Nontrivial Characteristics Embedded in Microarray Coexpression Profiles

Chuanyang Yin\textsuperscript{1}, Huijie Yang\textsuperscript{2}, Guimei Zhu\textsuperscript{3}, Bing-hong Wang\textsuperscript{4,*}

\textsuperscript{1} College of Information and Control, Nanjing University of Information Science & Technology, Nanjing 210044, China
\textsuperscript{2} School of Management, University of Shanghai for Science and Technology, Shanghai 200093, China
\textsuperscript{3} Department of Physics and Centre for Computational Science and Engineering, National University of Singapore, Singapore 117542
\textsuperscript{4,*} Department of Modern Physics and Centre for Nonlinear Science, University of Science & Technology of China, Hefei 230026, China

E-mail: bhwang@ustc.edu.cn

Abstract. We propose the biological robustness as a key criterion in inferring functional networks from microarray coexpression profiles. The networks are constructed by filtering out the small values in the coexpression correlation matrices. The robustness refers to there exists a wide region of the filtering criterion value, in which the constructed networks have almost the same properties, such as the degree distribution, the motif pattern, the edge property and so on. This kind of networks are considered biologically meaningful.

1. Introduction

Understanding gene regulatory mechanism at genome-wide level is one of the essential topics in systems biology \cite{1, 2}. Microarray technology can measure simultaneously the expressions of tens of thousands of genes, which makes it possible to unveil gene expression networks at large scale. However, there exist two challenges in inferring of the networks, namely, the noises and the dimensionality. The noises come from the stochasticity of the gene expression events and the substantial noises incurred by technologies. Generally, we can only conduct limited number of replications in typical microarray experiments. The dimensionality refers to the number of available data points is much smaller than the number of genes. The key to the problems is to introduce restrictions from biological information\cite{5, 52}.

Various methods have been proposed to infer the underlying networks \cite{3, 5, 6}, such as the genetic algorithms \cite{4}, neural networks \cite{7}, Bayesian models \cite{8–10}, Boolean dynamical models \cite{11–13} and ordinary differential equations-based reverse-engineering algorithms \cite{14–19}. But these methods can not overcome the data shortage, i.e. the dimensionality problem, and/or the computational inefficiency. A predominant and fruitful approach is to study the correlations between gene expression experiments \cite{20–25}. The basic assumptions for the method are that the genes performing associated biological functions should exhibit similar expression patterns and the interactions between genes are linear. This approach allows a relatively scale-free estimate of the gene expression similarities and the identification of positive and negative correlations between genes. However, there are still some debates on the assumptions. And
the range-restriction-effect confines the analysis to high-variance-variables, which leads to the indeterminacy of the measure of relatively small fluctuations of gene expressions. Overall, the mentioned methods have to be supplemented by extra restrictions, based upon biological knowledge, to reconstruct correctly the functional networks.

Besides the biological information such as phenotypes, functional classification, and cellular responses, the other resource of biological knowledge is to find the nontrivial characteristics embedded in the correlations between gene expressions. The nontrivial characteristics shed light on the regulatory mechanisms obeyed in the biological processes.

The relations between different elements of a complex system can be described with complex networks. The nodes and edges represent the elements and the relations, respectively. Application of complex network theory to the analysis of biological data has provided insight into the topology of biological networks [26–28].

In this paper, with the concepts in network theory we try to find biologically meaningful characteristics embedded in the correlations of gene expressions, which may provide more clues and restrictions to the reconstruction of the functional networks.

2. Materials and Methods
2.1. Materials
We study two microarray expression profiles of Yeast. One is the microarray expression data of 287 Yeast mutants [29]. The genes expressed in most mutants are considered, resulting a total of 6209 genes. The other is the response expressions of Yeast cells to environment temperature changes [30]. Totally 6090 genes expressed in most of the experiments are selected.

2.2. Gene Expression Correlations
The Pearson correlation coefficient is widely used as a metric to calculate the similarity estimations between different genes. We denote the expression measurements with $X_{nm}$, where $m = 1, 2, \ldots, M$ are the responses to the totally $M$ different conditions, $n = 1, 2, \ldots, N$ the labels of the genes. $X_n = \{X_{n1}, X_{n2}, \ldots, X_{nM}\}$ is called the expression series of the gene $i$. The coexpression correlation matrix can be constructed as,

$$R_{ij} = \frac{\sum_{m=1}^{M} (X_{im} - \bar{X}_i) (X_{jm} - \bar{X}_j)}{\sqrt{\sum_{m=1}^{M} (X_{im} - \bar{X}_i)^2} \cdot \sqrt{\sum_{m=1}^{M} (X_{jm} - \bar{X}_j)^2}},$$

where $\bar{X}_i$ and $\bar{X}_j$ are the averages of the expressions for genes $i$ and $j$, respectively. To filter out the positive failures due to noises in $R$, we can introduce a threshold $r_c$. It is widely accepted that the correlation values larger than $r_c$ correspond to edges with high confidence. In this way, we can construct the functional network $A$, whose adjacency matrix reads,

$$A_{ij} = \begin{cases} 0 & i = j, \\ 1 & R_{ij} \geq r_c, \\ -1 & R_{ij} \leq -r_c, \\ 0 & \text{otherwise}. \end{cases}$$

How to determine the criterion $r_c$ is the key task in the procedure. In literature [31, 32], a special value is assigned to $r_c$, at which the degree distribution of the constructed network obeys a power-law and the number of edges is as large as possible. In Ref. [33], it is found that with the increase of the filtering value there exists a critical point at which the distribution function of the nearest neighbor level spacings for the spectrum of the constructed network changes from Wigner to Poisson distribution (the two extremes $\beta = 1$ and $\beta = 0$ for the distribution function, $\sim s^{\beta} \cdot e^{-s^{\beta+1}}$). This critical point is proposed as the criterion $r_c$. Carter et al propose a $P$-value
criterion to construct the networks [34]. Independently permuting the components of each gene expression series and recalculating all the correlations, the resulting correlation distribution can be used to estimate a $P$-value for each correlation coefficient in $R$. The $P$-value threshold is $10^{-5}$. While in a recent study [35], a maximum value of $r_c$ is used at which the network keeps connected.

This kind of characteristic transition-based strategies can help us to find useful information on the regulatory mechanisms, but different criteria may lead to different yet artificial results and conclusions. Actually, we have not enough evidences to support the idea that the gene regulatory networks should exist at the transition states. In this paper, we consider the evolutions of the properties of networks constructed with different values of $r_c$. We examine systematically the degree distribution, the survival edges and the motif patterns of the constructed networks. The nontrivial characteristics in the evolutions may provide more biologically meaningful restrictions, which may lead to a much more reasonable criterion $r_c$.

2.3. Robustness

Besides the characteristic transition-based criterion, the robustness may be a much more reasonable benchmark for constructing networks. Regarding the correlation coefficient $R_{ij}$ as the essentiality of the corresponding edge between genes $i$ and $j$. It is reasonable to assume that the real networks, from high essentiality core as a seed to the final real functional network, obey the same set of biologically meaningful rules. The fractal structures of many kinds of biological networks support this assumption [36–38]. Namely, a constructed network reflecting correctly the regulatory relations should be criterion-independent. The impacts of the criterion value, i.e., the artificial characteristics, should be negligible.

The value of $r_c$ determines the characteristics of the constructed functional network. If $r_c$ is extremely small, the biological meaningful patterns will be submerged by the artificial edges from weak correlations. With the increase of $r_c$, the artificial edges are filtered out step by step and the biological meaningful patterns tend to dominate the network structures. However, an unreasonable large value of $r_c$ will destroy the patterns by eliminating most of the meaningful edges. Hence we should expect a wide region of $r_c$ in which the constructed networks can capture the characteristics of the real functional networks. If the networks in this region have almost same characteristics, we can conclude that the constructed networks are governed by the same law, i.e., criterion-independent. This range of $r_c$ indicates robustness of organism.

The degree distribution function (DDF) of the constructed networks is used as the measure to detect the criterion-independent characteristics. With the increase of the criterion $r_c$ the width of degree distribution, denoted with $K$, will decrease monotonically. The criterion-independent implies that the form of DDF keeps unchanged. This is a kind of scale-invariance (the scale is the width $K$). The DDF can then be expressed as [39–41],

$$P(k) \sim \frac{1}{K^\sigma} \cdot F \left( \frac{k}{K^\sigma} \right),$$

where $k$ is the degree and $\sigma$ the self-similarity exponent. For a power-law distribution, it equals to the power-law exponent, while for a Gaussian distribution the self-similarity property is trivial, i.e., $\sigma = 0.5$.

2.4. Survival Edges

With the increase of the criterion $r_c$ more and more edges with low confidence are discarded and the survival edges are supposed to be biologically meaningful. Hence, from this evolution process we can separate the structure patterns in the real functional network from the artificial edges. And then we can find the statistical characteristics of the artificial edges and the structure patterns of the real network.
For the edge $A_{ij} = 1$, we can describe its property by using the properties of nodes $i$ and $j$ at the two ends of the edge. Herein, we introduce four quantities,

$$
\begin{align*}
    k_{i-j} &= |k_i - k_j|, k_{i+j} = k_i + k_j; \\
    C_{i-j} &= |C_i - C_j|, C_{i+j} = C_i + C_j,
\end{align*}
$$

where $k_i$, $k_j$ and $C_i$, $C_j$ are the degrees and the clustering coefficients of the nodes $i$ and $j$, respectively. $k_{i-j}$, $k_{i+j}$, $C_{i-j}$ and $C_{i+j}$ are called edge degree difference (EDD), edge degree summation (EDS), edge clustering difference (ECD) and edge clustering summation (ECS), respectively.

2.5. Motifs

Motifs are some special subgraphs containing several connected nodes [42]. These subgraphs occur with significant high probabilities compared with that in the corresponding randomized networks. This concept is firstly introduced by Milo et al to describe the local structures of complex networks [43]. Detailed works show that the gene transcription, protein-protein interaction and metabolic networks can be characterized with different kinds of motifs. And the motifs can have different dynamical characteristics, which in turn can act as bricks in biological processes. In this paper we consider the motifs in the constructed networks with different values of $r_c$. The changes of the motif patterns imply the changes of biological functions for the networks.

We consider all the subgraphs containing three nodes. For each constructed network, we detect all the possible three-node subgraphs in the networks and reckon the numbers of the subgraph patterns, respectively. Then generating an ensemble of complete random networks with same numbers of edges and nodes, from which we can obtain the average numbers of the subgraphs. In order to find occurring significance of a certain subgraph, a statistical quantity $Z$-score is often used, which reads,

$$
Z_i = \frac{(N_i^{\text{real}} - \langle N_i^{\text{rand}} \rangle)}{\text{std}(N_i^{\text{rand}})},
$$

where $N_i^{\text{real}}$ is the occurring number of subgraph $i$ in the real network, $\langle N_i^{\text{rand}} \rangle$ and $\text{std}(N_i^{\text{rand}})$ the average and the standard deviation of the occurring numbers of subgraph $i$ in the ensemble of randomized networks, respectively. If the $Z$-score of a subgraph is high, we call the subgraph a motif of the network. For convenience, $Z$-score is normalized to $z_i$ as [44],

$$
z_i = \frac{Z_i}{\sqrt{\sum_j Z_j^2}}.
$$

The summation covers $Z$-scores of all the possible subgraphs.

3. Results

At different values of the criterion $r_c$ we construct the networks for the cells response and the mutants profiles, respectively. Fig.1 presents several typical DDFs for the cells response networks at different $r_c$. With the increase of $r_c$ we can find a wide region of $r_c$ in which the DDF obeys power-law. The region is $r_c \in (0.42, 0.92)$. For the mutants networks, we have similar results as shown in Fig.2. The region where the DDF obeys power-law is a slightly different as, $r_c \in (0.47, 0.92)$. Out of the regions the DDF will depart significantly from power-law behavior.

Figure 3 presents the slopes of the DDFs for networks at different $r_c$. We can find that though the DDFs have power-law behaviors, at the sub-regions near the upper and lower borders the slope changes significantly from 0.55 to 1.25. For the cells response networks there exists a
Figure 1. (Color online) Typical DDFs for cells response networks constructed with different values of $r_c$. In a wide region of $r_c \in [0.42, 0.92]$ the DDFs have power-law behaviors.

Figure 2. (Color online) Typical DDFs for mutants networks constructed with different values of $r_c$. In a wide region of $r_c \in [0.47, 0.92]$ the DDFs have power-law behaviors.

region, $r_c \in [0.63, 0.88]$, in which the slope keeps almost unchanged, $\sigma = 1.00 \pm 0.05$, i.e., criterion-independent. The region for mutants networks is, $r_c \in [0.62, 0.82]$. That is to say, the networks in the regions have self-similarity characteristics. These robust networks should capture the real functional patterns.

With the increase of $r_c$, the artificial edges are filtered out step by step. Using the quantities of EDD, EDS, ECD and ECS we try to find the characteristics of the biological meaningful edges (survival edges) and the artificial edges. The simulations, which start with the same network constructed by the Pearson correlation coefficient, are performed in two methods, i.e., our $r_c$ based filtering process and randomly discarding edge process. At the beginning of the simulations, we build an initial network with a very small $r_c$, say, $r_c=0.05$. The filtering processes are performed with the ascending values of $r_c$, so we can get a series of networks and
Figure 3. (Color online) The relation between the slope $\sigma$ and the criterion $r_c$. Near the borders $\sigma$ changes significantly from 0.55 to 1.25. We can find an region in which $\sigma$ keeps almost unchanged as 1.00 $\pm$ 0.05. The regions for the cells response and mutants networks are, [0.63, 0.88] and [0.62, 0.82], respectively.

the number of abandoned edges in the previous step, respectively. Meanwhile, the randomly filtering processes followed as a reference, which discard the same number of randomly selected edges. At each of the step, we calculate EDD, EDS, ECD, and ECS for each removed edge and calculate the distribution function of the results.

Fig. 4 shows several typical results for cells response networks. It is found that at a small value of $r_c$, the filtering process is significantly different from the randomly discarding process. The edges with large values of $EDD$ and $ECD$ prefer to be filtered out firstly compared with the corresponding random process. That is, an edge between two nodes with close properties (degree and clustering coefficient) tends to survive, which may indicate these edges have high confidence of biological meaning. We can find also that edges with large $EDS$ tend to be conserved. We can not find distinctive differences between the ECS distributions for the reconstruction procedure and the random one.

When the value of $r_c$ becomes large enough, e.g., $r_c = 0.8$ and 0.9, the filtering process tends to become indistinguishable from the random one. At this stage, the biologically meaningful edges are discarded, which will destroy the real network patterns.

Similar conclusions can be reached for the mutants networks (not shown).

There are totally 7 patterns for the subgraphs containing three nodes, as shown in Fig.5a. The red dashed and black solid edges denote the positive and negative relations between two genes. Fig. 5b and Fig. 5c exhibit the $Z$-scores, $z_i$, for the subgraphs numbered ID-1 to ID-7 in the cells response and mutants networks, respectively.

Generally, the subgraphs numbered ID-4 and ID-6 are motifs in the wide robust region of $r_c$. When $r_c$ reaches the upper border of the robust region, graph ID-6 are not motif for the cells response networks. And even the original motifs are replaced by the graphs ID-1, ID-2 and ID-3 as new motifs in the mutants networks. This characteristic not only verifies the existence of a wide robust region of $r_c$ in which the constructed networks are biologically meaningful, but also provides a criterion to determine the upper border of $r_c$. 
4. CONCLUSION

The reconstruction of regulatory networks from microarray data is one of the essential problems in systems biology. To get useful information, we must overcome two challenges, the noises and the dimensionality. Various methods have been proposed, but no method can reach valuable results without the supplementary of biological knowledge. Besides the biological information from experiments, finding nontrivial characteristics embedded in microarray data can also provide much more useful information on regulatory mechanisms.

In this paper, the measures of complex networks such as the degree, edge properties and
Motif patterns are used to find the biologically meaningful characteristics embedded in the gene coexpression correlation matrices. These measures provide characteristics of the constructed networks from microscopic to macroscopic scales.

We find that there exists a wide region of the criterion $r_c$, in which the constructed networks are robust, namely, the DDF keeps unchanged. This kind of scale-invariance is criterion-independent and biologically meaningful.

In the robust region, with the increase of $r_c$ the filtered out edges are significantly different from the ones randomly selected. This nontrivial filtering procedure implies its biological meaning. When $r_c$ reaches the upper border of the robust region, the significance of this procedure tends to be indistinguishable from the random one.

The existence of the robust region is confirmed by the motif patterns in the constructed networks. In the main part of the region, the motifs keep unchanged, i.e, the subgraphs ID-4 and ID-6 are motifs. Near the upper border, ID-6 will not be a motif for cells response networks and even the graphs ID-1, ID-2, ID-3 become motifs instead of the original motif patterns ID-4 and ID-6 for the mutants networks.

Figure 5. (Color online) (a) All the possible types of subgraphs containing three connected nodes. The red dashed and black solid edges denote up-regulatory and down-regulatory relations respectively; (b) $Z$-scores for the total 7 subgraphs in the cells response networks. Near the upper border of the robust region only the subgraph ID-4 survives as motif; (c) $Z$-scores for the total 7 subgraphs in the mutants networks. Near the upper border of the robust region the subgraphs ID-4 and ID-6 are replaced by the subgraphs ID-1, ID-2 and ID-3 as motifs.

To conclude, based upon the biological robustness of organism, the determination of criterion $r_c$ is a natural process and the robust region $r_c$ should lead much more reliable networks.
our method may provide a new perspective in referring functional networks from microarray coexpression data.

Acknowledgments
This work is partially supported by the National Natural Science Foundation of China (Grant No. 61273229), the Research fund of Nanjing University of Information Science & Technology (20110395). The authors are grateful to B. W. Li for useful discussions.

References
[1] Cho K, Choo S, Jung S, Kim J, Choi H and Kim J 2007 IET Syst. Biol. 1 149
[2] Sprinzak D and Elowitz M 2005 Nature 438 443
[3] Bansal M, Belcastro V, Ambesi-Impiombato A and Bernardo D 2007 Mol. Syst. Biol 3 78
[4] Wadde M and Hertz J 2000 BioSystems 55 129
[5] Hecker M, Lambecka A, Toepferb S, Somerenc E and Guthkea R 2009 BioSystems 96 86C103
[6] Werhli A, Husmeier D 2008 Journal of Bioinformatics and Computational Biology 6 543-72
[7] D’Haeseleer P, Liang S and Somogyi R 2000 Bioinformatics 16 707
[8] Hartemink A, Gifford D, Jaakkola T and Young R 2001 in Pac. Symp. Biocomputing 6 422
[9] Yu J, Smith V, Wang P, Hartemink A and Jarvis D 2004 Bioinformatics 20 35949
[10] Dojer N, Gambin A, Mizera A, Wilczynski B and Tiuryn J 2006 BMC Bioinformatics 7 249
[11] Akutsu T, Miyano S and Kuhara S 1999 Pac. Symp. Biocomput. 4 17
[12] Akutsu T, Miyano S and Kuhara S 2000 Bioinformatics 16 727
[13] Martin S, Zhang Z, Martino A and Paulon J 2007 Bioinformatics 23 866
[14] Someren E, Wessels L and Reinders M 2000 Proc. Int. Conf. Intell. Syst. Mol. Biol. 8 355
[15] Yeung M, Tegne J and Collins J 2002 Proc. Natl. Acad. Sci. USA 99 6163
[16] Gardner T, Bernardo D, Lorenz D and Collins J 2003 Science 301 102
[17] Someren E, Vaes B, Steegenga W, Sijbers A, Dechering K and Reinders M 2006 Bioinformatics 22 477
[18] Bonneau R,Reiss D,Shannon P, Facciotti M, Hood L, Baliga N and Thorsson V 2006 Genome. Biol. 7 R36
[19] Davidson E, McClay D and Hood L 2003 Genome Res. 13 773
[20] Crescenzi M and Giuliani A 2001 FEBS Lett. 507 114
[21] Basso K, Margolin A, Stolovitzky G, Klein U, Dalla-Favera R and Califano A 2005 Nat. Genet. 37 382
[22] Butte A and Kohane I 2000 Pac. Symp. Biocomput. 5 418
[23] Faith J, Hayete B, Thaden J,Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins J and Gardner T 2007 PLoS Biol. 5 e8
[24] Watts D 1999 Small Worlds (Princeton: Princeton University Press)
[47] Yang H, Zhao F and Wang B 2006 Physica A 364 544
[48] Yang H, Zhao F and Wang B 2006 Chaos 16 043112
[49] McGraw N and Menzinger M 2007 Phys. Rev. E 75 027104
[50] Bandyopadhyay J and Jalan S 2007 Phys. Rev. E 76 026109
[51] Ivanov P, Amaral L, Goldberger A, Havlin S, Rosenblum M, Struzik Z and Stanley H 1999 Nature 399 461
[52] Wang Y, Wu C, Ji Z, Wang B, Liang Y 2011 PLoS ONE 6(5) e20060