International Journal of Genomics, Proteomics, Metabolomics & Bioinformatics (IJGPMB) ISSN: 2577-4336

Protein Eight Secondary Structure Classes Prediction Using Artificial Neural Networks

Research Article

Hanan Hendy*, Wael Khalifa, Mohamed Roushdy, Abdel-Badeeh M. Salem.

Teaching Assistant, Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.

Abstract

Protein is considered the backbone of any human being. Protein is responsible for many functionalities in the human body, these functionalities differ according to the way protein amino acids (amino acids are the raw elements of protein) bond together. Then the protein forms its secondary,tertiary and quaternary structures from the amino acid structure (primary sequence) by forming hydrogen bonds. Many machine learning techniques have been used through the past decade to try to predict the protein secondary structure. The most commonly used paradigm was the Artificial Neural Networks. A lot of research was conducted in this field. This paper presents the usage of Artificial Neural Networks to predict the protein secondary structure. The difference this paper proposes is predicting the eight classes of secondary structure not only the three main classes named: alpha, beta and coil. The maximum accuracy reached is 71% which is better that other discussed methods.

Keywords: Artificial Neural Networks; Bioinformatics; Machine Learning; Protein Secondary Structure Prediction.

Introduction

The importance of protein lies in that it is responsible for all body functions. The protein performs many functions such as alcohol dehydrogenase oxidizes alcohols to aldehydes or ketones, hemoglobin carries oxygen, insulin controls the amount of sugar in the blood and much more. The protein secondary structure extraction is a very complex task and requires a lot of tools for scientists to be able to perform it. Moreover, it is important for detecting disorders as well as helps in the studies done on tertiary structures [1].

The first used ANN was presented by Ning Qian and Terrence J. Sejnowski. They used a non-linear neural network. They reached an accuracy of 64.3% [2]. Then, John-Marc Chandonia and Martin Karplus used neural networks to predict not only the secondary structure but also the structural class of the primary amino acid. They reached 62.64% of accuracy [3]. Later, Gianluca Pollastri et al., used a bidirectional neural network to reach an accuracy of 78% [4]. In this paper we show the difference when training the ANN to predict the three main classes of protein secondary structure versus predicting the eight classes. The conducted experiments performed with and without post processing to group the results. This is shown the results as well.

Data preprocessing

The discussed experiments use a dataset extracted from the Protein Data Bank (PDB) [5]. The dataset is named CB513 [6, 7]. This dataset consists of 513 different protein sequences. The length of protein sequences ranges from 20 amino acids to 754 amino acids with average 164 amino acid. This inconsistency in the protein length raised a problem. This is because the protein sequences cannot be used directly in the artificial neural network triaging. The way to handle this issue is to choose a window size and split the data accordingly. Then the data is converted to a matrix like form to be ready for ANN training.

The given dataset contains files for each sequence. These files are appended together then split based on the window size. This process resembles the process of predicting the mid amino acid in the chosen window. So, if the chosen window size is 7, then the split process predicts the amino acid number 4 in respect to the 7

***Corresponding Author**:
Hanan Hendy,
Teaching Assistant, Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.
E-mail: hanan.hendy@cis.asu.edu.eg

Received: August 26, 2016
Accepted: September 16, 2016
Published: September 21, 2016

Citation: Hanan Hendy, Wael Khalifa, Mohamed Roushdy, Abdel-Badeeh M. Salem (2016) Protein Eight Secondary Structure Classes Prediction Using Artificial Neural Networks. *Int J Genomics Proteomics Metabolomics Bioinformatics.* 1(4), 17-19. **doi:** http://dx.doi.org/10.19070/2577-4336-160004

amino acids in the window.

For example,

151703022006121812150803171706011602091701091301517030220061218121508031717060116020917010913015170302200612181215080317170601160209170109130

After this stage the input and output matrices of the dataset can be processed. The input matrix is the combination of the windows split in the previous step column based and the output matrix is a group of columns with zeros and one at the place of the corresponding secondary structure.

Methods and Results

After the data is ready. Artificial Neural Network [8] is built to be trained and tested with this dataset. MATLAB R2013a neural network toolbox was used for implementing the ANN. The experiments were conducted with different paradigms:

- Window size: 17, 19, 25 and 31.
- Number of Hidden Layer and Neurons per layer.
    - Single hidden layer with different number of neurons. (10 and 3).
    - Multiple hidden layer with different number of neurons (10, 3 and 3, 10).
- Input format: Numeric format and Binary format.

The build ANN was used to predict the three main classes of protein secondary structure, then used to predict the eight classes of protein secondary structure, last it was used as the second experiment but then post processing was applied to combine the results back same as the first experiment. The combining process (post-processing) is done based on Table 1.

The results are summarized in table 2 and 3 where they show the difference between using binary and numeric encoding for input amino acids sequence respectively. The highest among the three experiments in each row is highlighted.

It is seen from the tables that when using the numeric encoding experiment 2 and 3 always shows a better accuracy unlike

Table 1. Protein secondary structure classes

Letter Code	Secondary Structure	Encoding	Combined Encoding
G	3-turn helix	01	01
H	4-turn helix (α helix)	02	
I	5-turn helix (π helix)	03	
E	extended strand in parallel and/or anti-parallel β-sheet conformation	04	02
B	residue in isolated β-bridge	05	03
S	bend	06	
T	hydrogen bonded turn	07	
C	coil	08	

Table 2. Prediction accuracy results using numeric input

Window size	Number of Hidden layers	Experiment 1	Experiment 2	Experiment 3
17	10	48.76%	53.62%	**57.01%**
	3	48.99%	53.70%	**56.93%**
	10 3	49.00%	53.73%	**56.98%**
	3 10	49.16%	**53.56%**	52.01%
19	10	46.16%	53.70%	**57.24%**
	3	46.58%	53.61%	**56.88%**
	10 3	46.30%	52.02%	**55.86%**
	3 10	46.37%	53.51%	**56.80%**
25	10	48.46%	53.79%	**57.11%**
	3	49.04%	53.64%	**56.91%**
	10 3	49.06%	53.68%	**56.83%**
	3 10	47.59%	52.84%	**56.36%**
31	10	48.59%	53.77%	**57.20%**
	3	46.45%	53.87%	**57.03%**
	10 3	46.57%	52.01%	**55.72%**
	3 10	46.43%	53.55%	**56.61%**

Table 3. Prediction accuracy results using binary input

Window size	Number of Hidden layers	Experiment 1	Experiment 2	Experiment 3
17	10	**64.71%**	52.01%	55.73%
	3	65.75%	63.06%	**70.89%**
	10 3	**65.83%**	60.41%	64.56%
	3 10	**64.71%**	56.87%	55.75%
19	10	59.79%	63.42%	**70.75%**
	3	**64.40%**	58.40%	63.39%
	10 3	**59.53%**	52.01%	55.72%
	3 10	**59.79%**	52.01%	55.72%
25	10	65.57%	60.34%	**66.80%**
	3	**67.84%**	52.01%	55.72%
	10 3	**66.98%**	59.49%	61.40%
	3 10	**65.57%**	52.01%	55.72%
31	10	**64.55%**	59.83%	62.52%
	3	**64.66%**	52.01%	55.72%
	10 3	**64.53%**	59.07%	62.48%
	3 10	**64.55%**	52.32%	55.84%

using the binary encoding. This is because when using numeric encoding the input nodes always have a probability to be used as features within the prediction process unlike the binary in which it is take or leave. That is why when we predict the eight classes accuracy increases. Moreover, experiment two did not show any significant change in the accuracy and this concludes to the point that it cannot be used unless there is a need for prediction the eight classes which is not the case. The common thing is that we need to predict the three main classes.

Conclusion

This paper discusses how artificial neural networks can be used to predict the protein secondary structure. Three experiments were conducted and compared one to other. The difference between the experiments was the prediction output. The highest accuracy reached was in the third experiment with binary format and a neural network with single hidden layer with 3 nodes. The accuracy reached was 71% which is better than 64.3% [2] and 62.64% [3]. However, it is lower than the one discussed in [4] that reached 78%. But this comparison is not accurate as the used environments and datasets vary. From the above findings, ANN suits to be used as a prediction paradigm for protein secondary structure, but predicting the eight classes on its own doesn't show a proper added value so it is recommended to predict the three classes directly or predict the eight classed and then combine them back to the three classes.

References

[1]. R. O. Esquivel, Carolina Barrientos, Catalina Soriano, Frank Salas, Jesús S. Dehesa, et al., (2013) Decoding the building blocks of life from the perspective of quantum information. Advances in Quantum Mechanics, Intech Open minds.

[2]. Ning Qian, Terrence J Sejnowski (1988) Predicting the secondary structure of globular proteins using neural network models. J Mol Biol. 202(4) : 865-884.

[3]. Chandonia, John-Marc, Martin Karplus (1995) Neural networks for secondary structure and structural class predictions. Protein Science. 4(2): 275-285.

[4]. Pollastri Gianluca, Darisz Przybylski, Burkhard Rost , Pierre Baldi (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins. 47(2): 228-235.

[5]. "PDB".http://www.wwpdb.org.

[6]. "Cuff and Batron Dataset". http://comp.chem.nottingham.ac.uk/disspred/datasets/CB513.

[7]. Chopra Paras, Andreas Bender (2006) Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature. In Silico Biol. 7(1): 87-93.

[8]. Leverington David (2015) A Basic Introduction to Feed forward Back propagation Neural Networks.