COMPARATIVE ANALYSIS OF TOPOLOGICAL PATTERNS IN DIFFERENT MAMMALIAN NETWORKS

BJOERN GOEMANN[1]
bjoern.goemann@bioinf.med.uni-goettingen.de

ANATOLIJ P. POTAPOV[1]
apo@bioinf.med.uni-goettingen.de

MICHAEL ANTE[1]
michael.ante@bioinf.med.uni-goettingen.de

EDGAR WINGENDER[1,2]
ewi@bioinf.med.uni-goettingen.de

[1] *Department of Bioinformatics, University Medical Center Goettingen, Georg August University Goettingen, Goldschmidtstr. 1, D-37077 Goettingen, Germany*
[2] *BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbuettel, Germany*

We have systematically analyzed various topological patterns comprising 1, 2 or 3 nodes in the mammalian metabolic, signal transduction and transcription networks: These patterns were analyzed with regard to their frequency and statistical over-representation in each network, as well as to their topological significance for the coherence of the networks. The latter property was evaluated using the *pairwise disconnectivity index*, which we have recently introduced to quantify how critical network components are for the internal connectedness of a network. The 1-node pattern made up by a vertex with a self-loop has been found to exert particular properties in all three networks. In general, vertices with a self-loop tend to be topologically more important than other vertices. Moreover, self-loops have been found to be attached to most 2-node and 3-node patterns, thereby emphasizing a particular role of self-loop components in the architectural organization of the networks. For none of the networks, a positive correlation between the mean topological significance and the *Z*-score of a pattern could be observed. That is, in general, motifs are not *per se* more important for the overall network coherence than patterns that are not over-represented. All 2- and 3-node patterns that are over-represented and thus qualified as motifs in all three networks exhibit a loop structure. This intriguing observation can be viewed as an advantage of loop-like structures in building up the regulatory circuits of the whole cell. The transcription network has been found to differ from the other networks in that (*i*) self-loops play an even higher role, (*ii*) its binary loops are highly enriched with self-loops attached, and (*iii*) feed-back loops are not over-represented. Metabolic networks reveal some particular topological properties which may reflect the fact that metabolic paths are, to a large extent, reversible. Interestingly, some of the most important 3-node patterns of both the transcription and the signaling network can be concatenated to subnetworks comprising many genes that play a particular role in the regulation of cell proliferation.

Keywords: network topology, network motif analysis, transcription network, signaling network, metabolic network, pairwise disconnectivity index

1   Introduction

Triggered by the increasingly better defined paradigms of Systems Biology, biological systems are described most appropriately as the sum of processes they exert rather than comprehensive catalogs of objects they constitute of. It has been proven for the various biological processes that it is most feasible to represent them as networks or, more formally, as graphs, which comprise the participating components as nodes and their relations as edges. Once the architecture of a network, or its wiring diagram, is known, it

is possible to add quantitative information to the edges in order to proceed to dynamical simulations of the system's behavior under certain conditions. This is still hard to achieve for a complete system, as the holistic approach of systems biology would demand. However, it can be done for defined subsystems. In the most basic case, these are topological patterns which describe the connection between a few nodes. Amongst them, motifs are believed to be the simplest building blocks in biological networks. Consequently, they have been explored intensively in the last years and have been investigated for their behavior in diverse kinds of networks [1].

Different functions of a living cell can be represented by different networks such as the metabolic, the signaling or the gene regulatory network. The latter is mostly considered as a mere transcription network, but inclusion of post-transcriptional mechanisms as well becomes increasingly feasible [1]. Holistic modeling of a system would obviously require an integrated view of these different networks, but before that task can be tackled, we have to make ourselves aware about their particularities. It is therefore the goal of this contribution to characterize the three mentioned networks, reconstructed for mammalian cells, with regard to their topological patterns and the impact of these substructures for the whole respective networks.

For this purpose, we have proposed a new topological parameter, the *pairwise disconnectivity index* [3], which is useful in identifying network components that are most critical for the coherence of a network. The methodology may be applied on single nodes or edges, or whole subgraphs such as motifs, as evidenced for transcription networks [4].

Here, we apply this logic to topological patterns of different sizes on the mammalian metabolic, signal transduction and transcription network to investigate whether these networks differ significantly in content and impact of certain topological patterns such as self-edges and 3-vertex-patterns.

2 Methods

2.1 Construction of the Networks

The mammalian transcription network was retrieved from the TRANSFAC® database, release 11.3 [5], and the TRANSPATH® database, release 8.3 (BIOBASE, Wolfenbuettel, Germany) [6]. In this network, the nodes represent transcription factor (TF) genes, and the edges the genetic interactions between them, i.e. comprising expression of each gene and trans-activation/-repression of the target genes of its product. The TRANSPATH database was also used to reconstruct the signal transduction network for mammalian (mostly human, mouse and rat) cells; for this, we extracted "semantic" reactions only which focus on the essential components between which information is actively forwarded, and did so on the level of "orthogroups" [7]. Both networks therefore represent "reference networks", i.e. superpositions of all reactions and paths that have been identified in any mammalian species, in any tissue / cell type. The transcription

network includes 279 nodes and 658 edges, while the signaling network is made by 1571 nodes and 3425 edges.

The mammalian metabolic network was reconstructed from Ligand section of the KEGG database [8], comprising all genes encoding metabolic enzyme activity in mammalian (more precisely: human, mouse and rat) systems. To keep this network comparable with the other two, we chose a gene-centric view here as well, so that the nodes represent genes encoding metabolic enzymes, and the edge semantics is to forward a metabolite produced by one enzyme to one that consumes it. The metabolic network consists of 1793 nodes and 5538 edges.

2.2 Computation of the Pairwise Disconnectivity Index

In a directed graph $G(V, E)$ with V as the set of vertices and E as the set of edges, a pattern is the joint feature of every n connected vertices and describes the way how they are linked together. Such a pattern always comprehends all existing edges between n vertices. Furthermore, none of the n vertices is isolated from the others, i.e. each of the n vertices must be directly attached to at least another one.

The total set of distinct n-vertex patterns in G is given by $P^n = \{P_1^n, P_2^n, \cdots, P_i^n\}$. The uniqueness of the i-th pattern is due to its structure and the particular set of n-vertex subgraphs in G, $P_i^n = \{P_{i,1}^n, P_{i,2}^n, \cdots, P_{i,j}^n\}$, whose vertices are exactly connected to each other as described by the pattern. A subgraph $P_{i,j}^n$ is also denoted as the j-th *instance* of pattern P_i^n and there is no other subgraph in G that consists of the same set of vertices and edges than $P_{i,j}^n$ (Figure 1). Importantly, an edge e of a pattern instance $P_{i,j}^n$, $e \in E_{i,j}^n$ where $E_{i,j}^n \subseteq E$, is incident only to vertices of this instance and denoted as an *intrinsic* edge of the pattern $P_{i,j}^n$. Other edges in G, $e \in E \backslash E_{i,j}^n$, do not account for the coherence between the vertices of $P_{i,j}^n$ and are called *extrinsic* edges.

For the global context of a graph G we propose to evaluate the importance of pattern instances based on their participation on the existing connections in G as generally introduced in [3]. More precisely, we estimate how crucial such an instance is for sustaining the connection between the existing pairwise linked vertices in G by destroying the coherence within the instance. The latter is accomplished by removing all *intrinsic* edges of a pattern instance since they essentially reflect a particular pattern. The more pairwise connected vertices in G become disconnected due to the elimination of the *intrinsic* edges the higher is the importance, i.e. topological significance, of a pattern instance $P_{i,j}^n$. This influence is quantified by the *pairwise disconnectivity index* of a pattern instance [4] which is defined as

$$Dis(P_{i,j}^n) = 1 - \frac{N'}{N} \tag{1}$$

Equation 1 depicts the fraction of those initially connected ordered pairs of vertices in G which have become disconnected upon the removal of all intrinsic edges of pattern instance $P_{i,j}^n$. Hence, N is the number of ordered pairs that are linked by at least one path

in *G* and *N'* is the number of ordered pairs in *G'(V,E')* with $E' = E \backslash E_{i,j}^n$. *G'* is thus the subgraph of *G* that results from removing the intrinsic edges of the pattern instance $P_{i,j}^n$ from *G*.

The topological significance of a pattern P_i^n is determined by the topological impact of multiple representatives of this pattern on the connectivity in *G*. The respective influences of all of its instances need to be considered equally to avoid a misleading result which might occur by choosing a procedure as in [9]. This can be achieved by averaging over all instances of a pattern, i.e.

$$Dis(P_i^n) = \overline{Dis(P_{i,j}^n)} = \frac{1}{J}\sum_{j=1}^{J} Dis(P_{i,j}^n) \tag{2}$$

Similar to Eq. 1, the *pairwise disconnectivity index of a pattern* varies between 0 and 1 whereas $Dis(P_i^n) = 0$ means that none of its *J* instances is crucial for the connection between any pairwise linked vertices. Consequently, $Dis(P_i^n) = 1$ refers to the case where no pair is connected anymore.
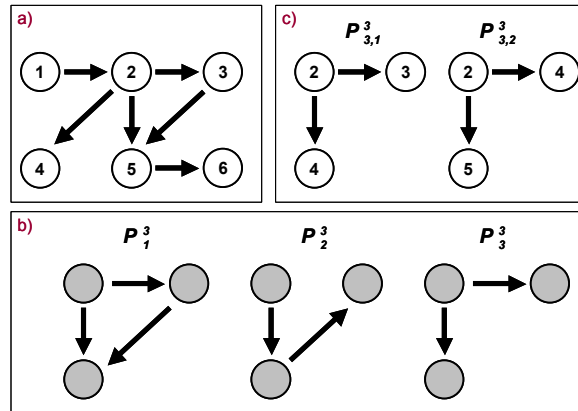

Figure 1: Topological patterns and their instances in a network. (a) A toy network with the set of vertices *V* = {1,...,6}; (b) the entirety of all 3-vertex patterns $P_3^3 = \{P_1^3, P_2^3, P_3^3\}$ in the toy network; (c) the two instances of the pattern P_3^3. Each instance is characterized by its individual sets of intrinsic and extrinsic edges. Thus, in regard to instance $P_{3,1}^3$, edges (2,3) and (2,4) are intrinsic and all other edges are extrinsic.

The topological significance of a pattern instance can be comprehended from the example of a feed-forward loop (FFL) included in Figure 1a. The respective FFL-instance is given by the edges 2 → 3, 2 → 5 and 3 → 5. Other edges, that may also start and/or end in the vertices 2,3,5 are the extrinsic edges of this feed-forward loop instance.

Whether this pattern instance may have any influence on the connection between pairwise linked vertices depends on how such a connection is built. If all directed paths between two vertices consist only of extrinsic edges, the pattern instance can hardly have an impact, i.e. the connection is independent from the FFL-instance. For example in

Figure 1a, this is the case for the pair {1, 4}. Note that similar applies even if an intrinsic edge is part of a path that links two vertices but still another path is present that does not contain any intrinsic edge.

Thus, all paths between a connected pair of vertices must contain at least one of the intrinsic edges so that the FFL-instance is critical for this connection. For the example in Figure 1a, this is the case for the pair {1, 6}, which depends on the feed-forward loop instance. The respective pairwise disconnectivity index is 0.67, i.e. the connection of eight of the twelve pairwise connected nodes critically depends on the FFL-instance.

3   Results and Discussion

3.1 *Autoregulation as a Feature of the Most Important Nodes*

The simplest topological pattern consists of one node and one self-edge, i. e. a gene or molecule is acting on itself: In the transcription network, such self-looping is provided by a TF-gene, the product of which binds to its own promoter; in the signaling network, a signaling molecule may "autocatalytically" activate itself (in most cases, the subunits of a homomeric complex cross-activate each other); in the gene-centric view of the metabolic network, a self-loop usually represents a reversible reaction (an enzyme consumes its own product).

For self-loops, we focus on the properties of the respective nodes since a self-edge cannot have a topological impact on the network coherence. Therefore, we compare the nodes with and without self-loops in the three networks based on their frequencies, inout-degrees, betweenness centrality and the pairwise disconnectivity index applied on vertices. Particularly the latter two metrics are useful for estimating the impact onto a whole network, but depict different properties and maybe complementarily used in topological analyses [3]. In contrast to betweenness centrality that considers the total number of shortest paths going through the given vertex [10][11], the pairwise disconnectivity index does not imply any simplifying assumptions about the significance of paths' length, but rather quantifies how crucial the given vertex is for sustaining the communication ability between connected pairs of other vertices in a network [3].

Table 1. Comparison of nodes with/without self-loops in different mammalian networks.

|	*Transcription*		*Signaling*		*Metabolic*	
	Self-loop	*Other*	*Self-loop*	*Other*	*Self-loop*	*Other*
Frequency	63	216	53	1518	94	1699
Z-Score	39.4	-	34.3	-	51.6	-
Mean inout-degree	10.2	3.1	16	3.9	17.5	5.5
Mean betweenness	0.0021	0.0002	0.004	0.0006	0.0025	0.0005
Mean disconnectivity	0.0261	0.0085	0.0056	0.0021	0.0043	0.0017

The total number of nodes with self-loops in these mammalian networks (Table 1) significantly exceeds the number that might be expected in the corresponding random network of the same size. The high values of Z-score calculated as it was suggested in [12] indicate that self-loops are a network motif there.

These three networks clearly differentiate: First, the relative frequency of nodes with self-edges is significantly higher in the transcription than in the two other networks (29% vs. 3.5% and 5.5%, respectively). Second, these nodes also show a significantly higher mean inout-degree, betweenness centrality and pairwise disconnectivity index. The betweenness centrality is most distinctive in the transcription network: about 10.5-fold more shortest paths pass each autoregulatory node (on average) than each node that has no self-edge in the transcription network, whereas the ratios are 6.7- and 5-fold for signaling and metabolic networks, respectively. The corresponding ratios for the pairwise disconnectivity index are 3.1-fold in transcription, 2.7-fold in signaling and 2.5-fold in metabolic networks. The picture becomes even clearer when considering the maximal values observed: the self-regulating node with maximal betweenness centrality or pairwise disconnectivity index exhibits a much higher value than the corresponding node without self-edge in the transcription network. In contrast, these maximum values are the same, or even lower, for autoregulatory nodes in the other two networks. Altogether, autoregulation seems to be a property of a node that directly goes along with a high topological importance of this node in a network.

3.2 The Mutual Regulation of Two Nodes is a Motif

Two kinds of very basic regulation are conceivable in two-node patterns: first, one node is under the control of another one, represented by a directed edge; second, mutual regulation of the two nodes as indicated by two edges in opposite direction ("binary loop"). Surprisingly, the second pattern is found more frequently than it would be expected by chance, as indicated by the positive Z-scores, and therefore is a motif in all three networks (Table 2). The first pattern is correspondingly "under-represented".

Table 2: Two-node patterns in different mammalian networks.

| Pattern | *Transcription* | | | *Signaling* | | | *Metabolic* | | |
	Freq.	*Z-Score*	$\overline{Dis}$	*Freq.*	*Z-Score*	$\overline{Dis}$	*Freq.*	*Z-Score*	$\overline{Dis}$
●——→●	576	-7.73	0.0026	3316	-15.88	0.0004	5212	-73.5	0.0002
●←——→●	18	**7.73**	0.0032	55	**15.88**	0.0015	232	**73.5**	0.0007

The column *Pattern* depicts the respective pattern, the column *Frequency* gives the number of occurrences of a pattern in a network, a positive *Z-score* indicates statistically significant over-representation. The column $\overline{Dis}$ gives the mean *pairwise disconnectivity index* of all instances of a pattern.

The highest over-representation is found in the metabolic network, which indicates the presence of many antagonistic enzymatic activities such as kinase-phosphatase pairs, where one enzyme consumes (e.g., a phosphoric acid ester) what the other produces (e.g., the free alcoholic hydroxyl group) and *vice versa*. The binary loop between two neighboring nodes in signaling pathways is much rarer than in metabolic pathways, and may indicate mere undirected physical interactions, e.g. between heteromeric subunits of a receptor, or cross-activations between such subunits, etc. The feed-back activation (or repression) in the transcriptional network, also occurs more frequently than statistically expected, but to a lesser extent than in the other two networks.
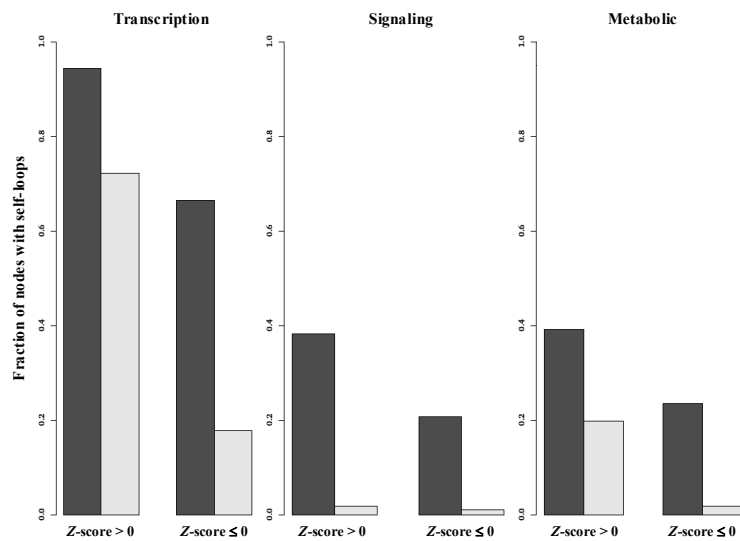
Figure 2: The occurrence of self-loop nodes among 2-node patterns in different mammalian networks. For each network, the patterns (shown in Table 2) are divided according to their frequency: *Z*-score > 0 (over-representation) and *Z*-score ≤ 0. Black and gray bars give the percentage of instances that include at least one node with a self-loop or exactly two self-loop nodes, respectively.

The mean *pairwise disconnectivity index* values are highest in the transcription network, probably due to its smaller size. It is noticeable, however, that these values differ much more between the two 2-node patterns in the metabolic (with a ratio of the *Dis* value for the one-sided pattern to the binary loop of 3.5)   and signaling network (ratio 3.8) than in the transcription network (ratio 1.2).

When analyzing the occurrence of self-loop nodes among these two 2-node patterns, we have observed that appear much more frequently in the motif (the 2-node loop, see Table 2) than in the non-motif pattern (or even "anti-motif"), the one-sided pattern (Figure 2). This effect is even more pronounced when comparing motifs and anti-motifs where both nodes possess a self-loop. The higher frequency of self-loops in the transcription network corresponds to a higher percentage of 2-node patterns with self-loops in the transcription network compared with the other two networks. However, the enrichment of self-loops attached to both nodes of binary loops in the transcription

and the metabolic network is evident, whereas it is much lower in the signaling network (Figure 2).

For the metabolic network, this observation may indicate that there are many mutually connected reversible enzymatic reactions so that whole parts of metabolic paths are reversible. This coincides with our knowledge about biochemical pathways which frequently comprise only few reactions that are *de facto* irreversible under the thermodynamic constraints of a living cell. Particularly interesting seems the observation for the transcription network, where the mutual control of two TFs is frequently accompanied by autoregulation of one or both of the participating TFs. This finding deserves further investigation.

3.3 Three-Node Patterns in the Networks Analyzed

We next analyzed the three networks for the recently most intensively studied kind of patterns, i.e. patterns made of three vertices. Of particular interest here is the feed-forward loop (FFL) pattern which has been shown previously to be a motif in several transcription networks [4]. By computing the frequency of FFL patterns in the networks analyzed here, we obtained significantly positive Z-scores for this pattern in the transcription, signaling and metabolic network (FFL, pattern 38 in Table 3). The same is true for derived patterns 46 and 166, each of them showing one mutual interaction between two of the three nodes and may thus be comprehended as two superimposed FFLs. Also pattern 102 may be considered as motif in all three networks, although its Z-score is very low in the transcription network; this pattern may be viewed as one FFL superimposed with one feed-back loop.

In contrast, the feed-back loop (FBL, pattern 140) is a motif only in the signaling and the metabolic network, but not in the transcription network. The low frequency of FBL in the *E. coli* transcription network was already reported by Alon and colleagues [1], and was observed by us additionally for the transcription networks of yeast and mammals [4]. This may also be a reason why the superposed FFL-FBL motif (pattern 102) has such a low Z-score in the transcription network.

Other observations may be more likely explicable by lack of present knowledge rather than reflect genuine features of the respective network. For instance, the high Z-score of pattern 36 (two nodes acting on a third one) in the metabolic and the low Z-score of the same pattern in the signaling network maybe real: There are certainly many metabolic enzymes that accept different substrates, produced by different enzymes, or one and the same substrate is produced by several other enzymes. In contrast, in signal transduction, convergent information flows resulting in a similar pattern topology are statistically under-represented, although there is a considerable number of such instances (13,606) in the network analyzed. However, we have to assume that the transcriptional regulation of TF-genes, like that of any other gene, is normally exerted by more than just one (other) transcription factor. Therefore, the statistical under-representation of pattern 36 in the transcription network is most likely due to the lack of knowledge about many of

the edges that exist in reality.

It is of interest that a common feature of all those patterns that we have identified as motifs (*Z*-score > 0) in all three networks, is that all their nodes are connected with each other thus forming loop structures of different configurations.

In general, the mean pairwise disconnectivity index indicates that on average only a few percent (mostly around or below 1%; up to 2.2% in case of pattern 102) of all pairwise connections are disrupted when taking out one of these pattern instances (Table 3, $\overline{Dis}$). The average impact of any 3-node pattern is higher in the transcription than any of the two other networks, which may be mainly due to its smaller size. This shows that the networks are rather robust. In most cases, by removing the intrinsic edges of an individual pattern instance, only a very limited number of the existing connections in a network are destroyed and no strong impact at the global scale is observed.

Table 3: Three-node patterns in different mammalian networks.

| Pattern | ID | Transcription | | | Signaling | | | Metabolic | | |
		Freq.	Z-Score	$\overline{Dis}$	Freq.	Z-Score	$\overline{Dis}$	Freq.	Z-Score	$\overline{Dis}$
	6	1916	-0.39	0.0023	11774	-8.92	0.0007	10390	-82.38	0.0003
	12	1068	-1.67	0.011	11865	-9.67	0.0009	14208	-44.75	0.0009
	14	73	-10.47	0.0079	881	-9.48	0.0024	1118	-34.37	0.0016
	36	1620	-2.91	0.0044	13606	-6.91	0.0005	47485	102.03	0.0002
	38	129	5.68	0.0050	496	14.24	0.0006	1627	68.15	0.0003
	46	17	11.18	0.0042	29	8.76	0.0001	85	27.47	0.0003
	78	4	-10.85	0.0099	49	-5.41	0.0011	336	-69.02	0.0014
	102	3	0.35	0.0224	23	8.78	0.0018	101	29.04	0.0044
	140	1	-0.88	0.0137	38	4.64	0.0009	29	6.27	0.0029
	164	197	-6.52	0.0051	722	-10.15	0.0012	4228	-69.65	0.0009
	166	20	7.12	0.0058	21	8.38	0.0001	492	82.09	0.0003
	174	6	7.31	0.0109	9	6.83	0.0001	71	23.75	0.0005
	238	1	-	0.0073	-	-	-	49	729.30	0.0001

The column *Pattern* depicts the respective pattern, to each of them an identifier (*ID*) was assigned. The column *Frequency* gives the number of occurrences of a pattern in a network, a positive *Z-score* indicates statistically significant over-representation. The column $\overline{Dis}$ gives the mean *pairwise disconnectivity index* of all instances of a pattern. The values for the transcription network have been taken from [4].

For none of the networks, a positive correlation between the mean *pairwise disconnectivity index* and the *Z*-score of a pattern could be observed, i.e. motifs are not *per se* more important for the overall network coherence than patterns that are not over-represented.

In spite of this unremarkable feature of the average values, individual instances can be identified for each pattern that show a remarkably high *pairwise disconnectivity index*. In this respect, non-motif patterns exhibit at least as many outliers as motifs (Figure 3). This has already been observed earlier for transcription networks [4], and has been confirmed here for signaling and metabolic networks (Figure 3). It is worth to note that nearly all 3-vertex pattern instances with the highest topological significance in the metabolic network are non-motif patterns and do not exhibit any of the loop structures.
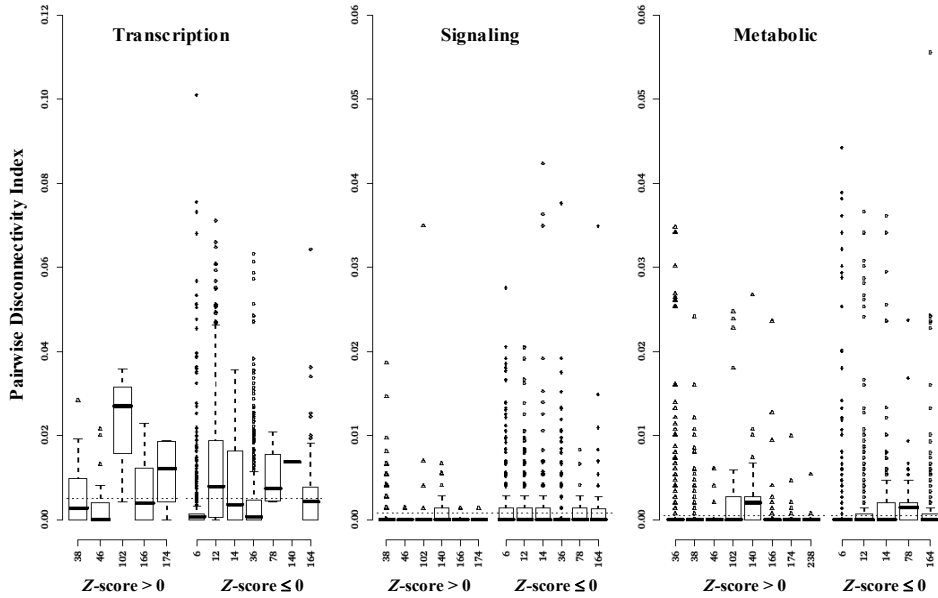

Figure 3: Topological significance of 3-vertex pattern instances in various mammalian networks. For each network, the left and right group of boxplots give the distribution of the *pairwise disconnectivity index* among the 3-vertex motifs or patterns that are not over-represented, resp. The patterns are denoted by their *ID*s (Table 3). The dotted line indicates the average *pairwise disconnectivity index* of all pattern instances in the corresponding network. Note that one point of outliers may represent several pattern instances.

As already described for the 2-node patterns (see 3.2), we also observed an enrichment of vertices with self-loops in 3-node patterns, in particular for the transcription network, but also for the metabolic networks (Figure 4). The difference to the signaling network becomes particularly obvious when focusing on those 3-node patterns where two or even all three vertices have a self-loop. Consistent with our observations about a preferential inclusion of self-loops with binary loops (Figure 2), all 3-node patterns that comprise such a binary loop seem to be particularly rich in self-loops as well.
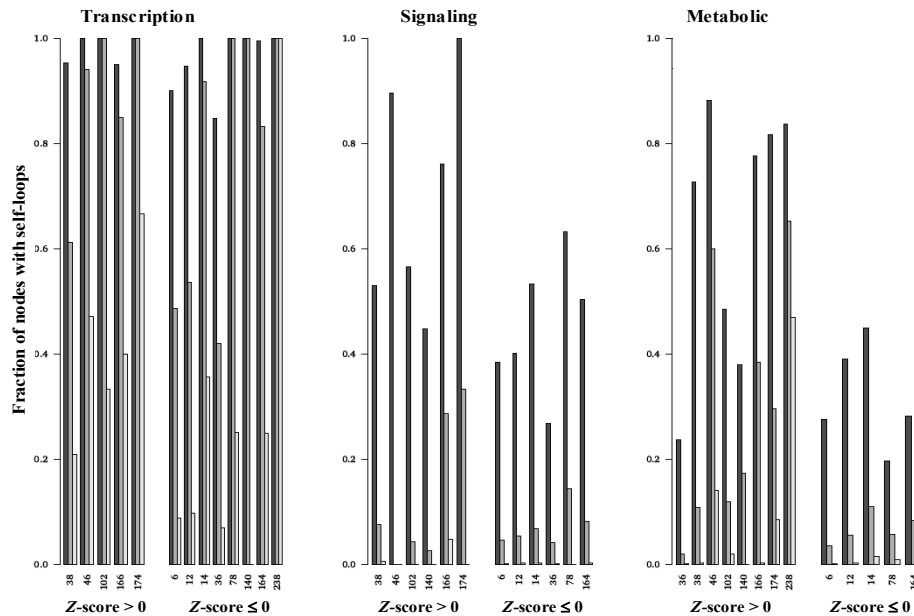
Figure 4: The occurrence of self-loop nodes among 3-node patterns in different mammalian networks. For each network, the patterns (denoted by their IDs according to Table 3) are divided according to their frequency: Z-score > 0 (over-representation) and Z-score ≤ 0. Black, gray and white bars give the fraction of instances that include at least one, at least two or exactly three self-edge vertices, respectively.

3.4 Large Important Subnetworks Derived from Pattern Analysis

We have noticed that many genes / nodes are particularly frequently re-used among those pattern instances that exhibit a very high *pairwise disconnectivity index*. Most of these instances form subgraphs with interesting biological features:

The largest subgraph thus identified in the transcription network comprises 15 TF-genes (Fig. 5a). A number of them are known to be involved in proliferation (*E2F1*, *NSEP1*, c-*myc*, *HMGA1*, c-*fos*, N-*myc*, *POU1F1*) and/or differentiation events (*CEBPA*, *PAX3*, *MITF*, *RUNX2*, *POU1F1*), altogether targeting at the c-*fos* protooncogene. Also, three nuclear steroid receptors are part of this subgraph (*NR3C1* / glucocorticoid receptor, *NR1I3* / constitutive androstane receptor, *ER1* / estrogen receptor 1), rendering it probable that the cell cycle regulatory effects this subnetwork may exert are also under some hormonal control. From c-*fos*, only two edges point back to other TF-genes, namely to c-*myc* and *HMGA1*. These three genes form one of the three instances of pattern 102, the superimposed FFL/FBL motif. It seems noteworthy that 7, i.e. nearly half, of the 15 nodes of this subgraph exhibit autoregulatory edges (Fig. 5a). These are twice as many as in the whole transcription network, where only 29% of all nodes exhibit a self-edge

(Table 1).

In the signaling network, the largest connected subgraph composed of three-node patterns with high disconnectivity is the one depicting the neighborhood of p53, the central regulator of cell cycle and apoptosis (Fig. 5b). Here as well, 11 out of the 12 connected molecules are involved in cell cycle regulation (Cdk1, cyclin B, APC2, p300, Plk1, Pin1, securin, RSK2, p53, Bcl-xL, Aurora-A). Five out of the 19 edges in this subnetwork are known to have a negative sign, i.e. represent an inhibition. Only one node, RSK2, possess an autoregulatory edge (~8%, compared with the 3.5% in the whole signaling network), which might be in the range of what was to be expected. RSK2 is known to autophosphorylate its serine 386 [13].

Concatenation of three-node patterns with high disconnectivity from the metabolic network does not lead to any prominent subnetwork. The only one is a tree-like structure where nucleoside-diphosphate kinase is the "donor" and many NTP-consuming enzymes are targets (not shown).
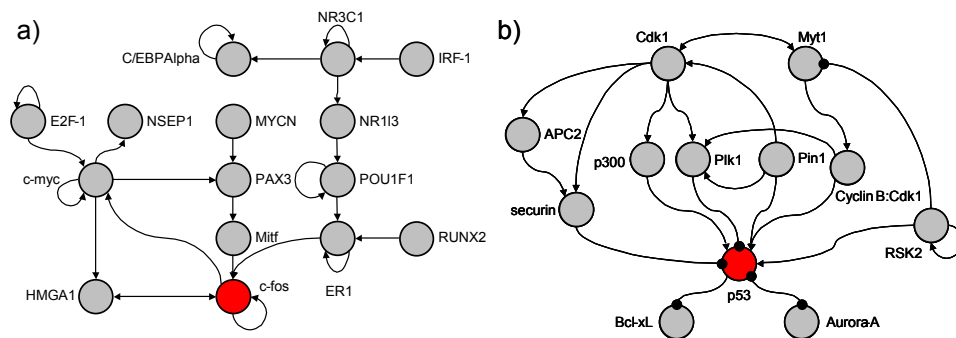

Figure 5: Subnetworks obtained from concatenation of patterns with highest *pairwise disconnectivity index* and common nodes/edges. (a) From the transcription network, a subgraph around the c-*fos* protooncogene was obtained; in addition, known self-loops have been included; (b) from the signaling network, a p53-centered subnetwork was retrieved; line with a dot-end represent inhibitory arcs, the arrows indicate activating interactions.

4   Conclusions

Enrichment analysis and assessment of the topological importance of small network patterns have been proven to complementarily reveal interesting properties of topological patterns and their specific instances. The three kinds of mammalian networks investigated in this study differ characteristically with regard to content and role of their patterns.

The 1-node pattern with a self-edge, termed here "self-loop", has been found to occur in a particularly high frequency in transcription networks. Nodes with a self-edge on average exhibit a higher *inout*-degree, a higher *betweenness centrality*, and a higher *pairwise disconnectivity index* than vertices without a self-loop. All 2- and 3-node patterns that are over-represented and thus qualified as motifs in all three networks exhibit a loop structure. This intriguing observation can be viewed as an advantage of loop-like structures in building up the regulatory circuits of the whole cell. It is in

accordance with the expected role of the loop structure in synchronizing the dynamical processes in scale-free networks [14].

Analysis of the two 2-node patterns relevant in the networks analyzed here also revealed that the binary loop is over-represented, whereas the linear pattern is under-represented. Likewise among the 3-node patterns, all over-represented structures (i.e., motifs) exhibit a loop structure. Both the 2- and the 3-node loops are additionally enriched by attached self-loops, suggesting that loop structures have adopted an important role during evolution of these networks.

The previously introduced *pairwise disconnectivity index* [3] has proven useful to provide a metric of the importance of individual topological patterns for the coherence of the network. While abundance and average importance of patterns may go along with each other in the case of 1- and 2-node patterns, no positive correlation has been observed for 3-node patterns. Non-motif patterns exhibit at least as many outliers with highly elevated *pairwise disconnectivity index* as motifs do. 3-Node pattern instances with the highest pairwise disconnectivity values usually do not belong to any of the loop structures.

Some of these 3-node patterns showing highest importance for the respective network revealed overlapping components and could be concatenated to larger subgraphs. This was possible for the transcription and the signaling network, and both subgraphs turned out to be in connection with the regulation of cell proliferation. We regard this as prove for the biological relevance of our disconnectivity analysis.

Summarizing the comparison of the different networks, we have noticed that the transcription network differs from the other networks in that (*i*) self-loops play an even higher role, (*ii*) its binary loops are highly enriched with self-loops attached, and (*iii*) feed-back loops are not over-represented. Metabolic networks reveal some particular topological properties which may reflect the fact that metabolic paths are, to a large extent, reversible.

Acknowledgments

References

[1] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., Network motifs: Simple building blocks of complex networks, *Science*, 298:824-827, 2002.

[2] Shalgi, R., Lieber, D., Oren, M., and Pilpel, Y., Global and local architecture of the mammalian microRNA-transcription factor regulatory network, *PLoS Comput. Biol.*, 3(7):e131, 2007.

[3] Potapov, A.P., Goemann, B., and Wingender, E., The pairwise disconnectivity index as a new metric for the topological analysis of regulatory networks, *BMC Bioinformatics*, 9:227, 2008.

[4] Goemann, B., Wingender, E., and Potapov, A.P., An approach to evaluate the topological significance of motifs and other patterns in regulatory networks, *BMC Syst. Biol.*, 3:53, 2009.

[5] Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., and Wingender, E., TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res.*, 34(Database issue):D108-D110, 2006.

[6] Krull, M., Pistor, S., Voss, N., Kel, A., Reuter, I., Kroneberg, D., Michael, H., Schwarzer, K., Potapov, A., Choi, C., Kel-Margoulis, O., and Wingender, E., TRANSPATH®: An information resource for storing and visualizing signaling pathways and their pathological aberrations, *Nucleic Acids Res.*, 34(Database issue):D546-D551, 2006.

[7] Choi, C., Crass, T., Kel, A., Kel-Margoulis, O., Krull, M., Pistor, S., Potapov, A., Voss, N., and Wingender, E., Consistent re-modeling of signaling pathways and its implementation in the TRANSPATH database, *Genome Inform.*, 15(2):244-254, 2004.

[8] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y., KEGG for linking genomes to life and the environment, *Nucleic Acids Res.*, 36(Database issue):D480-D484, 2008.

[9] Dobrin, R., Beg, Q.K., Barabási, A.L., and Oltvai, Z.N., Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network, *BMC Bioinformatics*, 5:10, 2004.

[10] Freeman, L.C., A set of measures of centrality based on betweenness, *Sociometry*, 40:35-41, 1977.

[11] Girvan, M. and Newman, M,E., Community structure in social and biological networks**, *Proc. Natl. Acad. Sci. USA*, 99: 7821-7826, 2002.

[12] Alon, U., *An Introduction to Systems Biology. Design Principles of Biological Circuits*, Chapman & Hall/CRC, Boca Raton, 2007.

[13] Frödin, M., Jensen, C.J., Merienne, K., and Gammeltoft, S., A phosphoserine-regulated docking site in the protein kinase RSK2 that recruits and activates PDK1, *EMBO J.*, 19(12), 2924-2934, 2000.

[14] Ma, X., Huang, L., Lai, Y.C., and Zheng, Z., Emergence of loop structure in scale-free networks and dynamical consequences, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys*., 79(5 Pt 2):056106, 2009.