Data Integration for the Study of Protein Interactions

Fengzhu Sun*, Ting Chen*, Minghua Deng, Hyunju Lee, Zhidong Tu

Molecular and Computational Biology Program

Department of Biological Sciences

University of Southern California

1042 West 36th Place, DRB 155

Los Angeles, CA 90089-1113

Tel: (213) 740-2413

Fax: (213) 740-2437

Email: fsun@usc.edu or tingchen@usc.edu

∗corresponding author

running title: data integration for protein interactions

Keyword: Protein interaction, Markov random field, protein function, EM algorithm, Kernel

Abstract

With the development of genomic technologies, enormous amount of biological data have been and are continually being generated. They include genomic sequence data, gene expressions, protein-protein interactions, protein structures, protein localizations, protein functions, etc. For biological problems of interest, each data source contributes partially to the understanding of the problems. An important issue is how to integrate the different data sources to obtain a more complete understanding of the problems. In addition, most of the data sources from the high throughput experiments contain many false positive and false negative errors. Statistics plans an essential role in understanding the reliability of the observed biological data as well as to choose a more reliable data set from the observed ones. Statistics and machine learning techniques can help the integration of different data sources to understand the biological problems. We present two examples: to study the reliability of observed protein interaction data sets, and to predict protein functions combining different data sources.

1 Introduction

In recent years, an increasing number of genomes of model organisms have been sequenced. Using these genomic sequences, researchers have been able to make tremendous progress in the study of genomes, such as numerous successes in the identification of genes, the detection of protein-binding DNA motifs, and the determination of gene regulation. Beyond these successes is the far more challenging and rewarding task of understanding proteomes by means of, e.g., (1) discovering signal transduction pathways, (2) determining protein structures, (3) detecting protein-protein, protein-DNA, and protein-metabolite interactions, (4) detecting post-translational modifications of proteins, and ultimately (5) elucidating the functions of genes and their protein products.

Unlike a genome, which is a stable feature of an organism, a proteome varies with the state of the development, the tissue, and the environment. Among many features of a protein, the interaction with other proteins is one of the most important aspects of its function. Traditionally, protein interactions have been studied individually by biochemical techniques. However, the speed of discovering new interactions increased dramatically in the last couples of years; several high-throughput techniques have produced a total of about 80,000 interactions between yeast proteins, which constitute a rough view of the actual protein-protein interaction network. The successful methods include yeast two-hybrid assays (Uetz et al. 2000; Ito et al. 2000; Ito et al. 2001), protein complex purification-mass spectrometry (Gavin et al. 2002; Ho et al. 2002), microarray gene expression profiles (Eisen et al. 1998), genetic interactions (Tong et al. 2002; Mewes et al. 2002), and computationally predicted protein associations (Enright et al. 1999; Marcotte et al. 1999a; Marcotte et al. 1999b). These protein interactions will be very useful to study gene regulatory networks, pathways, as well as functions of proteins. To understand the interaction network and its applications for protein function prediction, it is essential to design a joint approach using tools from mathematics, statistics, computer science, and molecular biology. In recent years, several groups have developed computational tools to analyze and compare the different interaction data sets.

Two issues are important in assessing the usefulness of an experimentally observed protein-

protein interaction data set. One is the *reliability* which is defined as the fraction of real protein-protein interactions in the observed interactions and the other is the *coverage* which is defined as the fraction of real interactions in the observed data over all the real interactions. A database of high coverage is not very useful if its reliability is low. Results of comparative analysis of multiple data sets have shown significantly different coverage and reliability for each technique (Deane et al. 2002; Mering et al. 2002; Mrowka et al. 2001). In this paper we review methods to study the following problems:

1. Estimate the reliability of a putative observed interaction data set;

2. Give a score that a pair of proteins interact by combining different data sources.

Assigning functions to novel proteins is one of the most important problems in the post-genomic era. Many researchers have undertaken the task of functionally analyzing one of the most well-studied species, the yeast genome, comprising approximately of around 6400 proteins, of which roughly one-third do not have known functions (Mewes et al. 2002), and the other two-thirds, most likely, have many other unknown functions. The annotation of the yeast genome will have a great impact on genomes of higher organisms such as the human: new genes can be annotated through their homologous yeast genes.

Several approaches have been applied to assign functions to genes, including analyzing gene expression patterns, phylogenetic profiles, protein fusions and protein-protein interactions. Gene expression analysis can cluster genes based on similar expression patterns. This makes it possible to assign a biological function to genes, depending on the knowledge of the functions of other genes in the cluster (Eisen et al. 1998). However, expression profiling gives an indirect measure of a gene product's biological and cellular function, because many cellular processes and biochemical events are ultimately achieved by interactions of proteins. A more complete study of protein functions can be achieved by looking at not only the mRNA levels but also the protein interaction network. We will review the following methods for protein function prediction:

1. A Markovian random field (MRF) model for assigning functions to proteins using highly

reliable protein-protein interaction data and other data sources including gene expression profiles, protein sequence similarities, and features of individual proteins, and correlations of protein functions.

2. The use of support vector machine (SVM) for protein function prediction combining different data sources;

3. A kernel-based MRF model for protein function prediction.

The paper is organized as follows. We first provide the data sources for the studies. Then we divide the paper into two major sections: estimating the reliability of observed putative protein interactions and predicting protein functions based on reliable protein interactions and other data sources. We then discuss the connections of the two topics and future research questions.

2 Data Sources

Protein interactions have traditionally been studied individually by genetic, biochemical, and biophysical techniques. However, these techniques are generally labor intensive and cannot keep up with the speed new proteins are discovered. Recently, several high-throughput methods for the detection of protein interactions have been developed. These include the yeast two-hybrid assays (Ito et al. 2000; Ito et al. 2001; Uetz et al. 2000), mass spectrometry (Gavin et al. 2002) and gene knockouts (Tong et al. 2002). *In silico* (computational) methods for interaction prediction include the chromosomal proximity method (Overbeek et al. 1999), the gene fusion method (Enright et al. 1999; Marcotte et al. 1999a), the phylogenetic method (Pellegrini et al. 1999), and the combined method (Marcotte et al. 1999b; Pavlidis et al. 2001; Zheng et al. 2002). Several databases have been developed to collect different sources of protein interaction data including the Munich Information Center for Protein Sequences (MIPS: http://mips.gsf.de/) (Mewes et al. 2002), Database of Interacting Proteins (DIP: http://dip.doe-mbi.ucla.edu/) (Xenarios et al. 2002), Biomolecular Interaction Network Database (BIND: http://www.bind.ca/)(Bader et al. 2003), and the General Repository for Interaction Datasets (GRID: http://biodata.mshri.on.ca/grid)(Breitkreutz et al. 2003).

Gene Expressions are widely used to study the relationship between proteins. It is generally believed that a pair of interacting protein pair are more likely to be co-expressed than random protein pairs and thus gene expression data can be useful for evaluating the reliability of protein interaction data as well as the probability that two proteins interact. It is also generally believed that if two proteins are highly correlated, they are more likely to have similar functions. Therefore, gene expression data can also be useful for protein function prediction. For this study, we use the gene expression data from Spellman et al. (1998). Other gene expression data can also be used.

Protein localizations. Proteins belong to different localizations in the cell and proteins within the same locations are more likely to interact. Therefore, protein localization data can be useful for predicting protein interactions. We use the protein localization data of Huh et al. (2003) in this study.

Domains. The amino acid sequence of a protein is extremely important for the proteins function. A proteins sequence determines its secondary and tertiary structure and thus, determines its interaction partners and its biological functions. Protein domains are conserved regions of peptide sequences with relatively independent tertiary structures and represent important features for understanding protein function. We use Pfam domains as the source of domain information. The SwissPfam (ver7.5) (ftp://ftp.genetics.wustl.edu/pub/pfam/) defines the mapping between proteins SWISS-PROT/TrEMBL accession numbers and Pfam domains.

Gene Ontology (GO) (http://www.geneontology.org/) (GO Consortium 2001) describes gene products (proteins or RNA) based on three principles: Cellular component, Molecular function, and Biological process. GO has a directed acyclic graph (DAG) structure. The high level categories are more general and contain many more genes than low level categories. For protein function prediction, we base on the known gene annotation given in GO.

All the databases listed above are publicly available.

3 Assessing the reliability of protein interaction data

Many protein interaction data sets generated from various laboratories using different techniques are available. It is difficult to compare different interaction data because different conditions and experimental techniques may not detect the same type of interactions. Another difficulty comes from the fact that the true interaction data is unknown. Two issues need to be considered in comparing different interaction data sets. One is the *reliability* of the observed interaction data set defined as the overlap between the true interactions and the observed interactions over all the observed interactions. The other is the *coverage* defined as the overlap between the true interactions and the observed interactions over the true interactions. Without knowing the true interaction data, it is difficult to study the coverage of a certain observed interaction data set. On the other hand, it is possible to study the reliability of an observed interaction data set using gene expressions and localizations.

Mrowka et al. (2001) first observed that the distribution of correlation coefficients of gene expressions for true interacting protein pairs is stochastically larger than that for random protein pairs. The distribution of gene expression correlation coefficients for observed interacting protein pairs from high-throughput yeast-two-hybrid assays is between that for random protein pairs and that for true interaction pairs. The observations indicate that the set of observed protein interactions from high-throughput experiment is a mixture of random protein pairs and true interaction pairs. Several problems are of interests:

1. How do we choose the true interaction set (the gold standard)?

2. How do we estimate the fraction of true interactions among a set of observed interactions?

3. Is it possible to give a reliability score for an individual observed interaction?

3.1 Estimating the reliability of putative protein interactions based on gene expressions

There is no consensus choosing the gold standard set of true protein interactions. Mrowka et al. (2001) used MIPS physical interactions (excluding those from high-throughput experiments) as the gold standard. They used a bootstrap method to count how many random pairs need to be added to the reference data such that it has the same statistical behavior of gene expression correlation coefficients as that of the observed protein-protein interaction data, and then estimate the reliability using the sampling data. On the other hand, Deane et al. (2002) used INT, a subset of DIP interactions which are derived from small-scale experiments, as the gold standard for real interactions. They formalized the above idea assuming that the distribution of the square of Euclidian distance between expression profiles of putative interacting pairs is a mixture of that for the real interacting pairs and that of random pairs. They then used a least square approach to estimate the reliability of the putative protein interaction data. Deng et al. (2003) further extended the idea in Deane et al. (2002) and used a maximum likelihood estimation (MLE) approach to estimate the reliability of a putative interaction data set. Similar to Mrowka et al. (2001), they used MIPS physical interactions as a reference set for true interactions. The same approach can be applied to estimate the fraction of protein pairs that belong to the same complex in an observed complex data set. The method can be briefly described as follows.

Let α be the reliability of a given set of putative protein interactions. Let $O_e(\cdot)$, $T_e(\cdot)$ and $R_e(\cdot)$ be the distribution of the correlation coefficients for gene pairs based on gene expressions for the given set of putative protein interactions, the true protein interaction set, and the random protein pairs, respectively. Then

$$O_e(\cdot) = \alpha T_e(\cdot) + (1-\alpha) R_e(\cdot). \tag{1}$$

$T_e(\cdot)$ and $R_e(\cdot)$ can be approximated based on the correlation coefficients for pairs of proteins within the golden standard set of protein interactions and the correlation coefficients of all the protein pairs, respectively.

Deng et al. (2003) split the values of correlation coefficients into $K = 20$ bins. Let n_k be the number of observed interaction pairs in the k-th bin. Let p_k and q_k be the fractions of real interactions and random pairs in the k-th bin, respectively. Then the likelihood function can be defined as:

$$L(\alpha) = \prod_{k=1}^{K} (\alpha p_k + (1-\alpha) q_k)^{n_k}. \tag{2}$$

$L(\alpha)$ is a convex function and a classical gradient algorithm can be used to estimate the parameter α, $\widehat{\alpha}$, by maximizing $L(\alpha)$.

The following equation was used to calculate the variance of $\widehat{\alpha}$,

$$\mathtt{Var}(\widehat{\alpha}) = \left(\sum_{k=1}^{K} n_k \frac{(p_k - q_k)^2}{(\widehat{\alpha} p_k + (1-\widehat{\alpha}) q_k)^2} \right)^{-1}. \tag{3}$$

3.2 Estimating the reliability of putative protein interactions based on gene expressions and protein localizations

Huh et al. (2003) generated a large-scale protein localization map of yeast and showed that protein interactions are strongly enriched among co-localized proteins and proteins between specific cellular locations. Therefore we can use both gene expressions and localizations for reliability estimation. Again we model the putative interaction data set as a mixture of true interactions and random pairs. Let $\theta_{ll'}$ and $\delta_{ll'}$ be the probability that a true interacting pairs and random protein pair belong to locations (l, l'), respectively. Let $n_{kll'}$ be the number of observed protein pairs within the putative interaction data set with correlation coefficient in the k-th bin and with localizations (l, l'). Combining gene expression data and protein localization data results in the following likelihood function

$$L(\alpha) = \prod_{k=1}^{K} \prod_{l,l'=1}^{L_0} (\alpha p_k \theta_{ll'} + (1-\alpha) q_k \delta_{ll'})^{n_{kll'}}, \tag{4}$$

where L_0 is the number of locations being considered. α can again be estimated by maximizing $L(\alpha)$.

The following equation was used to calculate the variance of $\widehat{\alpha}$,

$$\mathtt{Var}(\widehat{\alpha}) = \left(\sum_{k=1}^{K} \sum_{l,l'=1}^{L_0} n_{kll'} \frac{(p_k \theta_{ll'} - q_k \delta_{ll'})^2}{(\widehat{\alpha} p_k \theta_{ll'} + (1-\widehat{\alpha}) q_k \delta_{ll'})^2} \right)^{-1}. \tag{5}$$

3.3 Applications to protein interactions from high throughput experiments

We applied the above methods to protein interaction data sets from several high throughput experiments. Two groups of interaction data sets were studied (Lee et al. 2005). The first group includes pairwise physical interactions including the MIPS, DIP, Uetz's (Uetz et al. 2000) and Ito's (Ito et al. 2000; Ito et al. 2001) interaction data sets. The Ito$i$IST indicates the set of protein pairs that are observed to interact i times. The MIPS physical interactions are used as a true interaction data set. The estimated reliability together with their standard deviations of the estimates using gene expressions and protein localizations alone or combined are given in Table 1.

The second group includes the protein complexes such as the MIPS complex data, the TAP complex data, and the HMS-PCI complex data. Any pair of proteins within the same complex are considered interacting. We treat the MIPS complex data as a true protein complex data set. Table 1 gives the estimated reliability and the corresponding standard deviation for the various protein complex data. The standard deviation of the estimate using gene expression alone is very large with the estimated reliability showing irregular patterns. For example, the estimated reliability for Ito4IST (0.895) is much higher than the estimated reliability of Ito6IST (0.676) contradicting with our intuition. The standard deviation of the estimated reliability using localization alone is much smaller and the estimated reliability for Ito$i$IST increases as i increases consistent with our intuition. Finally the standard deviation of the estimate based on the combined data is smaller than that using gene expressions or protein localizations alone.

3.4 Estimating the probability of interaction for individual protein pairs

The above approach can only estimate the fraction of true interactions in a putative interaction data set. However, it does not give a reliability score for a particular observed interaction. Saito et al. (Saito et al. 2003) proposed the criterion "interaction generality" to assess the reliability of a particular interaction protein pair based on the idea that a protein cannot interact with too many interacting partners. If a protein interact with a large number of proteins, it is most likely

	Localization		Gene Expression		Both	
Data	Reliability	Standard Err.	Reliability	Standard Err.	Reliability	Standard Err.
			Physical Interactions			
DIP	0.587	0.0082	0.815	0.0244	0.619	0.0076
Uetz	0.685	0.0273	0.529	0.0843	0.699	0.0257
Ito1IST	0.268	0.0140	0.167	0.0383	0.293	0.0133
Ito2IST	0.411	0.0259	0.558	0.0831	0.470	0.0253
Ito3IST	0.532	0.0345	0.753	0.1144	0.611	0.0321
Ito4IST	0.552	0.0397	0.895	0.1436	0.640	0.0366
Ito5IST	0.547	0.0429	0.964	0.1567	0.640	0.0394
Ito6IST	0.556	0.0491	0.676	0.1768	0.641	0.0451
Ito7IST	0.608	0.0544	0.791	0.1942	0.682	0.0492
Ito8IST	0.614	0.0572	0.878	0.2054	0.684	0.0514
			Complexes			
TAP	0.4544	0.0063	0.585	0.0081	0.516	0.0056
HMS-PCI	0.1975	0.0042	0.248	0.0053	0.205	0.0037

Table 1: Reliability of the protein physical interaction data (Uetz's, DIP, and Ito's with different IST hits), and the protein complex data (the TAP and the HMS-PCI) using the protein localization data, the gene expression data and both data sets.

a "stick" protein and the observed interactions associated with this protein does not have real functional associations. Recently, Troyanskaya et al. (2003) and Jansen et al. (2003) developed Bayesian approaches to give a reliability measure for a particular putative interaction based on the observations that interacting protein pairs are more likely to have similar functions, to have similar gene expression patterns, and to be in the same location. Troyanskaya et al. (2003) gave a reliability score for two proteins to be functionally related and Jansen et al. (2003) gave a reliability score for two proteins to be in the same complex. In a more recent paper, (Jaimovich et al. 2005) proposed a Markov random field (MRF) model for predicting protein interactions. They assumed a MRF model for the interaction network based on the theory of random graphs (Frank and Strauss 1986). Conditional on the true interaction network, they assumed probability models for the observed data. Machine learning approaches were used to estimate the parameters as well as to predict the posterior probability of interactions for protein pairs conditional on the observations from different data sources. More details can be found in (Jaimovich et al. 2005).

4 Protein function prediction using protein interaction data

It has been observed that interacting proteins are more likely to have similar functions (Mering et al. 2002). Therefore, protein interaction networks can be useful for protein function prediction. For a given protein, all the proteins interacting with the given protein form its neighbors. Fellenberg et al. (Fellenberg et al. 2000) and Schwikowski et al. (Schwikowski et al. 2000) developed a neighbor counting method for protein function prediction. For an unknown protein, they counted the number of known proteins of its neighbors for each function of interest and assigned the unknown protein with the function category having the highest frequency. One problem with this approach is that it does not consider the frequency of the proteins having certain functions of interest. Hishigaki et al. (Hishigaki et al. 2001) developed a χ^2-statistic based approach for protein function prediction. For an unknown protein and a function of interest, a χ^2-statistic is calculated by comparing the observed frequency with the expected frequency of neighbors having the function of interest. The unknown protein is assigned the function with the highest χ^2 statistic. Both the counting method and the χ^2 method do not consider unknown protein neighbors. Several novel methods have been developed for protein function prediction based on interaction networks and other data sources. In this section we review these approaches.

Suppose a genome has N proteins $P_1, \cdots, P_N$. Let $P_1, \cdots, P_n$ be the unknown proteins and $P_{n+1}, \cdots, P_{n+m}$ be the known proteins, $N = n+m$. A protein may have several different functions. To simplify the problem, we study each functional category separately. For a function of interest, let $X_i = 1$ if the i-th protein has the function and 0 otherwise. The problem is to assign values to $X = (X_1, \cdots, X_n)$ conditional on the protein interaction networks, other pairwise relationships, features of individual proteins, and the functions of the known proteins.

4.1 A Markov Random Field (MRF) model for protein function prediction

Based on the idea of guilty-by-association, Deng et al. (Deng et al. 2002) first developed a MRF model for protein function prediction. The basic idea is to assign a prior probability for $X =$

$(X_1, \cdots, X_{n+m})$, the configuration of function labelling based on the protein interaction network. Under this model, they calculated the posterior probability distribution for $(X_1, \cdots, X_n)$ conditional on the network and $(X_{n+1}, \cdots, X_{n+m})$. The key is how to assign the prior probability distribution. Different priors give different accuracy for protein function prediction.

4.1.1 A MRF model based on one network

In (Deng et al. 2002), they assigned the prior as follows. Let π be the probability of a protein having the function of interest. Without considering the interaction network, the probability of a configuration of X is proportional to

$$\prod_{i=1}^{N} \pi^{x_i}(1-\pi)^{1-x_i} = \left(\frac{\pi}{1-\pi}\right)^{N_1}(1-\pi)^N, \tag{6}$$

where $N_1 = \sum_{i=1}^{N} x_i$.

Deng et al. (Deng et al. 2002) then considered one interaction network. Let S denote all the interacting protein pairs. The probability of the functional labelling conditional on the network is proportional to

$$\exp(\beta N_{01} + \gamma N_{11} + \kappa N_{00}), \tag{7}$$

where $N_{ll'}$ is the number of (l, l')-interacting pairs in S, and

$$\begin{aligned}
N_{11} &= \sum_{(i,j)\in S} x_i x_j \\
&= \#\{(1 \leftrightarrow 1) \text{ pairs in S}\}, \\
N_{10} &= \sum_{(i,j)\in S} (1-x_i)x_j + (1-x_j)x_i \\
&= \#\{(1 \leftrightarrow 0) \text{ pairs in S}\}, \text{and} \\
N_{00} &= \sum_{(i,j)\in S} (1-x_i)(1-x_j) \\
&= \#\{(0 \leftrightarrow 0) \text{ pairs in S}\}.
\end{aligned} \tag{8}$$

Therefore, the total probability of the functional labelling is proportional to $\exp(-U(x))$, where

$$\begin{aligned} U(x) &= -\alpha N_1 - \beta N_{10} - \gamma N_{11} - \kappa N_{00} \\ &= -\alpha \sum_{i=1}^{N} x_i - \beta \sum_{(i,j)\in S} x_i x_j \\ &\quad - \gamma \sum_{(i,j)\in S} (1-x_i)x_j + (1-x_j)x_i \\ &\quad - \kappa \sum_{(i,j)\in S} (1-x_i)(1-x_j), \end{aligned} \tag{9}$$

and $\alpha = \log(\frac{\pi}{1-\pi})$.

$U(x)$ is referred as the *potential function* in the field of MRF and defines a global Gibbs distribution of the entire network,

$$\Pr(X \mid \theta) = \frac{1}{Z(\theta)} \exp(-U(x)), \tag{10}$$

where $\theta = (\alpha, \beta, \gamma, \kappa)$ are parameters and $Z(\theta)$ is a normalized constant calculated by summing over all the configurations:

$$Z(\theta) = \sum_{x} \exp(-U(x)).$$

$Z(\theta)$ is called the partition function.

Several other approaches for protein function prediction based on one interaction network have been developed. In particular, Vazquez et al. (2003) considered multiple function categories and proposed to maximize the number of interactions within the same function categories. For one function of interest, it is equivalent to maximize

$$N_{00} + N_{11}$$

where N_{00} and N_{11} are defined as above. The (Deng et al. 2002) model differs from the (Vazquez et al. 2003) model in two significant ways. (1) Vazquez et al. (2003) used only the interaction network and did not consider the fraction of proteins having the function of interest in the known proteins. (2) Vazquez et al. (2003) gave an equal weight to intra-function class interactions. Letovsky and Kasif (Letovsky and Kasif 2003) proposed a model to assign functions to proteins based on a probabilistic analysis of graph neighborhoods in a protein-protein interaction network, which is fundamentally a MRF model, and the belief propagation algorithm was used to assign function probabilities for proteins in the network.

4.1.2 A Markov Random Field (MRF) model for multiple networks

Deng et al. (Deng et al. 2003b) further extended the above model to multiple networks and to include features of individual proteins. Assume that L sources of protein pairwise relationships that may be useful for protein function prediction are available. A network can be built based on each pairwise relationship denoted as $\mathrm{Net}_1,\ \mathrm{Net}_2, \cdots, \mathrm{Net}_L$, respectively. The entire network we consider is the union of all the networks denoted as S.

Similar to equation (7), our belief for the functional labelling of all the proteins based on network Net_l is proportional to

$$P\{\text{ labelling } |\mathrm{Net}_l\} \propto \exp(\beta_l N_{10}^{(l)} + \gamma_l N_{11}^{(l)} + \kappa_l N_{00}^{(l)}), \tag{11}$$

where $(N_{10}^{(l)}, N_{11}^{(l)}, N_{00}^{(l)})$ are defined similarly as equation (8).

Multiplying over all the networks, our belief for the functional labelling of all the proteins is proportional to

$$\begin{aligned} P\{\text{ labelling } |\text{networks }\} &\propto \prod_{l=1}^{L} \exp(\beta_l N_{10}^{(l)} + \gamma_l N_{11}^{(l)} + \kappa_l N_{00}^{(l)}) \\ &= \exp \sum_{l=1}^{L} \left(\beta_l N_{10}^{(l)} + \gamma_l N_{11}^{(l)} + \kappa_l N_{00}^{(l)}\right). \end{aligned} \tag{12}$$

Our total belief for the functional labelling of all the proteins is proportional to the multiplication of equations (6) and (12).

Then an MRF over all the functional labelling is defined by

$$P\{\text{labelling, networks}\} = \exp(-U(x))/Z(\theta), \tag{13}$$

where

$$U(x) = -\sum_{i=1}^{n+m} x_i \alpha - \sum_{l=1}^{L} \left(\beta_l N_{10}^{(l)} + \gamma_l N_{11}^{(l)} + \kappa_l N_{00}^{(l)}\right), \tag{14}$$

θ indicates the vector of parameters, and $Z(\theta)$ is the summation of $\exp(-U(x))$ over all the functional labelling. Under the above model, all the parameters $(\kappa_1, \kappa_2, \cdots, \kappa_L)$ are redundant and are set to 1. In the terminology of MRF, $U(x)$ is called the potential function.

4.1.3 Incorporating features of individual proteins

In addition to protein pairwise relationships, features of individual proteins can be very important for protein function prediction. A feature refers to an observation about a protein. It can be the presence or absence of a motif signal, the protein's conservation and localization, the protein's isoelectric point, its absolute mRNA expression level, or mutant phenotypes from experiments about the sensitivity or resistance of disruption mutants under various growth conditions. Several investigators have developed protein function prediction methods based on features of individual proteins (Clare et al. 2002; Gupta et al. 2002; Hegyi et al. 1999; Jensen et al. 2002; Kell et al. 2000; King et al. 2001; Stawiki et al. 2002; Drawid et al. 2000). Deng et al. (Deng et al. 2003b) integrated features into the MRF models for protein function prediction.

Suppose we have M features of interest, $F_1, F_2, \cdots, F_M$. The m-th feature can take values $0, 1, 2, \cdots k_m - 1$ where k_m is the number of categories for the m-th feature. Let the feature vector corresponding to protein P_i be $f_i = (f_{i1}, f_{i2}, \cdots, f_{iM})$, where f_{im} is the index for the m-th feature of the i-th protein. For the m-th feature, let $p_{1m}(k)$ ($p_{0m}(k)$) be the conditional probability that a protein has feature index k given that a protein has (does not have) the function of interest. For simplicity, we assume that all the features contribute independently to the functions of proteins.

For a given feature vector $f = (f_1, f_2, \cdots, f_M)$, define

$$P_1(f) = \prod_{m=1}^{M} p_{1m}(f_m),$$
$$P_0(f) = \prod_{m=1}^{M} p_{0m}(f_m).$$

The probability of the features of all the proteins given the functional labelling is

$$P\{\text{features} \mid \text{labelling}\} = \prod_{i:X_i=1} P_1(f_i) \times \prod_{i:X_i=0} P_0(f_i). \tag{15}$$

Multiplying equations (13) and (15), we have the following probability model

$$\begin{aligned} P\{\text{labelling, networks, domain features}\} = \\ P\{\text{labelling, networks}\} \times P\{\text{domain features} \mid \text{labelling}\}. \end{aligned} \tag{16}$$

Deng et al. (Deng et al. 2003b) described methods to estimate the posterior distribution of the functions of the unknown proteins given the features of all the proteins, the different sources of protein pairwise relationship, and the annotations of the known proteins.

4.1.4 Computational Issues

Given the above models, the problem is to estimate the posterior probability distribution given the annotation of the known proteins, the features of all the proteins, and the network. The parameters are also unknown. Using equation (16), it can be shown that

$$\begin{aligned}&\log \frac{Pr(X_i = 1 \mid F, X_{[-i]}, \theta)}{1 - Pr(X_i = 1 \mid F, X_{[-i]}, \theta)}\\ =&\alpha_i + \sum_{l=1}^{L} (\beta_l - 1) M_0^{(i)}(l) + (\gamma_l - \beta_l) M_1^{(i)}(l),\end{aligned} \tag{17}$$

where F is the feature information for all the proteins, $X_{[-i]} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_{n+m})$, $\alpha_i = \log \frac{\pi P_1(f_i)}{(1-\pi)P_0(f_i)}$, $M_0^{(i)}(l)$ and $M_1^{(i)}(l)$ are the numbers of neighbors of protein P_i labelled with 0 and 1 according to the l-th network, respectively. The parameters can be estimated based on the network consisting of the known proteins using the pseudo-likelihood idea (Li 1995) by an S-plus routine (Venables et al. 1996) using equation (17).

Once all the parameters have been defined, Gibbs sampler (Liu 2001) can be used to estimate the posterior probability distribution of $(X_1, \cdots, \cdots, X_n)$. The algorithm can be described as follows:

1. Randomly set the value of missing data $X_i = \lambda_i, i = 1, \cdots, n$ with probability π.

2. For each protein P_i, update the value of X_i using equation (17).

3. Repeat step 2 T times until all the posterior probabilities $\Pr(X_i \mid D,\ X_{[-i]}, \theta)$ are stabilized.

4.2 Kernel-based methods for protein function prediction

In the MRF formulation, we only consider immediate neighbors for proteins. The protein interaction network can be used to define similarity between any pair of proteins using the diffusion kernel (Kondor and Lafferty 2002). In the following we first briefly describe kernel based methods of Lanckriet et al. (Lanckriet et al. 2004a; Lanckriet et al. 2004b; Lanckriet et al. 2004c) to combine different data sources for protein function prediction. Then we describe our effort to combine the idea of kernel based method with the MRF model.

4.2.1 Support vector machine (SVM) and semidefinite programming (SDP)

In a series of recent papers, Lanckriet et al. (Lanckriet et al. 2004a; Lanckriet et al. 2004b; Lanckriet et al. 2004c) developed kernel-based methods for protein function prediction using SVM. Suppose that there are L data sources such as protein interactions, gene expressions, domains, localizations, etc. For the l-th data source, a kernel matrix K_l (semi-positive definite) is defined. For continuous data such as gene expressions, the Gaussian diffusion kernel can be used. For protein interactions, diffusion kernel on graphs can be used (Kondor and Lafferty 2002). Several other kernel matrixes have been developed for different sources of data structures in (Lanckriet et al. 2004a; Lanckriet et al. 2004b; Lanckriet et al. 2004c). To integrate different data sources, Lanckriet and colleagues considered the linear combinations of the kernel matrixes

$$K = \sum_{l=1}^{L} \mu_l K_l$$

where $\mu_l \geq 0$, $l = 1, 2, \cdots$ are parameters to be determined.

They used SVM with 1-norm soft margin to build a classifier. The problem can then be solved by solving the following constraint maximization problem:

$$\begin{aligned} &\max_{\alpha,t} 2\alpha^T e - ct \\ \text{subject to } &t \geq \frac{1}{r_i}\alpha^T \text{diag}(y) K_l \text{diag}(y)\alpha, \quad l = 1, 2, \cdots, L \\ &\alpha^T y = 0, \\ &C \geq \alpha \geq 0, \end{aligned} \tag{18}$$

where $r_i = \text{trace}(K_i)$, $c = \mu^T r$ and y is the annotation of the known proteins. This problem is a quadratically constraint quadratic program (QCQP) problem (Boyd and Vandenberghe 2001) and can be solved using standard software such as SeDuMi (Sturm 1999). The computational time is $O(n^3)$, where n is the number of proteins in the training set.

4.2.2 Combining kernel with the MRF model for protein function prediction

Lanckriet et al. (2004a) showed that SVM described above outperformed the MRF approach in almost all the function categories considered. One of the main reasons probably is due to the inclusion of multiple level neighbors in the kernel based methods. Note that $K_l(i,j)$ defines a similarity between protein P_i and protein P_j based on the l-th data source. Similar to equation (11), the probability of the labelling based on the l-th network N_l can be modelled as

$$\exp(\beta_l D_{10}(l) + \gamma_l D_{11}(l) + \kappa_l D_{00}(l)) \tag{19}$$

where β_l, γ_l, and κ_l are constants, and

$$\begin{aligned} D_{11}(l) &= \sum_{i<j} K_l(i,j) I\{x_i = 1, x_j = 1\}, \\ D_{10}(l) &= \sum_{i<j} K_l(i,j) I\{(x_i = 1, x_j = 0) \text{ or } (x_i = 0, x_j = 1)\}, \\ D_{00}(l) &= \sum_{i<j} K_l(i,j) I\{x_i = 0, x_j = 0\}. \end{aligned} \tag{20}$$

The summations are over all the protein pairs. Multiplying equation (6) and equation (19) for $l = 1, 2, \cdots, L$, we obtain the the total probability proportional to

$$\exp\left(\alpha N_1 + \sum_{l=1}^{L} (\beta_l D_{10}(l) + \gamma_l D_{11}(l) + \kappa_l D_{00}(l))\right) \tag{21}$$

From equation (21), it can be shown that

$$\begin{aligned} &\log \frac{\Pr(X_i = 1 \,|X_{[-i]}, \theta)}{1 - \Pr(X_i = 1 \,|X_{[-i]}, \theta)} \\ =&\alpha + (\beta_l - \kappa_l) K_0^{(i)}(l) + (\gamma_l - \beta_l) K_1^{(i)}(l). \end{aligned} \tag{22}$$

where

$$\begin{aligned} K_0^{(i)}(l) &= \sum_{j \neq i} K_l(i,j) I\{x_j = 0\}, \\ K_1^{(i)}(l) &= \sum_{j \neq i} K_l(i,j) I\{x_j = 1\}. \end{aligned}$$

Note that if we let $K_l(i,j) = 1$ when protein i interacts with protein j and $K_l(i,j) = 0$ otherwise in the l-th network, this new model is the same as the MRF model of Deng et al. (Deng et al. 2002). We can similarly develop a MCMC approach to approximate the probability that an unknown protein having the function of interest. We refer the above approach as kernel-based MRF (KMRF)

4.3 Applications to real data

All the methods described above have been applied to predict protein functions. The MRF model has been used for protein function prediction first based on the MIPS function classification (Deng et al. 2002; Deng et al. 2003b) and later were extended to functions defined in GO (Deng et al. 2004). The SVM approach has been used to predict protein functions based on MIPS (Lanckriet et al. 2004a), to predict ribosomal proteins and memberane proteins (Lanckriet et al. 2004c). A summary paper for protein function prediction based on SVM is given in (Lanckriet et al. 2004b). The new KMRF method has been applied for protein function prediction based on GO (Lee et al. 2005b) and for prediction of protein essentiality (Tu et al. 2005). The KMRF approach can be easily extended to incorporate correlated functions. For most functions that have been considered so far, the SVM approach outperformed the MRF approach. The KMRF approach has similar performance as the SVM approach. For example, for predicting protein essentiality, the receiver operating characteristic (ROC) scores for the MRF, SVM, and the KMRF approaches are 0.804. 0.812, and 0.831, respectively, based on the core interaction data set. Integrating protein function based on cellular processes, conservation, and localizations into the model increased the ROC score of the KMRF model to 0.869.

5 Discussion

Enormous amount of biological data have been generated and stored in public and private databases. These data sources are extremely important for biological studies. However the data are generally noisy and contain many false positive and false negative errors. There are no systematic statistical tools to choose the most reliable data from the noisy data. The various data sources can most likely contribute to our understanding of the biological problems of interest. However the data sources are usually correlated and their contributions to our understanding of the biological problems are not independent. An important issue is how to integrate the usually noisy and correlated data sources to understand the biological problems.

In this paper, we review our recent efforts in integrating different data sources for biological studies. First we describe likelihood based methods for estimating the reliability of putative interaction data sets. We show that the localization data give more accurate estimation of the reliability than using the gene expression data. Integrating the localization and gene expression data can give even more accurate estimates of the reliability of the different data sets. Other statistical methods for estimating the probability of two proteins being interact integrating different data sources have also been developed.

Second we describe methods for protein function prediction based on interaction networks, genetic interactions, other pairwise relationships, as well as features of individual proteins. These approaches include MRF, SVM, and KMRF. As far as we know, the combination of kernels with MRF is novel in protein function prediction. The simplicity of KMRF and its high accuracy in protein function prediction warrant further studies of the this approach in other fields.

In protein function prediction, we assume implicitly that the networks under consideration, such as the protein interaction network and genetic interaction network, are highly reliable. Therefore we used the core interaction data in DIP in all our studies on protein function prediction. We tried to use all the interactions (not reliable) in DIP for protein function prediction and, as expected, the prediction accuracy is lowered. A problem is how best to use all the interactions for protein function prediction. The effect of incompleteness of the interaction data on protein function prediction is also unknown.

In summary, we show the power of integrating multiple data sources for biological studies. Significant questions remain as to how to integrate noisy and incomplete data in biological studies. It is also important to develop methods to evaluate the dependence among the different data sources and to integrate the correlated data sources for biological studies.

Acknowledgments

The research was partly supported by NIH/NSF joint mathematical biology initiative DMS-0241102.

References

Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Research, 31:248-250.

Breitkreutz, B.J., Stark, C. and Tyers M. (2003). The GRID: The General Repository for Interaction Datasets. Genome Biology, 4:R23.

Clare, A. and King, R.D. 2002. Machine learning of functional class from phenotype data. Bioinformatics, 18:160-166.

Deane, C.M., Salwinski, L., Xenarios, I. and Eisenberg, I. (2002). Protein interactions: Two methods for assessment of the reliability of high-throughput observation. Molecular and cellular proteomics, 1:349-356.

Deng, M., Zhang, K., Mehta, S., Chen, T. and Sun, F.Z. (2002). Prediction of protein function using protein-protein interaction data. In Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB2002):197-206.

Deng, M.H., Sun, F.Z., and Chen, T. (2003). Assessment of the reliability of protein-protein interactions and protein function prediction. Pacific Symposium of Biocomputing (PSB2003):140-151.

Deng, M.H., Chen, T. and Sun, F.Z. (2003b). An integrated probabilistic model for functional prediction of proteins. In Proceedings of the Seventh International Conference on Computational Molecular Biology (RECOMB2003):95-103.

Deng, M.H., Tu, Z.D., Sun, F.Z. and Chen T (2004). Mapping gene ontology to proteins based on protein-protein interaction data. Bioinformatics 20:895-902.

Drawid, A. and Gerstein, M. (2000). A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. Journal of Molecular Biology, 301:1059-1075.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Bostein D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences USA, 95:14863-14868.

Enright, A.J., Iliopoulos, I., Kyrpides N.C. and Ouzounis, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature, 402:86-90.

Fellenberg, M., Albermann, K., Zollner, A., Mewes, H.W. and Hani, J. (2000). Integrative analysis of protein ineraction data. In Proc. of the Eighth Int. Conf. on Intelligent System for Molecular Biology (ISMB2000):152-161.

Frank O., Strauss D. (1986). Markov graphs. Journal of American Statistical Association, 81:832-842.

Gavin, A., Böche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A., Cruciat, C. et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415:141-147.

The Gene Ontology Consortium. (2001). Creating the gene ontology resource: design and implementation. Genome Research, 11:1425-1433.

Gupta, R. and Brunak, S. (2002). Prediction of glycosylation across the human proteome and the correlation to protein function. Pacific Symposium of Biocomputing (PSB2002):310-322.

Hegyi, H. and Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to yeast genome. Journal of Molecular Biology, 288:147-164.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein-protein interaction data. Yeast, 18:523-531.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K., and Boutilier, K., et al. (2002). Systematic identification of protein complexes in *Saccharomyces Cerevisiae* by mass spectrometry. Nature, 415:180-183.

Huh, W.K., Falvo, J.V., Gerke, L.C. Carroll, A.S., Howson, R.W., Weissman, J.S. and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature, 425:686-691.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. and Sakaki, Y. (2000). Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proceedings of the National Academy of Sciences USA, 97:1143-1147.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001). A comprehensive two hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences USA, 98:4569-4574.

Jaimovich, A., Elidan, G., Margalit, H. and Friedman, N. (2005). Towards an integrated protein-protein interaction network. In Proceedings of the Ninth International Conference on Computational Molecular Biology (RECOMB2005):14-30.

Jansen, R., Yu, H.Y., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S.B., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 302:449-453.

Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Stærfeldt, H.H, Rapacki,K., and Workman, C., et al. (2002). Prediction of human protein function from post-translational modifications and localization features. Journal of Molecular Biology, 319:1257-1265.

Kell, D.B. and King, R.D. (2000). On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. Trends Biotechnology, 18:93-98.

King, R.D., Karwath, A., Clare, A. and Dehaspe, L. (2001). The utility of different representations of protein sequence for predicting functional class. Bioinformatics, 17:445-454.

Kondor, R.I. and Lafferty, J. (2002). Diffusion kernels on graphs and other discrete inpute spaces. In Proceedings of International Conference on Machine Learning, 315-322.

Lanckriet, G.R.G., Deng, M.H., Cristianini, N., Jordan, M.I. and Noble, W.S. (2004a). Kernel-based data fusion and its appliation to protein function prediction in yeast. Pacific Symposium on Biocomputing (PSB2004):300-311.

Lanckriet, G.R.G., Cristianini, N., Jordan, M.I. and Noble, W.S. (2004b). Kernel-based integration of genomic data using semidefinite programming. In Schölkopf B, Tsuda K, and Vert J.-P (eds) Kernel Methods in Computational Biology, MIT press, Cambridge, MA pp. 71-92.

Lanckriet, G.R.G., Bie, T.D., Cristianini, N., Jordan, M.I. and Noble, W.S. (2004c). A statistical framework for genomic data fusion. Bioinformatics, 20:2626-2635.

Lee, H.J., Deng, M.H., Sun, F.Z. and Chen, T. (2005a). Assessment of the reliability of protein-protein interactions using protein localization and gene expression data. Technical Report.

Lee, H.J., Tu, Z.D., Deng, M.H., Sun, F.Z. and Chen, T. (2005b). Diffusion kernel based logistic regression models for protein function prediction. Technical Report.

Letovsky, S. and Kasif, S. (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics, 19 (Suppl.1):197-204.

Li, S.Z. (1995). Markov random field modeling in Computer vision. Springer-Verlag, Tokyo.

Liu, J.S. (2001). Monte Carlo Strategies in Scientific Computing. Springer-Verlag, New York.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice,D.W., Yeates, T.O. and Eisenberg, D. (1999a). Detecting protein function and protein-protein interactions from genome sequences. Science, 285:751-753.

Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. Nature, 402:83-86.

Mering, C.V., Krause, R. Snel, M., Oliver, S.G., Fields, S. and Bork, P. (2002). Comparative assessment of large scale data sets of protein-protein interactions. Nature, 417:399-403.

Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B. (2002). MIPS: a database for genomes and protein sequences. Nucleic Acids Research, 30:31 - 34.

Mrowka, R., Patzak, A., and Herzel, H. (2001). Is there a bias in proteome research? Genome Research, 11:1971-1973

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (2000). The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences USA, 96:2896-2901.

Pavlidis, P. and Weston, J. (2001). Gene functional classification from heterogeneous data. In Proceedings of the Fifth International Conference on Computational Molecular Biology (RECOMB2001):249 - 255.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences USA, 96:4285-4288.

Saito, R., Suzuki, H. and Hayashizaki, Y. (2003). Construction of reliable protein-protein interaction networks with a new interaction generality measure. Bioinformatics, 19:756-763.

Schwikowski, B., Uetz, P. and Fields, S. (2000). A network of protein-protein interactions in yeast. Nature Biotechnology, 18:1257-1261.

Stawiki, E.W., Mandel-Gutfreund, Y., Lowenthal, A.C. and Gregoret, L.M. (2002). Progress in predicting protein function from structure: unique features of O-Glycosidases. Pacific Symposium of Biocomputing (PSB2002):637-648.

Sturm, J.F. (1999) Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. Optimization Methods and Software, 11-12:625-653.

Tong, A.H.Y., Drees, B., Nardelli, G., Bader G.D., Brannetti, B., Castagnoli, L. Evangelista, M., Paoluzi, S., Quondam, M., Zucconim A, et al. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science, 295:321-324.

Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman,R.B. and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proceedings of the National Academy of Sciences USA, 100:8348-8353.

Tu, Z.D., Lee, H.J., Deng, M.H., Chen, T. and Sun, F.Z. (2005). Understanding protein essentiality - linking genomic information with phenotype. Technical Report.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, et al. (2000). A Comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae.* Nature, 403:623-627.

Vazquez, A., Flammini, A., Maritan, A. and Vespignani, A. (2003). Global protein function prediction from proteinCprotein interaction networks. Nature Biotechnology, 21:697-700.

Venables, W.N. and Ripley, B.D. (1996). Modern Applied Statistics with S$-$Plus. Springer-Verlag, New York.

Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S. and Eisenberg, D. (2002). DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. Nucleic Acids Research, 30:303-305.

Zheng, Y., Roberts R.J. and Kasif, S. (2003). Genomeic functional annotation using co-evolution profiles of gene clusters. Genome Biology, 3:1-9.