A SVM-Based System for Predicting Protein-Protein Interactions Using a Novel Representation of Protein Sequences

Zhuhong You[1], Zhong Ming[1], Ben Niu[1], Suping Deng[2], and Zexuan Zhu[1]

[1] College of Computer Science and Software Engineering, Shenzhen University
Shenzhen, Guangdong 518060, China
[2] Department of Computer Science and Technology, Tongji University
Shanghai 201804, P.R. China
zhyou@szu.edu.cn

Abstract. Protein-protein interactions (PPIs) are crucial for almost all cellular processes, including metabolic cycles, DNA transcription and replication, and signaling cascades. However, the experimental methods for identifying PPIs are both time-consuming and expensive. Therefore, it is important to develop computational approaches for predicting PPIs. In this article, a sequence-based method is developed by combining a novel feature representation using binary coding and Support Vector Machine (SVM). The binary-coding-based descriptors account for the interactions between residues a certain distance apart in the protein sequence, thus this method adequately takes the neighboring effect into account and mine interaction information from the continuous and discontinuous amino acids segments at the same time. When performed on the PPI data of *Saccharomyces cerevisiae*, the proposed method achieved 86.93% prediction accuracy with 86.99% sensitivity at the precision of 86.90%. Extensive experiments are performed to compare our method with the existing sequence-based method. Achieved results show that the proposed approach is very promising for predicting PPI, so it can be a useful supplementary tool for future proteomics studies.

Keywords: protein-protein interaction, binary coding, support vector machine, protein sequence, local descriptor.

1 Introduction

Proteins are crucial for almost all of functions in the cell, including metabolic cycles, DNA transcription and replication, and signaling cascades. Usually, proteins rarely perform their functions alone; instead they cooperate with other proteins by forming a huge network of protein-protein interactions (PPIs). PPIs are responsible for the majority of cellular functions. In the past decades, many innovative techniques for detecting PPIs have been developed [1-3]. Due to the progress in large-scale experimental technologies such as yeast two-hybrid (Y2H) screens [2, 4], tandem affinity purification (TAP) [1], mass spectrometric protein complex identification

D.-S. Huang et al. (Eds.): ICIC 2013, LNCS 7995, pp. 629–637, 2013.

(MS-PCI) [3] and other high-throughput biological techniques for PPIs detection, a large amount of PPIs data for different species has been accumulated [1-6]. However, the experimental methods are costly and time consuming, therefore current PPI pairs obtained from experiments only covers a small fraction of the complete PPI networks [7, 8]. In addition, large-scale experimental methods usually suffer from high rates of both false positive and false negative predictions [9-12]. Hence, it is of great practical significance to develop the reliable computational methods to facilitate the identification of PPIs [7].

A number of computational methods have been proposed for the prediction of PPIs based on different data types, including phylogenetic profiles, gene neighborhood, gene fusion, and sequence conservation between interacting proteins, literature mining knowledge. There are also methods that combine interaction information from several different data sources [13]. However, these methods cannot be implemented if such pre-knowledge about the proteins is not available. Recently, a couple of methods which derive information directly from amino acid sequence are of particular interest [14, 15]. Many researchers have engaged in the development of sequences-based method for discovering new PPIs, and the experiment results showed that the information of amino acid sequences alone is sufficient to predict PPIs. Among them, one of the excellent works is a SVM-based method developed by Shen et al [15]. In the study, the 20 amino acids were clustered into seven classes according to their dipoles and volumes of the side chains, and then the conjoint triad method abstracts the features of protein pairs based on the classification of amino acids. When applied to predict human PPIs, this method yields a high prediction accuracy of 83.9%. Because the conjoint triad method cannot takes neighboring effect into account and the interactions usually occur in the discontinuous amino acids segments in the sequence, on the other work Guo et al. developed a method based on SVM and auto covariance to extract the interactions information in the discontinuous amino acids segments in the sequence [7]. Their method yielded a prediction accuracy of 86.55%, when applied to predicting *saccharomyces cerevisiae* PPIs. In our previous works, we also obtained good prediction performance by using autocorrelation descriptors and correlation coefficient, respectively [16,17].

In this study, we report a new sequence-based method for the prediction of interacting protein pairs using SVM combined with binary coding. More specifically, we first represent each protein sequence as a vector by utilizing a binary-coding-based representation of protein sequence which provides us with a chance to mine interaction information from the continuous and discontinuous amino acids segments at the same time [18]. The effectiveness of binary-coding-based descriptors depends largely on the correct selection of amino acid grouping [18]. By grouping amino acids into a reduced alphabet, we can create a more accurate protein sequence representation. Here, we adopted the amino acids grouping according to the successful use of classification in [14]. Then we characterize a protein pair in different feature vectors by coding the vectors of two proteins in this protein pair. Finally, an SVM model is constructed using these feature vectors of the protein pair as input. To evaluate the performance, the proposed method was applied to *Saccharomyces cerevisiae* data. The experiment results show that our method achieved 86.93% prediction accuracy with 86.99% sensitivity at the precision of 86.90%.

2 Materials and Methodology

In this paper, we have presented a new approach to predict PPIs using support vector machine (SVM) from protein sequences. Our method for predicting the PPIs depends on three steps: (1) Generation of the PPI dataset; (2) Feature vector extraction; (3) Classification using SVM.

2.1 Generation of the Data Set

We evaluated the proposed method with the data from yeast used in the study of Guo et al. [7]. The PPI dataset was collected from *Saccharomyces cerevisiae* core subset of Database of Interacting Proteins (DIP). After the redundant protein pairs which contain a protein with fewer than 50 residues or have ≥40% sequence identity were remove, the remaining 5594 protein pairs comprise the final positive dataset. The 5594 non-interacting protein pairs were generated from pairs of proteins whose sub-cellular localizations are different. The whole dataset consists of 11188 protein pairs, where half are from the positive dataset and half are from the negative dataset. Note that we have used exactly the same non-redundant dataset as used in Guo et al. [7]. Four-fifths of the protein pairs from the positive and negative dataset were respectively randomly selected as the training dataset and the remaining one-fifths were used as the test dataset.

2.2 Feature Vector Extraction

To use machine learning methods to predict PPIs from protein sequences, one of the most important computational challenges is to extract feature vectors from protein sequences in which the important information content of proteins is fully encoded. In this section, we adopt a novel sequence representation model by using binary coding based descriptors.

There are three types of local descriptors used in the aforementioned studies: Composition, Transition and Distribution, which are computed based on the variation of occurrence of functional groups of amino acids within the primary sequence of the protein. In this study, the 20 amino acids were firstly clustered into seven functional groups based on the dipoles and volumes of the side chains. The functional groups used were: Cluster_1 (amino acids A,G,V), Cluster_2 (amino acids C), Cluster_3 (amino acids D,E), Cluster_4 (amino acids F,I,L,P), Cluster_5 (amino acids H,N,Q,W), Cluster_6 (amino acids K,R) and Cluster_7 (amino acids M,S,T,Y). In total there would be 63 features (7 composition, 21 transition, 35 distribution) if they were computed from the whole amino acid sequence.

In order to extract the interaction information of protein sequences, we split the protein sequences into fifteen different regions of varying length and composition to describe multiple overlapping continuous and discontinuous interaction patterns within a protein sequence. we first divided the entire protein sequence into four equal length regions (A-D). Then a novel binary-coding-based method was adopted to construct a couple of continuous and discontinuous regions on the basis of above partition (A-D). Here the continuous regions are composed of residues which are local in the polypeptide sequence, while discontinuous regions consist of residues from different parts of the sequence, brought together by the folding of the protein to its native structure.

More specifically, a protein sequence was encoded as the combination of 4-bit binary digits (0 or 1), which means we need fifteen different combinations (0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111). It should be noticed that here 0 or 1 denote one of the four equal length region A-D is excluded or included in constructing the continuous or discontinuous regions respectively. For example, 0011 denotes a continuous region constructed by C and D (the final 50% of the sequence). Similarly, 1011 represents a discontinuous region constructed by A, C and D (the first 25% and the final 50% of the sequence). These regions are illustrated in Figure 1. For each region the 63 local descriptors are extracted, resulting in a 63*15=945 feature vector. Then the PPI pair is characterized by concatenating the two vector spaces of two individual proteins. Thus, an 1890-dimentional vector has been constructed to represent each protein pair and used as a feature vector for input into SVM classifier.

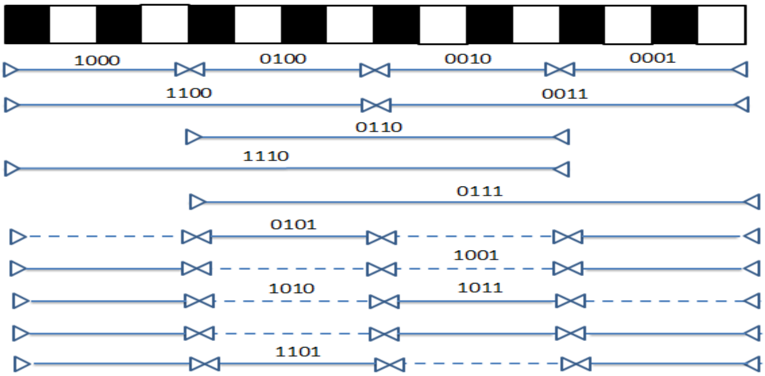


Fig. 1. The Schematic diagram for constructing fifteen descriptor regions for a hypothetical protein sequence

2.3 Support Vector Machine

Support Vector Machine (SVM) is a classification and regression paradigm first developed by Vapnik [19]. It has attracted much research attention in these years due to its improved generalization performance over other techniques in many real world applications including bioinformatics. The SVM originated from the idea of the structural risk minimization theory [19]. The main difference between this technique and many other conventional classification techniques including neural networks is that it minimizes the structural risk instead of the empirical risk. The principle is based on the fact that minimizing an upper bound on the generalization error rather than minimizing the training error is expected to perform better. SVM training always seeks a global optimized solution and avoids over-fitting, so it has the ability to deal with a large number of features. A complete description to the theory of SVMs for pattern recognition is in Vapnik's book [20].

The basic idea of utilizing SVM model for classification can be stated briefly as follows. Firstly, map the original data X into a feature space F with high dimensionality through a linear or non-linear mapping function, which is relevant with the selection of the kernel function. Then, within the feature space from the first step, seek an optimized linear division, i.e. construct a hyperplane which separates the data into two classes.

Given a training dataset of instance-label pairs $\{x_i, y_i\}, i = 1,2,....,N$ with input data $x_i \in R^n$ and labeled output data $y_i \in \{+1,-1\}$. The classification decision function implemented by SVM is represented in the following equation:

$$y(x) = sign\left[\sum_{i=1}^{N} y_i \alpha_i \cdot K(x, x_i) + b\right] \tag{1}$$

where the coefficients α_i are obtained by soving the following convex Quadratic Programming (QP) problem:

$$\text{Maximize} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K(x_i, x_j) \tag{2}$$

$$\text{Subject to} \quad 0 \le \alpha_i \le C \tag{3}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \; i = 1,2,...,N. \tag{4}$$

In the equation (3), C is a regularization parameter which controls the tradeoff between margin and misclassification error. These x_j are called Support Vectors only if the corresponding $\alpha_j > 0$.

In this work, Radial Basis Functions (RBF) kernel, $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, is applied, which has better boundary response and most high-dimensional data sets can be approximated by Gaussian like distributions. In the experiment the well-known software LIBSVM (http://www.csie.ntu.edu.te/~cjlin/libsvm) was employed to do classification.

3 Experiments and Results

3.1 Evaluation Measures

To evaluate the prediction performance of the proposed method, Sensitivity (Sens), Precision (PE), Matthews's correlation coefficient (MCC), and overall accuracy (Accu.) were calculated. The definitions of these measures are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$PE = \frac{TP}{TP + FP} \tag{7}$$

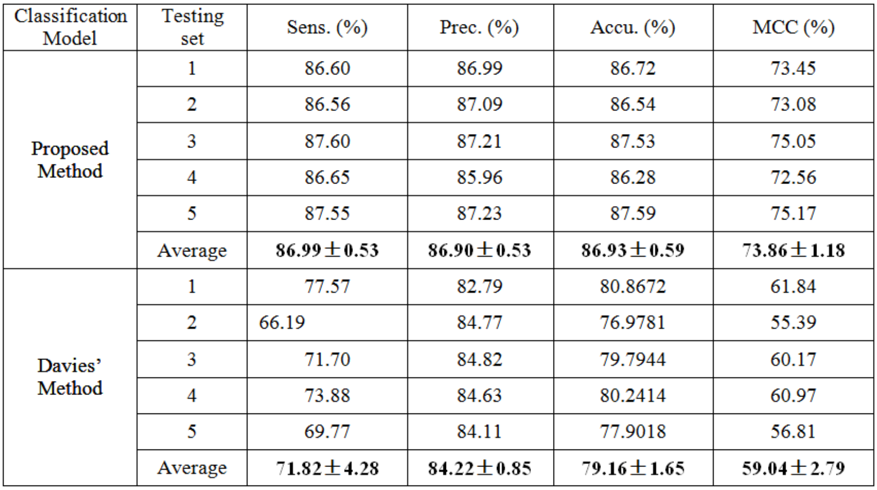
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{8}$$

where true positive (TP) is the number of true PPIs that are predicted correctly; false negative (FN) is the number of true PPIs that are predicted to be non-interacting pairs; false positive (FP) is the number of true non-interacting pairs that are predicted to be PPIs, and true negative (TN) is the number of true non-interacting pairs that are predicted correctly. MCC denotes Mathew's correlation coefficient.

3.2 Prediction Performance of Proposed Model

We evaluated the performance of the proposed approach using the DIP PPIs data as investigated in Guo *et al*. [7]. In order to reduce the bias of training and testing data, a 5-fold cross-validation technique is adopted. More specifically, the dataset is divided into 5 subsets, and the holdout method is reiterated 5 times. Each time four of the five subsets are put together as the training dataset, and the other one subset is utilized for testing the model. Thus five models were generated for the five sets of data. The prediction results of SVM prediction models with proposed representation of protein sequence are shown in Table 1.

Table 1. The prediction result of the test dataset using proposed method



Classification Model	Testing set	Sens. (%)	Prec. (%)	Accu. (%)	MCC (%)
Proposed Method	1	86.60	86.99	86.72	73.45
	2	86.56	87.09	86.54	73.08
	3	87.60	87.21	87.53	75.05
	4	86.65	85.96	86.28	72.56
	5	87.55	87.23	87.59	75.17
	Average	**86.99±0.53**	**86.90±0.53**	**86.93±0.59**	**73.86±1.18**
Davies' Method	1	77.57	82.79	80.8672	61.84
	2	66.19	84.77	76.9781	55.39
	3	71.70	84.82	79.7944	60.17
	4	73.88	84.63	80.2414	60.97
	5	69.77	84.11	77.9018	56.81
	Average	**71.82±4.28**	**84.22±0.85**	**79.16±1.65**	**59.04±2.79**

It can be observed from Table 1 that for all five models the precisions are$\geqslant$85.96%, the sensitivities are$\geqslant$86.56%, and the prediction accuracies are $\geqslant$86.28%. On average, proposed method yields a PPI prediction model with an accuracy of 86.93$\pm$0.59%. To better investigate the practical prediction performance of proposed method, we also calculated the MCC value. From table 1, we can see that proposed method gives good prediction performance with an average MCC value of 73.86%. Further, it can also be seen in the experiments that the standard deviation of sensitivity, precision, accuracy and MCC are as low as 0.53%, 0.53%, 0.59% and 1.18% respectively. From the results, it can be concluded that proposed method is an accurate and robust method for the prediction of PPIs.

Many other sequence-based methods have been used for predicting of PPIs. In order to evaluate the prediction ability of the SVM prediction model using binary coding, extensive experiments are performed to compare our method with state-of-the-art techniques Davies' work [21]. Table 1 gives the average prediction results of 5-fold cross-validation over there two methods. From Table 1, we can see that the model based on Davies' work gives poor results with the average sensitivity, precision and accuracy of 71.82%, 84.22% and 79.16%, respectively. The results illustrate that our method outperforms other sequence-based methods such as Davies' method. All the analysis shows that our model is an accurate and fast method for the prediction of PPIs.

4 Discussion and Conclusions

With the large amount of protein sequences information provided by genome sequencing project, there is a growing demand for developing advanced computational methods for predicting potential PPIs using sequence information alone. In this study, we proposed a novel sequence-based approach for PPIs prediction using SVM combined with a binary-coding-based method. The binary-coding-based method was implemented to extract sequence information of proteins, and then an SVM algorithm was employed to construct the prediction model. The proposed representation of protein sequence descriptor account for the interactions between residues in both continuous and discontinuous regions of a protein sequence, so this method enables us to draw more PPI information from the protein sequence. When performed on the PPI data of S.cerevisiae, the method achieved 86.93% prediction accuracy with 86.99% sensitivity at the precision of 86.90%. Given the complex nature of PPIs, the performance of our method is promising and it can be a helpful supplementary for PPIs prediction.

Acknowledgments. This work is supported in part by the National Science Foundation of China, under Grants 61102119, 61170077, 71001072, 61272339, 61272333, 61001185, in part by the China Postdoctoral Science Foundation, under Grant 2012M520929.

References

1. Gavin, A.C., Bosche, M., Krause, R., Grandi, P.: 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. Proceedings of the National Academy of Sciences of the United States of America 98(8), 4569–4574 (2001)
3. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415(6868), 180–183 (2002)
4. Krogan, N.J., Cagney, G., Yu, H.Y., Zhong, G.Q.: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440(7084), 637–643 (2006)
5. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403(6770), 623–627 (2000)
6. Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y.: A protein interaction map of Drosophila melanogaster. Science 302(5651), 1727–1736 (2003)
7. 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 Research 36(9), 3025–3030 (2008)
8. 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 Bioinformatics 11 (2010)
9. 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), 2744–2751 (2010)
10. Xia, J.F., You, Z.H., Wu, M., Wang, S.L., Zhao, X.M.: Improved method for predicting pi-turns in proteins using a two-stage classifier. Protein and Peptide Letters 17(9), 1117–1122 (2010)
11. 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 Bioinformatics 13 (2012)
12. You, Z.-H., Li, L., Yu, H., Chen, S., Wang, S.-L.: Increasing reliability of protein interactome by combining heterogeneous data sources with weighted network topological metrics. In: Huang, D.-S., Zhao, Z., Bevilacqua, V., Figueroa, J.C. (eds.) ICIC 2010. LNCS, vol. 6215, pp. 657–663. Springer, Heidelberg (2010)
13. Qi, Y.J., Seetharaman, J.K., Joseph, Z.B.: Random forest similarity for protein-protein interaction prediction from multiple sources. In: Pac. Symp. Biocomput., pp. 531–542 (2005)
14. Yang, L., Xia, J.F., Gui, J.: Prediction of Protein-Protein Interactions from protein sequence using local descriptors. Protein and Peptide Letters 17(9), 1085–1090 (2010)
15. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., Jiang, H.: Predictina protein-protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America 104(11), 4337–4341 (2007)
16. Shi, M.G., Xia, J.F., Li, X.L., Huang, D.S.: Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset. Amino Acids 38(3), 891–899 (2010)

17. Xia, J.F., Han, K., Huang, D.S.: Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. Protein and Peptide Letters 17(1), 137–145 (2010)
18. Tong, J.C., Tammi, M.T.: Prediction of protein allergenicity using local description of amino acid sequence. Frontiers in Bioscience 13, 6072–6078 (2008)
19. Herrera, L.J.: Recursive prediction for long term time series forecasting using advanced models. Neurocomputing 70(16), 2870–2880 (2007)
20. Cortes, C., Vapnik, V.: Support vector network. Machine Learning (1995)
21. Davies, M.N., Secker, A., Freitas, A.A., Clark, E., Timmis, J., Flower, D.R.: Optimizing amino acid groupings for GPCR classification. Bioinformatics 24(18), 1980–1986 (2008)