2006

A New Approach to Identify Functional Modules Using Random Matrix Theory

Mengxia Zhu
Southern Illinois University Carbondale, menxia@cs.siu.edu

Qishi Wu
University of Memphis

Yunfeng Yang
Oak Ridge National Laboratory

Jizhong Zhou
University of Oklahoma Norman Campus

Published in Zhu, M., Wu, Q., Yang, Y. & Zhou, J. (2006). A new approach to identify functional modules using random matrix theory. IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB '06, 1-7. doi: 10.1109/CIBCB.2006.330980

A New Approach to Identify Functional Modules Using Random Matrix Theory

Mengxia Zhu
Computer Science Dept
Southern Illinois University
Carbondale, IL 62901
mengxia@cs.siu.edu

Qishi Wu
Computer Science Dept
University of Memphis
Memphis, TN 38152
qishiwu@memphis.edu

Yunfeng Yang
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831
yangy@ornl.gov

Jizhong Zhou
Botany & Microbiology Dept
University of Oklahoma
Norman, OK 73019
jzhou@ou.edu

Abstract— **The advance in high-throughput genomic technologies including microarrays has generated a tremendous amount of gene expression data for the entire genome. Deciphering transcriptional networks that convey information on members of gene clusters and cluster interactions is a crucial analysis task in the post-sequence era. Most of the existing analysis methods for large-scale genome-wide gene expression profiles involve several steps that often require human intervention. We propose a random matrix theory-based approach to analyze the cross correlations of gene expression data in an entirely automatic and objective manner to eliminate the ambiguities and subjectivity inherent to human decisions. The correlations calculated from experimental measurements typically contain both "genuine" and "random" components. In the proposed approach, we remove the "random" component by testing the statistics of the eigenvalues of the correlation matrix against a "null hypothesis" — a truly random correlation matrix obtained from mutually uncorrelated expression data series. Our investigation on the components of deviating eigenvectors using varimax orthogonal rotation reveals distinct functional modules. We apply the proposed approach to the publicly available yeast cycle expression data and produce a transcriptional network that consists of interacting functional modules. The experimental results nicely conform to those obtained in previously published literatures.**

Keywords: Random Matrix, Microarray, Pearson correlation, Eigenvalue, Eigenvector, Varimax orthogonal rotation.

I. Introduction

The exponential growth of genomic sequence data starting in early 1980s has spurred the development of computational tools for DNA sequence similarity searches, structural predictions, and functional predictions. The emergence of high-throughput genomic technologies in the late 1990s has enabled the analysis of higher order cellular processes based on genome-wide expression profiles such as oligonucleotide or cDNA microarray. Genes now can be affiliated by their co-regulated expression waveforms in addition to sequence similarity and proximity on the chromosome as in gene content analysis. Genes ascribed to the same cluster are usually responsible for a specific physiological process or belong to the same molecular complex. Such transcriptome (mRNAs) datasets deliver new knowledge and insights to the existing genome (genes) datasets, and can be used to guide proteome (proteins) and interactome research that aims to extract key biological features such as protein-protein interactions and subcellular localizations more accurately and efficiently.

However, organizing genome-wide gene expression data into meaningful function modules remains a great challenge. Many computational techniques have been proposed to conjecture the cellular network based on microarray hybridization data. Examples include Boolean network methods, differential equation-based network methods, Bayesian network methods, hierarchical clustering, K means clustering, self-organizing map (SOM), and correlation-based association network methods.

Boolean network method [7], [3] is a coarse simplification of gene network to determine the gene state as either 0 or 1 from the inputs of many other genes. Differential equation-based network models [4] gene networks as a set of nonlinear differential equations that can indicate the gene rate change without the assumption of discrete time steps. Bayesian network gives a graphical display of dependence structure based on conditional probabilities among genes. In hierarchical clustering, a dendogram is constructed by iteratively grouping together genes with the highest correlation, which is essentially a greedy algorithm achieving local optimality and disregards negative association [10]. K means clustering [8] serves as an improved approach to hierarchical clustering but requiring a subjective specification on the number of clusters. SOM [11] is a neural network-based iterative clustering method and also requires the user to estimate the initial cluster number. The correlation-based association network technique has been commonly adopted to identify cellular networks due to its computational simplicity and the nature of microarray data (typically noisy, highly dimensional and significantly undersampled). However, the association network method relies on arbitrarily assigned thresholds for link cutoff, which inevitably introduces subjectivity in network structure and topology. A novel technology, which can determine the structure of transcriptional networks and uncover biological regularities in a computerized and unbiased way, has been under active study by biological scientists.

We propose and develop a system to construct and analyze various aspects of transcriptional networks based on random matrix theory (RMT). Correlation matrix for yeast genome demands a significant amount of computing cycles to calculate all eigenvalues and eigenvectors. High performance computing resources such as supercomputer and Linux cluster as well as

parallel programming techniques should be utilized to address this problem. We aim to tackle computing problems with tens or hundreds of thousands genes within short period of computing time.

The rest of the paper is organized as follows: mathematical model and data preparation are discussed in Section II. In Section III, we present the statistics of correlation matrix. Discussion on deviating eigenstates based on random matrix theory is given in Section IV. Genes are clustered and functional modules are identified in Section V. Experimental results on yeast cycle data are presented in Section VI to demonstrate the effectiveness of our method. We conclude our work in Section VII.

II. Problem Formulation

We define the expression signal of gene $i = 1,...,N$ in various samples $s = 1,...,K$ as:

$$W_i(s) \equiv \ln\left(\frac{Es_i(s)}{Ec_i(s)}\right), \tag{1}$$

where $Es_i(s)$ denotes the expression signal of sample s for gene i, and $Ec_i(s)$ is the corresponding control signal. Due to the various levels of expression signal shown by different genes, we normalize the data as:

$$w_i(s) \equiv \frac{W_i(s) - \langle W_i \rangle}{\sigma_i}, \tag{2}$$

where $\sigma_i \equiv \sqrt{\langle W_i^2 \rangle - \langle W_i \rangle^2}$ represents the standard deviation of W_i, and $\langle W_i \rangle$ stands for the average over different samples for gene i. From this normalized $N \times K$ data matrix M, we calculate the cross-correlation matrix C according to

$$C \equiv \left(\frac{1}{K}\right) MM^T. \tag{3}$$

Pearson correlation coefficient C_{xy} between gene x and y, each with k data series, can also be calculated from Eq. 4:

$$C_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(K-1)s_x s_y}, \tag{4}$$

where s_x and s_y denotes the standard deviations. Pearson correlation ranges from 1 as perfect correlation to -1 as perfect anti-correlation. When $C_{ij} = 0$, no correlation exists between genes i and j.

However, conducting direct study on these empirical cross-correlation coefficients is rather difficult due to the unique properties of microarray experiments. Firstly, the cross-correlation between any pair of genes may not be constant: such co-regulations can fluctuate over time or under different sample conditions. Secondly, the limited number of samples that a microarray is typically conducted upon, may introduce significant "measurement noise" that compromises the accuracy of the underlying correlations.

In order to filter out randomness contained within the empirical cross-correlation matrix, we test the eigenstates of this correlation matrix against those of a controlled counterpart, a truly random correlation matrix generated by computer random generator. Statistic properties that conform to the truly random matrix are labeled as noise contributions; on the other hand, any deviating eigenstates are treated as genuine correlations, which will be amplified and analyzed for transcriptional network construction.

III. Statistics of Correlation Matrix

A. Distribution of correlation coefficients

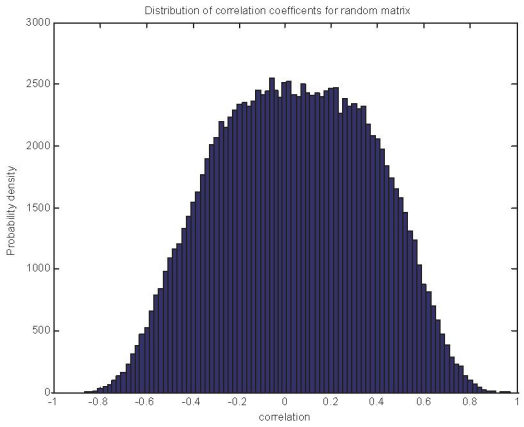
We contrast the distribution $P(C_{ij})$ for cross-correlation matrix C with $P(R_{ij})$, where R denotes a random correlation matrix constructed from a series of mutually uncorrelated data with zero mean and unit variance generated by a computer. Fig. 1(a) shows that $P(R_{ij})$ demonstrates a Gaussian distribution with zero mean, which indicates complete randomness within the data. However, $P(C_{ij})$ as shown in Fig. 1(b) is asymmetric and centered around a positive value in contrast to $P(R_{ij})$, which implies that positive correlation is more pronounced than negative correlation among genes.

B. Distribution of eigenvalues

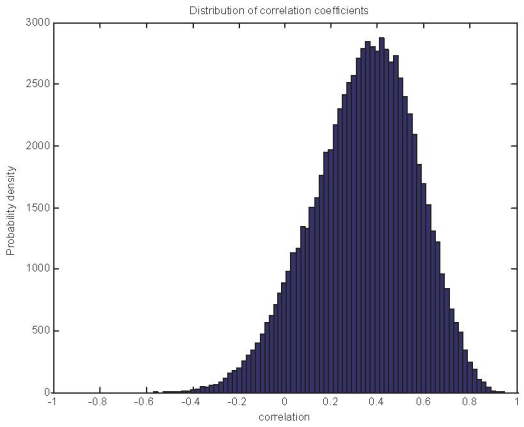
We further compare the probability distributions $P^C(\lambda)$ and $P^R(\lambda)$ of the eigenvalues λ_i calculated from the cross-correlation matrix C and the random matrix R, respectively. Eigenvalues are arranged in a descreasing order such that $\lambda_i > \lambda_{i+1}$. The probability distributions $P^C(\lambda)$ and $P^R(\lambda)$ are plotted in Fig. 2(a) and Fig. 2(b). It has been observed that a set of the eigenvalues of C fall within the well-defined range of $[\lambda_-, \lambda_+]$ calculated from R, with a few deviating from the upper and lower bounds conveying the true correlation information. This observation enables us to separate the real correlation from the randomness. Such denoising process is necessary since microarray data is extremely undersampled and may introduce significant measurement noise. Interestingly, Kwapien *et al.* [6] found that increasing the length of time series or number of samples would cause eigenvalues to deviate more from the random matrix eigenvalue bounds. They declared that the bulk of the correlation matrix is not pure noise as conventionally thought to be. Based on their results, it is possible that more subtle and less prevalent co-regulated gene groups could be squeezed out of the noise segment if we are able to acquire a larger sample size K. However, experimental results are still needed to validate this assumption. In practice, a large sample size K from the perspective of mathematical view is not always feasible for most biological datasets due to the considerable time and material resources involved in bio-related experiments.

C. Distribution of nearest-neighbor eigenvalues

The comparison made above between $P^C(\lambda)$ and $P^R(\lambda)$ alone is not sufficient to show that the majority of the eigenvalue spectrum of C is random. In general, matrices with the same eigenvalue distribution may have different eigenvalue correlations, and vice versa [9]. Hence, we also need to examine the correlation in the eigenvalues of C to determine if it conforms to that of a random matrix.

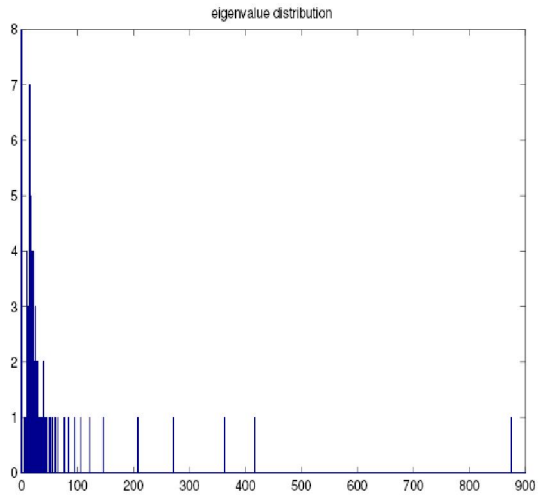


(a) Distribution of random correlation coefficients.

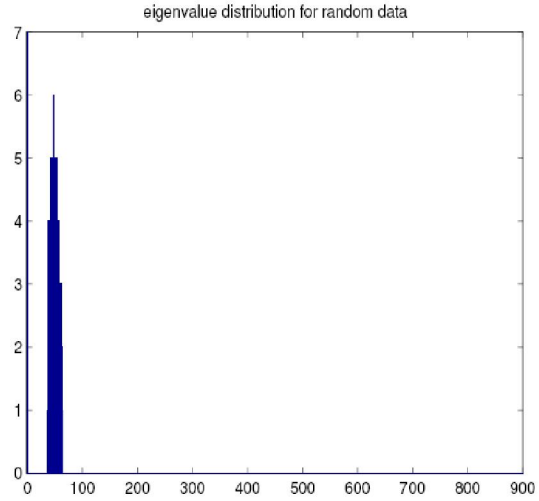


(b) Distribution of gene expression correlation coefficients.

Fig. 1. Comparison of correlation coefficient distributions.



(a) Distribution of eigenvalues of gene expression correlation matrix.



(b) Distribution of eigenvalues of random system.

Fig. 2. Comparison of eigenvalue distributions: x axis represents the eigenvalues and y axis represents the probability densities.

RMT makes two universal predictions for real symmetric matrices: the nearest neighbor spacing distribution (NNSD) of eigenvalues follows Gaussian orthogonal ensemble(GOE) statistics as in Eq. 5, if there exists correlation between nearest-neighbor eigenvalues, while follows Poisson if there is no correlation. In order to ensure the uniform average value for eigenvalue spacing throughout the spectrum, we map the eigenvalues to new unfolded eigenvalues, whose distribution is uniform. Unfolding procedure guarantees that eigenvalue spacing is represented in units of local mean eigenvalue spacing. To realize this, one can replace λ_i by the unfolded spectrum $\lambda_i^u = f_{av}(\lambda_i)$, where $f_{av}(\lambda_i)$ is the smoothed integrated density of eigenvalues obtained by fitting the original integrated density to a cubic spline or by local density average. We compute the nearest neighbor spacing distribution $P(n)$, $n = \lambda_i^u - \lambda_{i-1}^u$. We know that $P(n)$ for random matrix follows the Wigner-Dyson distribution. Our experiments show that the NNSD $P(n)$ for C conforms well with $P_{GOE}(n)$.

$$P_{GOE}(n) \approx \frac{\pi n}{2} \exp\left(\frac{-\pi n^2}{4}\right). \tag{5}$$

These results support the assumption that the majority of the eigenvalues are of randomness in nature both from the perspective of eigenvalue distribution and eigenvalue correlation distribution. Thus, random matrix theory serves as an ideal mathematical tool to investigate microarray datasets that typically have a significant amount of noise and errors.

IV. Deviating Eigenvalues and Eigenvectors

A. Deviating eigenvalue

We consider the set of eigenvalues that deviate from the eigenvalue range of the random matrix as genuine correlation. The amount of variance contributed by each eigenvector (factor) can be reflected by the proportion of eigenvalue over the

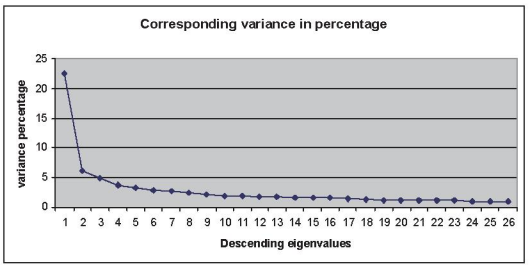


Fig. 3. Variance explained by sorted descending eigenvalues.

sum of all eigenvalues based on principle component analysis (PCA). In other words, principle factors are responsible for the majority of variation within the data. Thus, only large eigenvalues, usually greater than 1, and their corresponding eigenvectors are retained for further treatment and gene group analysis. The rest of the eigenstates contain either insignificant or noisy information. We can see from Fig. 3, a plot of variance versus eigenvalue, that a large proportion of variation is picked up by the first several large eigenvalues.

B. Deviating eigenvector components

Deviating eigenvalues naturally lead us to the investigation of their corresponding deviating eigenvectors. There are N eigenvectors u^i in total, $i = 1...N$. Each eigenvector u^i has N components corresponding to N gene variables. All eigenvectors are perpendicular(orthogonal) to each other and are normalized to length of 1. The probability distribution of eigenvector components for different eigenvalues are plotted and compared against that of a random matrix, which follows Gaussian distribution with zero mean and unit variance.

$$\beta(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right). \tag{6}$$

The probability distribution of the eigenvector components with the corresponding eigenvalue λ_k from the bulk $\lambda_- \leq \lambda_k \leq \lambda_+$ shows a good agreement with Gaussian distribution as indicated by the lower right graph in Fig. 4. The deviating eigenvector components demonstrate a significant deviation from the Gaussian distribution as shown by the upper and lower left graphs in Fig. 4. It has been also observed that the distribution curve is gradually reforming to approximate the shape of a Gaussian distribution when eigenstates approach the characteristic region represented by a random matrix.

V. Functional Modules Identification

A. Collective behavior from the largest eigenvalue

The observation also shows that if the majority of gene expression correlations are co-regulated, the eigenvector components corresponding to the largest eigenvalue with contribution from almost all genes have the same sign as shown in Fig. 5. Such eigenvector components distribution can be commonly found in a specific gene cluster, where most of the genes are co-regulated with a few to be anti-co-regulated.

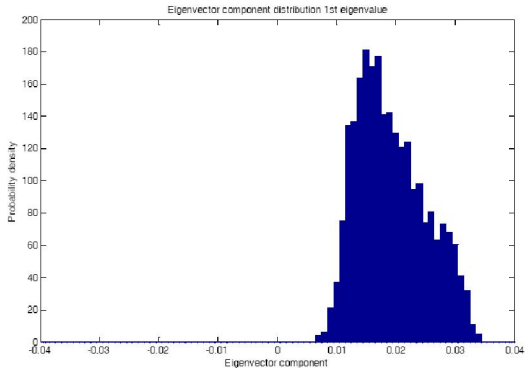


Fig. 5. Distribution of eigenvector component for u^1.

It further supports the existence of housekeeping genes in a significant number, which are constitutively expressed to carry out basic cell functions needed for the sustenance of the cell. Similar phenomena are also observed in financial stocks data, and can be interpreted as a common influence to all stocks by certain stimuli such as newsbreaks of interest rate increase [9]. We quantify the alike collective behavior of the entire genome by eigensignals computed as the scalar product of the sample series on the first eigenvector u^1:

$$z^1(s) \equiv \sum_{i=1}^{N} u_i^1 W_i(s). \tag{7}$$

We have the following when the common influence effect is considered:

$$W_i(s) = \alpha_i + \beta_i z^1(s) + \varepsilon_i(s), \tag{8}$$

where $z^1(s)$ is common to all genes, and α_i and β_i are gene-specific constants, which can be estimated by least squares regression. The largest eigenvalue is usually an order of magnitude larger than the rest of the eigenvalues. Such strong eigensignal can significantly suppress the effect of other eigenvalues because of the fact that $\sum_1^N \lambda_i = Tr(C) = N$. We want to remove the effect of λ_1 in order to augment the impact of the remaining eigenvalues for easy and reliable study. From Eq. 8, we calculate the residuals $\varepsilon_i(s)$ as the matrix elements to construct a new correlation matrix C'. The eigenstates of C' are then analyzed to build transcriptional networks capable of revealing some subtle gene clusters that might have been masked by the largest eigenvalue λ_1.

B. Loading factor and orthogonal rotation

After acquiring a set of normalized eigenvectors, we transform the eigenvector components to loading factors by taking the multiplication of vector components and the square root of corresponding eigenvalue. Each eigenvector represents one factor leading to one gene cluster. A larger loading factor indicates that the corresponding gene "load" more on that eigenvector, or that gene is more expression-dominating for that cluster. To simplify the eigenvector structure and make

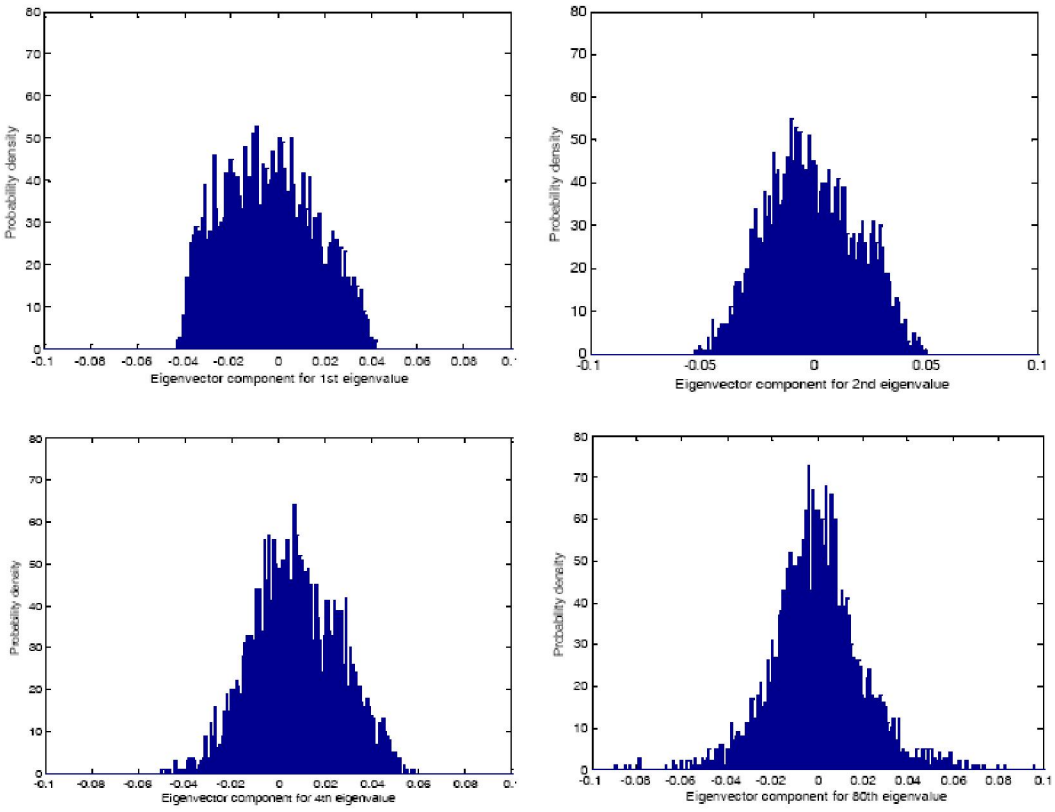


Fig. 4. Distribution of eigenvector component for u^1, u^2, u^4, u^{80}.

the interpretation of gene clusters easier and more reliable, we apply orthogonal rotation to the retained eigenvectors. Since the rotation is performed in the subspace of the entire eigenstates, the total variation explained by the newly rotated factors are always less then the original space, but the variation within the subspace remains the same before and after rotation, only the partition of variation changes.

VARIMAX [5] is a simple and popular rotation method that transforms the principle data axes such that each eigenvector will contain a small number of large loadings and a large number of zeroes or small loadings. Biologically speaking, each gene tends to load heavily on only one or a few gene clusters. Thus, gene clusters consist of a reduced number of genes compared with pre-rotation results. The rationale behind VARIMAX is that a rotation (linear combination) of the original factors is searched in order to maximize the variance within factor loadings. A rotation matrix R can be determined to specify such rotation as following:

$$R = \begin{bmatrix} \cos\theta_{i,i} & \cos\theta_{i,j} \\ \cos\theta_{j,i} & \cos\theta_{j,j} \end{bmatrix},$$

where $\theta_{i,j}$ is the rotation angle from old axis i to new axis j. The graphical representation for a 2D orthogonal rotation is illustrated in Fig. 6 with dotted lines representing new axes.

C. Stability of gene clusters in samples

The stability of gene clustering based on our eigenstates analysis can be evaluated in terms of variance of total expression signals denoted by Z^i for eigenvector i among different samples and time series. The variances are directly associated with the corresponding eigenvalues as one of the most important properties of eigensignals [6] in Eq. 10. The gene cluster derived from eigenvector with larger eigenvalue is more unstable compared with gene cluster associated with smaller eigenvalue. Note that variance levels indicate the consistence of gene members across different samples.

$$z^i(s) \equiv \sum_{k=1}^{N} u_k^i W_k(s) \tag{9}$$

$$Stab(u^i) = \sigma^2(Z^i) = (u^i)^T C u^i = \lambda_k \; where \; i = 1, \ldots N \tag{10}$$

VI. Experimental Data Analysis

The program in this work is implemented in C++ and currently runs on a single workstation[1].

The components of a deviating eigenvector with large values are identified as gene members belong to a specific functional

[1] We are now in the process of transitting our system from a workstation to a supercomputer or Linux cluster running ScaLAPACK [1] for parallel eigenstates computation.

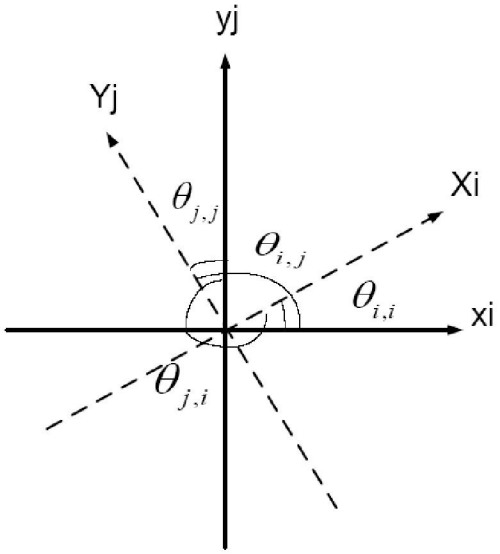


Fig. 6. A 2D orthogonal rotation.

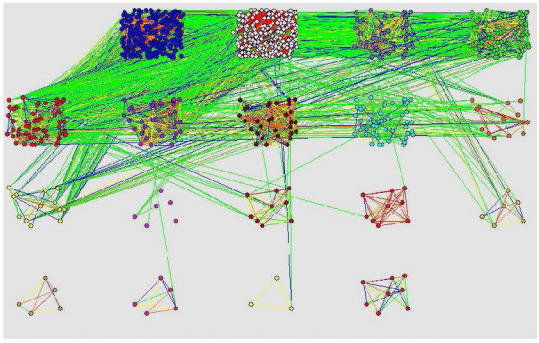


Fig. 7. Yeast transcriptional network.

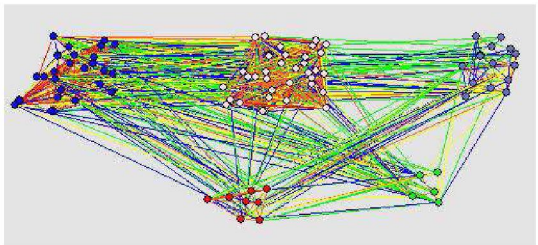


Fig. 8. Subgroups within one group.

Functional Modules	Gene Members	Functional Modules	Gene Members
Protein biogenesis	RPL5B RPL34B RPS19A RPS26B RPL7B RPL14A RPS6B RPL33B RPS1B RPL9A RPL11B RPS24A RPL26B RPS15 RPL24A RPL12B RPS0A RPS19B RPL19B RPL27B RPS29B RPL13A RPS4B RPL33A RPS29A RPL18B RPL11A RPS21A RPS7A RPL24B RPL9B RPL6B RPS20 RPL17B... (omit the rest)	Heat shock protein	SSA2 SSA1 ECM10 SSA4 KAR2 STI1 SIS1 SUR2 YME1 SRP21 FCY1 SUI1 VAM3 PFD4
		Histone	HHT1 HTA1 HTB1 HHF2 HTA2 HHT2 HHF1 HTB2 HHO1
DNA replication and repair	CDC45 KIN2 MSH2 DUN1 POL30 SMC3 RFA2 RNR1 SWE1 POL2 POL32 RFA1 SEN34 PIF1 SPH1 OGG1 CDC21 RHC18 RFC3 RAD27 TUB4 ZDS2 ASF1 HOM1 MSH6 RFC5 SAS2 BIN4 CDC2 PMS1 ARP1 POL1 TOP1 RAD51 FCP1 PMT5 PCH1 RNR3 MRE11 ALG2 PRI1 RFA3 BUD2 DUT1 ASF2 ADK2 CDC9 ECM25 RFC4	TCA cycle	CYT1 SDH2 MDH1 SDH4 QCR6 ACH1 COR1 SDH3 RIP1
		allantoin utilization	DAL2 DAL1 DAL3 DAL5 MEP2
Protein degradation	SCL1 PUP2 PRE2 PRE5 PRE1 PRE3 RPN6 PRE4 RPT6 PRE9 RPT1 RPN8 RPT4 RPN10 RPN11 PRE7 PUP1 RPN3 PRE8 PRE10 UFD1 UMP1 PRE6 RPN12 PUP3 RPN7 RPT3 RPN2 RPT5 GSH2 QRI8 STE24	Energy metabolism	MRPL9 NONE MRPL13 ATP12 PET123 COX12 MRPL35 NONE MRP2 ATP11 MRP17 DIB56 NONE QCR8 MRP51 RPA2 HXT4 COX17 NONE RML2 MRP35 MRPL38 MRP7 MRPL25 ATP14 MSS51 FMC1 MSM1 NONE MRPL8 CYT2 COX14 MEF1 NONE MRPL32 MSN4 CBP4 PHB1 MSF1 MRPL16 MRPL6 IMG2 MRP13 NAM2 MAS1 MRPL36 NAM9 CYC3 NONE KIM4 CBP3 ST51 COX10 MBA1 MRPL31 ECM19 STP4 CAF16 COX13 CAP1

Fig. 9. Some functional modules and their gene members.

Functional Modules	Gene Members
Glycolysis	TDH2 ENO2 TDH3 PGK1 TPI1 TDH1 FBA1 GPM1 ADE3 ERG20 PDC5 TKL1 TFB2 CPH1 NIP29 CYS3 CSE4 HXK2 NONE TYE7 ACS2 PDS1 PFK1
Cell cycle	HXT10 CDH1 SOL1 FET5 SUT1 GIP1 CDC27 UBC1 PCT1 ALG9 NONE CLB3 MIP6 YCK3 REV7 NMT1 CDC20 CDC10 GPI8 PES4 CDC16 CDC26 MET4 YUH1 VPS9 PIG2 EHD2 PAP1 CCC1 CRN1 CDC3 SPC42 DIN7 KAR1 CNM67 VPS35 BBP1

Fig. 10. Sub-functional modules and their gene members.

module involved in a similar cellular pathway. Yeast cell cycling data, as used in the manuscript, is one of the best known microarray dataset and has been extensively evaluated. Therefore, the structure of the network has been quite well understood. By applying RMT method on this dataset, we have demonstrated that our results are consistent with available biological knowledge, which allows us to reach the conclusion that we have identified functional modules. The entire yeast genome is partitioned into a large number of functional modules sharing similar expression patterns based on public available microarray data downloaded from yeast cell cycle project at [2].

Fig. 7 shows 17 distinct modules such as protein biogenesis, DNA replication and repair, energy metabolism, protein degradation, heat shock protein, TCA cycle, protein folding, allantoin mechanism, and histone. Various colors of the edge represent different ranges of correlation value between two genes (vertices). It can be visually observed that correlations within groups are much higher represented by red or orange links than correlations between different groups indicated by blue or green links, which strongly indicates the effectiveness of our clustering approach. For groups with a large number of genes, we recursively apply the same method to identify subgroups within large groups. Fig. 8 is the refined submodular

results for the first group in blue with 230 genes in Fig. 7. Two major submodules are identified as glycolysis and cell cycle.

VII. Conclusions

High-throughput genomic technologies such as microarrays have provided gene expression data at the transcription level. Its unprecedented power for the study of gene expression of thousands of genes simultaneously can be potentially used to unveil the topology and functions of transcriptional networks. In this paper, we explored random matrix theory and orthogonal rotation techniques to dissect transcriptional networks and identify various functional modules.

Luo et al [**?**] also proposed a random matrix theory-based approach to infer transcriptional networks based on microarray data. However, their analysis is mainly focused on eigenvalues. In addition, their method require more computation cycles to calculate eigenvalues for many different correlation matrices. In our approach, we only need to compute eigenstates for one correlation matrix.

Most previous clustering methods partition members into non-overlapping groups. However, in our method, one gene is allowed in multiple groups, which is a legitimate assumption from the biological perspective since a single gene may get involved in different pathways. Transitively co-regulated genes, which are not directly correlated but both of which have correlation with the same gene, can also be detected and grouped. Our method is computationally efficient, objective without human intervention, and robust to high levels of noise. Function of unknown genes are conjectured and explored through their associated function modules.

Since our computational analysis is solely based on a single microarray dataset, we only obtain rough structure of functional modules. If genes in the same functional module do not show significant correlation in expression pattern, we will not be able to identify them using RMT method. It is likely that genes in the same functional module show significant correlation under one condition but not under another condition (For example, module of heat shock proteins are rarely identified in other yeast microarray dataset). By consolidating results from multiple microarray datasets, we could improve the integrity of functional modules. The authors will work toward this direction in the future.

In general, we think that "all the subjective factors induced by humans built into the microarray data itself" can be divided into two categories: systematic subjective factors (e.g. overlook low-density signals of spots if signal/background signal is set to be high, which will impact every slide of the whole dataset) and random subjective factors. RMT method, or any existing clustering method (e.g. hierarchical clustering, K-means, SOM, etc.), is unable to deal with systematic subjective factors. On the other hand, RMT method is capable of removing the random subjective noise, which normally lead to low correlation between genes.

It would be our future interest to apply this method to human genome data with 30k genes. A highly parallel implementation of our algorithm is needed to address large-scale biological applications. Our code can easily migrate to supercomputers or cluster machines to utilize high performance computing resources. Some advanced visualization techniques will also be introduced at a later stage to aid data comprehension and inspire discoveries.

References

[1] Scalapack. http://www.netlib.org/scalapack/scalapack_home.html.

[2] Yeast. http://genome-www.stanford.edu/cellcycle.

[3] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under theboolean network model. *Pacific Symp. Biocomp*, 4:17–28, 1986.

[4] T. Chen, H.L. He, and G.M. Church. Modeling gene expression with differential equations. *Pacific Symp. Biocomp*, 4:29–40, 1999.

[5] H.F. Kaiser. The varimax criterion for analytic rotation in factor analysis. volume 23, pages 187–200, 1958.

[6] J. Kwapien, P. Oswiecimka, and S. Drozdz. The bulk of the stock market correlation matrix is not pure noise. *Physica A*, 359:589–606, 2006.

[7] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symp. Biocomp*, 3:18–29, 1998.

[8] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1998.

[9] Vasiliki. Plerou, Gopikrishnan. Parameswaran, Rosenow. Bernd, Lus A. Nunes. Amaral, and Thomas. Guhr. Random matrix approach to cross correlations in financial data. *Physical Rev*, 65(066126):1–18, 2002.

[10] R.R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci Bull*, 38:1409–1438, 1958.

[11] P. Toronen, M. Kolehmainen, G. Wong, and E. Castre'n. Analysis of gene expression data using self-organizing maps. *FEBS L.*, 451:142–146, 1999.