Eur. Phys. J. B **50**, 255–258 (2006)
DOI: 10.1140/epjb/e2006-00103-4

THE EUROPEAN PHYSICAL JOURNAL B

Statistical indicators of collective behavior and functional clusters in gene networks of yeast

J. Živković[1], B. Tadić[2], N. Wick[1,3], and S. Thurner[1,a]

[1] Complex Systems Research Group HNO, Medical University of Vienna, Währinger Gürtel, 18-20, 1090 Vienna, Austria
[2] Department for Theoretical Physics, Jožef Stefan Institute, PO Box 3000, 1001 Ljubljana, Slovenia
[3] Clinical Institute for Pathology AKH Vienna, Währinger Gürtel, 18-20, 1090 Vienna, Austria

Received 30 September 2005 / Received in final form 18 November 2005
Published online 12 April 2006 –

Abstract. We analyze gene expression time-series data of yeast (*S. cerevisiae*) measured along two full cell-cycles. We quantify these data by using q-exponentials, gene expression ranking and a temporal mean-variance analysis. We construct gene interaction networks based on correlation coefficients and study the formation of the corresponding giant components and minimum spanning trees. By coloring genes according to their cell function we find functional clusters in the correlation networks and functional branches in the associated trees. Our results suggest that a percolation point of functional clusters can be identified on these gene expression correlation networks.

PACS. 87.10.+e General theory and mathematical aspects – 89.75.-k Complex systems – 89.75.Hc Networks and genealogical trees

1 Introduction

Gene regulatory networks describe the effective interactions between genes. The activity of a gene, i.e., its current rate of being transcribed into RNA molecules, can have effects on the activity levels of other genes, which will as a result become up- or down- regulated. The sum of all up- and down- regulation relations in the whole genome is the gene regulatory network. The complete knowledge of the gene network would reveal a large portion of an understanding of life. However, this goal is far from being achieved. With present DNA-chip technology it is possible to measure the transcription rates at a given point in time of an entire genome, but even these technologies only allow a glimpse on the structure of the underlying network, due to the underdeterminedness of the problem [1]. This situation got the physics community interested, to statistically characterize the available data and to (crudely) estimate the structure of the complex networks governing gene dynamics. A step toward an identification of potential gene interaction networks is to identify and quantify meaningful *statistical* indicators of gene cooperative behavior, which is the main purpose of the present work. The idea is that fluctuations of gene expressions over time, e.g., during a cell-cycle, can be considered as an output of an interacting gene collective forming a structured network. The hope is that a network structure estimate can be inferred from statistical properties. At least it should be possible to statistically characterize the types of potential candidate networks.

We consider the time-course expression data $x_i(t)$ for the genome of yeast *S. cerevisiae* [2]. We determine some statistical indicators of collective dynamical behavior of genes, such as the q-exponential fit of the cumulative distribution, a ranking distribution and a mean-variance analysis of differential gene expressions. We construct and estimate the expression-correlation network from time increments of expression data and analyse clusters and spanning trees. We identify biological functions of genes with use of a yeast database [3]. We find that the resulting, correlation based clusters match considerably well with specific biological functions of genes in the cell.

2 Scale-invariance in gene expression levels

The genome-wide gene expression data in [2] are given in the form of a matrix $x_i(t)$ in which every row represents one of $N = 6406$ yeast genes and each column contains the time evolution of gene expression of that gene i. Gene expressions are measured at 17 time points, taken every 10 min which covers approximately two full cell-cycles. We first properly normalize the gene expressions for each of the 17 measurements separately by dividing each gene expression value by the average value of gene expression for

[a] e-mail: thurner@univie.ac.at

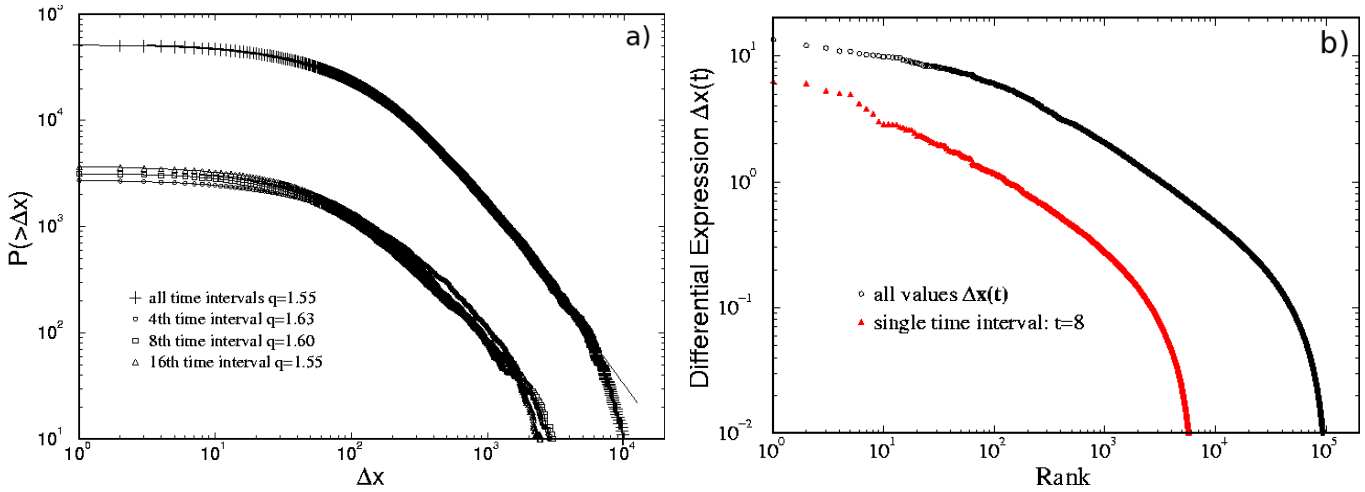

Fig. 1. Histograms (a) and ranking of differential gene expressions (b).

that corresponding column. In order to avoid systematic trends in the time series we use *differential* expression data defined as $\Delta x_i(t) = x_i(t) - x_i(t-1)$ for each gene i. We determine the cumulative distribution $P(>\Delta x)$ for each time-interval separately and also all measurements (all entries in matrix). The results are given in Figure 1a. This distribution can be fitted to a q-exponential form [4],

$$P(\Delta x) = B_q \left[1 - (1-q)\frac{\mid \Delta x \mid}{\mid \Delta x_0 \mid}\right]^{\frac{1}{1-q}}; \quad q \neq 1, \qquad (1)$$

where q represents the non-extensivity parameter. The fitted values of q for the various time-intervals are in the range 1.52−1.63. The average over all times yields $q = 1.55$, potentially indicating a non-trivial collective behavior of genes along the cell-cycle. In Figure 1b the ranking distribution is shown for genes according to their differential expressions at a particular time (lower curve) and for all measurements (upper curve). In both cases these curves exhibit approximate power-law regions, i.e. Zipf's law [5]. The occurrence of Zipf's law has been found in the ranking of expression data of many other species [6]. The results in Figure 1b indicate that the characteristic form of the distribution, in particular its slope, more or less persists even when ranking is averaged over all-time measurements.

Gene expression levels fluctuate during a cell-cycle. We calculate its temporal mean, $\langle x_i \rangle_t = \frac{1}{17}\sum_t x_i(t)$ and its variance $\sigma_i = \sqrt{\langle {x_i}^2 \rangle_t - \langle x_i \rangle^2{}_t}$, for all genes $i = 1, \cdots N$. In many dynamical systems a relation between those quantities is found to be of the form

$$\sigma_i \sim \langle x_i \rangle^{\mu}. \qquad (2)$$

In the case of driven dynamical systems on networks the scaling relation equation (2) holds when the values of μ depend on both, the network topology and the driving conditions. In particular, many real networks seem to fall into two "universality classes" [7]: $\mu = 1$, for example for

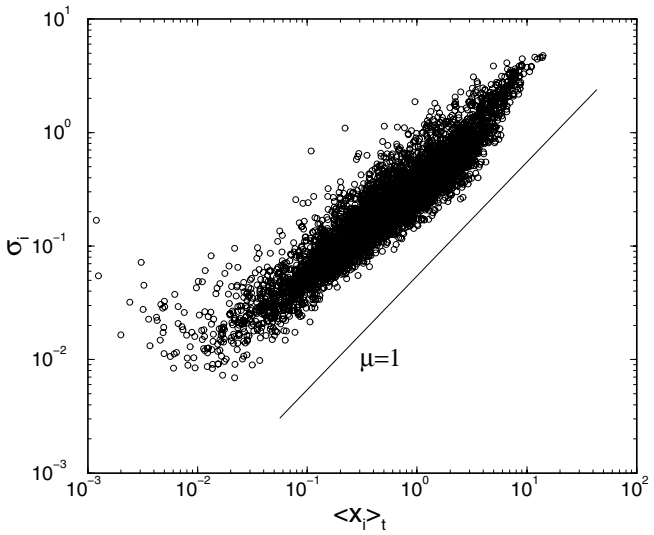

Fig. 2. Time fluctuation σ_i plotted against time-averaged gene expression $\langle x_i \rangle_t$, for all N genes.

scale free tree graphs and cyclic structures, and $\mu = 1/2$, often found in weakly driven cyclic graphs. In Figure 2 the temporal variance σ_i of the expression level $x_i(t)$ is plotted against its temporal mean $\langle x_i \rangle_t$ for each gene. The data yields a slope of $\mu \sim 0.89$ which suggests a heterogeneous network of genes with highly driven dynamics.

3 Gene expression networks

3.1 Construction of gene networks

Measures of the statistical indicators, do not identify the network topology, however, they suggest that some collective phenomena seem to occur, which could be thought of grouped up- or down- regulations within "clusters" of genes.

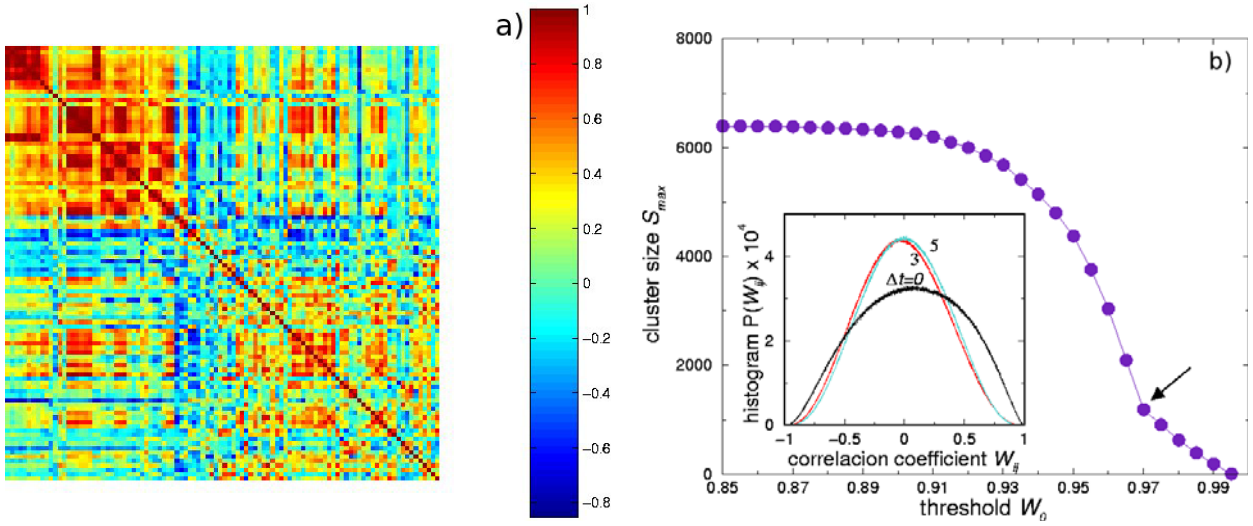

Fig. 3. Section of the correlation matrix W_{ij} (a). Size of the giant cluster S_{max} as a function of the threshold W_0 for equal-time correlations, $\Delta t = 0$, (b). Inset: histogram of W_{ij} for different time lags, $\Delta t = 0, 3$ and 5.

As a first attempt we construct a "gene expression network" from correlation coefficients of temporal differential gene expressions $\Delta x_i(t)$,

$$W_{ij}(\Delta t) = \frac{\sum_t (\Delta x_i(t) - \langle \Delta x_i \rangle)(\Delta x_j(t+\Delta t) - \langle \Delta x_j \rangle)}{\sigma_i \sigma_j}. \tag{3}$$

A section of this correlation matrix W_{ij} is shown in Figure 3a. The histogram of the correlation coefficients W_{ij} for all pairs of genes are shown in the inset to Figure 3b for several time lags, $\Delta t = 0, 3$ and 5. These distributions of correlation coefficients clearly exhibit a non-Gaussian character. To define a network we select a threshold W_0. A link is defined to exist between genes i and j if their correlation exceeds the threshold, $W_{ij} > W_0$. By systematically decreasing W_0 we observe the formation of a giant component, whose size S_{max} is plotted against the threshold W_0 in Figure 3b. The conditions for the formation of the giant cluster [8] $\langle k^2 \rangle - 2\langle k \rangle > \sum_k k(k-2)P(k) > 0$ are fulfilled at rather large values of the threshold. The size of the giant cluster increases first linearly by decreasing W_0, until an inflection point is reached at $W_0 \approx 0.97$ (arrow in Fig. 3b). The steep increase below this point resembles a percolation-like behavior in which the network gradually becomes complete in the range $0.95 \lesssim W_0 \lesssim 0.97$.

3.2 Clusters and trees

A particular way to statistically characterize the network topology is to study different types of connected clusters supported by that network. In Figures 4a and 4c we show all clusters remaining at thresholds of $W_0 = 0.93$ and $W_0 = 0.90$, respectively. A minimum cluster size of 10 nodes was chosen. Individual genes are nodes, colored according to their cell function [3]; the color map is described in the caption of Figure 4. In Figures 4b and 4d the *minimum spanning trees*, which are constructed from the "distance" $d_{ij} \equiv \sqrt{2(1 - W_{ij})}$ are shown (see e.g. [9]). The maximum spanning trees computed from W_{ij} directly lead to very similar trees (not shown), indicating that most of the dynamics is driven by positive correlations. For the threshold values in the range $0.9 \lesssim W_0 \lesssim 0.97$, apart from the giant cluster a number of smaller clusters is present. By color-coding according to the biological functions in the cell [10], (of which a large fraction is known for yeast [3]), grouping of genes into clusters occurs, suggesting that the gene expression correlations in equation (3) captures functionally similar genes. By varying the threshold W_0 in the range between $0.9 \lesssim W_0 \lesssim 0.97$, we detect the appearance of a color-grouping shortly below the inflection point $W_0 \approx 0.97$. Color-groupings then increases with lowering the threshold. For comparison, in Figures 4c and 4d we show the situation at $W_0 = 0.9$, where, apart from very small clusters which are removed from the figure, many new genes joined the giant component. Its minimum-distance spanning tree is also shown in Figure 4d. Genes with certain functions, in particular the "protein synthesis" and "cellcycle" function seem to appear in rather cohesive subgroups of the network. Genes with predominant "metabolism" functions, appear more dispersed over different branches. All these plots are obtained for $\Delta t = 0$. For $\Delta t > 0$ the functional clusters and branches remain for a while before they gradually disappear in the noise. This is in agreement with the observed character of the correlation distributions in Figure 2, where smaller deviations from a Gaussian distribution are found for $\Delta t > 0$.

4 Conclusions

In conclusion, we made several observations about the statistical nature of gene expression data which seem to

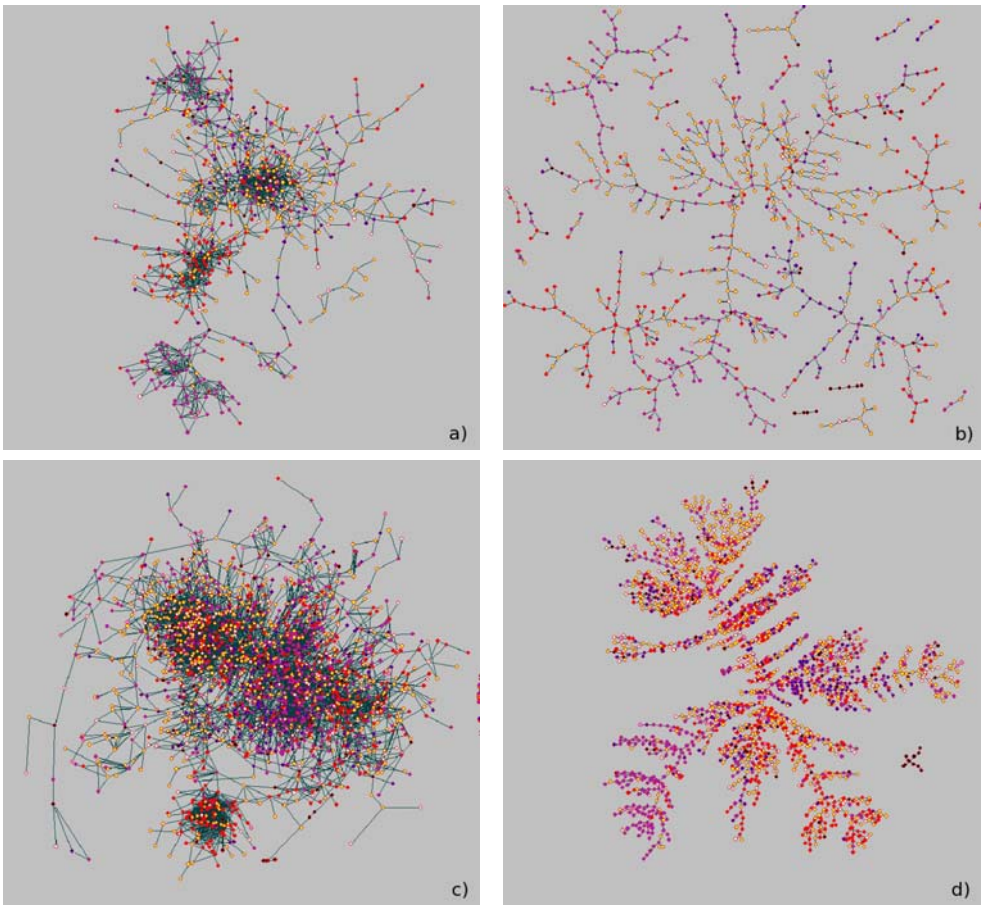

Fig. 4. Gene expression networks ((a) and (c)) for $W_0 = 0.93$ (top) with a minimum cluster size of 10 nodes and $W_0 = 0.90$ (bottom). Minimum spanning trees ((b) and (d)) for the distance measure d_{ij} and same values of W_0. Genes are colored according to their functions they fulfill in the cell [3]: yellow–metabolism; pink–energy household ; red–cellcycle; blue–transcription; purple–protein synthesis white–cellular transport/rescue; black–celltype/development; green–unknown.

suggest that at least a significant fraction of genes is up/down regulated in a highly collective manner. Indicators pointing in this direction are: (*i*) the cumulative distribution of differential gene expressions can be fitted to q-exponentials, with a non-trivial $q \sim 1.55$; (*ii*) an approximate Zipf's law holds in the ordering distribution of differential expressions; (*iii*) an almost linear mean variance dependence with $\mu = 0.89$ signals tightly driven dynamics; (*iv*) the correlation matrix element distributions are non-Gaussian and non-Poisson and finally, (*v*) even crude correlation coefficient networks display the emergence of clusters and functional branches in minimum spanning trees, which seem to be biologically relevant.

B.T. acknowledges support by the Project No. P1-0044, S.T. by the WWTF Life Science grant LS139. We acknowledge partial support by the bilateral project SI-AT/01/04-05.

References

1. S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, G.M. Church, Nature Genet. **22**, 281 (1999); W. Wang, J.M. Cherry, D. Botstein, H. Li, Proc. Natl. Acad. Sci. **99**, 16893 (2002); K. Rho, H. Jeong, B. Kahng, e-print `arXiv:cond-mat/0301110`; F. Li, T. Long, Y. Lu, Q. Ouyang, C. Tang, Proc. Natl. Acad. Sci. **101**, 4781 (2004); D. Balcan, A. Erzan, Eur. Phys. J. B **38**, 253 (2004)
2. R.J. Cho, Campbell, et al., Molecular Cell. **2** 65 (1998); `http://arep.med.harvard.edu/cgi-bin/ExpressDByeast`
3. `http:// mips.gsf.de`
4. C. Tsallis, J. Stat. Phys. **52**, 479 (1988)
5. G.K. Zipf, *Psycho-Biology of Languages* (Houghton-Mifflin, 1935; MIT Press, 1965)
6. C. Furusawa, K. Kaneko, Phys. Rev. Lett. **90**, 088102 (2003)
7. M. Argollo de Menezes, A.-L. Barabasi, Phys. Rev. Lett. **92**, 028701 (2004)
8. S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of Networks* (Oxford Univ. Press, 2003)
9. G. Bonanno, G. Caldarelli, F. Lillo, R.N. Mantegna, e-print `arXiv:cond-mat/0211546` (2002)
10. In the available data-bases of yeast, apart from genes with unknown functions, often multiple functions can be assigned to a single gene. To keep one color of each gene we select the first listed function in the database. We also have grouped several fuctions related to protein synthesis to a single one in order to keep the number of colors visually distinguishable