PTOMSM: A modified version of Topological Overlap Measure used for predicting Protein-Protein Interaction Network

Xun Huang
School of Life Sciences, Tsinghua University, Beijing 100084, China;
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China;
E-mail: bioxun@gmail.com.

Abstract

A variety of methods are developed to integrating diverse biological data to predict novel interaction relationship between proteins. However, traditional integration can only generate protein interaction pairs within existing relationships. Therefore, we propose a modified version of Topological Overlap Measure to identify not only extant direct PPIs links, but also novel protein interactions that can be indirectly inferred from various relationships between proteins. Our method is more powerful than a naïve Bayesian-network-based integration in PPI prediction, and could generate more reliable candidate PPIs. Furthermore, we examined the influence of the sizes of training and test datasets on prediction, and further demonstrated the effectiveness of PTOMSM in predicting PPI. More importantly, this method can be extended naturally to predict other types of biological networks, and may be combined with Bayesian method to further improve the prediction.

Keywords: network, Bayesian method, integration, PPI, Topological Overlap Measure

Introduction

Over the past decade, newly developed high throughput technologies have generated tremendous amounts of biological data, including protein interaction data, genomic data, and expression data. Several computational methods have been proposed to integrate protein relationships extracted from heterogeneous data sources for the purpose of inferring PPI links. Most of these methods are based on integrating the confidence scores of each relation subset[1, 2]. Confidence score represents the correlation between the subset link value and the existence of a corresponding gold standard PPI link. Specifically, when calculating the total score for a link between protein A and protein B, all of the confidence scores of the link between A and B in each subset are provided as input to an integrating function. For example, the Bayesian method has been widely used in the integration of different data sources, and has been shown as an effective approach[3-6]. Several integrative databases of protein-protein interactions have been generated with this integration approach(e.g. STRING[6-8], pSTIING[9], and PIPs[10]). Though these databases are quite useful for biological research, they have an obvious shortcoming: the methods used to generate the datasets can only identify candidate PPI links that belong to at least one of the heterogeneous relational subsets. Motivated by this limitation, this study aims to expand the scope of interaction discovery and identify links beyond those contained in the relational subsets.

The topological overlap measure(TOM) is a useful approach in network neighborhood analysis[11-13], that can be used to filter spurious or missing connections between network nodes[14]. TOM values could reflect the relative interconnectedness between two nodes, an attribute which is not only determined by the direct link between two nodes, but is also influenced by indirect paths between the two nodes. Therefore, the use of TOM could potentially aid in discovery of indirect links.

To discover links other than the known relations contained in the data sources, we extended the TOM by combining known

PPI data with a similarity matrix, defining Parameterized Topological Overlap Measure with Similarity Matrix (PTOMSM) as a score for predicting PPI links. PTOMSM is designed based on the conception that if protein A1 and B interact with each other, and protein A2 is similar to A1, the probability of the interaction between A2 and B tends to be higher than normal. This approach is somewhat similar to homology-driven identification of protein-protein interaction[15].

In this study, we choose the yeast Saccharomyces cerevisiae as a model system for our method. We transformed various preliminary similarities into a probabilistic score via an evaluation process, and integrated different similarities into a joint similarity. We next calculated the PTOMSM for each similarity (including joint similarity). Our study demonstrates that PTOMSM using the joint similarity is more efficient than other PTOMSMs (with individual similarities) at PPI prediction. Furthermore, we also compared PTOMSM with a Naive Bayesian approach[16, 17] for integrating the heterogenous datasets and generating the probabilistic PPI network. Our results indicate that using the same data sources, our PTOMSM approach is more effective than the naive Bayesian integration at prediction of PPIs. All of our studies were performed on different sizes of training and testing datasets. We also illustrate that PTOMSM is particularly more effective when the training dataset is smaller than the testing dataset.

Materials and methods

Abbreviations of some basic terms used in this study:

TOM*: Topological Overlap Measure*

PTOMSM*: Parameterized Topological Overlap Measure with Similarity Matrix*

PTOMSMj*: PTOMSM with joint similarity*

PTOMSM*(X)*: *PTOMSM with similarity X(e.g. cellular localization, co-expression)*

PTOM*: PTOMSM*(PPI for training)*

LPMB*: Likelihood ratios calculated by Probabilistic Model of Bayesian method*

LPMNBI*: Likelihood ratios Integrated by Probabilistic Model of naive Bayesian method*

LPMB*(X)*: *LPMB for dataset X(e.g. cellular localization, co-expression)*

BPCT*: Bayesian Probability to Complete a Triangle*

Gold standard positive and gold standard negative

Gold standard positive PPI links for training and validation were obtained from the DIP[18], BioGrid[19] and IntAct[20] datasets.

All the protein pairs not present in the gold standard PPI dataset were used as gold standard negative links. Despite the simplicity of this organization, it yields an unbiased estimate of the true distribution of positive and negative interactions, and is likely to be more effective than choosing gold standard negative based on different cellular localization[21].

Definition of preliminary similarities

Based on experiential knowledge, we choose relationships from those datasets that tend to exhibit some positive correlation with PPI related similarity. Since the preliminary similarity values will later be transformed into final uniform similarities, the absolute values of the preliminary similarities and the difference between them will have little effect on the final results. The most important aspect of these datasets is the order of the preliminary similarity.

Preliminary similarity of protein interaction

If there is a PPI link, this preliminary similarity is 1, otherwise it is 0. Remarkably, if the PPI for training is used as this similarity, PTOMSM is equivalent to PTOM.

Preliminary similarity of homology

Homology groups were obtained from the HomoloGene at NCBI, the Inparanoid[22], and the Kog[22] databases. The preliminary similarity value of homology was assigned as 1 for all proteins that belong to the same homology group and all others were assigned a value of 0.

Preliminary similarities of three GO terms

We use the co-occurrence in the biological process category of the Gene Ontology[23] to define the biological process related preliminary similarity(S_{p0}) via GO term of biological process in Equation (1), where SSBP is the smallest shared biological process GO term[17]. Likewise, we define the cellular localization related preliminary similarity (S_{l0}), and molecular function related Preliminary similarity(S_{f0}) in Equation (2 and 3), where SSGC is the smallest shared cellular component GO term, and SSGM is the smallest shared molecular function GO term. The β in Equation (1, 2 and 3) were set to 50 for this study..

$S_{p0} = 1/(1+SSBP/\beta)$ (1)

$S_{c0} = 1/(1+SSGC/\beta)$ (2)

$S_{f0} = 1/(1+SSGM/\beta)$ (3)

Preliminary similarity of co-occurrence in text

Co-occurrence analysis is commonly applied in mining gene relations from literature[24]. We extracted co-occurrence relations by searching genes that are cited in the same literature abstracts. In this study, we only focus on binary co-occurrence relations from literature. The abstracts are all from the abstracts stored in the PubMed of NCBI. We use terms 'Saccharomyces AND gene' to search PubMed, and collected 847364 abstracts by September 2009.

Preliminary similarity of co-expression

Pearson correlation coefficient (PCC) has been commonly used in previous studies for gene co-expression analysis[25]. We adopted the PCC to measure this similarity. Yeast microarray data were obtained from the Saccharomyces Genome Database in ratio format[26]. In each case, microarray based gene expression analysis was previously reported to have been performed for the analysis of DNA Damage[27], Cell Cycle[28], Environmental Response[29], and Essential Genes[30].

Transformation of preliminary similarity to final similarity into PTOMSM using a Bayesian-fashioned method

Because different types of preliminary similarity might contribute differently to the prediction of protein interactions, some method of standardization is necessary in order to integrate the various preliminary similarities. Here we use Bayesian probability to evaluate the preliminary similarity, and transform those similarities into their relevant Bayesian probabilities.

First, we calculated the probability to complete a triangle of every link in the network. This value represents the probability that one link can correctly predict a new link with the assistance of another neighboring link. The mechanism of this prediction is illustrated in Figure 1. The Bayesian probability is defined in Equation (5 and 6), and we denote this probability as the BPCT (Bayesian Probability to Complete a Triangle).

$$\mathrm{PPI2} = \mathrm{PPI0} \cdot \mathrm{PPI0} \quad (5)$$

$$BPCT_{i,j} = \frac{2 \cdot PPI2_{i,j}}{\sum_{u \neq j} PPI0_{i,u} + \sum_{u \neq i} PPI0_{u,j}} \quad (6)$$

$$\overline{BPCT} = \frac{\sum_i \sum_j PPI2_{i,j}}{(n-1)\cdot \sum_i \sum_j PPI0_{i,j}} \quad (7)$$

$$\mathrm{Sf}_{i,j,k} = \begin{cases} BPCT_{i,j,k} & \text{if NprU\$} < \$10 \\ Ffit(S_{i,j,k}) & \text{if NprU\$} \geq \$10 \\ \overline{BPCT} & \text{if } \mathrm{S}_{i,j,k} \text{ is undefined} \end{cases} \quad (8)$$

$$Lr_{i,j,k} = \frac{Sf_{i,j,k}\cdot(1-\overline{BPCT}_k)}{\overline{BPCT}_k\cdot(1-PR_{i,j,k})} \quad (9)$$

$$BPCTt = \sum_{k=1}^{n} \frac{\overline{BPCT_k}}{n} \quad (10)$$

$$St_{i,j} = \frac{1}{1+\frac{BPCTt\cdot\prod_k Lr_{i,j,k}}{1-BPCTt}} \quad (11)$$

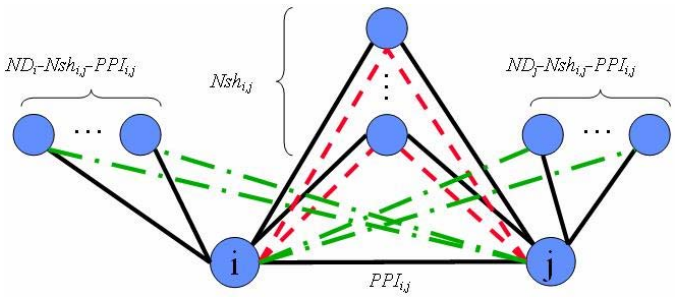

Figure 1: **Bayesian Probability of link to complete a triangle.**

$$BPCT_{i,j} = \frac{\text{Numberofrightpredictions}}{\text{Numberoftotalpredictions}} = \frac{2\cdot Nsh_{i,j}}{ND_i + ND_j - 2\cdot PPI_{i,j}}$$

PPIi,j represents the link between proteins i and j; NDi is the Node Degree of Node i; Nshi,j is the number of nodes that link both node i and j; Solid lines denote primary PPIs. Dashed lines denote correctly those rightly predicted links. Dashed lines denote incorrectly predicted links.

The PPI0 variable in Equation (5 and 6) denotes the adjacent matrix that represents the gold standard links of PPIs in the training dataset. Because we do not consider interactions between two copies of the same protein, the elements of the diagonal of this matrix are all zero. NprU in Equation (8) is the number of unique values in the sub-dataset. If there are few unique values in the sub-dataset (NprU < 10), the Sf of a link is defined as the average value of BPCT of those links with the same preliminary similarity. If there are many unique values in the dataset (e.g. GO group-related Similarity and PCC), the relation curve between the Bayesian probability and the preliminary similarity is plotted and fitted with a linear function. Ffit in Equation (8) denotes the fitting function. The different 'k's in Equations (8, 9, 10 and 11) refer to the different types of similarity. $\mathrm{Sf}_{i,j,k}$ in Equation (8) represents the specific similarity used in PTOMSM. Likewise, St in Equation (11) represents the joint similarity, which will also be used directly in calculating PTOMSM.

Integration of different datasets using a naive Bayesian method

In order to illustrate the effect of our method, we have to compare it with some standard method, and we chose Bayesian

method. The naive Bayesian method described in [16, 17] was used to integrate various datasets. Preliminary similarity in each dataset was binned into discrete intervals. The likelihood ratio (Lr) of each interval can be calculated from the positive and negative matches to the gold standard. Next, the integrated LrT between two proteins is calculated as the product of all the Lr values of the link in the heterogenous datasets. The calculation of LrT can be expressed as shown in Equation (4) , in which f_i denotes the feature value(binned preliminary similarity) in the i-th dataset. LPMB is used to denote the Lr in predicting PPI, and LPMNBI is used to denote the LrT in predicting PPI.

$$LrT(f_1 \ldots f_N) = \prod_{i=1}^{N} Lr(f_i) = \prod_{i=1}^{N} \frac{P(f_i \mid pos)}{P(f_i \mid neg)} \qquad (4)$$

Developing a modified version of the Topological Overlap Measure

Topological Overlap Measure

The pairwise topological overlap measure (TOM) reflects the relative interconnectedness between the two nodes[31]. This property is determined by both the direct link between the two nodes and the indirect paths between the nodes. The TOM is defined in equation (12). The TOM is generally used on unweighted network[11, 32]. However, it can be applied to unweighted networks as well (e.g. MTOM)[12]. Because MTOM is somewhat complicated, this study only compared PTOMSM to the weighted TOM.

$$\mathrm{TOM}_{ij} = \begin{cases} \dfrac{\sum_{u \neq i,j} a_{iu} a_{ju} + a_{ij}}{min\{\sum_{u \neq i} a_{iu} - a_{ij}, \sum_{u \neq j} a_{ju} - a_{ij}\} + 1} & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \qquad (12)$$

Topological Overlap Measure with Similarity Matrix

Motivated by the TOM[31], we defined the PTOMSM as shown in Equation (13). In Equation (13), PPI_{ij} is the primary protein-protein interaction value between protein i and protein j, S_{uj} is the similarity between protein j and protein u. α is a constant that is optimized in the training. We also defined the Parameterized Topological Overlap Matrix without any additional Similarity Matrix (PTOM) as shown in Equation (14).

$$\mathrm{PTOMSM}_{ij} = \begin{cases} \dfrac{\sum_{u \neq i,j} PPI_{iu} S_{uj}}{\alpha * \sum_{u \neq i,j} S_{uj} + 1} + \dfrac{\sum_{u \neq i,j} PPI_{ju} S_{ui}}{\alpha * \sum_{u \neq i,j} S_{ui} + 1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \qquad (13)$$

$$\mathrm{PTOM}_{ij} = \begin{cases} \dfrac{\sum_{u \neq i,j} PPI_{iu} PPI_{uj}}{\alpha \sum_{u \neq i,j} PPI_{uj} + 1} + \dfrac{\sum_{u \neq i,j} PPI_{ju} PPI_{ui}}{\alpha \sum_{u \neq i,j} PPI_{ui} + 1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \qquad (14)$$

Evaluation

Definition of indicator of prediction effectiveness

A better prediction should provide more relations with higher reliability. We define a variable, R, in Equation (15 and 16) as an indicator of the effectiveness of prediction, where AUC is the area under the precision-recall curve(PRC), NposT is the total number of all gold standard positive PPI links in the network, Precision0 is the average Precision in the network, and n

is the total number of nodes in the network.

$$R = AUC \cdot \frac{NposT}{Precision0} \quad (15)$$

$$Precision0 = \frac{NposT}{n \cdot (n+1)} \quad (16)$$

Cross-validation

The primary PPI are divided into N sets of the same size (N-fold cross-validation). In order to compare the influence of the relative size of the samples on predictive strength, we used two kinds of validation under several different values of N:

1) Using 1 of the N sets for training, the (N-1) remaining sets for testing (denoted as Sets(1:N-1), Rsize=1/(N-1));
2) Using (N-1) of the N sets for training, the 1 remaining set for testing (denoted as Sets(N-1:1), Rsize=N-1).

To succinctly present our results in this paper, we mainly showed the results from the studies on Sets(6:1), and the result for the other Ns are presented in the Supplementary Materials.

Results

Fitting the relationship between Preliminary Similarity and BPCT

We found that the Preliminary Similarity of cellular localization displays an obvious positive correlation with BPCT. Moreover, the BPCT value for the Preliminary Similarity of cellular localization is much higher than $\overline{BPCT}$ (see supplemental data). This suggests that the similarity matrix of cellular localization may improve the prediction of indirect links via PTOMSM. This result is also in accordance with the results presented in Table 1.

Table 1. Prediction efficiency of PTOMSM and Bayesian integration (Sets(6:1)). The results at the optimal value of α are presented for PTOMSM.

Type of matrix	R	AUC	NposT	Precision0
TOM	233100	0.01511	92160	0.005974
LPMB*(Co-Text)	56810	0.00899	59990	0.009497
LPMB*(Co-expression)	101300	0.02045	32310	0.006523
LPMB*(Cellular localization)	195100	0.04134	50060	0.0106
LPMB*(Molecular function)	93870	0.01361	57240	0.008302
LPMB*(Biological process)	187100	0.01753	78720	0.007377
LPMB*(Homology)	19230	0.00738	18560	0.007118
LPMNBI	1090000	0.07497	90850	0.00625
PTOM	540300	0.03502	92160	0.005974
PTOMSM*(Co-Text)	735100	0.04765	92160	0.005974
PTOMSM*(Co-expression)	754000	0.04887	92160	0.005974
PTOMSM*(Cellular localization)	2015000	0.1306	92160	0.005974
PTOMSM*(Molecular function)	1153000	0.07475	92160	0.005974
PTOMSM*(Biological process)	1607000	0.1041	92160	0.005974
PTOMSM*(homology)	249100	0.01615	92160	0.005974
PTOMSMt	2523000	0.1635	92160	0.005974

Optimization of the parameter α

In the training process, the parameter α in Equation (13) is adjusted in order to achieve optimal prediction. We obtained the

PRC curves of PTOMSMs with different 'α's, and calculated the AUC and R of each PRC curve (Such PRC curves can be seen in supplemental material). We found that there is positive correlation between the R value for the training data and R value for the testing data (Figure 2). In other words, if the value of R for a training dataset has a high value, the value of R for a testing dataset with the same α tends to be higher. Specifically, the value of α that provides the maximum value of R for the training data is likely to be similar to the α value that provides the maximum value of R for the testing data. This characteristic of the R-α curve can serve as the basis for choosing an optimal value of α with the training data.

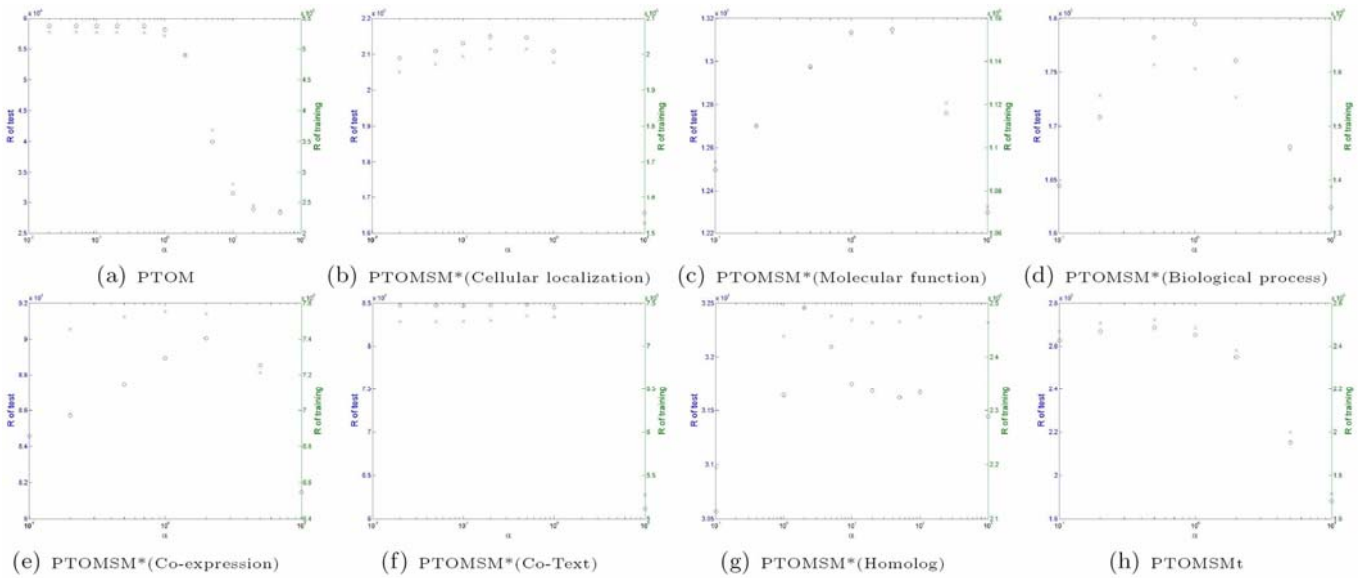

Figure 2: **The influence of the value of α in PTOMSM on the value of R in training and in testing(Sets(6:1)).** The circles denotes evaluation with the training dataset, and the crosses denote evaluation with the test dataset.

Effect of different similarities

As shown in Table 1 and Figure 4, the PTOMSM with joint similarity (PTOMSMj) is the most efficient at predicting PPI, indicating the validity of integration. What's more, the similarities based on cellular localization, biological process, molecular function and co-expression, have a greater effect than other similarities on PPI prediction by PTOMSM. This finding is unsurprising, as the relationship between Co-expression, Co-Text and PPI-related similarity is less direct, so there tends to be a greater level of 'noise' in these two datasets. Interestingly, the PTOMSM with homology-generated similarity performed the worst in PPI prediction. This may be due to the scarcity of links in the similarity matrix, which could result in a sparse PTOMSM matrix. Consequently, this sparse matrix could not cover very many true PPI links. The performance of this similarity metric may also be due to a lack of homology between many proteins (or that the homology has not been identified).

Comparison of PTOMSM and LPMNBI

We calculated the LPMBs and LPMNBI and evaluated them with a test PPI dataset. The indicators of prediction effect are presented in Table 1 and Figure 4, from which we can see that the value of R is the largest for the case of PTOMSMj, indicating that PTOMSMj is more effective at PPI prediction than LPMNBI.

To examine PPI discovery by PTOMSMj and LPMNBI in more detail, we selected PPI candidates according to different cutoffs of final score assigned by the PTOMSMj and LPMNBI analyses. As shown in Figure 3a, PTOMSMj discovers many more PPI links at the same precision or TP/FP than does LPMNBI. Moreover, PTOMSMj discovers many more total PPI links than LPMNBI (Figure 3b). We also found that both methods discover the majority of the PPI links at a high cutoff(Rank > 0.9), indicating that best discovery efficiency (more True Positive and less False Positive) can be achieved by

choosing a high cutoff value.

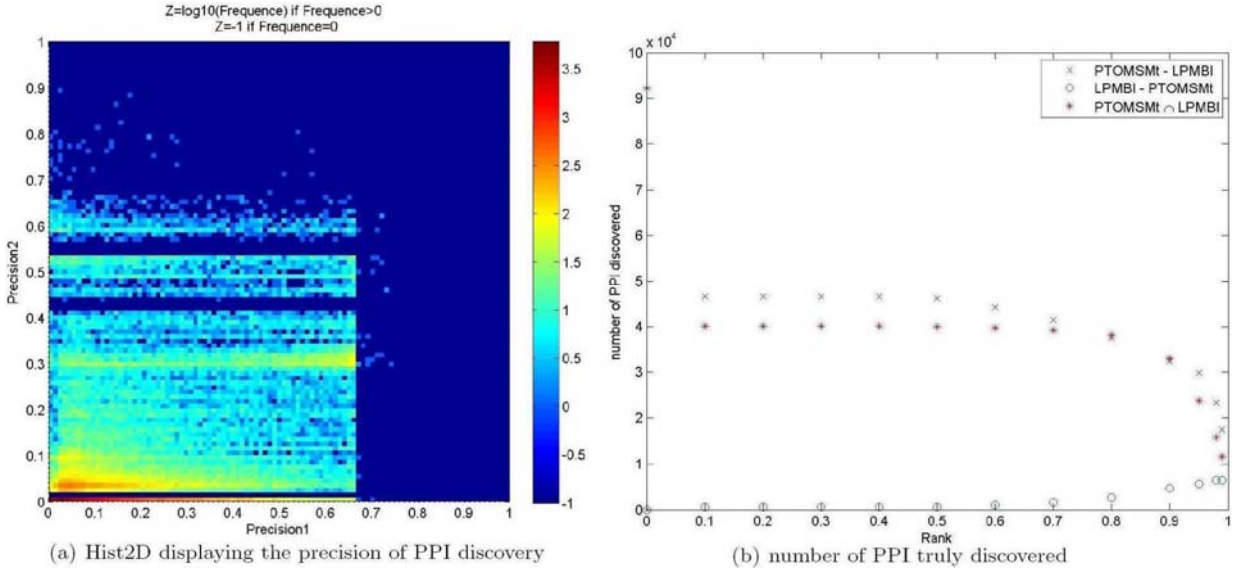

Figure 3: **Comparing PTOMSMt and LPMNBI(Sets(6:1)).**
(a) Two Dimensional Histogram displaying the precision at which the gold standard positive PPIs were discovered. The colors denote the logarithm of density of correctly discovered PPI links. Precision1 indicates the precision of PTOMSM, and Precision2 indicates the precision of LPMNBI. (b)Crosses in the figure denote PPIs discovered only by PTOMSMt. Circles denote PPI discovered only by LPMNBI. Asterisks denote PPI discovered by both PTOMSMt and LPMNBI. The rank in abscissa indicates that when a link is assigned a score by LPMNBI(or PTOMSM) that is higher than x percent of all the links, the rank of that link is x. When a cutoff of rank is chosen, those links with a higher rank than the cutoff are regarded as positive in our prediction.

The influence of the sizes of the training and testing datasets

A paradox in statistics-related prediction is that if the training sample is comparatively large, the prediction tends to be more accurate, but less valuable. On the other hand, if the training sample is comparatively small, the prediction tends to be less accurate, but more valuable. Therefore, we endeavored to examine the influence of dataset size on prediction.

As can be seen in Figure 4, on the whole, cellular localization, biological process, and molecular function-related PTOMSMs exceed Co-Text, Co-expression, and Homology related PTOMSM in PPI prediction. The rank of PTOM (PPI as similarity) among the other PTOMSMs appears quite different from the other metrics. If the training dataset is relatively small, PTOM behaves as one of the worst in PPI prediction; conversely, if the training dataset is relatively large, PTOM is one of the best predictors of PPI. We assume that when there are fewer 'seed' links (PPI links for training), and 'fruit' PPI links (PPI links for test) are more abundant, there is a greater chance for the similarity links (like 'antennae') to grasp the PPI links in the test set. On the other hand, in PTOM, both the 'seed' links and the 'antennae' links are PPI links used in the training. As a result, if few PPI links are used for training, there will be less chance to form cliques with more than three proteins, and PTOM will perform poorly at discovering the remaining PPI links (PPI for test). This effect of the relative size of the training and test dataset additionally illuminates the advantage of PTOMSMj over PTOM, as it is generally more valuable to be able to discover new PPI links when comparatively few PPIs are known.

The sizes of the training and testing datasets also have significant influence on Bayesian integration. As Figure 4 shows, both PMBLs and PMNBIL decrease in their utility of prediction as Rsize increases. On the other hand, the R values for PTOMSMj were significantly larger than the R values of PMNBIL, suggesting that PTOMSMj is more effective at PPI prediction than PMNBIL regardless of the change of the sizes of the training and testing datasets.

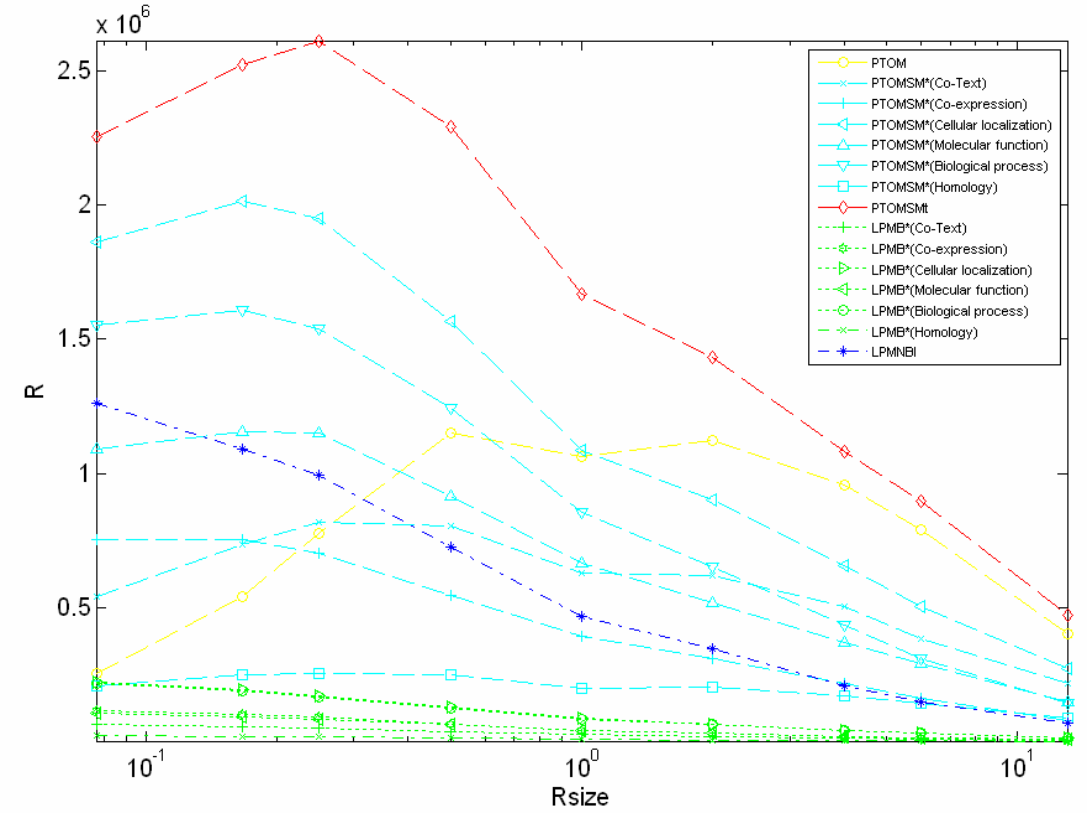

Figure 4: **The influence of Rsize on prediction.**
All the PTOMSMs presented in the figure use optimal values of α. R in ordinate is an indicator of the effectiveness of prediction, and is de fined in Equation (15).

Discussion

Traditional approaches to PPI prediction that integrate diverse datasets can only identify candidate PPI links that already exist in at least one of the integrated sub-datasets. Although many of these methods effectively integrate a large mass of heterogenous data into high-confidence network, many true links remain undiscovered when using these approaches. To increase the scope of selection, we proposed a method called PTOMSM, which is a modified version of the TOM[32]. PTOMSM combines known PPI links with heterogenous relationships and identifies candidate PPI links that do not exist in either the gold standard PPI dataset or in the heterogenous relationship datasets.

We studied the influence of the training and testing dataset size on prediction, finding that almost all of the PTOMSMs and Bayesian integration perform better in predicting PPI when the training dataset is small. The only exception is PTOM, which performs extraordinarily well when the training dataset is big. Because PTOM only employs the information of gold standard PPI links, it can be inferred that PPI related similarity provides better predictive power to PTOMSM for PPI prediction than do other similarity metrics when the size of training dataset is big. On the other hand, when the training dataset is small, PTOM perform much worse than PTOMSMj, indicating that the utilization of a variety of information sources, in addition to PPI alone, is necessary to enhance the prediction of PPIs when there are much more unknown PPIs than known PPIs.

In summary, our PTOMSM approach is effective in PPI prediction in two respects. First, PTOMSM using the joint similarity performs better than PTOMSM using any individual similarity, indicating the effectiveness of integration. Second, when using the six relationship datasets in combination with PPI dataset, the PTOMSM method shows better predictive power than the traditional naive Bayesian method, because PTOMSM could identify more reliable PPI links. In all

likelihood, this method could also be applied to prediction of other regulatory links, such as transcriptional regulation, a field of inquiry that is part of our future work. Additionally, this prediction method may be generalized to discover additional indirect links in a manner similar to the generalization of TOM to GTOM[32]. Although our studies demonstrate that PTOMSMj exhibits some merits in predicting PPI, our findings do not imply that PTOMSMj will perform better than LPMNBI in all cases. It is possible that the utilization of other datasets, such as orthology interaction and domain interaction, may enable LPMNBI to outperform PTOMSMj. Therefore, we expect an approach combining PTOMSM and LPMNBI could identify more reliable candidate links.

Acknowledgements

The authors wish to thank Chaozu He for instruction in biological research and helpful financial support, Lei Zhang for instruction in programming, Junling Wang for discussions of network related research and Weixiang Liu for discussion of Mathematics.

References

1. Li, J., X. Li, H. Su, H. Chen, and D.W. Galbraith, *A framework of integrating gene relations from heterogeneous data sources: an experiment on Arabidopsis thaliana.* Bioinformatics, 2006. **22**(16): p. 2037.
2. Jansen, R., N. Lan, J. Qian, and M. Gerstein, *Integration of genomic datasets to predict protein complexes in yeast.* Journal of Structural and Functional Genomics, 2002. **2**(2): p. 71-81.
3. Srinivasan, B.S., N.H. Shah, J.A. Flannick, E. Abeliuk, A.F. Novak, and S. Batzoglou, *Current progress in network research: toward reference networks for key model organisms.* Briefings in Bioinformatics, 2007. **8**(5): p. 318.
4. Lu, H., B. Shi, G. Wu, et al., *Integrated analysis of multiple data sources reveals modular structure of biological networks.* Biochemical and Biophysical Research Communications, 2006. **345**(1): p. 302-309.
5. Troyanskaya, O.G., K. Dolinski, A.B. Owen, R.B. Altman, and D. Botstein, *A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae).* Proceedings of the National Academy of Sciences, 2003. **100**(14): p. 8348.
6. von Mering, C., L.J. Jensen, B. Snel, et al., *STRING: known and predicted protein-protein associations, integrated and transferred across organisms.* Nucleic acids research, 2005. **33**(Database Issue): p. D433.
7. Jensen, L.J., M. Kuhn, M. Stark, et al., *STRING 8--a global view on proteins and their functional interactions in 630 organisms.* Nucleic Acids Research, 2009. **37**(Database issue): p. D412.
8. Mering, C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, *STRING: a database of predicted functional associations between proteins.* Nucleic Acids Research, 2003. **31**(1): p. 258.
9. Ng, A., B. Bursteinas, Q. Gao, E. Mollison, and M. Zvelebil, *pSTIING: a'systems' approach towards integrating signalling pathways, interaction and transcriptional regulatory networks in inflammation and cancer.* Nucleic Acids Research, 2006. **34**(Database Issue): p. D527.
10. McDowall, M.D., M.S. Scott, and G.J. Barton, *PIPs: human protein-protein interaction prediction database.* Nucleic Acids Research, 2009. **37**(Database issue): p. D651.
11. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Statistical Applications in Genetics and Molecular Biology, 2005. **4**(1): p. 1128.
12. Li, A. and S. Horvath, *Network neighborhood analysis with the multi-node topological overlap measure.* Bioinformatics, 2007. **23**(2): p. 222.
13. Horvath, S., B. Zhang, M. Carlson, et al., *Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target.* Proceedings of the National Academy of Sciences, 2006. **103**(46): p. 17402.
14. Yip, A.M. and S. Horvath, *Gene network interconnectedness and the generalized topological overlap measure.* BMC bioinformatics, 2007. **8**(1): p. 22.

15. Frech, C., M. Kommenda, V. Dorfer, T. Kern, H. Hintner, J.W. Bauer, and n.K. O, *Improved homology-driven computational validation of protein-protein interactions motivated by the evolutionary gene duplication and divergence hypothesis.* BMC bioinformatics, 2009. **10**(1): p. 21.
16. Jansen, R., H. Yu, D. Greenbaum, et al., *A Bayesian networks approach for predicting protein-protein interactions from genomic data*. 2003.
17. Rhodes, D.R., S.A. Tomlins, S. Varambally, et al., *Probabilistic model of the human protein-protein interaction network.* Nature biotechnology, 2005. **23**(8): p. 951--959.
18. Salwinski, L., C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, and D. Eisenberg, *The database of interacting proteins: 2004 update.* Nucleic acids research, 2004. **32**(Database Issue): p. D449.
19. Stark, C., B.J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, *BioGRID: a general repository for interaction datasets.* Nucleic acids research, 2006. **34**(Database Issue): p. D535.
20. Kerrien, S., Y. Alam-Faruque, B. Aranda, et al., *IntAct--open source resource for molecular interaction data.* Nucleic acids research, 2007. **35**(Database issue): p. D561.
21. Ben-Hur, A. and W. Noble, *Choosing negative examples for the prediction of protein-protein interactions.* BMC bioinformatics, 2006. **7**(Suppl 1): p. S2.
22. O'Brien, K.P., M. Remm, and E.L.L. Sonnhammer, *Inparanoid: a comprehensive database of eukaryotic orthologs.* Nucleic acids research, 2005. **33**(Database Issue): p. D476.
23. Ashburner, M., C.A. Ball, J.A. Blake, et al., *Gene Ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25--29.
24. Jenssen, T.K., A. L|[aelig]|greid, J. Komorowski, and E. Hovig, *A literature network of human genes for high-throughput analysis of gene expression.* Nature Genetics, 2001. **28**(1): p. 21-28.
25. Lee, I., S.V. Date, A.T. Adai, and E.M. Marcotte, *A probabilistic functional network of yeast genes*. 2004.
26. Cherry, J.M., C. Ball, S. Weng, et al., *Genetic and physical maps of Saccharomyces cerevisiae.* NATURE-LONDON-, 1997: p. 67-74.
27. Gasch, A.P., M. Huang, S. Metzner, D. Botstein, S.J. Elledge, and P.O. Brown, *Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p.* Molecular biology of the cell, 2001. **12**(10): p. 2987.
28. Spellman, P.T., G. Sherlock, M.Q. Zhang, et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Molecular biology of the cell, 1998. **9**(12): p. 3273.
29. Gasch, A.P., P.T. Spellman, C.M. Kao, et al., *Genomic expression programs in the response of yeast cells to environmental changes.* Molecular biology of the cell, 2000. **11**(12): p. 4241.
30. Mnaimneh, S., A.P. Davierwala, J. Haynes, et al., *Exploration of essential gene functions via titratable promoter alleles.* Cell, 2004. **118**(1): p. 31-44.
31. Ravasz, E., A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabasi, *Hierarchical organization of modularity in metabolic networks*. 2002.
32. Yip, A.M. and S. Horvath, *The generalized topological overlap matrix for detecting modules in gene networks.* Biocomp, Las Vegas, NV, USA, 2006: p. 451--457.