ABSTRACT

Identifying the protein-protein interactions is essential for applications ranging from rational drug design to analysis of metabolic and signal transduction networks. Computational methods for identifying protein-protein interactions from primary structure are becoming increasingly important for the reason that the number of experimentally determined protein-protein interactions is small, time-consuming and labor-intensive. Behind protein-protein interactions there are protein domains interacting physically with one another to perform the necessary functions. Therefore it is relevant to use the domain structure of proteins as the main feature for the learning system. The support vector machines (SVM) has been trained to recognize and predict protein-protein interaction within Saccharomyces Cerevisiae using datasets from the Database of Interacting Proteins (DIP). The learning system generates a binary decision about potential protein–protein interactions, based only on protein domains structure. The accuracy of the system to predict protein-protein interactions in a 10-fold cross-validation experiment was 88%. This suggests 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.

KEYWORDS
Protein-protein interactions, Support Vector Machines, Protein primary structure.

1. Introduction

The majority of functions in cells are accomplished by proteins. Therefore, assigning functions to the proteins encoded by a genome is one of the crucial steps in gaining understanding of the organism. The sequencing of entire genomes has moved the attention from the study of single proteins or small complexes to that of the entire proteome. In this context, the search for reliable methods for proteins’ function assignment is increasingly important. Although, most methods annotating protein function utilise sequence homology to proteins of experimentally known function, about 40% of proteins cannot be assigned a function based only on similarity because they do not have statistically significant sequence similarity to functionally known proteins [1]. Therefore, computational approaches not only based on similarity have been introduced to predict proteins functions. These approaches include phylogenetic patterns, gene expression, and protein-protein interactions.

Part of the reason why it is difficult to relate the chemical function of a protein to its biological purpose using homology-based annotation is that proteins do not function alone. To understand the function of a protein, it must be considered in its proper cellular context, for example by appreciating how the cell would behave without it [2]. Many proteins are parts of larger complexes, which are the functional units that fulfill a role in the cell [3]. In this case it can be argued that all the proteins that form the complex should also have the same function. Since a protein does not perform its function alone but in the context of many other proteins as well as other bio-molecules, it is highly relevant to study the interaction partners of a protein in order to understand its function [4,5].

Most protein-protein interactions have been discovered by laboratory techniques such as yeast two-hybrid system that can detect all possible combinations of interactions. However, these findings can be superfluous and the number of experimentally determined structures for protein-protein interactions is still quite small. As a result, methods for computational prediction of protein–protein interactions are becoming increasingly important. Therefore the aim of this paper is to use the protein primary structure data to predict protein-protein interactions.

More recently, multi-protein interaction complexes have been elucidated by tandem-affinity purification and mass spectrometry [3,4] (Gavin et al., 2002; Ho et al., 2002). The best-studied organism in terms of...
protein–protein interactions is the budding yeast *Saccharomyces cerevisiae*. The DIP yeast interaction catalog currently contains information on roughly 15,000 interaction pairs obtained both from literature sources and large-scale protein interaction mapping projects [6]. Information gained from reconstruction of protein interaction networks based on experimental data is already providing invaluable clues to protein function [7].

Another possibility to delineate 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 [8] and data mining techniques, such as association rule mining [9]. The most common sequence attribute used for this purpose is the presence of well-defined sequence motifs corresponding to structural domains. The rationale for this choice is that molecular interactions are typically mediated by a great variety of widespread interaction domains that are frequently combined in proteins in a complicated mosaic fashion [10]. It is thus logical to assume that the patterns of domain occurrence in interacting proteins provide useful information for training PPI prediction methods (see Figure 1). More recently, several groups reported progress in reconstructing protein interactions based on protein domain composition using various statistical scoring functions [11,5].

The 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, the conclusions and some ideas on future directions are provided in Section 5.

![Figure 1. The protein domain architectures facilitate the interactions among proteins.](image)

2. Methods

2.1. Support Vector Machines

The Support Vector Machine (SVM) is a binary classification algorithm. As such 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 [12]. Suppose we have a set of labeled training data \( \{x_i, y_i\}, i = 1, \ldots, n, y_i \in \{1,-1\}, x_i \in R^d \), and have the separating hyperplane \( (w \cdot x) + b = 0 \), where feature vector: \( x \in R^d \), \( w \in R^d \) and \( b \in R \). In the linear separable case the SVM simply looks for the separating hyperplane that maximizes the margin by minimizing \(|w|^2/2\) subject to the following constraint:

\[
y_i(w \cdot x_i + b) \geq 1 \quad \forall i, i = 1, \ldots, n \quad (1)
\]

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

\[
y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n. \quad (2)
\]

The dual optimization is solved here by introducing the Lagrange multipliers \( \alpha_i \) 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:

\[
\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (3)
\]

subject to \( 0 \leq \alpha_i \leq C, \ i = 1, \ldots, n \) & \( \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (4)
\]

SVM has the following advantages to process biological data [8]: (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 starting point of any supervised learning process is the construction of an appropriate feature space to describe training examples. Since proteins domains are highly informative for the protein-protein interaction [13], 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 [14] (Bateman et al., 2004). 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 only 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:

\[ x = [d_1, d_2, ..., d_i, ..., d_n] \quad (5) \]

where \( d_i = m \) when the protein \( p \) has \( m \) pieces of domain \( d_i \), and \( d_i = 0 \) otherwise.

The effect of multiple domains can be taken into account by using this formula to construct the feature vector. 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).

3. 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 comprised 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 2 shows a part of DIP, where each row represents a pair of interacting proteins (the third and the sixth columns represent proteins names).

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 (Saccharomyces Genome Database). Given that sequence file, we then run InterProScan [15] 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 3.

From the result file of InterProScan, we list up all PFAM domains that appear in yeast proteins and index them. Figure 4 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 but we keep it through the whole procedure. The number of all

Proceedings of the Postgraduate Annual Research Seminar 2005 158
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. Figure 5 shows the format of the feature vectors to be used by SVM.

Figure 4. An example of protein domains that appears in yeast genome.

Figure 5. Feature vector format.

4. Results and Discussion

In this study, we used the LIBSVM software (Chang and Lin, 2001) as the classification tool. The standard radial basis function (RBF) as available in LIBSVM was selected as a kernel function. Different values of $\gamma$ 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 show poor performance for overlapping classes; in our case, no a priori knowledge was available whether classes overlap.

Ten-fold cross-validation was utilized to obtain the training accuracy. The entire set of training pairs was split into 10 folds such that each fold contained approximately equal number of positive and negative pair. 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.

Table 1. Comparison with the previous work.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Technique</th>
<th>ROC score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gomez et al. 2003</td>
<td>SVM</td>
<td>0.674</td>
<td></td>
</tr>
<tr>
<td>[16]</td>
<td>Attraction-Repulsion</td>
<td>0.818</td>
<td></td>
</tr>
<tr>
<td>Dohkan et al. 2004</td>
<td>SVM</td>
<td>0.94</td>
<td>79.5%</td>
</tr>
<tr>
<td>Deng et al. 2002</td>
<td>Maximum Likelihood Estimation</td>
<td>77.6%</td>
<td></td>
</tr>
<tr>
<td>Bock &amp; Gough 2001</td>
<td>SVM</td>
<td></td>
<td>80.96%</td>
</tr>
<tr>
<td>Sato et al. 2003</td>
<td>Partial Correlation Coefficient</td>
<td></td>
<td>80%</td>
</tr>
<tr>
<td>Chung et al. 2004</td>
<td>SVM</td>
<td></td>
<td>94%</td>
</tr>
<tr>
<td>Our Experiment</td>
<td>SVM with RBF and parameter selection</td>
<td>0.9723</td>
<td>88.72%</td>
</tr>
</tbody>
</table>

The receiver operating characteristic (ROC) is also used to evaluate the results of our experiments (Figure 6). An ROC 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.

Figure 6. The ROC plot.

5. 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. The most difficult thing 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 our future research. 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 classification accuracy for both positive and negative training examples of 88% achieved in this work is promising. This suggests 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