Predicting Proteins Interactions from Protein Sequence Features using Support Vector Machines

Hany Alashwal, Safaai Deris and Razib M. Othman
Artificial Intelligence and Bioinformatics Laboratory
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Abstract *- Computational methods to predict protein-protein interactions are becoming increasingly important. This is due to the fact that most of the interactions data have been identified by high-throughput technologies like the yeast two-hybrid system which are known to yield many false positives. In this paper we investigate the use of two protein sequence features, namely, domain structure and hydrophobicity properties. The support vector machines (SVM) has been used as a learning system to predict protein interactions based only on protein sequence features. Protein domain structure and hydrophobicity properties are used separately as the sequence feature. Both features achieved accuracy of about 80%. But domains structure had receiver operating characteristic (ROC) score of 0.8480, while hydrophobicity had ROC score of 0.8159. These results indicate that protein-protein interaction can be predicted from domain structure with relatively better accuracy than hydrophobicity.*

Keywords: Bioinformatics, Support Vector Machines, Protein-Protein Interactions.

1.0 Introduction

A major challenge in bioinformatics is assigning function to newly discovered proteins. Although, most methods annotating protein function utilize sequence homology to proteins of experimentally known function, such a homology-based annotation transfer is problematic and limited in scope [1]. This is due to the fact that proteins work in the context of many other proteins and rarely work in isolation. The more we know about molecular biology the more we realize that protein-protein interactions affect almost all processes in a cell [2, 3]. For example structural proteins need to interact in order to shape organelles and the whole cell, molecular machines such as ribosome or RNA polymerases are hold together by protein-protein interactions, and the same is true for multi-subunit channels or receptors in membranes. It is estimated that even simple single-celled organisms such as yeast have about 6000 proteins interact by at least 3 interactions per protein, i.e. a total of 20,000 interactions or more [4]. By extrapolation, there may be on the order of ~100,000 interactions in the human body.

As a result, identifying protein-protein interactions (PPI) represents a crucial step in understanding proteins functions. Most of the interactions data was identified by high-throughput technologies like the yeast two-hybrid system, which are known to yield many false positives [5]. In addition, *in vivo* experiments that identify protein-protein interaction are still time-consuming and labor-intensive; besides, they identify a small number of interactions. As a result, methods for computational prediction of protein-protein interactions based on sequence information are becoming increasingly important.

Over the past few years, several computational approaches to predict protein-protein interaction have been proposed. One of the earliest techniques was based on the assumption that protein–protein interactions are evolutionary conserved. It involves orthology-based mapping of a known reference interaction network to another, target organism [6].

Other methods relies on exploration of similarity of expression profiles to predict

interacting proteins [7], coordination of occurrence of gene products in genomes, description of similarity of phylogenetic profiles [8] or trees [9], and studying the patterns of domain fusion [10]. However, it has been noted that these methods predict protein–protein interactions in a very general sense, meaning joint involvement in a certain biological process, and not necessarily actual physical interaction [11].

Another possibility to computationally predict interacting proteins is to correlate experimental data on interaction partners with computable or manually annotated features of protein sequences using machine learning approaches, such as support vector machines (SVM) [12] and data mining techniques, such as association rule mining [13].

The most common sequence feature used for this purpose is the protein domains structure. The motivation for this choice is that molecular interactions are typically mediated by a great variety of interaction domains [14]. It is thus logical to assume that the patterns of domain occurrence in interacting proteins provide useful information for training PPI prediction methods.

In a recent study, Kim et al. [15] introduced the notion of potentially interacting domain pair (PID) to describe domain pairs that occur in interacting proteins more frequently than would be expected by chance Assuming that each protein in the training set may contain different combinations of multiple domains, the tendency of two proteins to interact is then calculated as a sum over log odd ratios over all possible domain pairs in the interacting proteins. Using cross-validation, the authors demonstrated 50% sensitivity and 98% specificity in reconstructing the training data set.

Gomez et al. [16] developed a probabilistic model to predict protein interactions in the context of regulatory networks. A biological network is represented as a directed graph with proteins as vertices and interactions as edges. A probability is assigned to every edge and non-edge, where the probability for each edge depends on how domains in two corresponding proteins "attract" and "repel" each other. The regulatory network is predicted as the one with the largest probability for its network topology. Using the database of interacting proteins, DIP [17], as the standard of truth and PFAM domains as sequence features, the authors built a probabilistic network of yeast interactions and reported very high true positive and true negative rates of 93 and 90%, respectively.

Another sequence feature that has been used to predict PPI in-silico is the hydrophobicity properties of the amino acid residues. Chung et al. [18] have used SVM learning system to recognize and predict PPI in yeast Saccharomyces cerevisiae. They selected only the hydrophobicity properties as sequence feature and combine it to the amino acid sequence of interacting proteins. According to their experiments, they reported 94% accuracy, 99% precision, and 90% recall in average. Although they achieved better results than the previous work using only hydrophobicity feature, their method of generating a negative dataset (i.e. non-interacting proteins pairs) is different from the previous work. They constructed the negative interaction set by replacing each value of the concatenated amino acid sequence with a random feature value. This approach may simplify the learning task and artificially raise classification accuracy for training data. There is no guarantee, however, that the generalized classification accuracy will not degrade if the predictor is presented with new, previously unseen data which are hard to classify. Therefore, in this study we proposed a better and more realistic method to construct the negative interaction set. Then we compared the using of domain structure and hydrophobicity properties as the protein features for the learning system. The choice of these two features is motivated by the previous discussed literature.

This paper is organized as follows. Section 2 gives a general description of our method to design feature space, select training data, and conduct learning. Section 3 describes protein interaction data sets used in this work as the standard of truth and the implementation of our predictor. In Section 4 we present and discuss experimental results of this work. Finally, some ideas on future directions are provided in Section 5.

2.0 Methods

2.1 Support Vector Machines

The Support Vector Machine (SVM) is a binary classification algorithm. Hence, it is well suited for the task of discriminating between interacting and non-interacting protein pairs. The SVM is

based on the idea of constructing the maximal margin hyperplane in the feature space [19]. Suppose we have a set of labeled training data $\{\boldsymbol{x}_i, y_i\}$, $i = 1,\ldots, n$, $y_i \in \{1,-1\}$, $\boldsymbol{x}_i \in R^d$, and have the separating hyperplane $(\boldsymbol{w} \cdot \boldsymbol{x}) + b = 0$, where feature vector: $\boldsymbol{x} \in R^d$, $\boldsymbol{w} \in R^d$ and $b \in R$. In the linear separable case the SVM simply looks for the separating hyperplane that maximizes the margin by minimizing $\|\boldsymbol{w}\|^2/2$ subject to the following constraint:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 \qquad \forall i\ , i = 1,\ldots, n \tag{1}$$

In the linear non-separable case, the optimal separating hyperplane can be found by introducing slack variables ξ_i, $i = 1,\ldots, n$ and user-adjustable parameter C and then minimizing $\|\boldsymbol{w}\|^2/\ 2 + C\ \Sigma_i\ \xi_i$, subject to the following constraints:

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \ i = 1,\ldots,n. \tag{2}$$

The dual optimization is solved here by introducing the Lagrange multipliers α_i for the non-separable case. Because linear function classes are not sufficient in many cases, we can substitute $\Phi(x_i)$ for each example x_i and use the kernel function $K(x_i,x_j)$ such that $\Phi(x_i).\Phi(x_j)$. We thus get the following optimization problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{3}$$

subject to

$$0 \leq \alpha_i \leq C,\ i = 1,\ldots,n.\ \ \&\ \ \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{4}$$

SVM has the following advantages to process biological data [12]: (1) SVM is computationally efficient and it is characterized by fast training which is essential for high-throughput screening of large protein datasets. (2) SVM is readily adaptable to new data, allowing for continuous model updates in parallel with the continuing growth of biological databases. (3) SVM provides a principled means to estimate generalization performance via an analytic upper bound on the generalization error. This means that a confidence level may be assigned to the prediction, and avoids problems with overfitting inherent in neural network function approximation.

2.2 Feature Representation

The initial step of any supervised learning process is the construction of an appropriate feature space to describe training examples. In the context of protein-protein interactions, it is believed that the likelihood of two proteins to interact with each other is associated with their structural domain composition [13, 14, 15]. It is also assumed that the hydrophobic effects drive protein-protein interactions [3, 18]. For these reasons, this study investigates the applicability of the domain structure and hydrophobicity properties as protein features to facilitate the prediction of protein-protein interactions using the support vector machines.

The domain data was retrieved from the PFAM database. PFAM is a reliable collection of multiple sequence alignments of protein families and profile hidden Markov models [19]. The current version 10.0 contains 6190 fully annotated PFAM-A families. PFAM-B provides additional PRODOM-generated alignments of sequence clusters in SWISSPROT and TrEMBL that are not modeled in PFAM-A.

When the domain information is used, the dimension size of the feature vector becomes the number of domains appeared in all the yeast proteins. The feature vector for each protein was thus formulated as:

$$\boldsymbol{x} = [d_1, d_2, \ldots, d_i, \ldots, d_n] \tag{5}$$

where $d_i = m$ when the protein p has m pieces of domain d_i, and $d_i = 0$ otherwise. This formula allows the effect of multiple domains to be taken into account. Another representation is by using domain scores. In our case, each training example is a pair of interacting proteins (positive example) or a pair of proteins known or presumed not to interact (negative example).

In a similar approach, the amino acid hydrophobicity properties can be used to construct the feature vectors for SVM. The amino acids hydrophobicity properties are obtained from (Hopp & Woods, 1981). The hydrophobicity features can be represented in feature vector as:

$$\boldsymbol{x} = [h_1, h_2, \ldots, h_i, \ldots, h_n] \tag{6}$$

where k is the number of amino acid in the protein x, $h_i = 1$ when the amino acid is hydrophobic and $h_i = 0$ when the amino acid is hydrophilic.

DIP:255N	AAC1	YMR056C	DIP:1189N	APG12	YBR217W	DIP:11374E
DIP:255N	AAC1	YMR056C	DIP:1330N	LSM1	YJL124C	DIP:3367E
DIP:255N	AAC1	YMR056C	DIP:4443N	PUF3	YLL013C	DIP:6745E
DIP:255N	AAC1	YMR056C	DIP:2429N	RAD3	YER171W	DIP:13079E
DIP:6289N	AAC3	YBR085W	DIP:1189N	APG12	YBR217W	DIP:11375E
DIP:6289N	AAC3	YBR085W	DIP:5003N	BUD32	YGR262C	DIP:11546E
DIP:6289N	AAC3	YBR085W	DIP:1361N	HAP2	YGL237C	DIP:12245E
DIP:6289N	AAC3	YBR085W	DIP:963N	LAS17	YOR181W	DIP:12547E
DIP:6289N	AAC3	YBR085W	DIP:2429N	RAD3	YER171W	DIP:13080E
DIP:6289N	AAC3	YBR085W	DIP:2068N	RAD59	YDL059C	DIP:13131E

Fig. 1. A part of DIP database

3.0 Materials and Implementation

3.1 Data sets

Protein interaction data can be obtained from the Database of Interacting Proteins (DIP; http://www.dip.doe-mbi.ucla.edu/). At the time of our experiments, the database comprises 15117 entries representing pairs of proteins known to mutually bind, giving rise to a specific biological function. Here, *interacting* mean that two amino acid chains were experimentally identified to bind to each other. Each interaction pair contains fields linking to other public protein databases, protein name identification and references to experimental literature underlying the interactions. Figure 1 shows a part of DIP, where each row represents a pair of interacting proteins (the third and the sixth columns represent proteins names).

The proteins sequences files were obtained for the Saccharomyces Genome Database (SGD; http://www.yeastgenome.org/). The SGD project collects information and maintains a database of the molecular biology of the yeast Saccharomyces cerevisiae. This database includes a variety of genomic and biological information and is maintained and updated by SGD curators. The SGD also maintains the S. cerevisiae Gene Name Registry, a complete list of all gene names used in S. cerevisiae. This task was transferred to the SGD by Dr. Robert Mortimer in early 1994. We have also compiled a set of general guidelines to gene naming that may be of help to researchers who are naming new S. cerevisiae genes.

The proteins sequence information is needed in this research in order to elucidate the domain structure of the proteins involved in the interaction and to represent the amino acid hydrophobicity in the feature vectors.

3.2 Data Preprocessing

Since proteins domains are highly informative for the protein-protein interaction, we used domain data as the main feature for protein sequence. We focused on domain data retrieved from the PFAM database, a reliable collection of multiple sequence alignments of protein families and profile hidden Markov models. In order to elucidate the PFAM domain structure in the yeast proteins, we first obtain all sequences of yeast proteins from SGD. Given that sequence file, we then run InterProScan [20] to examine which PFAM domains appear in each protein. We used the stand-alone version of InterProScan. Apart from the result file is shown in Figure 2.

Fig. 2. A part from the protein domains file.

From the result file of InterProScan, we list up all PFAM domains that appear in yeast proteins and index them. Figure 3 shows an example of protein domains that appears in yeast genome. The first column represents a protein whereas the following columns represent the domains that appear in the protein. The order of this list is not important as long we keep it through the whole procedure. The number of all domains listed and indexed in this way is considered the dimension size of the feature vector, and the index of each PFAM domain within the list now indicates one of the elements in a feature vector.

Next we focus on the protein pair to be used for SVM training and testing. The assembling of feature vector for each protein pair can be done by concatenating the feature vectors of proteins constructed in the previous step.

When hydrophobicity is used, each amino acid will be replaced by 1 if it is hydrophobic and 0 if it is hydrophilic. Two separate training sets

for domain and hydrophobicity features have been constructed.

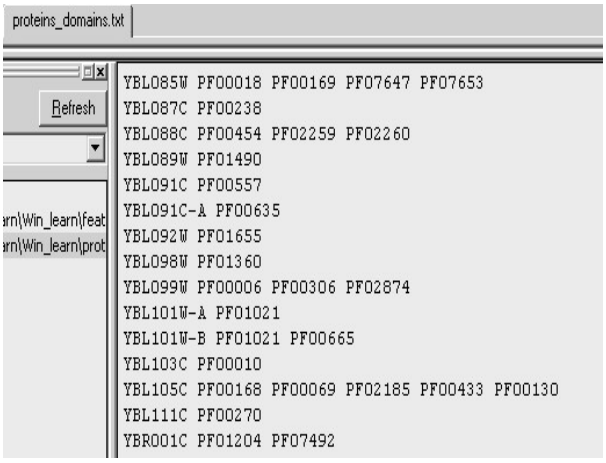

Fig. 3. An example of protein domains that appear in yeast genome.

4.0 Results and Discussion

In this study, we used the LIBSVM software [21] as a classification tool. The standard radial basis function (RBF) as available in LIBSVM was selected as a kernel function. Different values of γ for the kernel $K(x, y) = exp(-\gamma \|x-y\|^2)$, $\gamma>0$ were systematically tested to optimize the balance between sensitivity and specificity of the prediction. It is important to emphasize that in all our experiments we used only soft margin SVM. They are better suited for most real-world applications than hard margin SVM because the latter shows poor performance for overlapping classes; in our case, no priori knowledge was available on whether classes overlap or not.

Ten-fold cross-validation was utilized to obtain the training accuracy. The entire set of training pairs was split into 10 folds so that each fold contained approximately equal number of positive and negative pairs. Each trial involved selecting one fold as a test set, utilizing the remaining nine folds for training our model, and then applying the trained model to the test set.

The receiver operating characteristic (ROC) is also used to evaluate the results of our experiments. It is a graphical plot of the sensitivity (fraction of true positives - TP) vs. 1-specificity (the fraction of false positives - FP) for a binary classifier system as its discrimination threshold is varied. The sensitivity can be defined as: TP / (TP + FN) where TP and FN stand for true positive and false negative, respectively. The specificity can be defined as: TN / (TN + FP) where TN and FP stand for true negative and false positive, respectively. The area under the ROC curve is called ROC score.

In Table 1, a comparison between domains structure and hydrophobicity as the protein feature is presented. The cross-validation accuracy results indicate that there is no significant difference when using domain structure or hydrophobic properties as the protein feature. However, ROC score indicates that domain structure is noticeably better than hydrophobic properties (Figure 4). Another aspect is the running time for both features. Evidently, when domain structure used, the data set is much smaller than the data set for the hydrophobic properties. Consequently, the running time required for domain structure training data is much less than the time running required for the hydrophobic training data as shown in Table1.

TABLE I
THE PERFORMANCE OF SVM FOR PREDICTING PPI USING DOMAIN AND HYDROPHOBICITY FEATURES.

Feature	Accuracy	ROC score	Running time
Domain	79.4372 %	0.8480	34 seconds
Hydrophobicity	78.6214 %	0.8159	20,571 seconds (5.7 hours)

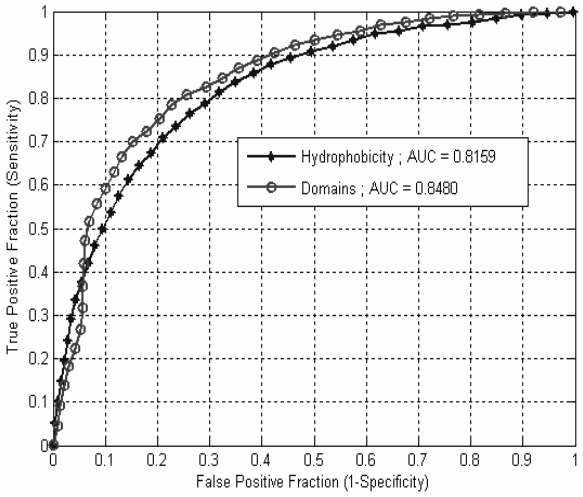

Fig. 4. The ROC plot for domain and hydrophobicity features.

5.0 Conclusion

The prediction approach reported in this paper generates a binary decision about potential protein-protein interactions based on the domain structure of the interacting proteins. One difficult challenge in this research is to find negative examples of interacting proteins, i.e., to find non-interacting protein pairs. For negative examples of SVM training and testing, we use a randomizing method. However, finding proper non-interacting protein pairs is important to ensure that prediction system reflects the real world. Discovering interacting protein patterns using primary structures of known protein interaction pairs may be subsequently enhanced by using other features such as secondary and tertiary structure in the learning machine. In conclusion the result of this study suggests that protein-protein interactions can be predicted from domain structure with reliable accuracy and acceptable running time. Consequently, these results show the possibility of proceeding directly from the automated identification of a cell's gene products to inference of the protein interaction pairs, facilitating protein function and cellular signaling pathway identification.

6.0 References

[1] B. Rost, J. Liu, R. Nair, K.O. Wrzeszczynski, and Y. Ofran. Automatic prediction of protein function. Cell. Mol. Life Sci. 60, 2637–2650. 2003.

[2] H. Lodish, A. Berk, L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, Molecular Cell Biology, 4th ed., W.H. Freeman, New York, 2000.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, Molecular Biology of the Cell, 4th edition, Garland Science, 2002.

[4] P. Uetz, and C.S.Vollert, "Protein-Protein Interactions", Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine (ERGPMM), Springer Verlag, 2005.

[5] E.M. Phizicky, and S. Fields, "Protein-protein interactions: Method for detection and analysis", Microbiological Reviews, 1995, pp.94-123.

[6] J. Wojcik, and V. Schachter, "Protein-Protein interaction map inference using interacting domain profile pairs", Bioinformatics, 2001, 17:S296-S305.

[7] E.M. Marcotte, M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg, "A combined algorithm for genome-wide prediction of protein function", Nature 1999, 402, pp:83–86.

[8] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles", Proc Natl Acad Sci USA 1999, 96, pp:4285–4288.

[9] F. Pazos and A. Valencia, "Similarity of phylogenetic trees as indicator of protein-protein interaction", Protein Eng. 2001, 14(9), pp:609-614.

[10] A.J. Enright, I. N. Ilipoulos, C. Kyrpides and C.A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events", Nature, 1999, 402, pp:86–90.

[11] D. Eisenberg, E.M. Marcotte, I. Xenarios, and T.O. Yeates, "Protein function in the post-genomic era", Nature, 2000, 405 pp:823-826.

[12] J.R. Bock and D.A. Gough. "Predicting protein-protein interactions from primary structure", Bioinformatics. May 2001, 17(5), pp:455-460.

[13] T. Oyama, K. Kitano, K. Satou, and T. Ito, "Extraction of knowledge on protein-protein interaction by association rule discovery", Bioinformatics, 2002, 18 no 5, pp:705-714.

[14] T. Pawson and P. Nash, "Assembly of cell regulatory systems through protein interaction domains", Science, 2003, 300, pp:445-452.

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

[16] S.M. Gomez, W.S. Noble, and A. Rzhetsky, "Learning to predict protein-protein interactions from protein sequences", Bioinformatics, 2003, Vol.19 no.15, pp:1875-1881.

[17] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions", Nucl. Acids. Res, 2002, 30(1), pp:303- 305.

[18] Y. Chung, G. Kim, Y. Hwang, and H. Park. Predicting Protein-Protein Interactions from One Feature Using SVM. In proceedings of IEA/AIE pp:50-55. 2004.

[19] V.N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.

[20] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, Griffiths-Jones, S. Khanna, A. Marshall, S.E. Moxon, L.L. Sonnhammer, D.J. Studholme, C. Yeats, and S.R. Eddy, "The Pfam: Protein Families Database", Nucleic Acids Research: Database Issue, 2004, 32, pp:D138-D141.

[21] N.J. Mulder, R. Apweiler, T.K. Attwood, et al., "The InterPro Database brings increased coverage and new features", Nucleic Acids Research, 2003, 31, pp:315-318.

[22] C.C. Chang and C.J. Lin, "LIBSVM : a library for support vector machines", 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm