CONTEXT SPECIFIC PROTEIN FUNCTION PREDICTION

NAOKI NARIAI[1] nariai@bu.edu
SIMON KASIF[1,2] kasif@bu.edu

[1]*Bioinformatics Program, Boston University, 44 Cummington St., Boston, MA 02215, USA*
[2] *Department of Biomedical Engineering, Boston University, 44 Cummington St., Boston, MA 02215, USA*

Although whole-genome sequencing of many organisms has been completed, numerous newly discovered genes are still functionally unknown. Using high-throughput data such as protein-protein interaction (PPI) information to assign putative protein function to the unknown genes has been proposed, since in many cases it is not feasible to annotate the newly discovered genes by sequence-based approaches alone. In addition to PPI data, information such as protein localization within a cell may be employed to improve protein function prediction in two ways: 1) By using such localization information as a direct indicator of protein function (e.g. nucleolus localized proteins might be involved in ribosome biogenesis), and 2) by refining noisy PPI data by localization information. In the latter case, localization information may be used to distinguish different types of PPIs: Namely, interactions between co-localized proteins (more reliable), and interactions between differently localized proteins (potentially less reliable). In this paper, we propose a probabilistic method to predict protein function from PPI data and localization information. A Bayesian network is used to model dependencies between protein function, PPI data and localization information. We showed in our cross-validation experiment that in some cases, our method (conditioning PPI data by localization information) significantly improves prediction precision, as compared to a simple Naive Bayes method that assumes PPI data and localization information are conditionally independent given protein function. Finally, we predicted 57 unknown genes as "ribosome biogenesis" proteins.

Keywords: protein function prediction; protein-protein interaction; Bayesian networks.

1. Introduction

One of the challenges in computational biology is to annotate the thousands of unknown genes that are gleaned from the newly sequenced genome of many organisms. Sequence similarity based methods such as BLAST [1], and protein motif (domain) based approaches such as PFAM [2] have been widely used for protein function prediction. However, these sequence-based approaches often fail, when applied to the unknown proteins due to lack of orthologous proteins in other organisms or weak sequence similarity to other known proteins. Recently, high-throughput technologies have produced massive amount of genomic information, such as protein-protein interactions (PPIs), protein localization, and gene expression data. Several types of methods have been proposed to use these genome-wide data to predict protein function. One successful method uses PPI data to assign protein function, based on the assumption that interacting proteins tend to share the same function [3,4]. However, since PPI data produced from high-throughput analyses is known to be noisy, combining other types of genome-wide data may additionally improve protein function prediction methods.

Bayesian network methodologies have been proposed for integrating multiple types of genome-wide data, such as localization information, gene expression data, and co-essentiality to predict PPIs [5]. Moreover, combining such heterogeneous data to predict a functional linkage graph [6], in which an edge between two nodes (genes) represents functional similarity with a reliability score, has been extensively studied [3,7-10]. Instead of producing a functional linkage graph, assigning protein functions to each gene directly from genome-wide data has also been proposed, such as the methods based on Markov random fields (MRFs) [4,11], and other machine learning methods such as support vector machines [12]. A Bayesian method to combine different types of functional linkage graphs and other genomic features (e.g. protein localization, protein motif, etc.) has been shown to improve prediction coverage and accuracy significantly compared to using single source of data [13]. However, the majority of these Bayesian networks methodologies assume conditional independence between genomic features (i.e. PPI data, gene expression data, localization information, and protein motifs) given a class label (i.e. protein function).

In this paper, we use a more sophisticated Bayesian network structure to capture dependencies between genomic features (PPI data and localization information) and class label (protein function) for protein function prediction. Fig. 1 represents the difference between the proposed Bayesian network and Naive Bayes structure. In our context specific Bayesian network model, we condition PPI data by localization information. In other words, we differentiate PPIs into two types: 1) PPIs between co-localized proteins, and 2) PPIs between differently localized proteins. The assumption here is that PPIs between co-localized proteins should be more reliable than PPIs between differently localized proteins (which might be false positives). Hence, in our model, we can assign different weights probabilistically according to the PPI type when predicting protein function given localization information and interacting proteins (and their functions). Our method is applied to protein function prediction in the yeast *Saccharomyces cerevisiae.* In order to assign protein functions, we use the Gene Ontology [14] "biological process" terms as function labels. We show in our 5-fold cross-validation study that our method works significantly better than the Naive Bayes method, when predicting certain functional classes, such as the "ribosome biogenesis" GO term. However, in other cases such as the "mitotic cell cycle" GO term, we find that the simple Naive Bayes method works equally well or even much better than our proposed method. We analyze the results and hypothesize that the more sophisticated model works better when the assumptions made in our model are biologically appropriate for a specific function of interest: For example, when PPI patterns in a subset of co-localized proteins are characterized by a distinct topology or probability of interaction (e.g. "ribosome biogenesis" proteins tend to have PPIs within same localization). Finally, we annotated 57 unknown genes as "ribosome biogenesis" at the estimated precision of 50% (a complete gene list is available in Supplementary Information and available online at http://genomics10.bu.edu/nariai/context/).

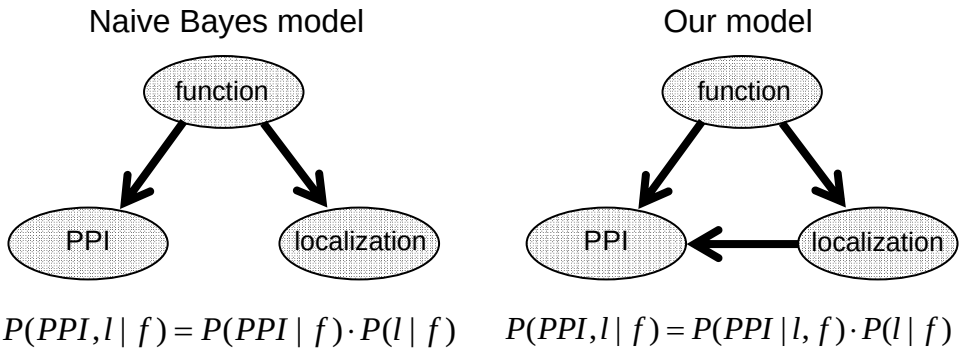

Fig. 1. Context specific protein function prediction: Naive Bayes model (left) and our proposed model (right). In our model, there is a dependency between PPI data and localization information. In the equations, *l*, and *f* represent localization and function, respectively.

2. Methods

2.1. *Data preparation*

Physical protein-protein interactions (PPIs) of *Saccharomyces cerevisiae* are collected from the GRID database [15], as of 01/03/2007. After eliminating redundant interactions and self-self interactions, 31202 PPIs among 5151 genes are obtained. Protein localization information is obtained from the MIPS database [16], as of 11/14/2005. In total, 5191 protein-localization associations are obtained, in which 4076 proteins are associated with at least one of 41 cellular localization categories. For each protein, a feature vector $\mathbf{l} = (l_1, l_2, ..., l_L)^{\mathrm{T}}$ is defined, where l_i is a random variable to show localization ($l_i = 1$ if the protein localizes in l_i, and $l_i = 0$ otherwise), and *L* is the total number of localization features (41 in this case). The Gene Ontology (GO) "biological process" terms are obtained from the Yeast SGD database [17], as of 06/03/2006. For each gene-term association, we expanded and included all 'is-a' and 'part-of' ancestors of the GO label. In total, 107636 gene-term assignments are obtained, in which 6289 genes have at least one of 1965 GO terms. GO terms that appear more than 300 times or less than 5 times among the 6289 genes are subsequently discarded, since we believe that such overly broad or narrow functional terms are not very useful for further experimental validation. From the PPI data collected, we construct a functional linkage graph [6], in which nodes represent genes (proteins) and edges represent PPIs between nodes. From the protein localization information, edges (PPIs) can be divided into two types: 1) PPIs between proteins that share the same localization, and 2) PPIs between proteins that do not share any localization. More precisely, since some proteins do not have localization information at all, there is another type of PPIs: 3) neither 1) nor 2). We call each type of PPIs as 1) co-localized PPIs, 2) cross-localized PPIs, and 3) other PPIs. Generally speaking, it is expected that co-localized PPIs are more reliable than others, since it is usually the case that PPI occurs within the same localization, and other types of PPI

might be rare cases or just false positives from high-throughput analyses. For each type of PPI and each GO term *t*, we calculate p_1, the probability that a protein has term *t*, given that the interacting partner has the label *t*, and p_0, the probability that a protein has term *t*, given that the interacting partner does not have the label *t*. Here, a χ^2 test is performed to ensure that p_1 and p_0 are statistically different using a Bonferroni-corrected *p*-value of 0.001. It is expected that p_1 is higher than p_0, which we are going to take advantage of to make function prediction, given a functional linkage graph. For convenience, we use notations $p_1^{(co)}$ and $p_0^{(co)}$ for the co-localized PPI, $p_1^{(\overline{co})}$ and $p_0^{(\overline{co})}$ for the cross-localized PPI, and $p_1^{(others)}$ and $p_0^{(others)}$ for the other PPI.

2.2. *Posterior probability of function given data*

For each protein and GO term, a Boolean random variable $f_{i,t}$ is associated, where $f_{i,t} = 1$ if the protein *i* is associated with the GO term *t*, and $f_{i,t} = 0$ otherwise. We calculate a posterior probability for all combinations of proteins and GO terms, given PPI data and localization information as $P(f_{i,t} = 1 \mid N_i, k_i, \mathbf{l}_i)$, where N_i is the total number of neighbors of protein *i* in the functional linkage graph (PPI network), k_i is the total number of neighbors of protein *i* which are annotated with *t*, and $\mathbf{l}_i$ is a feature vector for localization information of protein *i*. Applying Bayes' theorem (with omitting subscripts),

$$
\begin{aligned}
P(f \mid N, k, \mathbf{l}) &= \frac{P(k, \mathbf{l} \mid f, N) \cdot P(f \mid N)}{P(k, \mathbf{l} \mid N)} \\
&= \frac{P(k, \mathbf{l} \mid f, N) \cdot P(f \mid N)}{P(k, \mathbf{l} \mid f, N) \cdot P(f) + P(k, \mathbf{l} \mid \overline{f}, N) \cdot P(\overline{f})} \\
&= \frac{P(k \mid \mathbf{l}, f, N) \cdot P(\mathbf{l} \mid f, N) \cdot P(f \mid N)}{P(k \mid \mathbf{l}, f, N) \cdot P(\mathbf{l} \mid f, N) \cdot P(f \mid N) + P(k \mid \mathbf{l}, \overline{f}, N) P(\mathbf{l} \mid \overline{f}, N) \cdot P(\overline{f} \mid N)} \\
&= \frac{P(k \mid \mathbf{l}, f, N) \cdot P(\mathbf{l} \mid f) \cdot P(f)}{P(k \mid \mathbf{l}, f, N) \cdot P(\mathbf{l} \mid f) \cdot P(f) + P(k \mid \mathbf{l}, \overline{f}, N) P(\mathbf{l} \mid \overline{f}) \cdot P(\overline{f})},
\end{aligned}
$$

where we assume that *f* and $\mathbf{l}$ are independent of the number of neighbors *N*, and hence, $P(f \mid N) = P(f)$, and $P(\mathbf{l} \mid f, N) = P(\mathbf{l} \mid f)$.

In the function above, $P(k \mid \mathbf{l}, f, N) = P(k_{co}, k_{\overline{co}}, k_{others} \mid f, N_{co}, N_{\overline{co}}, N_{others})$, where $N_{co}, N_{\overline{co}}, N_{others}$ are the number of co-localized neighbors, cross-localized neighbors, and others, respectively (please see Section 2.1. for the notations), and $k_{co}, k_{\overline{co}}, k_{others}$ are the those neighbors that are annotated with *t*. We assume a multinomial distribution and calculate this probability as

$$P(k_{co}, k_{\overline{co}}, k_{others} \mid f, N_{co}, N_{\overline{co}}, N_{others}) = \frac{N!}{\prod_{i=1}^{6} x_i!} \cdot \prod_{i=1}^{6} \theta_i^{x_i},$$

where

$$x_1 = k_{co},\ x_2 = N_{co} - k_{co}, x_3 = k_{\overline{co}},\ x_4 = N_{\overline{co}} - k_{\overline{co}}, x_5 = k_{others}, x_6 = N_{others} - k_{others},$$

$$\theta_1 = p_1^{(co)} \cdot P(PPI_{co} \mid f),\ \theta_2 = (1 - p_1^{(co)}) \cdot P(PPI_{co} \mid f),$$

$$\theta_3 = p_1^{(\overline{co})} \cdot P(PPI_{\overline{co}} \mid f),\ \theta_4 = (1 - p_1^{(\overline{co})}) \cdot P(PPI_{\overline{co}} \mid f),$$

$$\theta_5 = p_1^{(others)} \cdot P(PPI_{others} \mid f),\ \theta_6 = (1 - p_1^{(others)}) \cdot P(PPI_{others} \mid f),$$

$$\left(x_i \geq 0, \sum_{i=1}^{6} x_i = N,\ \sum_{i=1}^{6} \theta_i = 1. \right)$$

where $P(PPI_{co} \mid f), P(PPI_{\overline{co}} \mid f), P(PPI_{others} \mid f)$ are prior probabilities (fractions) of each type of PPI given a term *t*, which are pre-calculated from a training set.

Similarly,

$$P(k \mid \mathbf{l}, \overline{f}, N) = P(k_{co}, k_{\overline{co}}, k_{others} \mid \overline{f}, N_{co}, N_{\overline{co}}, N_{others}) = \frac{N!}{\prod_{i=1}^{6} x_i!} \cdot \prod_{i=1}^{6} \theta_i'^{x_i},$$

where

$$\theta_1' = p_0^{(co)} \cdot P(PPI_{co} \mid \overline{f}),\ \theta_2' = (1 - p_0^{(co)}) \cdot P(PPI_{co} \mid \overline{f}),$$

$$\theta_3' = p_0^{(\overline{co})} \cdot P(PPI_{\overline{co}} \mid \overline{f}),\ \theta_4' = (1 - p_0^{(\overline{co})}) \cdot P(PPI_{\overline{co}} \mid \overline{f}),$$

$$\theta_5' = p_0^{(others)} \cdot P(PPI_{others} \mid \overline{f}),\ \theta_6' = (1 - p_0^{(others)}) \cdot P(PPI_{others} \mid \overline{f}).$$

$$\left(\sum_{i=1}^{6} \theta_i' = 1. \right)$$

Finally, $P(\mathbf{l} \mid f)$ and $P(\mathbf{l} \mid \overline{f})$ are calculated as

$$P(\mathbf{l} \mid f) = \prod_{i=1}^{L} P(l_i \mid f), P(\mathbf{l} \mid \overline{f}) = \prod_{i=1}^{L} P(l_i \mid \overline{f}),$$

where $P(l_i \mid f)$ = (# of t –labeled genes that have a localization at l_i) / (# of t –labeled genes), $P(l_i \mid \overline{f})$ = (# of genes that are not labeled with t and have a localization at l_i) / (# of genes that are not labeled with *t*). $P(f)$ and $P(\overline{f})$ are prior probabilities, which are fractions of *t*-labeled genes, and genes that are not labeled with t in a training set, respectively.

3. Results

We applied our method to *Saccharomyces cerevisiae* protein function prediction (data preparation is described in Section 2.1.) and evaluate its performance through 5-fold cross validation. Since it is expected that the performance varies from one GO term to another, we choose "ribosome biogenesis" and "mitotic cell cycle" GO terms for our targets. In our 5-fold cross validation experiment, 6289 genes are first divided into five equally-sized sets. Following the standard conventions, four gene sets are selected and treated as a training set, and the remaining one is used as a test set. This second step is repeated until all gene sets have been chosen as a test set. Fig. 2 shows the prediction precision, #TP / (#TP+#FP), for varying posterior probability threshold by our method (conditioning), Naive Bayes method, and the method using PPI data alone. Error bars in graphs show the standard deviation of precision from 10 independent 5-fold cross validation experiments. In the case of predicting the "ribosome biogenesis" GO term, our method is significantly better than the Naive Bayes method and using PPI data alone (t-test, significance < 0.01). However, in the case of the "mitotic cell cycle" GO term, the proposed method is significantly worse than other methods (but Naive Bayes method is significantly better than using PPI data alone). Other than these cases, we found that for predicting the "generation of precursor metabolites and energy" GO term, our method works equally well compared to the Naive Bayes method (data not shown). These results show that whether our conditioning method works better than a Naive Bayes method or not depends on which GO term we are predicting. We explain why the prediction performance is so different depending on GO terms. Since our method weights PPI differently according to the localization (co-localized PPI, cross-localized PPI, and others), our method is most effective when positives (proteins that are annotated with the function of interest) have different values of $p_1^{(co)}, p_1^{(\overline{co})}, p_1^{(others)}$, and tend to have consistent frequencies for each type of PPI compared to negatives (proteins that are not annotated with the function). Note that $p_1^{(co)}, p_1^{(\overline{co})}, p_1^{(others)}$ are 0.62, 0.24, 0.41, respectively for "ribosome biogenesis", and 0.25, 0.20, 0.26, respectively for "mitotic cell cycle". We see that these values are quite different from each other for the "ribosome biogenesis" GO term, but not for "mitotic cell cycle". This means that co-localized PPI is more reliable for predicting the "ribosome biogenesis" GO term compared to others, hence is helpful to improve precision. Fig. 3 shows the number of neighbors annotated with the same function (*x*-axis) and the number of co-localized neighbors annotated with the same function (*y*-axis). A diagonal pattern is apparent for "ribosome biogenesis", but not for the "mitotic cell cycle" GO term. Since proteins annotated with "ribosome biogenesis" tend to have more co-localized PPI than other types of PPI compared to negatives, and $p_1^{(co)}$ is much higher than $p_1^{(\overline{co})}$, our method could successfully distinguish positives from negatives better than a Naive Bayes method.

From Fig. 2, we estimated the threshold probability 0.10 as the 50% precision point. We newly annotated 57 unknown genes as "ribosome biogenesis" (a complete gene list is available in Supplementary Information, http://genomics10.bu.edu/nariai/context/).

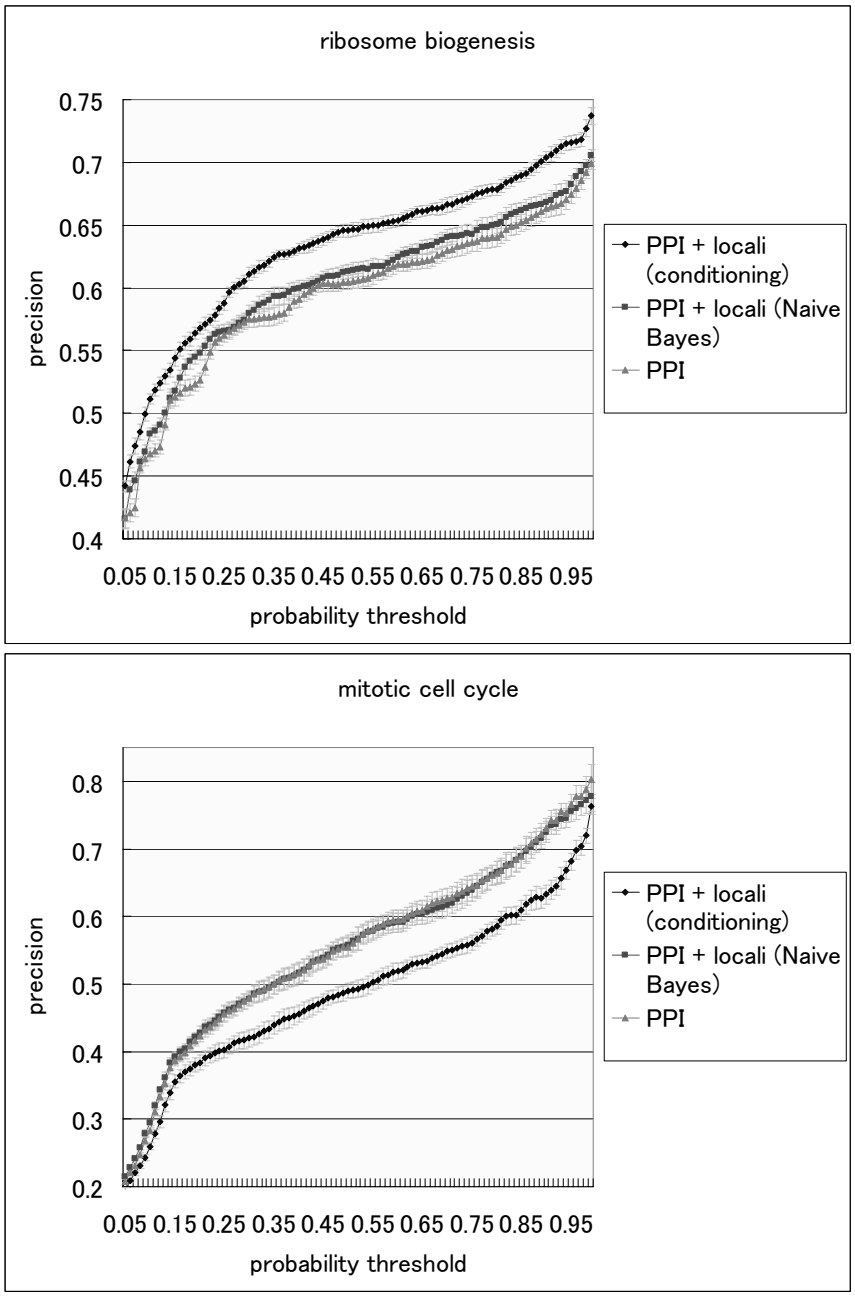

Fig. 2. Prediction performance for GO terms “ribosome biogenesis” (top) and “mitotic cell cycle” (bottom). Precision is defined as (#TP) / (#TP+#FP).

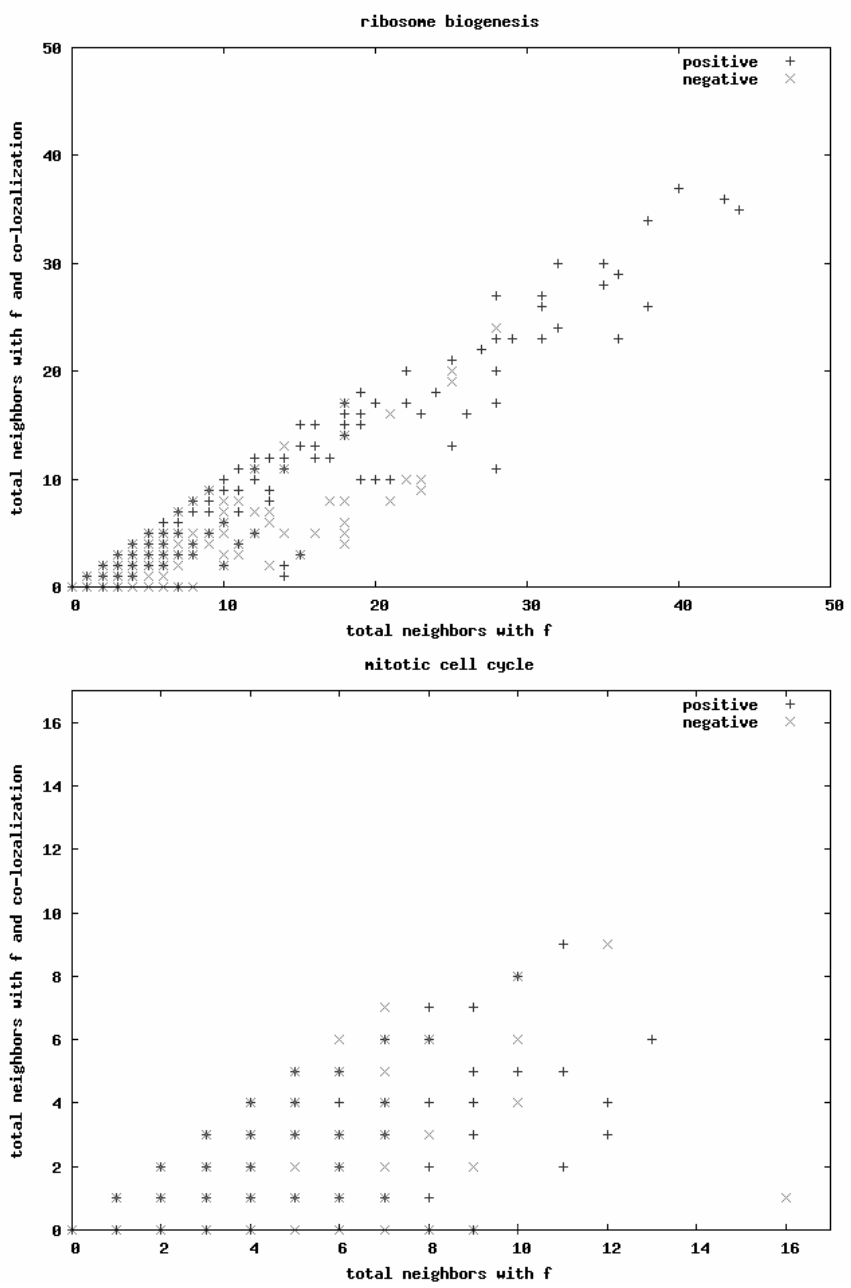

Fig. 3. The number of neighbors with the same function (*x*-axis) and the number of co-localized neighbors with the same function (*y*-axis). We can see a diagonal pattern for the GO term “ribosome biogenesis” (top), but not for “mitotic cell cycle” (bottom).

4. Discussion

In this paper, we propose a probabilistic method to predict protein function from PPI data and localization information under a Bayesian network structure, in which PPI data is conditioned by localization information. The assumption here is that treating PPI data differently (co-localized PPI and cross-localized PPI) will lead to better prediction performance. We showed in our 5-fold cross validation experiment that our method successfully improved prediction precision compared to a simple Naive Bayes method in some cases. In other cases, conditioning PPI data by localization did not improve prediction performance. However, even in these cases where the method does not provide a statistically significant improvement, it allows us to obtain deeper insight into gene function. In particular, it allows us to identify proteins that tend to interact in a similar fashion across multiple localizations. We analyzed the results and hypothesize that if the fraction of co-localized PPI and cross-localized PPI is not consistent for proteins annotated with a specific GO term, then a Naive Bayes method may work better than the proposed method. One limitation of our method is that we only distinguish between two types of PPIs: Co-localized PPIs and cross-localized PPIs. Ideally, every type of PPIs should be treated differently according to a specific localization: PPIs between a protein localized in A (such as nucleus) and a protein localized in B (such as ER). When more PPI data and localization information become available, our method can be extended to model additional types of PPIs. It might also be possible to take other contextual information into account, such as time (e.g. using time-series gene expression data during cell cycle) and biochemical context (e.g. interactions mediated or inhibited by specific protein domains or small molecules). We may then be able to determine the set of biological contexts [18] where PPIs actually take place for a specific functional category. We believe that such a tailored prediction model for each functional category is a key to improve prediction performance and obtain insights into biology. However, learning such a comprehensive context model would require a significant amount of data, while the currently available data remains sparse.

Acknowledgments

We thank Manway Liu and Dr. Martin Steffen for constructive comments. This work was supported by NSF grant ITR-048715 and NHGRI grant R01HG003367-01A1.

References

[1] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25(17):3389-3402, 1997.

[2] Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L., and Bateman, A., Pfam: clans, web tools and services, *Nucleic Acids Res.*, 34(Database issue):D247-251, 2006.

[3] 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. USA*, 101(9):2888-2893, 2004.

[4] Letovsky, S. and Kasif, S., Predicting protein function from protein/protein interaction data: a probabilistic approach, *Bioinformatics*, 19(Suppl 1):i197-204, 2003.

[5] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M., A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, 302(5644):449-453, 2003.

[6] Yanai, I. and DeLisi, C., The society of genes: networks of functional links between genes from comparative genomics, *Genome Biol.*, 3: research0064, 2002.

[7] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D., A combined algorithm for genome-wide prediction of protein function, *Nature*, 402(6757): 83-86, 1999.

[8] Lee, I., Date, S.V., Adai, A.T., and Marcotte, E.M., A probabilistic functional network of yeast genes, *Science*, 306(5701):1555-1558, 2004.

[9] Lu, L.J., Xia, Y., Paccanaro, A., Yu, H. and Gerstein, M., Assessing the limits of genomic data integration for predicting protein networks, *Genome Res.*, 15(7):945-953, 2005.

[10] Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., and Botstein, D., A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae), *Proc. Natl. Acad. Sci. USA*, 100(14):8348-8353, 2003.

[11] Deng, M., Chen, T., and Sun, F., An integrated probabilistic model for functional prediction of proteins, *J. Comput. Biol.*, 11(2-3):463-475, 2004.

[12] Lanckriet, G.R., De Bie, T., Cristianini, N., Jordan, M.I., and Noble, W.S., A statistical framework for genomic data fusion, *Bioinformatics*, 20(16): 2626-2635, 2004.

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

[14] 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. The Gene Ontology Consortium, *Nat. Genet.*, 25(1):25-29, 2000.

[15] Breitkreutz, B.J., Stark, C., and Tyers, M., The GRID: the General Repository for Interaction Datasets, *Genome Biol.*, 4(3):R23, 2003.

[16] Mewes, H.W., *et al*., MIPS: analysis and annotation of proteins from whole genomes, *Nucleic Acids Res.*, 32(Database issue);D41-44, 2004.

[17] Dwight, S.S., *et al*., Saccharomyces genome database: underlying principles and organisation, *Brief. Bioinform.*, 5(1):9-22, 2004.

[18] Rachlin, J., Cohen, D.D., Cantor, C., and Kasif, S., Biological context networks: a mosaic view of the interactome, *Mol. Syst. Biol.*, 2:66, 2006.