The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Uncovering differentially expressed pathways with protein interaction and gene expression data

Yu-Qing Qiu[1,2,*] Shihua Zhang[1,2] Xiang-Sun Zhang[1,†]

[1]Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing 100190, China
[2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China

Abstract The identification of genes and pathways involved in biological processes is a central problem in systems biology. Recent microarray technologies and other high-throughput experiments provide information which sheds light on this problem. In this article, we propose a new method to identify differentially expressed pathways via integration of gene expression and interactomic data in a sophisticated and efficient manner. Specifically, by using signal to noise ratio to measure the differentially expressed level of networks, this problem is modeled as a mixed integer linear programming problem (MILP). The results on yeast and human data demonstrate that the proposed method is more accurate and robust than previous ones.

Keywords Molecular interaction network; gene expression; pathway; mixed integer linear programming; signal to noise ratio.

1 Introduction

With the rapid development of high-throughput technologies, a huge number of experiment datasets have been generated, including gene expression data [1], protein-protein interaction [12] and protein-DNA interaction [14] data, etc. These data give insights into gene functions and cellular molecular mechanisms, which are essential topics in systems biology.

Cells achieve biological functions by pathways composed of genes or proteins and their chemical and physical interactions which can be considered as molecular interaction networks. Interactomic data and gene expression data are well known information sources to study the cellular regulation mechanism among genes and reveal the functional pathways. Previous works mainly analyzing only one kind of data may lead to bias or extract information with little biological meaning[15]. Recently, a new and more reasonable trend of studying biological pathways is to integrate various biological information sources, such as experiment datasets or prior established biological knowledge [18]. Some pathway analysis approaches, such as MAPPFinder [4] and GSEA [17], were

*Email: yqqiu@amss.ac.cn.
†Corresponding author. Email: zxs@amt.ac.cn.

developed to detect differentially expressed pathways from GO function categories [2], or pathway databases such as KEGG [9] and GenMAPP [9], via scoring enrichment of differentially expressed genes within a pathway [3]. Although these methods can detect subtle but coherent gene expression changes, a major drawback is that they cannot discovery new pathways correlated to phenotypes or diseases which have no record in pathway database. On the other hand, another kind of methods is to integrate interactome with gene expression data to identify pathways [8]. Interactome data including protein-protein interaction, protein-DNA interaction data etc. is represented as a network, with nodes corresponding to genes or gene products, and edges corresponding to physical interactions between genes. These methods implement a score function to evaluate differentially expressed levels of subnetworks between different conditions or phenotypes. Then, the connected subnetworks with high sores are found from whole network via optimization techniques, such as simulated annealing [8]. These identified subnetworks are considered as pathways response to specific phenotypes or conditions. However, the simulated annealing is a random optimization method which depends seriously on the initial solution and is hard to determine the optimal parameters.

In this paper, to overcome the above mentioned problems, we propose a novel method to detect differentially expressed pathways from molecular networks based on both interactomic data and gene expression data. Our approach assumes that the majority of genes in a differentially expressed pathway are more likely differentially expressed and these genes are connected as a subnetwork by molecular interaction. These assumptions are supported by the work of [5] where genes with similar expression profiles are shown to be more likely to encode interacting proteins. We model the pathway identification problem by utilizing the signal to noise ratio (SNR) which is a nonparametric statistical measure of differentially expressed level to score a subnetwork or pathway. To detect high scoring subnetworks, an exact searching strategy based on mixed integer linear programming (MILP) is proposed. We test the proposed method or MILP method on yeast molecular interaction networks with microarray profile data. Comparison with other methods shows that the proposed approach is robust and more accurate, and does not depend on the initial solution. The resulted subnetworks cover significantly more known genes corresponding to conditions, and GO function enrichment analysis indicates that the identified subnetworks are more related to conditions. These results demonstrate the effectiveness and efficiency of the method for extracting differentially expressed pathways.

2 Materials and Methods

In this paper, we model a pathway as a connected subnetwork in a molecular interaction network with the following procedure. Firstly, the signal to noise ratio measuring differentially expressed level is calculated for each gene based on gene expression data. Then, by taking each gene as a root, the subnetworks including the root gene are identified from the molecular interaction network by a mixed integer linear programming model. Next, the density distributions of the scores of subnetworks of different sizes are estimated using a non-parameter kernel density estimation method, and a percentile is calculated to distinguish the significant subnetworks. Finally, all significant subnetworks are mapped to the original network to generate an integrated differentially expressed pathway.

2.1 Data

Microarray dataset corresponds to the experiment of GAL80 perturbation [7] was used to test the present method. A small yeast molecular network containing 331 genes and 362 protein-protein and protein-DNA interactions with the mRNA expression data was used to study the galactose utilization pathway.In this paper, we do not distinguish a gene from its product protein. Whatever an interaction occurs in protein-protein or protein-DNA is all considered as a molecular interaction.

2.2 Statistics and Significant Testing

The score of a differentially expressed gene is measured by absolute signal to noise ratio (SNR) metric[6]:

$$t_i = \frac{|\mu_{i1} - \mu_{i2}|}{\sigma_{i1} + \sigma_{i2}}, \tag{1}$$

where μ_{i1} and μ_{i2} are the means of the expression levels of gene i in sample set 1 and sample set 2 respectively, and σ_{i1} and σ_{i2} are the standard deviations of gene i in sample set 1 and sample set 2 respectively. The up-expressed genes and down-expressed genes are both regarded as differentially expressed. Other difference measurement, such as t-statistic can also be implemented. However, the SNR statistic reflects the correlation structure of the data without assumption of any hypothesis about the statistical distribution of the samples which would have to be verified [13], and it can be computed empirically if the selected attributes are meaningful in a statistical sense by a hypothesis contrast.

For a subnetwork containing a gene set $S = \{g_1, g_2, \ldots, g_k\}$, the score of differentially expressed level is calculated as follows:

$$W(S) = \frac{1}{k} \sum_{i=1}^{k} t_i. \tag{2}$$

A subnetwork with a high W value indicates that it expresses differentially. This statistic is not related to the size of gene set. All statistic values of subnetworks with different sizes can be considered to follow the same density distribution.

To obtain the significant subnetworks, we can calculate the p-value representing significant levels of subnetworks identified. We can also apply a nonparametric permutation test method, which estimates the distribution of the statistics $W(S)$ using the permutations of genes or sample labels to compute the p-values. However, in this paper we found that both of them cannot reliably extract significant networks, i.e. they got large subnetworks which are not significant. In other words, these two methods are not appropriate to obtain a significant subnetwork. Thus, instead we implemented the kernel density estimation method to obtain the density function of W and $(1-\alpha)$ $percentile$, and then reported the subnetworks with scores higher than the $(1-\alpha)$ $percentile$. As the density functions are multimodal (see Figure 1), nonparametric kernel density estimation method provided in Matlab (http://www.mathwork.com/) is applied. The $(1-\alpha)$ $percentile$ $w_{(1-\alpha)}$ is calculated via solving the following equation

$$\frac{\int_0^{w_{(1-\alpha)}} Pr(w)dw}{\int_0^{\infty} Pr(w)dw} = 1-\alpha, \tag{3}$$

where $Pr(w)$ represents the density function of W. Since the density function of W has two peaks, i.e. one is near 0 and the other is bigger than 0 (see Figure 1), significantly lager W values should lie on the right side of the second peak. Thus, α is selected empirically such that the $(1-\alpha)$ $percentile$ is higher than the value which reaches the second density peak.

2.3 A mixed integer linear programming model

The molecular interaction network can be represented as an undirected graph $G = (V,E,T)$, where V represents the set of N genes, E represents the set of interactions between nodes and T is the set of weights which are assigned by the SNR value t_i of each gene. The goal is to search connected subgraphs with the highest W score of (2).

Searching the connected subgraph with the maximum score is an NP-hard problem [8]. However, Lee and Dooly studied the constrained maximum-weight connected graph problem [10] and proposed several algorithms which can be applied to this problem. As the objective function W is a linear function of the nodes' weights given the number of nodes in the subgraph, the problem of finding the connected subgraph with the maximum score can be modeled by a constrained maximum-weight connected graph problem. Hence, given a specified root node v_1 and specified number R of nodes in a subgraph, we propose a new method to find the largest scoring connected subgraph of size R including the root node v_1, which is a constrained maximum-weight connected graph problem and can be solved by a mixed integer linear programming (MILP) [10],

$$max \qquad W = \frac{1}{R}\sum_{i=1}^{n} t_i x_i, \tag{4}$$

$$s.t. \qquad \begin{cases} \sum_j c_{1j} = R-1, & \\ \sum_j c_{ji} - \sum_{j\neq 1} c_{ij} = x_i, & i = 2,\cdots,n, \\ c_{ij} \leq (R-1)x_i, & i,j = 1,\cdots,n, \\ x_i \in \{0,1\}, & i = 1,\cdots,n, \end{cases} \tag{5}$$

where x_i represents if node v_i is selected $(x_i = 1)$ or not $(x_i = 0)$ in the subgraph, c_{ij} are dummy variables representing the flow between selected nodes. The constraints (5) ensure the connectivity of the selected nodes, which is a major advantage of the proposed method.

Ideally, running MILP in every node in the graph from $R = 1$ to $R = N-1$, we can find all possible connected subgraphs with the highest scores of all sizes, thereby finding the largest one with a significantly high level.

However, it is not tractable or feasible to do this due to the NP hard nature of the MILP. On the other hand, by exploiting the small world property and modularity of molecular interaction networks, we can expect to obtain high quality solutions by searching the subgraph locally. Thus, we run the MILP with $R \leq K$ at each node in a subgraph containing all nodes not far away from it. The algorithm is illustrated in Table 1.

For each gene, a subgraph G_i of its searching area is spanned from the root node V_i in G by the breadth-first strategy. If there are nodes with the same distance to the root, they are added to G_i in the ascending order of the t_i value. Since MILP cannot

Table 1: Algrithm for searching subnetworks.

```
for i = 1 to N
   find subgraph G_i based on breadth-first strategy and SNR order
   for R = 1 to K
      run MILP with G_i, R and V_i
   end
end
```

be solved in polynomial time, to reduce the complexity of MILP we set an upper-bound m of the number of nodes in G_i. Nodes are added into G_i until it contains m nodes. Generally, in small networks where the number of nodes is less than 500, m can be the size of G. While in large networks there are thousands of nodes, such as protein-protein interaction networks of human, if m is set to the size of G, MILP does not work well due to the complexity of space and time. Thus, we can set m to a lower value to obtain the solutions for each restricted MILP problem in acceptable time. Finally, we obtain maximum score subgraphs of various sizes from 1 to K including the root node in G_i. W values of subgraphs with different sizes are used to estimate the density function of W. K is also used to control the computational complexity. Significantly high scoring subgraphs in these small subgraphs which are filtered using the kernel density estimation method are combined together to constitute a differentially expressed subnetwork. There are three parameters K, m and α in our method. However, the two examples in the results section indicate that the present method can obtain results superior to other methods. The solution of mixed integer linear programming is calculated by an open source software lp_solve (http://lpsolve.sourceforge.net/). The model is solved directly without relaxation.

3 Results

The MILP model with SNR metric denoted as MILP_S was applied to yeast molecular interaction networks. For the purpose of comparison, we searched differentially expressed networks using the jActiveModules plug-in of Cytoscape [16] which is implemented in [8]. The jActiveModules assigns a Z-score to measure the differentially expressed level of each gene and searches active subnetworks in networks using the simulated annealing (SA) technique. The results of two runs of SA are denoted as SA1 and SA2 respectively. The MILP with other metrics (the MILP model with t-statistic metric and Z-score named as MILP_t and MILP_Z respectively) were also tested to illustrate its effectiveness.

We implemented the present method to search differentially expressed subnetworks in a small yeast network (see Figure 2a) which is used to study galactose utilization pathway [7] by perturbing GAL80. We set m equal to 331 which is the total size of this small network and $K = 10$. Then by running the MILP_S with $\alpha = 0.3$, we obtained the differentially expressed subnetwork which contains the identified 90 genes as demonstrated in Figure 2b. The $(1-\alpha)$ $percentile$ is higher than the value which reaches the second density peak (see Figure 1). We ran the jActiveModules twice with the default parameters and got two different active subnetworks (see Figure 2c with 48 genes and Figure 2d with 67 genes). On the other hand, MILP_Z method with $\alpha = 0.15$ uncovered differentially expressed subnetwork including 92 genes (see Figure 2e).

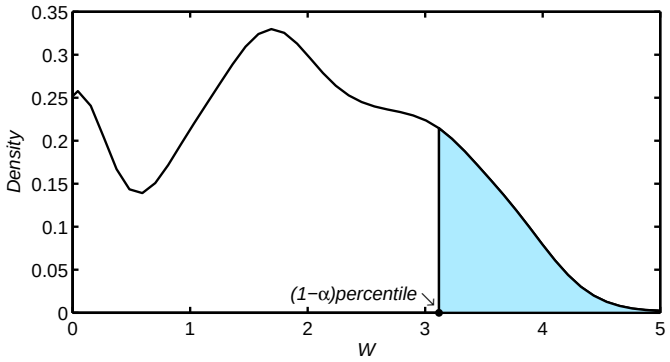

Figure 1: The density function of W estimated by non-parameter kernel method and the $(1-\alpha)$ *percentile* in the yeast GAL80 perturbation experiment. The blue area charts the density of significant W.

We found that the two active subnetworks by using jActiveModules (Figures 2c and 2d) are different from each other. However, they are both included in the subnetworks obtained by MILP$_S$ (Figure 2b) and MILP$_Z$ (Figure 2e) except some small isolated connected components containing only one or a few nodes. The reason is that simulated annealing algorithm is a random optimization method, for which the optimal solution is sensitive to the parameters and initial solution. In contrast, MILP is a linear and deterministic method without any random factor. The results of MILP$_S$ (Figure 2b) and MILP$_Z$ (Figure 2e) have only a few differences, and furthermore the results of MILP$_t$ (not shown) are almost as same as the one of MILP$_S$. Thus, the results demonstrate that the MILP model is robust and accurate besides the theoretical background. Note that the Z-score is calculated from the inverse normal cumulative distribution function. The SNR metric is nonparametric statistic and does not require norm distribution assumption.

Three genes GAL3, GAL4 and GAL10 included in the galactose utilization pathway are presented in the subnetwork identified by MILP$_S$, while none included in the subnetworks were found by jActiveModules. The regulatory genes GAL3, GAL4 and GAL80 which is perturbed in this experiment exert tight transcriptional control over the galactose transporter, the enzymes and to a certain extent, each other [7]. The function enriched GO categories [2] of these four subnetworks were calculated using BiNGO [11] (see Table 2). The result of MILP$_S$ is included in more GO categories than the two results of jActiveModules such as transcription regulator activity, biding and response to stimulus which correspond to the galoctose metabolic pathway. Some GO categories are present in one SA result but absent in another were identified by the proposed method such as cell and extracellular region. Moreover, several genes that are present in our subnetworks but absent in others (circled in red in Figure 2b, denoted by subnetwork$_d$) are enriched in GO terms of regulation of biological process, cell communication and extracellular region which are not significant in the two resulted subnetworks of SA1 and SA2.

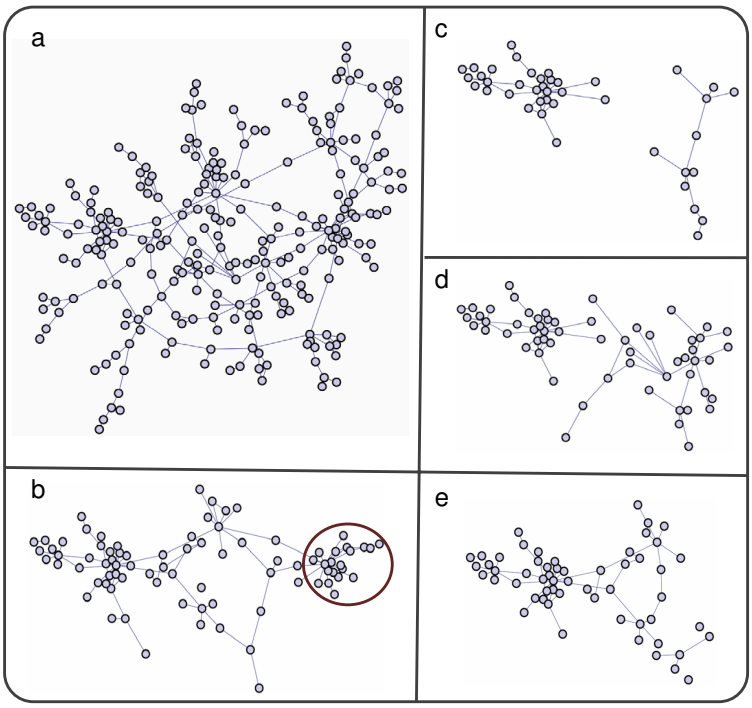

Figure 2: The active networks identified from a small yeast molecular interaction network to study pathways in GAL80 knockout experiment [8]. From the whole network (a), four active subnetworks are detected by MILP$_S$ (b), MILP$_Z$ (e), SA1 (c) and SA2 (d) respectively. The red circle in (b) charts the special part subnetwork$_d$ that is not uncovered by other methods.

Table 2: Enriched GO categories for yeast GAL80 knockout experiment.

ID	description	MILP$_S$	MILP$_Z$	subnetwork$_d$	SA1	SA2
30528	transcription regulator activity	1.3E-06	2.65E-05			
5623	cell	0.00588	0.010987			0.00142
51869	response to stimulus	0.011451				
5488	binding	0.005031	0.13037			
5576	extracellular region	0.002128		0.002232		
6519	amino acid and derivative metabolic process	0.006614	0.002388			
8152	metabolic process	9.34E-06	2.35E-10		5.53E-07	1.3E-07
9056	catabolic process		0.011456		0.005549	
3824	catalytic activity	0.13991	0.045357		0.080097	0.20126
5737	cytoplasm					0.000924
3676	nucleic acid binding	0.000154	0.004012			
50791	regulation of biological process	7.25E-07		2.36E-08		0.008636
5198	structural molecule activity	0.000181	1.33E-06		1.77E-07	2.69E-06
7154	cell communication	8.93E-05		0.000182		
30234	enzyme regulator activity			0.000368		0.001305
9058	biosynthetic process	6.33E-06	2.23E-09		1.77E-07	2.4E-07
8151	cellular process	4.45E-06	5.36E-06	0.034299	3.64E-05	2.23E-06
16829	lyase activity	1.38E-05	8E-06		0.000205	0.001083
43170	macromolecule metabolic process	0.000449	1.73E-05		7.33E-05	8.55E-07

4 Discussion and Conclusion

In this article, we proposed a new algorithm (MILP) based on a mixed integer programming model to identify differentially expressed pathways between two experimental conditions or disease states by utilizing information of gene expression data and molecular interaction network. Specifically, the MILP approach is able to effectively identify locally differentially expressed molecular interaction subnetworks and efficiently infers the global differentially expressed pathways. In particular, the proposed algorithm has theoretic background, and can ensure connectivity of the identified pathway. Firstly, a nonparametric scoring method is employed to evaluate the difference score of a subnetwork in the molecular network according to the gene expression information. Then differentially expressed subnetworks are detected from molecular interaction network based on the mixed integer linear programming model. Finally, the differentially expressed pathway is constructed by significantly high score subnetworks. We applied the proposed method on yeast data set to test its effectiveness.

The numerical experiments show that the proposed method outperforms the existing methods, e.g. simulated annealing based methods [8] in terms of identifying accurate pathways which cover more genes and GO categories associated to conditions. In general, proteins which response to the conditions and have interactions are more likely to present in the same differentially expressed pathway. The proposed method captures this property and identifies differentially expressed pathway containing more literatures verified genes. In summary, we proposed a new MILP model for identifying cellular differentially expressed pathways responding to external conditions by using the molecular network topology and gene expression information.

Acknowledgements

This work was partly supported by the Ministry of Science and Technology, China, under Grant No. 2006CB503905, National Natural Science Foundation of China under Grant No. 10631070. The authors would like to thank Prof. Luonan Chen for helpful discussions and suggestions.

References

[1] P. O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21(1 Suppl):33–7, 1999.

[2] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[3] R.K. Curtis et al. Pathways to the analysis of microarray data. *Trends in Biotechnology*, 23(8):429–435, 2005.

[4] S.W. Doniger et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology*, 4(1):R7, 2003.

[5] H. Ge et al. Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nature Genetics*, 29(4):482–6, 2001.

[6] T. R. Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531, 1999.

[7] T. Ideker et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929, 2001.

[8] T. Ideker et al. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(Suppl. 1):233–240, 2002.

[9] M. Kanehisa et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database Issue):D354, 2006.

[10] H.F. Lee and D.R. Dooly. Algorithms for the constrained maximum-Weight connected graph problem. *Naval Research Logistics*, 43:985–1008, 1996.

[11] S. Maere et al. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, 21(16):3448–3449, 2005.

[12] S. Peri et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363, 2003.

[13] S. Ramaswamy et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA*, 98(26):15149, 2001.

[14] B. Ren et al. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–9, 2000.

[15] E. Segal et al. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19(Suppl. 1):264–272, 2003.

[16] P. Shannon et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498, 2003.

[17] A. Subramanian et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, 102(43):15545–15550, 2005.

[18] X.M. Zhao et al. Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Research*, 36(9):e48, 2008.