Prediction Protein-Protein Interactions with LSTM

Zheng Tao[1], Jiahao Yao[1], Chao Yuan[1], Ning Zhao[1], Bin Yang[2](✉), Baitong Chen[3], and Wenzheng Bao[1]

[1] School of Information Engineering, Xuzhou University of Technology, Xuzhou 221000, China
[2] School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277160, China
[3] Xuzhou No. 1 People's Hospital, Xuzhou 221000, China

Abstract. As the basis and key of cell activities, protein plays an important role in many life activities. Protein usually does not work alone. Under normal circumstances, most proteins perform specific functions by interacting with other proteins, and play the greatest role in life activity. The prediction of protein-protein interaction (PPI) is a very basic and important research in bioinformatics. PPI controls a large number of cell activities and is the basis of most cell activities. It provides a very important theoretical basis and support for disease prevention and treatment, and drug development. Because experimental methods are slow and expensive, methods based on machine learning are gradually being applied to PPI problems. We propose a new model called BiLSTM-RF, which can effectively predict PPI.

Keywords: Deep learning · Protein-protein interaction · LSTM

1 Introduction

In 2020, the outbreak of the COVID-19 epidemic ignited the enthusiasm of pharmaceutical companies for the protein [1, 2]. With the inactivation of domestic COVID-19 and the approval of adenovirus and recombinant protein vaccines, the progress of vaccine has become the focus of attention [3]. On November 9,2020, Pfizer of the United States announced that the mRNA novel coronavirus vaccine BNT162B2, developed in collaboration with BioNTech in Germany, is effective in healthy subjects not infected with 2020 and has contracted novel coronavirus and is more than 90% effective. This was followed by that, on November 16, 2020, novel researches released the results of phase 3 clinical experiments with the COVID-19 vaccine, mPROTEIN-1273, indicating that the vaccine has 94.5% protective efficacy [4–6]. In June 2021, the mRNA vaccine ARCov, developed by Amber Bio, is about to become the first mRNA vaccine in China to enter the phase 3 clinical stage. According to statistics, more than there are 200 COVID-19 vaccines in development worldwide, 48 of which have entered the clinical research stage. Now, PROTEIN therapy has become a hot investment spot of heavy capital bets, and more and

D. Jiang and H. Song (Eds.): SIMUtools 2021, LNICST 424, pp. 540–545, 2022.
https://doi.org/10.1007/978-3-030-97124-3_41

more academic research institutions and pharmaceutical companies are making efforts in this field [7–9]. Thus, protein has become a research direction that cannot be ignored in the current market.

Within organisms, protein has very important cellular functions, including synthetic proteins, catalytic reactions, regulating gene expression, regulating innate immunity, and sensing small molecules. On the one hand, protein is an important link of genetic information expression [10–12]. On the other hand, protein also modulates some important life activities. In recent years, with the researchers' research on PROTEIN structure, various scientific research achievements have come in. For example, scientists from institutions like the University of Toronto have developed a new technique called CHyMErA (Cas Hybrid for Multiplexed Editing and Screening applications) that is applied to any mammalian cells while systematically targeting DNA fragments at multiple sites. Professor Gao Caixa of the Institute of Genetic and Developmental Biology of the Chinese Academy of Sciences and his team optimized a guided editing system (prime editing system, PPE) to produce the required point mutations, insertions, and deletions in the two major cereal crops. The main components of the PPE system are the Cas9 incision ase-RT fusion protein and protein [13–16].

In this paper, a new prediction framework was constructed to predict different proteins according to their characteristic numbers. We compare the proportion of positive and negative samples, and test the accuracy of the proportion of positive and negative samples. The results show that, although the model has satisfactory prediction performance, it has some shortcomings, such as embedding more advanced super parameter selection scheme in the system, further optimizing the proposed deep learning framework, Deep learning methods are time-consuming, and need to pay more attention to deep learning methods based on PU and heterogeneous computing, which can be strengthened in future research.

2 Methods and Materials

The data should contain more characteristic values, and the total data should not be too small. In that case, the network of training is not fit well, and the classification effect will not be very good. The physical and chemical properties of the proteins can reflect the differences between different proteins and can be used as the characteristic number. Then the proteins are roughly divided into two categories, and the positive and negative labels are taken as the last column of the data. Finally, the characteristic number of proteins is 406, and 407 columns of data are obtained by adding tag column. For classification problems.

The performance of a model is determined by many things. One of the most important factors is feature selection. Extraction method. Feature extraction can change our Convert raw data into features that better represent the data. Improving the prediction accuracy of unknown data. It directly affects the prediction results of the model. At present, researchers have proposed many feature extraction methods which are dedicated to extracting the most effective features from data for classification and identification.

To minimize training errors, Gradient descent methods, such as applying a time series back-propagation algorithm, can be used to modify the weights of each time based on errors. The main problem with gradient descent in recurrent neural networks (RNN) was first discovered in 1991 when the error gradient disappeared exponentially with the length of time between events. When LSTM blocks are set, the error is also calculated backwards, from output to each gate in the input stage until the value is filtered out. Normal retransmitting nerves are therefore an effective way to train LSTM blocks to remember long-term values.

After obtaining a group of data with positive and negative sample labels, ten-layer cross validation is carried out to obtain the training set and test set, and then the labels and data are separated and processed respectively. Here, a group of positive and negative sample ratio of 1:10 is taken as an example. In order to get the best proportion of training, we test it from 1:1 to 1:7, and judge its accuracy by ACC value and ROC curve. Acc reflects the model's ability to classify positive samples correctly.

Therefore, training data should be converted to cell array, and training data label should be converted to category type.

The basic model structure of LSTM can be made with deep network designer in MATLAB. The characteristic number of the data is determined, and the corresponding parameters are written according to the model structure of LSTM. Based on this model structure, we can clearly see the internal structure and data function of LSTM at each layer. The parameters of each layer should be modified according to the actual data.

Sequence input layer (inputsize): sequence input layer, specify input dimensionlstm-Layer: Determine the number of LSTM hidden units.

Lstmlayer (numhiddenunits, 'outputmode', 'last'): bidirectional LSTM layer. It specifies the hidden node, and the output mode is 'last', that is, the last classification value is output.

Fully connected layer (numclasses): full connected layer, specifying the number of output classes.

Softmax layer: this layer outputs the probability of each category.

Classificationlayer: classification layer, which outputs the final classification results, similar to probability competition voting.

Determined the relevant parameters of training options, including optimization function, initial learning rate and iteration times. LSTM does not support multi-core operation, so in the running environment parameter, only CPU or GPU can be selected. But GPU support for convolutional neural networks requires a GPU device with compute capability 3.0 or higher.

The problem of fitting should be considered in the training process. If the correct rate of training is very high, but the correct rate applied to the test set is very low, then the practicability of this network will be very low, that is, there is the problem of over fitting.

Through the training set to get the network, and then put the test set into the network to test, get the final prediction result through the network, and compare the result with the label of the test set to get the correct rate. Adjust the proportion of relevant parameters and training set test set in time.

3 Results

The main analysis method of ROC is to draw this characteristic curve. The abscissa of the curve is the false positive rate (FPR), and N is the number of true negative samples, FP is the number of positive samples predicted by the classifier in n negative samples. The ordinate was true positive rate (TPR), P is the number of real positive samples, TP is the number of positive samples predicted by the classifier.

If we want to make a quantitative analysis of the model, we need to introduce a new concept, namely AUC (area under ROC curve). This concept is very simple, that is, the area under the ROC curve. To calculate the AUC value, we only need to integrate along the transverse axis of the ROC. In real scenes, the ROC curve is generally above this line, so the AUC value is generally between 0.5 and 1. The larger the AUC, the better the performance of the model (Table 1).

Table 1. The performances of each ratio of positive and negative data.

pos:neg	1:1	1:2	1:3	1:4	1:5	1:6	1:7	1:8
Acc	0.7451	0.7360	0.7690	0.7546	0.7806	0.7530	0.7879	0.7506

Through the ten layer cross validation method to divide the training set and test set can also effectively give the network sufficient data for training. It can be seen from the results that the classification effect of LSTM is mostly better. It can be seen that the ACC value is generally above 0.7, that is to say, the overall classification effect of LSTM is relatively good. Through the different proportion of positive and negative samples, we can find that the ACC value is the largest at 1:7, that is to say, the network classification effect is the best at 1:7. Of course, for the classification problem, the influencing factors are not only the positive and negative sample ratio, but also the number of features of the data. When the number of features is less, the classification result is poor. When the number of features is more, the classification result is good, but the running speed will become worse. Therefore, we need to find the appropriate positive and negative sample ratio and accurate data eigenvalues. The effect of LSTM classification of such data can best meet people's needs.

The shortcomings of LSTM. LSTM allows information to be stored across arbitrary time lags, and error signals to be carried far back in time. This potential strength, however, can contribute to a weakness in some situations: the cell states S often tend to grow linearly during the presentation of a time series (the nonlinear aspects of sequence processing are left to the squashing functions and the highly nonlinear gates). If we present a continuous input stream, the cell states may grow in unbounded fashion, causing saturation of the output squashing function, H. This happens even if the nature of the problem suggests that the cell states should be reset occasionally, e.g. at the beginnings of new input sequences (whose starts, however, are not explicitly indicated by a teacher). Saturation will make H derivative vanish, thus blocking incoming errors, and make the cell output equal the output gate activation, that is, the entire memory cell will degenerate into an ordinary BPTT unit, so that the cell will cease functioning as a memory.

The problem did not arise in the experiments reported by Hochreiter and Schmidhuber because cell states were explicitly reset to zero before the start of each new sequence.

4 Conclusions

In this paper, a new prediction framework was constructed to predict different proteins according to their characteristic numbers. We compare the proportion of positive and negative samples, and test the accuracy of the proportion of positive and negative samples. The results show that, although the model has satisfactory prediction performance, it has some shortcomings, such as embedding more advanced super parameter selection scheme in the system, further optimizing the proposed deep learning framework, Deep learning methods are time-consuming, and need to pay more attention to deep learning methods based on PU and heterogeneous computing, which can be strengthened in future research.

Acknowledgement. This work is supported by the fundamental Research Funds for the Central Universities, 2020QN89, Xuzhou science and technology plan project (KC19142), the talent project of 'Qingtan scholar' of Zaozhuang University, Jiangsu Provincial Natural Science Foundation, China (SBK2019040953), Youth Innovation Team of Scientific Research Foundation of the Higher Education Institutions of Shandong Province, China (2019KJM006), the Key Research Program of the Science Foundation of Shandong Province (ZR2020KE001), the PhD research startup foundation of Zaozhuang University (2014BS13) and Zaozhuang University Foundation (2015YY02), the Natural Science Foundation of China (61902337), Natural Science Fund for Colleges and Universities in Jiangsu Province (19KJB520016), Xuzhou Natural Science Foundation KC21047 and Young talents of science and technology in Jiangsu.

References

1. Brohee, S., Van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinform. **7**(1), 1–19 (2006)
2. Sugaya, N., Ikeda, K.: Assessing the druggability of protein-protein interactions by a supervised machine-learning method. BMC Bioinform. **10**(1), 1–13 (2009)
3. Shen, J., et al.: Predicting protein–protein interactions based only on sequences information. Proc. Natl. Acad. Sci. **104**(11), 4337–4341 (2007)
4. Zhang, Q.C., et al.: Structure-based prediction of protein–protein interactions on a genome-wide scale. Nature **490**(7421), 556–560 (2012)
5. Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T.P., Hautaniemi, S.: Integrated network analysis platform for protein-protein interactions. Nat. Methods **6**(1), 75–77 (2009)
6. De Las Rivas, J., Fontanillo, C.: Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput. Biol. **6**(6), e1000807 (2010)
7. Zhang, Y.P., Zou, Q.: PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning. Bioinformatics **36**(13), 3982–3987 (2020)
8. Shen, Z., Lin, Y., Zou, Q.: Transcription factors–DNA interactions in rice: identification and verification. Brief Bioinform. **21**(3), 946–956 (2020)

9. Liu, G.H., Shen, H.B., Yu, D.J.: Prediction of protein–protein interaction sites with machine-learning-based data-cleaning and post-filtering procedures. J. Membr. Biol. **249**(1), 141–153 (2016)
10. Sato, T., et al.: Interactions among members of the BCL-2 protein family analyzed with a yeast two-hybrid system. Proc. Natl. Acad. Sci. **91**(20), 9238–9242 (1994)
11. Schwikowski, B., Uetz, P., Fields, S.: A network of protein–protein interactions in yeast. Nat. Biotechnol. **18**(12), 1257–1261 (2000)
12. Coates, P.J., Hall, P.A.: The yeast two-hybrid system for identifying protein–protein interactions. J. Pathol.: A J. Pathol. Soc. Great Br. Ireland **199**(1), 4–7 (2003)
13. Free, R.B., Hazelwood, L.A., Sibley, D.R.: Identifying novel protein-protein interactions using co-immunoprecipitation and mass spectroscopy. Curr. Protoc. Neurosci. **46**(1), 5–28 (2009)
14. Kim, Y., Subramaniam, S.: Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. Proteins: Struct. Funct. Bioinform. **62**(4), 1115–1124 (2006)
15. Zhang, S.W., Hao, L.Y., Zhang, T.H.: Prediction of protein–protein interaction with pairwise kernel support vector machine. Int. J. Mol. Sci. **15**(2), 3220–3233 (2014)
16. Burger, L., Van Nimwegen, E.: Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. Mol. Syst. Biol. **4**(1), 165 (2008)
17. You, Z.H., Zhu, L., Zheng, C.H., Yu, H.J., Deng, S.P., Ji, Z.: Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. BMC Bioinform. **15**(15), 1–9 (2014)
18. Cui, G., Fang, C., Han, K.: Prediction of protein-protein interactions between viruses and human by an SVM model. BMC Bioinform. **13**(7), 1–10 (2012)
19. Bradford, J.R., Westhead, D.R.: Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics **21**(8), 1487–1494 (2005)
20. Guo, Y., Yu, L., Wen, Z., Li, M.: sing support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res. **36**(9), 3025–3030 (2008)
21. Koike, A., Takagi, T.: Prediction of protein–protein interaction sites using support vector machines. Protein Eng. Des. Sel. **17**(2), 165–173 (2004)
22. Yi, H.C., You, Z.H., Wang, M.N., Guo, Z.H., Wang, Y.B., Zhou, J.R.: RPI-SE: a stacking ensemble learning framework for ncRNA-protein interactions prediction using sequence information. BMC Bioinform. **21**(1), 1–10 (2020)
23. Du, X., Sun, S., Hu, C., Yao, Y., Yan, Y., Zhang, Y.: DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. J. Chem. Inf. Model. **57**(6), 1499–1510 (2017)
24. Sun, T., Zhou, B., Lai, L., Pei, J.: Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinform. **18**(1), 1–8 (2017)
25. Zhang, L., Yu, G., Xia, D., Wang, J.: Protein–protein interactions prediction based on ensemble deep neural networks. Neurocomputing **324**, 10–19 (2019)
26. Kong, M., Zhang, Y., Xu, D., Chen, W., Dehmer, M.: FCTP-WSRC: protein–protein interactions prediction via weighted sparse representation based classification. Front. Genet. **11**, 18 (2020)
27. Ma, W., Cao, Y., Bao, W., Yang, B., Chen, Y.: ACT-SVM: prediction of protein-protein interactions based on support vector basis model. Sci. Program. **2020** (2020)