Identifying Protein Secondary Structures and Detection of Complexities using Deep Neural Network

Subhendu Bhusan Rout[1]

Department of CSE&A, IGIT Sarang, Odisha, India

subhendu.as@gmail.com

Dillip Kumar Swain[2]

Department of CSE&A, IGIT Sarang, Odisha, India

dillipswain.41@gmail.com

Sasmita Mishra[3]

Department of CSEA, IGIT Sarang, Odisha, India

sasmita.mishra@gmail.com

Abstract

The Protein molecule is known as the large biological molecule in a living organism. The protein performs several works like transporting molecules, catalysing metabolic reaction, responding to stimuli etc in a human body. Protein Structure analysis and prediction is very much essential to make any research about the same protein molecule. The basic intention of protein structure prediction (PSP) is to predict the three dimensional structure that generate by the amino acid sequence. The very peculiar matter is only twenty amino acid found in a living body where as approximately one lakh protein molecules can be framed from the same amino acid compositions in different percentages. The three dimensional structure framed by the amino acid compositions generally changes its shape and size due to the effect of external agents or medicines that comes in contact with these protein molecules. The basic intention behind the prediction of structure of the protein is to design new drugs or medicines. From the structures the medicine researchers working for the development of medicines may easily detect the changes in the living body or the requirement of drugs or medicines. The detection of the structure and the prediction of perfect structure is always a challenging task. The protein structure is basically a three dimensional structure in its secondary transformation. The structure may be in the form of α Helix, β sheets or loop etc. In this paper the identification of the secondary structures and the percentages of α Helix, β sheets or loop structures are being predicted and the probable complexities that may occur during the prediction is discussed. Deep neural network is a deep structured learning process is an application of the broader family machine learning. Deep learning architectures has a number application in various fields

like medical science, bioinformatics, medical image analysis etc. A novel method is being proposed in this research article for the detection, correction and removal of various complexities during prediction using deep neural network. This technique will be helpful for different researchers working in the field for drug design and medicine research.

Key Words- *Protein, Deep Neural Network, PSP Problems, Amino Acids, α Helix, β sheets, Bioinformatics.*

I. INTRODUCTION

Protein Structure prediction is an important task of bioinformatics and bioinformatics is the applications of computer aided technologies in the field of medical science. Protein structure prediction is the basic work that is the prediction of the three dimensional structures that is commonly generated by the amino acid sequences. The three dimensional structure framed by the amino acid compositions generally changes its shape and size due to the effect of external agents or medicines that comes in contact with these protein molecules. The idea behind the prediction of structure of the protein is to design new drugs or medicines. From these structures the medicine researchers working for the development of drugs; can easily detect the affects, diseases and its requirements of medicines. The protein structure generally creates different shapes and structure in its three dimensional structures. The sequences may be in the form of α Helix, β sheets or loop etc. Similarly twin structure removal during protein structure prediction is one more important task because due to similar structures many times it creates problem in detection of exact percentage of different structures.

Deep Neural Networks (DNN) is one of the popular method for machine learning and is being widely used in many fields. DNN is basically inspired by structure of mammalian visual system. Generally the mammalian visual system contains many layers of neural network. It processes information from retina to visual center layer by layer. After this it extracts edge feature, part feature, shape feature and eventually forms the overall concept of the picture [1]. Basically the depth of DNN is greater than or equal to four. So on the other way a Multilayer Perceptron (MLP) with more than 1 hidden layer may be a DNN framework. The basic works of DNN is to extracts feature layer by layer and combine low-level features to regenerate the high-level features. A 3-layer deep learning algorithm is used to train the deep neural network model.

This Paper is organized as follows. In section-II we have discussed about Protein Structure prediction with its necessity, real time application and developments. It provides a brief idea about the detailed background of Protein Structure prediction. In section-III we have discussed Deep Neural Network and its properties. The proposed work is being discussed in section-IV in detection of complexities. The section-V concludes with conclusion and future work.

II. PROTEIN STRUCTURE PREDICTION

Protein Structure prediction is the problem deals with the prediction of the three dimensional structure from the amino acid sequences of a protein molecule. In general the proteins are the large biological

molecules available in a living body which contain large number of amino acid sequence with a maximum of 20 types of amino acids. The amino acid sequence in contact with external agents changes its shape in their secondary, territory and quaternary stages. Though this process is a biological process but the analysis of this PSP problem may helpful for the design of new drugs and medicine during the medical researches [2].

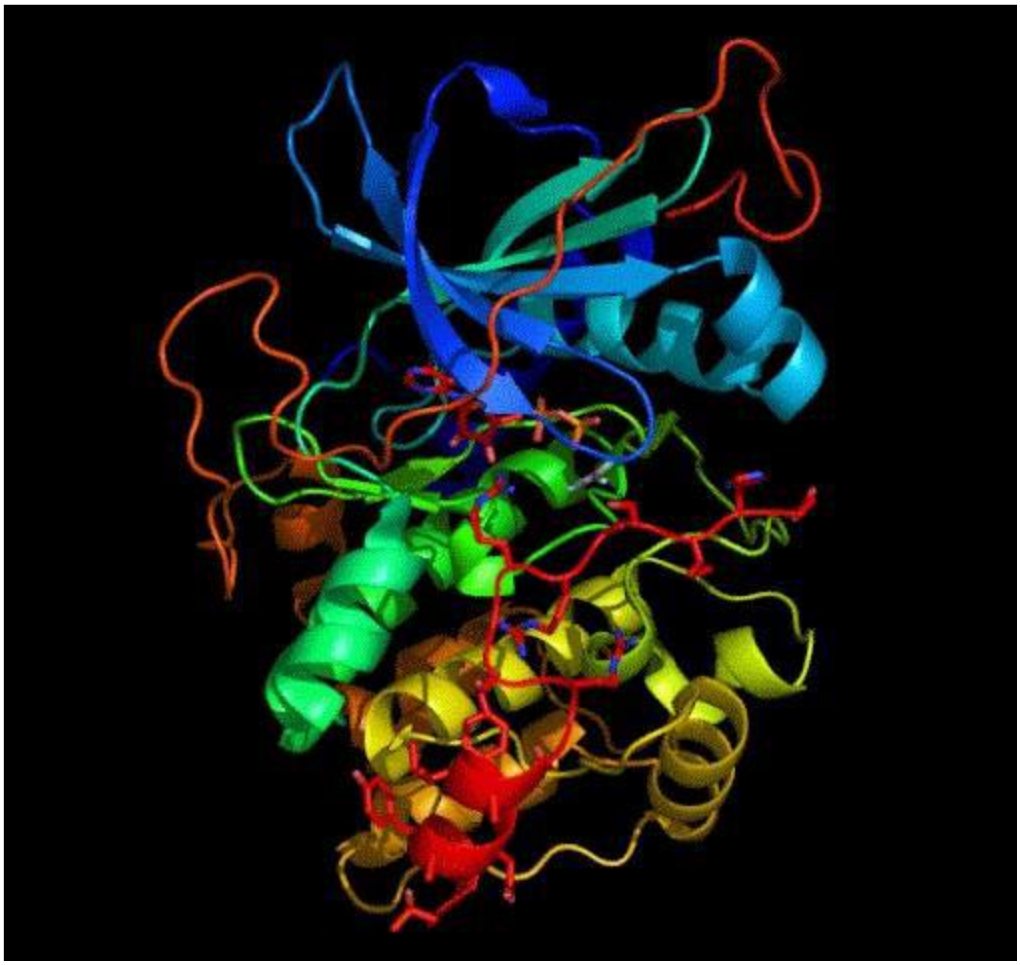


Fig 1. 3D structure of a protein

The Fig.1 represents the three dimensional structure of the protein. These shapes changes from time to time. This structure is having three type of structures like α Helix, β sheets or loop The twenty amino acid composition generally creates different shapes and structure in its three dimensional structures. This typical structure consists of different percentage of α Helix, β

sheets or loop structures. Similarly in a protein structure there may consist of number of twin structures which one more important task to remove the twin structure. It is very much important for removal of twins as well as the matching structures. In [3] M. T. Hoque et al. provided an algorithm for the purpose of twin structure removal. They have also proposed one more technique to provide near similar cases from the population. Twin removal is a part of protein structure prediction regarding the prediction of successful structure. Twin removal [2][3] algorithm has a big application for the problems of removing similar individuals from the population, with applying different searching techniques which makes an obstruction regarding the production of optimized results in this matter. The same proposed scheme is also applicable for the removal of twins during the collections of information or genetic molecules from a particular group of population.

GA+ is the application of genetic algorithm framework for PSP problems [6]. It uses HP model and FCC lattice. HP model is nothing but the Hydrophobic Polar energy model that is proposed in [7]. In this model amino acids are divided into two groups like hydrophobic and Hydrophilic. Each amino acid sequence is as a string‘s’ of the alphabet H & P. Similarly lattice (L) points are generated upon the bases of some vectors ranging from v1.....vk. Two lattice points p, q ϵ L are said to be the neighbors of each other if p = q + vi for some vectors of vi in basis of the Lattice (L). The Protein conformations generally represented by a sequence of basis vectors which is also one lattice that represent in Fig.2 [6].

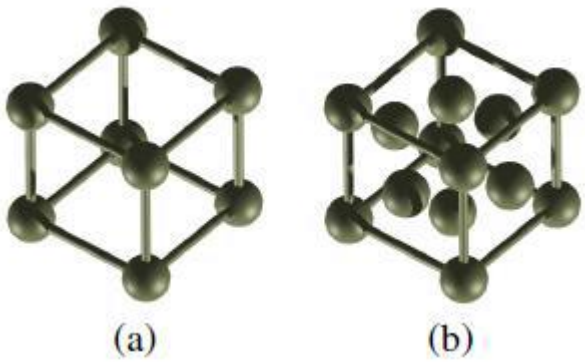


Fig 2. (a) Cubic Lattice, (b) FCC lattice

In molecular biology protein structure describes the various levels of organization of protein molecules. According to size Proteins are nanoparticles that contain polymers of amino acids. Each protein polymer (also known as a polypeptide) consists of a sequence formed from 20 possible L-α-amino acids, also referred to as residues. For chains under 40 residues the term

peptide is frequently used instead of protein. Protein structures range in size from tens to several thousand residues. Very large aggregates can be formed from protein subunits. A protein may undergo reversible structural changes in performing its biological function. The alternative structures of the same protein called as different conformations and transitions between them are called conformational changes [2].

III. DEEP NEURAL NETWORK AND ITS PROPERTIES

The traditional neural network needs to learn all the time to solve various tasks in a specified manner. It also used various methods to provide better and better result. As it receives new information in any system it again starts learning with the new situation.

The learning process becomes deeper when tasks become more complex. Deep neural network represents the type of machine learning when the system uses many layers with the high level functions from the input information. Fig. 2 shows the level of deep learning in comparison to Artificial Intelligence and Machine Learning.

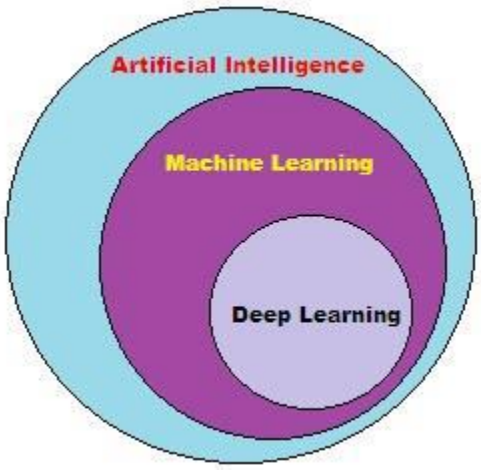


Fig 3. Deep Learning

The basic intention of deep learning is to transform the data into a more creative with an abstract component. It has several applications in data processing and optimized result analysis due to the presence of several multiple layers. A model with a deep 1D convolutional neural network that consists of three 1D convolution layers, two max-pooling layers, two dropout layer, one spatial dropout layer and four fully connected layers is proposed in [5]. Instead of the type of cell from which the signal is recorded the generalized deep learning model used to classify gene expression based on the histone modification signals. The proposed method

automatically performs feature extraction using 1D convolutional layers which are used further for the establishment of relationships among the learned features and to make accurate predictions. According to V. Chaubey et. al. [5] this model is able to make predictions for all cell types when the training of the same completed only once. It also performs better in comparison with the predictions made for different kinds of cells and the computational resources required [5].

The deep neural network is roughly a subset of artificial intelligence and machine learning. Fig. 3 shows a deep neural network with number of hidden layers for the data analysis. In [8] S. kamis et al. proposed a technique for sentiment analysis with twitter data. In this research a comparison and evaluation occurs with the combinations of CNN and a long shortterm memory (LSTM) networks. They have also compared different word embedding systems such as Word2Vec and GloVe models [8]. Several tests and combinations are applied and accordingly the best scoring values for each model are compared in terms of their performance. Several experiments are being carried out for this sentiment analysis. In the first level a single 1-dimensional CNN layer is being used which convolved with 12 kernels with size 1 × 3 and it performs better in comparison with other kernel configurations. Similarly the test is being carried out with single LSTM network, individual CNN and LSTM network, single 3 layers CNN and LSTM network and finally multiple CNN and LSTM network [8].

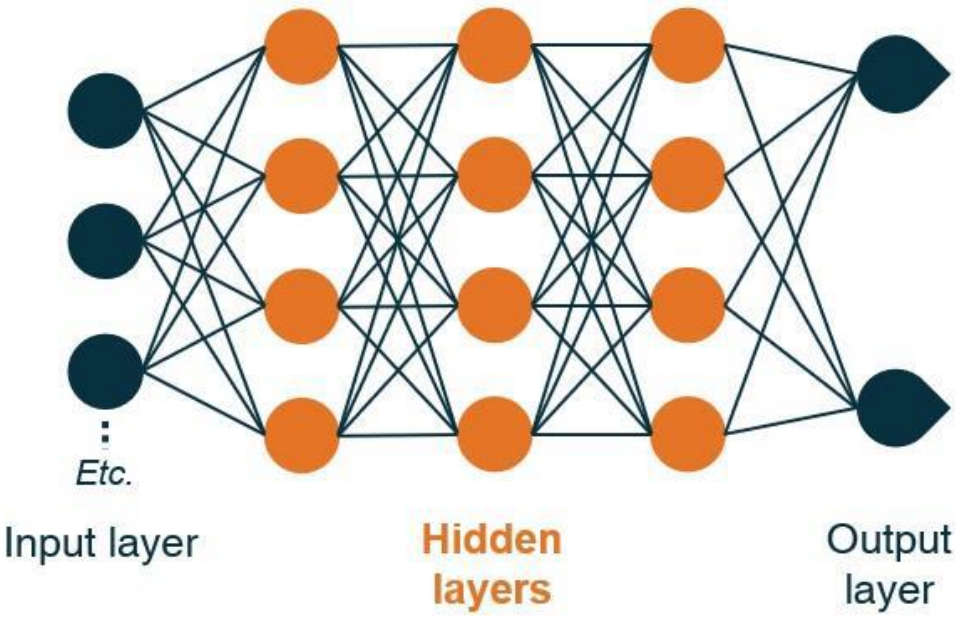


Fig 4. The deep neural network with more number of hidden layers

Deep Convolutional Neural Network has now achieved a great position in the field of computer vision and image recognition. CNN is successful because of the hidden layers which are not fully connected with the previous layers. Artificial Neural networks are being proved as most efficient in Deep Learning mainly because it can handle large number of datasets. The most widely used subset of Artificial Neural Network is the Convolutional Neural Network (CNN). It is basically useful for computer vision, pattern recognition and Natural Language Processing (NLP). CNN has a huge application because comparing over the traditional Neural Networks, it reduces number of parameters and focus more on domain specific features. There are various CNN architectures that are proposed, and proved effective such as such as LeNet, AlexNet, GoogleNet etc [9]. All techniques are having their own advantages for different targeted purposes.

IV. DEEP NEURAL NETWORK IN DETECTION OF COMPLEXITIES IN PSP PROBLEM

PSP problem is one of the common tasks of bioinformatics. Basically in each protein molecule twenty amino acids found with different percentages. In our experiment we have taken a rat dataset available at NCBI (National Centre for Biotechnology Information). The twenty amino acids with their percentages are shown in table 1. In this data set the total proteins available is about 28847. The Leucine (L) is having the most percentage among the amino acid percentage i.e. 10.0015% and selenocysteine(U) is having lowest percentage in amino acid composition i.e. 0.0002%.

Table 1 Amino acid percentages

Number of proteins in database: 28,847			
Amino Acid	**Percentage**	**Amino Acid**	**Percentage**
F	3.8007%	S	8.3817%
T	5.4828%	N	3.6516%
K	5.7199%	E	6.9139%
Y	2.7344%	V	6.2044%
Z	0.0000%	Q	4.6777%

M	2.2258%	C	2.2550%
L	10.0015%	A	6.8256%
W	1.1962%	X	0.0042%
P	6.0531%	H	2.5580%
D	4.8436%	I	4.5223%
R	5.5233%	G	6.4239%
U	0.0002%		

Enzyme cleaved fragments:

Non redundant fragments 2,659,114

Redundant fragments 1,151,680

Out of range fragments 1,408,675

Total time: 118 sec

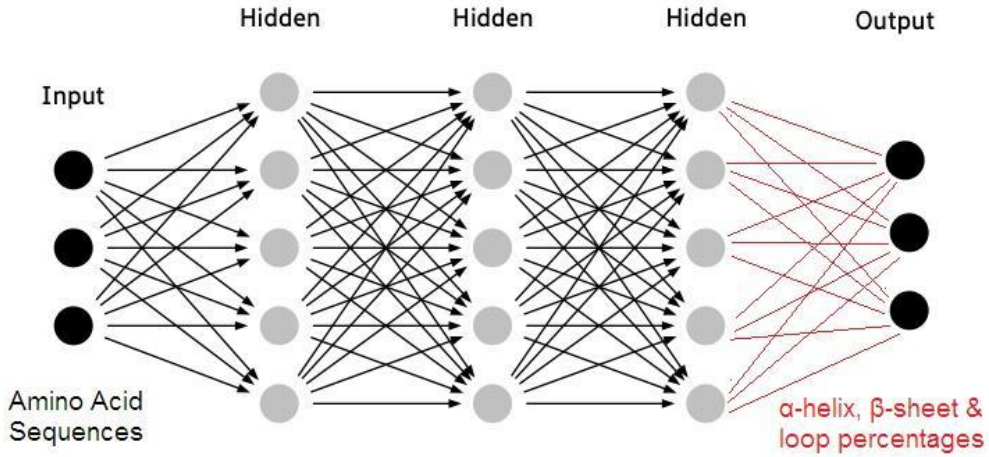
As it a big data set so 28847 proteins available in the same data set. The sequence of the amino acid composition of a single protein is as follows.

>SP|A0A4Z3|A3LT2_RAT ALPHA- 1,3-GALACTOSYLTRANSFERASE 2 OS = RATTUS NORVEGICUS GN=A3GALT2 PE=1 SV=1

MALEGLRAKKRLLWRLFLSAFGLLGLYHYWFKIFRLFEVFIPMGICPMAIMPLLKDN
FTGVLRHWARPEVLTCTSWGAPIIWDETFDPHVAEREARRQNLTIGLTVFAVGRYLE
KYLEHFLVSAEQYFMVGQNVVYYVFTDRPEAVPHVALGQGRLLRVKPVRREKRWQ
DVSMARMLTLHEALGGQLGREADYVFCLDVDQYFSGNFGPEVLADLVAQLHAWH
FRWPRWMLPYERDKRSAAALSLSEGDFYYHAAVFGGSVAALLKLTAHCATGQQLD
REHGIEARWHDESHLNKFFWLSKPTKLLSPEFCWAEEIGWRPEIHHPRLIWAPKEYA
LVRT

These type of genomic data and protein data is now a days very big in size. The secondary structure of a protein is a combination of α-helix, β-sheet and loop structures. The percentage

of these structures varies in their different stages by the application of drugs, medicines or on the presence of any external agents. Similarly twin structures [3] are also available in many cases so removal of twin structure is also essential. As these types of data are very big in size so application of deep neural network plays a vital role in detection and prediction of the protein secondary structures.



the Convolution Neural Network was basically proposed by Yann Lecun [10] now working as VP and Chief AI Scientist of Facebook. It is basically a CNN model that can use back propagation algorithm directly. The convolution kernel detects all shares in the input map and realizes the weight of all input map. For extracting different feature from the available in put separate kernel may be used. For n1 × n2 input x using m1 convolution kernel of m2 × m3 to convolute it. Obtaining $(n_1 – m_2 + 1) × (n_2 – m_3 +1)$ feature with mapping y. The expression may be [1].

$$y_i = b_i + W_i * x \text{ ---------} \quad (1)$$

where $i \in \{1, m_1\}$ and * is the two-dimensional discrete convolution operator [1].

Deep learning is being extensively used for various tasks like image classification, segmentation, audio video generation etc. Most of the application of CNN employ 2d convolution which effectively map the complex feature of an image but performs les Effectively in case of one dimensional data. Any application of two dimensional convolution requires a conversion of the one dimensional signal to two dimensional which is less effective. To over come this problem several technologies New technology has been proposed [11] and our coming research focuses on these.

V. CONCLUSION AND FUTURE WORK

Though several techniques have been developed for the purpose of protein structure prediction and analysis but still PSP problem is still a challenging task due to the complexities of protein. Due to the revolution in gene science and mutation of protein molecules every time new protein molecule comes forward with a new and complex task. As these data set are very large in size so precise and updated technologies will be helpful for processing of these huge amount data. This proposed technique is a genuine, novel and upgraded technology to challenge and accept new task that comes in the field of bioinformatics. This proposed technique in detection of protein structure and complexities will be helpful for different researchers and medicine designers for the improvement in drug design and disease detection and transformations. Several large data set of protein can be process and the complexities can be detected during prediction which will helpful for avoiding tiny to tiny mistakes during any medical and medicine research. Our upcoming research focuses upon the two dimensional convolutional neural network with different type of medical image analysis.

REFERENCES

[1] Y. Huang, D. Xiusheng, S. Shiyu, C. Zhigang, "A Study on Deep Neural Networks Framework" Proc. of Int. conf. on Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), IEEE, 2016.

[2] S. B. Rout, S. Mishra, D. K. Swain, "Protein Structure Prediction using Brute Force Search and Genetic Algorithm", International Journal of Research and Analytical Reviews, Vol. 6(2), 2019.

[3] M. T. Hoque, M. Chetty, A. Lewis, A. Sattar, "Twin Removal in Genetic Algorithms for Protein Structure Prediction using Low-Resolution Model," Transactions on Computational Biology and Bioinformatics, Vol. 8(1), pp. 234–245, 2011.

[4] P.A. Bates, L.A. Kelley, R.M. MacCallum, M.J. Sternberg, "Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM", Journal of Proteins, Vol.45(5), pp.39–46, 2001.

[5] V. Chaubey, M. Nair, G.N. Pillai, "Gene Expression Prediction Using a Deep 1D Convolution Neural Network" Proc. of Symposium Series on Computational Intelligence, IEEE, 2019.

[6] M. A. Rashid, M. Hoque, M. Newton, D. Pham, A. Sattar, "A new genetic algorithm for simplified protein structure prediction," Advances in Artificial Intelligence, Lecture Note, 2012.

[7] K. F. Lau, K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins," Macromolecules, Vol. 22(10), pp. 3986–3997, 1989.

[8] D. Goularas, S. Kamis, "Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data" Proc. of Int. Conf. on Deep Learning and Machine Learning in Emerging Applications, 2019.

[9] D. Arora, M. Garg, M. Gupta, "Diving deep in Deep Convolutional Neural Network" Proc. of 2nd Int. Conf. on Advances in Computing, Communication Control and Networking, 2020.

[10] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, "Gradient based learning applied to document recognition" Procceding of the IEEE, pp. 2278-2324, 1998.

[11] S. Kiranyaz, T. Ince, R. Hamila, M. Gabbouj, "Convolutional Neural Netwoks for patient-specific ECG classification", Proc. of Annual Int. Conf. on Medical Biology, 2015.

[12] G. Dahl, T. Sainath, G. Hinton. "Improving deep neural networks for lvcsr using rectified linear units and dropout," Proc. of Int. Conf., British columbia: IEEE, pp. 8609-8613, 2013..

[13] H. Zen, A. Senior, M. Schuster, "Statistical parametric speech synthesis using deep neural networks," Proc. of Int. Conf., British columbia: IEEE, pp. 7962-7966, 2013.

[14] M. Zayan. "Satellite orbits guidance using state space neural network," Proc of Aerospace Conference, IEEE, 2006.

[15] Z. Ling, L. Deng, D. Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," Audio, Speech, and Language Processing, vol. 21(10), pp. 2129-2139, 2013.

[16] A. Bautu, H. Luchian, "Protein structure prediction in lattice models with particle swarm optimization", In Int. Conf. on Swarm Intelligence, IEEE, pp. 512-519, 2010.

[17] D. A. Pelta, N. Krasnogor, "Multimeme algorithms using fuzzy logic based memes for protein structure prediction. In Recent Advances in Memetic Algorithms, pp. 49-64, 2005.

[18] X. Zhao "Advances on protein folding simulations based on the lattice HP models with natural computing" Journal of Applied Soft Computing, Vol. 8 (2), pp. 1029-1040, 2008.