

# A method of classification and recognition of blue copper protein

著者	Purqon Acep, Sugiyama Ayumu, Takamatsu Yuichiro, Nagao Hidemi, Nishikawa Kiyoshi
journal or publication title	AIP Conference Proceedings
volume	832
number	5
page range	574-577
year	2006-05-01
URL	<a href="http://hdl.handle.net/2297/3451">http://hdl.handle.net/2297/3451</a>

# A Method of Classification and Recognition of Blue Copper Protein

Acep Purqon, Ayumu Sugiyama, Yuichiro Takamatsu, Hidemi Nagao,  
and Kiyoshi Nishikawa

*Division of Mathematical and Physical Science, Graduate School of Natural Science and Technology,  
Kanazawa University, Kakuma, Kanazawa 920-1192, JAPAN*

**Abstract.** Some proteins in blue copper proteins have similar properties. In some cases it is not easy to distinguish the proteins each other. The study to recognize and classify in blue copper proteins has important roles to recognize the difference of similar properties, for examples, structures and residue sequences in blue copper proteins. There are many methods being developed to predict protein structure from many approaches, which one still not satisfactory yet. Therefore it is a challenge for scientists to develop or improve their methods. One of promising method is artificial neural networks (ANN). ANN is learning machine methods consisted of input, hidden and output layer. ANN is tested to recognize secondary structure in blue copper protein. It is found that ANN can distinguish for 7-type of secondary structure and recognize 72% secondary structure in blue copper protein.

**Keywords:** blue copper protein, Artificial Neural Networks

**PACS:** 87.15.Aa; 87.15.Cc

## INTRODUCTION

Blue copper protein is one of metalloproteins containing copper ion and showing blue color in EPR spectrum. Blue copper protein consists of various type of protein and can be found in biological system for example in animal, human, plants and so on. Until now this group is being explored. One of interesting problem in blue copper protein is similarity in structure and amino acid sequences. Sometime it is hard to distinguish protein each others. Therefore the study to recognize among the protein has important roles.

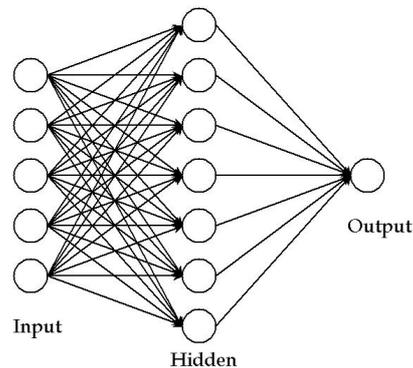
In this study we use artificial neural networks (ANN) method to predict and recognize the secondary structure in blue copper protein. As we know ANN has many various methods and applications in many areas especially for classification, recognition, prediction, simulation, analysis and so on. In this problem, we use ANN as classification and recognition methods. Some calculation using ANN for Secondary structure prediction can be found in some papers[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11].

## METHOD

ANN is simply consists of input, hidden and output layers as shown in Fig.1. In this case we use ANN using backpropagation algorithm. This algorithm is as follow[12]. First, it is weight initialization which set all weights and node thresholds to small random numbers.

Second, it is calculation of activation determined by

$O_j = F(\sum W_{ji}O_i - \theta_j)$ , where  $W_{ji}$  is the weight from input  $O_i$ ,  $\theta_j$  is the node threshold, and  $F$  is sigmoid function  $F(a) = 1/(1 + \exp(-a))$ .



**FIGURE 1.** Schematic diagram of ANN with input, hidden and output layer. Input for this case is primary structure of blue copper protein and output is secondary structure of blue copper protein

Third, it is weight training which start at the output units and work backward to the hidden layers recursively. Adjust weights is determined by  $W_{ji}(t+1) = W_{ji}(t) + \Delta W_{ji}$  where  $W_{ji}(t)$  is the weight from unit  $i$  to unit  $j$  at time  $t$  (or  $t$ th iteration) and  $\Delta W_{ji}$  is weight adjustment. The next is weight change which is computed by  $\Delta W_{ji} = \eta \delta_j O_i$  where  $\eta$  is a trial-independent learning rate ( $0 < \eta < 1$  e.g 0,3) and  $\delta_j$  is the error gradient at unit  $j$ . The convergence is sometimes faster by adding a momentum term  $W_{ji}(t+1) = W_{ji}(t) + \eta \delta_j O_i + \alpha [W_{ji}(t) - W_{ji}(t-1)]$  where  $0 < \alpha < 1$ .



**TABLE 3.** Vector for output. All characters/strings from 7 secondary structure are converted to be value/number

Letter	Secondary Structure	Output vector
H	Helix	1 0 0 0 0 0
B	Residue	0 1 0 0 0 0
E	Extended Beta Strand	0 0 1 0 0 0
G	310 Helix	0 0 0 1 0 0
I	Pi Helix	0 0 0 0 1 0
T	Hydrogen Bonded Turn	0 0 0 0 0 1
S	Bend	0 0 0 0 0 1

**TABLE 4.** Training data: (75% data) representing animal, plant, and bacterium in blue copper protein

1RKR	1ETJ	1PMY	2TSA	1AG6
5AZU	2CBP	1BAW	1BYO	2PLT
2AZA	1BQ5	1ZIA	1PAZ	1AAC
1JER	1NDT	8PAZ	1RCY	1KDJ
1URI				

the input of the network is then a vector with  $5 \times 20$  elements.

To make ANN work we train this algorithms and make preprocessing in order to reduce time and avoid overfitting. Training data consists of 2851 residue sequences and testing data 751 residue sequences. in matrix form there are  $100 \times 2851$  for training data and  $7 \times 751$  for testing data. The output pattern is much simpler because we only need to encode only 7-vector as shown in table 3. the 7-vector of output are Helix (H), Residue (B), Extended Beta strand (E), 310 Helix (G), Pi Helix (I), Hydrogen Bonded Turn (T) and Bend (S). The consideration and determination of hidden layer number depend on situation. In this case to make comparison, we use 50 hidden layer and 150 hidden layer. Figure 3 shows the architecture of the corresponding 5 window neural networks for recognizing and distinguishing 7-type of secondary structure.

In this case we to make prediction based on input, we

**TABLE 5.** Testing data: (25% data). The testing data are chosen so that the output data representing animal, plant, and bacterium in blue copper protein

PDB	type	PDB	type
1ARN	Animal	1IUZ	Plant
1NWP	Plant	1PLC	Plant
1A4C	Animal	7PCY	Plant
1PCS	Bacterium		

**TABLE 6.** Recognition using 150 hidden layer. Mark (\*) has two meaning categories: the secondary structure prediction is out of pattern and the others one is false interpretation

1ARN (80%)	OEEEEEO TTSOB*O**E EE*TTTOSEE* EEEEEOSSOO HHHHO*O*EE EE***HH*H* HHH*T*T*** TTTTTT**B SEEOO*OTT OEEEEEEG* **TTOOE*E EOOSTTTTT SEEEEE
1NWP (60%)	OEEE**O T**O**OS*E E**TTTOSEEE EEEE*OS**O HHHHO*O*EE EE* ***** **TT*TB *EEOO*OTT OEEEEEEG*GTTT** EEE *OOSTT*** *EEEE
1A4C (85%)	OEEEEEO **SO*SOSEE EE*TTTOSEEE EEEEEOSSOO HHHHOOBEE EEG***HHH HHHH**TGGG TTTTTT*** SEEOO*OTT OEEEEEEG* G*T*TOEEEE EOOSTTTTT* *EEEE
1PCS (57%)	*EE*** *OO**EES*E EEE**T**EE EEEE**OBO *EEO*****HHH*** *****TOE*EE *EOSOEEEE EEO***TTT OEE*EE
1IUZ (74%)	*EE**OTT *OOSEESS*E EE*TT*E*EE EEOSSOBO* EEO*****TT **HH***OS O**STTOEEE EE*O*OEEEE EEO***** *EEEE
1PLC (74%)	E*ESOTTO O**EES*EEE **TT*EEEE EEOSSOBO*E EO***OTTT *****OOTT OOB*S*T**E EEE*OS*EEE EEO*GGTTT TOEEEE
7PCY (73%)	EEE**OTT OO*SEE*S*E EE*TTTOEEEE EEOSS*OBO* EEO*****TT ***HHHO*** O**S*TO*EE EEOOEEEE EEO*TTTTT *EEEE

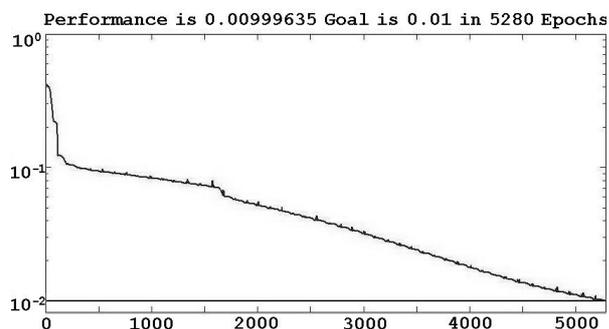
start from no. 1 to 5 of primary structure of the sequence to predict secondary structure no.3. Then, no.2 to 6 to predict secondary structure no.4, and so on. Therefore we do not involved two data in the first secondary structure and two data in the last secondary structure as part of prediction (see again Fig.2). We divide 28 proteins of blue copper protein into 21 protein which involve azurin, plastocyanin an so on for training data and 7 proteins for testing data which also involve azurin, plastocyanin an so on(see table 4 and 5).

## RESULTS AND DISCUSSION

The performance of prediction and recognition using ANN for each protein is shown in Table 6. ANN can recognize each letter or 7-types of secondary structure with the exception using the mark (\*) which means they have two categories. The first category, the secondary

**TABLE 7.** Performance of recognition between 50 hidden layer with SSE=0.01 and 150 hidden layer with SSE=0.0001

PDB	50 hidden	150 hidden
1ARN	76%	80%
1NWP	58%	60%
1A4C	85%	85%
1PCS	59%	57%
1IUZ	55%	74%
1PLC	73%	74%
7PCY	60%	73%
<b>AVERAGE</b>	<b>66.4%</b>	<b>72%</b>



**FIGURE 4.** Example of SSE calculation using parameter 150 hidden layer and SSE=0.01 which can be reached after 5280 epoch

structure prediction is out of pattern and the second one is false interpretation.

Table 7 shows The prediction result for each protein is varied, for example, starting from the highest prediction respectively are *A. denitrificans* (M121H) azurin, *A. xylooxidans* azurin II, *P. nigra* plastocyanin, *U. pertusa* plastocyanin, *E. proliferans* plastocyanin, *P. putida* azurin, and *S.sp* PCC 6803 plastocyanin. On the other hand *A. denitrificans* (M121H) azurin is the highest prediction result with 85%. It is meant that sequence pattern in this protein have been recognized by ANN. Meanwhile *S.sp* PCC 6803 plastocyanin is the lowest prediction result with 57%, probably the pattern in this protein can not well be recognized or there are new pattern in this sequence residue of the protein. However the result still can be improved using variation of hidden layer number, addition of blue copper protein data, and then we make better preprocessing/conversion and make better rules in neighbourhood rules and so on.

The average of prediction can be compared between ANN using 50 hidden layer with 66.4% and 150 hidden layer with 72%. In this case increasing hidden layer can improve the prediction result. we make comparison

with another paper which the prediction result is 64% [10]. In addition, in this paper we calculate the output of prediction for 7 types of secondary structure. Meanwhile in another paper [3, 9, 10] use 3 types, there are Helix, Beta strand and Coil. Although prediction of 7 types of secondary structure is more difficult, but the ANN still can predict for this problem.

## CONCLUDING REMARKS

From this result, we can make conclusion that ANN method can be used to make prediction and recognition for secondary structure in blue copper protein. This method can be alternative method for recognition and prediction problem. The result for this case shows that ANN can distinguish for 7-type of secondary structure and recognize 72% secondary structure in blue copper protein.

## ACKNOWLEDGMENTS

H.N. is grateful for financial support from the Ministry of Education, Science and Culture of Japan (grant 15550010).

## REFERENCES

1. T. Head-Gordon, and F. H. Stillinger, *Phys. Rev. E* **48**, 1502–1515 (1993).
2. F. H. Stillinger, T. Head-Gordon, and C. L. Hirshfeld, *Phys. Rev. E* **48**, 1469–1477 (1993).
3. R. C. Yu, and T. Head-Gordon, *Phys. Rev. E* **51**, 3619–3627 (1995).
4. M. Compiani, P. Fariselli, and R. Casadio, *Phys. Rev. E* **55**, 7334–7343 (1997).
5. H. G. Bohr, K. Frimand, K. J. Jalkanen, R. M. Nieminen, and S. Suhai, *Phys. Rev. E* **64**, 021905 (2001).
6. N. Qian, and T. J. Sejnowski, *J. Mol. Biol.* **202**, 865–884 (1998).
7. B. Rost, C. Sander, and R. Schneider, *J. Mol. Biol.* **235**, 13–16 (1994).
8. S. Muskal, and S. Kim, *J. Mol. Biol.* **225**, 713–727 (1992).
9. P. Stolorz, A. Lapedes, and Y. Xia, *J. Mol. Biol.* **225**, 363–377 (1992).
10. X. Zhang, J. P. Mesirov, and D. L. Waltz, *J. Mol. Biol.* **225**, 1049–1063 (1992).
11. H. Gray, B. Malmstrom, and R. William, *J. Biol. Inorg. Chem.* **5**, 551–559 (2000).
12. L. M. Fu, *Neural Networks in Computer Intelligence McGraw-Hill*, 80–91 (1994).
13. E. Adman, *Advance in Protein Chemistry* **42**, 145–197 (1991).
14. W. Kabsch, and C. Sander, *Biopolymers* **22**, 2577–2637 (1983).
15. J. Procter, [www.zbh.uni-hamburg.de/teaching/WS2002/00.905/](http://www.zbh.uni-hamburg.de/teaching/WS2002/00.905/) (2002).