ISSN:0975-9646
Saswati Mahapatra et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, 1400-1402

A Domain-Based Frequency Count Approach for Protein-Protein Interaction Prediction using Support Vector Machine

Saswati Mahapatra, Tripti Swarnkar
Department of Computer Applications, Institute of Technical Education and Research
Siksha O Anusandhan University, Bhubaneswar-30, Orissa, India.

***Abstract*— Proteins are involved in many essential processes within cell. Uncovering the diverse function of proteins and their interactions within the cell may improve our understanding of protein functions. Several high-throughput techniques employed to decipher PPI are erroneous and are limited by the lack of coverage. Computational techniques are therefore sought to predict genome-wide PPI. In this paper, domain structure is used as a feature for computational prediction of PPI and support vector machines (SVM) as a learning system. We have used both, existing method and frequency count (FC) method for feature representation of protein domains and carried our experiment using SVM with different kernels. Both the methods achieved accuracy of about 78% for RBF kernel. But frequency count method reduced the storage requirement by half. These results indicate that PPI can be predicted from domain structure using frequency count method with reliable accuracy and reduced storage requirement.**

***Keywords* - *Protein-protein interactions, Support Vector Machines, Domain structure, Frequency Count Method*.**

I INTRODUCTION

Living systems are made up of various molecular entities. Chief among them are proteins, whose role in sustaining life has been known for more than 150 years, that is to say, long before the structure of DNA was unravelled. Proteins are involved in many essential processes within the cell such as metabolism, cell structure, immune response and cell signaling [1]. Research in [2] has suggested that the functionality of unknown proteins could be identified from studying the interaction of unknown proteins with a known protein target with a known function. Thus, the determination of protein-protein interactions (PPIs) is an important challenge currently faced in computational biology [3].

Large-scale high-throughput experiments have assisted in defining PPIs within the interactome(all possible PPIs in a cell). However, data generated by these experiments often contain false positives, false negatives, missing values with little overlap observed between experimentally generated datasets . This may suggest that the data are erroneous, incomplete or both [4].

Due to the limitations of experimental data computational methods (for example, statistical and machine learning techniques) have been applied at various stages in the inference of PPI networks, for instance, the integration of diverse heterogeneous datasets, the prediction of potential PPIs, the evaluation of predictions, and the analysis of inferred PPI networks [5],[6].

Several proposed computational 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] , and studying the patterns of domain fusion [9].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 .

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) [10] and data mining techniques, such as association rule mining [11].

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 [12]. It is thus logical to assume that the patterns of domain occurrence in interacting proteins provide useful information for training PPI prediction methods[13].

In a recent study, Kim et al. [14] 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 dataset.

Gomez et al. [15] developed a probabilistic model to predict protein interactions in the context of regulatory networks. Using the database of interacting proteins, DIP[16], 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 of93 and90%,respectively.

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 Section5.

II METHODS

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 hyper plane in the feature space [17]. Suppose we have a set of labelled training data $\{xi, yi\}$, $i = 1,\ldots, n$, $yi \in \{1,-1\}$, $xi \in R^d$, and have the separating hyper plane $(w \cdot x) + b = 0$, 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 hyper plane that maximizes the margin by minimizing $||w||2/2$ subject to the following constraint:

$$yi(w \cdot xi + b) \geq 1 \quad \forall i , i = 1,\ldots, n \qquad (1)$$

Taking the Lagrange multipliers *αi and the* the kernel function *K(xi, xj)* such that *Φ(xi).Φ(xj),* the dual optimization is solved the following optimization problem:

$$\max \sum_{i=1}^{n} \alpha i - \frac{1}{2} \sum_{i,j=1} \alpha i \alpha j \, yi \, yj K(xi, xj) \qquad (2)$$

subject to $0 <= \alpha i <= C$, $i = 1, \ldots, n$. &

$$\sum_{i=1}^{n} \alpha i \, yi = 0 \qquad (3)$$

SVM has the following advantages to process biological data [10],[17] (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 over fitting 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 [11], [18]. For these reasons, this study investigates the applicability of the domain structure 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].

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:

$$x = [d_1, d_2, \ldots\ldots d_i, \ldots\ldots d_n] \qquad (4)$$

if a protein has a domain with label i, then the ith number of the feature vector is 1, otherwise 0. 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).

III MATERIALS AND IMPLEMENTATION

A. *Data Sets*

We obtained the protein interaction data from the Database of Interacting Proteins (DIP)[16]. The DIP database provides sets of manually curated protein-protein interactions in *Saccharomyces cerevisiae*. The current version contains 5051 proteins involved in 23860 interactions for which there is domain information. We have taken 1000 protein interaction pairs.

DIP does not include any pair of non interacting proteins. So we randomly generate a set of non interacting protein pairs of size comparable to the number of the interacting protein pairs. Protein pairs which do not contain any domain pair in the training set are deleted because proteins with no domain information are of no use .

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 proteins sequence information is needed in this research in order to elucidate the domain structure of the proteins involved in the interaction .

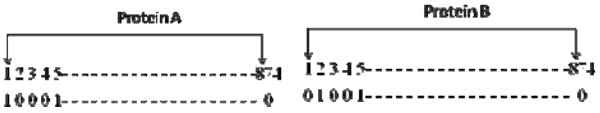
B. *Data Preprocessing*

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 [19] to examine which PFAM domains appear in each protein. We used the stand-alone version of InterProScan. From the output file of InterProScan, we list up all PFAM domains that appear in yeast proteins and index them. The number of all domains listed and indexed 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.

1) *Feature Vector Construction:*

We have used Frequency Count (FC) Method for the construction of feature vector. 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. If there are N distnict domains exist each protein will be represented by feature vector of size N. 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.

In FC method size of the feature vector for each protein pair(A-B) is N. If a domain indexed as 5 is found in both the proteins(A &B) in an interacting pair , we replace 5th entry with 2.If a domain is found in either in protein A or protein B(domain1 or domain 2), that entry is replaced by 1 otherwise 0.



Feature Vector for interaction pair A-B (FC Method)

1 2 3 4 5- - - - - - - - - - - - - - 874

1 1 0 0 2- - - - - - - - - - - - - - - - - - 0

Fig1: Feature representation for a pair of proteins.

In contrast to the FC method, the size of the feature vector in the existing method[20] of representation is 2N.In this context FC method reduces the storage requirement for the feature representation by half.

IV RESULTS & DISCUSSION

In this study, we used MATLAB 7.10.0.499 (R2010a) version for classification . The standard radial basis function (RBF) is taken as a kernel function. Different values *of y* for the kemel *K(x, y) = exp(-y* □x-y□), *y>O* 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.

In Table 1, a comparison between FC method and Existing method of feature representation using domain as protein feature is presented. The cross-validation accuracy results indicate that no significant difference is there in accuracy of both the results. However, when FC method is used, the memory requirement of the dataset is much smaller than the memory requirement by the existing method of feature representation.

TABLE 1
The performance of SVM for predicting PPI using FC method and the existing method of feature representation

Feature Used	Model Used	Accuracy	Required Storage Space	
			Freq. Count Method	Existing Method
Domain	SVM	79%	6960000 bytes	13920000 bytes

V 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 have used 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 machine learning. In conclusion the result of this study suggests that protein-protein interactions can be predicted from domain structure using Frequency Count method of representation with reliable accuracy and moderate storage requirement.

REFERENCES

[1] B.Alberts, Essential Cell Biology: An Introduction to the Molecular Biology of the Cell, Garland, New York, NY, USA,1998.
E.M.Phizick, S.Fields "Protein-protein interactions:methods for detection and analysis," *Microbiological Reviews*,vol. 59, no. 1, pp. 94–123, 1995.S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[2] P.Bork, L.J.Jensen,; C. von Mering, A.K.Ramani, I.Lee,E.M.,Marcotte "Protein interaction networks from yeast to human," *Current Opinion in Structural Biology*, vol. 14, no. 3,pp. 292–299, 2004.

[3] C.Mering, R.Krause, B.Snel, et al., "Comparative assessment of large-scale data sets of protein-protein interactions,"*Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[5] R.Jansen,M.Gerstein "Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction," *Current Opinion in Microbiology*, vol. 7, no. 5, pp. 535–545, 2004.

[6] R.Jansen,H.Yu, D.Greenbaum, et al., "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003.

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

[8] M.Pellegrini, E.M.Marcotte, MJ.Thompson, D.Eisenberg, TO.Yeates, "Assigning protein fimctions by comparative genome analysis: protein phylogenetic profiles", Proc Nat! Acad Sci USA 1999,96, pp:4285-4288.

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

[10] J.R.Bock, D.A.Gough, "Predicting proteinprotein interactions from primary structure",*Bioinformatics.*May2001 , 17(5), pp:455-460.

[11] T.Oyama,K.Kitano, K.Satou, T.Ito,"Extraction of knowledge on protein-protein interaction by association rule discovery",*Bioinformatics,2002,* 18no5, pp:705-714.

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

[13] R.Roslan, R.M.Othman, Z.AS.Shah, S.Kasim, et al." Incorporating multiple genomic features with the utilization of interacting domain patterns to improve the prediction ofprotein–protein interactions", Information Sciences 180 (2010) 3955–3973.

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

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

[16] I.Xenarios, L.Salwinski, XJ.Duan,P. Higney,et al. "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.

[17 Vapnik,V.N.; *The Nature of Statistical Learning Theory.* Springer, 1995.

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

[19] A.Bateman, L.Coin, R.Durbin, R.D.Finn,et al., "The Pfam: Protein Families Database," *Nucleic Acids Research*: Database Issue, vol. 32, pp: D138-D141, 2004.

[20] X.W.Chen, M.Liu, " Domain-Based Predictive Models for Protein-Protein Interaction Prediction", EURASIP Journal on Applied Signal Processing, Volume 2006, Article ID 32767, Pages 1–8.