ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 1387 -1391
doi: 10.18178/wcse.2017.06.240

Prediction of Secondary Structures of Hemoglobin Using Clonal Selection Algorithm

Burcu ÇARKLI YAVUZ [1 +], Cengiz SERTKAYA [2] and Nilüfer YURTAY [2]

[1] Department of Information Systems Engineering, Sakarya University, 54187, Sakarya, TURKEY
[2] Department of Computer Engineering, Sakarya University, 54187, Sakarya, TURKEY

Abstract. Protein structure prediction is one of the most important research areas in bioinformatics. Knowing the structure of the protein provides significant information about the function of protein. However, it is a challenging process to identify the 3-dimensional structure of a protein. Methods such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy require long duration and high costs, furthermore, these methods are not appropriate for every protein. Artificial intelligence and heuristic methods are preferred in the prediction of 3-dimensional structures of proteins recently because of their significant contribution. In this paper, the use of the Clonal Selection Algorithm (CSA) method for protein secondary structure prediction is studied and the results are compared to other artificial intelligence and heuristic methods. Artificial immune system (AIS) based CSA predicted hemoglobin secondary structure with a high prediction accuracy of 88.38 %.

Keywords: Artificial immune system, Clonal Selection Algorithm, Hemoglobin, Protein secondary structure prediction.

1. Introduction

Proteins are the most complex and functional molecules known and indicate an effective activity in every biological function in a cell. There exist 20 types of amino acids which have different chemical characteristics. Protein molecules consist of a long amino acid chain that is connected to each neighbor with a covalent peptide bond. Existence or absences of these bonds, folding shape or formations identify the structure of proteins. Biologists define protein structure on four levels: primary, secondary, tertiary and quaternary [1].

In the hierarchical classification of protein structure, the secondary structure is a significant level and to identify protein features for the fold recognition secondary structure is used. To predict three-dimensional structure accurately prediction of secondary structure of proteins is essential [2].

The first attempts to predict secondary structure are seen in the 1970s. Chou-Fasman [3] and GOR methods [4] were mainly built on statistical analysis of single residue.

Prediction accuracy dramatically increased with the start of using the machine learning algorithms such as Artificial Neural Networks [5], Support Vector Machines [6,7] and Hidden Markov Models [8] in predicting of protein secondary structure. Heuristic methods are commonly preferred due to difficulties in predicting protein structures in laboratory environments [9]. State-of-the-art methods that are exclusively or partly based on neural networks are PHD [10], PHDpsi [11], PSIPRED [12].

Within the last decade, new computational intelligence models have been proposed which are based on the natural immune system. Of particular interest is the AIS model, based on the natural immune system's process of saving the body from viruses and unfamiliar invaders.

+ Corresponding author, Tel: +90 506 532 65 11
E-mail address: bcarkli@sakarya.edu.tr.

AIS-based algorithm approaches, which are characterized by natural problem learning and solving, are widely used in a variety of applications such as learning, anomaly detection and optimization. Some specific examples include: Robot control [13], Diagnosis of diabetes [14], Outbreak detection [15].

There are many techniques utilized to develop AIS-based algorithms, with the Clonal selection based algorithm being the most popular. Clonal selection based theory was first described in 1959 [16], and states that cells with receptors circulate the host organism, then these receptor cells recognize the antigen and proliferate via duplication (cloning). The resultant increase of cells causes the organism to develop a more specific response to the antigen, triggering the immune system to produce antibody cells. These antibody cells recognize and decide whether to save or destroy the antigen.

In this study CSA, a recent method of machine learning algorithms is used to predict protein secondary structure. In literature CSA is not commonly employed in protein secondary structure prediction, like prediction of protein cellular localization sites [17], therefore possible higher prediction accuracies by using CSA algorithm may be an important step in protein prediction studies [18]. Hemoglobin protein is taken as the first input protein for the method developed in this paper. After the prediction accuracy by using CSA is obtained it will be compared to the other studies which used hemoglobin as the input protein.

2. Materials and Methods

Protein structure prediction is a significant subject in medical and biological sciences, as mentioned in section 1. With the help of heuristic methods, it is possible to predict protein secondary structure faster, more accurate and more economical. Beginning from this point of view application of CSA for hemoglobin protein is realized as indicated in the flowchart given in Fig. 1.

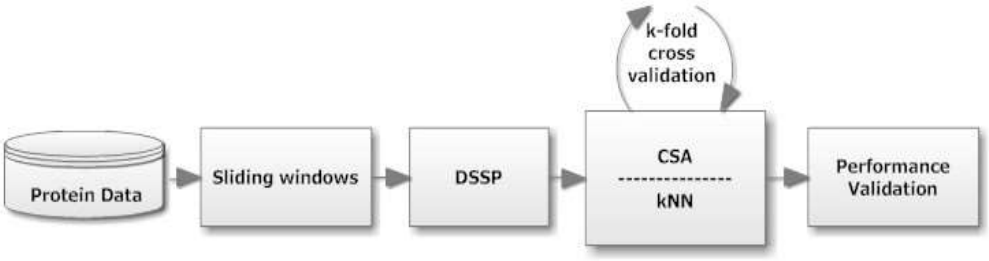

Fig. 1: Flowchart of prediction of hemoglobin secondary structure

In this study, the applicability of CSA is tested on a protein that contains 22 different structures and sequences of Hemoglobin protein dataset which consist of 3336 amino acid residues. Protein structures are presented to researchers in Protein Data Bank (PDB) and, first and secondary structures of Hemoglobin protein are gathered from PDB [19].

To assign the secondary structure to the amino acids of a protein DSSP [20] method can be used. The DSSP algorithm (Dictionary of Secondary Structure of Proteins) is used to reduce secondary structure classes from 8 to 3 which are helix (H), strand (E), and Coil (C) because this is the most widespread method preferred.

From an information processing point of view, the natural immune system is a learning and solving system. Here the antigen (Ag) is a problem to solve and Antibody (Ab) is a solution. In this study, an artificial immune model based on clonal selection theory is presented.

In order to describe the algorithm well, the notations of proposed algorithm should be defined as follows,

The antigen is the objective function to be optimized. Ag can be represented by a single dimensional array of attributes.

$$A_g = \{ a_1, a_2, a_3, \dots a_k \} \tag{1}$$

The antibody is the representation of solution candidates. Ab can be represented by a single dimensional array of attributes.

$$A_b = \{ a_1, a_2, a_3, \ldots a_k \} \quad (2)$$

In Ag and Ab, a is an attribute (amino acid) value and k is attribute count in one single row data.

Antibody population (P) is n-dimensional group of antibody. Here, n is the number of antibodies in population.

$$P = \{ Ab_1, Ab_2, \ldots Ab_n \} \quad (3)$$

Affinity is the fitness measurement between antigen and antibody. The purpose of this process is to select antibody with the highest affinity. In this study, the affinity's selection is identified through Hamming Distance.

$$Hamming\ (Ag, Ab) = \sum_{i=1}^{n} |Ag_i - Ab_i| \quad (4)$$

After the affinity values are calculated, antibody with larger affinity will clone to produce good antibody population and accelerate the convergence rate of the algorithm. Antibodies in this cloned population are subjected to a mutation process. In mutation process, two antibodies in the same class are selected and at least two random attributes are exchanged among these antibodies.

After the clone and mutation process, the best examples of this group are selected by using affinity measurement in Eq. (4) to compose for the next generation.

If the number of samples in population is changed, all these steps are repeated. In the absence of change in the number of samples, the training process has been completed. This is then used to evaluate and predict the new antigen dataset.

Fig. 1 indicates the flowchart of CSA. It can be summarized as follows [21]:

Step 1: Initialization: generate antibody population (P) from training dataset and choose random attribute count that will be used in mutation.

Step 2: Find antibodies which best matched with antigen by using affinity measurement.

Step 3: Clone these selected antibodies and randomly mutate between them.

Step 4: Calculate affinity of antigen with newly generated antibodies. If these antibodies are better than the old ones, then they replace the old ones in permanent population and go to step 2 otherwise go to step 5.

Step 5: CSA is ready for predicting test dataset.

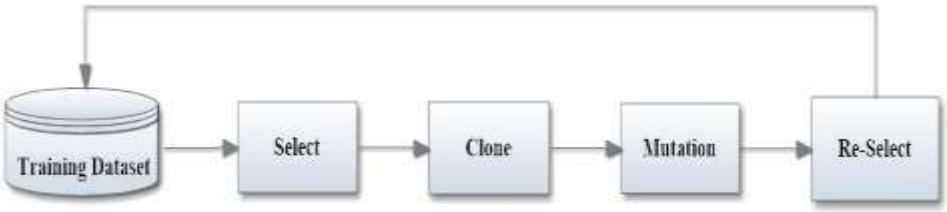

Fig. 2: Flowchart of CSA

In this study, k-fold cross validation technique is used and full dataset is randomly split into 10 folds. AIS algorithm is trained and tested 10 times. In each case, one fold is taken as test dataset and the remaining 9 folds are added to form training dataset for AIS algorithm. Thus 10 different test results are obtained. The average of these results gives the accuracy of the algorithm.

AIS model performances are evaluated by the three-state per-residue accuracy method (Q3) which is commonly used to measure of protein secondary structure prediction performance worldwide. In Q3 method ratio of number of residues in state (x) to correctly predicted the number of residues for H, E and C by AIS (y). Higher value of Q3 shows a better result. Q3 equation is given in Eq. (5).

$$Q_3 = \frac{y}{x} * 100 \quad (5)$$

3. Results and Discussion

The proposed AIS method is implemented in C#.NET by using Framework 4.5. System accuracy measured through the Hemoglobin data.

Hemoglobin protein's structures are divided into window sizes such as 11, 13, 15, 17, 19 and 21. Simulating the algorithm given in this study for all window sizes following results are obtained. The highest ratio of success in prediction of hemoglobin secondary structure using CSA method's training and tests with different window sizes is reached with 15 window sizes.

Resulting prediction accuracies and comparison of these to other methods used in different studies related to hemoglobin protein are given in Tab. 1.

Table I: Comparison Of Classification Performance between CSA and other Methods To Predict Hemoglobin Secondary Structure

Method	**Classification accuracy**	**Method**	**Classification accuracy**
SOGR (Self Organizing Global Ranking) [5]	87.75	Gaussian Radial Basis Kernel [7]	77.64
SOM (Self Organizing Maps) [5]	86.99	C-F (The Chou and Fasman) [22]	78
GRNN (Generalized Regression NN) [5]	85.93	GOR (The Garnier, Osguthorpe, and Robson) [22]	62
MLP (Multilayer Perceptron) [5]	80.94	PHD (The EMBL Profile NN) [22]	86
PNN (Probabilistic Neural Network) [5]	86.08	PSA (The Protein Sequence Analysis) [22]	60
Sigmoid Kernel [7]	70.05	**CSA**	**88.38**
Polynomial Kernel [7]	73.25		

In this paper, clonal selection principle based artificial immune model is outlined to obtain a satisfactory approach to estimate the secondary structure of proteins. The applied model is useful to obtain simulations under different window sizes. This AIS model achieves better results than classical neural network models due to the benefits of a population of solutions, the evolutionary selection pressure and mutation.

Hemoglobin is an appropriate protein that used in secondary structure prediction studies, and it is possible to compare the results with the previous studies. Hemoglobin secondary structure prediction is made with CSA and a satisfying result is obtained.

As a future work, it will be tried to increase the secondary structure prediction accuracy with making some changes in original CSA algorithm and to create a hybrid method. More complex protein structures can be worked with AIS models. An improved training technique such as attribute weighting procedure can be added to AIS and the accuracy can be improved.

4. Acknowledgments

This study was supported by Research Fund of the Sakarya University. Project Number: 2013-50-02-029.

Authors are also grateful to Assoc. Prof. Numan Çelebi from Sakarya University and Dr. Lillian Clark from the University of Portsmouth, for their valuable comments that greatly improved the manuscript.

5. References

[1] B. Alberts, AJ. Lewis, M. Raff, K. Roberts, P. Walter. *Hücrenin Moleküler Biyolojisi*, Garland Science, Taylor and Francis, 2008.

[2] YY. Ji, YQ. Li. The role of secondary structure in protein structure selection, *Eur Phys J E Soft Matter* 32 (1):

103–107, 2010.

[3] PY. Chou, GD. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence, *Adv Enzymol Relat Areas Mol Biol* 47: 45–148, 1978.

[4] J. Garnier, DJ Osguthorpe, B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *J Mol Biol* 120: 97-120, 1978.

[5] E. Atar. *Yapay Sinir Ağları ile Proteinlerin İkincil Yapılarının Kestrimi*, MSc Thesis, Yildiz Technical University, 2005.

[6] B. Yang, Q. Wu, Z. Ying, H. Sui. Predicting protein secondary structure using a mixed-modal SVM method in a compound pyramid model, *Knowledge-Based Systems* 24(2): 304-313, 2011.

[7] T. Ibrikci, A. Cakmak, I. Ersoz, OK. Ersoy. Hemoglobin secondary Structure Prediction with Four Kernels on Support Vector Machines, *ICSC Congress on Computational Intelligence Methods and Applications*, 2005.

[8] K. Asai, S. Hayamizu, K. Handa. Prediction of protein secondary structure by the hidden Markov model, *Comput Appl Biosci*. 9: 141-6, 1993.

[9] B. Çarklı Yavuz, N. Yurtay, HB. Dinçtürk Botofte, MŞ. Arslan. Proteinlerin İkincil Yapılarının Tahmininde Heuristik Yöntemler, *International Symposium on Innovative Technologies in Engineering and Science*, Sakarya, Turkey, 2013.

[10] B. Rost, C. Sander. Prediction of Protein Secondary Structure at Better than 70% Accuracy, *Journal of Molecular Biology* 232(2): 584–599, 1993.

[11] D. Przybylski, B. Rost. Alignments grow, secondary structure prediction improves, *Proteins* 46(2): 197-205, 2002.

[12] DT. Jones. Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol*. 292(2): 195-202, 1999.

[13] HYK. Lau, VWK. Wong, ISK. Lee. Immunity-based autonomous guided vehicles control, *Applied Soft Computing* 7(1): 41–57, 2007.

[14] MR. Bozkurt, N. Yurtay, Z. Yılmaz, C. Sertkaya. Comparison of Different Methods For Determining Diabetes, *Turkish Journal of Electrical Engineering & Computer Sciences* 22: 1044-1055, 2014.

[15] M. Mousavi, A. Abu Bakar, S. Zainudin, Z. Awang Long, M. Sahani, M. Vakilian. Negative selection algorithm for dengue outbreak detection, *Turkish Journal of Electrical Engineering & Computer Sciences* 21: 2345-2356, 2013.

[16] FM. Burnet. *The clonal selection theory of acquired immunity*, Vanderbilt University Press, 1959.

[17] I. Turkoglu. A hybrid method based on artificial immune system and k-NN algorithm for better prediction of protein cellular localization sites, *Applied Soft Computing* 9(2): 497-502, 2009.

[18] V. Cutello, G. Morelli, G. Nicosia, M. Pavone, G. Scollo. On discrete models and immunological algorithms for protein structure prediction, *Natural Computing* 10 (1): 91-102, 2011.

[19] J. Zhang, Z. Hua, JR. Tame, G. Lu, R. Zhang, X. Gu. The crystal structure of a high oxygen affinity species of haemoglobin (bar-headed goose haemoglobin in the oxy form), *Journal of Molecular Biology* 255(3): 484–93, 1996.

[20] W. Kabsch, C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* 22(12): 2577–637, 1983.

[21] LN. De Castro, FJ. Von Zuben. *Artificial Immune Systems: Part I – Basic theory and applications*, Technical report, 1999.

[22] SR. Krystek, WJ. Metzler, J. Novotny. Protein Secondary Structure Prediction, *Current Protocols in Protein* Science, 2001.