Combining Hierarchical Inference in Ontologies with Heterogeneous Data Sources Improves Gene Function Prediction

Xiaoyu Jiang[1], Naoki Nariai[2], Martin Steffen[3,4], Simon Kasif[2,4], David Gold[1], Eric D. Kolaczyk[1]
[1]Department of Mathematics and Statistics [2]Bioinformatics Program
[3]Department of Genetics and Genomics [4]Department of Biomedical Engineering
Boston University, Boston MA, USA
xiaoyu@math.bu.edu, nariai@bu.edu, steffen@bu.edu, kasif@egnc.bu.edu,
dlgold@math.bu.edu, kolaczyk@math.bu.edu

Abstract

The study of gene function is critical in various genomic and proteomic fields. Due to the availability of tremendous amounts of different types of protein data, integrating these datasets to predict function has become a significant opportunity in computational biology. In this paper, to predict protein function we (i) develop a novel Bayesian framework combining relational, hierarchical and structural information with improvement in data usage efficiency over similar methods, and (ii) propose to use it in conjunction with an integrative protein-protein association network, STRING (Search Tool for the Retrieval of INteracting Genes/proteins), which combines information from seven different sources. At the heart of our work is accomplishing protein data integration in a concerted fashion with respect to algorithm and data source. Method performance is assessed by a 5-fold cross-validation in yeast on selected terms from the Molecular Function ontology in the Gene Ontology database. Results show that our combined use of the proposed computational framework and the protein network from STRING offers substantial improvements in prediction. The benefits of using an aggressively integrative network, such as STRING, may derive from the fact that although it is likely that the ultimate gene interaction matrix (including but not limited to protein-protein, genetic, or regulatory interactions) will be sparse, presently it is still known only incompletely in most organisms, and thus the use of multiple distinct data sources is rewarded.

1. Introduction

An understanding of the functional roles of proteins is central in biology, for purposes ranging from general knowledge to the development of targeted medicine and diagnostics. Protein function prediction methods can take many forms. For biological process and pathway annotation, the use of protein interaction relationships in terms of functional linkage graphs has been a popular choice in recent years. Markov Random Field model are well suited to model such relationships, and are commonly applied under a Bayesian framework [6, 9, 11]. Many such relationships and annotations are stored in various databases, such as BIOGRID, a protein-protein interaction (PPI) dataset; MIPS, a database for genome/protein sequences, and the Gene Ontology (GO) database, a rigorous vocabulary for biological functions and available for computation. Concurrently, gene functionality prediction by information integration has become a major focus. Various genome-wide data have been employed in Bayesian frameworks [8, 11, 13], Markov Random Field models [4], and machine learning approaches [7]. Notably, however, a common feature of these methods is that they predict protein functions in a "flat" fashion, without capitalizing on the ontological structures among functions from the GO database.

Ontology structures, essentially, are hierarchies, with certain top to bottom annotation criterion, the *true-path rule*, which protein function predictions should in principle follow. Many methodologies have recently been proposed to combine protein data and the ontology structures, [1, 2, 5, 12]. However, importantly, all of these that predict at multiple depths in the GO hierarchy take a separate step to correct inconsistent predictions, rather than producing them directly in a probabilistically coherent way. This problem is tackled in [6], but the methodology proposed therein is limited in that it uses only a simple PPI network as input.

In this paper, we propose a new framework for protein function predication – PHIPA (Probabilistic Hierarchical Inference of Protein Activity) – that uniquely incorporates integrative aspects at the level of both statistical method-

ology and data input. Our proposed methodology combines protein relational information and different protein feature data (such as protein motif (domain) and cellular localization information), together with the Gene Ontology (GO) hierarchical structure. At the same time, rather than encoding protein relational information through a standard PPI network, we use a network derived from an integrative protein database, containing known and predicted protein association information from multiple sources, i.e., STRING (Search Tool for the Retrieval of INteracting Genes/proteins) [10]. Interestingly, little work appears to have been done to date to take advantage of the integrated information in STRING to predict protein functions. Overall, through our proposed approach we achieve greater data usage efficiency and are able to produce predictions that are inherently consistent with the *true-path rule*. The merit of our work is to carry out the over-riding theme of information integration through a combination of data and methodology, rather than merely either of them, to best infer proteins' functional roles.

2. Methodology

2.1. Assumptions and Notations

For a given protein i, we want to predict whether it has a function G, a term from the Gene Ontology (GO) database, given the relational information from protein networks, protein categorical feature information, and the GO hierarchical structure. The *true-path rule* for the gene ontologies requires that if a child term (i.e., more specific term) describes the gene product, then all its ancestor terms (i.e., less specific terms) must also apply to that gene product.

Ontologies are structured as directed acyclic graphs (DAG's), where a child term may have multiple parent terms. To avoid the NP-hard problem of assigning values to variables in a DAG of size N given their conditional probabilities on the arcs, we first apply a minimal spanning tree (MST) algorithm, to transform a DAG into a tree-structured hierarchy, as a routine approximation of probability distributions on DAG's [3]. As a result, term G has one parent in the tree, denoted as $pa(G)$.

We propose to build a classifier based on the use of hierarchical conditional probabilities of the form

$$P(Y_G^{(i)} = 1 | Y_{pa(G)}^{(i)} = 1; \mathcal{X}_{net}^{(i)}; \mathcal{X}_{feat}^{(i)}). \quad (1)$$

The notations are explained as follows.

- The binary variable $Y_G^{(i)} = 1$ indicates that protein i has function G; otherwise, it takes the value -1.
- The notation $\mathcal{X}_{net}^{(i)}$ denotes all the neighborhood information for protein i from protein networks, such as protein-protein interaction (PPI) network, gene co-expression and others, as used in [7, 11]. To simplify the notation, we do not use subscripts to distinguish different networks. More specifically, $\mathcal{X}_{net}^{(i)} = \{k_G, k_{pa(G)}\}$, the number of protein i's neighbors labeled with term G and $pa(G)$ from a given network, respectively. We do not need to consider the neighborhood size because only neighbors labeled with $pa(G)$ can be further annotated with G and hence affect protein i's label, due to the *true-path rule* [6].
- The notation $\mathcal{X}_{feat}^{(i)}$ denotes the categorical feature information for protein i, for instance, protein motif (domain), protein localization and phenotype information [11]. Different feature information is not indexed by subscripts here. Each feature may consist of multiple categories, forming a feature vector. For example, there are 373 protein domains used as the protein motif vector $\boldsymbol{m}$ in this paper, $\boldsymbol{m} = (m_1, \ldots, m_{373})$; 33 cellular location categories are used as the localization feature, $\boldsymbol{l} = (l_1, \ldots, l_{33})$.

Motivated by reasonable empirical evidence and to further aid the tractability of probabilistic calculations, we apply an assumed Markov property on the protein networks and the GO hierarchy. That is, we assume that protein i's functional label is independent of the others given its neighborhood status, and that a GO term is independent of other terms in the hierarchy given its parent. In addition, a Naive Bayes assumption is used to separate protein networks and features, i.e., information from networks is assumed independent of that from protein features, given the protein's functional annotations.

2.2. Local Conditional Probability

It can be derived by Bayes rule that the target probability (1) has the following form under our model assumptions:

$$\begin{aligned} & P(Y_G^{(i)} = 1 | Y_{pa(G)}^{(i)} = 1; \mathcal{X}_{net}^{(i)}; \mathcal{X}_{feat}^{(i)}) \\ = & \frac{\Pi_{j=1}^{N_{net}} \alpha_j \Pi_{k=1}^{N_{feat}} \beta_k}{1 + \Pi_{j=1}^{N_{net}} \alpha_j \Pi_{k=1}^{N_{feat}} \beta_k} \end{aligned} \quad (2)$$

where N_{net} and N_{feat} are the numbers of different protein networks and features used, respectively. We explain the key components α_j and β_k in detail below.

The notation α_j is the ratio of the probabilities of neighborhood information from network j given protein i is labeled and NOT labeled with the target function G. Employing the Hierarchical Binomial-Neighborhood (HBN) assumption from [6], we can show that for a given network (omitting network index j),

$$\alpha = \frac{Binomial(k_G, k_{pa(G)}; p_1) \cdot f}{Binomial(k_G, k_{pa(G)}; p_0) \cdot (1 - f)},$$

where k_G and $k_{pa(G)}$ are explained before, parameter p_1 (p_0) is the probability with which neighbors of protein i are independently labeled with G and $pa(G)$, given i is labeled (NOT labeled) with G. We estimate them from the training data using a standard pseudo-likelihood approach. The parameter $f = P(Y_G^{(i)} = 1 | Y_{pa(G)}^{(i)} = 1)$ is estimated by

$$f = w f_{global} + (1 - w) f_{nbhd},$$

where f_{global} and f_{nbhd} are relative frequencies of term G given its parent on the entire training set and the neighborhood of protein i, respectively; and the weight w can be predetermined or estimated by the pseudo-likelihood method.

The parameter f used in our framework is smoothed in the above fashion towards a balance point between the global and local conditional relative frequencies, borrowing information of G from both the whole training set and the specific protein neighborhood. In the work of [6], the parameter f is estimated simply by the global empirical frequency of G given $pa(G)$ on the training set i.e., w=1. Note that f is term-specific, which can sometimes lead to an estimation issue due to a lack of data for rare terms. Some rare terms with low frequencies on the entire network may have local enrichment [9]. In these cases, using a smaller global relative frequency may decrease the predictive probability for those proteins and hence increase false negatives. For example, for term *GO:0003774, motor activity*, its global conditional relative frequency f_{global} given its parent *GO:0003674, molecular function*, is 0.0033. It has a local enrichment on the neighborhood of gene YOR035C, with a local conditional relative frequency $f_{nbhd} = 0.4545$. Obviously, using f_{global} as f is misleading in characterizing the gene YOR035C.

The notation $\boldsymbol{\beta_k}$ is the ratio of the probabilities of feature k given protein i being labeled and NOT labeled with the target function G. Assuming that feature information is independent with broader functions (parent terms), given more specific information below them (child terms), we have (omitting i, k)

$$\begin{aligned} \beta &= \frac{P(\mathcal{X}_{FEAT} | Y_G = 1, Y_{pa(G)} = 1)}{P(\mathcal{X}_{FEAT} | Y_G = -1, Y_{pa(G)} = 1)} \\ &= \frac{P(c_1, \ldots, c_m | Y_G = 1)}{P(c_1, \ldots, c_m | Y_G = -1)}, \end{aligned}$$

where c_j is the j-th category in feature $\boldsymbol{c}$, m is the number of categories for $\boldsymbol{c}$.

Naive Bayes is a common technique in this scenario. For example, [11] applied Naive Bayes assumptions to factorize β in the standard manner, i.e.,

$$\beta = \Pi_{k=1}^{m} \frac{P(c_k | Y_G = 1)}{P(c_k | Y_G = -1)}.$$

However, proteins may carry information from two feature categories that heavily overlap. Such redundancy among the feature categories is not uncommon. Take protein motif (domain) categories as an example. There are 16 proteins associated with motif type *IPR002041, Ran_GTPase*, among the 5132 yeast genes we studied, which is entirely covered by the subset of the 31 proteins associated with motif type *IPR003574, GTPase_Rho*. Using Naive Bayes here will cause inflated likelihoods of feature components, and hence lead to low predictive accuracy.

To solve this problem, we develop a greedy search algorithm to find the maximally informative bins of feature categories, and use the Naive Bayes assumption upon the bins, in order to reduce redundancy. More specifically, consider a categorical feature $\boldsymbol{c} = (\vec{c}_1, \ldots, \vec{c}_m)$, where $\vec{c}_j$ takes the form of an $n \times 1$ binary vector for category j, and n is the number of proteins. The i-th entry in $\vec{c}_j$ being 1 denotes that protein i is assigned to the j-th category; 0 otherwise. We compute the correlation coefficient for two binary vectors as

$$r = \frac{n_{11} n_{00} - n_{10} n_{01}}{\sqrt{(n_{11} + n_{10})(n_{11} + n_{01})(n_{00} + n_{10})(n_{00} + n_{01})}},$$

where n_{pq} is the number of entry pairs where the first entry takes value p in the first vector and the second entry takes value q in the second vector, where $p, q = 1$ or 0.

For a pre-chosen threshold t for the correlation, and separately for the proteins' label status in each term G (i.e., $Y_G = 1$ or $Y_G = -1$), we first find the largest subset of categories where the correlation for any pair is at least t and set this subset as the first bin, B_1. For the other categories, we repeat the same procedure until all categories are analyzed. Bins are allowed to contain individual categories.

After binning all the categories, the ratio of the feature components becomes

$$\beta = \frac{\Pi_{k=1}^{N_{bins}} P(B_k | Y_G^{(i)} = 1)}{\Pi_{k=1}^{N'_{bins}} P(B_k | Y_G^{(i)} = -1)}.$$

The above binning process is label-specific. Since we use two sets of proteins, one labeled and the other not labeled with term G, it is possible to get different bins of feature categories, i.e., N_{bins} may not equal N'_{bins}.

2.3. Global Conditional Probability

As mentioned, the *true-path rule* implies $P(Y_{G_d}^{(i)} = 1, Y_{pa(G_d)}^{(i)} = -1) = 0$. With the local conditional probability for a term G_d at the d-th level below the root term G_0 of a given GO hierarchy, the global conditional probability for a protein i, $P(Y_{G_d}^{(i)} = 1 | \mathcal{X}_{NET}^{(i)}; \mathcal{X}_{FEAT}^{(i)})$ has the

following form

$$\begin{aligned} & P(Y_{G_d}^{(i)} = 1 | \mathcal{X}_{NET}^{(i)}; \mathcal{X}_{FEAT}^{(i)}) \\ = \; & \Pi_{j=1}^{d} P(Y_{G_j}^{(i)} = 1 | Y_{G_{j-1}}^{(i)} = 1; \mathcal{X}_{NET}^{(i)}; \mathcal{X}_{FEAT}^{(i)}) \\ \leq \; & P(Y_{G_{d-1}}^{(i)} = 1 | \mathcal{X}_{NET}^{(i)}; \mathcal{X}_{FEAT}^{(i)}), \end{aligned} \quad (3)$$

where G_{j-1} is the parent term of G_j along the path from G_d to the root G_0. The probability of a more specific term will be no more than that of any of its ancestors, which guarantees to produce threshold-based GO term label assignments that comply with the *true-path rule*. This is an advantage of our method. Most existing methods using terms from the gene ontology as functions allow inconsistency to happen and take a separate step to post-process [1, 9].

2.4. STRING

As part of our overall framework, we use as input on protein relations a network based on STRING (Search Tool for the Retrieval of INteracting Genes/proteins) [10]. STRING is an integrative protein-protein association database, containing known and predicted associations from 7 evidence sources: *database imports*[1], *high-throughput experiments, co-expression, homology based on phylogenetic co-occurrence, homology based on gene fusion events, homology based on conserved genomic neighborhood, and text mining.*[2]

STRING simplifies the access to protein association by providing a comprehensive collection of protein-protein associations for a large number of organisms. A score S is assigned to each interacting pair of proteins by bench-marking against the KEGG pathway. The score is calculated by $1 - S = \Pi_i(1 - S_i)$, where i indicates the individual evidence type described above, and S_i is the score from the i-th source.

We refer to our overall Bayesian framework – incorporating the GO hierarchy, protein categorical features and STRING – as *Probabilistic Hierarchical Inference of Protein Activity (PHIPA)*.

3. Results

3.1. Data Preparation

- STRING: Associations of Yeast (*Saccharomyces cerevisiae*) genes are extracted from the STRING database [10]. 5132 genes are used, after deleting isolated ones, based on which a functional linkage graph is built, where an edge is added to two nodes (proteins) if there is a non-zero STRING score for them.
- Protein motif information: Protein motif categories are extracted from MIPS database. 373 categories are used, after deleting non-informative ones (motif categories with less than 5 proteins assigned). Completely redundant categories are eliminated, wherein two categories are judged to be thus if they have an identical subset of proteins assigned to them.
- Protein localization information: Protein cellular locations are extracted from the MIPS database. 33 categories are used after performing the same data cleaning step as above.
- GO terms: 12 terms are selected from the Molecular Function ontology as listed below. These terms were chosen (i) to focus mainly on DNA binding and signaling, (ii) to check certain other basic metabolic areas, in case protein motifs are particularly useful in some, but not all GO categories, and (iii) to explore algorithmic performance at various depths in the hierarchy.
  - terms related to DNA binding: *GO:0003677, DNA binding; GO:0016874, ligase activity; GO:0004518, nuclease activity; GO:0004386, helicase activity; GO:0003700, transcription factor activity.*;
  - terms related to signaling: *GO:0016887, ATPase activity; GO:0004672, protein kinase activity; GO:0003924, GTPase activity*;
  - terms related to other types of molecules, including proteins, sugars, membrane ion channels *GO:0008233, peptidase activity; GO:0015075, ion transporter activity; GO:0004407, histone deacetylase activity; GO:0051119, sugar transporter activity.*
- Protein-protein interaction (PPI): PPI data is extracted from the GRID database, for the purpose of comparison. The same 5132 genes as in the STRING network are used and a functional linkage graph is built based on their interactions.

3.2. Overall Performance Comparison

In order to obtain a sense of the overall performance gains offered by the various components of our proposed method, we compared it to two other related methods proposed recently in the literature: the hierarchical Binomial-neighborhood (HBN) method [6] and the heterogeneous Binomial-neighborhood (HeteroBN) method [11]. Each of

[1] PPI and pathway databases. Please refer to [10] for more explanation on the evidence sources.

[2] In the following text, we simplify the names of the 7 evidence sources as *database, experiment, co-expression, co-occurrence, gene fusion event, neighborhood and text mining*.

these methods was referred to earlier and differs from our PHIPA method in important aspects of integration. Specifically, (i) HBN integrates only the GO hierarchy with protein interaction data, (ii) HeteroBN integrates protein interaction data with protein motif and localization data, and (iii) both utilize only a standard PPI interaction network to encode information on protein interactions. In contrast, PHIPA integrates the protein interaction data with both the GO hierarchy and protein motif and localization data, and additionally utilizes STRING to encode protein interactions.

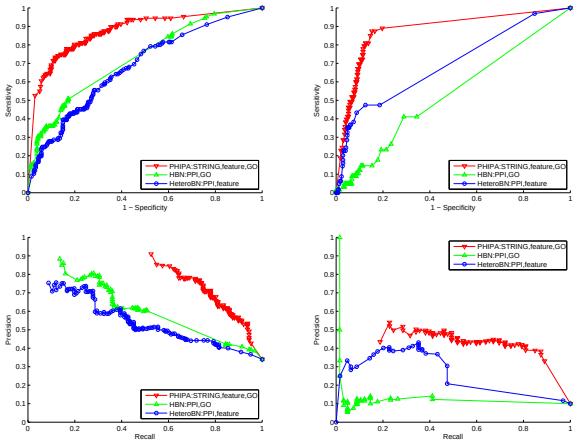

Figure 1. Comparison of protein function prediction accuracy for PHIPA versus previously published algorithms. [Left/right]: GO:0015075, ion transporter activity and GO:0003700, transcription factor activity; [Top/bottom]: ROC curves and precision-recall plots.

A 5-fold cross-validation study was performed on the 12 selected *molecular function* terms using the three methods. Method performance is evaluated here by (a) ROC curves, (b) precision-recall plots, where the curves are functions of a common threshold applied to the probabilities output by each method, as the threshold varies from 0 to 1. Sensitivity, specificity, precision and recall are calculated by averaging the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) counts over the 5 folds for varying thresholds.

PHIPA outperforms HBN by an outstanding margin in all 12 terms, and shows substantial advantage over HeteroBN on most of the terms. Interestingly, protein motif and localization information appear to be highly important in predicting terms such as *GO:0016887, APTase activity*, *GO:0004672, protein kinase activity*.

Due to space limitations, we show the ROC curves and the precision-recall plots only for the terms *GO:0015075, ion transporter activity* and *GO:0003700, transcription factor activity*, which are representative (Figure 1). Please refer to the supplementary materials at *http://math.bu.edu/people/xiaoyu* for all plots and tables for this paper. The significant gain of PHIPA over HBN and HeteroBN directly reflects the benefit of effectively integrating the STRING network information, protein motif and localization information, together with the GO hierarchy into the construction of the classifier.

3.3. Network Comparison

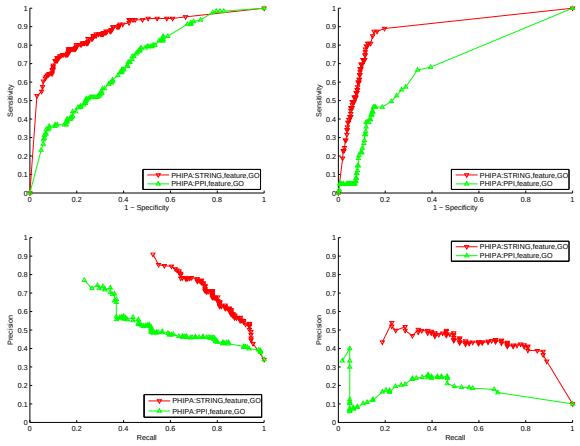

Figure 2. The network dependence of protein function prediction accuracy. [Left/right]: GO:0015075, ion transporter activity and GO:0003700, transcription factor activity; [Top/bottom]: ROC curves and precision-recall plots.

To examine the effect of choice of protein interaction network i.e., STRING vs PPI, we compared PHIPA on STRING to PHIPA on PPI. Note that protein motif and localization information and the GO hierarchy are utilized in both cases. ROC curves and precision-recall plots were generated for all 12 terms, again under 5-fold cross-validation. See Figure 2 for two representative sets of plots. These results indicate that STRING, as an integrative protein association network, offers more information useful to protein function prediction, than PPI, one of the most commonly used protein network in this field.

3.4. Other Protein Information

To study the contribution of protein feature (motif and localization) information and the GO hierarchical structure, four models were implemented on the STRING network:

(1)PHIPA with protein features (motif and localization), (2) PHIPA without protein features (called PHIPA in the legend in Figure 3), (3) HeteroBN, and (4) BN (the Binomial-Neighborhood method from [9], essentially based on a standard Markov random field model).

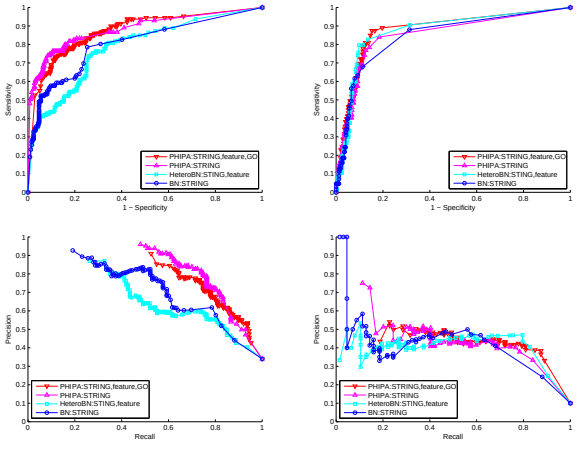

Figure 3. Effect of protein motif and localization information, and the GO hierarchy to prediction accuracy. [Left/right]: GO:0015075, ion transporter activity and GO:0003700, transcription factor activity; [Top/bottom]: ROC curves and precision-recall plots.

Again, ROC curves and precision-recall plots were generated for all 12 terms. Two representative sets of plots are shown in Figure 3. Overall, protein motif and localization information have a small but positive effect on prediction accuracy when using the STRING network. An intriguing observation is that the benefit of incorporating the GO hierarchy varies by terms. For some functions, such as *ion transporter activity* and *DNA binding*, the GO hierarchical structure improves the prediction accuracy significantly; while for others, such as *transcription factor activity* and *helicase activity* , its contribution can be negligible.

4. Discussion

A unified Bayesian Markov Random Field framework, PHIPA, is proposed in this paper, integrating the protein-protein association information from the STRING network, protein motif and localization features, as well as the GO hierarchical structure. The core of our work is information fusion through coherent collaboration of methodology and data usage to improve predictive capabilities.

The results of the previous section show that the proposed PHIPA framework, with STRING, provides a powerful platform for integrating different protein information for inference of protein function. The STRING network is seen to be a major source of the improvements we witness over other methods. The addition of protein features shows a more modest performance contribution and, for certain terms, inclusion of the GO hierarchy demonstrates potential for noticeable advantages. Further analysis can be conducted in a regression framework to study the effects of different STRING evidence types.

References

[1] Z. Barutcuoglu, S. R. E, and T. O. G. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22:830–836, 2006.

[2] L. Blockeel, H an dSchietgat, J. Struyf, and A. D. S. Clare. Hierarchical multilabel classification trees for gene function prediction. *Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB)*, 2006.

[3] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.

[4] M. Deng, T. Chen, and F. Sun. An integrated analysis of protein function prediction. *Journal of Computational Biology*, 11:463–475, 2004.

[5] R. Eisner, B. Poulin, D. Szafron, P. Lu, and R. Greiner. Improving protien function prediction using the hierarchical structure of the gene ontology. *IEEE Symposium on computational Intelligence in Bioinformatics*, 2005.

[6] X. Jiang, N. Nariai, M. Steffen, S. Kasif, and E. D. Kolaczyk. Integration of relational and hierarchical network information for protein function prediction. *BMC Bioinformatics*, 9:350, 2008.

[7] G. R. G. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635, 2004.

[8] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306:1555–1558, 2004.

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

[10] C. V. Mering, L. J. Jensen, B. Snel, and et al. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33:D433–D437, 2005.

[11] N. Nariai, E. D. Kolaczyk, and S. Kasif. Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE*, 2(3):e337, 2007.

[12] B. Shahbaba and M. Neal. Gene function classification using bayesian models with hierarchy-based priors. *BMC Bioinformatics*, 7:448, 2006.

[13] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and B. D. A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). *Proc natl Acad Sci USA*, 100:8348–8353, 2003.