International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Function Prediction from Protein Primary Structure using Deep Learning LSTM Algorithm

Anjna Jayant Deen, Manasi Gyanchandani

Abstract: ***Biological information of protein primary structure is responsible for finding the protein function, extracting features and function of a protein in the biology lab is challenging and time-consuming. Identification of protein function provides essential information for the treatment of various diseases and drug design. Therefore, extracting the protein knowledge from primary structure alone has been a diverse field in the study of bioinformatics data mining and computational biology. This study aimed to function prediction of protein primary structure using the LSTM methods. PRNP(prion protein)most of the nervous system tissues express by prion protein, this is generally to protease-resistant from disease, due to this reasons, the human codon PRNP is most closely associated with Alzheimer disease. The PRNP protein data trained with Hemo sapiens PRNP selection, classification was implemented with network layer perceptron. The learning algorithms are frame by the nervous system. The training results observation indicate that the learning success of prion protein classification leads positively.***

I. INTRODUCTION

All living organism is changing fast [1]. As a result of this, the identification of protein to protein cell communication is the main issues in computational biology. Protein identifies by their structural properties, functional behaviour, crystallography and NMR(nuclear magnetic resonance). Protein sequence are in primary structure(amino acid sequence), secondary structure consider eight hydrogen bond patterns (H,G,I,E,B,T,S and C),further these can be reduced to three (H,G) to (H) α-helices, (E,B) to (E) β-sheets and (I,T,S,C) to (C) coil. Sometimes protein-protein interaction holds dysfunctional information, and this causes to leads the misfolding in protein binding site [3], so result that few common neuron disorders like Alzheimer's and Parkinson's. Another factor like age and mutations also leads to the misfolding of proteins. Correctly Identify the protein fold and misfold collected from an amino acid sequence. Misfold protein will provide information about genetic disorders and diseases, which helps to new treatments.

Correct fold tells about the when and how normal conditions returned. Folding creates a protein secondary structure. The amino acid sequence fold decided its 1D,2D and 3D structure, whereas this 3D structure found in the protein binding site and their function [2]. Identification of the 3D structures of proteins is a costly and time consuming, and also very challenging process in the lab experiments due to the size of their small structures. Finds correct function to protein, there are three methods commonly used in practice:- 1.X-ray diffraction, it applies to diffract many specific directions of protein crystals, the diffraction patterns, measured at an angle of φ strong streaks of varying intensity originating at the Bragg peaks with an angle of φ to the various direction.,2.NMR (nuclear magnetic resonance) and 3. Electron crystallography. Practice these methods are the most popular and still very hard to implement. The researcher tries to find exact and new testing results in solving by uses of computational methods. Protein folding parameter and boundary frequency determine a first time in their hierarchical order were discuss in paper[4]. In article [10] says every protein have a two-state fibrous and globular, which was first to describe the significance of the amyloid proteins. Amyloid prion protein identifies many diseases and Alzheimer. However, prion protein function is found in the amyloid state[12]. The classification method is fast and accurate for applying to find different protein cell information. Due to large protein datasets, various kinds of protein structure matching problem have generated, The artificial neural system used for sequence encoding to store a large number of sequence patterns[24].

1.1. Protein fold function selection

In this study protein sequence based features extracted :

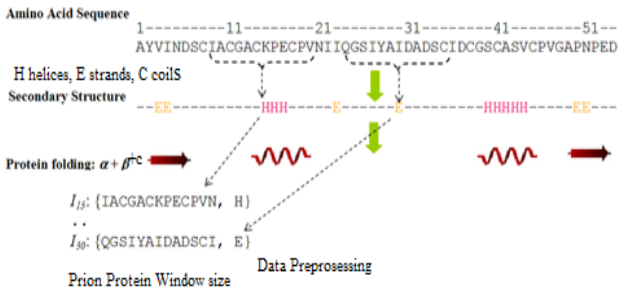


Figure 1. Feature selection diagram of Primary structure to secondary structure

1.2. Amino acid composition based features: 20-dimensional vector space in each protein sequence length are in a different size. This feature has used with the subcellular localization of proteins [20], classification of nuclear receptors [19] and protein fold recognition [18].

Manuscript published on November 30, 2019.
* Correspondence Author
Anjna Jayant Deen* is currently pursuing a PhD degree program in Computer Science & Engg Department, Moulana Azad National Institute of Technology, Bhopal, India,
Manasi Gyanchandani is currently working as an Assist. Professor in Computer Science & Engg Department, Moulana Azad National Institute of Technology, Bhopal, India,

The AAC(amino acid composition) calculated i as in eq.(1) [16].

$$f_i = \frac{ni}{L} \text{.....eq.(1)}$$

where fi = frequency in amino acid composition (i);
ni = Number of protein sequence in amino acid composition (i);
L = Total number of amino acid residues found in that protein
and i =1 to 20 (ARNDWVCEYQGHILKTSMFP).

1.3. Amino Acid Composition :

AAC an nth sequence order in sequence pair depicted for essential feature, the interaction between the ith and $(i+n)^{th}$, (n >0) were n is number of amino acid residues, and gives the sequence order information and composition of amino acids[16]. Amino acid pair composition is a (20×20) dimensional vector of protein residues, which has been used to determine many problems, such as subcellular localization of proteins [20], classification of G-protein-coupled receptors [19], etc. The n^{th} order of amino acid composition residue in a protein is calculated according to eq.(2)[16].

$$f(D^{i,(j+n)})j = \frac{n(Di,(j+n))j}{L-n} \text{ .....eq.(2)}$$

where f(Di,(j+n))j = is the frequency of an nth order amino acid pair j; n(Di,(j+n))j = is the number of nth order amino acid pair j; n is the order of amino acid pair and j = 1 to (20×20).

1.4. Structure-based features: Secondary structural state (H helices, E strands, C coils) frequencies of amino acids: these are the frequencies of amino acids collectively represented as a (20×3) dimensional vector. The frequency is calculated using the parameter given in eq.(3)[16].

$$f^k_i = \frac{nk\,i}{L} \text{.....eq.(3)}$$

Here, k = (H helices, E strands, C coils),
fki = is the frequency of amino acid i occurring in the secondary structural state k,
nki = is the number of amino acid i occurring in the secondary structural state k.

The calculation has been based on secondary structure information.

The predictions were made using protein structure prediction server allows users to submit a protein sequence [17]. And only those with confidence level >=1 were considered for calculations[16].

II LSTM BASED PROTEIN FUNCTION PREDICTION

Long Short Term Memory networks[14] LSTM (input gate, forget gate and output gate) shown in fig.2 are a specific type of neural network, capable of learning long-term dependencies, LSTM work tremendously well on a wide range of problems, and are now widely used in prediction network. A novel, adaptive "forget gate" that enables an LSTM cell to learn to reset itself at fitting instances, thus allow to flow freely in internal resources[22].

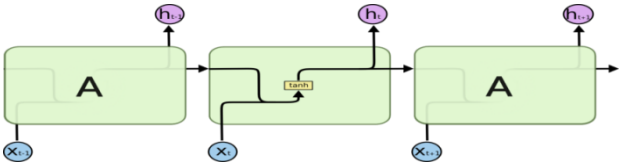


Fig 2. Block Diagram of LSTM network

The LSTM recurrent neural network cell is connected recurrently to each other, replacing the usually hidden units of ordinary recurrent networks. An input feature has computed with a regular artificial neuron unit. Its value can be assemble into the state if the sigmoid input gate allows it. The state unit has a linear self-loop whose weight is controlled by the forget gate. The output of the neuron cell can be shut off by the output gate. Gate value depending upon the linear and non-linear activation, sigmoid, tanh, softmax, and ReLU function. It is used to determine the output of hidden layers, and that output will be used as the input for the following node, resulting in values maps in between 0 to 1 or -1 to 1 etc. The state unit can also use as an extra input to the gating units. LSTM based deep learning-based prediction model, as shown in fig 3.

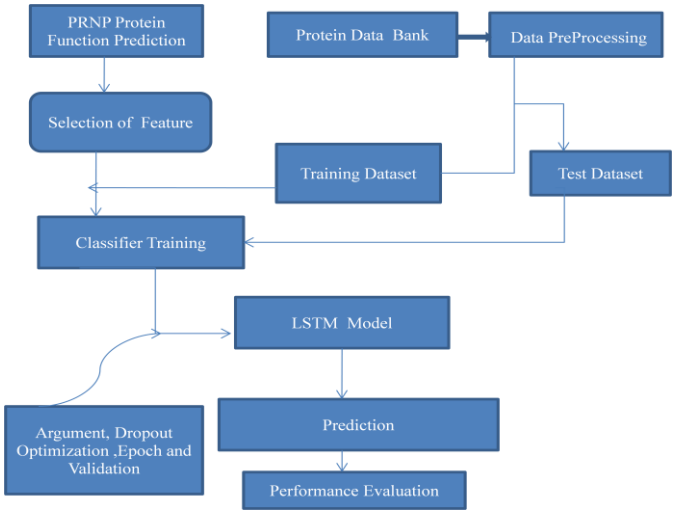
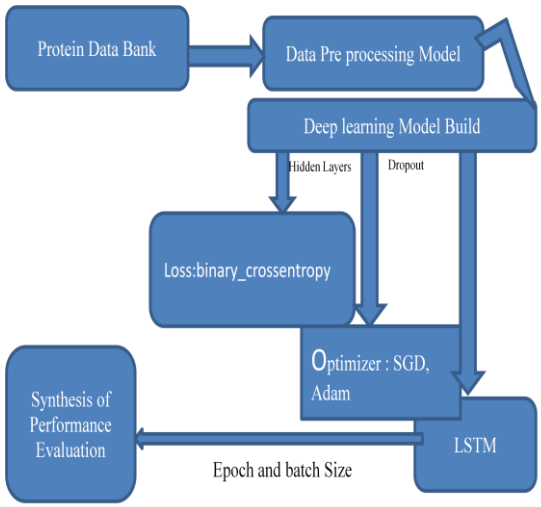


Fig 3.PRNP Protein Function Prediction Methodology

2.1. Dropout, Stochastic gradient descent and Batch size :

The fundamental of dropout is that training a network with the stochastic model and process predictions by averaging over multiple stochastic outcomes, implements a form of bagging with parameter distribution, and any random change is admissible. Deploy of model families that allow a fast estimate inference rule. Constructing new inputs by noise can be regularization by using dropout. Dropout trains each hidden layers must be able to perform well unconcerned of which other hidden layers are in the model. Hidden layer contains vector value were each layer in a model dimensionality determines the width of the network. Hidden units must be set to be swapped and interchanged between models[11].

Stochastic gradient descent, presented in usually batch sizes driven by the successive consideration:- (a.) Larger batches bring a higher accurate estimate of the gradient, but with less than linear returns; (b.) If all pattern in the batch is to be processed in parallel, then needed a spare amount of memory cell. The overall runtime can be very high as the need to take extra steps, being these reasons it weaker learning to consider, and it takes further steps to view the entire training set; (c.) Small batches can offer a regularizing effect [25]. Training with such a small batch size might desire a low learning rate to keep stability due to the high variance in the estimate of the gradient. LSTM workflow model shown in fig. (4). Distinct kinds of algorithms use various types of protein primary sequence from the batch in different ways. Thus, the experience that stochastic gradient descent reduces generalization error in the learning model.



Deep learning LSTM Model

Fig 4. Work flow diagram of Deep learning Model

III EXPERIMENTAL SETUP

In this study, data collected from NCBI data collected 73850 PRNP(PRioN) primary protein sequences, namely Homo sapiens (640 protein sequence). These benchmarks sequence data is downloads from www.ncbi.nlm.nih.gov(protein data bank). In this study, we implement in Python(HP Z80 workstation) Keras model for LSTM: Recurrent network consisting sequential way in a different hidden layer in a stack, layer passing in layer instances to the constructor specifying shape in the first layer. An argument passing in the first layer will shape it and followed by others hidden layers, In LSTM training model need to configure learning steps, these steps are like an optimizer, a loss function and a list of metrics. Models were trained on Numpy array of input data and labels.

3.1. Evaluation of Performance: Accuracy(Q3)

ALHEASGPSVILFGSDVTVPPASNAEQAK⟶Amino acid sequence

hhhhhooooeeeeoooeeeoooooohhhhh ⟶Actual Secondary Structure

ohhhhooooeeeeoooooeeeooohhhhhh ⟶ Q3=23/29=79% If n=29,

Random prediction of Q3 is approx 35%,secondary structure assignment in real proteins is uncertain to about 1-10%; Therefore, a protein function prediction would have Q3=89.9 % to 98.9%.The results were observed according to the Q3 success standard, the percentage total number of residues accurately count as in eq.(4)

$$Q3 = \frac{QH+ QE+ QC}{n} * 100 \quad \text{.....eq.(4)}$$

for Q_H = helices , Q_E = strands and Q_C = coils states[11]. The results were observed by changing the number of hidden layers from 1 to 6 and the number of neurons in each hidden layer from 1 to 200 in this study, and the results obtained for 10, 50, 80, and 200 epochs LSTM prediction accuracy were given in the Tables 1.

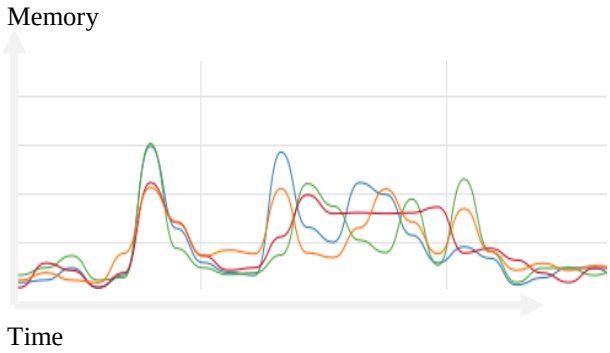


Fig. 5 CPU and memory utilization graph during Autoencoder training

IV RESULT AND DISCUSSION

An experimental method is in these work aiming to predict the function of PRNP protein. The homo sapiens prion protein data were improved with a deep learning model, whereas the trained data has classified for the next stage. The deep learning experimental results obtained and indicate that there is a significant increase in the prediction success. The LSTM network with SGD gradient and ADAM based classification observed success rate in PRNP Homo sapiens 640; Protein sample has been iterations in hidden layers. It found that the LSTM with Adam and Gradient-Based Optimization Classification success is highest with 98.96% in two hidden layers with 200 epochs. For each iteration, data trained, and test data classified separately. As seen in Table1, the data iterated many times, the percentage of classification success occurred between 96.09% to 97.06% for 10 epochs, 97.09% to 98.20%f for 50 epochs, 97.09% to 98.06% for 80 epochs and 97.09% to 98.96% for 200 epochs, their graph shown in fig 6(a, b, c, d).

Table 1.LSTM Success Change Based On The Number Of Hidden Layers

Epoch	Hidden layer 1	Hidden layer 2	Hidden layer 3	Hidden layer 4	Hidden layer 5	Hidden layer 6
10	96.093	97.093	97.093	97.09	97.093	97.093
50	97.093	96.22	96.89	96.627	96.62	98.20
80	97.093	98.06	96.40	97.79	97.62	97.20
200	97.093	98.95	98.34	96.62	97.46	98.96

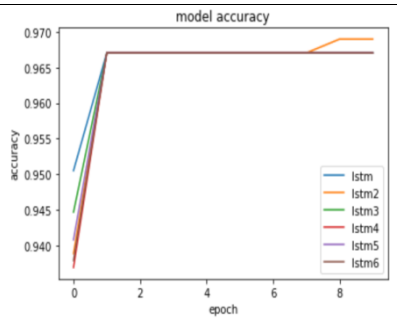


Fig.6a 10 epochs accuracy graph in lstm layers

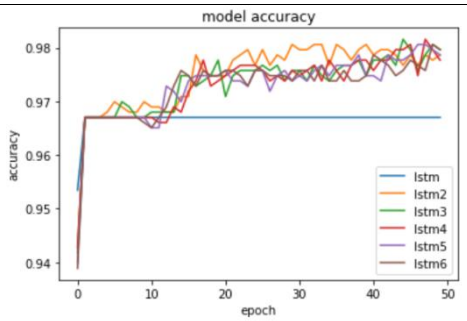


Fig.6b 50 epochs accuracy graph in lstm layers

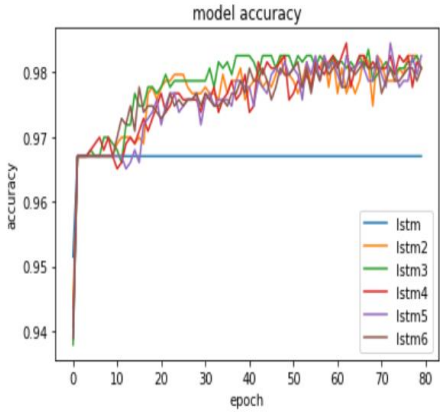


Fig.6c 80 epochs accuracy graph in lstm layers

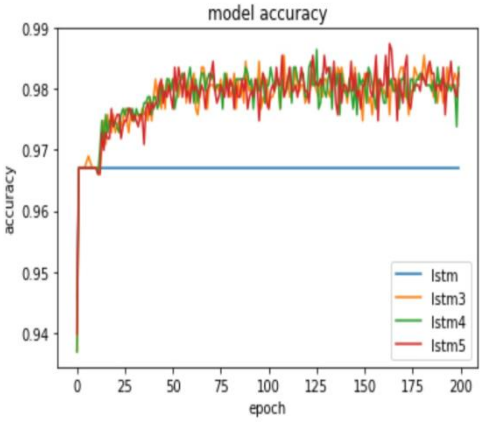


Fig.6d 200 epochs accuracy graph in lstm layers

V. CONCLUSION

Directly predict the protein function from their primary sequences has been a challenging task to find accurate prediction results. Many computational tools and learning algorithms have applied towards a classification of protein function from primary sequence without alignments. Deep learning LSTM based models containing autoencoder trained on PRNP protein datasets. In conclusion, using Adam and Gradient-Based optimization prior to classification is recommended for a higher prediction success. LSTM less number of hidden layers are required achieved high performance for prediction. Python programming supports deep learning techniques are so powerful today. It can build simple neural networks and generate predictions with them.

REFERENCES

1. L. A. Allison, "Fundamental Molecular Biology", 2nd ed. USA, Wiley, 2012.
2. H. F. Lodish, "Molecular Cell Biology", 7th ed. New York, NY, USA, 2013.
3. E. Reynaud, et.al, ‘‘Protein misfolding and degenerative diseases’’, Nature Educ., vol. 3, no. 9, pp. 28-32, 2010.
4. P.Y.Chouand ,G.D.Fasman, ‘‘Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins’’ , Biochemistry, vol. 13, no. 2, pp. 211–222, Jan1974.
5. J. Garnier, D. J. Osguthorpe, et.al, ‘‘Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins’’ J.Mol.Biol.,vol.120, no.1, pp.97–120, Mar1978.
6. N. Qian ,T. J. Sejnowski, ‘‘Predicting the secondary structure of globular proteins using neural network models’’, J. Mol. Biol., vol. 202, no. 4, pp. 865–884, Aug1988.
7. W. Pirovano , J. Heringa, ‘‘Protein secondary structure prediction in Data Mining Techniques for the Life Sciences'" Methods in Molecular Biology, vol. 609. Clifton, NJ, USA: Humana Press, pp. 327–348,2010.

8. L. N. De Castro , F.J. Von Zuben, ‘‘Artificial immune systems Part I - Basic theory and applications,’’ FEEC,UNICAMP, Campinas, Brazil, Tech. Rep. RT DCA vol.01,pp.99, 1999.
9. M. B. Rodrigues et. al, ‘‘Health of things algorithms for malignancy level classification of lung nodules,’’ IEEE Access, vol. 6, pp. 18592–18601, 2018.
10. EvangeliaI, Zacharaki , "Prediction of protein function using a deep convolution neural network ensemble" ,PeerJComput.Sci., pp.1-9, 2017.
11. P. Ghanty, N. R. Pal, et.al, ‘‘Prediction of protein secondary structure using probability based features and a hybrid system,’’ J. Bioinf. Comput. Biol., vol. 11, no. 5, pp. 135, Oct. 2013
12. Burce,Nilufer and Ozhan Ozkan "Prediction of Protein Secondary Structure With Clonal Selection Algorithm and Multilayer Perceptron" Digital Object Identifier, Vol. 6, pp.45256-45261,2018.
13. Sepp Hochreiter, Jurgen Schmidhube, "LONG SHORT-TERM MEMORY ", Neural Computation 9(8),1735-780, pp.1-32,1997.
14. Leandro Nunes de Castro Fernando José et.al, "Artificial Immune Systems",6th International Conference, ICARIS, Santos, Brazil, Proceedings pp.26-29, 2007
15. Mohammad Tabrez ,Anwar Shamim, et.al,"Structural bioinformatics Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs" bioinformatics Vol. 23 no. 24, pages 3320–3327, 2007.
16. McGuffin,L.J. et. al, "The PSIPRED protein structure prediction server". Bioinformatics, 16, pp.404–405,2000.
17. Ding,C.H. and Dubchak,I," Multi-class protein fold recognition using support vector machines and neural networks", Bioinformatics, 17,pp. 349–358,2001.
18. Karchin,R. et al. "Classifying G-protein coupled receptors with support vector machines". Bioinformatics, 18, pp.147–159,2002.
19. Guo,J. et al,"GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins". Proteomics,vol. 6,pp. 5099–5105,2006.
20. Ian Goodfellow, Yoshua Bengio et.al,"Deep Learning" Book website www.deeplearningbook.org
21. FelixA.Gers, Jurgen Schmidhuber et.al, "Learning to Forget: Continual Prediction with LSTM" Neural Computation, Volume 12 , Issue 10 , October , pp.2451-2471,2000.
22. Yehong Chen “Long sequence feature extraction based on deep learning neural network for protein secondary structure prediction” IEEE 978-1-5090-5363-6/17/2017.
23. Cathy ,George Whitson, et.al " Protein classification artificial neural system" ,Protein Science, Cambridge University Press ,vol.I, pp .667-677,1992.
24. Wilson and Martinez "Reduction Techniques for Instance-Based Learning Algorithms" ,Journal Machine learning, vol.38 Issue 3, pp. 257-286,March 2000.

AUTHORS PROFILE

Anjna Jayant Deen working as Assistant Professor in UIT, RGPV Bhopal, she is having more than 22 years experience in teaching and research. She is currently a PhD student in MANIT Bhopal. Her research interest is in Data science, Network security, Bioinformatics, Artificial Intelligence, Machine learning and Neural network. Her publications in more than 16 research papers in Scopus-journal, International Conference and International journal. She is a member of IEEE..

Manasi Gyanchandani is working as Assistant Professor in MANIT Bhopal. She is having more than 24 years of experience in teaching and research. Her educational qualification is a PhD in Computer Science and Engineering. Dr Manasi Gyanchandani's area of specialization in big data, big-data privacy and security, Machine learning, data mining, privacy-preserving, artificial intelligence, expert system, neural network intrusion detection & information retrieval. Dr Manasi Gyanchandani has more than 32 research papers in SCI/ Scopus/ Peer-Reviewed journals, International Conference, book chapter, and National Conference. She is a Life Member of ISTE.