Application of Artificial Neural Network (ANN): Development of Central-based ANN (CebaANN)

Su Yeon Jeong[1] , Tae Seon Yoon[2] and Chae Yoon Jeong[1]

[1]*Hankuk Academy of Foreign Studies, Student, 17035 Gyeonggi-do Yongin, Korea*
[2]*Hankuk Academy of Foreign Studies, Science and Information Department, 17035 Gyeonggi-do Yongin, Korea*

Abstract. Nowaday, the number of known protein structures is significantly less than the number of known amino acid sequences. It is because the regularity of amino acid depend on structure is not clear and the number of thermodynamic conditions are too many. There are some cases that discovering protein structure by experiment. However, It needs much time and cost for increasing the number of amino acid sequences, thus, there is less efficiency. So the empirical method which predict theoretically the structure of protein has been developed. We suggest Central-Based Artificial Neural Network as prediction method of protein structure. CebaANN can analyze similarity more detail by making part of center that affect outcome bigger. In experiment we got 85% of prediction probability at E structure, but we got 34% of probability at total.

1 Introduction

Nowaday, the number of known protein structures is significantly less than the number of known amino acid sequences. It is because the regularity of amino acid depend on structure is not clear and the number of thermodynamic conditions are too many. There are some case that discovering protein structure by experiment. However, It needs much time and cost for increasing the number of amino acid sequences, thus, there is less efficiency.

Empirical methods which theoretically predict the structure of protein has been proposed to overcome this limitation. First is ab initio method based on the thermodynamic hypothesis of Anfinsen. Second is method that using protein which is already discovered as template. First method is possible in a brief time, however, accuracy of the structure is unable to give us a satisfactory result. Therefore, second method which means two protein which have similar structure is usually used.

Because it is unable to compare existing protein sequences and structure, one by one, algorithm such as artificial neural network and Support Vector Machine is usually used at analyzing link between sequence and structure and deriving similarity by comparing result of general structure and result of input structure.

Artificial Neural Network is algorithm which was invented from neural network and consisting of input, hidden, output nodes. In this paper, we developed cebaANN which is application of ANN, and propose it is algorithm for protein structure prediction. CebaANN has existing structure [input - hidden - output]. However, it can analyze similarity more detail by making part of center that affect outcome bigger. In experiment we got 85% of prediction probability at E structure, but we got 34% of probability at total.

2 Relative research

2.1 Artificial neural network

An artificial neural network is an interconnected group of nodes, akin to the vast network of neurons in a brain. Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one neuron to the input of another. G and H structure is numbering by P(a), In G and H structure is numbering by P(a), machine learning and cognitive science, artificial neural networks (ANNs) are a family of statistical learning models inspired by G and H structure is numbering by P(a), biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of input and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.

For example, a neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function

(determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read.

Like other machine learning methods - systems that learn from data - neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition.

2.2 DSSP classification

The Protein structure is the biomolecular structure of a protein molecule. Proteins are polymers — specifically polypeptides — formed from sequences of amino acids. Each unit of a protein is called an amino acid residue because it is the residue of every amino acid that forms the protein by losing a water molecule. By convention, a chain under 40 residues is often identified as a peptide, rather than a protein.[1]To be able to perform their biological function, proteins fold into one or more specific spatial conformations, driven by a number of non-covalent interactions such as hydrogen bonding,ionic interactions, Van der Waals forces, andhydrophobic packing. To understand the functions of proteins at a molecular level, it is often necessary to determine their three-dimensional structure.

The Dictionary of Protein Secondary Structure, in short DSSP, is commonly used to describe the protein secondary structure with single letter codes. The secondary structure is assigned based on hydrogen bonding patterns as those initially proposed by Pauling et al. in 1951 (before any protein structure had ever been experimentally determined). There are eight types of secondary structure that DSSP defines:

- G = 3-turn helix (310 helix). Min length 3 residues
- H = 4-turn helix (α helix). Min length 4 residues
- I = 5-turn helix (π helix). Min length 5 residues
- T = hydrogen bonded turn (3, 4 or 5 turn)
- E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues
- B = residue in isolated β-bridge (single pair β-sheet hydrogen bond formation)
- S = bend (the only non-hydrogen-bond based assignment)
- C = coil (residues which are not in any of the above conformations)

In this paper, we only use 6 structures (G, H, I, T, E, B)

2.3 Chou-fasman parameters

The Chou-Fasman method of secondary structure prediction depends on assigning a set of prediction values to a residue and then applying a simple algorithm to those numbers. We were in the process of converting a short for each amino acid to a number, using this table.

Table 1. Chou-fasman parameters.

Protein Name	Chou-fasman parameters						
	P (a)	P (b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.07	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.11	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.05	0.117	0.128
Glutamic Acid	151	37	74	0.056	0.06	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.19	0.152
Histidine	100	87	95	0.14	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.07
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.12	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

3 Central-based artificial neural network (CebaANN)

Artificial neural networks are learning algorithm obtained an idea from the structure of the biology of neural networks. Artificial neural network is a model with problem solving that artificial neurons which form a network by binding synapse change weights in progress of learning. In here, it can compare binds of synapse to connection between nodes, and compare strength of bind to weight.

CebaANN is the application of artificial neural network that the structure is changed for making input value's affect more powerful as close to center input node. Unlike typical neural network that each input node is connected each hidden node, CebaANN has network that center input node is connected all of hidden nodes and as going outside, the number of hidden nodes which is connected input node decreases. For the structure, network must be constituted with the specific condition:

- output node should be one, and (n = the number of input nodes, m = the number of hidden nodes) m should be [(n+1)/2]

- threshold is (average of input node values / m) * sum of weights between input and hidden * sum of weights between hidden and output.
- threshold should be constant. so while in the progress, the sum of weight between input and hidden and between hidden and output should be constant.
- CebaANN is teacher learning method (if the target really have the result, base value is 1, and if not, base value is 0). So it have two case that apply the algorithm. Case 1 is output>threshold and base value is 0, Case 2 is output<threshold and base value is 1.

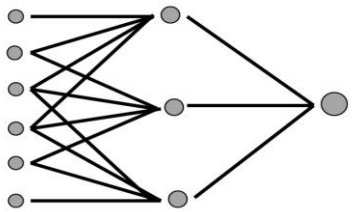

Figure 1. The structure of CebaANN_type 1.

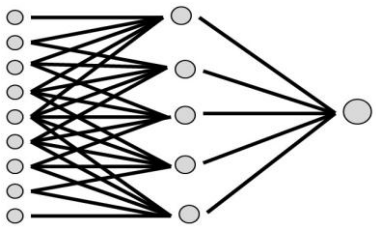

Figure 2. The structure of CebaANN_type 2

3.1. The method how change weights (Balancing method)

Balancing method :

If Case 1, follow 3.1.1 and 3.1.2 and make new output that applying changed weights. Repeat these three steps until output < threshold.

If Case 2, increase weights on the route of maximum input value (opposite of 3.1.1) and decrease weights on the route of minimum input value (opposite of 3.1.2) and make new output that applying changed weights. Repeat these three steps until output > threshold.

During the progress, if the number of Maximum or Minimum aren't one, decide the maximum for using as maximum which is close to center. if minimum is in same situation, decide the minimum like same, too. Because as go center, it's affect is more big.

3.1.1 Decreasing weights on the route of maximum input value

Decreasing input-hidden weight and hidden-output weight of the route whose input-hidden weight * hidden-output weight is minimum among the routes' that maximum input value passes. And increasing other input-hidden weights which is connected same hidden node as much as 1/(2*[m/2]) of change amount. As result, the sum of input-hidden weights is constant, and output is decreased.

3.1.2 Increasing weights on the route of minimum value

Increasing input-hidden weight and hidden-output weight of the route whose input-hidden weight * hidden-output weight is maximum among the routes' that minimum input value passes as much as change amount of a). And decreasing other input-hidden weights which is connected same hidden node as much as 1/(2*[m/2]) of change amount. As result, the sum of input-hidden weights and the sum of hidden-output weights are constant, and output is decreased.

4. Experiment (Using cebaANN)

We did experiment that getting weights which can predict protein structure when protein sequence pass the CebaANN.

4.1. Making material of experiment

Target protein sequence dataset : RS126. We use 2/3 of dataset as making special weights, and use others as measuring correct percent of prediction.

4.1.1 Separating target protein sequences

Seperating target protein sequence as amino acid sequence whose length is the number of input node (In this experiment, nine) from first amino acid. and increase the position of start amino acid as one. Now, we say this as input sequence.

4.1.2 Numbering each elements

Each element of input sequnces is numbering by chou-fasman parameters. G and H structure is numbering by P(a), and E and B structure is numbering by P(b), and T and S structure is numbering by P(turn). It's because G and H are helixs, and E and B are β–sheet or bridge, and T and S are turn or bend. And then, we can get three numbered sequence at one input sequence. Now, we say these three sequences as input group.

4.1.3 Matching input group with structure.

Matching input group with input sequence's disclosed structure. only one structure among six or 'none' is matching with input group. We say this matching as material. After, at the progress of running algorithm, elements' of input groups become input, and matched structure become base value.

4.2. Running CebaANN

Classifying materials into six groups of structure. And then, run CebaANN at each six group. During this progress, if material has targeted structure, base value become 1, and if not, base value become 0. After running is finished, special weights of each six group were made. First element of input group is runned at G, H group, second element of input group is runned at T, S group, Third element of input group is runned at E, B group.

4.3 Predicting structure.

At first, make protein sequence (which want to predict structure) as input group (4.1.1, 4.1.2). And then, passing input group at special weights of six groups.

If all of output is smaller than threshold, it has no structure. But if some of output is bigger, structure is decided by list of the highest-to-lowest parameter value of middle amino acid among structure whose output is bigger. If the number of highest parameter value isn't one, it would be decided as structure which has bigger output value. total.

5. Result

5.1. Special weights

5.1.1 Hidden-Output weights

special weights on the route between Hidden and Output.

Table 2. Hidden-Output weights.

(hidden)	1	2	3	4	5
G	9	236	260	200	295
H	328	180	188	172	132
E	67	109	209	272	343
B	244	152	144	164	296
T	648	337	148	-35	-98
S	396	-795	276	220	903

5.1.2 Input-Hidden weights

special weights on the route between Input and Hidden.

Table 3. Input-Hidden weights.

	1	2		(Input node) 3		
	(hid)1	1	2	1	2	3
G	-11078	3864	148	4200	241	193
H	353	105	187	438	449	359
E	-586	-250	-177	365	200	28
B	395	205	319	115	133	139
T	7966964	1966195	3835343	-3240436	-343323	457980
S	1.1E+180	-3E+179	1.7E+182	-3E+179	-2E+180	1E+179

(Input node) 4				5	
(hidden)1	2	3	4	1	2
2372	151	118	172	1637	112
44	78	75	79	43	137
990	352	198	120	474	214
205	224	224	188	83	69

-2957871	-4678738	-6698963	33440169	-3733851	461165
-3E+179	3.6E+179	1E+179	8.8E+178	-3E+179	4.4E+178

(Input node) 5			6			
3	4	5	hidden)2	3	4	5
119	27	63	351	333	427	257
92	18	48	152	145	158	204
153	59	98	409	213	125	107
50	43	1	254	279	238	135
2907195	-4.6E+07	46248755	726552	823860	27577772	-4.3E+07
1E+179	8.8E+178	3.7E+178	-2E+182	-4E+179	-4E+179	-1E+179

(Input node) 7			8		9
(hidden)3	4	5	4	5	5
238	247	157	127	82	442
330	308	421	438	605	-266
412	274	222	423	338	239
309	283	159	248	219	483
2510930	-2.1E+07	1888390	5975314	21967080	-2.8E+07
1E+179	8.8E+178	3.7E+178	8.8E+178	3.7E+178	3.7E+178

5.2. Correct percent of prediction

Correct percent : (the number of sequence which predicted structure equals the target structure) / (the number of sequence which has the target structure)

Table 4. percent of prediction

structure	Total	G	H	E	B	T	S	None
(%)	34.58	0.0	0.0	85.36	0.0	0.0	0.0	8.69

6. Conclusion

Nowaday, the prediction method of protein structure is developed so much. However the necessity of exact prediction method of protein is still exist. There are many amino acid sequence which doesn't discovered its structure. And because protein structure is clue to find protein's function, prediction method can help many people as diverse way. CebaANN's prediction is only act at E structure, but it's percent is high. It can prove that CebaANN has possibility to be accurate prediction method. We need to find why it only act at E structure and why it doesn't work at other structure. So we would expand experiment ahead. The limitation was it only do work at E structure. And We will make improvements at that part. This experiment was meaningful that we do new challenge.

References

1. Machin, David. "Biomathematics." (1976)
2. Taylor, William Ramsay. "The Classification of Amino Acid Conservation." *Journal of Theoretical Biology* 119.2 (1986): 205-18

3. P. Prevelige, Jr. and G. D. Fasman, Chou-Fasman Prediction of Secondary Structure, in Prediction of Protein Structure and the Principles of Protein Conformation, ed. G. B. Fasman (ISBN 0-306-43131-9, Plenum, New York)