Topology of mammalian transcription networks

Anatolij P. Potapov[1], **Nico Voss**[2],
apo@bioinf.med.uni-goettingen.de nvo@biobase.de

Nicole Sasse[1] **Edgar Wingender**[1,2]
nsa@bioinf.med.uni-goettingen.de ewi@bioinf.med.uni.goettingen.de

[1] Department of Bioinformatics, UKG, University of Göttingen, Goldschmidtstr. 1, D-37077 Göttingen, Germany
[2] BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany.

Abstract

We present a first attempt to evaluate the generic topological principles underlying the mammalian transcriptional regulatory networks. Transcription networks, *TN*, studied here are represented as graphs where vertices are genes coding for transcription factors and edges are causal links between the genes, each edge combining both gene expression and *trans*-regulation events. Two transcription networks were retrieved from the TRANSPATH® database: The first one, *TN_RN*, is a 'complete' transcription network referred to as a reference network. The second one, *TN_p53*, displays a particular transcriptional sub-network centered at *p53* gene. We found these networks to be fundamentally non-random and inhomogeneous. Their topology follows a power-law degree distribution and is best described by the scale-free model. Shortest-path-length distribution and the average clustering coefficient indicate a small-world feature of these networks. The networks show the dependence of the clustering coefficient on the degree of a vertex, thereby indicating the presence of hierarchical modularity. Clear positive correlation between the values of betweenness and the degree of vertices has been observed in both networks. The top list of genes displaying high degree and high betweennes, such as *p53, c-fos, c-jun* and *c-myc,* is enriched with genes that are known as having tumor-suppressor or proto-oncogene properties, which supports the biological significance of the identified key topological elements.

Keywords: regulatory network, gene network, transcription, gene expression, topology

1 Introduction

The way how an extracellular signal is transduced inside a cell to reach its target within a specific transcriptional regulation is one of the most fundamental biological control mechanisms since it is the main step that governs the transformation of one-dimensional codification of when, where and under which conditions a certain gene is expressed into cellular reality. This is achieved by transcription factors which recognize specific DNA sequence elements in promoters, enhancers and other regulatory regions of the genome. The genes encoding transcription factors (TFs) are themselves subject to control by other and the same TFs. From this, a transcriptional network emerges which is responsible for controlling essential biological processes such as morphogenesis, cell proliferation and differentiation, homeostasis and responses to environmental stimuli.

From the very beginning, it is not self-evident what the topological properties of the transcriptional network are, i.e. whether it is modularly and hierarchically organized, whether it has scale-free and small world properties, etc. These properties have been investigated for several other biological networks, such as metabolic and protein-protein interaction networks [1,14,22] as well as for gene expression networks [6,11]. Transcription networks, which by their nature are expected to exert some particularities, have been investigated so far only for *E. coli* [16] and *S. cerevisiae* [8,11]. However, the tasks a transcriptional network has to fulfill in higher eukaryotes, in particular governing the coordinated tuning of gene expression between trillions of specialized cells, imposes much greater demands on the whole system than encountered in a simple unicellular organism. Therefore, and having the databases TRANSPATH® and TRANSFAC® at hand, we decided to investigate the organization of transcription networks in higher eukaryotes.

This work presents a first attempt to evaluate the generic topological principles underlying the mammalian transcriptional regulatory networks. The networks studied here do not refer to those transcription networks which deal with protein-protein interactions between transcription factors. Instead, they focus on the causal relationship between genes coding for transcription factors.

2 Methods

2.1 Data sets and networks' retrieval

The whole analysis described here is based on the contents of the TRANSPATH® Professional database release 4.4 [3,10], in which the transcription factor-target gene information was imported from the TRANSFAC® database [12]. Since TRANSPATH® focuses on signaling pathways involved in the regulation of transcription factors and gene expression in vertebrates, working with this database opens the possibility of extending our analysis to the upstream signaling pathways at a later stage. Moreover, it facilitated to perform our analyses at the level of "ortholog abstraction", at which all species-specific data that refer to mammalian genes/gene products have been summarized to corresponding generic entries. A public version of TRANSPATH® is freely available to users from non-profit organizations [3,10,26].

To extract transcriptional regulatory networks, the whole network represented in the TRANSPATH® database was transformed into a unipartite graph at ortholog abstraction level. Then, the network was reduced by removing molecules and genes that have no relation to transcription factors. Then, paths of the kind 'gene->molecule->gene' were displayed as 'gene->gene' edges.

To retrieve the set of compounds related to *p53*, a shortest-path graph algorithm was applied to both upstream and downstream searches. Shortest paths were calculated by means of breadth-first search implemented in a C^{++} program.

2.2 Graph representation and analysis

Transcriptional regulatory networks are represented here as directed graphs, the vertices of which denote genes coding for transcription factors and the edges denote combined expression and *trans*-activation actions. The directedness of edges corresponds to the causal relation of the regulatory events within regulatory networks. The in-degree, k_{in}, is the number of incoming edges, the out-degree, k_{out}, is the number of outgoing edges, and the inout-degree, k_{inout}, is the number of incoming and outgoing edges of a vertex.

Degree distribution, $P(k)$, gives the probability that a vertex in the network is connected to k other vertices. Thus, $P(k_{in})$, $P(k_{out})$, and $P(k_{inout})$ denote the distribution of the in-, out-, and inout-degree, respectively.

The clustering coefficient of a vertex is defined as $c_i = 2n/k_i(k_i\,\text{-}1)$, where n is the number of edges between the k_i nearest neighbors of vertex i, disregarding the directedness of its edges [21]. The clustering coefficient of the network C is obtained by averaging c_i over all vertices of a network. The clustering coefficient of a classical random graph, C_{random}, is calculated as $C_{random} = 2E/(V(V\text{-}1))$, where V and E are the numbers of vertices and edges, respectively [4]. $C(k)$ gives the average clustering coefficient of all vertices with k links.

Betweenness centrality of vertex j is the fraction of those directed shortest paths between all pairs of vertices that pass through vertex j [7]. It is computed as

$$BC = \sum_{s \neq j \neq t} \frac{\delta_{st}(j)}{\delta_{st}},$$

where δ_{st} is the total number of shortest paths between nodes s and t, and $\delta_{st}(j)$ those of them that pass through node j.

Working with the relational database, we calculated the above mentioned parameters by means of Pajek software [24] and an SQL program. Shortest paths were calculated as described above (2.1). The length of the mean shortest path was found as the average over the shortest paths between all connected pairs of vertices.

3 Results and Discussion

3.1 Scale-free properties of transcription networks

Two transcription networks were retrieved from the TRANSPATH® database: The first one, *TN_RN*, is a ‘complete’ transcription network referred to as a reference network. It consists of 121 vertices and 212 edges. The second one, *TN_p53*, comprises a particular transcriptional sub-network centered at the *p53* gene and consists of 44 vertices and 80 edges.

Being displayed as directed graphs, these regulatory networks are characterized by relatively small shortest path length values. The average shortest path length, that represents the statistical diameter of a network, is found to be 2.2 and 2.4 in *TN_RN* and *TN_p53*, respectively. The maximal shortest path length, that represents the actual diameter of a network, is 5 for both networks. That does not exclude the presence of longer paths, which were not in the focus of our analysis. Such small shortest paths signify the possibility of faster propagation of the regulatory communication between TF-genes.

The degree distribution, $P(k)$, is the primary characteristic of the architecture of a network. It gives the probability that a randomly selected vertex has exactly k links to its nearest neighbors. The topology of both networks has been found to be far from following a bell-like Poisson distribution that would be expected for a random and statistically homogeneous model. Instead, it appears to follow a power-law distribution and is best described by the scale-free model. In the double-logarithmic plots, the distribution could be well approximated by straight lines (Fig.1; Tab. 1) that fit the power-law $P(k) \sim k^{-\gamma}$, which is a signature of a scale-free network. This means that most TF-genes are connected sparsely, while few of them (hubs) are connected with many others and play an important role in sustaining the integrity of the transcription networks. This kind of distribution was found when analyzing either incoming signaling (k_{in}), outgoing signaling (k_{out}), or their combination (k_{inout}). An example of inout-degree distribution is presented in Fig. 1.

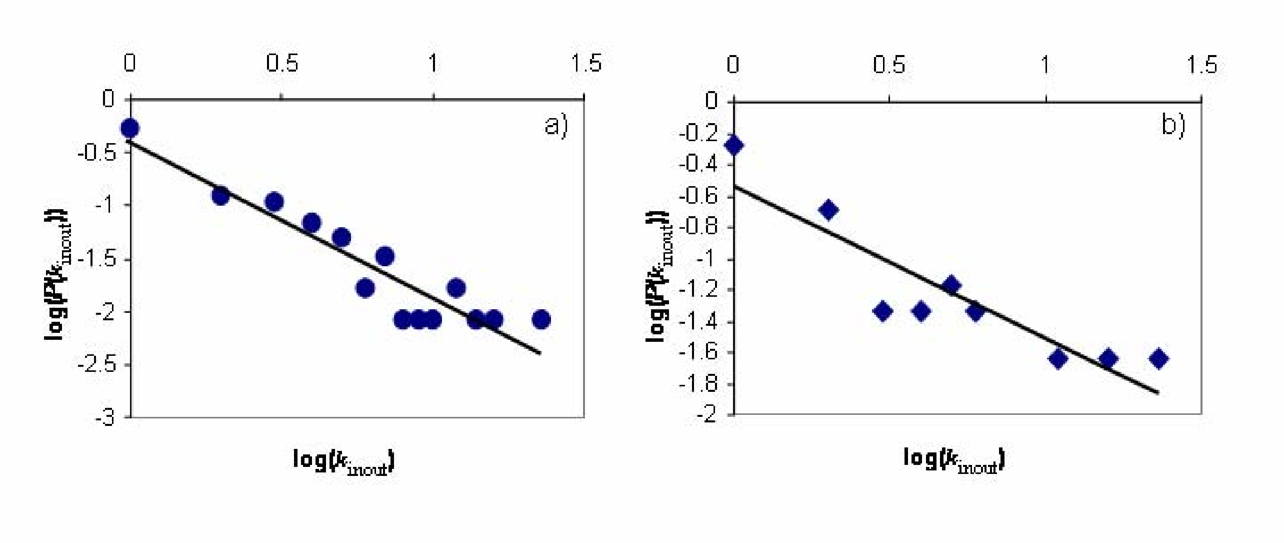

Figure 1: The degree distribution, $P(k)$, in *TN_RN* (a) and *TN_p53* (b) networks shows their scale-free topology. An example of incoming and outgoing degrees (kinout) is presented. The straight lines are made by linear regression; log means log_{10}. The largest TF-gene hubs are in contact with a large fraction of all TF-genes.

Table 1: Summary of the topological properties of *TN_RN* and *TN_p53* transcription networks. γ, ω: slopes of straight lines in double logarithmic plots calculated by linear regression. Standard deviation and Pearson correlation coefficient, *R*, are indicated when necessary.

Parameter	*TN_RN*	*TN_p53*
Degree distribution		
γ-in	1.34 ± 0.28 (R^2 = 0.78)	0.56 ± 0.26 (R^2 = 0.51)
γ-out	1.78 ± 0.14 (R^2 = 0.95)	1.50 ± 0.16 (R^2 = 0.92)
γ-inout	1.46 ± 0.21 (R^2 = 0.87)	0.97 ± 0.21 (R^2 = 0.81)
Average clustering coefficient		
C	0.134	0.241
C/C_{random}	4.6	2.8
Clustering coefficient distribution		
ω	0.66 ± 0.25 (R^2 = 0.40)	1.1 ± 0.24 (R^2 = 0.77)

3.2 Modularity of transcription networks

The clustering coefficient of a vertex shows the level of interconnectivity between the direct neighbours of the vertex. The results of this analysis are presented in part in Tab. 2. High values of the clustering coefficient of TF-genes, such as *p53*, *Egr1*, c-*jun*, c-*myc*, and *HNF4A*, indicate greater local communication between each of them and their nearest neighbouring TF-genes.

The mean clustering coefficients of all vertices were found as 0.134 for *TN_RN* and 0.241 for *TN_p53* (Tab. 1). These values are larger than those expected for classical random graphs with the same number of vertices and edges (C/C_{random}; cf. Tab. 1), thus demonstrating the increased clustering feature of the transcription networks. Together with the rather short paths found (see above), that indicates the small-world architecture of the networks analyzed.

Modular organization is considered to be a hallmark of biological regulatory systems with each module performing its special functional task, separable from the functions of other modules [9,15]. The modularity of networks can be expressed and displayed in terms of their topology and the scale-free topology can be reconciled with that within the framework of a hierarchical modularity [1,14]. The hierarchical modularity does not automatically arise from the scale-free topology and it implies that sparsely connected vertices are part of highly clustered areas, with communication between the different highly clustered neighborhoods being maintained by a few hubs. The hierarchical topology does not relate to the coexistence of relatively independent groups of vertices; instead, it relates to many small clusters that are densely interconnected [13]. A high clustering coefficient is known to be suggestive of modular organization [1,14]. Accordingly, the increased clustering feature of *TN_RN* and *TN_p53* networks can be considered as indicating the possibility that these transcription networks might have a hierarchical and modular architecture.

However, the most important signature of hierarchical modularity is the scaling of the clustering coefficient, which follows $C(k) \sim k^{-1}$ a straight line of a slope -1 on log-log plot [1,5,14]. As follows from our evaluations, $C(k)$ depends on k in both networks analyzed and this dependence can be well approximated with a power low $C(k) \sim k^{-\omega}$ (Fig. 2). Moreover, the scaling exponent ω reasonably approaches the value one (Tab. 1). Thus, the topology of both *TN_RN* and *TN_p53* transcription networks tends to have a hierarchical modular structure. Such a hierarchical and modular organization might enable these networks to operate not only on separate TF-genes but on ready made and tuned groups of such genes as well, thus increasing the efficiency of transcriptional regulation in response to the needs of the cell.

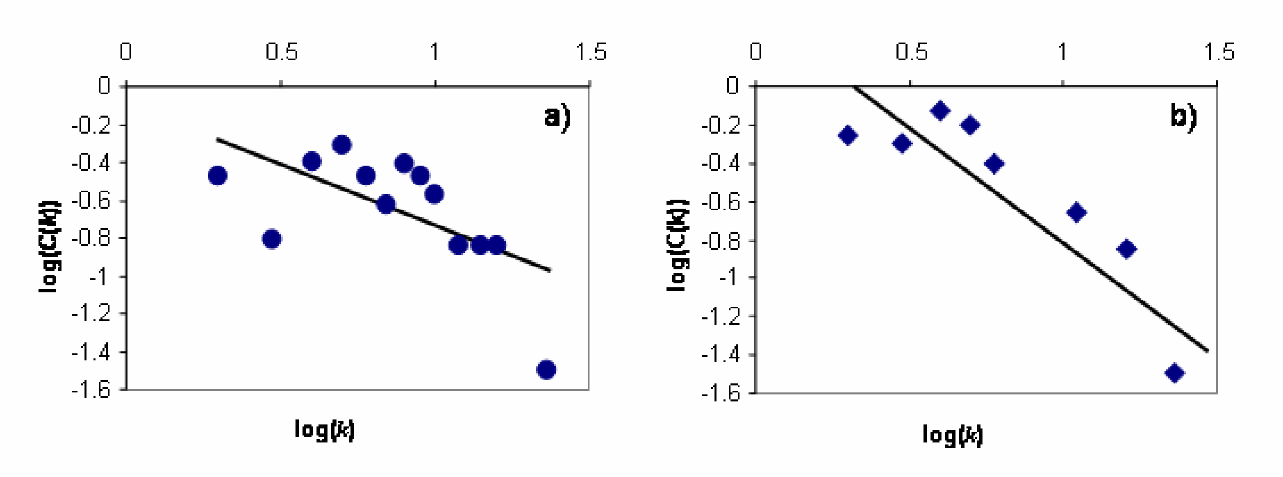

Figure 2: Scaling of the clustering coefficient in *TN_RN* (a) and *TN_p53* (b) networks displays the presence of hierarchical modularity in them. Each value *C*(*k*) gives the average clustering coefficient of all vertices with *k* links. The straight lines are computed by linear regression; log means log_{10}.

The betweenness centrality, *BC*, is a measure of the intermediary role of each individual element in the communication between all other elements: *BC* of vertex *j* is the fraction of those shortest paths between all pairs of vertices that pass through vertex *j* [7]. Thus, it allows quantify how influential a given TF-gene in a whole network is. A clear positive correlation between the *BC* values and the degree (in-, out- or inout-degrees) of vertices has been observed in both *TN_RN* and *TN_p53* networks. An example of such an in-degree analysis is presented in Fig. 3. This observation is of special importance as it proves the interrelation between very local (degree) and rather integral (*BC*) topologies of the transcription networks. It strongly supports the view that hubs (i.e., elements with highly enriched local topology) represent the most influential elements of a network [2] and tend to be essential for sustaining the integrity of a network [1,23].

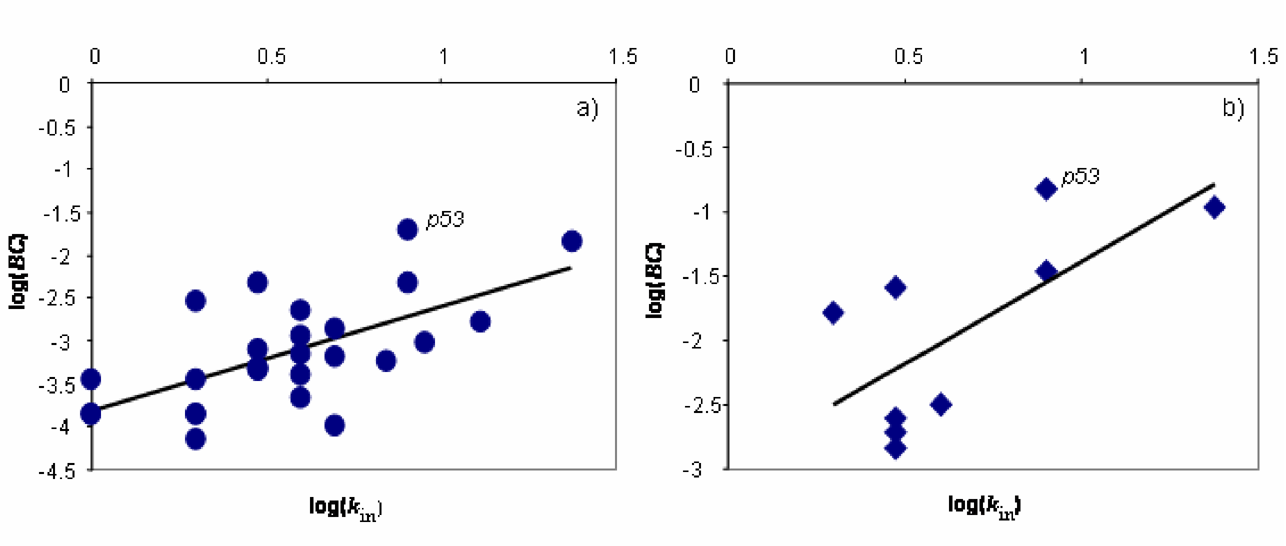

Figure 3: Positive correlation between the values of betweenness centrality, *BC*, and the in-degree of vertices in *TN_RN* (a) and *TN_p53* (b) networks. The position of *p53* gene is indicated. The straight lines are computed by linear regression; log means log_{10}.

Table 2: Topological features of some individual TF-genes in the reference network. The top part of the complete list is presented. Here, *k* – degree, *c* - the clustering coefficient, *BC* – betweenness centrality of a given TF-gene.

Name	k_{in}	k_{out}	k_{inout}	***c***	***BC***
p53	8	10	18	0.142	0.0188
c-fos	24	3	27	0.032	0.0139
Egr1	3	6	9	0.333	0.0047
c-jun	8	5	13	0.197	0.0046
WT1	2	1	3	0	0.0029
SRF	4	8	12	0.333	0.0022
c-myc	13	3	16	0.143	0.0016
HNF4A	5	5	10	0.333	0.0014
HOXA1	4	1	5	0.5	0.0011
RAR-beta	9	4	13	0.393	0.0009
ONECUT1	3	2	5	0	0.0007
HOXB1	4	2	6	0.5	0.0007
Egr2	5	1	6	0.1	0.0006
IRF-1	7	1	8	0	0.0005
TCF1	3	6	9	0.5	0.0004
ELK1	3	4	7	0.7	0.0004
CRE-BP1	3	5	8	0.333	0.0004
JUND	4	1	5	0.4	0.0003
STAT3	2	4	6	0.5	0.0003
RARA	1	6	7	0.5	0.0003
C/EBPalpha	4	2	6	0.4	0.0002

3.3 Identification of TF-genes of high impact

Tab. 2 displays several topological characteristics of some individual TF-genes in *TN_RN* which are arranged according their *BC* values. The high betweenness centrality of *p53* (0.0188) indicates that this TF-gene is an important intermediary in this regulatory network. The top ranking TF-genes, such as *p53*, c-*fos*, c-*jun*, *SRF*, and c-*myc*, are attributed with relatively high values of in-, out- and inout-degrees, and as a rule with high clustering coefficient. It is interesting to note that many of them are known to be involved in regulation of cell proliferation and have features of tumor-suppressors or proto-oncogenes [17,18].

However, some exception from the above mentioned correlation can be found as well (Tab. 2). For instance, although *WT1* is characterized by the *BC* value about 0.0029, the in-, out-, and inout-degrees of this gene are rather small and the clustering coefficient is even zero. Such a TF-gene might be an internal part of a linear path that connects two modules each consisting of several TF-genes. In any case, the relatively high *BC* value measured by us for *WT1* is consistent with its known biological activity as a tumor-suppressor gene [18].

Among several topological characteristics of the transcriptional networks studied here, the betweenness centrality, *BC*, appears to be the most representative in regard to the biological significance of distinct elements.

3.4 What we know and what we need

The transcription networks studied here are a kind of abstraction: They are master networks that focus on the common features of transcriptional regulation in different vertebrates and emphasize their evolutionary constant constructions. The *TN_RN* and *TN_p53* networks provide a genome-wide view above the level of species (ortholog abstraction) and may serve as a framework for further analyses in more depth.

It must be taken into account that the analysis presented here was done on the basis of the existing experimental evidence as it is represented in the databases TRANSPATH® and TRANSFAC®. Rough estimations show that we may have knowledge about approximately 1% of all existing transcription factor binding sites in the human genome (with similar figures for mouse and rat genomes). However, the analysis done here does not depend on knowledge about precise binding sites. The information whether a gene is regulated by a certain transcription factor may be sufficient. This kind of data is increasingly provided by high-throughput data about TF-target genes relations, e.g., from ChIP-on-chip experiments which may render our transcription network graph much more extended and interlinked. The recent inclusion of this kind of data into the TRANSFAC® database may help in this regard [25]. The same may be achieved by applying target gene predictions which, however, implies more reliable predictions of TF binding sites than most tools presently provide. Combinatorial analysis of TF binding sites and derivation of predictive models may improve this situation in near future.

4 Conclusions

A significant part of regulatory events in a living cell is mediated through transcription of different genes. TF-genes form a specific genetic sub-network which is a very basic part, i.e., a skeleton, of intracellular regulation. In this work, we could show that mammalian transcription networks are fundamentally non-random and inhomogeneous and exhibit scale-free, hierarchical and modular properties. The coexistence of error-tolerant scale-free and hierarchical topology, found recently in some biological networks [13,19,20,22] appears to be an emerging feature of transcriptional regulatory processes in mammalia.

Acknowledgements

This work has been supported by grant 503568 (COMBIO) within the 6th Framework Programme for Research, Technological Development and Demonstration of the European Commission.

References

[1] Albert, R. Jeong, H. and Barabási, A.-L.: Error and attack tolerance of complex networks, *Nature*, 406(6794):378-382, 2000.

[2] Barabási, A.-L., and Oltvai, Z.N., Network biology: understanding the cell's functional organization, *Nat. Rev. Genet.*, 5(2):101-113, 2004.

[3] Choi, C., Mathias Krull, M., Kel-Margoulis, O., Pistor, S., Potapov, A., Voss, N., and Wingender, E., TRANSPATH® – a high quality database on signal transduction, *Comparative Functional Genomics*, 5(2):163-168, 2004.

[4] Dorogovtsev, S.N. and Mendes, J.F.F., Evolution of networks, *Adv. Phys.*, 51(4):1079-1187, 2002.

[5] Dorogovtsev, S.N., Goltsev, A.V., and Mendes, J.F., Pseudofractal scale-free web, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 65(6 Pt 2):066122, 2002.

[6] Featherstone, D.E. and Broadie K., Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network, *Bioessays*, 24(3):267-274, 2002.

[7] Goh, K.I., Oh, E., Jeong, H., Kahng, B., and Kim, D., Classification of scale-free networks. *Proc. Natl. Acad. Sci. USA*, 99(20):12583-12588, 2002.

[8] Guelzim, N., Bottani, S., Bourgine, P., and Kepes F., Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.*, 31(1):60-63, 2002.

[9] Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W., From molecular to modular cell biology, *Nature*, 402:C47-C52, 1999.

[10] Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A., and Wingender, E., TRANSPATH®: an integrated database on signal transduction and a tool for array analysis, *Nucleic Acids Res.*, 31(1):97-100, 2003.

[11] Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M.Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308-312, 2004.

[12] Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele S., and Wingender, E., TRANSFAC® : transcriptional regulation, from patterns to profiles, *Nucleic Acids Res.*, 31(1):374-378, 2003.

[13] Ravasz, E. and Barabási, A.-L., Hierarchical organization in complex networks, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 67(2 Pt 2):026112, 2003.

[14] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., and Barabási, A.-L., Hierarchical organization of modularity in metabolic networks, *Science*, 297(5586):1551-1555, 2002.

[15] Spirin, V., and Mirny, L.A., Protein complexes and functional modules in molecular networks, *Proc. Natl. Acad. Sci. USA*, 100:12123-12128, 2003.

[16] Thieffry, D., Huerta, A.M., Perez-Rueda, E., and Collado-Vides, J., From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli, *Bioessays*, 20(5):433-440, 1998.

[17] Vogelstein, B., Lane, D., and Levine, A.J., Surfing the p53 network. *Nature*, 408(6810):307-310, 2000.

[18] Vogelstein, B. and Kinzler, K.W., Cancer genes anb the pathways they control. *Nat.Med.*, 10(8):789-799, 2004.

[19] Wagner, A., The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol.Biol.Evol.*, 18(7),1283-1292, 2001.

[20] Wagner, A. and Fell, d.A., the small world inside large metabolic networks. *Proc.Biol.Sci.*, 268(1478):1803-1810, 2001.

[21] Watts, D.J., and Strogatz, S.H., Collective dynamics of 'small-world' networks, *Nature*, 393(6684):440-442, 1998.

[22] Yook, S.H., Oltvai, Z.N., and Barabási, A.-L., Functional and topological characterization of protein interaction networks, *Proteomics*, 4(4):928-942, 2004.

[23] Yu, H., Greenbaum, D., Xin Lu H., Zhu, X. and Gerstein, M., Genomic analysis of essentiality within protein networks, *Trends Genet.*, 20(6):227-223, 2004.

[24] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[25] http://www.biobase.de/pages/products/statistics.html#transfac

[26] http://www.gene-regulation.com/pub/databases.html#transpath