World Academy of Science, Engineering and Technology
International Journal of Biomedical and Biological Engineering
Vol:1, No:7, 2007

Comparison of Domain and Hydrophobicity Features for the Prediction of Protein-Protein Interactions using Support Vector Machines

Hany Alashwal, Safaai Deris and Razib M. Othman

***Abstract*—**The protein domain structure has been widely used as the most informative sequence feature to computationally predict protein-protein interactions. However, in a recent study, a research group has reported a very high accuracy of 94% using hydrophobicity feature. Therefore, in this study we compare and verify the usefulness of protein domain structure and hydrophobicity properties as the sequence features. Using the Support Vector Machines (SVM) as the learning system, our results indicate that both features achieved accuracy of nearly 80%. Furthermore, domains structure had receiver operating characteristic (ROC) score of 0.8480 with running time of 34 seconds, while hydrophobicity had ROC score of 0.8159 with running time of 20,571 seconds (5.7 hours). These results indicate that protein-protein interaction can be predicted from domain structure with reliable accuracy and acceptable running time.

***Keywords*—**Bioinformatics, Protein-protein interactions, Support Vector Machines, Protein Features

I. Introduction

One of the major challenges in bioinformatics is assigning function to newly discovered proteins. Most methods annotating protein function utilize sequence homology to proteins of experimentally known function. However, such a homology-based annotation transfer is problematic and limited in scope [1]. This is due to the fact that proteins work in networks of many other proteins and rarely work in isolation. The recent studies of molecular biology realize that protein-protein interactions affect almost all processes in a cell [2], [3]. 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]. It is also estimated that, there may be nearly 100,000 interactions in the human body.

Therefore, 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. Some of the earliest techniques were based on the similarity of expression profiles to predict interacting proteins [6], coordinatation of occurrence of gene products in genomes, description of similarity of phylogenetic profiles [7] or trees [8], and studying the patterns of domain fusion [9]. However, it has been noted that these methods predict protein–protein interactions in a general sense, meaning joint involvement in a certain biological process, and not necessarily actual physical interaction [10].

Another recent method has been introduced based on the assumption that protein–protein interactions are evolutionary conserved. It involves the use of high-quality protein interaction map with interaction domain information as input to predict an interaction map in another organism [11].

Meanwhile, another approach to computationally predict protein-protein interactions is by associating experimental data on interacting proteins with 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 interacting domains [14]. It is thus logical to assume that the patterns of domain occurrence in interacting proteins provide useful information for training PPI prediction

Manuscript received March 21, 2006. This work was supported in part by the Ministry of Science, Technology and Environment, Malaysia, under Grant 74289.

Hany Alashwal is a Ph.D. student at the Faculty of Computer Science and Information Systems, Univeristi Teknologi Malaysia, 81310 Skudai, Johor, Malaysia, (phone: +607-5537791; fax: +607-5565044; e-mail: hany@siswa.utm.my).

Safaai Deris is a Prof. at the Software Engineering Department at the Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia, (e-mail: safaai@fsksm.utm.my).

Razib M. Othman is a lecturer at the Software Engineering Department at the Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia, (e-mail: razib@fsksm.utm.my).

methods.

In a recent study, the notion of potentially interacting domain pair (PID) was introduced to describe domain pairs that occur in interacting proteins more frequently than would be expected by chance [15]. 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. In a similar approach, [20] developed a scoring scheme which takes into account both experimental PPI data and interaction pairs derived computationally based on domain fusion analysis.

Reference [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 (Protein Families database) domains as sequence features, the authors built a probabilistic network of yeast interactions and reported an ROC score of 0.818.

Another sequence feature that has been used to predict PPI in-silico is the hydrophobicity properties of the amino acid residues. Reference [18] 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. 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. As they mention in their conclusion, this approach simplify the learning task and artificially raise classification accuracy for training data. However, there is no guarantee 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 use of domain structure and hydrophobicity properties as the protein features for the learning system. The choice of these two features is motivated by the above discussed literature.

This paper is organized as follows. The second section gives a general description of our method to design feature space, select training data, and conduct learning. The third section describes protein interaction data sets used in this work and the implementation of our predictor. In the forth section we present and discuss experimental results of this work. Finally, some ideas on future directions are provided in the fifth section.

II. Method

In order to compare two protein sequence features for the prediction of protein-protein interactions, we applied the same process on both features, as shown in Fig. 1.

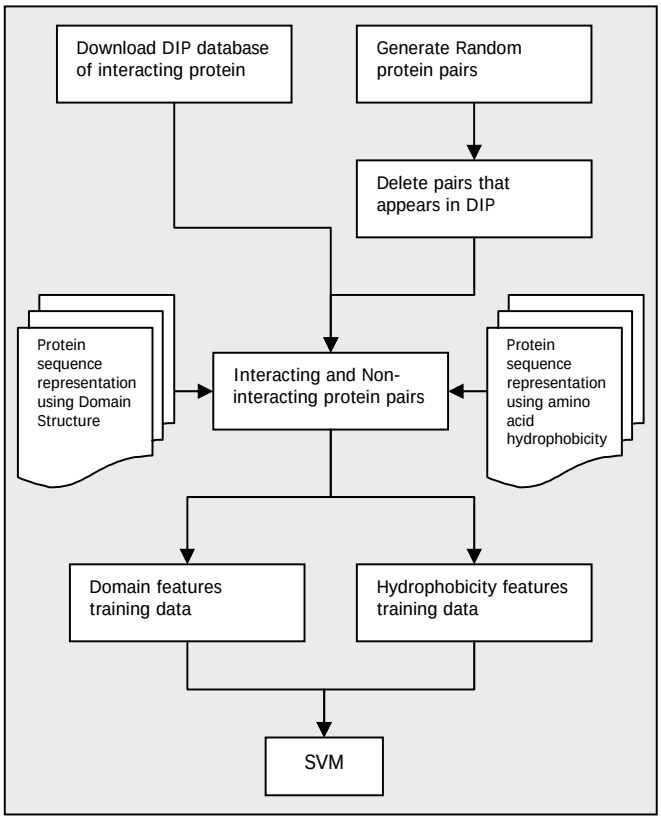

Fig. 1 The general operational framework

This process starts by generating a dataset of interacting and non-interacting proteins pairs. For the interacting pair, it is simply obtained from the Database of Interacting Protein (DIP). But, there is no dataset of experimentally identified non-interacting proteins. Therefore we use a random method to generate proteins pairs, and then delete all pairs that appear in the DIP. 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. The Support Vector Machines have been used as the learning system. It has been trained to distinguish between interacting and non-interacting protein pairs using domain and hydrophobicity training sets. The following sections give some details about the methods that were used in this work.

A. Support Vector Machines

The Support Vector Machine (SVM) is a binary classification algorithm. Thus, 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:

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

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 such that $K(x_i,x_j) = \Phi(x_i).\Phi(x_j)$. 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

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

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.

B. Feature Representation

The construction of an appropriate feature space that describes the training data is essential for any supervised machine learning system. 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 [14], [15], [20]. It is also assumed that the hydrophobic effects drive protein-protein interactions [4], [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 [21]. 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 that is calculated by PFAM. In this case d_i can be calculated as following:

$$\sum_{j=1}^{k} S_j(d_i) \tag{6}$$

where $S_j(d_i)$ is the score of d_i in the allocation j, and k is the number of the occurrence of d_i in the protein p. In order to scale the feature value to the interval [-1,1], we use the following formula.

$$\sum_{j=1}^{k} (6 - (\ln(S_j(d_i) + 0.1)) \tag{7}$$

In the same manner, the amino acid hydrophobicity properties can be used to construct the feature vectors for SVM. The amino acids hydrophobicity properties are obtained from [22]. The hydrophobicity features can be represented in feature vector as:

$$\boldsymbol{x} = [h_1, h_2, \ldots, h_i, \ldots, h_n] \tag{8}$$

where k is the number of amino acid in the protein x, hi = 1 when the amino acid is hydrophobic and hi = 0 when the amino acid is hydrophilic. We also consider the case where the hydrophobicity scale can be included in the feature vector by replacing the amino acid with its correspondent hydrophobicity value.

Using the above described four feature representations, we constructed four training set (domains, domains score, hydrophobicity, hydrophobicity scale). Each training example

is a pair of interacting proteins (positive example) or a pair of proteins known or presumed not to interact (negative example).

III. Materials and Implementations

A. Data sets

We obtained the protein interaction data from the Database of Interacting Proteins (DIP). The DIP database was developed to store and organize information on binary protein–protein interactions that was retrieved from individual research articles. The DIP database provides sets of manually curated protein-protein interactions in *Saccharomyces cerevisiae*. The current version contains 4749 proteins involved in 15675 interactions for which there is domain information. DIP also provides a high quality core set of 2609 yeast proteins that are involved in 6355 interactions which have been determined by at least one small-scale experiment or at least two independent experiments and predicted as positive by a scoring system [23]. Table I shows detailed description of the datasets that are comprised by DIP.

TABLE I
THE PROTEIN INTERACTIONS OF YEAST S. CEREVISIAE IDENTIFIED BY WET LAB EXPERIMENTS

Number of Proteins	Number or Interactions	Number of Experiments	Number of Interactions
4749	15675	1	13653
		2	1278
		3	407
		4	167
		5	84
		6+	101

The proteins sequences files were obtained for the *Saccharomyces* Genome Database (SGD) [24]. 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 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.

B. Data Preprocessing

Since proteins domains are highly informative for the protein-protein interaction, we used the domain structure of a protein as the main feature of the sequence. We focused on domain data retrieved from the PFAM database which is 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 [25] to examine which PFAM domains appear in each protein. We used the stand-alone version of InterProScan. A part of the result file is shown in Fig. 2.

```
<protein id="Q0065" length="544" crc64="A77CD9ADBDCA6465" >
<interpro id="IPR000883" name="Cytochrome c oxidase, subunit I"
type="Family">
   <child_list>
      <rel_ref ipr_ref="IPR004677"/>
   </child_list>
   <match id="PF00115" name="COX1" dbname="PFAM">
      <location start="5" end="339" score="8.2e-67" status="T"
evidence="HMMPfam" />
   </match>
</interpro>
<interpro id="IPR001982" name="Homing endonuclease, LAGLIDADG/HNH"
type="Domain">
   <match id="PF00961" name="LAGLIDADG_1" dbname="PFAM">
      <location start="316" end="403" score="6.4e-22" status="T"
evidence="HMMPfam" />
      <location start="422" end="515" score="3.2e-11" status="T"
evidence="HMMPfam" />
   </match>
</interpro>
```

Fig. 2 A part from the protein domains file

From the output file of InterProScan, we list up all PFAM domains that appear in yeast proteins and index them. Fig. 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.

The next step is to construct a feature vector for each protein. For example, if a protein has domain A and B which happened to be indexed 12 and 56 respectively in the above step, then we assign "1" to the 12th and 56th elements in the feature vector, and "0" to all the other elements. 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.

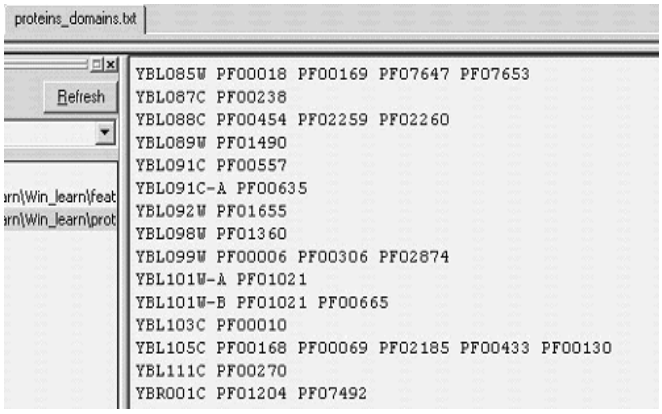

Fig. 3 An example of protein domains that appear in yeast genome

IV. Results and Discussion

We developed programs using Perl for parsing the DIP databases, control of randomization and sampling of records and sequences, and replacing amino acid sequences of interacting proteins with its corresponding feature. To make a positive interaction set, we represent an interaction pair by concatenating feature vectors of each proteins pair that are listed in the DIP-CORE as interacting proteins. For the domain feature we include only the proteins that have structure domains. The resulting positive set for domain feature contains 1879 protein pairs. But when using hydrophobicity feature, all protein in DIP-CORE were included which yielded 3002 protein pairs.

Constructing a negative interaction set is not an easy task. This is due to the fact that there are no experimental data in which protein pairs have confirmed to be non-interacting pairs. As a result, using a random approach to construct the negative data set is an avoidable at this moment. Furthermore, for the purposes of comparing prediction algorithms or feature representation, the resulting inaccuracy will be approximately uniform with respect to each computational method or feature representation. For these reasons, the negative interaction set was constructed by generating random protein pairs. Then, all protein pairs that exist in DIP were eliminated. This random approach can generate as many as 20202318 potentially negative candidates. Hence, the number of positive protein pairs is quite small compared to that of potentially negative pairs. The excessive potentially negative examples in the training set may lead to yield many false negatives because many of the positive examples are ambiguously discriminative from the negative examples in the feature space. For this reason, a negative interaction set was constructed containing the same number of protein pairs as for the positive interaction set for domain and hydrophobicity features.

In this study, we used the LIBSVM software [26] 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. Ten-fold cross-validation was used 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. Then the cross-validation accuracy is calculated in each run as the number of corrected prediction divided by the total number of data ((TP+TN)/(TP+FP+TN+FP)). Then the average is calculated for the 10 folds.

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.

The results of our experiments are summarized in Table II. All experiments reported in this work, run in Redhat Enterprise Linux AS release 3.2 on 1.8 GHz SMP CPUs with 2 GB of memory.

TABLE II
The overall performance of SVM for predicting PPI using domain and hydrophobicity features

Feature	Accuracy	ROC score	Running time
Domain	79.4372 %	0.8480	34 seconds
Domain Scores	76.397 %	0.8190	38 seconds
Hydrophobicity	78.6214 %	0.8159	20,571 seconds (5.7 hours)
Hydrophobicity Scales	79.1375 %	0.7716	34,602 seconds (9.6 hours)

When only domain structure was considered as the protein feature without information on domain appearance score, the cross-validation accuracy and ROC score were respectively 79.4372% and 0.8480. When domain scores were included the cross-validation accuracy and ROC score were decreased to 76.397% and 0.8190 respectively. These results indicate that it is not important to include the domains score information to the feature representation of the protein pairs. It is informative enough to consider only the existence of domains structure in the protein pairs. It is important here to note that the performance of the prediction algorithm is far better than an absolute random approach which has ROC score of 0.5. This indicates that the difference between interacting and non-interacting protein pairs can be learned from the available data.

In the case of hydrophobicity dataset, the cross-validation and ROC score were respectively 78.6214% and 0.8159. We

can see from these results that both domain dataset and hydrophobicity dataset have little difference in terms of cross-validation accuracy. On the other hand, ROC score indicates that domain structure is noticeably better than hydrophobic properties (see Fig. 4). Another aspect is the running time for both features. Clearly, 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 running time required for the hydrophobic training data as shown in Table II.

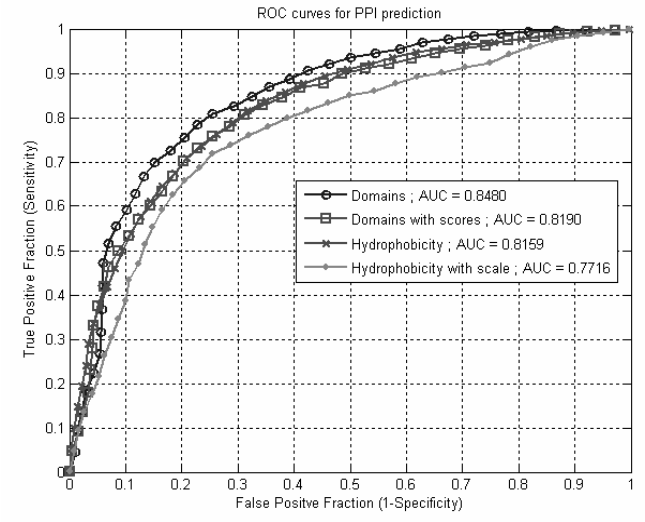

Fig. 4 ROC curves and scores for predicting protein-protein interactions

These results are better and came aligned with the results that have been obtained by [16] who reported ROC score of 0.818. Whereas our predictor achieved ROC score of 0.848 for domains feature dataset. However, Chung et al. (2004) reported accuracy of 94% by using hydrophobicity as the protein feature. The reason behind this big difference between our result and their results lies in the approach of constructing the negative interaction dataset. They assign random value to each amino acid in the protein pair sequence. This leads to get new pairs that considered negative interacting pairs and greatly different from the pairs in the positive interaction set. This leads to 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. In our work we constructed the negative interactions set by randomly generating non-interacting protein pairs which would be more difficult to distinguish from the positive set than entirely randomizing features values. This makes the learning problem more realistic and ensures that our training accuracy better reflects generalized classification accuracy.

V. Conclusion

The prediction approach reported in this paper generates a binary decision regarding potential protein-protein interactions based on the domain structure or hydrophobicity properties 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 protein 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.

References

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

[2] H. Lodish, A. Berk, L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular cell biology* (4th edition). 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*, pp.94-123, 1995.

[6] 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*, vol. 402, pp:83–86, 1999.

[7] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *In the proceedings of National Academy of Sciences*, USA, vol. 96, pp. 4285–4288, 1999.

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

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

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

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

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

[13] T. Oyama, K. Kitano, K. Satou, and T. Ito, "Extraction of knowledge on protein-protein interaction by association rule discovery," *Bioinformatics*, vol. 18(5), pp: 705-714, 2002.

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

[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*, vol. 13, pp: 42-50, 2002.

[16] S. M. Gomez, W. S. Noble, and A. Rzhetsky, "Learning to predict protein-protein interactions from protein sequences," *Bioinformatics*, vol. 19(15), pp: 1875-1881, 2003.
[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," *Nucleic Acids Research*, vol. 30(1), pp: 303- 305, 2002.
[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] S. K. Ng, Z. Zhang, S. H. Tan, and K. Lin, "InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes," *Nucleic Acids Research*, vol. 31, pp: 251–254, 2003.
[21] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D.J. Studholme, C. Yeats, and S. R. Eddy, "The Pfam: Protein Families Database," *Nucleic Acids Research*: Database Issue, vol. 32, pp: D138-D141, 2004.
[22] T. P. Hopp and K. R. Woods, "Predicting of protein antigenic determinants from amino acid sequences," *Proc. Natl Acad. Sci.* USA, 78, 3824-3828, 1981.
[23] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations," *Molecular & Cellular Proteomics*, vol. 1(5), pp: 349-56, 2002.
[24] Hong EL, Balakrishnan R, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Livestone MS, Nash R, Park J, Oughtred R, Skrzypek M, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Hitz B, Miyasato S, Schroeder M, Sethuraman A, Weng S, Dolinski K, Botstein D, and Cherry JM. "Saccharomyces Genome Database" http://www.yeastgenome.org/, (10th Oct 2005).
[25] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, et al., "The InterPro Database brings increased coverage and new features," *Nucleic Acids Research*, vol. 31, pp: 315-318, 2003.
[26] 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. (24th March 2005).