Prediction of Protein Function Using Gaussian Mixture Model in Protein-Protein Interaction Networks

A.M. Koura†, A. H. Kamal, I. F. Abdul-Rahman

†Faculty of Computers and Information, Cairo University

Summary

Predicting protein function is one of the most important problems in the post-genomic era. Recent high-throughput experiments have determined proteome-scale protein physical interaction maps for several organisms. In this paper, a new method, which is based on Gaussian Mixture Model, is introduced to predict protein function from protein-protein interaction data. In the proposed method, A global information are taken into account by representing a protein using all the functional annotations of all proteins assigned with that term and have a shortest path with target protein in the all protein interaction network. We apply our method to a constructed data set for yeast and fly based upon protein function classifications of GO scheme and upon the interaction networks collected from IntAct protein-protein interaction. The results obtained by leave-one-cross-validation test show that the proposed method can obtain desirable results for protein function prediction and outperforms some existing approaches based on protein-protein interaction data.

Key words:
Gaussian Mixture Model, protein function, protein-protein interactions, Bayesian classifier

1. Introduction

With the rapid growth of sequenced genomes, the gap between the number of protein sequences deposited in public databases and the experimental annotation of their functions is widened gradually. It is necessary to know the function of the proteins when we investigate cellular and physiological mechanisms of organisms. Although experimentally determining protein function is more accurate, it is labor intensive and time-consuming. Therefore, developing a reliable computational method for predicting protein function is very significant for genome research. Some computational methods have been developed to predict protein function. This approach uses different biology data type. Earlier methodologies focus on estimating the function based on genomic sequence analysis, for example, analyzing sequence similarity between proteins listed in the databases [1] using programs[2, 3], using the gene fusion method or 'Rosetta stone' to infer yet unknown functions for protein[4], exploring the principle on similarity of phylogenetic trees for protein function prediction[5]. With the development of high-throughput experimental techniques, various high-throughput biological data, such as microarray gene expression profiles and mutant phenotype, have also been used to assign functions to novel proteins [5, 6].

Protein-protein interaction (PPI) biology data type has been employed by some researches to predict protein function [7-10]. Proteins play an important role in many biological functions within a cell and many cellular processes, and proteins collaborate or interact with each other to perform special biochemical events. Therefore, it is possible to deduce functions of a protein through the functions of its interaction partners.

Many approaches based on protein-protein interaction have been proposed for the prediction of protein function [7-10]. These methods assign functions to novel proteins by utilizing topological interaction patterns of a protein-protein interaction network. Schwikowski et al. [7] applied a straight-forward approach and predicted the function of an unannotated protein to be the most common one among its neighbors. Hishigaki et al. [8] considered both directly and indirectly connected proteins in the PPI network and developed a method based on Chi-square statistics. Deng et al and Letovsky and Kasif applied the model of Markov random fields (MRF) and provided statistical frameworks for the prediction of protein functions. Letovsky proposed a probabilistic approach (PA) for protein annotation, which assumes that the number of neighbors of a protein that are annotated with a given function is binomially distributed and the distribution's parameter depends on whether the protein has that function [10].

In our research, the protein function prediction is formulated as a binary classification problem with novel feature representation and the Gaussian mixture model used to estimate the likelihood rate. The proposed method, rather than using information about the local neighborhood of the protein, using global information on the whole network is taken into account when making predictions. For each function we used a Bayesian approach to compute the posterior probability that protein has a

Manuscript received April 5, 2010
Manuscript revised April 20, 2010

function. This study attempts to answer the question "inclusion any additional information on the whole network will improve the prediction of function of unlabeled proteins. Extensive experimental compare our method with NC and PAP, and show that the proposed method has better ability for the protein function prediction.

The remainder of the paper is organized as follows. In Section 2, we present our GMM- based prediction model. Extensive experimental results and comparison with other methods are reported in Section 3. Discussion of the proposed work is introduced in section 4. The paper is concluded in Sections 5.

2. Theoretical Consideration

Our goal is to assign GO terms to proteins in protein interaction networks. We formulate our pattern-based protein function prediction as a (multi-class) classification problem: GO terms are classes; proteins are items to be categorized into classes; and network information of proteins corresponds to features of items.

Definition protein interaction network can be represented as an undirected graph G = (V, E, F) with a set of vertices V and a set of edges E. Each vertex $v \in V$ represents a unique protein, while each edge $(u,v) \in E$ represents an observed interaction between proteins u and v, F is a finite alphabet of (annotation) terms (from a function vocabulary, e.g., Gene Ontology GO).

2.1 Feature Extraction

We computed a shortest-path vector for each protein using Dijkstra's algorithm from protein interaction network. Each node v is then identified by an n-dimensional feature vector where n is the number of terms. The ith component of the vector is a function of the lengths of the shortest paths in the graph between v and all nodes labeled with the ith term. Let $I_A(t)$ denote the indicator function of a set A that determines whether t belongs to A. i.e.

$$I_A(t) = \begin{cases} 1 & t \in A \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Let T_p and T_q denote the set of terms assigned to proteins p and q respectively. In this research, we adopt a form of feature vector driven from the global information of the underlying network. The form exploits the observation that the degree of similarity in a certain function between any two proteins in the network depends on the distance between them in the network.The feature vector of a protein p is described as:

$$X_p = \{x_{p_1}, x_{p_2}, \ldots, x_{p_m}\}, m = |F|$$

With

$$x_{p_t} = \sum_{q \in V} \exp[-d_{min}(p,q)]\, I_{T_q}(t) \quad (2)$$

where $d_{min}(p,q)$ is the shortest path length between protein p and q.

Note that the contribution of a protein q to feature element x_{p_t} increases with the decrease in the length of the shortest path between q and p provided that q is annotated with t on the other hand, q has no effect on p if it is not annotated with t. This emphasizes the usefulness of using the above equation.

2.2 Feature Reduction

In biological data, feature vector is large, so feature reduction is essential. We applied the principle component analysis PCA for the purpose of dimensionality reduction [11]. Principal components analysis is a statistical technique that linearly transforms an original set of variables into a substantially smaller set of uncorrelated variables that represents most of the information in the original set of variables. Its goal is to reduce the dimensionality of the original data set. A small set of uncorrelated variables is much easier to understand and use in further analyses than a larger set of correlated variables.

2.3 The Proposed Method

To know the set of terms that might associate some unknown protein p, first, we define a scoring function $f_t(p)$ for every term $t \in F$. Terms are then sorted in descending order according to $f_t(p)$. The topmost terms are supposed to have high chance of being considered associating p. We define the score function as being the ratio between the posterior probability that $t \in T_p$ given the feature vector of the protein p and posterior probability that $t \notin T_p$ given the feature vector of the protein p. This is mathematically described as follows:

$$f_t(p) = \log\left(P(t \in T_p | X_p)\right) - \log\left(P(t \notin T_p | X_p)\right) \quad (3)$$

We adopted a Bayesian approach to estimate $P(t \in T_p | X_p)$ and $P(t \notin T_p | X_p)$ and utilize the whole structure information of the network for this purpose as follows:

$$P(t \in T_p | X_p) = \frac{P(X_p | t \in T_p) P(t \in T_p)}{P(X_p)} \quad (4)$$

$$P(t \notin T_p | X_p) = \frac{P(X_p | t \notin T_p) P(t \notin T_p)}{P(X_p)} \quad (5)$$

Notice that it is the product of the likelihood and the prior probability that is most important in determining

the posterior probability; the evidence factor, $P(X_p)$, can be viewed as merely a scale factor that guarantees that the posterior probabilities sum to one, as all good probabilities must.

The prior probability $P(t \in T_p)$ could be estimated using a given protein interaction network as:

$$P(t \in T_p) = \frac{n_t}{n} \quad (6)$$

Here n_t is the number of proteins that t annotates and n is the number of all proteins in a given protein interaction network. The prior probability $P(t \in T_p)$, is estimated as:

$$P(t \notin T_p) = 1 - P(t \in T_p) \quad (7)$$

We propose Gaussian mixture Model (GMM) for the likelihood probabilities $P(t \in T_p|X_p)$ and $P(t \in T_p|X_p)$. We randomly select a set of i.i.d. samples of features of proteins annotated with term t as a training data for GMM to model $P(t \in T_p|X_p)$and another set not annotated with term t as training data for GMM model to build models for the term $P(t \notin T_p|X_p)$.

2.4 Gaussian Mixture Model

Gaussian Mixture Models (GMMs) are commonly used as parametric models of the probability distribution of continuous measurements or features [12]. The GMM's probability density function is represented as a weighted sum of Gaussian component densities.

$$p(x|\lambda) \sum_{i=1}^{M} w_i \; g(x|\mu_i, \Sigma_i) \quad (8)$$

Where x is a D-dimensional continuous-valued data vector (i.e. measurement or features), w_i, i = 1, . . . ,M, are the mixture weights and $g(x|\mu_i, \Sigma_i), i = 1, \cdots, M$ are the component Gaussian densities. Each component density is a D-variate Gaussian function of the form

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2} (x\mu_i)' \sum_{i}^{-1} (x\mu_i) \right\}$$

(9)

With mean vector μ_i and covariance matrix $\sum_i$. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$. The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \qquad i = 1, \cdots, M.$$

The parameters to be estimated are the mixing coefficients w_i, the covariance matrix $\sum_i$, and the mean vector μ_j. The form of the covariance matrix can be spherical, diagonal, or full. Maximizing the data likelihood is often used as a training procedure for mixture models. GMM parameters could be estimated from training data using the iterative Expectation-Maximization (EM) algorithm. The expectation-maximization (EM) algorithm, a well-established and common technique, is used for maximum-likelihood parameter estimation [13, 14]. The EM algorithm iteratively modifies the model parameters starting from the initial iteration $K = 0$. EM guarantees a monotonically non decreasing likelihood, starting from a random set of parameter values. It is thus able to find a local maximum which depends on parameter initialization. Good initialization of parameters, however, results in a near optimum solution to the maximum likelihood problem.

3. Experimental Consideration

For our experiments, we built protein interaction networks from IntAct protein-protein interaction database [15]. We employed two dataset, namely FLY (Drosophila melanogaster) and YEAST (Saccharomyces cerevisiae). FLY dataset contain 6666 proteins and 19565 interactions YEAST database contains 5974 proteins and 25555 interactions. The most recent gene ontology GO functional classification scheme [16] is accepted as the functional annotation scheme of proteins.

3.1 Cross-validation of function prediction

We assessed the performance of our function prediction approach by the leave-one-out cross-validation method. For each protein in annotations, we assumed it is unannotated and predicted its function using its interaction information and the annotations of the other proteins. Then we compared the predicted functions with the true annotations. The prediction performance was evaluated using precision and recall (also called true positive rate). Let M_i be the set of functions from the actual annotation in IntAct for a protein p_i ,N_i be the set of functions predicted by our algorithm for p_i and K_i be the set of common functions of M_i and N_i Precision and recall are then described

$$Precision = \frac{\sum_i |K_i|}{\sum_i |N_i|}, \quad (10)$$

and

$$Recall (True\ Positive\ Rate) = \frac{\sum_i |K_i|}{\sum_i |M_i|}, \quad (11)$$

where $|K_i|$ is the size of the set of K_i and n is the total number of distinct proteins that are annotated on at least one functional. Since there is a tradeoff between having high precision and high recall, we evaluate the accuracy of different techniques by using the F-values of predictions, instead. We employed F-value which is defined [17] as the harmonic mean of precision and recall of a prediction set.

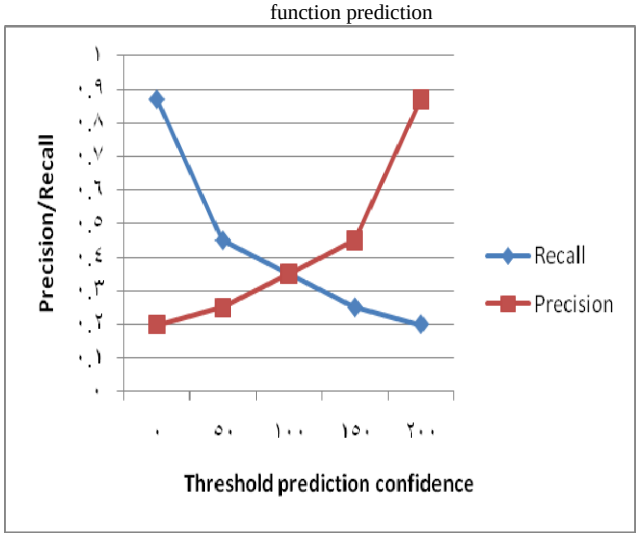
In table 1 shows number of component appropriate for our method. Table 1 shows the different component number of GMM to get accurate result.

Table 1: Accuracy F-values on each database

GMM Model Component	FLY	YEAST
2	.57	.52
3	.84	.82
4	.59	.56

Figure 1 illustrates the precision and recall plots with respect to the threshold of prediction confidence, which is a user dependent parameter in our algorithm. When we use 200 as the threshold of prediction confidence, our algorithm predicts no or a very few functions for each protein, but most of the functions are correctly predicted comparing to the actual annotations. It results in the precision of greater than 0.8. As a lower threshold is used, recall increases while precision decreases monotonically. Approximately, when the recall is 0.5 and 0.7, we had the precision of 0.4 and 0.2, respectively. In this experiment, we also merged data from the two species to observe how the prediction accuracy changes we tested our technique on FLY+YEAST using molecular function annotations only. Table 2 displays that the accuracy of our method did not decrease by the integration of cross-species information.

Figure 1: Precision and recall plots by cross-validation for protein function prediction



The performance of our function prediction algorithm was assessed by the leave-one-out cross-validation using the proteins that appear in the interaction data from IntAct and are annotated on the functional categories GO. As a higher threshold of prediction confidence is used, precision increases whereas recall decreases

Table 2: Accuracy F-values on merged database

	FLY	YEAST	FLY+YEAST
GMM-based	.84	.82	.87

3.2 Ontology Comparison

we tested the accuracy of GMM technique on FLY and YEAST datasets for Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) sub-ontologies of GO. Table 3 shows the results of this experiment on FLY and YEAST.

Table 3: Ontology comparison on FLY and YEAST

species	MF	BP	CC
FLY	.84	.72	.70
YEAST	.82	.86	.85

3.3 Comparison with other approaches

We compared our method to the neighbor counting (NC) technique [16] and pattern-based annotation prediction (PAP) method [17].For NC, PAP, and our technique, we computed the F-values of GO term predictions on FLY. We computed the F-value for each k value in top-k prediction tests. To sum up the prediction results, for each individual protein, we picked the k value that produces the highest F-value for that protein. Therefore F-values of techniques represent the highest possible accuracy of the technique, rather than the accuracy specific to the value of k. Figure 2 shows a plot of F-values against increasing values of k. Our method generates F-values higher than NC and PAP techniques for every value of k, except k =1.

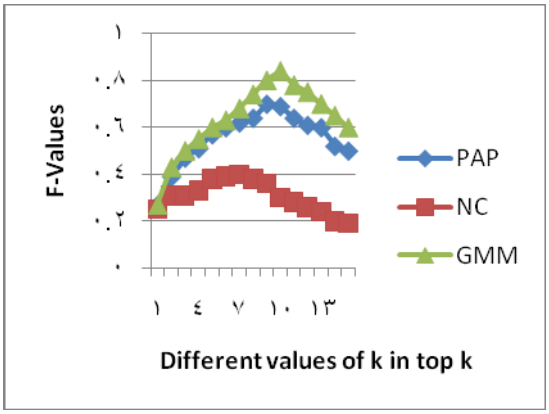


Figure 2: Accuracy of NC, PAP, and GMM for different values of k in top-k prediction experiments

4. Discussion

Through recent advances of high-throughput techniques, a significant amount of protein-protein interaction data has been accumulated. Protein function has been predicted from the interaction data because the evidence of interaction can be interpreted as functional links.

In this paper, we proposed a new method to predict protein functions based on Gaussian mixture model from protein-protein interaction data. Leave-one-out cross-validation method was performed on a constructed data set for fly and yeast to evaluate the prediction performance. One of the attractive advantages of the proposed method is that it considers the effect of global information on the whole network for the protein function prediction. Comparisons with the Neighbor Counting method (NC) and PAP method, The GMM has better performance and can be used to assign functions to novel discovered proteins as a supplementary method. NC method is chosen as a baseline in order to contrast with its assumption that interacting protein pairs have common annotations [18] while PAP methods are chosen to illustrate how much improvement is gained by the utilization of additional information from protein interaction network, since PAP employs only direct neighbors of proteins, and is shown to have reasonable accuracy in comparison with our methodology [19].To further improve the prediction accuracy, we can take into account the functional associations between indirect protein interactions and the topology of the protein interaction network to represent protein with the semantic knowledge in the Gene Ontology (GO) database for the purpose of improving the accuracy of function prediction.

5. Conclusion

The paper introduced a novel technique for prediction of protein functions. The technique used global features from the protein interaction networks and modeled various functions using a Gaussian mixture models. Bayesian decision was then taken to determine if a specific function is likely assigned to a given protein. Experimental results show that predication of protein function was improved by using the global information of protein interaction networks and the new modeling approach. Compared with other published technique, the proposed approach shows promising results in terms of the accuracy and recall.

References

[1] T. C. Hodgman, “A historical perspective on gene/protein functional assignment, “Bioinformatics, vol. 16, 2000, pp. 10-15.

[2] W. R. Pearson and D. J. Lipman, “Improved tools for biological sequence comparison,” Proc Natl Acad Sci U S A, vol. 85, 1988, pp.2444-2448.

[3] L. F. Wu, T. R. Hughes, A. P. Davierwala, M. D. Robinson, R Stoughton, and S J Altschuler, “Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters,” Nat Genet, vol. 31, 2002, pp. 255-265

[4] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, “Detecting protein function and protein-protein interactions from genome sequences,” Science, vol. 285, 1999, pp. 751

[5] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, “Protein interactions: two methods for assessment of the reliability of high throughput observations,” Mol Cell Proteomics, vol. 1, 2002, pp. 349-356.

[6] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T S Furey, M Ares, Jr., and D Haussler, “Knowledge-based analysis of microarray gene expression data by using support vector machines,” Proc Natl Acad Sci U S A, vol. 97, 2000, pp. 262-267.

[7] D. J. Reiss and B. Schwikowski, “Predicting protein-peptide interactions via a network-based motif sampler,” Bioinformatics, vol. 20 Suppl 1, 2004, pp. I274-I282.

[8] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi, “Assessment of prediction accuracy of protein function from protein-protein interaction data,” Yeast, vol. 18, 2001, pp. 523-531.

[9] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun,” Prediction of protein function using protein-protein interaction,” In Proceeding of IEEE Computer Society Bioinformatics Conference, pages 197–206, 2002.

[10] S. Letovsky, and S. Kasif,” Predicting protein function from protein/protein interaction data: a probabilistic approach,” Bioinformatics, 19:i197–i204, 2003

[11] H. Hotelling,” Analysis of a complex of statistical variables into principal components,” Journal of Educational

Psychology, 24:417-441,498-520,1933.
[12] R. C. Rose, E. M. Hosftetter, and D. A. Reynolds, “Integrated models of speech and background with application to speaker identification in nosie ,”IEEE Trans .Speech Audio Processing ,vol. 2,no.2,pp.245-257,Apr.1994.
[13] A .Dempster, N. Laird, and D.Rubin, “Maximum likelihood from incomplete data vai the EM algorithm,” J.Royal Stat .Soc., vol.39,pp.1-38, 1977.
[14] L.Baum et al.,”A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,” Ann.Math Stat.,vol.41,pp.164-171,1970.
[15] Henning Hermjakob et al., IntAct,” an open source molecular interaction database,” Nucleic Acids Research, 2004, Vol. 32.
[16] Gene Ontology Annotations Database, available http://www.geneontology.org/GO.curent.annotations.shtml
[17] W. M. Shaw , R Burgin , and P. Howell,” Performance standards and evaluations in IR test collections: Vector-space and other retrieval models,”. Info Proc Manag 1997, 33(1):15–36.
[18] B. Schwikowski , P. Uetz , and S. Fields,” A network of protein–protein interactions in yeast,” Nat Biotechnol 2000, 18:1257–1261.
[19] M. Kirac, G. Ozsoyoglu ,”Protein function prediction based on patterns in biological networks,”.Proceedings of 12th International Conference on Research in Computational Molecular Biology (RECOMB) 2008:197-213