A method to integrate, assess and characterize the protein-protein interactions

Fa Zhang[1], Lin Xu[1,2], Jingchun Chen[3], Zhiyong Liu[4] and Bo Yuan[3]
1, Key Laboratory of Compute r System and architecture, the Institute of Computing Technology, Chinese Academy of Sciences
2, Graduate School, Chinese Academy of Sciences
3, Department of Biomedical Informatics, The Ohio State University
4, National Natural Science Foundation of China

***Abstract*-Recently, large-scale protein-protein interactions were recovered using the similar two-hybrid system for the model systems. This information allows us to investigate the protein interaction network from a systematic point of view. However, experimentally determined interactions are susceptible to errors. A previous assessment estimated that only ~10% of the interactions can be supported by more than one independent experiment, and about half of the interactions may be false positives. These false positives might unnecessarily link unrelated proteins, resulting in huge apparent interaction clusters, which complicate elucidation for the biological importance of these interactions.**

Address this problem, we present an approach to integrate, assess and characterize all available protein-protein interactions in model organisms yeast and fly. We first integrate all available protein-protein interaction databases of yeast and fly, and merge all the datasets. We then use machine learning techniques to score the reliability for each interaction, and to rigorously validate the scoring scheme of yeast protein-protein interactions from different aspects. Our results show that this scoring scheme provides a good basis for selecting reliable protein-protein interaction dataset.

I. INTRODUCTION

Protein-protein interactions define the core of system biology, since they mediate the formations of protein complexes that exert certain cellular functions. Also, protein-protein interactions are responsible for modulating one protein's function by another protein, such as controlling an enzyme's activity.

Traditional molecular biology has witnessed enormous success on elucidating protein-protein interactions[1]. It is the application of high-throughput techniques (yeast two-hybrid, affinity chromatography, and mass-spectrometry) that makes it possible to identify and analyze protein-protein interactions at genomic scale. These techniques have been applied to investigate the protein-protein interaction maps of the budding yeast *Saccharomyces cerevisiae* [2] and fruit fly [3]. Along with these large-scale studies, databases were created to collect and annotate this large amount of information, such as MIPS[4], DIP[5], GRID[6], BIND[7], STRING[8]. These repositories greatly promote the scientific discoveries through protein-protein interactions.

However, two major issues arise, the reliability of the protein-protein interactions identified through these high-throughput techniques and the heterogeneity of the databases. Studies have shown that for yeast *Saccharomyces cerevisiae*, only small portion of the results obtained through different high-throughput techniques overlap [2]. This suggested that false positives are common among these interactions. Machine learning approaches and regression method were applied to gain insights into the confidence of yeast protein-protein interaction datasets, and encouraging results were obtained. However, incompleteness of the datasets covered

1-4244-0623-4/06/$20.00 ©2006 IEEE

in these studies largely limited their usage for further investigations in the scientific community. We believe this was at least in part caused by the second issue mentioned above, i.e., the heterogeneity nature of the protein-protein interaction databases. Yeast genes are referenced either by the standard Open Reading Frame (ORF), or by native names based on function. Interaction information is organized either by plain text format, XML format, or web-searchable database. Gene identification is cross- cross-referenced to different databases, which makes data integration even more difficult.

Our aims in this study are to assess the quality for each of the protein interaction dataset using an optimizing technique. We first comprehensively integrate and merge all available interaction databases of yeast and fly using cross-reference identification and sequence information. We then score the reliability for each interaction by an evolutionary-strategy algorithm. Experimental results show that this method can filter out a significant portion of reliable protein-protein interaction dataset.

II. MATERIALS AND METHODS

We first downloaded all available protein-protein interaction datasets of yeast and fly from 11 public accessible databases. Then we merged the datasets together and removed the redundancies using cross-reference identification and sequence information, thus each interaction was unique and is associated with one or more methods that identified the interaction. We next assigned a weight to each of the methods and obtained a score for each interaction using evolutionary strategy algorithm. We then filtered the reliable interaction dataset according to a series of confidence score cutoff. Final we used different criteria to assess the soundness of our scoring scheme.

A. Dataset

All the protein-protein interaction datasets in this study were obtained from public-accessible databases. As summarized in tableⅠ, for yeast, we downloaded 333,594 interactions

TABLE Ⅰ

SUMMARY OF THE PROTEIN-PROTEIN INTERACTIONS DATASETS IN THIS STUDY

	Methods	No. of Entries	No. of Unique Entries	Source
YEAST	HMS-PCI	39,609	35,542	BIND, GRID, MINT, DIP
	CO-IP	23,750	20,497	BIND, MINT, Mering
	Y2H	23,635	7,275	BIND, GRID, MINT, DIP
	Genetic	22,312	17,538	GRID, Curagen
	Co-expression	15,678	15,677	Mering
	Complex	75,493	75,493	MIPS, SGD
	Prediction	6,746	6,733	Mering
	STRING	102,164	102,164	STRING
	Other	4,728	4,616	BIND, DIP, MINT, GRID
	NA	19,479	19,431	DIP, GRID, Literature
	Total	333,594	**239,789**	10
FLY	Y2H	84,842	41,545	BIND, DIP, Curagen
	Genetic	11,368	7,978	GRID, FlyDB
	STRING	35,463	35,463	STRING
	NA	38,633	38,633	DIP, GRID, BIND, Literature
	Total	170,306	**86,295**	7

(239,789 unique interactions) from 10 individual databases, for fly, this datasets contains 170,306 interactions (86,295 unique) form 7 databases.

B. Interaction Confidence Score

In the integrated interaction dataset, there are total of 21 different approaches used to detect the interactions of yeast, and 8 different methods used for fly. Each approach is assigned a confidence score to represent the likelihood it actually occurs in cells, based on the types of methods that have detect a given interaction and the number of times it was detected. For each interaction we sum up the reliability scores for all the approaches that identified this interaction:

$$S_R = \sum R_i \quad (i = 1,2\ldots n)$$

Where n is the number of methods that detected the interaction, and R_i is the reliability score for the i^{th} method. For each dataset, the mean raw score μ and the standard deviation δ were calculated. The normalized confidence score of each interaction is the mahalanobis distance between its raw score to the mean value:

$$S = (S_R - \mu) / \delta$$

This normalized score will be assigned to each interaction as the confidence score.

C. Optimization by Evolutionary Strategy

Since the confidence score for each interaction is dependent on the reliabilities of the methods, we need to assign a reasonable reliability score to each of methods used to detect interactions. Here we use evolutionary strategy algorithm to optimize this assignment.

First we build a training dataset for this purpose. The positive training dataset includes the interactions that 1) detected by more than one independent method; 2) appear in at least four databases; 3) the two interacting genes co-localize in high-throughput localization experiments. The negative training dataset includes the interactions that 1) detected by only one method; 2) appear in only one database; 3) the two interacting genes do not co-localize in high-throughput localization experiments. This way for yeast we obtained 12,369 reliable interactions, 11,570 unreliable interactions. For fly we obtained 31,482 reliable interactions, and 44,740 unreliable interactions. The fitness function for a set of method reliability score is defined as the difference between the average confidence score of the reliable and unreliable datasets.

Finally evolutionary strategy algorithm was used to maximize the fitness function. Here, we used the ECJ package (http://cs.gmu.edu/~eclab/projects/ecj/) to implement this optimization.

III. Results And Discussions

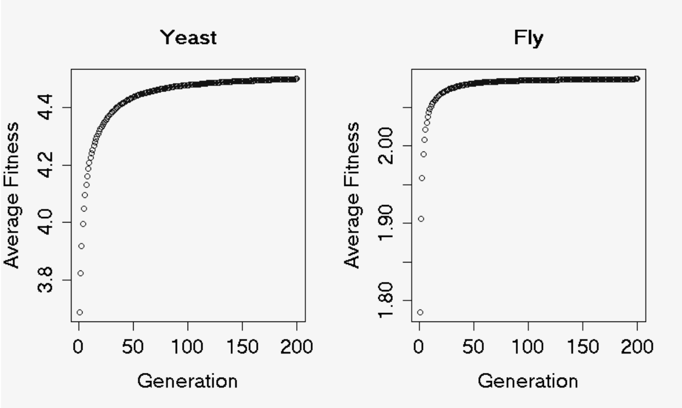

Fig.1.The optimized weights better distinguish reliable and unreliable interactions.

Varieties of methods have been used to identify gene-gene interactions. Each method has advantages and intrinsic drawbacks. This suggests that the likelihood that a specific interaction is real may be inferred from the types of methods that detected it and the number of times it was detected. Comparing to random assignment, we found that the optimized method reliability scores greatly increases the score difference between the reliable and the unreliable interactions, as shown in Fig.1. In this figure, X-axis means evolution generation, and y-axis is average fitness of the population at each generation. In yeast dataset, optimization process increases the score difference by almost one standard deviation. The improvement is less dramatic for fly dataset, due to the fact that fewer methods were used in fly dataset.

A. *Validation through network topology*

Recently many naturally occurring complex networks have been shown to be scale-free, such as yeast protein-protein interaction network[9]. Here, we evaluated the validity of confidence score assignment through scale-free network fitting. In principle, if assigned confidence scores are truly meaningful, then using a higher cutoff to select interaction dataset we should obtain a network that fits better into a scale-free model.

One of the most distinguishable feature of a scale-free network is that the degree distribution of the nodes follows a power law, $\rho(\mathrm{k}) \sim \mathrm{k}^{-\mathrm{r}}$[9]. That is, under logarithmic scale this distribution approximates a straight line, $ln(\rho(\mathrm{k})) \sim -\mathrm{r}ln(\mathrm{k})$. Therefore, we chose a series of cutoff score to select reliable dataset for yeast and fly, and fitted the degree distributions of the resulted network to straight lines. As shown in Fig. 2, the R-square value of the fitting is about 0.7 for the whole dataset of yeast. This poor fitting suggests that the overall dataset contains many false interactions. These false interactions tend to be random and change the scale-free topology of the real network. The quality of the fitting increases along with the increase of selection score cutoff, suggesting the increasing enrichment of true interactions in the dataset. Very good fitting is achieved if the cutoff score between 0.8~0.9, at which point the R-square of the fittings reaches over 0.95 for both the yeast and fly dataset. It is interesting to note that the quality of the fitting will decrease if we further increase the score cutoff. This suggests that the dataset with very high confidence scores is biased towards those highly conserved interactions. Therefore, using cutoff between 0.5 and 1.5 we obtained a highly reliable dataset for yeast, and for fly the good cutoff score interval is 0.3~1.6, as shown in Fig. 2A.

Another important characteristic feature of a scale-free network is the scaling exponent in the degree distribution, which can be calculated from the slope of the straight line in the degree distribution fitting[9]. As shown in fig 3B, using cutoff between 0.5 and 1.5 on yeast dataset we obtain the scaling exponent value between 2.1 and 2.3. For fly dataset, we obtain the scaling exponent between 2.3 and 2.6, corresponding to the cutoff score of 0.3~1.6. This further suggests that our confidence score assignment is valid and robust.

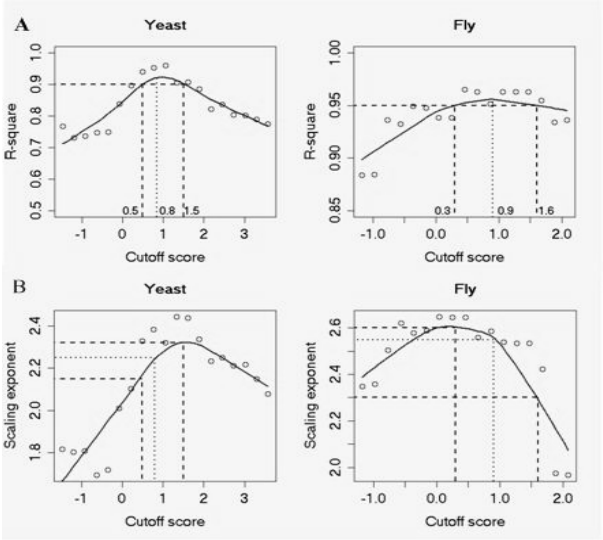

Fig.2. Validation of interaction confidence score through scale-free model fitting.

B. *Validation Through Gene Ontology Homogeneity*

After validating through topological modeling, we next checked the validity of the confidence score assignment through functional annotations. If the gene-gene interaction dataset is reliable, in most cases a gene's interacting partners are likely to participate in the same or closely related cellular processes. Gene Ontology (GO)[10] provides hierarchical annotation of cellular processes for genes, and the homogeneity of a group of genes can be assessed through information content of their common annotations[11].

We chose a series of score cutoffs to select reliable dataset. For each obtained network, we calculated the homogeneity index of each gene’s interacting partners. As shown in Fig.3, if the whole dataset is used in the network, the homogeneity index of the genes is very low. The higher the score cutoff is used, the higher the homogeneity index of the genes. The Pearson correlation coefficient between the cutoff score and the average homogeneity index is 0.972 for yeast, $\mathrm{p}<10^{-7}$, and 0.884 for fly, $\mathrm{p}<10^{-4}$. A jump of homogeneity index is observed at the cutoff score of around 0, suggesting the data quality greatly improved.

This is in agreement with the quality of scale-free fitting (Fig.2). Therefore, these results strongly suggest that our confidence score assignment closely represents the likelihood of an interaction actually exists in the cell.

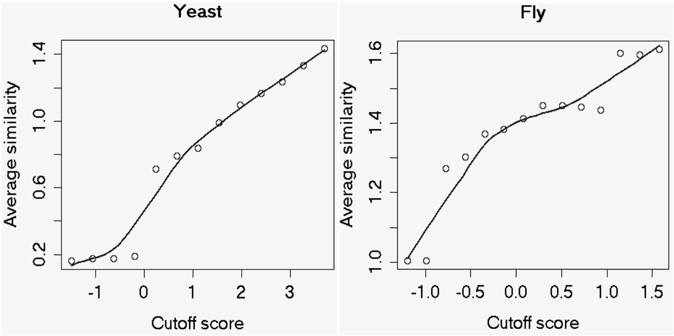

Fig.3. Validation of interaction confidence score through homogeneity of Gene Ontology annotation. It indicates the higher the cutoff score, the more likely a gene's interacting partners participate in the same cellular process.

C. Validation of yeast dataset through clustered interaction

Another way to validate interaction dataset at functional level is to cluster genes into different cellular processes and assess the pairwise interactions between all the cellular processes. In principle, a reliable dataset should contain interactions mostly within the same cellular process[12]. Both GO and MIPS database provide cellular process annotations for yeast, and the annotations can be clustered into high-level processes.

Here we compare the whole interaction dataset of yeast to the selected dataset with confidence score above 0.8, using both GO and MIPS clustering results. As shown in Fig.4, yeast genes were classified into groups based on the cellular processes, according to the annotation in GO (panel A, B) and MIPS (panel C, D). In each plot, x and y axes were the cellular processes, and the grey scale of each intersecting area shows the interaction density, which is the number of interactions between proteins from the two groups. The diagonal represents the interactions between proteins in the same cellular process.

In Fig.4, the majority of the interactions in the whole dataset are between genes from different cellular processes (off diagonal in panel A and C), that means most of such interactions are likely to be false positives. On the contrary, the majority of the interactions in selected dataset are between genes from the same cellular process (on diagonal in panel B and D). Most of the spurious interactions are removed. This result further confirms the validity of our confidence score assignment.

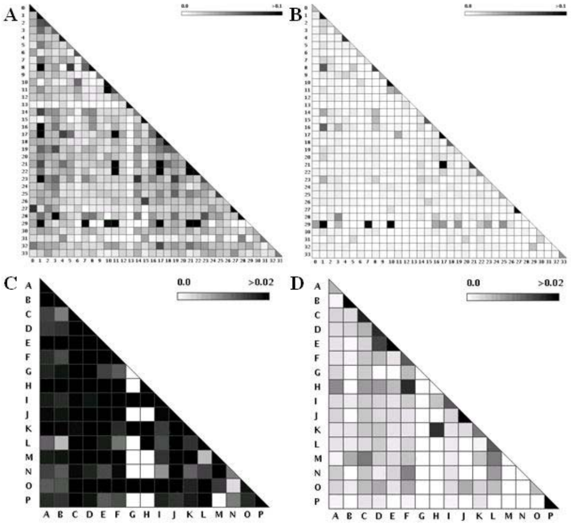

Fig.4. Validation of interaction confidence score through pairwise functional similarity

D. Validation of yeast dataset through frequent network motifs

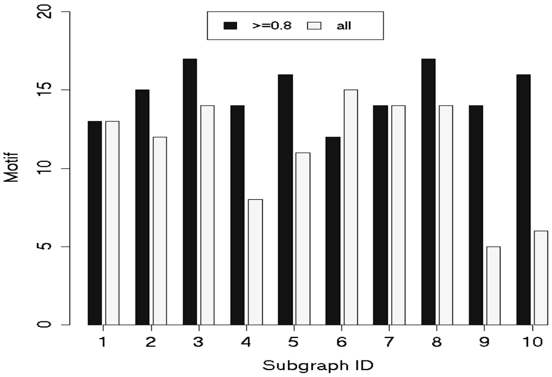

Fig.5. Validation of interaction confidence score through significant network motif mining

Small frequent motifs are considered the building blocks of complex networks [13]. Frequent motifs can be identified by comparing the occurrence of a motif in the network to the occurrence of the same motif in equivalent random networks obtained through permuting the original network. Only the motifs that occur significantly more often in the real network than in the random networks are considered frequent network

TABLE Ⅱ

SUMMARY OF THE STATISTICS FOR THE RELIABLE INTERACTIONS IN THE DATASET WITH SCORE >= 0.8

	Whole dataset	Dataset with score >=0.8
Number of genes	5,986	5,038
Number of interactions	23,9789	24,786
Database occurrence >= 2	39,149 (16.3%)	24,733 (99.8%)
Independent method >= 2	39,683 (16.6%)	24,775 (99.9%)
Same GO clustering	29,552/114,693 (25.8%)	5,524/12,675 (43.6%)
Same MIPS clustering	25,620/75,682 (39.5%)	4,676/7,233 (64.6%)
R^2 of scale-free fitting	0.78	0.95
Neighbor GO homogeneity	0.16	0.81

motifs [13]. Since the false interactions tend to be random, they represent noise in the whole dataset and may actually decrease the number of frequent motifs to be uncovered.

Here we used network motif detection to validation the confidence score assignment. First we constructed two yeast interacting networks based on the whole dataset and the selected reliable dataset (cutoff score >= 0.8). Then each network was partitioned into 10 subgraphs using graph partition software *hmetis*[14]. Next, each subgraph was mined for frequent network motifs of size 5 using the software *mfinder*[15]. We compared the number of identified network motifs in the selected dataset to the number in all dataset, and we found that the selected reliable dataset had more frequent motifs than all dataset in 7 out of the 10 subgraphs. Statistical analysis indicates that the network obtained using 0.8 cutoff score contains more frequent network motifs of size 5 than the network of the whole dataset, as shown in Fig.5. This again strongly suggests that the assigned confidence score is a reasonable metric that can be used to select real interactions from available datasets.

From the above analysis we believe that we can obtain a reliable gene-gene interaction dataset for yeast by using 0.8 as the score cutoff. As summarized in tableⅡ, this dataset contains 24,786 interactions between 5,038 genes. Essentially all interactions in this dataset were annotated by at least two databases, and were detected by at least two independent methods. The degree distribution of the reliable yeast dataset fits well with the scale-free model, and the GO homogeneity index of genes in the reliable dataset is much higher than that of the whole dataset. Also the interactions between the same cellular processes are largely enriched, comparing to the whole dataset.

IV. CONCLUSION

In this paper, we presented our approaches to integrate, assess and characterize the protein-protein interactions in model organisms yeast and fly. First we downloaded all the interaction datasets from 11 public accessible databases. We then merged the datasets using cross-reference identification and sequence information so that each interaction is unique and is associated with one or more methods that identified the interaction. Next we assigned a weight to each of the methods, optimized using an evolutionary strategy algorithm. For each interaction we summed up the weights of its methods to obtain a raw score. The raw scores of all interactions were normalized against the mean and standard deviation, and the normalized score was assigned to each interaction as its confidence score.

We used a series of criteria to assess the soundness of our scoring scheme. We first used a series of confidence score cutoff to select an interaction dataset and fitted the corresponding network with scale-free network model. We found that the higher the score cutoff used, the better the scale-free fitting. We studied the functional homogeneity index of the interacting partners of each protein, using both the GO and MIPS database annotations. We found that the higher the score cutoff used, the higher the average functional homogeneity of the genes in the network. Furthermore, we found that interactions with higher confidence score are more likely to be between genes of the same cellular pathway. We also used network motif detection to explore the whole yeast dataset and selected reliable dataset. We found that more frequent network motifs are statistically significant in the selected reliable network than in the whole dataset, indicating that the noise in the whole dataset is largely removed in the selected dataset. Overall, our studies show that our scoring scheme provides a good basis for selecting reliable protein-protein interaction dataset.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China project under 60503060 and key project under 90612019.

REFERENCE

[1]. Hartwell, L.H., et al., From molecular to modular cell biology. Nature, 1999. 402(6761 Suppl): p. C47-52.

[2]. Ito, T., et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A, 2001. 98(8): p. 4569-74. Epub 2001 Mar 13.

[3]. Giot, L., et al., A protein interaction map of Drosophila melanogaster. Science, 2003. 302(5651): p. 1727-36. Epub 2003 Nov 6.

[4]. Mewes, H.W., et al., MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res, 2004. 32(Database issue): p. D41-4.

[5]. Pereira-Leal, J.B., A.J. Enright, and C.A. Ouzounis, Detection of functional modules from protein interaction networks. Proteins, 2004. 54(1): p. 49-57.

[6]. Breitkreutz, B.J., C. Stark, and M. Tyers, The GRID: the General Repository for Interaction Datasets. Genome Biol, 2003. 4(3): p. R23. Epub 2003 Feb 27.

[7]. Bader, G.D., D. Betel, and C.W. Hogue, BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res, 2003. 31(1): p. 248-50.

[8]. von Mering, C., et al., STRING: a database of predicted functional associations between proteins. Nucleic Acids Res, 2003. 31(1): p. 258-61.

[9]. Barabasi, A.L. and Z.N. Oltvai, Network biology: understanding the cell's functional organization. Nat Rev Genet, 2004. 5(2): p. 101-13.

[10]. Harris, M.A., et al., The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res, 2004. 32(Database issue): p. D258-61.

[11]. Dormitzer, P.R., et al., Structural rearrangements in the membrane penetration protein of a non-enveloped virus. Nature, 2004. 430(7003): p. 1053-8.

[12]. von Mering, C., et al., Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 2002. 417(6887): p. 399-403. Epub 2002 May 8.

[13]. Milo, R., et al., Network motifs: simple building blocks of complex networks. Science, 2002. 298(5594): p. 824-7.

[14]. Karypis, G. and V. Kumar, hMetis: A Hypergraph Partitioning Package. Technical Report, 1998.

[15]. Kashtan, N., et al., Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics, 2004. 20(11): p. 1746-58.