern between the backbone carbonyl and NH groups [5)]. By DSSP [6)] (Dictionary of Secondary Structure assignment of Proteins), eight kinds of secondary structure classes are distinguished. These eight classes are often grouped into three states: H=helix, E=strand and L=non-regular structure. Typically, H includes H (α -helix), G (3_{10}-helix) and I (π -helix). E includes E (extended strand) and B (residue in isolated b-bridge). L includes T (turn), S (bend) and (blank=other).

When using neural networks to predict protein secondary structures, the estimation of their performance is significantly influenced by the choice of protein sequences. To avoid the misleading prediction of homologous proteins, protein sequences are usually required to have a low pair-wise identity. We adopted a database of the 126 non-homologous protein sequences shown in Table 1., which are proposed by Rost and Sander, 1993 [2)]. In this set, for the chains with a length of more than 80 residues (amino acids in a protein sequence are also called residues), the mutual pair-wise similarity is less than 25%. The total number of residues is 23942, in which 31% H, 22%E, 47%L are included.

Table 1. The database of non-homologous proteins

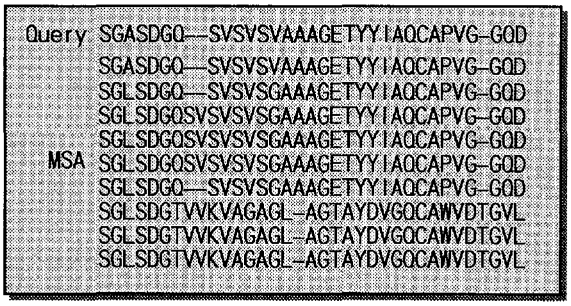
1	1acx	1azu	1bbp_A	1bds	1bmv_1
	1bmv_2	1cyo	256b_A	2aat	2ak3_A
	2alp	3ait	3blm	6acn	8abp
	8adh	9api_A	9api_B		
2	1cbh	1cc5	1cdh	1cdt_A	1crn
	1cse_I	2cab	2ccy_A	3cla	3cln
	4bp2	4cms	4cpa_I	4cpv	6cpa
	6cpp	6cts	7cat_A		
3	1a45	1dur	1eca	1etu	1fc2_C
	1fdl_H	1fkf	1fnd	1fxi_A	1g6n_A
	1iqz	2cyp	2fox	2gbp	3ebx5
	5cyt_R	5er2_E	6dfr		
4	1gd1_O	1gp1_A	1hip	1il8_A	1158
	1lap	2gls_A	2gn5	2hmz_A	2i1b
	3hmg_A	3hmg_B	3icb	4gr1	5hvp_A
	6hir	7icd	9ins_B		
5	1gdj	1lmb_3	1mcp_L	1ovo_A	1paz
	1pyp	2lhb	2ltn_A	2ltn_B	2mev_4
	2or1_L	2pcy	2phh	3pgm	4pfk
	5ldh	5lyz	9pap		
6	1mrt	1ppt	1r09_2	1rbp	1rhd
	1s01	1sh1	2mhu	2pab_A	2rsp_A
	3rnt	3sdh_A	4rhv_1	4rhv_3	4rhv_4
	4rxn	4sgb_I	7rsa		
7	1bks_A	1bks_B	1tgs_I	1tnf_A	1ubq
	2sns	2sod_B	2stv	2tgp_I	2tmv_P
	2tsc_A	2utg_A	2wrp_R	3tim_A	4ts1_A
	4xia_A	6tmn_E	9wga_A		

Furthermore, to exclude a potential dependency of evaluated accuracy on the particular test set chosen, we use seven-fold cross-validation testing to estimate the prediction accuracy of the method. The 126 protein sequences are randomized and separated into seven groups. Six groups of them are used for training the neural networks and the remaining group is used for testing. The tests are repeated cyclically seven times until each group of proteins is used once for testing. The average value over all seven tests shows a reasonable estimation for prediction.

2.2 Generation of sequence profiles

In the conventional prediction method, the orthogonal encoding is used to represent amino acids. But the prediction is limited in precision. According to the recent research, the alignment is very useful for improving prediction accuracy. Therefore, we use the multiple sequence alignment and position specific scoring matrices (PSSM) generated by BLOCK to make sequence profiles for amino acids substitution matrices.

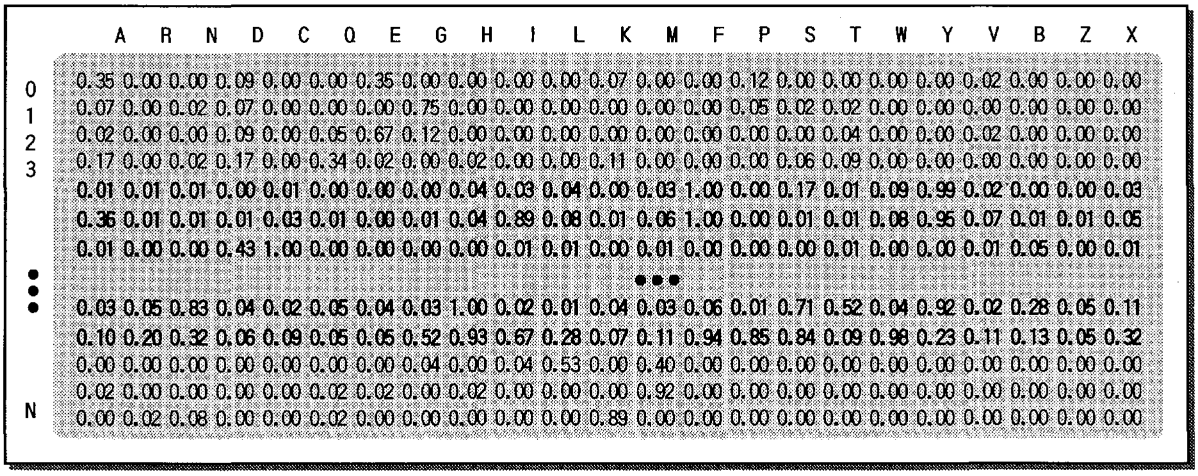
For each of the 126 sequences, we first perform the PSI-BLAST[7] with three iterations to search the homology sequences, which is a very powerful program for sequence searching. The sequences of the output, those are shorter than half of the query, are rejected. Then, for each query sequence and their homology sequences found by PSI-BLAST, we apply CLUSTALW [8] (Version 1.8) to generate multiple sequence alignments (MSA) with default parameters. CLUSTALW is a freely available program for MSA. For one query sequence that is of unknown structure and function, its alignments with its homology sequences, which are well characterized in both structure and function, reveal the structure and function of the test sequence [9]. Figure 1 gives a segment of the result. The first line is the query sequence. All of the sequences are arranged as a matrix. Each line means a protein sequence and each column corresponds to a position of the sequence. For each kind of amino acid, we calculate its frequency of occurrence at each position, namely, each column of the alignment. Thus, 23 real numbers in the range of 0-1 are generated, which are used to represent each position of the sequence. We call these numbers frequency-profile. The Figure 2 gives an example of the final sequence profile. The lines with non-bold numbers means frequency-profile. For a protein sequence with the length of N residues, the size of the frequency-profile is 23*N.



```
Query SGASDGQ--SVSVSVAAAGETYYIAQCAPVG-GQD
      SGASDGQ--SVSVSVAAAGETYYIAQCAPVG-GQD
      SGLSDGQ--SVSVSGAAAGETYYIAQCAPVG-GQD
      SGLSDGQSVSVSVSGAAAGETYYIAQCAPVG-GQD
      SGLSDGQSVSVSVSGAAAGETYYIAQCAPVG-GQD
  MSA SGLSDGQSVSVSVSGAAAGETYYIAQCAPVG-GQD
      SGLSDGQ--SVSVSGAAAGETYYIAQCAPVG-GQD
      SGLSDGTVVKVAGAGL-AGTAYDVGQCAWVDTGVL
      SGLSDGTVVKVAGAGL-AGTAYDVGQCAWVDTGVL
      SGLSDGTVVKVAGAGL-AGTAYDVGQCAWVDTGVL
```

Figure 1. A segment of result of MSA

Since the gaps in the query sequences are useless for prediction, we remove the gaps in the first line and the column below the gaps. For the remained part of MSA, we perform Multiple Sequence Alignment Processor of BLOCK [10] to generate position specific scoring matrices (PSSM). BLOCK is also a freely available program, which includes many sequence analysis methods. The Multiple Sequence Alignment Processor of BLOCK carves out one or several blocks from the MSA, which are no gapped multiple sequence alignments representing conserved protein regions, and generate PSSM.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
0	0.35	0.00	0.00	0.09	0.00	0.00	0.35	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
1	0.07	0.00	0.02	0.07	0.00	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
2	0.02	0.00	0.00	0.09	0.00	0.05	0.67	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.02	0.00	0.00	0.00
3	0.17	0.00	0.02	0.17	0.00	0.34	0.02	0.00	0.02	0.00	0.00	0.11	0.00	0.00	0.00	0.06	0.09	0.00	0.00	0.00	0.00	0.00	0.00
	0.01	**0.01**	**0.01**	**0.00**	**0.01**	**0.00**	**0.00**	**0.00**	**0.04**	**0.03**	**0.04**	**0.00**	**0.03**	**1.00**	**0.00**	**0.17**	**0.01**	**0.09**	**0.99**	**0.02**	**0.00**	**0.00**	**0.03**
	0.36	**0.01**	**0.01**	**0.01**	**0.03**	**0.01**	**0.00**	**0.01**	**0.04**	**0.89**	**0.08**	**0.01**	**0.06**	**1.00**	**0.00**	**0.01**	**0.01**	**0.08**	**0.95**	**0.07**	**0.01**	**0.01**	**0.05**
	0.01	**0.00**	**0.00**	**0.43**	**1.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**	**0.00**	**0.01**	**0.00**	**0.00**	**0.00**	**0.01**	**0.00**	**0.00**	**0.01**	**0.05**	**0.00**	**0.01**
● ● ●													●●●										
	0.03	**0.05**	**0.83**	**0.04**	**0.02**	**0.05**	**0.04**	**0.03**	**1.00**	**0.02**	**0.01**	**0.04**	**0.03**	**0.06**	**0.01**	**0.71**	**0.52**	**0.04**	**0.92**	**0.02**	**0.28**	**0.05**	**0.11**
	0.10	**0.20**	**0.32**	**0.06**	**0.09**	**0.05**	**0.05**	**0.52**	**0.93**	**0.67**	**0.28**	**0.07**	**0.11**	**0.94**	**0.85**	**0.84**	**0.09**	**0.98**	**0.23**	**0.11**	**0.13**	**0.05**	**0.32**
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.04	0.53	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.02	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.02	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	0.00	0.02	0.08	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 2. An example of the sequence profile

For a block with *n* width, the size of PSSM is 23**n*. Usually, the value of PSSM is in the range of ±25. We use a logistic function (shown as formula (1)) to scale each value within the range 0-1:

$$f(x)=\frac{1}{1+e^{-0.5x}} \qquad (1)$$

The final sequence profile is generated by embedding the PSSM of the blocks to the corresponding position of frequency-profile. In Figure 2, each line corresponds to a position of the sequence. The lines with bold numbers are PSSM.

3. Neural networks for prediction

3.1 Neural networks for single-state prediction

The first step of our work is to develop three single-state neural networks to predict three states of secondary structure respectively.

The structure of the single-state neural network (SNN) for H is a standard three-layer feed-forward neural network, as shown in Figure 3. The networks for E and L states are the same. The input layer is composed of a string of local consecutive residues. The sequence profile as shown in Figure 2 is used as amino acid substitution matrices to represent each residue of the window. The window width *w* is set to be 17 in our work (*w* is 5 in the Figure 3). In Figure 3, each square of the input layer represents one residue of the window, thus, it includes 23 numbers. Therefore, for *w*=17, the whole input pattern of the first layer extends to 23*17. The output layer includes only one node, which is corresponded to the secondary structure of the central residue of the window. In Figure 3, the assigned structure is simplified to be only two statuses: H and not H (* is not H), which are numerically represented by 1 and 0.

The window is shifted residue by residue along the sequence. Since the target output of the network is defined to be the central residue of the window, there will be no sufficient residues in both terminuses of each protein chain. In order to smooth over the default of the insufficient part, in the part of extending over the terminus, the 23 numbers to represent amino acids are set to be 0.

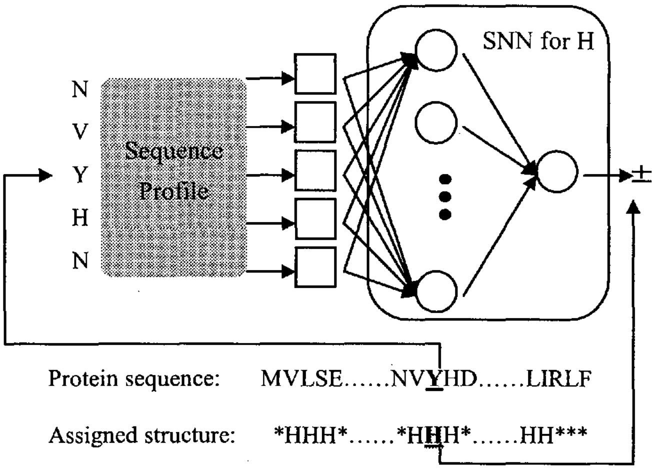

SNN for H
N
V
Y
H
N
Sequence Profile
±
Protein sequence: MVLSE......NVYHD......LIRLF
Assigned structure: *HHH*......*HHH*......HH***

Figure 3. A neural network for single-state prediction

3.2 Multi-model neural networks

To get the overall prediction, the second step of our work is to construct a multi-modal neural network (MNN).

The structure of the MNN is shown in Figure 4. After the single-state neural networks predicting the three states H, E and L independently, the results of them are input to a decision neural network (DNN) to obtain overall prediction.

The structure of the DNN is also a three-layer forward neural network. The input layer is composed of a segment of assigned secondary structure of proteins. The window width *w* is also set to be 17. Three secondary structure states are represented by three binary values as following: H →(1 0 0); E →(0 1 0); L →(0 0 1). Therefore, the whole input pattern becomes 3*17. The corresponding output of the neural network includes three nodes, which are corresponded to the assigned structure of the central position of the window.

The purpose of developing a DNN is to combine the single-state predictions to obtain the overall accuracy, but not to do the prediction itself. Because the assigned structure of the test group is the predicted target, which should be unknown, using the structure of the test group as the input data is not reasonable. In the procedure of combination, according to each position of the protein chains, the prediction results from the SNN are arranged in order of H-state, E-state and L-state to define the input pattern of the test group, and the network will give the overall prediction of the three states. According to our work, the training of the DNN is very swift.

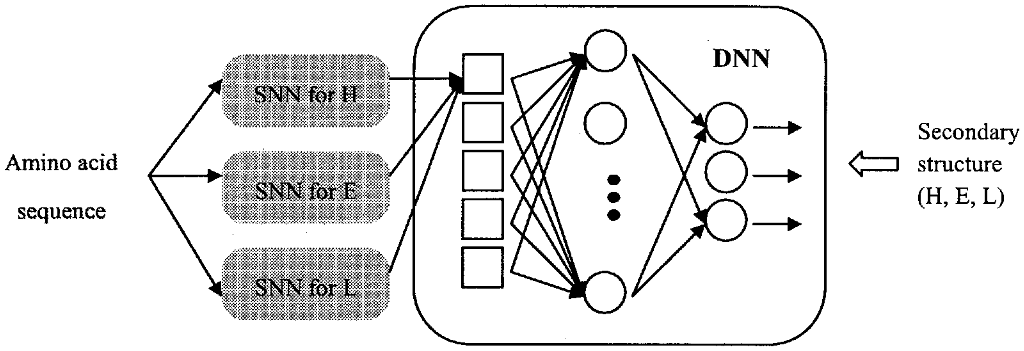


Figure 4. The structure of multi-model neural network

When the assigned structure of the protein sequences of the test group is fed, which is not for prediction, but for test, the network can give the corresponding structure state with 100% accuracy. That means the higher the prediction accuracy is, the better overall prediction the network can give. Therefore, in order to improve the prediction accuracy of protein secondary structure, we should improve the single-state prediction accuracies as high as possible.

3.3 Enhanced MNN with majority decision

In order to improve the single-state prediction, majority decision is introduced. Here, the concept of MNN is again applied to the part of single-state prediction: several neural networks are used to predict the single-state of proteins independently and the final decision is made by majority decision. We call the MNN with majority decision the enhanced MNN.

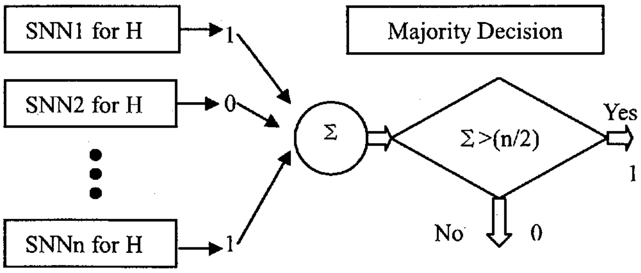


Figure 5. Majority decision for single-state prediction

Figure 5 gives a sketch map of majority decision for H-state prediction. E-state and L-state predictions are the same. SNN means a neural network for single-state prediction. The total number of SNN n is set to be 5. The outputs of SNNs are only two statuses: 1 and 0, namely, H or not H. We add the results and judge whether the sum is greater than (n/2) or not. If the sum is greater than (n/2), the final result is decided to be 1, i.e., the state in the position is predicted to be H. Otherwise the result is 0, i.e., not H.

The majority decision has been proved to be an effective tool to improve the prediction from the viewpoint of probability. Moreover, since the initial connection weights of the neural networks are taken randomly, by doing the prediction several times independently, the contingency of prediction can be reduced significantly. Therefore, using majority decision is also advantageous to improve the stability of prediction.

4. Model validation

4.1 Measure of accuracy for prediction

Several measures of prediction accuracy have been proposed to estimate the performance of the prediction methods. The most commonly used measure is the overall accuracy on three states Q_3 defined as the ratio of the total number of correctly predicted residues to the total number of residues in the database [2]:

$$Q_3 = \frac{1}{N}\sum_{i=1}^{3} p_i * 100 \ \% \qquad (2)$$

where, p_i is the number of the residues predicted correctly in state i (i=H, E or L), and N is the total number of residues in the database. Q_i gives the percentage of correctly predicted residues in state i:

$$Q_i = p_i \, / \, N_i * 100\% \qquad (3)$$

where, N_i is the number of observation residues in state i. Another complementary measure of prediction accuracy is the Matthews' correlation coefficients for each type of predicted secondary structure [2]:

$$C_i = (p_i n_i - u_i o_i) / \sqrt{(p_i + u_i)(p_i + o_i)(n_i + u_i)(n_i + o_i)} \quad (4)$$

where p_i is the number of correctly predicted residues in assigned state i; n_i is the number of those correctly predicted residues in not assigned state i; u_i is the number of underestimated residues and o_i is the number of overestimated residues in state i. The closer this coefficient is to a value of 1, the more successful the method for predicting a residue in state i.

4.2 MNN without majority decision

In the experiment of predicting the data shown in Table 1 by using MNN without majority decision: the window width of the single-state prediction neural networks is set to be 17 and the neuron numbers in three layers are set to be 391 (23*17), 60 and 1; the window width of DNN is also 17 and the neuron numbers are 51 (3*17), 20 and 3.

Figure 6 gives the prediction accuracies by the MNN without majority decision. According to cross-validation testing, the predictions are repeated 7 times and the average values are shown in the last group.

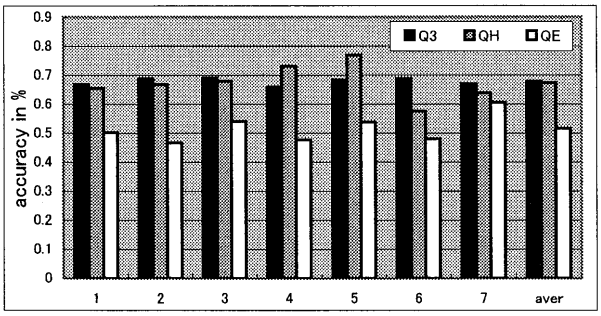


Figure 6. Prediction accuracy of MNN

Table 2. The prediction accuracy of MNN

	Q_3 (%)	Q_H (%)	Q_E (%)
1	66.8	65.5	50.2
2	68.8	66.8	46.6
3	69.0	67.9	54.1
4	65.9	73.1	47.7
5	68.4	76.9	53.8
6	68.8	57.6	48.0
7	66.9	63.9	60.5
Average	67.8	67.4	51.6

Table 2 gives the detailed values of the experiment. The averages of Q_3, Q_H, and Q_E are 67.8%, 67.3% and 51.6%. The average correlation coefficients are also calculated: C_H=0.58, C_E=0.43.

4.3 Enhanced MNN with majority decision

In the enhanced MNN, the majority decision is added to improve the single-state prediction.

Figure 7 gives the prediction accuracies by the enhanced MNN with majority decision. The neuron numbers in the enhanced MNN are the same with the MNN. Table 3 gives the detailed values of the experiment. By adding majority decision to the single-state prediction, the averages of Q_3, Q_H, and Q_E are improved to 70.2%, 71.1% and 57.3% respectively. Furthermore, we notice that the variance of Q_3 in cross-validation testing is from 68.5% to 71.4%. The deviation value to the average accuracy is about $\pm 1.5\%$, which shows that according to different data, prediction by the enhanced MNN is very stable.

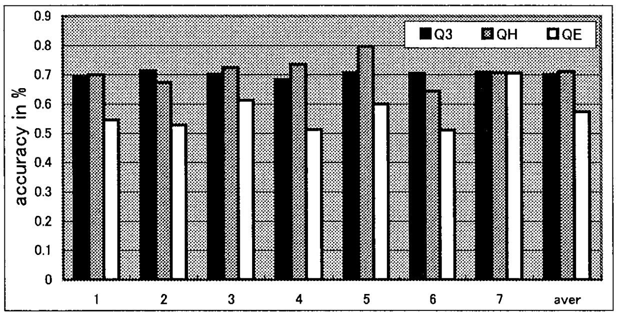


Figure 7. Prediction accuracy of enhanced MNN

Table 3. The prediction accuracy of enhanced MNN

	Q_3 (%)	Q_H (%)	Q_E (%)
1	69.2	70.0	54.5
2	71.4	67.3	52.8
3	70.2	72.4	61.3
4	68.5	73.4	51.3
5	70.7	79.6	60.0
6	70.5	64.5	51.1
7	70.9	70.7	70.6
Average	70.2	71.1	57.4

We compare the averages of Q_3, Q_H, Q_E, Q_L, C_H, and C_E with the two methods in Table 4. Obviously, all of the accuracies of the enhanced MNN are better than those of the MNN.

Table 4. Comparison of the enhanced MNN with the MNN

	Q_3	Q_H	Q_E	C_H	C_E
MNN	67.8%	67.4%	51.6%	0.58	0.43
Enhanced MNN	70.2%	71.1%	57.4%	0.61	0.48
Difference (ratio)	+2.4% (+3.5%)	+3.7% (+5.5%)	+5.8% (+11.2%)	+0.03 (+5.2%)	+0.05 (+11.6%)

4.4 Comparison with other methods

Moreover, we compare the accuracies of our methods with the claimed accuracies of conventional prediction methods.

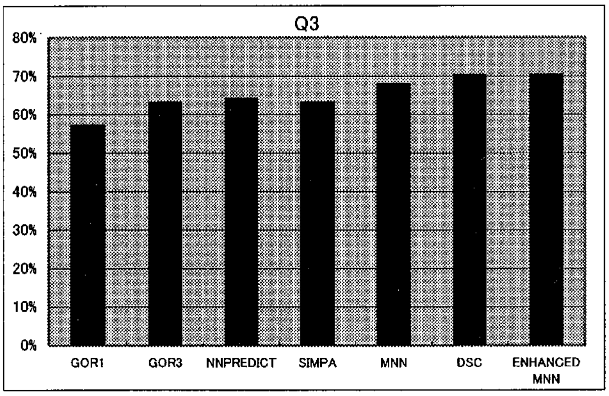


Figure 6. Comparison with other methods

Table 5. The detailed values of comparison

Method	Q_3 (%)
GOR1	57%
GOR3	63%
NNPREDIC	64%
SIMPA	63%
MNN	68%
DSC	70%
Enhanced MNN	70.2%

The accuracies of our methods: the MNN and the enhanced MNN are better than GOR1 [11], GOR3 [12], NNPREDICT [1] and SIMPA [13], where, GOR1 and GOR3 are based on Bayesian Statistics, NNPREDICT is based on neural networks and SIMPA is based on nearest neighbor. The Q_3 of the enhanced MNN is a little better than DSC [14], which is also based on Bayesian Statistics. Table 5 gives the detailed values of comparison.

5. Conclusions

We developed a multi-modal neural network to predict protein secondary structure. There are two steps in our work: the first is using several neural networks to predict single-state of proteins: α-helices, β-sheets or non-regular structure respectively; the second is to develop a multi-modal neural network to obtain the overall prediction. A special sequence profile is used to represent amino acid sequence instead of orthogonal encoding. To enhance the prediction ability of MNN, majority decision is introduced to the single-state prediction. The overall accuracy of the enhanced MNN is improved to 70.2%, which is 2.4% higher than that of MNN without majority decision.

Predicting the state of secondary structure by separate neural networks is easier than by the conventional neural network because the separate prediction decreases the complexity of the problem.

Comparing our method to several conventional methods, such as GOR1, GOR3, NNPREDICT and SIMPA, the results showed our method is superiority to them, but still lower than PSIPRED and PHD.

The neural networks show a great potential to predict the secondary structure of proteins. As we believe the MNN involves strong possibilities for better prediction, we continue to work to achieve better accuracy.

6. Acknowledgement

This research is supported in part by the 2002 grant of the Ministry of Education, Culture, Sports, Science and Technology in Japan (Grant No. 14015206).

7. Reference

1) Kneller, D. G., Cohen, F. E. and Langridge, R.,

"Improvements in protein secondary structure prediction by enhanced neural networks", *J. Mol. Biol.*, Vol. 214, pp. 171-182, (1990)

2) Rost, B. and Sander, C., "Prediction of Protein Secondary Structure at Better than 70% Accuracy", *J. Mol. Biol.*, Vol. 232, pp. 584-599, (1993)

3) David, T. Jones, "Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices", *J. Mol. Biol.*, Vol. 292, pp. 195-202, (1999)

4) Bernstein, F. C., et al., "The Protein Data Bank: a computer based archival file for macromolecular structures", *J. Mol. Biol.,* Vol. 112, pp. 535-542, (1977)

5) Burkhard Rost & Chris Sander, "Third Generation Prediction of Secondary Structures", *Method in Molecular Biology*, Vol. 143, pp. 71-95, (1998)

6) Kabsch, W. and Sander, C., Dictionary of Protein Secondary Strcutre: Pattern Recognition of Hydrogen Bonded and Geometrical Features, *Biopolymers,* Vol. 22, pp. 2577-2637, 1983

7) Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., W. & Lipman, D. J., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucl. Acids. Res.*, Vol. 25, pp. 3389-3402, (1997)

8) Thompson, J. D., Higgins, D. G., Gibson, T. J., "CLUSRAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice", *Nucl. Acids. Res.*, Vol. 22, pp. 4673-4680, (1994)

9) Desmond G. Higgins and William R. Taylor, "Multiple Sequence Alignment", Method in Molecular Biology, volume 143, pp. 1-18, 2000

10) Henikoff, S. Henikoff, J. G., Alford, W. J., Pietrokovski, S., "Automated construction and graphical presentation of protein blocks from unaligned sequences", *Gene.* Vol. 163, GC17-26, (1995)

11) Garnier, J., Osguthorpe, D.J., Robson, B., "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins", *J Mol Biol.*, Vol. 120, pp. 97-120, (1978)

12) Gibrat, J. F., Robson, B., Garnier, J., "Further developments of protein secondary structure prediction using information theory", *J Mol Biol.*, Vol. 198, pp. 425-443, (1987)

13) Levin, J. M, "Exploring the limits of nearest neighbor secondary structure prediction", *Protein Engineering*, Vol. 10, pp. 771-776, (1997)

14) King, R. D. and Sternberg, M. J. E., "Identification and application of the concepts important for accurate and reliable protein secondary structure prediction", *Protein Science*, Vol. 5, pp. 2298-2310, (1996)