Using Knowledge Driven Matrix Factorization to Reconstruct Modular Gene Regulatory Network

Yang Zhou[1] and **Zheng Li**[2] and **Xuerui Yang**[2] and **Linxia Zhang**[2]

and **Shireesh Srivastava**[2] and **Rong Jin**[1] and **Christina Chan**[1,2]

1. Department of Computer Science and Engineering

2. Department of Chemical Engineering and Materials Science

Michigan State University, East Lansing, MI 48824

Abstract

Reconstructing gene networks from micro-array data can provide information on the mechanisms that govern cellular processes. Numerous studies have been devoted to addressing this problem. A popular method is to view the gene network as a Bayesian inference network, and to apply structure learning methods to determine the topology of the gene network. There are, however, several shortcomings with the Bayesian structure learning approach for reconstructing gene networks. They include high computational cost associated with analyzing a large number of genes and inefficiency in exploiting prior knowledge of co-regulation that could be derived from Gene Ontology (GO) information. In this paper, we present a knowledge driven matrix factorization (KMF) framework for reconstructing modular gene networks that addresses these shortcomings. In KMF, gene expression data is initially used to estimate the correlation matrix. The gene modules and the interactions among the modules are derived by factorizing the correlation matrix. The prior knowledge in GO is integrated into matrix factorization to help identify the gene modules. An alternating optimization algorithm is presented to efficiently find the solution. Experiments show that our algorithm performs significantly better in identifying gene modules than several state-of-the-art algorithms, and the interactions among the modules uncovered by our algorithm are proved to be biologically meaningful.

Introduction

Reconstructing gene regulatory network from the micro-array data is important for understanding the underlying mechanism behind cellular processes. A number of computational methods have been developed or applied to automatically reconstruct gene networks from gene expression data. Clustering methods, such as hierarchical clustering, K-means and self-organizing map (Eisen *et al.* 1998), are commonly used to identify gene modules. The main disadvantage of clustering methods is that they are unable to uncover the interaction among different modules, which is crucial to the understanding of disease mechanisms. To overcome this problem, several studies have proposed to integrate clustering methods with structure learning algorithms. In (Toh & Horimoto 2002), the authors combined a clustering method with the Graphical Gaussian Model (GGM) for module network reconstruction. In (Segal *et al.* 2003), a Bayesian framework is presented to integrate a clustering method with Bayesian network learning. A disadvantage with these approaches is that they rely solely on gene expression data, which are noisy, in the analysis. Furthermore, as revealed by several studies (Husmeier 2003; Yu *et al.* 2004), structure learning methods tend to perform poorly when the number of experimental conditions is significantly smaller than the number of genes.

In the past, a large number of studies have been devoted to exploiting prior knowledge for gene network reconstruction to alleviate the problem that expression data are often sparse and noisy (Bar-Joseph *et al.* 2003; Berman *et al.* 2002; Hartemink *et al.* 2002; Ideker *et al.* 2001; Ihmels *et al.* 2002; Pilpel, Sudarsanam, & Church 2001; Li & Yang 2004). A typical approach is to construct a Bayesian prior for the directed arcs in the Bayesian network using the prior knowledge of regulator-regulatee relationships that are derived from other data such as location analysis data and protein interaction data. A problem with this type of approach is that it is often difficult to extend them to incorporate the co-regulation relationships that can be easily derived from the GO database. This is a shortcoming with Bayesian network analysis especially for mammalian systems, where interaction data is not as readily available, whereas GO information is. Therefore, developing a framework of knowledge driven analysis with high-throughput data that effectively exploits the prior knowledge of co-regulation relationships from GO could enhance the robustness of the network reconstruction from gene expression data.

The key challenge with using GO for network reconstruction is that the co-regulation relationships derived from GO may be noisy and inaccurate. In this paper, we propose a framework for gene modular network reconstruction based on the **Knowledge driven Matrix Factorization** (**KMF**) that is able to effectively exploit the prior knowledge derived from GO. The key features of the proposed framework are (1) it derives both the gene modules and their interactions from a combination of expression data and the GO database, (2) it incorporates the prior knowledge of co-regulation relationships into network reconstruction via matrix regularization, and (3) it presents an efficient learning algorithm that combines the techniques of non-negative matrix factoriza-

tion and semi-definite programming.

It is important to note that although our framework is closely related to other algorithms for matrix factorization (e.g., non-negative matrix factorization), they differ significantly in both their computational methods and goals. First, unlike the existing algorithms for matrix factorization that are designed either for clustering or for dimensionality reduction, our framework aims to learn a module network structure from gene expression data. Second, unlike other matrix factorization algorithms that solely depend on iterative algorithms for optimization, the proposed framework exploits both convex and non-convex optimization strategies for finding the optimal network structure.

A Framework for Knowledge Driven Matrix Factorization (KMF)

The following terminology and notations will be used throughout the rest of this paper. Let m be the number of experimental conditions, and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \in \mathbf{R}^m$ be the expression levels of the ith gene measured under m conditions. Let n be the number of genes, and $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ include the expression levels of all n genes. Given the expression data, we can estimate the pairwise correlation between any two genes. A number of statistical correlation metrics can be used for this purpose, such as Pearson correlation, mutual information, and chi-square statistics (Yang & Pedersen 1997). The computation results in a symmetric matrix $W = [w_{ij}]_{n \times n}$ where w_{ij} measures the correlation between gene $\mathbf{x}_i$ and $\mathbf{x}_j$.

The main idea behind the knowledge driven matrix factorization framework is to compute the network structure by factorizing the correlation matrix W into the matrices for gene modules (i.e., the module matrix) and the module network structure (i.e., the network matrix). We denote by M the module matrix, of which element M_{ij} represents the confidence of assigning the ith gene to the jth module. We denote by C the network matrix, of which element C_{ij} represents the interaction strength between module i and module j. Note that the computational problem addressed here is fundamentally different from the problems addressed by the previous studies of matrix factorization (Lee & Seung 2000) that mainly focused on dimensionality reduction and data clustering.

To determine the gene modules (i.e., M) and their network structure (i.e., C), we consider the following three criteria when formulating the framework of knowledge driven matrix factorization for module network reconstruction:

1. The module matrix M and the network structure matrix C should be combined to accurately reproduce the correlation matrix W. This is based on the assumption that gene correlation information can essentially be explained by the gene modules and their interactions.
2. The module matrix M is expected to be consistent with the prior knowledge collected from Gene Ontology. In particular, two genes that bear a large similarity in gene functions as described in GO are likely to be assigned into the same module.
3. The network structure matrix C is expected to be consistent with the hierarchical scale-free structure of gene networks. As suggested in (Barabasi & Albert 1999), a network with hierarchical scale-free structure tends to have a small number of linkages in total. We thus expect matrix C to be a sparse matrix with most of its elements being small and close to zero.

In the following subsections, we will first discuss how to capture the above three criteria in formulating the objective function, followed by the description of the full framework.

Matrix Reconstruction Error

Before discussing the reconstruction error, we need to first describe how to approximate the gene correlation matrix W by the gene modules and the module network. We assume that the correlation between two genes $\mathbf{x}_i$ and $\mathbf{x}_j$ arises because of the interaction between the two modules that are associated with the two genes. Hence, variable W_{ij} can be approximated by $\sum_{a,b} M_{ia} M_{jb} C_{ab}$ where M_{ia} and M_{jb} represent the association of genes to gene modules and C_{ab} represent the interaction between modules a and b. In the form of matrices, the above idea is summarized as to approximate W by the product $M \times C \times M^\top$. We denote by $l_d(W, M, C)$ the matrix reconstruction error between W and $M \times C \times M^\top$.

A number of measurements have been proposed to calculate the matrix reconstruction error. One general approach is to measure the norm of the difference between W and the reproduced matrix $MCM^\top$, i.e., $l_d(W, M, C) = \|W - MCM^\top\|$. Two types of matrix norms used in measuring the reconstruction errors are the entry-wise norms (e.g, Frobenius norm) and the induced norms (e.g., spectral norm). The key difference between these two types of norms is that the entry-wise norms measure the error by the entry-wise mismatch, while the induced norm measures the mismatch between two matrices by the difference in their eigenspetra. Here we adopt the Frobinius norm,

$$\begin{aligned} l_d(W, MCM^T) &= ||W - MCM^\top||_F^2 \\ &= \sum_{i,j=1}^{n} (W_{ij} - [MCM^T]_{ij})^2 \end{aligned}$$

Consistency with the Prior Knowledge from GO

Second, we exploit the prior knowledge from GO to regularize the solution of M. We encode the information within GO by a similarity matrix S, where $S_{ij} \geq 0$ represents the similarity between two genes in their biological functions (Jin *et al.* 2006). Furthermore we create a normalized combinatorial graph Laplacian L from the similarity matrix S. Then the disagreement between gene modules and collected gene information from GO is measured by $l_m(M, L) = \text{tr}(MLM^\top)$. To better understand this quantity, we expand $l_m(M, L)$ as follows:

$$\text{tr}(MLM^\top) = \sum_{i,j=1}^{N} S_{ij} \left(\sum_z (M_{iz} - M_{jz})^2 \right)$$

Note that term S_{ij} measures the similarity between gene i and j in gene functions described in GO, and term $\sum_z (M_{iz} - M_{jz})^2$ measures the difference in module memberships between two genes. Hence, the product of the two terms essentially indicates the disagreement between M and the gene information from GO. By minimizing this disagreement, we ensure that the gene modules are consistent with the prior information of genes.

Gene Module Network with Hierarchical Scale-free Structure

As revealed by previous studies (Bhan, Galas, & Dewey 2002; Joeng *et al.* 2000), the structure of many gene networks appears to be hierarchical and scale-free. In particular, genes are first clustered into modules, and the gene modules are then connected by scale-free networks. The most important feature of a scale-free network is that most nodes in the network are connected to a few neighbors, and only a small number of nodes, which are often called "hubs", are connected to a large number of nodes. Compared to other network structures, a scale-free network has a very skewed degree distribution, and therefore a relatively smaller number of edges (Barabasi & Albert 1999). This fact implies that the network structure matrix C should be a *sparse matrix* of which most elements are small or close to zero. Thus, to ensure a scale-free network structure, we regularize the sparsity of matrix C by term $l_c(C) = \|C\|_F^2$.

Matrix Factorization Framework for Hierarchical Module Network Reconstruction

Combining the measurements for the above three criteria, we have the final formulation for finding the optimal M and C, i.e.,

$$\begin{aligned} \min_{M,C} \quad & l_d(W, M, C) + \alpha l_m(M, L) + \beta l_c(C) \\ \text{s. t.} \quad & C \succeq 0, \quad C_{ii} = 1,\ i = 1, 2, \ldots, n \\ & C_{ij} \geq 0,\ i, j = 1, 2, \ldots, r \\ & M_{ij} \geq 0,\ i, j = 1, 2, \ldots, n \end{aligned} \tag{1}$$

where parameter α and β weight the contribution of terms l_m and l_c respectively. The constraint $C \succeq 0$ ensures that the interaction among modules complies with the triangular inequality, i.e., if module i has strong interactions with both module j and module k, then module j and k are also expected to have strong interactions. Unlike the Bayesian network based structure learning that requires solving a discrete optimization problem, (1) is an optimization problem of continuous variables and therefore can usually be solved more efficiently than Bayesian network.

Solving the Constrained Matrix Factorization

We solve the above optimization problem through alternating optimization. It alters the process of optimizing M with fixed C and the process of optimizing C with fixed M iteratively till the solution converges to a local optimum. We describe these two processes as follows:

Optimize M by fixing C: The related optimization problem is:

$$\begin{aligned} \arg\min_{M \in \mathbf{R}^{n \times r}} \quad & F_m(M) = \|W - MCM^\top\|_F^2 + \alpha \text{tr}(M^\top L M) \\ \text{s. t.} \quad & M_{ij} \geq 0,\ i, j = 1, 2, \ldots, n \end{aligned}$$

To find an optimal solution for M, we propose the following bound optimization algorithm. Let $\tilde{M}$ represent the solution of the previous iteration, and our goal is to find a solution of M for the current iteration. First, we consider bounding the first term in $F_m(M)$ by the following expression:

$$\begin{aligned} & \left(W_{ij} - \sum_{k,l=1}^{r} M_{ik} M_{jl} C_{kl} \right)^2 - W_{ij}^2 \\ \leq & \frac{1}{2} \sum_{k,l=1}^{r} [\tilde{M}C\tilde{M}^\top]_{ij} \tilde{M}_{ik} \tilde{M}_{jl} C_{kl} \left([\frac{M_{ik}}{\tilde{M}_{ik}}]^4 + [\frac{M_{jl}}{\tilde{M}_{jl}}]^4 \right) \\ & - 2 \sum_{k,l=1}^{r} W_{ij} \tilde{M}_{ik} \tilde{M}_{jl} C_{kl} \left(\log \frac{M_{ik}}{\tilde{M}_{ik}} + \log \frac{M_{jl}}{\tilde{M}_{jl}} + 1 \right) \end{aligned}$$

We then upper bound the second term in $F_m(M)$ such that:

$$\begin{aligned} & S_{ij}(M_{ik} - M_{jk})^2 \\ = \ & S_{ij} M_{ik}^2 + S_{ij} M_{jk}^2 - 2 S_{ij} M_{ik} M_{jk} \\ \leq \ & S_{ij} M_{ik}^2 + S_{ij} M_{jk}^2 \\ & -2 S_{ij} \tilde{M}_{ik} \tilde{M}_{jk} \left(\log \frac{M_{ik}}{\tilde{M}_{ik}} + \log \frac{M_{jk}}{\tilde{M}_{ik}} + 1 \right) \end{aligned}$$

Taking the derivative of the upper bound of $F_m(M)$ with respect to M_{ik} and set the derivative to be zero, we have the optimal solution for M as:

$$M_{ik} = \tilde{M}_{ik} \left(\frac{2 c_{ik}}{b_{ik} + \sqrt{b_{ik}^2 + 4 a_{ik} c_{ik}}} \right)^{\frac{1}{2}} \tag{2}$$

where

$$\begin{aligned} a_{ik} &= [\tilde{M}C\tilde{M}^\top \tilde{M}C]_{ik} \\ b_{ik} &= \alpha \tilde{M}_{ik} D_i \\ c_{ik} &= \alpha [S\tilde{M}]_{ik} + [W\tilde{M}C]_{ik} \\ D_i &= \sum_{k=1}^{n} S_{ik} \end{aligned}$$

Optimize C by fixing M: This corresponds to the following optimization problem:

$$\begin{aligned} \arg\min_{C \in \mathbf{S}^r} \quad & F_c(C) = \|W - MCM^\top\|_F^2 + \beta \|C\|_F^2 \\ \text{s. t.} \quad & C \succeq 0, \quad C_{ii} = 1,\ i = 1, 2, \ldots, r \\ & C_{ij} \geq 0,\ i, j = 1, 2, \ldots, r \end{aligned}$$

We expand the objective function $F_c(C)$ as follows:

$$\begin{aligned} F_c(C) = & \text{tr}(WW^\top) - 2\text{tr}(WMCM^\top) \\ & + \text{tr}(MCM^\top MCM^\top) + \beta \text{tr}(CC^\top) \end{aligned}$$

We then introduce auxiliary variable B and slack variables η, ξ and convert into the following optimization problem:

$$\begin{aligned} \arg\min_{C \in \mathbf{S}^r} \quad & \eta + \beta\xi - 2\text{tr}(M^\top WMC) \\ \text{s. t.} \quad & C \succeq 0, \quad C_{ii} = 1,\ i = 1, 2, \ldots, r \\ & C_{ij} \geq 0,\ i, j = 1, 2, \ldots, r \\ & \eta \geq \sum_{i,j=1}^{r} B_{ij} B_{ji}, \quad B = M^\top MC \\ & \xi \geq \sum_{i,j=1}^{r} C_{ij}^2 \end{aligned} \tag{3}$$

This optimization problem can be solved effectively using semi-definite programming technique.

Determining Parameters α and β

The regularizer parameter α and β significantly affect the outcome of the proposed algorithm: α balances the information from GO against the information from gene expression data, and β controls the sparseness of the interaction matrix C. Here we use the *stability analysis* to determine the value of α and β. The basic assumption of stability analysis is that if the parameters are set properly, then then algorithm runs with different random initialization should result in more or less similar results (Tibshirani, Walther, & Hastie 2000). We run our algorithm multiple times with a given setting of α and β, then evaluate each result to all other results using the evaluation metric defined in the next section, then we calculate the standard deviation of all these evaluation metrics. α and β are tuned to minimize this standard deviation.

Experimental Results and Discussion

Our experiments are designed to evaluate our proposed knowledge driven matrix factorization framework in reconstructing modular gene network, particularly in identifying gene modules and uncovering the interactions among gene modules.

Datasets

Two datasets are used in our experiments:

- *Gene expression data of yeast cell cycle system:* The gene expression data for 104 genes involved in yeast cell cycle were obtained from the Yeast Cell Cycle Analysis Project (`http://genome-www.stanford.edu/cellcycle/data/rawdata/`). These genes were divided into six groups based on their peak expression in the different phases of the cell cycle and the transcription factors that regulate them (Spellman *et al.* 1998).
- *Gene expression data of liver cell system:* Gene expression data was obtained for HepG2 cells exposed to free fatty acids (FFAs) and tumor necrosis factor (TNF-α) (Srivastava & Chan 2007). Gene expression data were obtained for 15 different conditions. The original data consisted of 19458 genes. The analysis of variance (ANOVA) was applied to the entire list of genes with $P < 0.01$ to compare the effect of treatment (e.g. FFA or TNF-α) and to determine whether a treatment had a significant effect. The expression levels of 830 genes were found to be significant due to either TNF-α or FFA (Li *et al.* 2007a) and this subset of genes are further analyzed in our experiment.

Evaluation of KMF based upon Yeast Cell Cycle Data

In this study, we focus on evaluating the capability of KMF in identifying gene modules because the interaction information among gene modules is not available. We compared the gene modules identified by KMF and three other baseline algorithms to the ground truth which includes six groups defined by (Spellman *et al.* 1998). The three algorithms used as baseline algorithms in our study to identify the gene modules in the yeast cell cycle genes include 1) Bayesian Module Network in Genomica (Segal *et al.* 2003), 2) Probabilistic Spectral Clustering (PSC) (Jin, Ding, & Kang 2006), and 3) Sparse Matrix Factorization (SMF) (Badea & Tilivea 2005). Bayesian Module Network in Genomica was chosen since it had used the yeast cell cycle genes to identify gene modules. PSC was chosen since it had been proved to be one of the state-of-the-art clustering algorithms. We evaluated three settings when applying PSC to identify the gene modules. In the first two settings, either the gene expression data or GO is used to construct the similarity matrix before applying PSC. In the third setting, the two similarity matrices based on gene expression and GO are combined linearly. SMF was chosen since it had been shown to identify gene modules using a non-negative factorization algorithm that combines gene expression data and transcription factor binding data. To obtain the cluster membership of the genes, we set a threshold on the membership matrix M. A natural choice for the threshold is $1/r$ where r is the number of clusters. The same method was applied to PSC and SMF to determine the binary cluster memberships.

To quantitatively evaluate the performance of the algorithms in our experiment, we use the Pairwise F-measure (PWF1) metric (Liu & Jin 2007). Let U be the set of gene pairs that share at least one cluster in the experiment, and T be the set of gene pairs that actually share at least one cluster, PWF1 is defined as

$$precision = \frac{|U \cap T|}{|U|}, \quad recall = \frac{|U \cap T|}{|T|}$$
$$PWF1 = \frac{2 \times precision \times recall}{precision + recall}$$

where $|\cdot|$ is the size operator on a set. The precision measures the accuracy in identifying co-regulated genes, and the recall measures the percentage of co-regulated genes that are correctly identified, where we assume that the genes within an original group (Spellman *et al.* 1998) were co-regulated. PWF1 combines these two factors by their harmonic mean.

Table 1 shows the PWF1 measure of different algorithms on the Yeast cell cycle data. We observe that KMF outperformed the other algorithms significantly. Note that KMF performed better than PSC using the combination of GO and

Algorithm	PWF1(mean±std)
KMF	**0.625±0.007**
Genomica	0.473
PSC (expression)	0.446±0.057
PSC (GO)	0.386±0.020
PSC (expression+GO)	0.571±0.013
SMF (expression+binding)	0.413±0.118

Table 1: PWF1 measure of the experimental results on the Yeast cell cycle dataset. Each algorithm except Genomica (which does not need initialization at the beginning of execution) was executed 10 times with different random initializations, and the mean and standard deviation of PWF1 from 10 runs are calculated.

gene expression data. This suggests that KMF is more effective in exploiting prior knowledge than PSC. In addition, KMF also showed a lower standard deviation on the PWF1 over multiple runs. This suggested that KMF performed robustly by utilizing the GO information to guide the modular network reconstruction from gene expression data.

Application to Identify Gene Modules and Modular Network in Liver Cells

We also applied KMF to gene expression data obtained from liver cells where the main objective was to identify the interactions between the modules.

In our experiment we manually set the number of clusters to be 30 according to the suggestions of biologists. We found that for most identified gene modules, genes with similar functions were enriched in their own separate modules/gene groups. For example, 7 out of the 11 genes in a module encode the 5 of the 6 enzymes involved in the TCA cycle. Similarly, one module consisted primarily of NADH dehydrogenases and one module consisted of the genes involved in the metabolism of ATP. 7 out of the 18 genes in a module encode different sub-complexes of cytochrome-c oxidase (complex IV). In general, most of the gene-groups could be assigned a particular function/process based upon the list of genes enriched in them. Due to the space limitation, we are unable to list all 30 modules and their enriched functions, so we only give a few examples above to illustrate the function enrichment of gene modules. We also note that a common practice in evaluating function enrichment by GO can not be applied here since we already utilize GO for the identification of gene modules and their interactions.

Next, we examined if KMF is able to correctly uncover the interactions among different modules by looking into the C matrix whose coefficients indicate the strengths of interactions, which is analogous to a correlation matrix. After sorting out the significantly higher values in C matrix, we found that module *complex III*, *complex IV*, *complex I*, *complex V* and *TCA and complex II* are closely connected as shown in Figure 1. From the aspect of molecular biology, most of the proteins in these modules are located in the mitochondria or on the mitochondria membrane, and these modules are indeed biologically connected. The modules of TCA cycle, electron transport chain (ETC) complex I, complex III, complex IV, and ATP synthase are related to energy production in the mitochondria, which are often referred to as intracellular powerhouse because they produce most of the energy used by the cell. The production of energy in the mitochondria is accomplished by two closely linked metabolic processes, the TCA cycle and oxidative phosphorylation. The TCA cycle converts carbohydrates, lipids, and amino acids into ATP and energy rich molecules, such as NADH. Oxidative phosphorylation generates ATP through the ETC consisting of five protein complexes embedded in the inner membrane of the mitochondria including complex I (NADH dehydrogenase), complex II (succinate dehydrognease of the TCA cycle), complex III (cytochrome-c reductase), complex IV (cytochrome-c oxidase), and complex V (ATP synthase). Thus, Figure 1 show a biological network involved in the production of energy in mitochondria reconstructed by KMF. For many other modules, their interactions are relatively weak according C matrix, and therefore their biological meaning is rather unclear from our study.

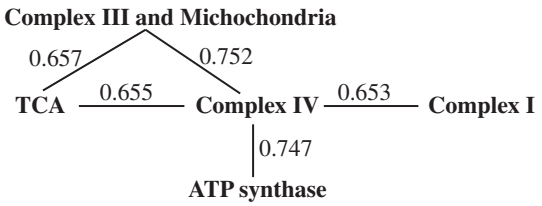

Figure 1: KMF uncovered the connections between energy production modules of TCA, Complex I, III, IV and V. Numbers on the edges are the interaction strengths between the modules from the interaction matrix (C matrix).

Therefore, KMF is able to identify highly enriched gene modules with distinct cellular functions and the interactions among the modules. In summary, KMF is an approach that can be applied to uncover pathways specific to a phenotype and potentially be used to elucidate mechanisms involved in diseases by integrating gene expression and *a priori* knowledge.

Readers are referred to our technical report (Li *et al.* 2007b) for the complete analysis of the experimental results. The full list of gene modules is available online at `http://www.chems.msu.edu/groups/chan/GO_KMF_genecluster.xls`.

Conclusions

In this paper, we propose a novel framework to meet the challenging problem of reconstructing gene networks from multiple information sources. The advantage of our proposed framework is that it derives both the gene modules and their interactions in a unified framework of matrix factorization, and it incorporates the prior knowledge of coregulation relationships from GO information into the network reconstruction process. We also present an efficient algorithm to solve the related optimization problem. Experiments show that our proposed framework performs significantly better in identifying gene modules than several state-of-the-art algorithms, and the interactions among modules

uncovered by our algorithm are proved to be biologically meaningful.

Acknowledgements

This work is supported in part by the National Science Foundation (IIS-0643494) and National Institute of Health (1R01-GM079688-01). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF and NIH.

References

Badea, L., and Tilivea, D. 2005. Sparse factorizations of gene expression data guided by binding data. In *Pacific Symposium of Biocomputing*.

Bar-Joseph, Z.; Gerber, G. K.; Lee, T. I.; Rinaldi, N. J.; Yoo, J. Y.; Robert, F.; Gordon, D. B.; Fraenkel, E.; Jaakkola, T. S.; Young, R. A.; and Gifford, D. K. 2003. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology* 21:1337–1342.

Barabasi, A. L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286:509–512.

Berman, B. P.; Nibu, Y.; Pfeiffer, B. D.; Tomancak, P.; Celniker, S. E.; Levine, M.; Rubin, G. M.; and Eisen, M. B. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. In *Proceedings of the National Academy of Sciences*, volume 99, 757–762.

Bhan, A.; Galas, D.; and Dewey, T. G. 2002. A duplication growth model of gene expression networks. *Bioinformatics* 18:1486–1493.

Eisen, M. B.; Spellman, P. T.; Brown, P. O.; and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of the National Academy of Sciences*, volume 95, 14863–14868.

Hartemink, A. J.; Gifford, D. K.; Jaakkola, T. S.; and Young, R. A. 2002. Combining location and expression data for principled discovery of genetic regulatory networks. In *Proc. Pacific Symposium on Biocomputing*, 437–449.

Husmeier, D. 2003. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics* 19(17):2271–2282.

Ideker, T.; Thorsson, V.; Ranish, J. A.; Christmas, R.; Buhler, J.; Eng, J. K.; Bumgarner, R.; Goodlett, D. R.; Aebersold, R.; and Hood, L. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292(5518):929–934.

Ihmels, J.; Friedlander, G.; Bergmann, S.; Sarig, O.; Ziv, Y.; and Barkai, N. 2002. Revealing modular organization in the yeast transcriptional network. *Nature Genetics* 31:370–377.

Jin, R.; Si, L.; Srivastava, S.; Li, Z.; and Chan, C. 2006. A knowledge driven regression model for gene expression and microarray analysis. In *EMBS '06*.

Jin, R.; Ding, C.; and Kang, F. 2006. A probabilistic approach for optimizing spectral clustering. In *NIPS*.

Joeng, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; and Barabasi, A. L. 2000. The large-scale organization of metabolic networks. *Nature* 407(6804):651–643.

Lee, D. D., and Seung, H. S. 2000. Algorithms for non-negative matrix factorization. In *NIPS*, 556–562.

Li, F., and Yang, Y. 2004. Recovering genetic regulatory networks from micro-array data and location analysis data. *Genome Informatics* 15(2):131–140.

Li, Z.; Srivastava, S.; Mittal, S.; Yang, X.; Sheng, L.; and Chan, C. 2007a. A three stage integrative pathway search (tips) framework to identify toxicity relevant genes and pathways. *BMC Bioinformatics* 8(202).

Li, Z.; Zhou, Y.; Zhang, L.; Srivastava, S. S.; Jin, R.; and Chan, C. 2007b. Using knowledge driven matrix factorization to reconstruct modular gene regulatory network. Technical Report MSU-CSE-07-200, Department of Computer Science, Michigan State University.

Liu, Y., and Jin, R. 2007. Boostcluster: boosting clustering by pairwise constraints. In *KDD*.

Pilpel, Y.; Sudarsanam, R.; and Church, G. M. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* 29:151–159.

Segal, E.; Shapira, M.; Regev, A.; Peér, D.; Botstein, D.; Koller, D.; and Friedman, N. 2003. Module networks: Identifying regulatory modules and their condition specific regulators from gene expression data. *Nature Genetics* 34(2):166–176.

Spellman, P. T.; Sherlock, G.; Zhang, M. Q.; Iyer, V. R.; Anders, K.; Eisen, M. B.; Brown, P. O.; Botstein, D.; and Futcher, B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* 9(12):3273–3297.

Srivastava, S., and Chan, C. 2007. Hydrogen peroxide and hydroxyl radicals mediate palmitate-induced cytotoxicity to hepatoma cells: relation to mpt. *Free Radical Research* 41(1):38–49.

Tibshirani, R.; Walther, G.; and Hastie, T. 2000. Estimating the number of clusters in a dataset via the gap statistics. Technical Report Technical Report 208, Dept. of Statistics, Stanford University.

Toh, H., and Horimoto, K. 2002. Inference of a genetic network by a combined approach of cluster analysis and graphical gaussian modeling. *Bioinformatics* 18(2):287–297.

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *Proc. Fourteenth International Conference on Machine Learning*.

Yu, J.; Smith, V. A.; Wang, P. P.; Hartemink, A. J.; and Jarvis, E. D. 2004. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* 3594–3603(18):20.