

Using Artificial Neural Network for Protein Secondary Structure prediction.

Pongsak Suvanpong

Department of Biological Sciences, Macquarie University, NSW 2109, Australia

Introduction

Protein secondary structure forms from a sequence of amino acid which is primary structure of the protein. It is generally a three dimension structure. A sub sequence string of a long amino acid string can be formed into 3 distinct structures namely: Alpha helix, beta sheet and random coil.

Some properties of a protein can be determined from Knowing the secondary structure of the protein. The known structure of proteins so far was done using a technique called X-ray diffraction patterns of crystallized then the data from the process is fed through the [DSSP algorithm](#)(Kabsch and Sander, 1983) to determine the exact protein structure. The process is time consuming and expensive. There are, however, so many possible combinations of amino acid that could fold into the distinct structures, that why automating predicting secondary structure of protein can be very useful.

The very first computer software that used machine learning approach for protein secondary structure prediction was [GOR](#)(Garnier J., Osguthorpe D. J. and Robson B, 1978). Named after the three scientists who developed it. The software used [Bayesian classifier](#) to predict the secondary structure of a protein. The scientists used the informations from [X-ray crystallography](#) to train their classifier.

An [artificial neural network](#) was first used to predict protein secondary structure in 1988, a research done at the Department of Biophysics at Johns Hopkins University (Ning Qian and Terrence J. Sejnowski, 1988). The neural network simulator they've used, uses [Back-propagation](#) training algorithm(Williams and Zipser, 1989), had achieved accuracy close to 70%. The neural network, as its name suggests, uses signal backward propagation to correct error from its input patterns during training. The recent advance in using artificial neural network simulator to predict secondary structure of a protein, has been able to achieve over 70% accuracy.

In this article I will describe YASPIN(Lin, Simossis, Taylor and Heringa, 2004). The software uses combination of artificial neural network simulator(ANN) and [Hidden markov model](#)(HMM), to predict secondary protein structure. YASPIN can be accessed online from this URL <http://ibivu.cs.vu.nl/programs/yaspinwww> or <http://mathbio.nimr.mrc.ac.uk>

YASPIN

YASPIN had 2 modules:ANN and HMM, each was trained separately. When running, however, the outputs from ANN module were fed into the HMM module for the final output.

The ANN had 315 inputs units and 7 output units. The input units had 15 groups and each group has 21 neurons, hence, $21 \times 15 = 315$. In each group, 20 neurons represented an amino acid in binary number and another neuron represented ending of amino acid sequence string. Therefore each input pattern contained string of 14 amino acids. The amino acids were encoded to binary string according to the order of the amino acid in table in figure 1. For example, to encode alanine, the binary value would look like this 10000000000000000000, hence, alanine was in the first entry of the figure 1, or arginine 01000000000000000000. The seven output units represented possible secondary structure of the input pattern. The output units labeled corresponding with seven local structure states(Kabsch and Sander, 1983): Hb-Helix beginning, H-Helix, He-Helix end, Eb-strand beginning, E-strand, Ee-strand end and C-coil. The ANN was trained using the Back-propagation algorithm using the dataset output from PDB25.

Amino acid	3 letter	1 letter
1) alanine	ala	A
2) arginine	arg	R
3) asparagine	asn	N
4) aspartic acid	asp	D
5) cysteine	cyc	C
6) glutamic acid	glu	E
7) glutamine	gln	Q
8) glycine	gly	G
9) histidine	his	H
10) isoleucine	ile	I
11) leucine	leu	L
12) lysine	lys	K
13) methionine	met	M
14) phenylalanine	phe	F
15) proline	pro	P
16) serine	ser	S
17) threonine	thr	T
18) tryptophan	trp	W
19) tyrosine	tyr	Y
20) valine	val	V

Figure 1. shows table of Amino Acids.

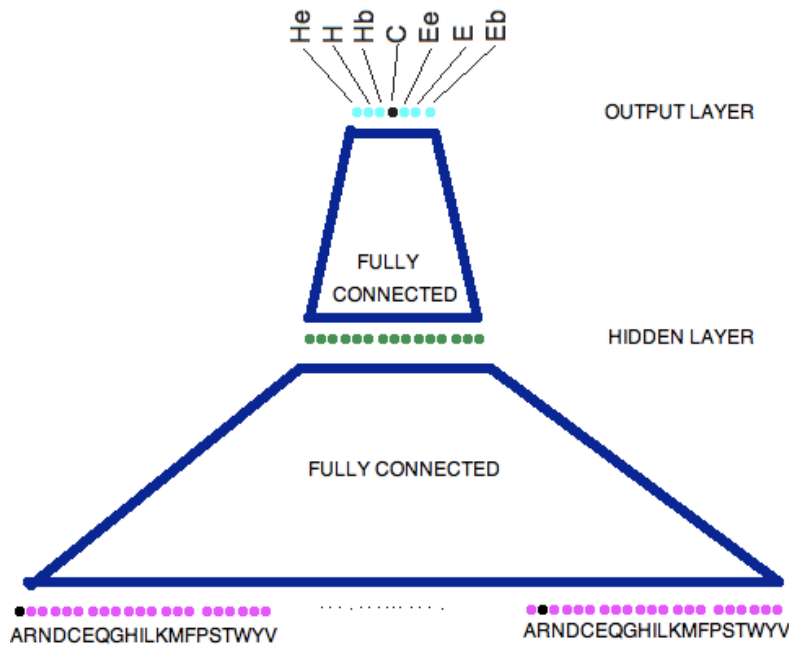


Figure 2. Shows Neural network module

The HMM module had 7 states, each labeled according to the outputs from the ANN module. The training data for the HMM module were from known DSSP output. The probabilities for transition between each states was calculated during training. [Viterbi Algorithm](#)(Durbin, 1998) was, then, used to determine the final output of the system when running. HMM generated 3 possible final output:H-Helix, E-strand and C for other

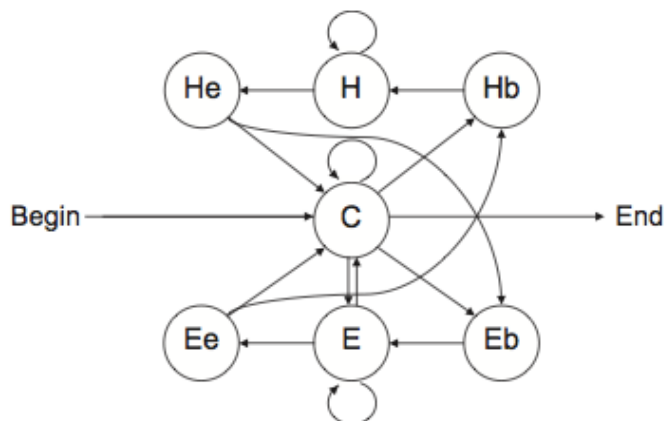


Figure 3. shows 7 states and transition flow of the HMM module. H for alpha-helix, E for strand and C for coil. 'b' for beginning and e for ends of the structure.(Image Source: Kuang Lin, Victor A. Simossis, Willam R. Taylor and Jaap Heringa, 2004).

Result

YASPIN was compared with many others secondary protein structure predictors. The accuracy was comparable, even though, YASPIN was trained without the test dataset while the other predictors were trained with the test data set. The accuracy

of YASPIN reached over 75% in some test. In addition YASPIN is much simpler, as the result, it could process data at much faster speed. YASPIN could predict protein secondary structure in seconds, much faster than other predictors.

Discussion

YASPIN was different to the other predictors, because of its 2 stages process. The ANN module produced 7 outputs label of possible structures, instead of 3 outputs like other predictors. With 7 outputs of the ANN module, it could capture the amino acid terminal sequence string of a structure, specially helix(Richardson and Richardson, 1988; Serrano and Fersht, 1989). The HMM module, in addition, provided the optimization for the final outputs from the ANN module. The HMM module looked at the output from ANN module, the Hb, He, Eb or Ee output would result in execution of [Viterbi Algorithm](#)(Durbin, 1998) to calculate the possibility of the final outputs: H, E or C.

References

Bishop,C.M. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford University Press, Oxford.

Durbin,R. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

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

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22,2577–2637.

Kuang Lin, Victor A. Simossis, Willam R. Taylor and Jaap Heringa. (2004) A simple and fast secondary structure prediction method using hidden neural networks. *BIOINFORMATICS*, Vol. 21 no. 2 2005, pages 152–159

Qian,N. and Sejnowski,T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*,202, 865–884.

Richardson,J.S. and Richardson,D.C. (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science*, 240, 1648–1652.

Williams, R. J. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1, 270-280

Serrano, L. and Fersht, A.R. (1989) Capping and alpha-helix stability. *Nature*, 342, 296-299.