Prediction of Protein-Protein Interactions Using Local Description of Amino Acid Sequence

Yu Zhen Zhou[1], Yun Gao[1], and Ying Ying Zheng[2]

[1] Computer Engineering Dept., JiangSu College of Information Technology,
Wuxi 214181, China
[2] Institute of Intelligent Machines, Chinese Academy of Sciences,Hefei 230031, China
zyz9508@sina.com, xiaoxiao@000416@yahoo.com.cn,
yingzheng1982@hotmail.com

Abstract. Protein-protein interactions (PPIs) are essential to most biological processes. Although high-throughput technologies have generated a large amount of PPI data for a variety of organisms, the interactome is still far from complete. So many computational methods based on machine learning have already been widely used in the prediction of PPIs. However, a major drawback of most existing methods is that they need the prior information of the protein pairs such as protein homology information. In this paper, we present an approach for PPI prediction using only the information of protein sequence. This approach is developed by combing a novel representation of local protein sequence descriptors and support vector machine (SVM). Local descriptors account for the interactions between sequentially distant but spatially close amino acid residues, so this method can adequately capture multiple overlapping continuous and discontinuous binding patterns within a protein sequence.

Keywords: Protein-protein interactions; Protein sequence; Local descriptors; SVM.

1 Introduction

Protein-protein interactions (PPIs) play important roles in most cellular processes, such as transcription regulation, signal transduction [1], and recognition of foreign molecules. Knowledge of PPIs can provide insight into protein functions [2, 3], lead to a better understanding of disease mechanisms and suggest novel methods for designing drugs that modulate specific disease pathways. In recent years, high throughput technologies have been developed for the large-scale PPI analysis, such as yeast two-hybrid screening methods [4], immunoprecipitation [5], and protein chips [6]. However, there are some disadvantages of existing experimental methods, such as time-intensive, high cost and a small fraction of the complete PPI network covered. In addition, these approaches suffer from high rates of both false negative and false positive predictions. Therefore, there is a strong motivation to develop reliable computational methods for inferring protein interactions [7], which provide an attracting perspective on predicting and understanding PPIs as complementary methods to experimental ones.

M. Zhou and H. Tan (Eds.): CSE 2011, Part II, CCIS 202, pp. 254–262, 2011.

A number of computational methods [7] have been developed for the prediction of PPIs based on various data types, including genomic information, protein domain and protein structure information. However, these methods are not universal, because the accuracy and reliability of these methods depend on the prior information of the protein pairs such as the information of protein homology [8, 9]. Moreover, compared to the rapid increase of the number of protein sequences, the protein three-dimensional structure data is scarce. So approaches that derive information directly from amino acid sequence information are of particular interest [8-14]. Many groups have engaged in the development of sequence-based method for predicting PPIs, and the preliminary results have demonstrated their feasibility. Specifically, Bock and Gough [10] tried to solve this problem by using a support vector machine (SVM) with several structural and physiochemical descriptors. Martin et al. [11] used a descriptor called signature product, which is a product of subsequences and an expansion of the signature descriptor from chemical information to predict PPIs. Nanni and Lumini [14] proposed a method to predict PPIs based on an ensemble of K-local hyperplane distance nearest neighbor classifiers, where each classifier is trained using a different physicochemical property of the amino acids. Shen et al. [8] developed a SVM model by combining a conjoint triad feature with S-kernel function of protein pairs to predict PPI network and yielded a high prediction accuracy of 83.93%. Guo et al. [9] proposed a sequence-based method by combining auto covariance descriptor with SVM, and when applied to predicting yeast PPIs, it achieved very promising prediction accuracy. In our previous study, we also obtained good prediction results by using correlation coefficient [15] and autocorrelation descriptor [16], respectively.

In this study, we present a sequence-based approach for the prediction of interacting protein pairs using support vector machine (SVM) combined with local descriptors [17, 18]. The utilization of the local descriptors provides us with a chance to mine interaction information from the continuous and discontinuous amino acids segments at the same time [17]. The effectiveness of local descriptors depends largely on the correct selection of amino acid grouping [18]. By grouping amino aids 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 [8]. To evaluate the performance, the proposed method was applied to *Saccharomyces cerevisiae* and *Helicobacter pylori datasets.* Empirical results have shown that our SVM prediction model with local descriptors yields good performance. We also evaluated the performance of our method by preparing four cross-species data as the independent test set, which further demonstrates the effectiveness of our method.

2 Methods

2.1 Data Set

We evaluated our method on publicly available S.cerevisiae dataset, which were extracted from S.cerevisiae core subset of database of interacting proteins (DIP) [19] by Guo et al [9]. After the protein pairs which contain a protein with fewer than 50 residues or have ≥40% sequence identity were removed, the remaining 5594 protein pairs comprise the final positive data set. The non-interacting pairs were generated from pairs of proteins whose subcellular localizations are different. The final data set consists of 11188 protein pairs, where half are from the positive data set and half from the negative data set. Three-fifths of the protein pairs which from the positive and negative data set were respectively randomly chosen as the training set, and the remaining two-fifths were used as the test set.

2.2 Local Protein Sequence Descriptors

To predict PPIs from sequences, one of the main computational challenges is to find the way to fully encode the important information content of proteins [11, 14]. In this study, each protein sequence is represented by local description of amino acid sequence [17, 18], and the PPI pair is characterized by concatenating the local descriptors of two proteins in this protein pair. To reduce the complexity inherent in the representation of the twenty standard amino acids and suit synonymous mutation, the amino acids were clustered into seven functional groups [8] based on the dipoles and volumes of the side chains (Table 1). Then the local protein descriptors abstract the features of protein pair based on the classification of amino acids. The process of generating local descriptors is described as follows.

Table 1. Division of amino acids based on the dipoles and volumes of the side chains

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group7
A, G, V	C	D, E	F, I, L, P	H, N, Q, W	K, R	M, S, T, Y

Firstly, for each protein sequence, every amino acid is replaced by the index depending on its grouping. For example, protein sequence AVDCNLSK is replaced by 11325476 based on this classification of amino acids. Secondly, we split the amino

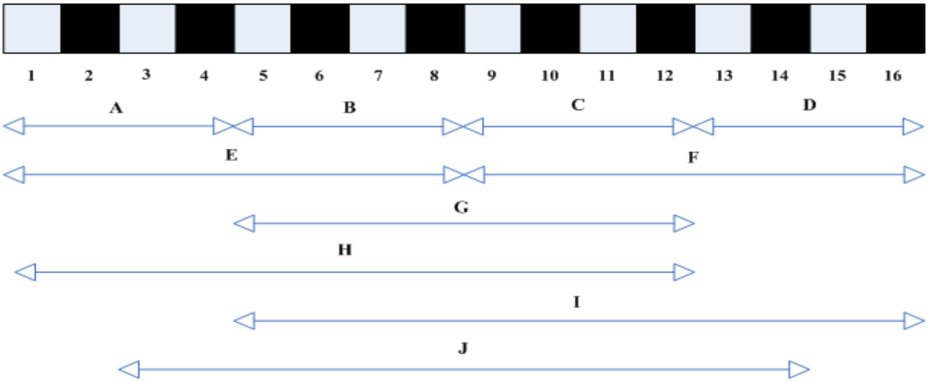

Fig. 1. Schematic diagram for constructing ten descriptor regions (A-J) for a hypothetical protein sequence. Adapted from Tong et al. [17] and Davies et al. [18]. The regions (A–D) and (E–F) are respectively generated by dividing the whole sequence into four equal regions and two equal regions. The region G, H, I and J stand for the central 50%, the first 75%, the final 75% and the central 75% of the entire sequence, respectively.

acid sequences into ten local regions of varying length and composition to describe multiple overlapping continuous and discontinuous interaction patterns within a protein sequence (see Figure 1). For each local region, three local descriptors, composition (C), transition (T) and distribution (D), are calculated. C stands for the composition of each amino acid group along a local region. T represents the percentage frequency with which amino acid in one group is followed by amino acid in another group. D characterizes the distribution pattern along the entire region by measuring the location of the first, 25, 50, 75 and 100% of residues of a given group.

For detailed descriptions of these descriptors, please refer to [17, 18]. Given that the amino acids are divided into seven groups in this instance, the calculation of the C, T and D descriptors generates 63 attributes in each local region (7 for C, 21 for T and 35 for D). The descriptors for all local regions were combined, resulting in 630 features representing the general characteristics of the protein sequence. Thus, a 1260-dimensional vector has been built to represent each protein pair and used as a feature vector for input into SVM.

2.3 SVM Optimization and Evaluation of Performance

The classification model for predicting PPIs was based on SVM. As a binary classification algorithm, SVM separates a given set of binary labelled training data (-1 and +1, in our case, non-binding and binding protein pairs) with a hyper-plane that is maximally distant from them (known as the maximal margin hyper-plane). The hyper-plane found by the SVM in feature space corresponds to a nonlinear decision plane in the input space. Each of the feature vector generated from the protein pair in the negative and positive dataset is assigned with a corresponding label of {-1} and {+1} respectively, indicating whether the pair is interacting with each other or not. The advantage of SVM is that there is no need to compute the coordinates of the data in the feature space, but instead simply computing the inner products between all pairs of data. This operation is often computationally cheaper than the explicit computation of the coordinates.

The LIBSVM package (http://www.csie.ntu.edu.te/~cjlin/libsvm) was employed in this work to do classification. A radial basis function (RBF) was selected as the kernel function. Two parameters, the regularization parameter C and the kernel parameters γ were optimized using a grid search approach. The prediction performance was evaluated by the overall prediction accuracy (ACC), sensitivity (SN), precision (PE) and Matthews correlation coefficient (MCC) [20]:

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}, \tag{1}$$

$$\mathrm{SN}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{2}$$

$$\mathrm{PE}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{3}$$

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

where TP, TN, FP and FN denote true positive, true negative, false positive and false negative, respectively. In addition, we also used the receiver operating characteristic (ROC) curve [21] to assess the prediction performance. An ROC curve is a graphical plot of the true positive rate (TPR) versus the false positive rate (FPR) for a binary classifier system as its discrimination threshold is varied. To summarize ROC curve in a single quantity, the area under an ROC curve (AUC) is used. The AUC score ranges from 0 to 1. When the AUC value of a predictor is larger than the area of other ROC curves, the predictor is regarded as a better one.

3 Results and Discussion

3.1 Assessment of Prediction Ability

In order to achieve good experimental results, the corresponding parameters for SVM were firstly optimized. Here, two parameters, C and γ were optimized using a grid search method within a limited range. Considering the numerous samples used in this work, 5-fold cross-validation was used to investigate the training set, which can minimize the overfitting of the prediction model. To test the robustness of the prediction model, five training sets and five test sets were prepared as described by the sampling method in Methods. Thus five models were generated for the five sets of data. The prediction results of SVM prediction models with local description of protein sequence are shown in Table 2. For all five models, the precisions are ≥88.66%, the sensitivities are ≥87.00%, and the prediction accuracies are ≥88.07%. On average, our method yields a PPI prediction model with an accuracy of 88.56±0.33%. To better investigate the practical prediction ability of our model, we also calculated the MCC and AUC values. From table 2, we can see that our method gives good prediction performance with an average AUC score of 95.07% and a MCC value of 77.15%. Further, it can also be seen in the experiments that the standard deviation of sensitivity, precision, accuracy, MCC and AUC are as low as 0.22, 0.60, 0.33, 0.68 and 0.39% respectively. The results illustrate that our model is an accurate and robust method for the prediction of PPIs.

Table 2. Prediction results of the test sets

Test set	SN (%)	PE (%)	ACC (%)	MCC (%)	AUC (%)
1	87.62	89.71	88.78	77.59	95.01
2	87.31	88.66	88.07	76.15	95.16
3	87.00	89.64	88.47	76.98	94.67
4	87.35	90.43	89.05	78.15	95.69
5	87.58	89.09	88.43	76.87	94.82
Average	87.37±0.22	89.50±0.60	88.56±0.33	77.15±0.68	95.07±0.39

There are two possible reasons that our SVM prediction model with local descriptors yields good performance. One is that the twenty standard amino acids have been divided into seven groups according to their related physicochemical properties of electrostatic and hydrophobic interactions[8]. Accordingly, the reduced dimension of vector space of protein sequence may partially overcome the overfitting problem [18]. In addition, the reduced but informative alphabet likely includes the information of synonymous mutations [8] for PPI because of similar characteristics within the same amino acid group. The other is that we used a series of local descriptors of varying length and composition to describe the physicochemical properties of proteins. Local descriptors account for the interactions between sequentially distant but spatially close amino acid residues [17]. As a result, such novel representation of local description of amino acid sequence enables our model to adequately capture multiple overlapping continuous and discontinuous binding patterns within a protein sequence.

3.2 Performance on Independent Dataset

As our method produced a good performance on the PPI data of S.cerevisia, we switched to evaluate the practical prediction ability of our final model against an independent dataset. Firstly, we constructed our final prediction model using the whole dataset (11188 protein pairs) with the optimal parameters (C = 32, γ = 0.03125). And then the prediction performance of the final predictor was evaluated using another dataset which is independent of the training dataset. Our model was trained on the S.cerevisia core subset in the DIP database; therefore we chose the other four species in this database as our independent test dataset. The performance of our method in predicting such samples is summarized in Table 3. The prediction performance in Caenorhabditis elegans, Escherichia coli, Homo sapiens, and Mus musculus achieved by our method is 75.73%, 71.24%, 76.27% and 76.68% respectively. It shows that the meta model can correctly predict the interacting pairs of three species with the accuracy of over 75% while the E. coli subset have a relatively lower accuracy which still >71%. It demonstrates that the SVM prediction model with local descriptors is able to achieve better performance towardscross-species dataset. We selected the PPIs data of S.cerevisiae to construct the final prediction model, so this model should represent the features of S.cerevisiae PPIs. At the same time, our model can also represent the features of C. elegans, E. coli, H. sapiens, and M. musculus, which is implied by the generalization ability of our model on these four species. Our findings indicate that our model may be applied to other organisms for which experimental data regarding PPIs may not be available.

Table 3. Prediction results on four species based on our model

Species	Test pairs	ACC (%)
C. elegans	4013	75.73
E. coli	6954	71.24
H. sapiens	1412	76.27
M. musculus	313	76.68

Interestingly, we found that there are some relationship between the prediction accuracies and the evolution of organisms. For example, S.cerevisiae and E. coli protein-protein interactions are not very closely related. That is to say, many proteins in the E. coli dataset were not presented in the S.cerevisiae dataset, and vice versa. As a result, when predicting E. coli from yeast, we had only limited success, as can be seen by the relative poor results reported in Table 3. On the other hand, if there is a close relation between species such as the yeast and M. musculus, our method can generate very promising results (see Table 3). These results agree well with the findings of Martin et al. [11]. At the same time, it should be pointed out that many PPIs in these four organisms dataset were not obtained by high-throughput proteome-wide methods and were small (for example, there were 1412 H. sapiens and 313 M. musculus interactions in the dataset). So it can be expected that with the increasing number of PPIs data, our model may not work as well.

3.3 Comparison with Other Methods

Many methods have been used in the prediction of PPIs. To compare prediction ability of the SVM prediction model using local descriptors with the existing methods, dataset H.pylori was constructed. The H.pylori dataset is comprised of 2916 protein pairs (1458 interacting pair and 1458 non-interacting pairs) as described by Maritin et al.[11]. Table 4 gives the average prediction results of 10-foldcross-validation over six different methods [10-14] on the H.pylori dataset. The methods of Bock and Gough ([10]), Martin et al. ([11]) and Nanni ([13]) are based on single classifer system to infer PPIs, while the methods of Nanni ([12]), Nanni and Lumini([14]) belong to ensemble classifier-based approach. From Table 4, we can see that the model based on SVM with local description of amino acid sequence gives good results with the average sensitivity, precision and accuracy of 0.851, 0.833 and 0.842, respectively. The results illustrate that our method outperforms other single classifier-based methods such as signature product method. It has pointed out that one single classification system cannot always provide high classification accuracy [22]. Instead, a multiple classifier system is proved to be more accurate and robust than an excellent single classifier [22]. However, it is remarkable that our prediction model obtain performance similar to those obtained by ensemble classifier-based methods. All these results demonstrate that the SVM classifier combined with local descriptors can improve the prediction accuracy compared with current state-of-the-art methods.

Table 4. Comparison of state-of-the-art methods on the *H.pylori* dataset

Mehods	SN	PE	ACC
Bock and Gough ([10])	0.698	0.802	0.758
Martin et al. ([11])	0.799	0.857	0.834
Nanni ([12])	0.806	0.851	0.83
Nanni ([13])	0.86	0.84	0.84
Nanni and Lumini([14])	0.867	0.85	0.866
Our method	0.851	0.833	0.842

4 Conclusions

In this paper, we present a simple and elegant sequenced-based approach to solve protein interaction problem. One particular feature of protein interaction is that the interactions usually occur in the discontinuous regions in the protein sequence, where distant residues are brought into spatial proximity by protein folding. However, litter literature tries to make use of such information. In the current study, a novel representation of local protein sequence descriptors was used to involve the information of interactions between distant amino acids in the sequence. A protein sequence was characterized by 10 local descriptors of varying length and composition. So this method is capable of capturing multiple overlapping continuous and discontinuous binding patterns within a protein sequence. As expected, experimental results show that our SVM-based predictive model with this encoding schemeis an important complementary method for PPI prediction.

We believe that the results can be further improved in the ways explained below. For example, the performance of local descriptors should be better enhanced if we optimize the amino acid grouping, i.e. we could extract more useful information by identifying the most efficient grouping. In addition, not all local descriptors are effective in the prediction. Some are less relevant to the prediction and some are redundant. We expect to improve the prediction accuracy by using some feature selection strategy, instead of using the whole local description features. Finally, it is evident that one single classification system cannot always provide high classification accuracy. Instead, a multiple classifier system is proved to be more accurate and robust than an excellent single classifier in many fields. Hence, using a multiple classifier learning approach could further improve the prediction accuracy.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. This research was supported by the grants of the National Science Foundation of China (Nos. 60472111 & 30570368).

References

1. Zhao, X., Wang, R., Chen, L.: Uncovering signal transduction networks from high-throughput data by integer linear programming. Nucleic Acids Research (36), 48 (2008)
2. Zhao, X., Wang, R., Chen, L., et al.: Gene function prediction using labelled and unlabeled data. BMC Bioinformatics 957 (2008)
3. Zhao, X., Wang, R., Chen, L., et al.: Protein function prediction with high-throughput data. Amino Acids 35, 517–530 (2008)
4. Ito, T., Chiba, T., Ozawa, R., et al.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA 98, 4569–4574 (2001)
5. Ho, Y., Gruhler, A., Heilbut, A., et al.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415, 180–183 (2002)
6. Zhu, H., Bilgin, M., Bangham, R., Hall, D., et al.: Global analysis of protein activities using proteome chips. Science 293, 2101–2105 (2001)
7. Skrabanek, L., Saini, H., Bader, G., Enright, A.: Computational prediction of protein–protein interactions. Molecular Biotechnology 38, 1–17 (2008)

8. Shen, J., Zhang, J., Luo, X., et al.: Predicting protein-protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA 104, 4337–4341 (2007)
9. 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, 3025–3030 (2008)
10. Bock, J.R., Gough, D.A.: Whole-proteome interaction mining. Bioinformatics 19, 125–134 (2003)
11. Martin, S., Roe, D., Faulon, J.L.: Predicting protein-protein interactions using signature products. Bioinformatics 21, 218–226 (2005)
12. Nanni, L.: Fusion of classifiers for predicting protein-protein interactions. Neurocomputing 68, 289–296 (2005)
13. Nanni, L.: Hyperplanes for predicting protein-protein interactions. Neurocomputing 69, 257–263 (2005)
14. Nanni, L., Lumini, A.: An ensemble of K-local hyperplanes for predicting protein-protein interactions. Bioinformatics 22, 1207–1210 (2006)
15. Shi, M., Xia, J., Li, X., Huang, D.: Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset. Amino Acids (2009), doi:10.1007/s00726-009-0295-y
16. Xia, J., Liu, K.H., Huang, D.: Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. Protein and Peptide Letters (2009) (in press)
17. Tong, J., Tammi, M.: Prediction of protein allergenicity using local description of amino acid sequence. Frontiers in Bioscience: A Journal and Virtual Library 13, 6072 (2008)
18. Davies, M., Secker, A., Freitas, A., Clark, E.: Optimizing amino acid groupings for GPCR classification. Bioinformatics 24, 1980–1986 (2008)
19. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D.: The Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 30, 303–305 (2002)
20. Matthews, B.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451 (1975)
21. Zweig, M., Campbell, G.: Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. 39, 561–577 (1993)
22. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley, New York (2004)