Efficient Framework for Predicting ncRNA-Protein Interactions Based on Sequence Information by Deep Learning

Zhao-Hui Zhan[1], Zhu-Hong You[2(✉)], Yong Zhou[1], Li-Ping Li[2], and Zheng-Wei Li[1]

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 21116, China
[2] The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi 830011, China
zhuhongyou@ms.xjb.ac.cn

Abstract. The interactions between proteins and RNA (RPIs) play a crucial role in most cellular processes such as RNA stability and translation. Although there have been many high-throughput experiments recently to detect RPIs, these experiments are largely time-consuming and labor-intensive. Therefore, it is imminent to propose an efficient computational method to predict RPIs. In this study, we put forward a novel approach for predicting protein and ncRNA interactions based on sequences information only. By employing the bi-gram probability feature extraction method and k-mer algorithm, the represent features from protein and ncRNA were extracted. To evaluate the performance of the proposed model, two widely used datasets named RPI1807 and RPI2241 were trained with the adoption of random forest classifier by using five-fold cross-validation. The experimental results with the AUC of 0.992 and 0.947 on dataset RPI1807 and RPI2241 respectively indicated the effectiveness of our experimental approach for predicting RPIs, which provided the guidance for reference for future research in the biological field.

Keywords: Protein-ncRNA interaction · Bi-gram · Deep learning
Stacked autoencoder · PSSM

1 Introduction

In recent studies in the field of biological knowledge, more and more experiments have shown that ncRNA plays a vital role in the complex cell processes such as cellular proliferation and differentiation [1], chromatin modification [2], cellular apoptosis and so on [3]. At the meantime, a large number of ncRNA have been discovered with the development of modern advanced science and technology while their functions are not yet exactly known [4]. Therefore, it is imminent to make clear the functions of these ncRNAs. To learn about functions of these ncRNAs, researchers are required to identify whether these ncRNAs were able to interact with other proteins in some process of biological reactions [5–11]. However, there still are some shortcomings and improved space in the current prediction methods. Therefore, extracting feature

D.-S. Huang et al. (Eds.): ICIC 2018, LNCS 10955, pp. 337–344, 2018.
https://doi.org/10.1007/978-3-319-95933-7_41

information from sequences is a necessary method which can well identify the interactions proved by large number of research between ncRNA and protein [12–18].

In this study, we put forward a sequence-based method using deep learning model Stacked-autoencoder network combined with Random Forests (RF) classifier. We used K-mers sparse matrices to represent RNA sequences, and then extracted feature vector from matrix by Singular Value Decomposition (SVD). Position Specific Scoring Matrix (PSSM) was used to obtain evolutionary information from each sequence while Bi-gram was further used to get feature vector from PSSM. Then data and label was fed into RF classifier to classify whether a pair of protein and ncRNA interact or not. Furthermore, to evaluate the performance of our approach, five-fold cross validated and two widely used dataset RPI1807 and RPI2241 was used. The experimental results show that our method achieved high accuracy and robustness of the protein-ncRNA interaction prediction tasks.

2 Materials and Methods

2.1 Datasets

We executed experiments on two widely used public datasets including RPI1807 and RPI2241. The dataset RPI1807 consists of 1807 positive ncRNA-protein interaction pairs and 1436 negative ncRNA-protein pairs, including 1078 RNA chains, 1807 protein chains, 493 RNA chains and 1436 protein chains, respectively [19]. It is established by parsing the Nucleic Acid Database (NAD) which provides the RNA-protein complex data and protein-RNA interface database. The RPI2241 dataset is constructed in a similar way, and contains 2241 interacting RNA-protein pairs.

2.2 Features Extraction

To extracted features from ncRNA sequences, k-mer sparse matrix approach was used. A two-dimensional matrix deformation memory to store the features of ncRNA which can express much more useful and significant information such as frequency and location information [20]. An input ncRNA sequence is converted into a $4^k \times (L-k+1)$ matrix M can be defined as follow.

$$M = \left(a_{ij}\right)_{4^k} \times (L-k+1) \tag{1}$$

$$a_{ij} = \begin{cases} 1, & if\ m_j m_{j+1} m_{j+2} m_{j+3} = k-mer(i) \\ 0, & else \end{cases} \tag{2}$$

After obtaining the corresponding two-dimensional matrix from the original sequence of ncRNA, we transform this matrix with large amounts of data by way of singular value decomposition (SVD) [21].

And as well, we extracted protein features from the PSSM matrix calculated from the original protein sequence instead using it directly, since the combinations of amino acid cannot all be found in the original protein sequence [22]. To extract the features

recognized from the protein fold, we proposed a bi-gram feature extraction technique computed through the representing information mainly contained from PSSM [23].

The bi-gram occurrence matrix B can be calculated as follows and $b_{m,n}$ be the element in the matrix B:

$$B = \left\{ b_{m,n}, 1 \le m \le 20, 1 \le n \le 20 \right\} \tag{3}$$

$$b_{m,n} = \sum_{i=1}^{r-1} p_{i,m} p_{i+1,n}, i \le m \le 20, 1 \le n \le 20 \tag{4}$$

where $b_{m,n}$ can be interpreted as the occurrence probability of the transition from m_{th} amino acid to n_{th} amino acid which is able to calculated from the element $p_{i,j}$ in its PSSM matrix [24]. Let F be the bi-gram feature vector of the protein fold recognition which is as follows:

$$F = \left\{ b_{1,1}, b_{1,2}, \cdots, b_{1,20}, b_{2,1}, \cdots, b_{2,20}, \cdots, b_{20,1}, \cdots, b_{20,20} \right\}^T \tag{5}$$

where the symbol T can be regarded as the transpose of the feature vector [25]. Then, the random forest classifiers were used to predict the interaction between ncRNA and protein.

2.3 Deep Learning Framework Based on Stacked Autoencoder

In order to improve the accuracy of the predicting performance, there had been many recent research which concentrated their attentions on automatic encoders and deep-learning networks [26–32]. In this study, we used the stacked auto-encoder network for deep learning and classification of training datasets to obtain an efficient deep learning network [33]. A complete stacked auto-encoder network consists of a sparse multilayer neural network auto-encoder which layer inputs can be obtained from the outputs of the previous layers [34]. With the hyper parameter optimization, we were able to get the best parameters of the stacked auto-encoder neural network suitable for our machine learning model [35]. The sparse auto-encoder network which was used to learn the feature changes is a single-layer automatic encoder as follows:

$$p_{(\alpha,\beta)}(x) = f\left(\alpha^T x\right) = f\left(\sum_{i=1}^{n} \alpha_i x_i + \beta_i\right) \tag{6}$$

where the input x can be interpreted as the d-dimension dataset and $f(x)$ is an activation function. And the auto-encoder network maps X into the output $p(X)$. And Sigmoid was selected as activation function as follows:

$$f(y) = \frac{1}{1 + e^{-y}} \tag{7}$$

And consequently, the loss function is as follows:

$$H(X, \alpha) = \|\alpha p - X\|^2 + \omega \sum_j |p(j)| \tag{8}$$

The stacked neural network architecture is composed of multiple neural network layers which outputs of the previous layers are the inputs of next layers [36]. At the meantime, the keras library from Internet was used to implement stacked auto-encoder and the parameters *batch_size* and *nb_epoch* both set to be 100 [37]. The details about keras can be found in website http://github.com/fchollet/keras.

2.4 Stacked Ensemble

In order to find out the solution of assembling mechanism implementing to integrate every individual output from classifiers to implement multi-classifier assembling and obtain an approximately optimal objective function [8, 38–40], we regarded the outputs of all level 0 classifiers as predicted probability scores while the successive level 1 classifiers as logistic regression classifiers. The experimental results shown that stacked assembling was equal to the average individual model results strategy when score weights of logistic regression of all individual level 0 classifiers were same.

$$P_w(y = \pm 1|s) = \frac{1}{1 + e^{-yw^T s}} \tag{9}$$

where s is predicted probability scores of all level 0 classifiers vector outputs and w is the weight vector of corresponding classifiers [41].

3 Experimental Results

The five-fold cross-validation method is used to evaluate the performance of our study, which randomly divides all the data set into five equal parts [42–45]. We followed the widely used evaluation measures to evaluate our method, including accuracy, sensitivity, specificity, precision and AUC [46–50]. The experimental results in dataset RPI1807 and RPI2241 were shown in Table 1.

Table 1. The experimental results in RPI1807 and RPI2241.

	Accuracy	Sensitivity	Specificity	Precision	AUC
RPI1807	0.9600	0.9344	0.9989	0.9117	0.9920
RPI2241	0.9130	0.8772	0.9660	0.8590	0.9470

According to the Table 1, our method achieved a decent performance with an accuracy of 0.9600, sensitivity of 0.9344, specificity of 0.9989, precision of 0.9117 and AUC of 0.9920 in testing dataset RPI1807 and an accuracy of 0.9130, sensitivity of 0.8772, specificity of 0.9660, precision of 0.8590 and AUC of 0.9470 in testing dataset RPI2241.

4 Conclusions

In this study, we proposed a sequence-based method using deep learning model Stacked-autoencoder network combined with RF classifier. By employing the k-mers sparse matrix and bi-gram algorithm, the represent ncRNA and protein features were extracted from the corresponding sequence information. In the process of experiments, our method has shown a satisfying performance for predicting RPIs on each reference dataset which thanks to the contribution of the Stacked ensemble autoencoder framework using deep learning. In general, our method tried to extract protein features and automatic learn the advanced features with the use of random forests classifiers, but still do not had a very good breakthrough achievement from the perspective of biology. In future research, we expect to design a better network architecture for extracting hidden advanced features from the perspective of biology.

References

1. Wapinski, O., Chang, H.Y.: Long noncoding RNAs and human disease. Trends Cell Biol. **21**(6), 354–361 (2011)
2. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P.: Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature **458**(7235), 223 (2009)
3. Yu, F., Zheng, J., Mao, Y., Dong, P., Li, G., Lu, Z., Guo, C., Liu, Z., Fan, X.: Long non-coding RNA APTR promotes the activation of hepatic stellate cells and the progression of liver fibrosis. Biochem. Biophys. Res. Commun. **463**(4), 679–685 (2015)
4. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S.: GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. **22**(9), 1760–1774 (2012)
5. Chen, X., You, Z.H., Yan, G.Y., Gong, D.W.: IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. Oncotarget **7**(36), 57919–57931 (2016)
6. Chen, X., Yan, C.C., Zhang, X., You, Z.H.: Long non-coding RNAs and complex diseases: from experimental results to computational models. Brief. Bioinform. **18**(4), 558 (2016)
7. Wang, Y.B., You, Z.H., Li, X., Jiang, T.H., Chen, X., Zhou, X., Wang, L.: Predicting protein-protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. Mol. BioSyst. **13**(7), 1336–1344 (2017)
8. Li, S., You, Z.H., Guo, H., Luo, X., Zhao, Z.Q.: Inverse-free extreme learning machine with optimal information updating. IEEE Trans. Cybern. **46**(5), 1229 (2016)
9. Lei, W., You, Z.H., Xing, C., Li, J.Q., Xin, Y., Wei, Z., Yuan, H.: An ensemble approach for large-scale identification of protein-protein interactions using the alignments of multiple sequences. Oncotarget **8**(3), 5149–5159 (2016)
10. Huang, Q., You, Z., Zhang, X., Zhou, Y.: Prediction of protein-protein interactions with clustered amino acids and weighted sparse representation. Int. J. Mol. Sci. **16**(5), 10855–10869 (2015)
11. Huang, Y.A., You, Z.H., Chen, X.: A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences. Curr. Protein Pept. Sci. **5**(19), 468–478 (2017)
12. You, Z.H., Huang, Z.A., Zhu, Z., Yan, G.Y., Li, Z.W., Wen, Z., Chen, X.: PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLoS Comput. Biol. **13**(3), e1005455 (2017)

13. Li, Z.W., You, Z.H., Chen, X., Li, L.P., Huang, D.S., Yan, G.Y., Nie, R., Huang, Y.A.: Accurate prediction of protein-protein interactions by integrating potential evolutionary information embedded in PSSM profile and discriminative vector machine classifier. Oncotarget **8**(14), 23638 (2017)
14. An, J.Y., You, Z.H., Chen, X., Huang, D.S., Yan, G., Wang, D.F.: Robust and accurate prediction of protein self-interactions from amino acids sequence using evolutionary information. Mol. BioSyst. **12**(12), 3702 (2016)
15. An, J.Y., You, Z.H., Chen, X., Huang, D.S., Li, Z.W., Liu, G., Wang, Y.: Identification of self-interacting proteins by exploring evolutionary information embedded in PSI-BLAST-constructed position specific scoring matrix. Oncotarget **7**(50), 82440–82449 (2016)
16. Lei, Y.K., You, Z.H., Ji, Z., Zhu, L., Huang, D.S.: Assessing and predicting protein interactions by combining manifold embedding with multiple information integration. BMC Bioinform. **13**(S7), S3 (2012)
17. You, Z.H., Lei, Y.K., Gui, J., Huang, D.S., Zhou, X.: Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. Bioinformatics **26**(21), 2744 (2010)
18. 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**(S15), S9 (2014)
19. Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.: Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. **33**(8), 831–838 (2015)
20. Pan, X., Fan, Y.X., Yan, J., Shen, H.B.: IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. BMC Genom. **17**(1), 582 (2016)
21. Chen, H., Huang, Z.: Medical image feature extraction and fusion algorithm based on K-SVD. In: Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC 2015, GuangDong, pp. 333–337 (2015)
22. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D.: The database of interacting proteins: 2004 update. Nucleic Acids Res. **32**, D449–D451 (2004)
23. Chatraryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L.: The BioGRID interaction database: 2015 update. Nucleic Acids Res. **43**, D470 (2015)
24. Suresh, V., Liu, L., Adjeroh, D., Zhou, X.: Revealing protein–lncRNA interaction. Brief. Bioinform. **17**, 106 (2015)
25. Paliwal, K.K., Sharma, A., Lyons, J., Dehzangi, A.: A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition. IEEE Trans. Nanobiosci. **13**(1), 44–50 (2014)
26. You, Z.H., Zhou, M.C., Xin, L., Shuai, L.: Highly efficient framework for predicting interactions between proteins. IEEE Trans. Cybern. **PP**(99), 1–13 (2016)
27. Huang, Y.A., Chen, X., You, Z.H., Huang, D.S., Chan, K.C.C.: ILNCSIM: improved lncRNA functional similarity calculation model. Oncotarget **7**(18), 25902–25914 (2016)
28. Zhu, L., You, Z.H., Huang, D.S., Wang, B.: t-LSE: a novel robust geometric approach for modeling protein-protein interaction networks. PLoS ONE **8**(4), e58368 (2013)
29. Zhu, L., You, Z.H., Huang, D.S.: Increasing the reliability of protein–protein interaction networks via non-convex semantic embedding. Neurocomputing **121**(18), 99–107 (2013)
30. You, Z.H., Yin, Z., Han, K., Huang, D.S., Zhou, X.: A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. BMC Bioinform. **11**(1), 1–13 (2010)

31. Xia, J.F., You, Z.H., Wu, M., Wang, S.L., Zhao, X.M.: Improved method for predicting phi-turns in proteins using a two-stage classifier. Protein Pept. Lett. **17**(9), 1117 (2010)
32. You, Z.H., Li, X., Chan, K.C.: An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers. Neurocomputing **228**, 277–282 (2017)
33. Li, J.Q., Rong, Z.H., Chen, X., Yan, G.Y., You, Z.H.: MCMDA: matrix completion for MiRNA-disease association prediction. Oncotarget **8**(13), 21187 (2017)
34. Mchugh, C.A., Russell, P., Guttman, M.: Methods for comprehensive experimental identification of RNA-protein interactions. Genome Biol. **15**(1), 203 (2014)
35. Yi, H.-C., You, Z.-H., Huang, D.-S., Li, X., Jiang, T.-H., Li, L.-P.: A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. Mol. Ther. Nucleic Acids **11**, 337–344 (2018)
36. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**(12), 3371–3408 (2010)
37. Dahl, G.E., Sainath, T.N., Hinton, G.E.: Improving deep neural networks for LVCSR using rectified linear units and dropout. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, pp. 8609–8613 (2013)
38. You, Z.H., Li, J., Gao, X., He, Z., Zhu, L., Lei, Y.K., Ji, Z.: Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines. Biomed. Res. Int. **2015**(2), 1–9 (2015)
39. You, Z.H., Chan, K.C.C., Hu, P.: Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. PLoS ONE **10**(5), e0125811 (2015)
40. You, Z.H., Li, S., Gao, X., Luo, X., Ji, Z.: Large-scale protein-protein interactions detection by integrating big biosensing data with computational model. Biomed. Res. Int. (2) (2014). https://doi.org/10.1155/2014/598129
41. Pedregosa, F., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**(10), 2825–2830 (2012)
42. Yuan, H., You, Z.H., Xing, C., Chan, K., Xin, L.: Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. BMC Bioinform. **17**(1), 184 (2016)
43. An, J.Y., You, Z.H., Meng, F.R., Xu, S.J., Wang, Y.: RVMAB: using the relevance vector machine model combined with average blocks to predict the interactions of proteins from protein sequences. Int. J. Mol. Sci. **17**(5), 757 (2016)
44. An, J.Y., Meng, F.R., You, Z.H., Fang, Y.H., Zhao, Y.J., Ming, Z.: Using the relevance vector machine model combined with local phase quantization to predict protein-protein interactions from protein sequences. Biomed. Res. Int. **2016**, 1–9 (2016)
45. Wong, L., You, Z.H., Ming, Z., Li, J., Chen, X., Huang, Y.A.: Detection of interactions between proteins through rotation forest and local phase quantization descriptors. Int. J. Mol. Sci. **17**(1), 21 (2015)
46. Wang, L., You, Z.H., Xia, S.X., Chen, X., Yan, X., Zhou, Y., Liu, F.: An improved efficient rotation forest algorithm to predict the interactions among proteins. Soft. Comput. **17**, 1–9 (2017)
47. Wang, L., You, Z.H., Chen, X., Yan, X., Liu, G., Zhang, W.: RFDT: a rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information. Curr. Protein Pept. Sci. **5**(19), 445–454 (2016)

48. Chen, X., Huang, Y.A., Wang, X.S., You, Z.H., Chan, K.C.: FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model. Oncotarget **7**(29), 45948 (2016)
49. Luo, X., You, Z., Zhou, M., Li, S., Leung, H., Xia, Y., Zhu, Q.: A highly efficient approach to protein interactome mapping based on collaborative filtering framework. Sci. Rep. **5** (7702), 7702 (2015)
50. Lei, Y.K., You, Z.H., Dong, T., Jiang, Y.X., Yang, J.A.: Increasing reliability of protein interactome by fast manifold embedding. Pattern Recognit. Lett. **34**(4), 372–379 (2013)