International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Stacked Bidirectional Long Short Term Memory Models To Predict Protein Secondary Structure

R.Thendral, AN.Sigappi

Abstarct: ***Protein Secondary Structure (PSS) is one of most complex problem in biology PSS is important for determining tertiary structure in the future, for studying protein fiction and drug design. However, Experimental PSS approaches are time consuming and difficult to implement, and its most essential to establish effective computing methods for predicting on protein sequence structure. Accuracy of prediction performance has been recently improved due to the rapid expansion of protein sequences and the design of libraries in deep learning techniques. In this research proposed a deep recurring network unit method called stacked bidirectional long-term memory (Stacked BLSTM) units to predict 3-class protein secondary structure from protein sequence information using a bidirectional LSTM layer. To evaluate the output of Stacked BLSTM, using publicly available datasets from the RCSB server. This study indicates that performance of our method is better than the of that latest stranded public dataset. The accuracy achieved is more than 89%.***

Keywords: amino acid sequence, Protein secondary structure, deep bidirectional LSTM.

I. INTRODUCTION

Proteins are the sequence of amino acids that are essential components of the structure and purpose with a wide range of tasks, with structural support, virus protection and muscle contraction. The structure of protein can be classified as primary refers to the linear sequence of amino acids and the secondary is regional sections of protein chain joined into Q3 or Q8 class secondary structures. In this research, concentrate on prediction the secondary protein structure of the primary sequences uses simple Q3 labelling method. This method again split into three subclasses called helix, string and loop or coil. Recently, the growth of neural network technique, comparable to machine learning prediction experiments. The following are the pervious deep learning various technique and the test accuracy are take from different literature reviews. They are deep Convolution Generative Stochastic Network, gets output of 66.4% accuracy[22],Long short term memory units got result of 67.4%[23], Convolution neural fields gaining output of 68.3%[25] and bidirectional LSTM with and without conditional random field(CRF)[26], achieving result of 69.4% and 73.3 %. accuracy respectively.

As a conceptual deep learning method for the processing of sequence data, LSTMs have proven to be capable of manipulating sequence data[5] and applying it to many real-world problems. like recognition of speech[6],captioning of image [7], composition of music [8] and human prediction of trajectory [9].Currently, LSTMs have been more effective in sequencing because of its long-term dependencies capability. Various research [19] have been performed to analyze the suitability of LSTMs in bioinformatics, and the outcome of LSTM is shown. From the scale view of predicting the protein structure, long sequence has become an important and challenging topic. Based on the complexity of the architecture the LSTM-based system, the design should able to fetch the structure prediction system's dynamic nature.

Recent studies proposed, LSTM prediction method have close to shallow architectures with one hidden layer to handle the time series data [15,12,16]. Early research [10,11] clearly shown deep LSTM units on multiple hidden layers will gradually increase the higher rates of data subsequent representation. However some research [13-14] used additionally more than one hidden LSTM layer and it necessary to compare and explain the impact of the number of LSTM on different LSTM-based models. All information presented in sequence information should be fully utilized in terms of dependency on prediction problems. Normally, the data set feeding to the LSTM model is arranged chronologically, resulting in the transmission of the information in the LSTMs from time step t to t-1 along the structure in a positive direction. The LSTM structure uses the forward dependencies to analyzing the periodicity of biological data from both forward and backward temporal viewpoints; in particular of repeated sequence patterns can enhance predictive efficiency [3,4]. However, study on our review of the literature, the backward dependence was used by few studies on sequence analysis. This research adopts a bi-directional LSTM (BLSTM) as a part of the network structure have potential to handle the both forward and backward dependencies. This present study, stacked LSTM with bidirectional (Stacked BLSTM) network for predicting the secondary structure. The proposed model can accommodate missing values for input data and is evaluated on both large-scale sequences. Results of the experiment show that with a high predictive accuracy, our model achieves secondary structure prediction.

II. METHODOLOGY

The proposed Stacked bi LSTM architecture are explained in detail and models illustrations in eventual sub-sections all take as examples the protein structure of prediction.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
R.Thendral*, Research Scholar, Computer Science and Engineering, Annamalai University, India.
AN.Sigappi, Ph.D, Computer Science and Engineering from Annamalai University, India.

2.1 Simple RNN

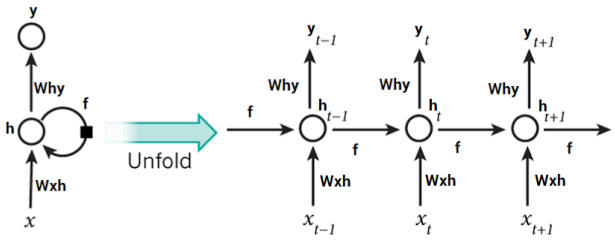

Fig. 1. Standard RNN architecture

The RNN is a powerful deep learning tools that utilize its internal memory to handle sequence data with loops. The RNN architecture, which is also the basic framework of LSTM and it shown in fig.1. The hidden layer input is vector X and it produce the output vector Y. The right side of the figure shows calculation, at each time iteration t, h_t was update by on x_t and h_{t-1} was updated by equation 1.

$$h_t = \sigma_h(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

W_{xh} is weight matrix , W_{hh} is weight matrix between h_{t-1}, h_t, the b_h is bias vector and b_h is activation function to generate hidden state, output

$$y_t = \sigma_y(W_{hy}h_t + b_y) \quad (2)$$

where W_{hy} is matrix of output layer, bias vector denotes b_y and activation variable is σ_y . The basic problem in RNN is that it is dealing with Vanishing Gradient problems. Vanishing Gradient problem occurs most of the time when we are concerned with large sequence data sets.

2.2 Long short term memory

To fix the above drawback of RNNs, various advanced recurrent neural units, including LSTM methods [2] are proposed on long-term dependencies sequence data.

However the number of common LSTM models have been introduced in previous studies although none of the variant will significantly gives the better LSTM architecture[17].Therefore, as part of the proposed network structure, the generic LSTM architecture with more layer is implemented in this analysis and hidden layer is only difference between this architecture [18],This referred in LSTM cell and it is shown in Fig.2. Here forgotten gate, helps LSTM cell and scalable technique is used for number of sequential data-related learning problems [17].

Similar to RNNs, for every iteration time t, the LSTM cell has input x_t and output h_t and for complex cell input state is $\check{C}_t$, output state C_t. The earlier LSTM cell output is $C_{t-1.,}$. Because of the gated design, LSTM able to manage long term memory dependencies to permit applicable sequence data shift to LSTM units. Here for time t, input gate t_t, forget gate f_t, and output gate o_t. it was shown in fig2.and gate calculation was done by below equations.

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (3)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (4)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (5)$$

$$\check{C}_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (6)$$

Here W_f, W_i, W_o, and W_c refers the weight matrices used for mapping also U_f, U_i, U_o, and U_c refer weight matrices that joined the pervious output to the next input state , Where b_f, b_i, b_o, and b_c are bias and σ_g is the activation function and tanh is tangent function. The pervious equation is used to calculate C_t and h_t with at each time iteration based on their results.

$$c_t = f_t * c_{t-1} + i_t * \check{C}_t \quad (7)$$

$$h_t = o_t * \tanh(c_t) \quad (8)$$

The final output is a vector and all the outputs represented by a sequence.

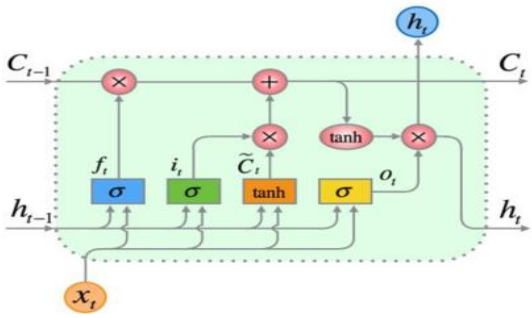

Fig. 2. LSTM Networks

2.3 Stacked Bidirectional LSTM

The fundamental aim of bidirectional recurrent neural network(BRNNs) proposed by Schuster and Paliwal, 1997 and Baldi et al., 1999. In this system to address the training sequence data both forward and backward uses more than one recurrent network.

It implies that the bidirectional RNN has increasing, subsequent data about all points before and after every point in a cretin series. Moreover, the target delay length bidirectional recurrent network is less consider by doing so the result is improved in sequence learning tasks, particularly in prediction of protein structure (Baldi et al., 2001 Chen and Chaudhari,2004). Prior studies [23] gives several hidden layer LSTM units with efficient sequence data effectively. Fig 3. shows that stacked bidirectional Long short term memory architecture.

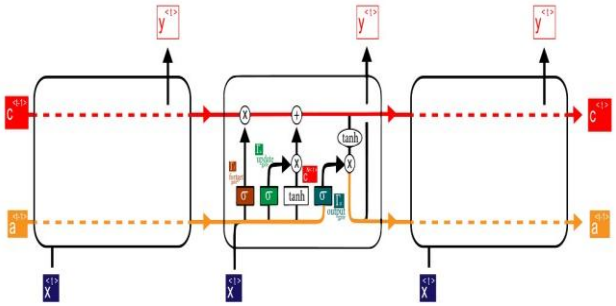

Fig. 3. Stacked bidirectional LSTM Networks

In this study deep Bidirectional LSTM architecture design with three layer, which is LSTM 1 and LSTM2,prediction layer that is called output layer with activation.

III. EXPERIMENTS

3.1 Dataset Description

There are several reference databases accessible on protein secondary structure prediction. The RCSB data set, consisting of a 93953 sequence of varying lengths, is the most complex data set dependent on the highest accuracy achieved using deep learning approaches. The database comprises of sets of amino acids labeled with a secondary structure. Sets and frameworks have been extracted from RCSB and annotated by DSSP [Kabsch & Sander, 1983]. The complete data set consists of 58673 non-homologic sequences and the data set is spilt into 46863 training, 5887 testing, and 5923 validation is used in this work with a limited length of 128 or less.

3.2 Implementation of Stacked Bidirectional LSTM

Tensorflow is the development of a conceptual system and the backend for Keras. First of all, we are designing our new Keras API Bidirectional LSTM. The stacked BLSTM is introduced by the Keras LSTM framework. We practice the model and change the parameters of the bidirectional LSTM using the Adaptive Moment Estimate (Adam) activation function. Table 1 indicates deep learning approaches usually have variable parameters.

3.3 Analysis of Experiment Results

The result is shown in Table 1 with difference turning hyper parameter from experiment select best model with high accuracy, result achieved learning rate 0.002 with 40 reach high accuracy showing positive for the parameter turning so further tuning the parameters and stacked BLSTM provide better result shown Table 3.However, we compare our method with DeepCNF(Wang et al.2018) proposed an next version of CNF (DeepCNF) by deep neural networks, which was an join the method between CNF and shallow convolution neural networks and DeepACLSTM[27].Table2 shows a comparison of the results with the accuracy of Q3 for proposed and baseline methods with deep learning techniques.

Table1.The Comparison of results with various parameters with stacked bidirectional LSTM

Model	Samples used for Training	Samples used Validation	Samples used Testing	Learning rate	epoch	Loss	Accuracy
BLSTM	46863	5887	5923	0.001	50	0.231	0.85
BLSTM	**46863**	**5887**	**5923**	**0.002**	**40**	**0.232**	**0.861**
BLSTM	46863	5887	5923	0.003	40	0.232	0.859

Table2. The highlights show the best accuracy of proposed method

Method	Accuracy %
DeepCNF	68.3
DeepACLSTM	74.2
SBLSTM	**89**

]Table3. The accuracy (%) of our method from experiments

Model	Samples used for Training	Samples used Validation	Samples used Testing	Learning rate	epochs	Loss	Accuracy
Stacked bidirectional LSTM	**46863**	**5887**	**5923**	**0.002**	**50**	**0.213**	**0.893**

The stacked bidirectional LSTM with the Leaning rate (learning rate=0.002) shows the best Q3 accuracy and fig. 4 shows learning curve with number of epoch and accuracy.

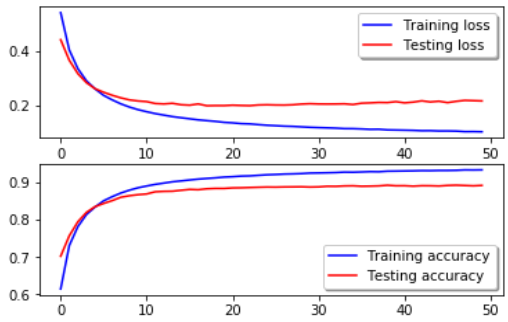

Fig. 4. The Learning Curve

IV CONCLUSIONS

A stacked bi-directional LSTM(Stacked BLSTM) units for prediction the protein secondary structure is proposed in this paper. Furtherance and benefaction in this study focus on two perspective 1) we propose stacked architecture, taking into account both forward and backward dependence on protein structure prediction 2) Limited sequence length 128 or less. Experiments have shown that Stacked BLSTM is best model to predict 3-class of protein secondary structure and it has the potential to gather high complicated sequence associations between amino acid residues. In future this work can be extended to prediction of super secondary structure.

REFERENCES

1. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
2. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2008). Time series analysis: forecasting and control John Wiley & Sons. Inc., Hoboken.
3. Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019.
4. Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015, June). An empirical exploration of recurrent network architectures. In International Conference on Machine Learning (pp. 2342-2350).
5. Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). IEEE.
6. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
7. Eck, D., & Schmidhuber, J. (2002). A first look at music composition using lstm recurrent neural networks. Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, 103, 48.
8. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 961-971).
9. Graves, A., Jaitly, N., & Mohamed, A. R. (2013, December). Hybrid speech recognition with deep bidirectional LSTM. In 2013 IEEE workshop on automatic speech recognition and understanding (pp. 273-278). IEEE.
10. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature 521.
11. Duan, Y., Lv, Y., & Wang, F. Y. (2016, November). Travel time prediction with LSTM neural network. In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC) (pp. 1053-1058). IEEE.

12. Chen, Y. Y., Lv, Y., Li, Z., & Wang, F. Y. (2016, November). Long short-term memory model for traffic congestion prediction with online open data. In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC) (pp. 132-137). IEEE.
13. Yu, R., Li, Y., Shahabi, C., Demiryurek, U., & Liu, Y. (2017, June). Deep learning: A generic approach for extreme condition traffic forecasting. In Proceedings of the 2017 SIAM International Conference on Data Mining (pp. 777-785). Society for Industrial and Applied Mathematics.
14. Fu, R., Zhang, Z., & Li, L. (2016, November). Using LSTM and GRU neural network methods for traffic flow prediction. In 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC) (pp. 324-328). IEEE.
15. Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies, 54, 187-197.
16. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.
17. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232.
18. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232.
19. Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015, June). An empirical exploration of recurrent network architectures. In International Conference on Machine Learning (pp. 2342-2350).
20. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
21. Johansen, A. R., Sønderby, C. K., Sønderby, S. K., & Winther, O. (2017, August). Deep recurrent conditional random field network for protein secondary prediction. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 73-78). ACM.
22. Jurtz, V. I., Johansen, A. R., Nielsen, M., Almagro Armenteros, J. J., Nielsen, H., Sønderby, C. K., ... & Sønderby, S. K. (2017). An introduction to deep learning on biological sequence data: examples and solutions. Bioinformatics, 33(22), 3685-3690.
23. Guo, Y., Li, W., Wang, B., Liu, H., & Zhou, D. (2019). DeepACLSTM: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. BMC bioinformatics, 20(1), 341.

AUTHORS PROFILE

Thendral R, Research Scholar, Computer Science and Engineering, Annamalai University, India. She finished Master Engineering (CSE) in sathiyabama University. She has teaching experience 15 years. Her interested areas are Artificial Intelligence, Computer Graphics, Data Mining.

Dr. AN. SIGAPPI, received her Ph.D in Computer Science and Engineering from Annamalai University in 2013. She did her Master Degree in Computer science and engineering from Anna University. Currently she is serving as a Professor in the Department of Computer Science and Engineering, Annamalai University, India. Her areas of interest include Image Processing, Machine Learning, and Data Analytics. She has published more than 25 research articles in international journals and conferences.