A LINEAR-TIME ALGORITHM FOR PREDICTING FUNCTIONAL ANNOTATIONS FROM PPI NETWORKS*

Yonghui Wu
Department of Computer Science and Engineering
University of California, Riverside CA 92521, USA
yonghui@cs.ucr.edu

Stefano Lonardi
Department of Computer Science and Engineering
University of California, Riverside CA 92521, USA
stelo@cs.ucr.edu

Recent proteome-wide screening efforts have made available genome-wide, high-throughput protein-protein interaction (PPI) maps for several model organisms. This has enabled the systematic analysis of PPI networks, which has become one of the primary challenges for the system biology community. Here we address the problem of predicting the functional classes of proteins (i.e., GO annotations) based solely on the structure of the PPI network. We present a maximum likelihood formulation of the problem and the corresponding learning and inference algorithms. The time complexity of both algorithms is linear in the size of the PPI network and our experimental results show that their accuracy in the functional prediction outperforms current existing methods.

Keywords: protein-protein interaction, protein function prediction, maximum likelihood estimation, Markov random field, gene ontology

1. Introduction

High-throughput protein-protein interaction (PPI) networks with various levels of proteome coverage are currently available for several model organisms, namely *S. cerevisiae*[20], *D. melanogaster*[8,7], *C.elegans*[13], *H. sapiens*[16] and *H. pylori*[15]. PPI data can be obtained through a variety of sophisticated assays, like co-immunoprecipitation, yeast two hybrid, tandem affinity purification and mass spectrometry. A PPI network is usually represented by a node-labeled undirected graph where vertices correspond to proteins and edges denote physical interactions.

Since the main mechanism by which cells are able to process information is through protein-protein interactions, PPI data has been essential to obtain new knowledge and insights in a wide spectrum of biological processes. In this paper, we focus on the problem of predicting the functional category of proteins *solely*

*A preliminary version of this work was presented at the *7th International Workshop on Data Mining in Bioinformatics*, San Jose CA, and included in its *Proceedings* (2007).

based on the topological structure of the PPI network. The rationale of this approach is based on the observation that a protein is much more likely to interact with another protein in the same functional class than with a protein with a different function[11,22,19,14]. The prediction of functional classes can be useful either for proteins for which there is little or non-existing functional information (e.g., for predicting the involvement of a protein in specific pathway), or to confirm existing annotations provided by other methods. Motivated by the expectation that in the near future massive PPI networks will be available, here we propose a *computationally efficient* method that accurately determines the functional categories and will be capable to scale gracefully with the size of the network.

A variety of algorithmic techniques have been proposed in the literature to solve the problem of functional prediction with a wide range of computational complexity. Perhaps the most computationally efficient algorithm is based on the *majority rule* where the function of an unknown protein is simply determined by the most common function among its interacting partners[18]. A slightly more sophisticated majority-based method is the χ^2-method[9]. At the other end of the computational complexity spectrum, Vazquez *et al.*[22,11] propose to assign proteins to functional classes so that the number of protein interactions among different functional categories is minimized. The optimization problem, known as *generalized multicut*, is NP complete[14].

The *functional flow* algorithm[14] lays somewhere in the middle of the complexity spectrum. The idea is to treat proteins with known function as infinite sources of (functional) flow. The flow is propagated through the network in a series of discrete steps. At the end, the function of unknown proteins is assigned based on the largest amount of flow received. Nabieva *et al.*[14] show that functional flow algorithm outperforms the generalized multicut algorithm, the majority rule-based algorithm and also its generalization to more distant neighbors[14]. Chua *et al.*[2] show that functional flow also outperforms the χ^2-method. Because of this, the performance of functional flow is the reference for our algorithm. Experimental results will show that our method achieves a better prediction accuracy than functional flow.

Perhaps the most similar method to the one we propose here is described in Deng *et al.*'s work[5,6], where the authors propose a probabilistic model based on the theory of Markov random fields. In their follow-up papers [3], Deng *et al.* show how to integrate in their Markov random field additional information, namely gene expression data, protein complex information, domain structures to increase the prediction accuracy. The relationship between this work and that of Deng *et al.*[5,6] will be discussed in greater detail later. Here, however, we want to emphasize that the method presented in this manuscript is computationally more efficient than Deng *et al.* Unfortunately, the accuracy of their prediction cannot be directly compared with ours because these methods predict multiple functional classes for each protein. The approach proposed by Letovsky *et al.*[12] is essentially similar to that of Deng *et al.*[6].

More recent papers tackle slightly different albeit related problems. Srinivasan

et al.[19] predict functional linkages between proteins based on the integration of four kinds of evidence, namely gene co-expression, gene co-inheritance, gene co-location and gene co-evolution. Jaimovich *et al.*[10] predict protein interactions based on the cellular localization of proteins.

2. Problem definition and model formulation

We denote by $G(V, E)$ the PPI network under analysis, where V represents the set of vertices(proteins) and E is the set of edges (interactions). For reason that will be clear later in the paper, we assume G to be directed (i.e., each undirected edge in the original PPI is represented by two directed edges of opposite directions, except for self-loops). We denote the set of k given functional classes as $\mathcal{F} = \{C_1, C_2, \ldots, C_k\}$. Each functional class can be thought as one of k possible colors that can be used to color the graph. Function $f : V \rightarrow \mathcal{F}$ captures the notion of functional class for all the proteins in V. When the function of a protein $v \in V$ is known, say C_i, then we will have $f(v) = C_i$. If the function of v is unknown, then $f(v) = \emptyset$. We define $W = \{v \in V : f(v) \in \mathcal{F}\}$ to be the set of proteins whose function is known and $U = V \setminus W$ to be the set of the proteins whose function is unknown. The functional annotation problem can be informally stated as follows. Given a PPI network $G(W \cup U, E)$ where W is annotated with functional classes, find the correct functional classes for the vertices in U.

The model used here to tackle the problem is entirely probabilistic and it is based on two simple observations. First, a simple statistical analysis on the available PPI data[17] and the associated GO functional annotations[1] reveals that the distribution associated with the functional classes is highly skewed. For example, in the *S. cerevisiae* network, the function "catalytic activity" is assigned to 1,514 proteins, whereas the function "protein tag" is only assigned to 5 proteins. This observation constitutes our prior knowledge on the probability of a randomly chosen protein to perform a certain function and can be captured by the notion of *prior distribution*. We denote the prior distribution by $\mathcal{P} : \mathcal{F} \rightarrow [0, 1]$, where $\mathcal{P}(C_i)$ is the probability of a randomly chosen protein to have function C_i.

Second, our model has to incorporate the connectivity structure of the PPI networks. As said, it is well-known that a protein is more likely to interact with another protein performing the same function[11,22,19,14]. We model this preference using conditional probability distributions. If protein $t \in W$ has function C_i and protein $s \in U$ interacts with t, then the probability that s performs function C_j is given by $\mathbf{P}(C_j|C_i)$. We expect $\mathbf{P}(C_i|C_i)$ to be higher than $\mathbf{P}(C_j|C_i), \forall j \neq i$, because s is more likely to perform the same function as t. This can be easily generalized to multiple interacting partners. Suppose we want to predict the function of protein $s \in U$ and that we know that $t_1, t_2, t_3, \ldots, t_m \in W$ interact with s, as well as their functions $f(t_1), f(t_2), f(t_3), \ldots, f(t_m)$. If we assume that $f(t_1), f(t_2), f(t_3), \ldots, f(t_m)$ are independent and distributed according to the conditional multinomial distribution $[\mathbf{P}(C_1|f(s)), \mathbf{P}(C_2|f(s)), \mathbf{P}(C_3|f(s)), \ldots, \mathbf{P}(C_K|f(s))]$, then the most likely

function for s is the one that maximizes

$$L(s) = \mathcal{P}(f(s)) \prod_{t\in\{t_1,t_2,\ldots,t_m\}} \mathbf{P}(f(t)|f(s)) = \mathcal{P}(f(s)) \prod_{t\in V:(s,t)\in E} \mathbf{P}(f(t)|f(s)).$$

We call $L(s)$ the *local likelihood* of protein s.

Note that a necessary condition to predict the functional class for $s \in U$ is to know the functional classes of the neighbors of s. Very often, however, the functions of the neighbors turn out to be unknown. Clearly, the assignment of a function to protein s may affect the prediction of the functions for the neighbors of s, and vice versa. Because of this, a purely local strategy is insufficient. To address this problem, we need to introduce the concept of *global likelihood* of a PPI Network as $L(G) = \prod_{v\in V} L(v)$.

The free variables in the global likelihood function $L(\cdot)$ are $f(u_i)$, for all proteins $u_i \in U$ with unknown function. We seek the assignment to $f(u_i)$ such that the global likelihood $L(G)$ is maximized, which is equivalent to maximizing

$$l(G) = \sum_{v\in V} \log(\mathcal{P}(f(v))) + \sum_{w\in V:(v,w)\in E} \log[\mathbf{P}(f(w)|f(v))]$$

Now we are ready to give a formal summary of the optimization problem associated with our model. We are given a directed PPI network $G(W \cup U, E)$ where U is the set of proteins with unknown functions and W is the set of proteins with known functions, a set of functions $\mathcal{F}$, a prior distribution $\mathcal{P}$ with $\sum_{C_i\in\mathcal{F}} \mathcal{P}(C_i) = 1$, and the conditional distributions $\mathbf{P}(C_i|C_j)$ such that $\sum_{C_i\in\mathcal{F}} \mathbf{P}(C_i|C_j) = 1, \forall C_j \in \mathcal{F}$. The problem is to predict the functional class $f(u)$ for each protein in set U, such that the global log likelihood $l(G)$ is maximized.

3. Relation to previous works

Our model implicitly defines a Markov random field (MRF), a probabilistic model which is also used in Deng *et al.*'s works[5,6]. In Deng *et al.*'s works[5,6], a distinct MRF is built for each functional class in $\mathcal{F}$. Each protein in the PPI network is associated to an indicator random variable for that function of interest. More specifically, each protein is associated with a unary potential $e^{\phi(X_i)}$, which has value $e^{\phi(1)}$ if the protein has that function and $e^{\phi(0)}$ otherwise. Each edge of the PPI graph is associated with a binary potential $e^{\psi(X_i,X_j)}$, which can take three possible values, namely $e^{\psi(1,1)}$ if both of the proteins have the function, $e^{\psi(0,1)}$ if one of the proteins has the function, and $e^{\psi(0,0)}$ if neither of the proteins has the function. Given the parameters $\theta = \{\phi(0), \phi(1), \psi(1,1), \psi(0,1), \psi(0,0)\}$, the global Gibbs distribution of the entire network is simply the product of the unary potentials and the binary potentials normalized by a constant factor depending on the parameters, as follows.

$$P\{X_1, X_2, X_3, \ldots, X_n|\theta\} = e^{\sum_{i=1}^{n} \phi(X_i) + \sum_{(i,j)\in E} \psi(X_i,X_j)}/Z(\theta)$$

Note that in our model, the prior probability $\mathcal{P}(f(v_i))$ corresponds to the unary potential in Deng's model, whereas the product $\mathbf{P}(f(v_i)|f(v_j))\mathbf{P}(f(v_j)|f(v_i))$ corresponds to the binary potential.

Despite the similarities, there are significant differences between Deng *et al.*'s model and ours. First, instead of building a distinct MRF for each function, we only have one unified probabilistic model for all the functions in $\mathcal{F}$ which allows us to capture the correlations between the functions. Second, the use of conditional distributions dramatically simplifies the process of estimating the parameters, which boils down to a simple count of relevant statistics (details to be explained in Section 4). The semantics of the conditional distributions also naturally give rise to the efficient iterative algorithm that we will develop later. Finally, since we are modeling from the conditional distributions, the normalization factor of the global Gibbs distribution in our model is always one irrespective of the parameters we use.

A less obvious connection can be established between our model and the generalized multicut approach by Vazquez *et al.*[22]. Recall that in this latter approach, the objective is to assign functional annotations to unknown proteins in such a way that one minimizes the number of times neighboring proteins have different annotations. A formal description of the generalized multicut problem follows. Let I be the standard indicator function which is equal to 1 if the boolean expression is true and 0 otherwise. Given a PPI network $G(U \cup W, E)$ we seek annotations to the proteins in U such that $\sum_{(u,v)\in E} I(f(u) \neq f(v))$ is minimized.

Fact 1. The generalized multicut problem is a special case of our optimization problem when the prior distribution is uniform and most of the mass of the conditional probabilities is concentrated around $\mathbf{P}(C_i|C_i)$.

Proof. Let us consider the following prior distribution and conditional distributions.

$$\begin{aligned} \mathcal{P}(C_i) &= 1/|\mathcal{F}| && \forall C_i \in \mathcal{F} \\ \mathbf{P}(C_j|C_i) &= \epsilon && \forall C_i, C_j \in \mathcal{F}, C_i \neq C_j \\ \mathbf{P}(C_i|C_i) &= 1 - (|\mathcal{F}| - 1)\epsilon && \forall C_i \in \mathcal{F} \end{aligned}$$

where $0 < \epsilon < 1$ is an arbitrarily small number.

Then, the global log likelihood for the graph can be written as

$$\begin{aligned} l(G(V,E)) &= \sum_{v\in V} \log(\mathcal{P}(f(v))) + \sum_{(v,w)\in E} \log(\mathbf{P}(f(w)|f(v))) \\ &= \sum_{v\in V} \log(1/|\mathcal{F}|) + \sum_{(v,w)\in E, f(w)\neq f(v)} \log(\mathbf{P}(f(w)|f(v))) \\ &\quad + \sum_{(v,w)\in E, f(w)=f(v)} \log(\mathbf{P}(f(w)|f(v))) \\ &= |V| \log(1/|\mathcal{F}|) + \sum_{(v,w)\in E, f(w)\neq f(v)} \log(\epsilon) \\ &\quad + \sum_{(v,w)\in E, f(w)=f(v)} \log(1 - (|\mathcal{F}| - 1)\epsilon) \end{aligned}$$

$$= |V| \log(1/|\mathcal{F}|) + |E| \log(1 - (|\mathcal{F}| - 1)\epsilon) + (\log(\epsilon) - \log(1 - (|\mathcal{F}| - 1)\epsilon)) \sum_{(v,w)\in E} I(f(v) \neq f(w)) \quad (1)$$

Note that the first two terms of (1) are constants and that the third term increases as the quantity $\sum_{(v,w)\in E} I(f(v) \neq f(w))$ decreases because $\log(\epsilon) - \log(1 - (|\mathcal{F}| - 1)\epsilon)$ is negative for a sufficiently small ϵ. Therefore, under this particular prior distribution and conditional distributions, maximizing the global log likelihood in our problem is equivalent to minimizing the objective function in the generalized multicut problem. □

The generalized multicut problem is NP complete[14] because it is a generalization of the multi-way cut problem[21], which is known to be NP complete. Since our problem is a generalization of the generalized multicut problem, it is NP complete as well.

4. Parameter learning

The prior distribution and the conditional distributions are multinomial distributions whose parameters can be learned from the structure of the give PPI and the functional annotations on W. We need to determine $k - 1$ parameters for the prior and $k(k - 1)$ parameters for the k conditional distributions. We obtain these parameters using the maximum likelihood estimation method.

Let $F(W, E')$ be the subgraph of $G(V, E)$ induced by the set W of known proteins, where $E' = \{(u, v)|(u, v) \in E, u \in W, v \in W\}$. The global likelihood for the subgraph $F(W, E')$ is defined as follows.

$$\begin{aligned} L(F(W, E')) &= \prod_{v\in W} \mathcal{P}(f(v)) \prod_{(u,v)\in E'} \mathbf{P}(f(u)|f(v)) \\ &= \prod_{C_i\in\mathcal{F}} \mathcal{P}(C_i)^{\sum_{v\in W} I(f(v)=C_i)} \\ &\quad \cdot \prod_{C_i\in\mathcal{F}} \prod_{C_j\in\mathcal{F}} \mathbf{P}(C_j|C_i)^{\sum_{(v_i,v_j)\in E'} I(f(v_i)=C_i, f(v_j)=C_j)} \end{aligned} \quad (2)$$

The first term in (2) is maximized when $\mathcal{P}(C_i) = \sum_{v\in W} I(f(v) = C_i)/|W|$ for all $C_i \in \mathcal{F}$. The second term in (2) is maximized when $\mathbf{P}(C_j|C_i) = \frac{\sum_{(v_i,v_j)\in E'} I(f(v_i)=C_i, f(v_j)=C_j)}{\sum_{(v_i,v_j)\in E'} I(f(v_i)=C_i)}$ for all $C_j \in \mathcal{F}$. Therefore, the maximum likelihood estimates for the parameters are

$$\mathcal{P}(C_i) = \sum_{v\in W} I(f(v) = C_i)/|W| \qquad C_i \in \mathcal{F}$$

$$\mathbf{P}(C_j|C_i) = \frac{\sum_{(v_i,v_j)\in E'} I(f(v_i) = C_i, f(v_j) = C_j)}{\sum_{(v_i,v_j)\in E'} I(f(v_i) = C_i)} \qquad C_i, C_j \in \mathcal{F}$$

As a common practice in Bayesian statistics, we apply (uniform) Dirichlet priors to our estimators. This prevents the problem of handling zero probabilities. The time complexity of the learning phase is $O(|E| + |W|)$, whereas the space complexity is $O(k^2)$.

5. Inference of functional classes

Since we determined that our problem is NP complete, it is rather unlikely that we will find a polynomial time algorithm that can solve the problem optimally. To this end, we designed a statistically based iterative algorithm (SBIA for short), which turns out to perform well in practice. Our algorithm consists of two phases, namely the initialization phase and the iterative phase. The initialization phase consists of two steps. In the first step, we estimate the parameters for the prior distribution and the conditional distributions as described in Section 4. In the second step, we assign an initial functional class to each protein in V, as follows.

For each unknown protein $v \in U$, we assign

$$f^0(v) = argmax_{C_i \in \mathcal{F}} \ \mathcal{P}(C_i) \prod_{(v,t) \in E, t \in W} \mathbf{P}(f^0(t)|C_i).$$

wherein $argmax$ is an operator that returns the optimal argument that maximizes the objective function that follows. In other words, we predict the initial function for v to be the one that maximizes the local likelihood of v (ignoring neighbors with unknown functions). If $v \in W$, then we set $f^0(v)$ to be the function corresponding to its annotation in the original data.

In the second phase, we iteratively re-evaluate our predictions. For clarity of exposition we use superscripts to denote the iteration number, i.e., $f^n(v)$ denotes the predicted functional class for v made in the n^{th} iteration. For each unknown protein $v \in U$, we set

$$f^n(v) = argmax_{C_i \in \mathcal{F}} \ \mathcal{P}(C_i) \prod_{(v,t) \in E} \mathbf{P}(f^{n-1}(t)|C_i).$$

That is, we adjust our prediction for protein v to be the function that maximizes the local likelihood with respect to the functions predicted for its neighbors in the previous step. Again, if $v \in W$, then $f^n(v) = f^{n-1}(v)$.

We stop the iterative process as soon as the difference between the values of the global likelihood in two consecutive steps drops below a given threshold. The pseudo-code in Figure 1 summarizes the algorithm. The time complexity of the algorithm is $O(d|E|)$, where d represents the number of iterations (usually $d \leq 5$ in our experiments).

6. Experimental results

The dataset used in our experimental studies consists of the two largest PPI networks available at the time of writing, namely the network for *S. cerevisiae* and the

PROTEIN_FUNCTION_PREDICTOR$(G, \mathcal{F}, f)$
Input
The PPI graph $G(U \cup W, E)$ where W is the set of known proteins and U is the set of proteins whose functions are to be determined.
The set of functional classes $\mathcal{F}$.
The functional annotations on the known proteins $f : W \rightarrow \mathcal{F}$.
Output
The predicted annotations on all the proteins in the graph $f : V \rightarrow \mathcal{F}$

Estimate $\mathcal{P}(C_i), \mathbf{P}(C_i|C_j)$ for all $C_i, C_j \in \mathcal{F}$ as described in Section 4.
for all $v \in V$ **do**
if $(v \in U)$ **then** $f(v) = argmax_{f(v)\in\mathcal{F}} \mathcal{P}(f(v)) \prod_{(v,t)\in E, t\in W} \mathbf{P}(f(t)|f(v))$
repeat
for all $v \in W$ **do** $f'(v) = f(v)$
for all $v \in U$ **do** $f'(v) = argmax_{f'(v)\in\mathcal{F}} \mathcal{P}(f'(v)) \prod_{(v,t)\in E} \mathbf{P}(f(t)|f'(v))$
$L(G) = \prod_{v\in V} \mathcal{P}(f(v)) \prod_{(v,w)\in E} \mathbf{P}(f(w)|f(v))$
$L'(G) = \prod_{v\in V} \mathcal{P}(f'(v)) \prod_{(v,w)\in E} \mathbf{P}(f'(w)|f'(v))$
for all v in V **do** $f(v) = f'(v)$
until $|L'(G) - L(G)| < \epsilon$
return f

Fig. 1. A sketch of our inference algorithm.

Table 1. The statistics of the PPI networks used in the experiments. $|V|$ is the number of proteins in the network, $|E|$ is the number of interactions, $|W|$ is the number of known proteins, and *naive expected* is the expected prediction accuracy of the naive approach (see text).

			17 functional classes		190 functional classes	
organism	$\|V\|$	$\|E\|$	$\|W\|$	*naive expected*	$\|W\|$	*naive expected*
yeast	4,959	17,511	3,022	0.5010	2930	0.1939
yeast high confidence	1,735	2,354	1,325	0.4286	1278	0.1979
fly	7,451	22,818	3,858	0.6016	3796	0.2832

one for *D. melanogaster*. The networks were obtained from the DIP database[17]. The yeast PPI network is composed of 4,959 proteins and 17,511 interactions, whereas the fly network consists of 7,451 proteins and 22,819 interactions. We also extracted a *high confidence* yeast PPI network, which is a subset of the yeast PPI network in which interactions that are confirmed by only a single experiment have been removed. This latter network has 1,735 proteins and 2354 interactions. The functional annotations were obtained from the Gene Ontology (GO) hierarchy[1].

We used cross validation to quantitatively evaluate the prediction accuracy of our algorithm and to compare its performance with other methods. In each experiment, we randomly removed the functional annotation from a percentage p of known proteins, where p ranges from 5% to 95%. This new set of "unknown" proteins served as the test set, called hereafter T. We use $W \setminus T$ to denote the set of known proteins after p% of them have been "un-labelled" and U to denote the

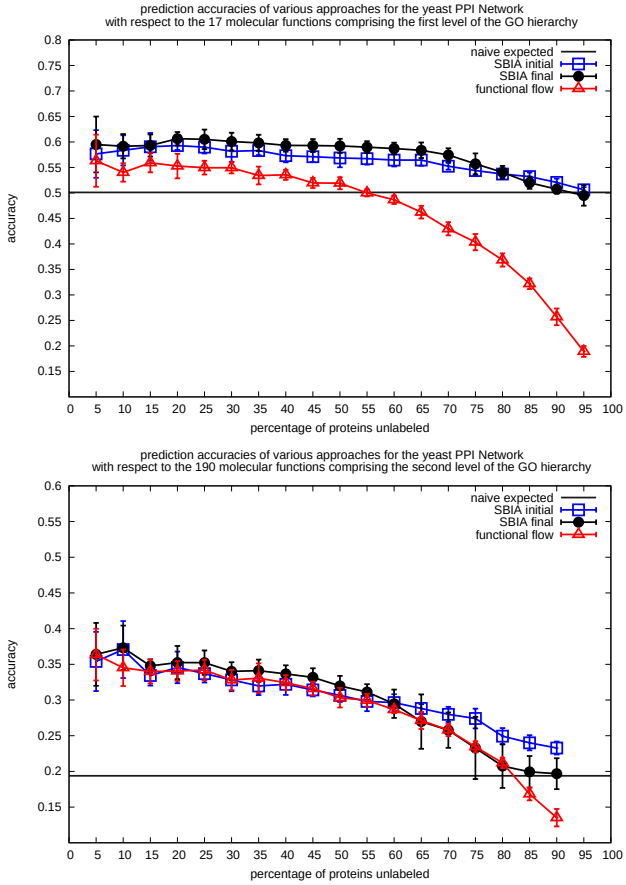

Fig. 2. Prediction accuracies on the yeast PPI network with respect to the 17 functional classes at the first level of the GO hierarchy (top) and 190 functional classes at the second level of the GO hierarchy (bottom). The x-axis represents the percentage of known proteins on which the algorithms are tested. The "naive expected" line indicates the expected prediction accuracy of the naive approach. "SBIA initial" refers to the accuracy of SBIA after the initialization phase, whereas "SBIA final" shows the final accuracy of SBIA. "Functional flow" denotes the prediction accuracy of the functional flow algorithm

set of the remaining unknown proteins. Clearly, the SBIA's learning phase (i.e., the computation of the prior and the conditional probabilities) is carried out only on the proteins in $W \setminus T$. Learning on the original set W would constitute "cheating".

So far, in our model we assumed that each protein can perform only one function. This is, however, not true for many proteins. A protein may participate in multiple biological processes and as a result, it will carry out multiple functions. In the yeast network, 488 proteins out of 3,022 are annotated with two or more top level functions. In the fly network, 1,961 proteins out of 3,858 are annotated with multiple

functions. To handle this issue, the nodes in W that are associated with multiple functions are replicated, so that each copy carries out exactly one of the annotated functions. Each copy has the same interaction partners of the original protein.

As said, the goal is to predict a function for each of the proteins in set $T \cup U$, based on the functional classes in $W \setminus T$ and the topology of the graph. For each protein in T, we declare a prediction to be correct if the predicted function is one of the functions the protein was originally assigned. The prediction accuracy is calculated as the ratio between the number of correct predictions and the total number of proteins in the set T. Since the prediction accuracy varies slightly every time we randomly select T, we replicate the same experiments ten times and compute the average accuracy. We also record the standard deviation, represented by the error bars in the figures.

We compared the accuracy of our method against that of functional flow[14] and against that of the *naive* approach. We chose to compare SBIA against the functional flow method because Chua *et al.*[2,14] report that functional flow outperforms both majority-rule based methods[18,9] as well as methods based on the generalized multicut[22,11]. As said, a direct comparison between our method and MRF-based methods[5,6,12] is not feasible because these latter approaches predict more than one functional class for each protein. The naive method simply predicts the function of a protein to be the most probable functional class according to the prior, i.e., $argmax_{C_i \in \mathcal{F}} \mathcal{P}(C_i)$. Clearly, the expected prediction accuracy of the naive approach is equal to the ratio between the number of proteins annotated with the most probable function and the total number $|W|$ of known proteins.

We carried out two sets of experiments. In the first set, we considered the seventeen top level molecular functions defined in GO. In the yeast PPI network, 3,022 proteins out of 4,959 are annotated with one or more top level functions. The most frequent function is "catalytic activity", which occurs 1,514 times. Thus, the expected prediction accuracy for the naive approach is 0.501 or 50%. In the high confidence yeast PPI network 1,325 proteins are annotated. The most frequent function in this network is again "catalytic activity", which is assigned to 568 proteins. In the fly PPI network, 3,858 protein out of 7,451 are annotated with one or more functions. The most prevalent function in this network is "binding", which appears 2,321 times. Hence, the expected prediction accuracy for the naive approach is 0.6016. The statistics of the networks constituting the dataset are summarized in Table 1.

Figure 2-top, 3-top, and 4-top summarize the results of the first set of experiments on the seventeen functional classes in the top level of the GO hierarchy. The figures show that SBIA always outperforms functional flow, especially when p is large. In the yeast network, the prediction accuracies of the functional flow algorithm even falls below that of the naive approach when p is greater than 55%. SBIA, however, still retains good prediction accuracy until p becomes higher than 70%, and then asymptotically converges to that of the naive approach. Notice that the initialization phase of SBIA already achieves a good prediction accuracy. When

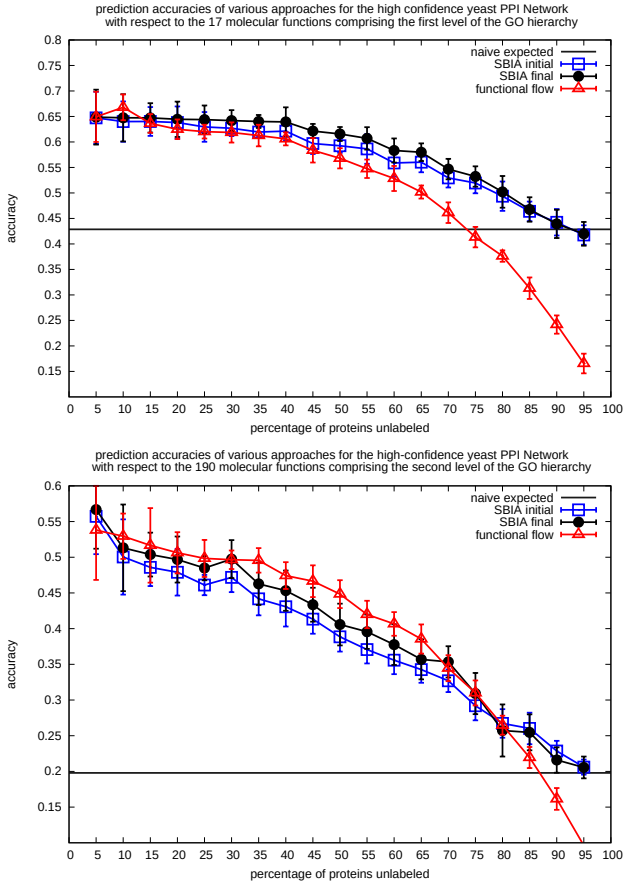

Fig. 3. Prediction accuracy on the yeast high confidence PPI network (see caption of Figure 2 for more details). TOP: 17 functional classes, BOTTOm: 190 functional classes.

p is less than 80%, the iterative phase improves the prediction accuracy even more, along with the global likelihood of the graph. The number of iterations executed is usually rather small, less than 5. When p is greater than 80%, the information left in the network is highly incomplete, and as expected the performance of our algorithm falls back to that of the naive approach.

Due to the higher quality of the data in the yeast high confidence network, the improvement in accuracy of our algorithm and functional flow relative to the naive approach is almost doubled. Rather surprisingly, in the fly network the naive approach performs the best. Our algorithm performs slightly inferior to the naive approach, but they are very close. The data in Table 2 shows that the information content in terms of functional association for the fly network is only about 1/33 of that of the yeast network and 1/86 of that of the high confidence yeast network. In

other words, in the fly network the knowledge of the functions of the neighbors of a protein does not help the prediction because the random variables associated with those functions are almost independent. The only reliable information we have is the prior distribution, whereas the structure and the annotations on the network do not provide significant additional information. This explains why our algorithm does not perform better than the naive approach. We postulate that the lack of the functional association in the fly network is due to the fact that this network was obtained mainly by high-throughput yeast-two-hybrid experiments[8,7], which are highly prone to false positive and false negative[23,4].

In the second set of experiments, we considered all the 190 molecular functions comprising the second level of the GO hierarchy. In the yeast network, 2,930 proteins out of 4,959 yeast proteins are annotated with one or more second level molecular functions. The most prevalent function is "hydrolase activity", which appears 568 times. Hence the expected prediction accuracy for the naive approach is 0.1939. In the high confidence yeast network, 1,278 out of 1,735 proteins are annotated. The most prevalent function is "protein binding", which is annotated to 253 proteins. In the fly network, 3,796 out of 7,451 proteins have been annotated with one or more second level functions. The most prevalent function is "nucleic acid binding", which appears 1075 times. Therefore, the expected prediction accuracy for the naive approach is 0.2832. The statistics are summarized in Table 1.

Figure 2-bottom, 3-bottom and 4-bottom summarize the second set of experimental results. In Figure 3-bottom the functional flow algorithm outperforms SBIA by 2-3% on average. We suspect that this is due to the relatively small size of the network (containing about 1,300 characterized proteins) under consideration and the large number of functions ($k = 190$). Recall that the number of parameters of our model is $\Theta(k^2)$. In this case, we believe that there is not enough data for the accurate estimation of the parameters for the prior distribution and the conditional distributions. For the two other PPI network, the results are similar to that in the previous set of experiments. On the yeast PPI Network, SBIA still outperforms functional flow, but the difference between the two approaches is not as strong as in the previous case. On the fly network, the functional association is still very low as reflected in Table 2 and not surprisingly the best predictor is again the naive approach.

7. A discussion on the functional association in PPI networks

The main assumption behind our probabilistic model is that directly interacting proteins are likely to share the same function. Following Chua *et al.*[2] we refer to this property of PPI networks as *direct functional association*. Clearly, the performance of our algorithm depends crucially on the degree of direct functional association. Functional association can be negatively affected by noise or incompleteness in the process of collecting PPI data and by potential inconsistencies or inaccuracies in the annotations of known proteins.

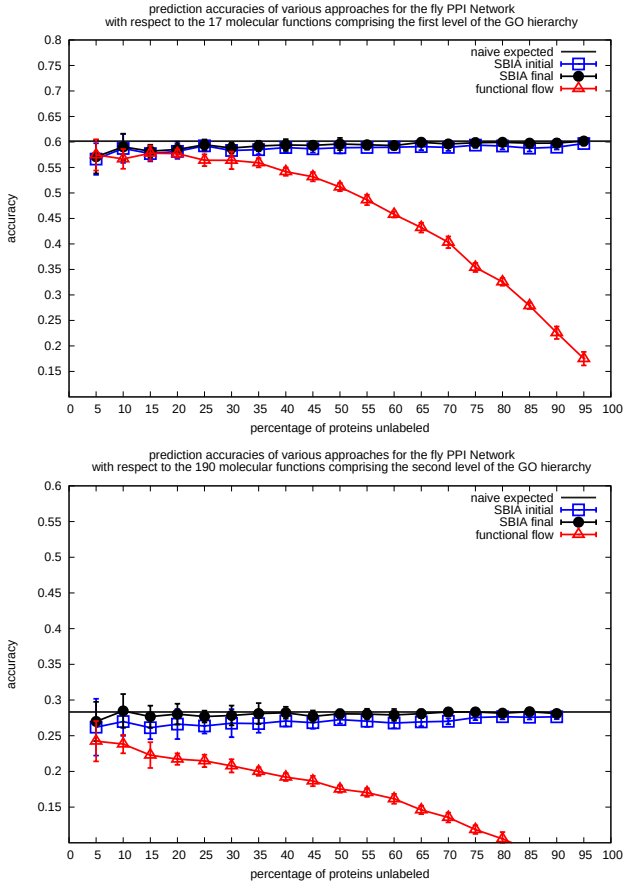

Fig. 4. Prediction accuracy on the fly PPI network (see caption of Figure 2 for more details). TOP: 17 functional classes, BOTTOM: 190 functional classes.

To quantify the degree of direct functional association, we propose a metric based on the notion of graph *entropy*. Graph entropy is a metric that is routinely employed to characterize the randomness of a graph[24]. Let (u, v) be an edge of the network under study. Let X be the discrete random variable associated with the functional class of u, and Y be the random variable associated with the functional class of v. The domain of X and Y is clearly the set $\mathcal{F}$. The *mutual information* $\mathcal{I}(X, Y)$ between X and Y is defined as $\mathcal{I}(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$, where $H(X)$, $H(Y)$, $H(Y|X)$ and $H(X|Y)$ are respectively entropies and conditional entropies of the corresponding random variables. The mutual information measures the reduction of uncertainty in one variable when we known the other. This reduction of uncertainty matches our intuitive notion of direct functional association between proteins u and v. Thus, we define a new metric Q that measures

the functional association of a PPI network as $Q = \mathcal{I}(X,Y)/H(X)$.

The value for Q ranges from zero to one depending on the degree of functional association between the two proteins. If one variable is independent from the other then Q is equal to zero. If one variable is completely determined once we are given the other, then Q is equal to one. We believe that Q relates directly to the ability of accurate prediction of any approach solely based on functional association. For example, when Q is close to zero, i.e., there is almost no mutual information between X and Y, the best predictor one can use is the naive approach based only on the prior.

Table 2 summarizes the value of Q for the three networks under study for the two levels of GO annotations. Note that the value of the functional association Q for the fly network is only 0.0018, which is about 33 smaller than that of the yeast network. Note also that Q is much higher in the high confidence yeast network, as one would expect. It is well-known that the quality of PPI data, in particular the one obtained in high-throughput systems (like the case of *Drosophila*) is rather low[23,4].

8. Conclusions

We developed an efficient algorithm to assign functional GO terms to uncharacterized proteins on a PPI network based solely on the topology of the graph and the functional labels of known proteins. The statistical model proposed in this paper is a generalization of the GenMultiCut model and resemble the MRF-based model by Deng *et.al.* The similarity with the work of Deng *et.al.* is, however, superficial as we discussed in details in the paper. In particular, the structure of our model allows one to obtain easily and efficiently the maximum likelihood estimation of the underlying parameters, which is tipically not possible for a general MRF. Based on our statistical model, we presented efficient learning and inference algorithms. Our inference algorithm is an iterative algorithm, where each iteration runs in time linear in the size of the input. According to our experimental results, our algorithm converges very quickly. More importantly, our method gives consistently better predictions when compared with previous known algorithms.

Table 2. A measure of functional association in the yeast and the fly PPI network. $H(X)$ is the entropy of X, $H(X|Y)$ is the conditional entropy of X given Y, $\mathcal{I}(X,Y)$ is the mutual information between X and Y, and Q is our measure of the direct functional association.

	17 functional classes				190 functional classes			
organism	$H(X)$	$H(X\|Y)$	$\mathcal{I}(X,Y)$	Q	$H(X)$	$H(X\|Y)$	$\mathcal{I}(X,Y)$	Q
yeast	1.692	1.589	0.103	0.0609	3.072	2.789	0.283	0.0921
yeast high confidence	1.725	1.458	0.267	0.1548	2.950	2.236	0.714	0.2505
fly	1.694	1.691	0.003	0.0018	3.211	3.153	0.058	0.0181

9. Acknowledgements

This project was supported in part by NSF CAREER IIS-0447773 and NSF DBI-0321756.

References

1. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene ontology: tool for the unification of biology. *Nature Genetics 25* (2000), 25–29.
2. Chua, H. N., Sung, W.-K., and Wong, L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics 22* (2006), 1623 – 1630.
3. Deng, M., Chen, T., and Sun, F. An integrated probabilistic model for functional prediction of proteins. *Journal of Computational Biology 11*, 2/3 (2004), 463–475.
4. Deng, M., Sun, F., and Chen, T. Assessment of the reliability of protein-protein interactions and protein function prediction. In *Pacific Symposium on Biocomputing* (2003), pp. 140–151.
5. Deng, M., Tu, Z., Sun, F., and Chen, T. Mapping gene ontology to proteins based on protein-protein interaction data. *Bioinformatics 20*, 6 (2004), 895–902.
6. Deng, M., Zhang, K., Mehta, S., Chen, T., and Sun, F. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology 10*, 6 (2003), 947–960.
7. Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., Jacq, B., Arpin, M., Bellaiche, Y., Bellusci, S., Benaroch, P., Bornens, M., Chanet, R., Chavrier, P., Delattre, O., Doye, V., Fehon, R., Faye, G., Galli, T., Girault, J.-A., Goud, B., de Gunzburg, J., Johannes, L., Junier, M.-P., Mirouse, V., Mukherjee, A., Papadopoulo, D., Perez, F., Plessis, A., Rosse, C., Saule, S., Stoppa-Lyonnet, D., Vincent, A., White, M., Legrain, P., Wojcik, J., Camonis, J., and Daviet, L. Protein interaction mapping: A drosophila case study. *Genome Res. 15*, 3 (2005), 376–384.
8. Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., Mcdaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., Dasilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., Mckenna, M. P., Chant, J., and Rothberg, J. M. A protein interaction map of *Drosophila melanogaster. Science 302*, 5651 (2003), 1727–1736.
9. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T. Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast 18*, 6 (2001), 523–531.
10. Jaimovich, A., Elidan, G., Margalit, H., and Friedman, N. Towards an integrated protein-protein interaction network. In *Proceedings of ACM RECOMB* (2005), pp. 14–30.

11. Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R., and Kasif, S. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A 101*, 9 (2004), 2888–2893.
12. Letovsky, S., and Kasif, S. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics 19*, 1 (2003), i197–i204.
13. Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J.-F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., van den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. A map of the interactome network of the metazoan *C. elegans*. *Science 303* (2004), 540–543.
14. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. In *Proceedings of ISMB* (2005), pp. 302–310.
15. Rain, J.-C., Selig, L., Reuse, H. D., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schchter, V., Chemama, Y., Labigne, A., and Legrain, P. The protein-protein interaction map of *Helicobacter pylori*. *Nature 409* (2001), 211–215.
16. Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala1, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature 437* (2005), 1173–1178.
17. Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. The database of interacting proteins: 2004 update. *Nucleic Acids Research 32* (2004), D449.
18. Schwikowski, B., Uetz, P., and Fields, S. A network of protein-protein interactions in yeast. *Nature Biotechnology 18* (2000), 1257 – 1261.
19. Srinivasan, B. S., Novak, A. F., Flannick, J. A., Batzoglou, S., and McAdams, H. H. Integrated protein interaction networks for 11 microbes. In *Proceedings of ACM RECOMB* (2006), pp. 1–14.
20. Uetz1, P., Giot1, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., , and Rothberg, J. M. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature 403*, 6770 (2000), 623–627.
21. Vazirani, V. V. *Approximation algorithms*. Springer-Verlag New York, Inc., New York, NY, USA, 2001.
22. Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. Global protein function prediction in protein-protein interaction networks. *Nature Biotechnology 21* (2003), 697.

23. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature 417*, 6887 (2002), 399–403.
24. Yang, Q., Siganos, G., Faloutsos, M., and Lonardi, S. Evolution versus intelligent design: Comparing the topology of protein-protein interaction networks to the internet. In *IEEE Computational Systems Bioinformatics Conference (CSB'06)* (Stanford, CA, 2006), pp. 299–310.