The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

A Novel Approach for Pathway Inference Based on Network Flow*

Xianwen Ren[1,2] Xiang-Sun Zhang[1,†]

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

Abstract Signal transduction pathways play important roles in various biological processes such as cell cycle, apoptosis, proliferation, differentiation and responses to the external stimuli. Efficient computational methods are of great demands to map signaling pathways systematically based on the interactome and microarray data in the post-genome era. In this study, we propose a novel approach to infer the pathways based on the network flow well studied in the operation research. We define a potentiality variable for each protein to denote the extent to which it belongs to the objective pathway. And the capacity on each edge is not a constant but a function of the potentiality variables of the corresponding two proteins. The total potentiality of all proteins is given an upper bound. The approach is formulated to a linear programming model and solved by the simplex method. Experiments on the yeast sporulation data suggest this novel approach recreated successfully the backbone of the MAPK signaling pathway with a low upper bound of the total potentiality. By increasing the upper bound, the approach successfully predicted all the members of the MAPK pathway responding to the pheromone. It also included the cross-talks between the MAPK pathway and the pathway controlling cell cycle which play important roles in the yeast sporulation. This simple but effective approach provides a very useful tool facilitating the biologists to mine the biological insight from the genomic data.

Keywords Pathway Inference; Network Flow; Linear Programming; Protein Interaction Network; Gene expression

1 Introduction

Signal transduction plays an important role in biological systems. It forms the basis of various biological phenomena, e.g. cell cycle, apoptosis, proliferation, differentiation and responses to the external stimuli. However, mapping the signal transduction pathways is very hard by the traditional biochemical methods. They are quite labor-intensive and time-consuming. With the large-scale genomic datasets available, especially the interactome and microarray data, computational methods become more and more powerful in mapping the signaling pathways transferring signals from a source to its targets.

*This work is supported by the NSFC grants 10631070, 60873205, the grant kjcx-yw-s7 from CAS, and 2006CB503905 from MST of China.
†Corresponding author, E-mail: zxs@amt.ac.cn

A few methods have been proposed to infer the pathways, e.g. the Color Coding Method [8], Netsearch [10] and the mixed integer linear programming model [12, 13]. These methods use different heuristics to predict the pathways. The Color Coding method assigns a value to each interaction by using logistic regression based on gene expression and interaction data. It searches the whole network to find the pathway with the highest score, which is defined as the product of the values assigned to its interactions. Netsearch provides a statistical method to score the paths of a certain length based on clustering of gene expression data. These two methods require predefining the pathway structure and pathway length. However, it is hard to get this type of information in advance for unknown pathways. The mixed integer linear programming model proposed by Zhao et al. does not require predefining the pathway structure or length. It searches the network to find a subnetwork with the highest weight sum that connects the source with the target. The weight of each edge is assigned based on the confidence scores of the interactions or the correlation coefficients of genes from gene expression data. The improved version of this method incorporated the flow concept to guarantee the connectivity and to control the size of the pathway [12].

In nature, signaling pathways are information flows. Network flow can simulate the principles of signaling pathways, not only a tool to guarantee the connectivity. In this study, we transform the interactome data into an edge-weighted network in which each edge is weighted by the absolute value of the Pearson correlation coefficient of the gene expression profiles of the two node proteins. Given the source and the target, the signaling pathway inference problem is treated as a maximum-flow problem. Maximizing the flow amount from the source to the target is reasonable and in accord with the biological law "survival of the fittest" because the efficiency in the signal transmission can award a survival advantage to the organisms. We define a potentiality variable for each protein to denote the extent to which the protein belongs to the pathway, and define the capacity of each edge as the product of the edge weight and the sum of the potentiality of the corresponding two proteins. This is different from the well-studied maximum-flow problem in which the capacity of each edge is constant. An upper bound is given to the total potentiality of all the proteins, which provides a parameter for the biological users to calibrate the model. This model is formulated to a pure linear programming problem and solved by the simplex method.

Compared to the previous methods, our approach is clearer and simpler conceptually. Because linear programming problems can be solved very efficiently, our approach is also expeditious computationally. Experiments on the yeast sporulation expression data showed the effectiveness of our method. The backbone of the MAPK pathway deposited in the KEGG database [6] was recreated successfully with a low upper bound of the total potentiality. And increasing the upper bound got more relevant members of the pathway. The cross-talks between the MAPK pathway and the pathway controlling cell cycle were also indicated, which is insightful because cell cycle must be arrested during the yeast sporulation.

2 Method

Our method aims to infer the pathways from the protein-protein interaction network and gene expression data. The protein-protein interaction network depicts the static topo-

logical structure of the interactome. Gene expression data contain the dynamic information of cellular responses to various conditions. Based on the gene expression data, the protein-protein network is transformed into an edge-weighted network by calculating the absolute values of the Pearson correlation coefficients for each edge.

Let $G(V,E,W)$ denote the edge-weighted network, where V is the vertex set, E is the edge set and W is the weight set. $e_{ij} \in E$ denotes that $i \in V$ interacts with $j \in V$. $w_{ij} \in W$ denotes the weight on the edge e_{ij}. Based on the edge-weighted network, the pathway-inference problem is transformed into a maximum-flow problem. The problem can be written as a pure linear programming model as follows:

$$\max \sum_j f_{jt} \tag{1}$$

Subject to

$$f_{ij} \le w_{ij}(p_i + p_j) \qquad \forall e_{ij} \in E \tag{2}$$

$$-f_{ij} \le w_{ij}(p_i + p_j \qquad \forall e_{ij} \in E \tag{3}$$

$$\sum_j f_{ij} = 0 \quad \forall i \in V \quad \text{except} \quad s,t \tag{4}$$

$$\sum_j f_{sj} = \sum_k f_{kt} \tag{5}$$

$$f_{ij} = -f_{ji} \qquad \forall e_{ij} \in E \tag{6}$$

$$p_i \ge 0 \qquad \forall i \in V \tag{7}$$

$$p_i \le 1 \qquad \forall i \in V \tag{8}$$

$$\sum_i p_i \le c \tag{9}$$

where f_{ij} and p_i are variables and c is a constant. s is the source and t is the target. f_{ij} denotes the flow from i to j. p_i is the potentiality variable of i. The objective is to maximize the flow the target receives, given by (1). (2) and (3) define the capacity of each edge. (4) and (5) confine that the net flow of each mediate node is zero and that the amount of the flow the target receives equals to the amount the source sends out. (6) restricts that the net flow from i to j must be the opposite of the net flow from j to i. (7) and (8) restrain the potentiality variables between zero and one. (9) gives the upper bound of the total potentiality c. The linear programming model is solved by the simplex method.

When the linear programming model is solved, edges through which the flow amount is not larger than zero are filtered. The remaining edges constitute the pathway from the source to the target.

3 Results

The sporulation is an important developmental process through which diploid cells of the budding yeast produce haploid cells. The sporulation and the normal cell cycle are two mutually exclusive developmental processes. The MAPK signaling pathways play key roles in the switch between these two developmental processes [5]. Time-series DNA microarrays have been applied to detect the changes of gene expressions during the yeast

sporulation [4]. We integrated the gene expression data of yeast during the sporulation and the protein-protein interaction data in DIP [11] to infer the active MAPK signaling pathways during yeast sporulation by our method. The linear programming model was solved by the linprog function in Matlab on a 1.86 GHz Pentium 4 PC.

We downloaded the gene expression data from the NCBI GEO database with the accession number GDS104 [4, 3]. It contains the gene expression data of seven time points during the sporulation (0, 30min, 2hrs, 5hrs, 7hrs, 9hrs and 11hrs). The DIP Core dataset of the yeast protein-protein interactions was downloaded on July 8th, 2008. Only the physical interactions are selected to infer the MAPK signaling pathways. Totally 4770 interactions of 2334 proteins were selected.

During the sporulation of yeasts, the membrane protein STE3 senses the pheromone MATa outside the cells and initializes the MAPK signaling pathways. The signal is transmitted into the nucleus and received by the transcription factor STE12 regulating the expressions of a series of executive genes. In this process, the receptor STE3 activates the heterotrimetric G protein GPA1+STE4+STE18, transmits the signal to STE20 through CDC42, and then triggers the MAPK cascade [2]. The MAPK cascade consists of MAPKKK STE11, MAPKK STE7 and MAPKs FUS3 and KSS1. STE11 phosphorylates and activates STE7 on two residues conserved on all MAPKKs. STE7 phosphorylates Fus3 and KSS1 in turn. STE5 acts as a scaffold to get the MAPKKK, MAPKK and MAPKs together, facilitating the sequential phosphorylation (Figure 1).

We applied our approach to search the pathway from STE3 to STE12. With the upper bound of the total potentiality being one, the backbone of the MAPK signaling pathway was recreated successfully (Figure 2). In the predicted pathway in Figure 2, STE3 first interacts with AKR1 which then interacts with STE4 and STE5. STE4 is a subunit of the heterotrimetric G protein. The MAPKKK (STE11), MAPKK (STE7) and one MAPK (KSS1) were all identified. AKR1 was included in the predicted pathway because STE3 has only one interactor (AKR1) in the interaction dataset.

The size of the predicted pathways increases when the upper bound of the total potentiality increases from one to six (Figure 3). All the MAPK pathway members except STE18 and MSG5 (15/17) are included in the predicted pathway where the upper bound is six (Figure 4). STE18 and MSG5 are not included because these two proteins are not included in the interaction dataset.

Besides the genes involved in the MAPK signaling pathway, the predicted pathway in Figure 4 also includes 23 other proteins. The functional enrichment analysis by BinGO [7], a plug-in of Cytoscape [9], shows that thirteen out of the 23 proteins involve in the cell division ($p = 1.0219e-8$). Four out of the remaining ten proteins are annotated as "response to pheromone" in the Gene Ontology database [1]. Thus, these proteins involve in the MAPK-related pathways that act in concert with the MAPK signaling pathway to promote the yeast sporulation. This is insightful and suggests that our method can identify the cross-talks between the related pathways.

4 Discussions and Conclusion

In nature, signal transduction pathways are ordered sequences of biochemical reactions. The components include signaling molecules (or ligands), cell-surface receptors, intracellular receptors, second messenger molecules, nuclear factors and other proteins.

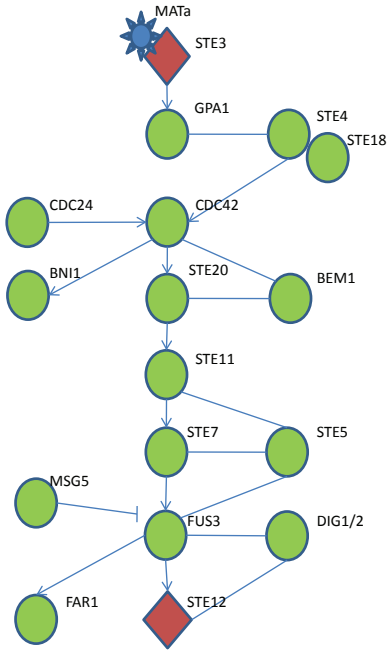

Figure 1: The MAPK signaling pathway responding to the pheromone MATa in KEGG

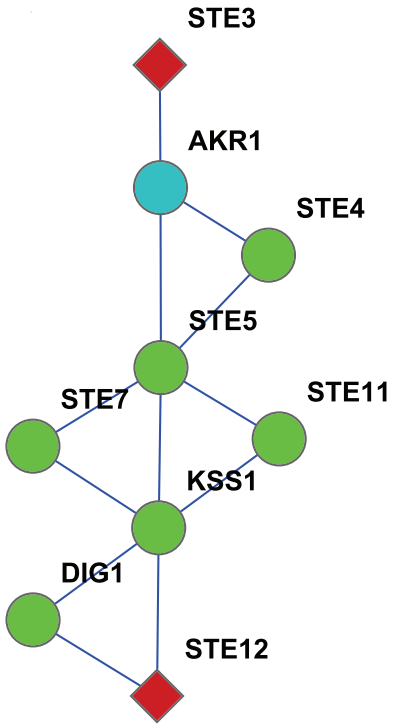

Figure 2: The backbone of the yeast MAPK signaling pathway recreated by our approach ($c = 1$)

Many ligands and the second messengers are not proteins. The information of signaling pathway provided by protein-protein interaction data is limited and partial. The current protein-protein interaction data are incomplete and noisy compared to the true interactome. This influences the accuracy of the predictions. Signal transduction is carried out by enzymes which are regulated through their abundance and activity. Microarray data only measured the approximate abundance of enzymes. Proteomic data will provide exact measures of both the abundance and the activity. Integration with proteomic data will improve the accuracy of predictions greatly when the data are available.

Despite those difficulties and defects, computational predicting the signaling pathways based on the interactome and expression data has achieved an appreciable accuracy. Combination with the traditional experimental methods will accelerate the progression of biomedical research greatly. In this study, we propose a conceptually clearer and simpler method to infer the signaling pathways based on the network flow. It is implemented as a pure linear programming model and is expeditious computationally. The upper bound of the total potentiality provides a valid parameter facilitating the biological users applying the approach to mine the publicly available genomic data to promote the signaling pathway mapping. With more types of genomic and proteomic data available, the approach

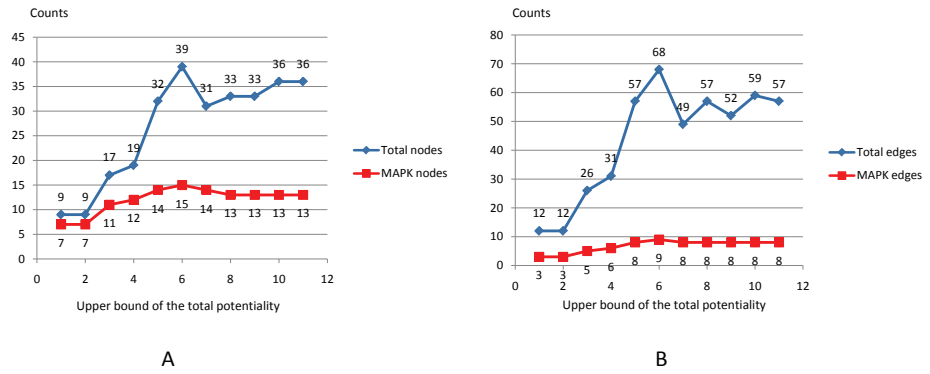

Figure 3: The size of the predicted pathways varies with the upper bound of the total potentiality. A, the number of nodes vs the upper bound; B, the number of edges vs the upper bound.

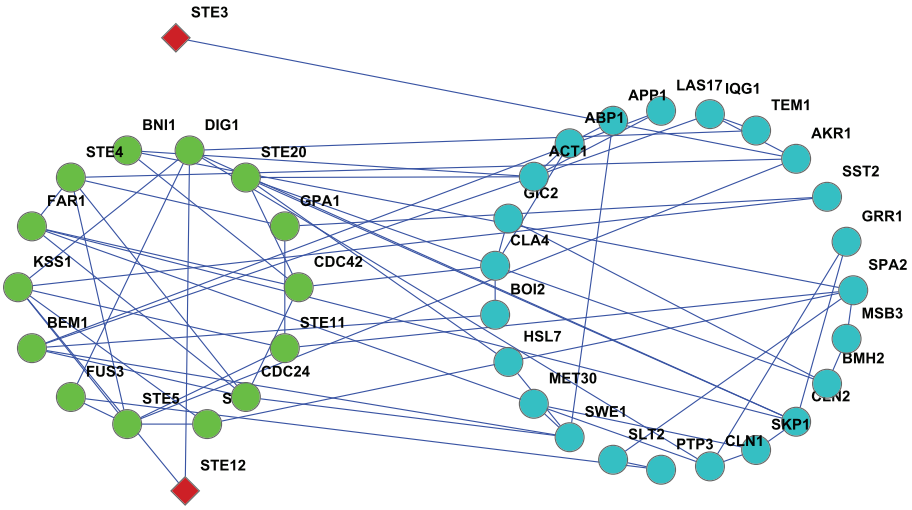

Figure 4: The predicted pathway including all the MAPK pathway members except STE18 and MSG5 ($c = 6$). Nodes in the left circle are all the MAPK pathway members. Nodes in the right circle involve the MAPK-related pathways regulating the cell division.

should evolve to utilize more information to improve the predicting accuracy.

Acknowledgements

The authors thank the anonymous critics who provided valuable advice to help improve the work and this manuscript.

References

[1] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000.

[2] Flora Banuett. Signalling in the yeasts: An informational cascade with links to the filamentous fungi. *Microbiol. Mol. Biol. Rev.*, 62(2):249–274, 1998.

[3] Tanya Barrett, Tugba O. Suzek, Dennis B. Troup, Stephen E. Wilhite, Wing-Chi Ngau, Pierre Ledoux, Dmitry Rudnev, Alex E. Lash, Wataru Fujibuchi, and Ron Edgar. Ncbi geo: mining millions of expression profiles–database and tools. *Nucl. Acids Res.*, 33(suppl 1):D562–566, 2005.

[4] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282(5389):699–705, 1998.

[5] Saul M. Honigberg and Kedar Purnapatre. Signal pathway integration in the switch from the mitotic cell cycle to meiosis in yeast. *J Cell Sci*, 116(11):2137–2147, 2003.

[6] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, and Akihiro Nakaya. The kegg databases at genomenet. *Nucl. Acids Res.*, 30(1):42–46, 2002.

[7] Steven Maere, Karel Heymans, and Martin Kuiper. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, 2005.

[8] J. Scott, T. Ideker, R. M. Karp, and R. Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *Journal of Computational Biology*, 13(2):133–144, 2006.

[9] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, 2003.

[10] Martin Steffen, Allegra Petti, John Aach, Patrik D'Haeseleer, and George Church. Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3(1):34, 2002.

[11] Ioannis Xenarios, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.*, 30(1):303–305, 2002.

[12] X. M. Zhao, R. S. Wang, L. Chen, and K. Aihara. Automatic modeling of signaling pathways by network flow model. *J Bioinform Comput Biol*, 7(2):309–22, 2009.

[13] Xing-Ming Zhao, Rui-Sheng Wang, Luonan Chen, and Kazuyuki Aihara. Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucl. Acids Res.*, 36(9):e48–, 2008.