The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009
Copyright © 2009 ORSC & APORC, pp. 198–206

Protein-Protein Interaction Detection By SVM From Sequence Information

Hong-Wei Liu[1,2,*]

[1]School of Information, Beijing Wuzi University, Beijing 101149, China
[2]Institute of Applied Mathematics Academy of
Mathematics and Systems Science, CAS, Beijing 100080, China

Abstract Proteins frequently bind together in pairs or larger complexes to take part in biological processes. Understanding such protein functions and biological processes in a cell across the entire genome is an important goal with diverse implications about protein function. In this paper, we propose a method to detect protein-protein interaction (PPI) based on sequence neighboring information and support vector machine (SVM). When applied on the currently available protein-protein interaction data for the yeast *Saccharomyces cerevisiae*, it yields a predictive accuracy of 87.98%. It is further evaluated on an independent PPIs with the test accuracy of 79.05%, which delivered the proposed method reasonable and promising.

Keywords Saccharomyces cerevisiae; protein-protein interactions (PPIs); support vector machine (SVM); feature extraction

1 Introduction

Detecting protein-protein interactions (PPIs) is a central problem in computational biology and aberrant such interactions may have implicated in a number of neurological disorders. As a result, the prediction of protein-protein interactions has recently received considerable attention from biologist around the globe. So many computational methods have been developed to facilitate the identification of novel PPIs. In recent years, many experimental techniques have been proposed for identifying the interaction of protein pairs. Most of these techniques that use protein properties for their ability to interact such as protein sequences [26], primary structures of proteins for this prediction [12] have therefore attracted considerable interest. Because experimental methods are time-consuming and expensive, current PPI pairs obtained from experiments only cover a small fraction of the complete PPI networks [10]. Hence, it is of great practical significance to develop the reliable computational methods to facilitate the identification of PPIs.

Most of the recent works focus on employing the protein domain knowledge to predict the protein-protein interaction[3, 9, 13, 17, 18, 24]. However, none of them consider all the sequence information to predict the protein-protein interaction. We understand that protein domains are highly informative for predicting protein-protein interaction as it reflects the potential structural relationships between proteins, however, other sequence

*E-mail address: ryuhowell@163.com.

parts (not currying any domain knowledge) may contribute to the information by showing how different two proteins are.

This paper presents an approach to pairwise protein interaction based on physico-chemical properties of amino acids aimed at addressing the ability of protein-protein interaction prediction. The interactions usually occur in the discontinuous amino acids segments in the sequence, and the information of these interactions may be able to further improve the prediction ability of the existing sequence-based methods. So the proposed method takes neighboring effect into account and makes it possible to discover patterns that run through entire sequences. The amino acid residues were translated into numerical values and then these numerical sequences were analyzed by SVM.

2 Materials and Method

2.1 Data set construction

The yeast organism is chosen primarily because there is more information about yeast protein interactions than about any other organism. For the interacting pair, it is simply obtained from the Database of Interacting Protein (DIP) [25]. The PPI dataset of budding yeast (*Saccharomyces cerevisiae*) is retrieved from DIP database in February 2007. The reliability of this subset has been tested by expression profile reliability and paralogous verification method. After removed the protein pairs that contained a protein less than 60 amino acids, the collected subset contained 5926 interactions pairs.

Since obtaining identified and standard non-interacting proteins pairs remains to be the concern of all researchers working in predicting protein-protein interaction, three mainly strategies for constructing negative data set are used in order to compare the effects of different training data sets on the performance. The first strategy is that the non-interacting pairs are generated by randomly pairing proteins appeared in the positive data set [20]. The second is based on such an assumption that proteins occupying different subcellular localizations do not interact. And the third strategy is used for creating non-interacting pairs composed of artificial protein sequences. Therefore, in our case we use a random method to generate proteins pairs, and then delete all pairs that appear in DIP. Consequently, we get the data set contained 11852 pairs, 5926 pairs in positive set and 5926 pairs in negative set. This is acceptable for the purposes of comparing the feature representation since the resulting inaccuracy will be approximately uniform with respect to each feature representation [1].

2.2 Support vector machine(SVM)

To discriminate between interacting and non interacting protein pairs, we employed SVM. SVM [7],[23] is a powerful classification algorithm and well suited the given task. It addresses the general problem of learning to discriminate between positive and negative members of a given class of n-dimensional vectors.The algorithm operates by mapping the given training set into a possibly high-dimensional feature space and attempting to learn a separating hyperplane between the positive and the negative examples for possible maximization of the margin between them [27]. The margin corresponds to the distance between the points residing on the two edges of the hyperplane. Having found such a plane, the SVM can then predict the classification of an unlabeled example. In fact, much of the SVM's power comes from its criterion for selecting a separating plane when many

candidate planes exist: the SVM chooses the plane that maintains a maximum margin from any point in the training set [16]. SVM classifiers do not require any complex parameters to be tuned and optimized, and they exhibit a great ability to generalize even when given a small number of training examples. The only significant parameters to be tuned are the choice of the kernel function and the soft-margin parameter (capacity or regularization parameter). The kernel projects the data to higher dimensional space to increase the computational ability.

To describe an SVM precisely, suppose the data are given as pairs $\{(x_i, y_i)\} \subset R^n \times \{\pm 1\}$, and the classifiers created by SVM algorithm are sequence patterns that can only give binary answers. In other words, given a sequence, each pattern answers either 'yes' (1) or 'no' (-1), as to whether the pattern matches parts of the sequence or not. Using this notation an SVM assumes the form $f(x) = \sum_i \alpha_i y_i k(x_i, x) + b$, where $f : R^n \to R$ is a decision function (x belongs to class 1 if $f(x)$ is greater than some threshold t, or to class -1 otherwise), $k : R^n \times R^n \to R$ is a kernel function, otherwise known as a dot product in some vector space, and the constants b and α_i are obtained by solving a quadratic programming problem (see [4] for details). The threshold t is typically 0, although it may be varied to obtain classifiers that are more or less accurate on positive predictions.

2.3 Feature extraction and measure

One of the main challenges in using SVMs for the prediction of PPIs in genome sequence is a suitable encoding of the genome sequences information in some vector space and requires a fixed number of inputs for training. However, there are often unequal length vectors because of protein sequences with different lengths. So a transformation is proposed that converts protein sequence into fixed-dimensional representative feature vectors, where each feature records the correlation of amino acids to the protein sequences of interest. These features are then used in conjunction with SVM to predict the possible interactions between proteins.

According to the above consideration, here physicochemical properties of amino acids were collected to reflected the interaction whenever possible and they are hydrophobicity [22], hydrophicility [11], volunes of side chains of amino acids [14], polarity [8], polarizability [5], solvent-accessible surface area [19] and net charge index of side chains of amino acids [28], respectively. The value of the seven physicochemical descriptors for each amino acid were normalized to zero mean and unit standard deviation.

In this paper, given a protein sequence P, a score describe the related interactions between residues. After translated each protein sequence into seven vectors with each amino acid by the physicochemical properties, deviations of the vectors X were computed. Then relation between residues with fixed distance d were calculated by the Equation (1) throughout the whole protein sequence, where j represents one descriptor, i the position in the sequence P, n the length of the sequence P. Here, d is the distance between one residue and its neighbor, a certain number of residues away, L requires to be optimally chosen.

$$\gamma_{dj} \propto \frac{1}{n-d} \sum_{i=1}^{n-d} x_{ij} * x_{(d+i),j}, \quad j = 1,2,\cdots,7, d = 1,2,\cdots,L \tag{1}$$

After each protein sequence is represented as a vector $V = (d_1, d_2, \cdots, d_{7L})$ by vectorizing γ_{dj} along with rows, there are two approaches to construct the feature vectors to

represent protein-protein pairs:

2.3.1 Concatenating the score vector

A protein pair (A,B) is represented by concatenating the score vectors V_A and V_B. That is the input feature vector C_{AB} for a protein pair (A,B) is calculated as follows:

$$C_{AB} = V_A \oplus V_B,$$

where $\oplus$ is the concatenation operator. We can obtain an additional improvement in the concatenating vector by enforcing symmetry in the protein-protein order. In other words, we can make a protein pair (A,B) equivalent to protein pair (B,A). This symmetry is easily achieved by training and testing on both C_{AB} and C_{BA}, and reporting the average predicted results in numerical experiments.

2.3.2 Distance on protein pairs

A protein pair (A,B) is represented by a distance vector. $d_k^{AB} = (d_k^A - d_k^B)^2$ is used to measure the distance between protein pair (A,B) with respect to d_k^A and d_k^B. So the input feature vector D_{AB} for a protein pair (A,B) is calculated as follows:

$$D_{AB} = (d_1^{AB}, d_2^{AB}, \cdots, d_{7L}^{AB})^T.$$

According to concatenating and distance operations, if the protein pair (A,B) is interacting it is placed in a positive set, otherwise, it is placed in a negative set.

2.4 Implementation

We use the proposed method to deal with the induced data set contained 11852 pairs and get the data_file satisfied the following format:

$$< line >=< target >< feature >:< value > ... < feature >:< value >$$

where the first entry$< target >= [+1|-1]$ gives the class labels(PPI or not), $< feature >= [integer]$ denotes the basis vector's index in the basis vector set and $< value >= [float]$ denotes the basis vector's weight satisfied that the summation of the weight's square in the same sequence is equal to 1. Note that the target value and each of the feature/value pairs are separated by a space character and feature/value pairs MUST be ordered by increasing feature number. Note again that indices start at 1.

2.5 Evaluation criterions

The performance of system is measured by how well a system can recognize interacting protein pairs. In order to analyze the evaluation measures in protein-protein interaction prediction, sub-sampling test and jackknife test are often used as two cross-validation methods [6]. Considering the numerous samples used in this work, 10-fold cross-validation was used to investigate the training set. To be precise, we first divided the sets of interactions and non-interactions (at random) into 10 roughly equal-sized non-overlapping subsets. We used each subset in turn as a test set, while we trained our method on the union of the remaining 9 subsets.

In order to evaluate the model on a positive and negative set of sequences, four statistics (counts) can be defined: the number of true positives (TP), false positives (FP), true

negatives (TN) and false negatives (FN). These represent the both predicted and observed, predicted but not observed, neither predicted nor observed, and not predicted but observed, respectively. We evaluated the performance of our classifier by computing accuracy (TP+TN)/(TP+FP+TN+FN), precision TP/(TP+FP), sensitivity TP/(TP+FN) and specificity TN/(TN+FP). Furthermore, the other criterions, such as receiver operating characteristic (ROC)curve [21], AUC(area under the ROC curve) is used.

3 Results and Discussions

3.1 Selecting suitable L

Using large value L will result in more quantities that account for interaction of amino acids with more distance apart in the sequence, and make the calculation expensive and time-consuming. However, if the value L is too small, the feature representations will not be well preformed. After some trials for selecting value L, we find that when $L = 25$ it can achieve a better characterization of the protein sequence.

3.2 Comparing the performance of C_{AB} with that of D_{AB}

After chosen suitable value L, a protein pair was converted into a 350-dimensional $(2 \times 25 \times 7)$ vector by concatenating operator with L of 25 amino acids. However, when distance on the protein pair was used, a protein sequence will be a vector of 175-dimensional vector. The final data set comprised of 11852 protein pairs,half from the positive data set and half from the negative data set. Here two-third of the protein pairs respectively from the positive and negative data set were randomly chosen as the training set (7900 protein pairs) and the remaining one-thirds were used as the test set (3952 protein pairs).

When the training set and the test set are prepared, we employ SVM to discriminate between the interacting and non-interacting proteins. SVMs have several advantages over other classifiers though we do not discuss them here. Instead, we refer to Vapnik [23] and Bennett and Campbell[2], among hers. To implement the SVMs in this paper, we used the software library named libsvm 2.89 (http:// www.csie.ntu.edu.tw/ Acjlin/ libsvm/) with Gaussian radial basis (RBF) kernel based on the induced data_file. The RBF kernel is used as it allows pockets of data to be classified which is more powerful way than just using a linear dot product. Two parameters, the regularization parameter C and the kernel width parameter γ were optimized using a grid search approach. A Python code which automates the process of SVM application is also available in this library package. This code tries to find optimal parameters for SVM application using RBF as the kernel function and returns an accuracy result on the clustering that is created by the SVM classification. In the numerical experiment, the penalty parameter $C = 5$ and the RBF kernel parameter $\gamma = 1$ which were determined by 10 fold cross validation.

Here, we used operator C_{AB} to represent the protein sequences and compared the performance of the model based on operator C_{AB} with that of the model base on operator D_{AB}. From Table 1, we can see that the model based on operation C_{AB} gives good results with the sensitivity, precision, accuracy and AUC of 90.05%,86.56%,87.98% and 0.863 respectively, whereas for operation D_{AB} ,84.43%, 76.85% , 81.15% and 0.774 ,respectively. These results imply that SVM with operation C_{AB} has the good generalization ability in prediction ability of PPIs on yeast.

Table 1: The performance of proposed methods

	TP	FN	TN	FP	Sensitivity	Precision	Accuracy	AUC
C_{AB}	1784	197	1693	278	90.05%	86.56%	87.98%	0.863
D_{AB}	1534	283	1673	462	84.43%	76.85%	81.15%	0.774

In addition to observations about specific classifiers, the accuracy, precision and sensitivity are useful for measuring the behavior of a classifier in general. In particular, the accuracy gives the overall performance of a classifier, the precision gives the percentage of positive predictions that are actually positive and the sensitivity gives the percentage of actual positives that are predicted. By looking at the precision and sensitivity statistics, we can determine if a classifier will identify positives correctly. If a classifier has a high precision and a low sensitivity, then it is likely to be correct when it makes a positive prediction, although it will make many false negative predictions. Conversely, a classifier with a low precision and a high sensitivity is likely to identify most true positives, even though many of its predictions will be false. In some sense, the first classifier is too conservative while the second is too optimistic.

3.3 Performance on the independent data set

In order to evaluate the practical prediction ability of the final prediction model, a large independent data set which is generalized by yeast two-hybrid experiments. After data preprocessing, Among the remaining 10138 protein pairs, 8014 PPIs are correctly predicted by the prediction model with C_{AB} and the success rate is 79.05%. We generate the negative set by using the positive data set randomly and get underlying 11012 non-interactions which can be incorporated into the test set. The result shows that the prediction model is able to correctly predict the non-interacting pairs with 74.36% accuracy. All these results demonstrate that this method is also capable of predicting well.

3.4 Comparing with other existing works

Comparing protein-protein interaction prediction systems with the other existing systems is always a difficult task. The reason is that, most of the authors used different type of data, experimental setup, and evaluation measures. In this section we will try to describe some of the good results achieved so far and compare them to our results. We will presents some of results achieved with an experimental work similar to ours in terms of the data used and experimental setup.

Kim et al [13] developed a statistical scoring system to measure the intractability between protein domains which could be used to predict protein-protein interaction. The prediction system gives about 50% sensitivity and more than 98% specificity.

Ng et al [17] developed an integrative approach to computationally derive putative domain interactions from multiple data sources. He reported true positive value of 58.97% and false positive value of 12.51%, which approximately yields sensitivity of 58.97%, specificity of 82.5% and accuracy of 73.23%.

Gomez et al [9] constructed an attraction-repulsion model associated with Pfam domains. The best result achieved in this study was a ROC score of 0.818. It's clear that our algorithm is outperformed most of the existing methods with cross-validation accuracy of 84.57% and ROC score reaches 0.8892.

4 Conclusion

Protein-protein interactions are operative at almost every level of cell function, in the structure of sub-cellular organelles, the transport machinery across the various biological membranes, packaging of chromatin, the network of sub-membrane filaments, muscle contraction, and signal transduction, regulation of gene expression, to name a few. In this article, the idea is to predict protein-protein interaction through sequence neighboring information. We have described a novel method that use SVMs, sequence neighbor information and experimental data to predict PPIs and explored its applicability by analyzing *Saccharomyces cerevisiae*. A proposed method is used for generating the score, which depends only on sequence neighbor information and allow us to perform transforming sequence information into physicochemical properties information. A data set of 11852 yeast PPIs was used to evaluate this prediction model and the prediction accuracy is 87.98%, which delivered the proposed method reasonable and promising. Our method also has the advantage of using a principled method (SVMs) to obtain our final classifier by statistical evaluation. Here we must to claim that more potential new PPIs will be predicted if we do not exploit too conservative attitudes towards dealing with the data set and model selection.

Efficient feature construction is important in determining the performance of a predictive method, thus future work can focus on how to improve feature extraction method, including optimizing the distance apart throughout the whole sequence. Future work can also be included to use more efficient and simple on imbalance classification problem to implement prediction task, such as SVM with an offset [15] and so on. Finally, the success of applying the proposed method on predicting protein-protein interaction encouraged us to plan future directions such as physicochemical properties discovering and finding related conserve information on the sequence.

Acknowledges

This work is partly supported by Chinese National Science Foundations (grant No. 10701080), Beijing Natural Science Foundations (grant No. 1092011) and Scientific Research Based on Beijing Wuzi University(No. WYJD200902).

References

[1] Alashwal,H. Deris,S. and Othman,R. (2006) Comparison of Domain and Hydrophobicity Features for the Prediction of Protein-Protein Interactions using Support Vector Machines,International Journal of Information Technology, Vol. 3, no. 1, 1305-2403.

[2] Bennett,K.P. and Campbell,C. (2000) Support vector machines: hype or hallelujah. ACM SIGKDD Explorations, 2, 1-13.

[3] Bock,J. and Gough,D. (2001) Predicting protein-protein interactions from primary structure. Bioinformatics, 17, 455-460.

[4] Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. Knowl. Discov. Data Mining, 2, 121-167.

[5] Charton,M. and Charton,B.I. (1982) The structure dependence of amino acid hydrophobicity parameters. J. Theor. Biol., 99, 629-644.

[6] Chou,K.C. and Zhang,C.T. (1995) Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol., 30, 275-349.

[7] Cristianini, N., and Shawe-Taylor,J. (2000) An introduction to Support Vector Machines, Cambridge, UK: Cambridge University Press.

[8] Grantham,R. (1974) Amino acid difference formular to help explain protein evolution. Science, 185, 862-864.

[9] Gomez,S.M., Noble,W.S. and Rzhetsky,A. (2003)Learning to predict protein-protein interactios from protein sequences, Bioinformatics, 19, 1875-1881.

[10] Han,J.D., Dupuy,D., Bertin,N., Cusick,M.E. and Vidal,M. (2005) Effect of sampling on topology predictions of proteinÍCprotein interaction networks. Nat. Biotechnol., 23, 839-844.

[11] Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl Acad. Sci. USA, 78, 3824-3828.

[12] Joel R. Bock, David A. Gough. (2001) Predicting protein- protein interactions from primary structure. Bioinformatics , 17 (5).

[13] Kim,W.K., Park,J. and Suh,J.K. (2002) Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair, Genome Informatics, 13, 42-50.

[14] Krigbaum,W.R. and Komoriya,A. (1979) Local interactions as a structure determinant for protein molecules: II. Biochim. Biophys. Acta, 576, 204-228.

[15] Li,B. Hu,J Hirasawa,K. Sun, P., Marko, K.(2006) Support vector machine with fuzzy decision-making for real world data classifcation. In IEEE World Congress on Computational Intelligence, Int Jiont Conf. on Nerual Networks,Canada.

[16] L. Liao, and W. S. Noble, (2003) Combining Pairwise Sequence Similarity and Support Vector Machines for Detecting Remote Protein Evolutionary and Structural Relationships, J. Comp. Biol., 10, pp: 857.

[17] Ng,S.K., Zhang,Z. and Tan,S.H. (2002) integrative approach for computationally inferring protein domain interactions, Bioinformatics, 19, 923-929.

[18] Pawson, T. and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains, Science, 300, 445-452.

[19] Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985) Hydrophobicity of amino acid residues in globular proteins. Science, 229, 834-838.

[20] Shen,JW., Zhang,J., Luo,X.M., Zhu,W.L., Yu,K.Q., Chen,K.X., Li,Y.X. and Jiang,H.L. (2007) Predicting protein-protein interactions based only on sequences information. Proc. Natl Acad. Sci. USA, 104, 4337-4341.

[21] Swets.(1988)Measuring the accuracy of diagnostic systems. Science,270, 1285-1293.

[22] Tanford,C. (1962) Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. J. Am. Chem. Soc., 84, 4240-4274.

[23] Vapnik,V. (1998) Statistical Learning Theory. Wiley Interscience, New York.

[24] Wojcik, J. and Schachter,V. (2001) Protein-Protein interaction map inference using interacting domain profile pairs, Bioinformatics, 17, 296-305.

[25] Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP: the database of interacting proteins. A research tool for studying cellular networks of protein interactions. Nucleic Acids Res., 30, 303-305.

[26] Yanzhi Guo, Lezheng Yu, Zhining Wen, Menglong Li. (2008)Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Research ,36(9),3025-3030.

[27] N. M. Zaki, S. Deris, and R. M. Illias, Feature Extraction for Protein Homologies Detection Using Markov Models Combining Scores, Int. J. on Comp. Intelligence and Appl., 1, pp: 1, 2004.

[28] Zhou,P., Tian,F.F., Li,B., Wu,S.R. and Li,Z.L. (2006) Genetic algorithm-base virtual screening of combinative mode for peptide/ protein. Acta Chim. Sinica, 64, 691-697.