Prediction of Protein Interactions by Combining Genetic Algorithm with SVM Method

Bing Wang, Lu-Sheng Ge, Wen-You Jia, Li Liu and Fu-Chun Chen

Abstract—This paper proposes a novel hybrid GA/SVM method that can predict the interactions between proteins intermediated by the protein-domain relations. Firstly, we represented a protein by the domains contained inside, which can consider the effects of Domain duplication. To simulate the combination of different domains, a transformation of the domain composition was taken subsequently. Further, a genetic algorithm was used to seek the optimized transformation, which had been adopted as the input vector of a predictor constructed using support vector machines method. Finally, experiment results validated the effectiveness and efficiency of our proposed approach by a better performance.

I. INTRODUCTION

Protein-protein interactions play a pivotal role in live biological cells by controlling the functions that proteins perform[1]. Only over the past few years, however, do a vast amount of protein data and the associated data, which last but not least benefited from rapid development of high-throughput biotechnology, make it possible to investigate the interactions between proteins. Various organisms as diverse as Helicobacter pylori [2] and Saccharomyces cerevisiae [3] have been used to investigate the protein interactions maps. Some complex multicellular organisms such as Caenorhabditis elegans [4] and Drosophila melanogaster [5] are involved most recently.

Recent years, some large-scale experimental methods have been developed to analyze protein-protein interactions, mainly including yeast two-hybrid assay [6-9], protein chips[10], mass spectrometry of purified protein complexes [11, 12]. But, such techniques are tedious, time-consuming and labor-intensive [13], and suffer from high rates of both false positive and false negative predictions [14, 15]. Mrowke et al. even estimated that there are 90% protein interactions obtained by Ito and Uetz, which are not correct [16]. Therefore, it is becoming more and more important for researchers to seek some good computational approaches, which are much faster and less expensive than most experimental analyses, to predict whether some proteins interact each other or not.

As a result, several computational techniques have been suggested for predicting potential protein interactions. These approaches attempt to deal with the predicting problem of protein-protein interactions based on different biological background knowledge, including genomic information, evolutionary relationships, three-dimensional features, sequence properties, and so on. Some methods are to focus on the functional relationships between proteins, such as the Gene fusion (Rosetta stone) method [13, 17] and the phylogenetic profiles [18, 19]. The others are to emphasize particularly on structural interactions and related information, such as probabilistic model methods [20, 21], evolutionary important residues clustering [22-24] and domain pair profile method[2, 25]. In 2001, Bock and Gough[26] proposed training a Support Vector Machines, utilizing the protein interaction data from the DIP database[27], to recognize different pairs of interacting proteins on the basis of primary structure and associated physicochemical properties.

In addition, the prediction method drawn from the relationships between proteins and domains is another way to study protein-protein interactions. Domains are modules of amino acid sequences within the proteins, which are evolutionarily conserved. Domain can be regarded as a structural or/and functional unit of protein, and all proteins can be characterized by a distinct domain or a combination of domains. This is the fundament of reducing a protein-protein interaction to a few domain-domain interactions. Recently, there are several investigators who have developed some methods that allowed predicting the interactions between proteins from their containing domains. Ng et al.[28] developed an automated interacting domain discovery system, referred to as InterDom, based on an integrative approach. Kim et al.[29] and Han et al.[30] presented several statistical methods, which are similar to the association method described by Sprinzak and Margalit[31]. These approaches ignore experimental errors, and thus the precision of results is not apparent. Deng et al. [32] applied the Maximum Likelihood Estimation (MLE) methods to infer the domain-domain interactions from protein-protein interactions, which had been shown to be of robustness in dealing with various experimental errors. Furthermore, Gomez et al.[31] constructed an attraction-repulsion model associated with Pfam domains. Although these studies showed that the domains composition...
of protein pairs can be used to infer protein-protein interactions, most of them assumed that two proteins interact if and only if at least one pair of domains from the two proteins interact and domain-domain interactions are independent with each other. However, it is apparent that multiple domains take part in physical interaction in multiple complex structures. How can we take the information of multiple domains into account is a key to predict protein-protein interaction from protein-domain relationships.

In this paper, we present a new GA/SVM method, which combines support vector machines (SVMs) with genetic algorithm (GA), to tackle the predicting problem of the protein interaction based on domain composition of protein. In the first step, a protein is characterized by the domains existing inside it. Here, we not only consider the types of domain, but also take the number of domains into account. For a protein pairs, a feature vector is constructed by concatenation of each feature of the two proteins. To take the effect of multiple domains into consideration and reduce the dimensions of input vector, a transformation of domain composition is adopted in our experiment. Specifically, a GA algorithm is used to seek an optimization of the transformation to enhance the prediction performance of our proposed SVM predictor. Experiments have shown that the domain composition indeed can be used to infer protein-protein interactions and that the optimized transformation results in a significant improvement in prediction results.

II. MATERIALS AND METHODS

A. Feature representation

It is known that domains are highly informative for the protein-protein interaction. We formulate the protein-protein interaction prediction problem as a two-class classification problem: each protein pair is a sample belonging to either ‘interaction’ class (the two proteins interact with each other) or ‘non-interaction’ class (the two proteins do not interact with each other). In our application, a protein is characterized by the domains existing in each protein. The feature vectors for each protein therefore can be formulated as

\[ p = [d_1, d_2, ..., d_i, ..., d_n] \]

where each feature corresponds to a kind of domain existing in the protein, the value of \(d_i\) is the number of this type of domain, and \(d_i = 0\) otherwise. The domain compositions of each protein are retrieved from the Pfam database by using InterProScan. The effects of domain duplication can be taken into account by using this formula to construct the feature vector. The full feature vector for a particular protein pair is constructed by concatenation of each feature of the two proteins. This can be written as \( s = p_1 \otimes p_2 \), the length of full feature vector therefore is \(2n\), where \(n\) is the number of all kinds of protein domains we used here.

B. Dataset preparation

In the experiments reported here, the whole dataset was formed by the positive subset and the negative subset at a 1:1 ratio. The positive dataset is downloaded from *Saccharomyces cerevisiae* core subset of DIP database. This dataset is validated by two methods described by Deane and colleagues. The first is to use the expression profile reliability index to estimate the biologically relevant fraction of protein interactions by comparing the RNA expression profiles of the proteins with expression profiles of known interacting and non-interacting pairs of proteins. The second is to use the paralogous verification method to test the reliability of a putative interaction pair by examining whether there is a known paralog that also interacts with its partner protein or not. The DIP database (the 02/04/2006 version) includes 5951 protein-protein interactions in yeast organism.

Domain information can be obtained from Pfam database[33]. Pfam contains a large collection of multiple sequence alignments and profile hidden Markov models (HMM) covering the majority of protein domains. There are 1943 Pfam domains in the current Pfam version 19.0 and 3611 protein pairs of the above interaction within the core subset of DIP database have at least one domain in both of proteins of protein-protein interaction pairs. After excluded the domains not found in the all protein pairs we obtained in the above, only 1874 Pfam domains are used in this study. Therefore, the positive interacting protein dataset used here includes 3611 protein pairs and a feature vector with domain information of \(2*1874\) dimensions represented each protein pair.

Since a non-interacting protein dataset is not readily available, a hypothetical non-interacting protein dataset is generated based on subcellular localization information and consists of protein pairs that do not co-localize together. The subcellular localization source is retrieved from Munich Information Center for Protein Sequences (MIPS) and only the four main types of localization are considered in this study: cytoplasm, nucleus, mitochondria and endoplasmic reticulum. The yeast proteins used in the positive dataset are assigned with the four types of localization information and those with multiple localizations are removed to minimize the introduction of possible noise in the training process. Four sets of proteins with respect to the four types of localization are generated and proteins from each set are subsequently paired with proteins from a different localization.

Due to the enormous amount of possible pairing, 5000 protein pairs are randomly selected and used in this work. After removing duplication and performing exclusion analysis of the whole DIP yeast interacting proteins, 4660 protein pairs are used as the hypothetical non-interacting dataset.
C. Hybrid GA/SVM method

Although the dimensionality of the feature vectors is very high, the vectors contain relatively a much smaller number of non-zero features. It is known that interaction domains are often used repeatedly by many multi-domain proteins, and the evolution and spread of domains through different protein families exists as a result of gene duplication. The original representation of protein does not take the similarity and evolution of domains into account and it may not bring a satisfying classification rate between interacting protein pairs and non-interacting pairs. Here, we adopt a hybrid GA/SVM classifier to predict interacting protein pairs.

The basic idea behind the hybrid GA/SVM is that we, in the firstly, transform original domain composition for enhancing classification rate and reducing the dimensionality of the domain features, then GA is used to select the optimized transformation, while SVM is adopted to evaluate the transformation. The procedure of the hybrid GA/SVM algorithm is summarized as follows:

Initialize: Population \( p(0), t=0; \)
Evaluate: Obtain the performance using SVM base \( p(0); \)
Repeat: \( t=t+1; \)
select \( p(t) \) from \( p(t); \)
perform crossover on \( p(t); \)
mutate \( p(t); \)
evaluate \( p(t) \) using SVM ;
Return optimized transformation set.

Support vector machines. SVM is a class of supervised learning algorithms introduced by Vapnik [34], which is based on the well developed statistical learning theory. An important feature of SVM is that due to Mercer’s conditions on the kernels, the corresponding optimization problems are convex and hence have no local minima [35]. In this section, an overview of SVM is presented.

Given a set of labeled training vectors \((x_i, y_i), i=1,2,\ldots,l\) where \(x_i \in \mathbb{R}^n, y_i \in \{-1,+1\}\), we focus on two-class classification. SVM maps the input vector \(x_i\) into a high dimensional feature space \(H\) by a mapping function \(\Phi(\cdot)\) and finds a hyperplane, which maximizes the margin, i.e. the distance between the hyperplane and the nearest data points of each class in the space \(H\). The decision function implemented by SVM can be written as

\[
f(x) = \text{sgn} \left( \sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b \right)
\]  (2)

with

\[
K(x_i, x) = \langle \Phi(x_i), \Phi(x) \rangle
\]  (3)

where \(\alpha_i\)'s are the coefficients obtained by solving the following optimization problem:

\[
\begin{align*}
\text{maximize} & \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\
\text{subject to} & \quad 0 \leq \alpha_i \leq C
\end{align*}
\]  (4)

where \(C\) is regularization parameter, controlling the tradeoff between the margin and the misclassification error.

Transformation of domain composition. In this paper, we take the value of each feature as a histogram and then transform the original protein representation into a relative lower dimensional vector to classify a protein pair into interacting set or non-interacting set.

This operation is similar to constructing a mapping \(f: \mathbb{R}^n \rightarrow \mathbb{R}^{(M<N)}\) and the reduction of dimensionality can be realized by selective merging of the value of original features[36]. It is apparent that there exist a number of different transformations which can result in different classification performance. It is also expected that there should exist a set of suitable transformations by which better classification results can obtained than the original ones. As a result, our objective is to pursuit the best transformation within this candidate set.

Selection of the optimized transformation using GA/SVM Genetic algorithm is a randomized search and optimization techniques guided by the principles of evolution and natural genetics. In order to find out the optimum solution of a problem, a GA starts from a set of assumed solutions (chromosomes) and evolves different but better sets (of solutions) over a sequence of iterations. In each generation (iteration) the objective function (fitness measuring criterion) determines the suitability of each solution and, based on these values, some genetic operations (selection/reproduction, crossover and mutation) are operated to produce the next generation. Often, GAs can rapidly locate good solutions, even for difficult search spaces. The parameters for GA we adopted are listed in Table I.

In this paper, the chromosome is encoded into a character string whose length is equal to the size of the population. As shown in Fig. 1, a ternary alphabet \(\Lambda=\{a,b,c\}\) is adopted for the strings. In the transformation process of domain composition, if the consecutive characters in the chromosome are identical, the corresponding positions in the original vector can be merged together and their values are combined to a numeral. The process of transforming is illustrated in Fig. 1.
TABLE I.
PARAMETERS FOR GA

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>50</td>
</tr>
<tr>
<td>Crossover type</td>
<td>Single point</td>
</tr>
<tr>
<td>Crossover prob.</td>
<td>0.8</td>
</tr>
<tr>
<td>Mutation prob.</td>
<td>0.01</td>
</tr>
<tr>
<td>Selector type</td>
<td>Roulette wheel</td>
</tr>
<tr>
<td>Scaling scheme</td>
<td>Linear</td>
</tr>
<tr>
<td>Elitism</td>
<td>Yes</td>
</tr>
<tr>
<td>Termination</td>
<td>Best score not changed over 50 generations</td>
</tr>
</tbody>
</table>

In the GA/SVM method, the evaluation of each candidate transformation is taken by SVM. According to the transformations, we recombine the domain compositions and take it as input vector of SVM predictor. In our experiments, the dataset is randomly partitioned into two subsets with almost same size. Consequently, two SVM classifiers are trained using each of them and tested on its complement subset. The fitness function then can be defined as the average classification rate of these two SVM classifiers.

III. RESULTS

A. Evaluation criteria

To evaluate the methods for predicting protein-protein interactions, we use three measures: specificity, sensitivity and accuracy. The specificity is generally defined as the ratio of the number of matched non-interaction protein pairs between the predicted set and the actual set over the total number of negative samples. The sensitivity is defined as the ratio of the number of matched interaction protein pairs over the total number of the positive samples in the observed set. Let TP be the number of true positives, i.e. protein pairs predicted to be interacting pairs that actually are interaction pairs, and FP be the number of false positives, i.e. protein pairs predicted to be interacting protein pairs that are in fact not interaction pairs. In addition, let TN be the number of true negatives, and FN the number of false negatives. Then the evaluation measures can be computed as follows:

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{6}
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{7}
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{8}
\]

B. Prediction performance

The general performances of our method are shown in Table 2. Here, to facilitate evaluation of our proposed method, the performance of random prediction is also listed in Table 2. To obtain the expected performance at random we shuffled the predictions and assigned them to the protein pairs in the test set. This process is important for it enables us to infer the significance of our results. According to Table II, we can find that the higher value (higher than 0.8 in three performance measures: sensitivity, specificity and accuracy) is in domain composition based predictors while the value around 0.5 is in random predictors. The higher value (over 0.62) in correlation coefficient demonstrated that our SVM predictor outperformed the random predictor very much. The better performance of domain composition-based SVM predictor indicates that the domain composition indeed contains protein function information and it can be helpful for the identification of protein-protein interactions.

From Table II, it is apparently seen that the transformation of domain composition can enhance prediction performance significantly. On the one hand, selecting the transformation of domain composition using GA led to an impressive improvement in performance compared to the original
domain representation: at least 1.2% increase in sensitivity, 10.8% increases in specificity, 6.5 in accuracy, and 13% in correlation coefficient. On the other hand, after the optimization process, the dimensionalities of input vector are reduced to 1151 (only 64.5% of original dimensionality). It can decrease computational complexity and make our method much faster in predicting new protein-protein interactions.

### TABLE II.

**The General Performance Of Prediction**

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>Correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM(_{random})</td>
<td>0.5209</td>
<td>0.5208</td>
<td>0.5029</td>
<td>0.0058</td>
</tr>
<tr>
<td>SVM(_{a})</td>
<td>0.8319</td>
<td>0.7959</td>
<td>0.8139</td>
<td>0.6282</td>
</tr>
<tr>
<td>GA/SVM(_{c})</td>
<td>0.8543</td>
<td>0.9045</td>
<td>0.8794</td>
<td>0.7589</td>
</tr>
</tbody>
</table>

a. SVM\(_{random}\) is the predictor whose prediction results are shuffled randomly;
b. SVM\(_{a}\) is the predictor whose attributes are original domain composition information;
c. GA/SVM\(_{c}\) is the predictor whose attributes are transformed domain composition information using GA algorithm.

### C. Prediction of protein-protein interactions

We also predicted the putative protein-protein interactions in yeast. As stated earlier, the negative sample was selected randomly in the training process, and there is inevitable that some putative interacting protein pairs were selected into 'negative' set. The experimental procedure were therefore repeated three times while varying the 'negative' sets, the pairs that were predicted as positive in at least two experiments were extracted as reliable interaction predictions.

As a result, we obtained many novel protein-protein interactions that are not detected by two-hybrid assay and other experimental methods. For example, to our knowledge, there is only a piece of interaction information in all current databases relevant to protein-protein interactions contained in the interactions between the protein Q03306 (the primary accession number of this protein in Swiss-Prot database) with the other proteins. But we found that there were many potential interactions related to this protein in our experiment. The top 142 predicted interaction partners of Q03306, whose values from SVM predictor are identical (1.7336), are listed in Table III. These proteins are arranged according to the primary accession number in Swiss-Prot database.

### IV. DISCUSSIONS

In this paper, we proposed a new method for predicting protein-protein interactions based on the protein-domain relationships. Unlike most of previous approaches which tackled the same problem using domain information, our method can consider the multiple domain effects.

Specifically, we adopted an optimized transformation of domain composition, which was selected using GA algorithm to construct the SVM predictor, and the experimental results showed that the prediction performance was significantly enhanced.

Many cellular processes are involved in proteins comprised of multiple domains. So the multiple domain effects should be considered when we are focus on the prediction problem of protein-protein interaction. On the other hand, it is expected that protein-protein interaction can be predicted more accurately using multiply domains information. This point had been demonstrated by our experiments. In our experiments, we characterized proteins by the domains existing in each protein. Specifically, a feature transformation process was taken to further consider the domain combination effects.

We here regarded the prediction of protein-protein interaction as a two-class classification problem. The positive samples were extracted from the *Saccharomyces cerevisiae* core subset of DIP database, while the negative samples were generated randomly. To obtain credible non-interacting protein pairs, we employed subcellular localization information which is retrieved from MIPS. The better experimental results indicated that our selected dataset was effective and our proposed SVM predictor could capture the difference between the interacting protein pairs with the non-interacting pairs.

Furthermore, the performance of our method can be further improved when the domain information is further and more reliably annotated. Currently, we focus on the core subset of DIP database, and assume that only 3611 yeast interactions are used in this study. The number of all
interactions in *Saccharomyces cerevisiae* is estimated to be ~20000-30000. Although it is not confirmed, our dataset is a portion of actual interactions, which results in our predictor can capture the general features in this incomplete data. It is expected that the results will be improved using a big reliable dataset or incorporation of other diverse organisms.

**REFERENCES**


