Using Weighted Extreme Learning Machine Combined with Scale-Invariant Feature Transform to Predict Protein-Protein Interactions from Protein Evolutionary Information

Jianqiang Li[1], Xiaofeng Shi[1], Zhuhong You[2(✉)], Zhuangzhuang Chen[1], Qiuzhen Lin[1], and Min Fang[1]

[1] Shenzhen University, Nanhai Road 3688, Shenzhen, China
{lijq,qiuzhlin}@szu.edu.cn, sxf7758258@sina.com, chenzhuangzh@qq.com, gracefangcs@163.com
[2] No. 40, Beijing South Road, Urumqi, China
zhuhongyou@ms.xjb.ac.cn

Abstract. Protein-Protein Interactions (PPIs) play an irreplaceable role in biological activities of organisms. Although many high-throughput methods are used to identify PPIs from different kinds of organisms, they have some shortcomings, such as high cost and time-consuming. To solve the above problems, computational methods are developed to predict PPIs. Thus, in this paper, we present a method to predict PPIs using protein sequences. First, protein sequences are transformed into Position Weight Matrix (PWM), in which Scale-Invariant Feature Transform (SIFT) algorithm is used to extract features. Then Principal Component Analysis (PCA) is applied to reduce the dimension of features. At last, Weighted Extreme Learning Machine (WELM) classifier is employed to predict PPIs and a series of evaluation results are obtained. In our method, since SIFT and WELM are used to extract features and classify respectively, we called the proposed method SIFT-WELM. When applying the proposed method on three well-known PPIs datasets of Y east, Human and Helicobacter.pylori, the average accuracies of our method using five-fold cross validation are obtained as high as 94.83%, 97.60% and 83.64%, respectively. In order to evaluate the proposed approach properly, we compare it with Support Vector Machine (SVM) classifier in different aspects.

Keywords: Protein-protein interactions · Scale-invariant feature transform Weighted extreme learning machine

1 Introduction

Protein-Protein Interactions (PPIs) get involved in many fundamental cellular functions, and the research on PPIs helps us to understand the molecular mechanisms of biological processes and to propose some new methods in practical medical field. So it is necessary and urgent to carry out the study of PPIs.

D.-S. Huang et al. (Eds.): ICIC 2018, LNCS 10954, pp. 527–532, 2018.
https://doi.org/10.1007/978-3-319-95930-6_49

Nowadays, a large amount of high-throughput methods have been developed to predict PPIs, such as yeast two-hybrid (Y2H) screening methods [1, 2], immunoprecipitation [3], and protein chips [4]. However, there are some shortcomings in these experiments, such as high cost and time-consuming. Moreover, these methods yield high false positives and false negatives, which result in difficulties to predict unknown PPIs by experimental methods.

In addition, there are many biological databases, such as BIND [5], DIP [6] and MINT [7]. Protein sequences occupy an overwhelming advantage in quantity in these databases, so in order to efficiently utilize these sequence data, it is necessary to develop computational methods to predict PPIs from protein sequences. In general, sequence-based computational methods have two main parts: feature extraction and sample classification [8–10].

In first part, Scale-Invariant Feature Transform (SIFT) [11, 12] is applied to extract features from Position Weight Matrix (PWM) [13]. In order to reduce the effect of noise and shorten training time, Principal Component Analysis (PCA) is used to reduce the dimension of features.

In second part, Weighted Extreme Learning Machine (WELM) [14, 15] is used to identify protein pairs' interacting or non-interacting based on SIFT features. WELM only needs to set two parameters, which is fast to get the best parameters. Moreover, WELM gets better performance in generalization.

In this paper, a novel computational method based on SIFT algorithm and WELM is proposed to predict protein-protein interactions, which helps to insight into the molecular mechanisms of cells and explain the causes of some disease, and it may propose some new treatment methods in practical medical field.

2 Materials and Methods

2.1 Datasets

In our experiment, we collect Yeast dataset from DIP [6]. After removing protein pairs whose sequence length less than 50 and filtering out protein pairs whose sequence identity bigger than 40%, we get 5594 positive protein pairs, and we construct 5594 negative sample according to the results in [16].

To demonstrate the generality of our approach, we collect 3899 protein pairs as positive dataset by removing sequence identity bigger than 25%, and we construct 4262 negative protein pairs according to the work in [17]. In addition, Helicobacter. pylori dataset consists of 1458 positive protein sequence and 1458 negative protein sequence according to the result of Martin et al. [18].

2.2 Scale-Invariant Feature Transform

Scale-Invariant Feature Transform (SIFT) is an algorithm widely used in the field of computer vision, which can be applied to extract local features from images. SIFT was firstly introduced by Lowe in [11], which was summarized and perfected in [12]. SIFT algorithm can be applied in different fields, such as face recognition, 3D modeling and

template matching because of its robustness to rotation, scaling, viewpoint and so on. In this paper, SIFT is used to extract features.

2.3 Weighted Extreme Learning Machine

Extreme Learning Machine (ELM) [14] is a single hidden layer feed-forward neural network (SLFN) algorithm, which is simple in theory but effective in practice. ELM just needs to set the hidden nodes in network before the use, and ELM produces the unique optimal result, so it gets fast in learning and achieves better performance in generalization. Weighted ELM (WELM) is proposed to process the data with imbalanced class distribution [15], which can maintain the advantages of original ELM, and extend to cost-sensitive learning according to user's needs.

2.4 Evaluation Criteria

In order to evaluate the performance of our method, we use the following evaluation criteria: accuracy, sensitivity, precision and Matthews correlation coefficient (MCC). They are calculated as

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \tag{4}$$

where true positive (TP) stands for the number of true interacting pairs that predicted correctly; true negative (TN) represents the number of true non-interacting pairs that predicted correctly; false positive (FP) is the number of true non-interacting pairs that predicted incorrectly and false negative (FN) is the number of true interacting pairs that predicted to be non-interacting pairs falsely.

3 Results and Discussion

3.1 Evaluation of the Proposed Method

In our experiment, we set the same parameters for three datasets—Yeast, Human and H.pylori, which are classified by WELM. Here, $L = 10000$ and $C = 2^5$, where L means the number of hidden neurons and C represents the trade-off constant [15]. Five-fold cross validation is employed to evaluate the performance of our method, which can

avoid over-fitting problem of our model and evaluate the stability of our model [19]. Results of our method are shown in Tables 1, 2 and 3.

Table 1. Five-fold cross validation results of our method applied on *Yeast* dataset.

Test Sets	Accu.(%)	Prec.(%)	Sen.(%)	Mcc.(%)	Auc.(%)
1	95.36	94.89	96.21	90.71	97.92
2	95.29	95.07	95.80	90.58	98.54
3	93.86	91.04	96.56	87.86	98.41
4	94.51	93.49	95.69	89.04	97.42
5	95.11	95.11	94.99	90.22	97.82
Average	94.83 ± 0.64	93.92 ± 1.74	95.85 ± 0.59	89.68 ± 1.21	98.02 ± 0.46

Table 2. Five-fold cross validation results of our method applied on *Human* dataset.

Test Sets	Accu.(%)	Prec.(%)	Sen.(%)	Mcc.(%)	Auc.(%)
1	97.06	95.06	99.05	99.05	99.48
2	97.61	96.01	99.25	95.28	99.63
3	98.17	97.15	99.03	96.34	99.60
4	96.97	95.40	98.48	93.99	99.33
5	98.17	97.47	98.90	96.34	99.59
Average	97.60 ± 0.57	97.60 ± 0.57	98.94 ± 0.29	95.23 ± 1.12	99.53 ± 0.12

Table 3. Five-fold cross validation results of our method applied on *H. pylori* dataset.

Test Sets	Accu.(%)	Prec.(%)	Sen.(%)	Mcc.(%)	Auc.(%)
1	85.02	89.29	78.13	70.35	90.74
2	82.02	87.20	77.30	64.64	89.32
3	86.14	88.28	83.70	72.40	90.06
4	83.52	84.17	80.16	66.92	87.93
5	81.48	86.18	76.26	63.51	89.88
Average	83.64 ± 1.97	87.02 ± 1.98	79.11 ± 2.94	67.56 ± 3.76	89.59 ± 1.06

From above tables, we can refer that WELM classifier combining with SIFT descriptors can predict PPIs effectively, and the low standard deviations of the results indicate that our approach is robust. The excellent results of our method lie in the following reasons: (1) When compared to sequence dataset, the corresponding PWM matrix can retain more prior information. (2) The SIFT descriptors extracted from datasets retain abundant information of protein pairs and have strong ability to resist noise. (3) WELM is faster than traditional neural network algorithm in training while guaranteeing the learning accuracy.

3.2 Comparison with SVM-Based Method

To further evaluate our method, we compare results of the proposed approach with the widely used SVM classifier LIBSVM, which is developed by professor Chih-Jen Lin of National Taiwan University [20]. From Table 4, we notice that WELM achieves better performance than SVM when proposing classification on Yeast, Human and H. pylori datasets. Thus we can conclude that WELM is superior to SVM.

Table 4. Performance comparison between the SIFT+WELM and the SVM prediction models

Dataset	Classifier	Accu.(%)	Prec.(%)	Sen.(%)	Mcc.(%)	Time (s)
Yeast	WELM	94.83 ± 0.64	93.92 ± 1.74	95.85 ± 0.59	89.68 ± 1.21	113.9 ± 1.6
	SVM	91.27 ± 1.06	90.39 ± 1.17	92.05 ± 0.55	82.55 ± 2.11	1033.2 ± 1.6
Human	WELM	97.60 ± 0.57	96.22 ± 1.06	98.94 ± 0.29	95.23 ± 1.12	56.7 ± 1.2
	SVM	96.55 ± 0.71	96.15 ± 1.49	97.12 ± 0.44	93.11 ± 1.41	403.7 ± 4.2
H.pylori	WELM	83.64 ± 1.97	87.02 ± 1.98	79.11 ± 2.94	67.56 ± 3.76	5.4 ± 0.3
	SVM	80.49 ± 1.40	77.79 ± 2.60	82.30 ± 2.72	61.11 ± 2.73	17.2 ± 0.1

4 Conclusions

The use of computational methods to predict PPIs is becoming more and more important because of its low cost and high efficiency when compared to the experimental methods. In this paper, we propose a novel prediction model by using scale-invariant feature transform and weighted extreme learning machine to predict PPIs. When compared to SVM-based methods, our method can increase the accuracy and shorten the training time greatly. The experimental results indicate that the proposed method is efficient, feasible and robust.

Acknowledgement. This work is supported in part by the Natural Science Foundation of SZU under Grant CYZZ20160304165036893 and Grant 2016048, in part by the National Natural Science Foundation of China under Grant U1713212, Grant 61572330, and Grant 61602319, and in part by the Technology Planning Project from Guangdong Province, China, under Grant 2014B010118005.

References

1. Gavin, A.-C., Bsche, M., Krause, R.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature **415**(6868), 141–147 (2002)
2. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA **98**(8), 4569–4574 (2001)
3. Ho, Y., Gruhler, A., Heilbut, A.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature **415**(6868), 180–183 (2002)
4. Snyder, M., Zhu, H., Bertone, P., Bidlingmaier, S.M., Bilgin, M., Casamayor, A.J., Gerstein, M., Jansen, R., Lan, N.: Global analysis of protein activities using proteome chips, p. 2101 (2004)

5. Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E.: The biomolecular interaction network database and related tools 2005 update. Nucleic Acids Res. **33**(Database issue), 418–424 (2005)
6. 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**(1), D449 (2004)
7. Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E.: MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. **35**(Database issue), 572–574 (2012)
8. 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)
9. You, Z.H., Lei, Y.K., Zhu, L., Xia, J., Wang, B.: Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. BMC Bioinf. **14**(S8), 1–11 (2013)
10. Huang, Y.A., You, Z.H., Gao, X., Wong, L., Wang, L.: Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. BioMed Res. Int. **2015**, 1–10 (2015)
11. Lowe, D.G.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, p. 1150 (2002)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
13. Stormo, G.D., Schneider, T.D., Gold, L., Ehrenfeucht, A.: Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. U.S. Department of Commerce, National Bureau of Standards: for sale by the Superintendent of Documents, U.S. Government Printing Office (1982)
14. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 Proceedings of the IEEE International Joint Conference on Neural Networks, vol. 2, pp. 985–990 (2005)
15. Zong, W., Huang, G.B., Chen, Y.: Weighted extreme learning machine for imbalance learning. Neurocomputing **101**(3), 229–242 (2013)
16. Guo, Y., Yu, L., Wen, Z., Li, M.: Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res. **36**(9), 3025–3030 (2008)
17. You, Z.H., Yu, J.Z., Zhu, L., Li, S., Wen, Z.K.: A MapReduce based parallel SVM for large-scale predicting protein-protein interactions. Neurocomputing **145**(18), 37–43 (2014)
18. Martin, S., Roe, D., Faulon, J.L.: Predicting protein-protein interactions using signature products. Curr. Opin. Struct. Biol. **21**(2), 218 (2005)
19. Jiao, Y., Du, P.: Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quant. Biol. **4**, 1–11 (2016)
20. Lin, C.H., Liu, J.C., Ho, C.H.: Anomaly detection using LibSVM training tools. In: International Conference on Information Security and Assurance, pp. 166–171 (2008)