Anant Vinjamoori
Biochem 218 Final Project
March 14, 2008

Using Network Motif Analysis to Explore Transcriptional Regulatory Systems

Biological processes are the products of highly complex and dynamical systems. As the molecular characterization of cellular activity has moved to the systems level, the various interactions between proteins, DNA, between genes have come to be appreciated within networks with intricate circuitry describing relationships between gene expression and the cell environment[1]. Recent study has suggested that these biological networks contain a small set of recurring regulation patterns, termed network motifs, which occur far more often than would be predicted by chance. These patterns are present in the regulatory networks of organisms ranging from simple microorganisms to plants and animals[1,2].

In the networks governing transcriptional regulation, transcription factors respond to biological signals and accordingly modulate the transcription rates of genes, enabling cells to make the proteins they need at the appropriate times and amounts. Embedded within these networks are motifs which help to carry out specific information-processing functions[1]. In this review, we will consider the detection and representation of network motifs as well as their applications in the analysis of transcriptional regulatory systems. We will start by discussing basic properties of network motifs, and present different techniques to detect them in transcriptional networks. The important issue of visualization and representation of network motifs will also be considered and discussed. Finally, we will also discuss some applications of network motif analysis in more dynamic contexts, such as network motif behavior and network motifs vis-a-vis evolution.

Examples of Network Motifs

Before the detection and analysis of network motifs is discussed, it is useful to consider some examples of biological network motifs. As would be expected, the network motifs found in eukaryotic cells are much more diverse and complex than the network motifs found in prokaryotic cells. In this section, 4 common network motifs found in the transcriptional regulation network of *E. coli* will be presented.

Autoregulation

The autoregulation motif is the simplest possible motif in a regulatory network **(Fig 1).** In this motif, a particular species up- or down-regulates its own expression and/or activity. In negative autoregulation, a transcription factor represses its own promoter. In positive autoregulation, a transcription factor activates its own promoter. Negative autoregulation accelerates the response time (time necessary to reach ½ steady state concentration), while positive autoregulation slows the response time[1,3].

Feedforward Loops (FFLs)

In the feedforward loop network motif, a general transcription factor X regulates the transcription of a second transcription factor Y, with both X and Y together regulating the expression of a structural gene Z. Feedforward loops may be further subdivided into coherent and incoherent feedforward loops **(Fig 2)**. In coherent feedforward loops, the direct of effect of the general transcription factor has the same sign as the net indirect effect via the specific transcription factor. By contrast, in incoherent feedforward loops, the direct and indirect effects have different signs[4]. The coherent FFL regulatory motif has been shown to result in sensitivity to changes in regulator concentration in one direction (ie OFF to ON), and insensitivity to concentration changes in the opposite direction (ie ON to OFF). This type of network functioning may confer a survival advantage in the context of a rapidly changing environment. By contrast, the incoherent FFL has been shown to result in increased response time of gene expression following stimulus steps[4,5].

Single Input Modules (SIMs)

In the single input module network motif, a single transcription factor may control the expression of a number of different structural genes. Single Input Module motifs are useful for controlling the timing of gene expression, with the temporal activation pattern resulting from the different activation thresholds of the various structural genes of interest **(Fig 3)**. Kalir et. al demonstrate that a single input module motif is responsible for the exact timing in the assembly of flagella in *E. coli.* This network motif enables a remarkably detailed temporal program of transcription associated with various multiple steps of flagella assembly[6].

Dense Overlapping Regulons

The dense overlapping regulon network motif is akin to a “multiple input, multiple output” module. In this motif, a layer of overlapping interactions between operons and a group of transcription factors enables different inputs to regulate many different outputs. The stress response system of E. coli encodes a dense overlapping regulon motif[1] (**Fig 4**).

Detecting Network Motifs

The detection of network motifs in a biological network generally consists of three tasks, all of which require substantial computational power. Firstly one must find which subgraphs occur in the input network and in what number. Second, one must determine of which these subgraphs are topologically equivalent (i.e isomorphic) and group them accordingly into subgraph classes. Finally, one must determine which subgraphs are displayed at much higher frequencies than in random graphs (under a specified random graph model)[7,8].

Tools Currently Available

Most existing algorithms for detecting network motifs function by enumerating all of the possible subgraphs with a particular number of network nodes. Unfortunately, the processing power and time required for such algorithms increases substantially as the network size increases[6,7]. An algorithm that rapidly samples subgraphs to more efficiently detect network motifs has been proposed by Kashtan *et. al*. This algorithm, named MFINDER, is based on a random sampling of specific subgraphs, and is thus capable of detecting network motifs with only a small number of samples. Effectively, the runtime of this algorithm is thus asymptotically independent of the network size. The effectiveness of this algorithm has been demonstrated in a wide variety of biological networks, including *E.coli* and yeast transcriptional networks, as well as *C. elegans* neuronal networks[9].

However, Wernicke et. al have demonstrated that the algorithm proposed by Kashtan et. al suffers from sampling bias and is only efficient when the network motifs are small. When subgraphs become large, the algorithm slows tremendously. To address these issues, a new improved detection tool, named FANMOD, based on the previous algorithm was developed (**Fig 5**). Unlike the algorithm of Kashtan et. al, FANMOD enumerates all subgraphs for a network of a given size, but then groups them into isomorphic subgraph classes, and then determines the frequency of these subgraph *classes* in a randomization of graphs whose number is specified by the user. The overall result of this approach is an algorithm which is more thorough (no sampling bias, unique capability to handle different kinds of interactions, ie protein-gene) and faster (runtime 10s vs. 620s for MFINDER to enumerate 5-node subgraph in *E. coli* transcriptional network [7,8]

Representing Network Motifs in Network Images

Once motif families have been identified to describe the biological network of interest, obtaining a proper image of the network is crucial to the appreciation of global network structure and to further analyses of the network. However, the tremendous size and complex nature of the data sets acquired from network motif analyses can make representation and visualization a challenge.

Milo et. al achieve a compact representation of the *E. coli* transcriptional network using symbols to represent particular network motif structures. These symbols are then arranged in a 2-dimensional space, with nodes representing operons and lines representing transcriptional regulation[10] (**Fig 6)**. However, a major problem with this method of representation is the sheer visual complexity of the network image; it is difficult to appreciate network connections and motifs from such a diagram. Alternative approaches suggest simplification of subgraphs depictions, but these can result in the loss of network information[11].

Huang et. al describe an alternative approach to network visualization that purports to address these issues. A "parallel plane layout" with the same symbology as before is proposed, with the various motifs identified in a network separated onto different plane layers in 3 dimensional space. Each plane represents a particular motif type. Each individual motif is placed into a cluster sphere, which is transparent to allow the viewer to see edges between motifs and the attendant relationships between them (**Fig 7)**. Relative to the method used by Milo et. al, this method provides an optimal amount of information whilst overcoming excessive complexity by separating the representation into planes in 3-dimensional space[11].

Exploring Dynamic Network Motif Behavior- SANDY

To this point, we have considered only the application of network motif analysis to static transcriptional regulatory networks. Approaching networks from a dynamic perspective spawns a number of interesting questions. If network motifs are thought carry out specific information processing functions, how might network architecture vary in response to different environmental stimuli? Can networks "rewire" their architecture? Luscombe et. al have pioneered a novel, effective approach to examining the dynamical behavior of a biological network. Specifically, they employ an approach termed Statistical Analysis of Network Dynamics (SANDY), which was capable of uncovering large changes in the underlying transcriptional network architecture of *S. cerevisiae*[12]. The SANDY technique is based on a previously described software tool named TopNet, which is used to correlate protein properties with topological statistics. Given an arbitrary undirected network and group of node classes as inputs, TopNet can compute important topological statistics, create sub-networks and draw power-law degree distributions for each sub-network[13]. SANDY begins by examining global characteristics that quantify network architecture, such as network size (i.e. number of transcription factors) and topological measures (In-degree, Out-degree, etc.). Importantly, SANDY also calculates the occurrence of specific network motifs, such as FFLs and DORs, within the network **(Fig 8)**. Once these global measures are quantified, the condition-dependency of the network architectures is then examined. This analysis technique enabled the authors of this study to make a number of striking findings. Specifically, it was found that in response to different stimuli, many transcription factors in the *S. cerevisiae* genome rewired the transcriptional network by altering their interactions with other network components to varying degrees. The SANDY approach is limited by existing datasets, which provide only gene expression data. As interaction data becomes more available, it will become possible to see how network topologies change when condition-specific interactions are integrated as well[12]. The authors anticipate that their findings will remain valid since their observations were robust to large perturbations in the system (perturbing static network by 30% through random addition, deletion and replacement of interactions). However, these claims are ultimately unfounded since purely stochastic perturbation of nodes in the network very conceivably may overlook crucial interactions and subgraphs in the network[10].

Understanding Network Motifs and Evolution

As we begin to understand how network motifs behave dynamically in biological systems, we can then ask how the genetic circuits that we observe evolved in the first place. It is well known that genes evolve most often by conservative evolution, wherein genes with similar functions originate from a common ancestor gene[1]. However, it does not seem likely that network motifs evolved in a similar fashion. Completely unrelated transcription factors can regulate similar output genes in different organisms. Consider two homologous genes in two organisms that are both regulated by SIMs in response to similar environmental cues. If the two SIMs had a common-ancestor SIM, the regulators in these systems would similarly be homologous. However, the sequences of the regulators can sometimes be so different that they are classified into different transcription factor families. Rather, independent convergent evolution on the same regulation circuit appears to present the most likely explanation[1,14,15].

Directly consistent with this hypothesis, Conant et. al present evidence that multiple types of transcriptional regulation circuitry in *E. coli* and *S. cerevisiae* have evolved independently and not through the duplication of ancestral circuits[14]. They consider 2 circuit classes in *E. coli* and 6 circuit classes in *S. Cerevisiae.* Since their approach requires uniform network topologies, they are not able to analyze the network for the dense-overlapping regulon. Moreover, they consider only regulatory genes and not downstream targets. Central to this study of circuit duplication is the study of gene duplication. Remarkably the authors use a relatively unsophisticated method to identify duplicate genes. They use a gapped BLAST with a threshold value of $E_{critical}$<10^-5. However, this liberal approach to identifying gene duplicates can lead only to the detection of even more duplicate circuits. Thus, the fact that their hypothesis of independent circuit evolution is affirmed in the presence of this bias lends even further support to their findings. However, the lack of inclusion of the dense-overlapping regulon motif in their analysis raises concern about validity of the *E. coli* findings, since the DOR was previously shown to be one of 3 representative network motifs in the *E. coli* transcriptional regulatory network[10].

Moreover, it has been suggested that even closely related organisms frequently have different network motifs. To prove this hypothesis, it was first established that orthologous transcription factors and transcription factors generally share the same regulatory interaction if the sequences of the regulators are sufficiently similar`[15,16]. The term “regulog” was coined to describe orthologous regulators with sequence identity generally greater than 30-60%, depending on protein family. This “regulog” concept is useful as a prediction tool, as regulatory interactions may be transferred between organisms so long as orthologous transcription factors and ortholgous target genes exist[15, 16,].

Applying the regulog principle in examining 1293 interactions in *E. coli* and a closely related pathogenic proteobacterium *P.aeruginosa*, Babu et. al find that whereas the conservation of genes and gene interactions is related to the phylogenetic distance between organisms, the conservation is of network motifs is not. Moreover, they also find that regulatory interactions in motifs were lost at the same rate as other interactions in the transcriptional network. These findings suggest that transcriptional regulatory networks evolve in an incremental fashion, with the loss and gain of individual interactions being more important than the loss and gain of whole motifs[16].

On a larger scale, Ihmels et. al demonstrate that the *use* of motifs within organisms can also evolve over time. Examining the transcriptional network of S. cerevisiae, they describe a large-scale modulation of the yeast transcription program connected to the emergence of the capacity for rapid anaerobic growth[17]. Specifically, they find that while genes coding for mitochondrial and cytoplasmic ribosome proteins display a strongly correlated expression pattern in the aerobic fungus *C. albicans*, this correlation is lost in the fermentative *S. cerevisiae* following an apparent genome duplication event. They then demonstrate that this change in gene expression is connected to the loss of a particular *cis*-regulatory element from dozens of genes[17]. While this study does not provide any direct analysis as to how particular network motifs are reorganized, it nonetheless opens a number of interesting questions for future investigation along these lines. That is, if a transcriptional regulatory network can be changed on such a large scale, what are the attendant consequences of the change on the network motifs within the network? Which motifs are conserved in such contexts? Are motifs re-arranged in specific patterns? Indeed, viewing networks through an evolutionary lens may uncover the presence of "evolutionary meta-motifs." Such motifs could be used to describe patterns in overall network evolution. The development of more sophisticated high-throughput gene expression and interaction tools as well as network analysis approaches should enable these questions to be answered in the near future.

Conclusion

Network motif analysis has revolutionized the study of transcriptional regulation in organisms of all types. In this review, we have discussed various approaches to detecting and representing network motifs, as well as their limitations. Moreover, we have also considered methods in which network motifs are analyzed in the context of a larger biological process, such as information processing or evolution. In both cases, we find that that applying network motif analysis enables more incisive and potent insights into biological processes that would have never been made with traditional binary experimental approaches. However, there remains much yet to be studied. How network motif function is influenced by its context within the surrounding networks is an important area that has not been well investigated. This is in part due to the fact that in the simple systems currently being studied, individual motif behavior does not appear to be influenced by other cellular networks[1]. Thus, as technologies and capabilities for systems-level scientific approaches grow, it will be important to apply the same network motif analysis to more complex higher order eukaryotic organisms, including humans.

References

1. Prill RJ, Iglesias PA, Levchenko A (2005) Dynamic properties of network motifs contribute to biological network organization. PLoS Biol 3(11): e343.

2. Alon U. Network motifs: theory and experimental approaches.Nat Rev Genet. 2007 June; 8(6): 450–461. doi: 10.1038/nrg2102.

3. Rosenfeld, N., Elowitz, M. B. & Alon, U. Negative autoregulation speeds the response times of transcription networks. J. Mol. Biol. 323, 785–793 (2002).

4. Puigjaner, Luis. CAPE: Computer Aided Process and Product Engineering. Wiley VCH Publishing, 2006. pp. 240-243

5. Mangan S, Zaslaver A, Alon U. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol. 2003 November 21; 334(2): 197–204.

6. Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, Leibler S, Surette MG, Alon U Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. Science. 2001 June 15; 292(5524): 2080–2083. doi: 10.1126/science.1058758.

7. Wernicke,S. (2005) A faster algorithm for detecting network motifs. In Proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI '05), Lecture Notes in Bioinformatics. Vol. 3692, pp. 165–177. (Long version submitted).

8. Wernicke S and Rasche FANMOD: a tool for fast network motif detection. Bioinformatics. 2006 May 1; 22(9): 1152–1153. Published online 2006 February 2. doi: 10.1093/bioinformatics/btl038.

9. Kashtan N, Itzkovitz S, Milo R, and Alon U Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics. 2004 July 22; 20(11): 1746–1758. Published online 2004 March 4. doi: 10.1093/bioinformatics/bth163.

10. Shen-Orr SS, Milo R, Mangan S, and Alon U Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet. 2002 May; 31(1): 64–68. Published online 2002 April 22. doi: 10.1038/ng881.

11. Huang, W., Murray, C., Shen, X., Song, L., Wu, Y. X. and Zheng, L. (2005) Visualization and Analysis of Network Motifs. In Proc. 9th International Conference on Information Visualization (IV'05), IEEE, 697-702.

12. Luscombe NM, Babu M, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature. 2004 September 16; 431(7006): 308–312. doi: 10.1038/nature02782.

13. Yu H, Zhu X, Greenbaum D, Karro J,Gerstein M TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. Nucleic Acids Res. 2004; 32(1): 328–337. Published online 2004 January 14. doi: 10.1093/nar/gkh164.

14. Conant, G. C. & Wagner, A. Convergent evolution of gene circuits. Nature Genet. 34, 264–266 (2003).

15. Yu H, Luscombe N, Lu H, Zhu X, Xia Y, Han J, Bertin N, Chung S, Goh C, Vidal M, Gerstein M: Annotation transfer for genomics: assessing the transferability of protein-protein and protein- DNA interactions between organisms.

16. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol. 2004 June; 14(3): 283–291. doi: 10.1016/j.sbi.2004.05.004.

17. Ihmels J, Bergmann S, Gerami-Nejad M, Yanai I, McClellan M, Berman J, and Barkai N. Rewiring of the yeast transcriptional network through the evolution of motif usage. Science. 2005 August 5; 309(5736): 938–940. doi: 10.1126/science.1113833.

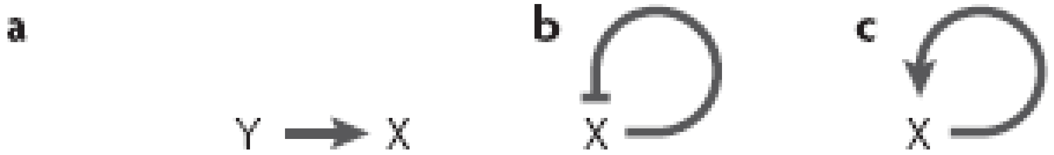


Figure 1: Autoregulatory Feedback Loops (from Alon 2007): a.) In simple regulation, transcription factor Y is activated by a signal Sy. When active, it binds the promoter of gene X to enhance or inhibit its transcription rate. **b.)** In negative autoregulation (NAR), X is a transcription factor that represses its own promoter. **c.)** In positive autoregulation (PAR), X activates its own promoter.

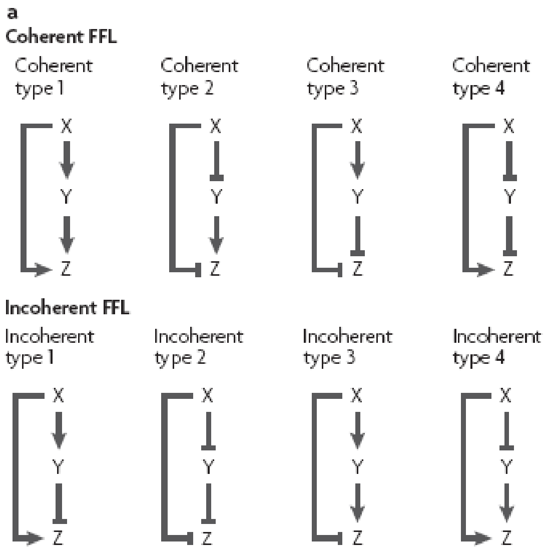


Figure 2: Coherent and Incoherent feedforward loops (from Alon 2007): The eight types of feedforward loops (FFLs) are shown. In coherent FFLs, the sign of the direct path from transcription factor X to output Z is the same as the overall sign of the indirect path through transcription factor Y. Incoherent FFLs have opposite signs for the two paths.

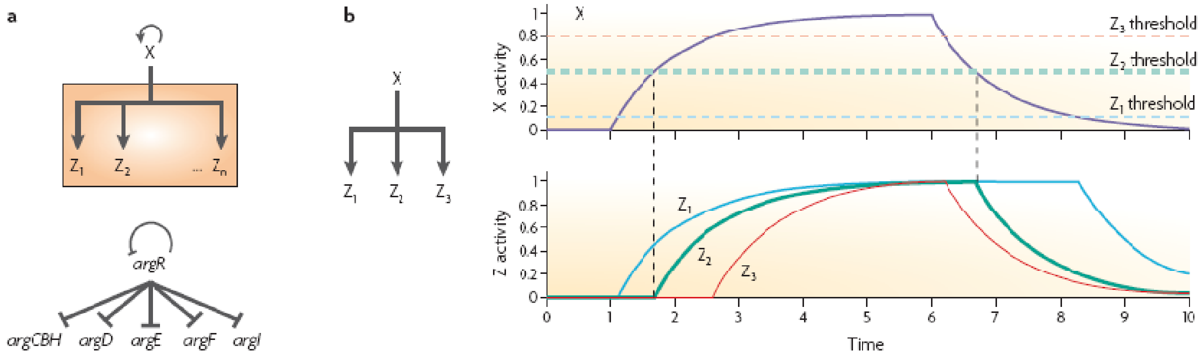


Figure 3: Single Input Modules (from Alon 2007): a.) The single-input module (SIM) network motif, and an example from the arginine-biosynthesis system. **b.)** Temporal order of expression in a SIM. As the activity of the master regulator X changes in time, it crosses the different activation threshold of the genes in the SIM at different times, generating a temporal order of expression.

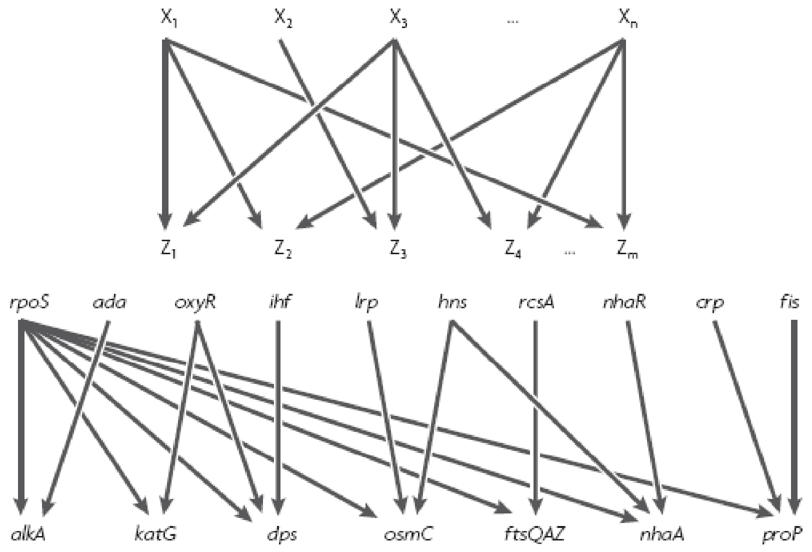
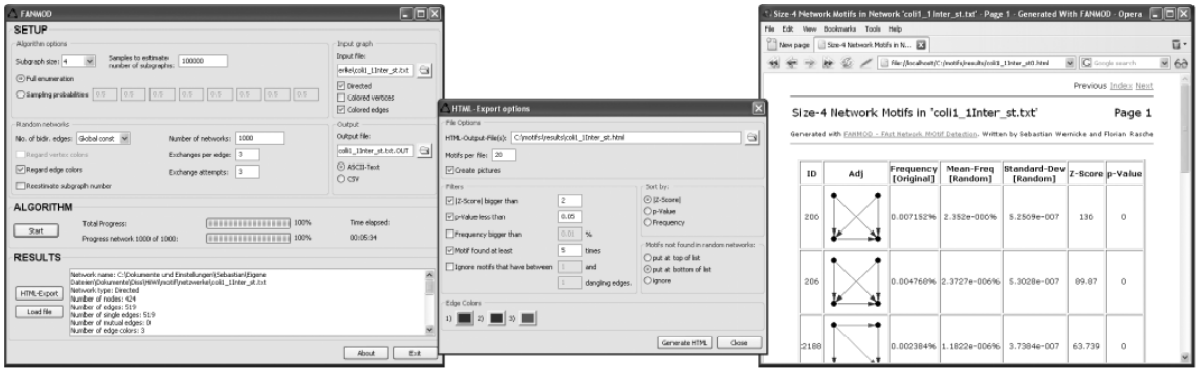


Figure 4: The dense overlapping regulon (DOR) network motif (from Alon 2007): In this motif, many inputs regulate many outputs (top panel). The bottom panel shows an example from the stress-response system of Escherichia coli.



FANMOD

SETUP

Algorithm options

Subgraph size: 4

Samples to estimate number of subgraphs: 100000

● Full enumeration

○ Sampling probabilities 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5

Input graph

Input file: erke\coli1_1Inter_st.txt

☑ Directed

□ Colored vertices

☑ Colored edges

Random networks

No. of bidir. edges: Global const

□ Regard vertex colors

☑ Regard edge colors

□ Reestimate subgraph number

Number of networks: 1000

Exchanges per edge: 3

Exchange attempts: 3

Output

Output file: coli1_1Inter_st.txt.OUT

● ASCII-Text

○ CSV

ALGORITHM

Total Progress: 100%

Progress network 1000 of 1000: 100%

Time elapsed: 00:05:34

RESULTS

```
Network name: C:\Dokumente und Einstellungen\Sebastian\Eigene
Dateien\Dokumente\Diss\HiWi\motif\netzwerke\coli1_1Inter_st.txt
Network type: Directed
Number of nodes: 424
Number of edges: 519
Number of single edges: 519
Number of mutual edges: 0
Number of edge colors: 3
```

HTML - Export options

File Options

HTML-Output-File(s): C:\motifs\results\coli1_1Inter_st.html

Motifs per file: 20

☑ Create pictures

Filters

☑ |Z-Score| bigger than 2

☑ p-Value less than 0.05

□ Frequency bigger than 0.01 %

☑ Motif found at least 5 times

□ Ignore motifs that have between 1 and 1 dangling edges.

Sort by:

● |Z-Score|

○ p-Value

○ Frequency

Motifs not found in random networks:

○ put at top of list

● put at bottom of list

○ ignore

Edge Colors

1) ■ 2) ■ 3) ■

Size-4 Network Motifs in Network 'coli1_1Inter_st.txt' - Page 1 - Generated With FANMOD - Opera

File Edit View Bookmarks Tools Help
New page
Size-4 Network Motifs in N...
Google search

file://localhost/C:/motifs/results/coli1_1Inter_st0.html

Previous Index Next

Size-4 Network Motifs in 'coli1_1Inter_st.txt' Page 1

Generated with FANMOD - FAst Network MOtif Detection. Written by Sebastian Wernicke and Florian Rasche

ID	Adj	Frequency [Original]	Mean-Freq [Random]	Standard-Dev [Random]	Z-Score	p-Value
206		0.007152%	2.352e-006%	5.2569e-007	136	0
206		0.004768%	2.3727e-006%	5.3028e-007	89.87	0
2188		0.002384%	1.1822e-006%	3.7384e-007	63.739	0

Figure 5: Using FANMOD (from Werincke 2006). Detecting size-4 network motifs with colored edges in the transcriptional network of E. Coli using the FANMOD interface (left). Via an export filter (middle), the obtained results can be exported to HTML (right).

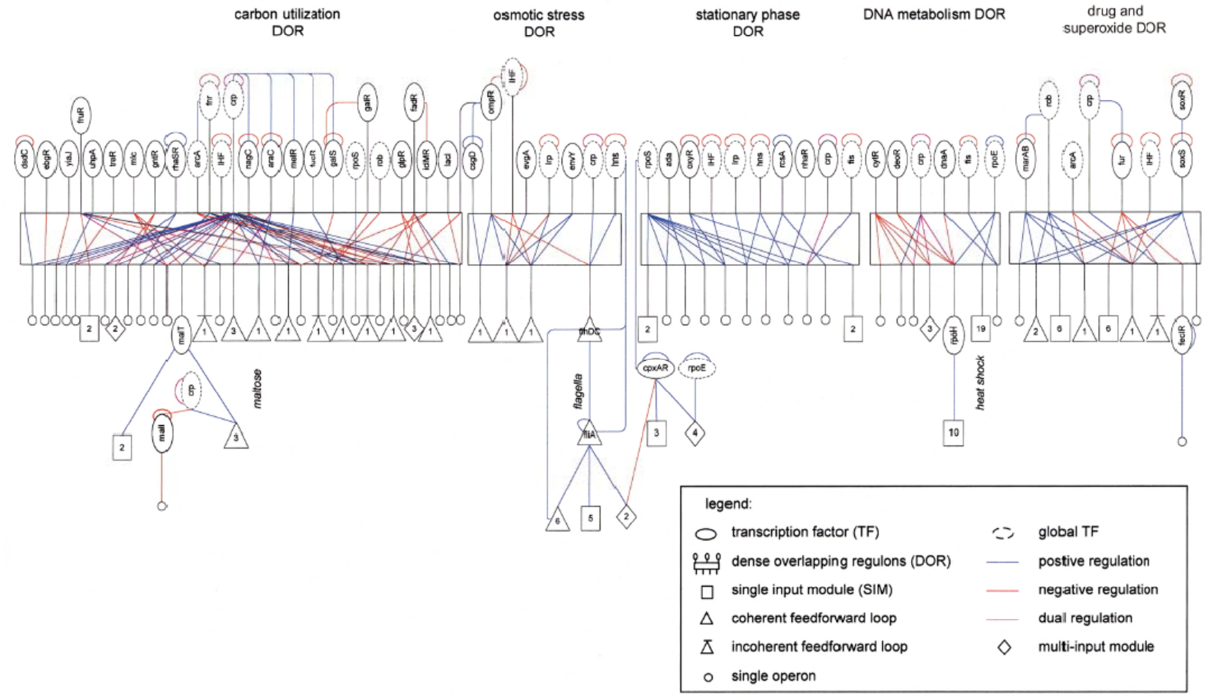


Figure 6 (From Milo 2002): Network image of the transcriptional network of *E. coli* as presented by Milo et. al.

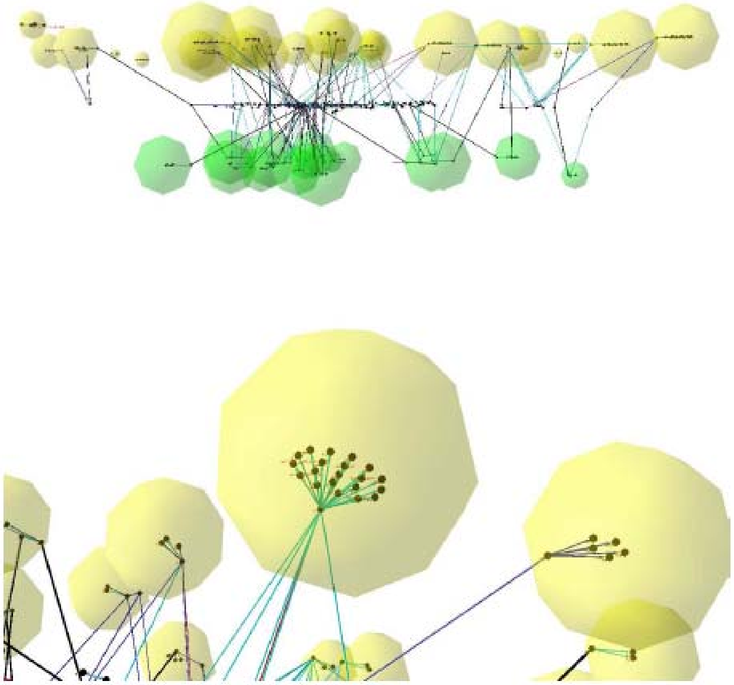
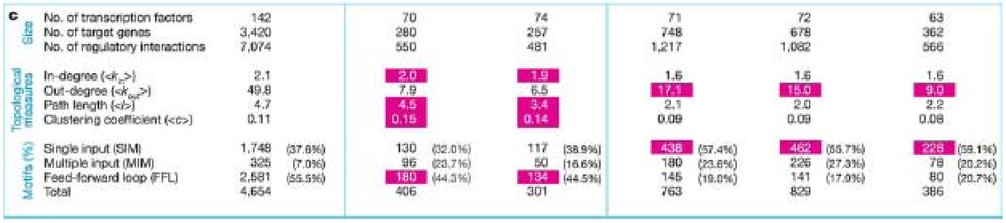


Figure 7 (From Huang 2005): Network image of the transcriptional network of *E. coli* as presented by Huang et. al. Top panel demonstrates the principle of layering of nodes. Bottom panel shows how the transparent spheres highlight the motifs while still showing their structure.



c

Size	No. of transcription factors	142		70		74		71		72		63	
	No. of target genes	3,420		280		257		748		678		362	
	No. of regulatory interactions	7,074		550		481		1,217		1,082		566	
Topological measures	In-degree ($<k_{in}>$)	2.1		2.0		1.9		1.6		1.6		1.6	
	Out-degree ($<k_{out}>$)	49.8		7.9		6.5		17.1		15.0		9.0	
	Path length ($<l>$)	4.7		4.5		3.4		2.1		2.0		2.2	
	Clustering coefficient ($<c>$)	0.11		0.15		0.14		0.09		0.09		0.08	
Motifs (%)	Single input (SIM)	1,748	(37.6%)	130	(32.0%)	117	(38.9%)	438	(57.4%)	462	(55.7%)	228	(59.1%)
	Multiple input (MIM)	325	(7.0%)	96	(23.7%)	50	(16.6%)	180	(23.6%)	226	(27.3%)	78	(20.2%)
	Feed-forward loop (FFL)	2,581	(55.5%)	180	(44.3%)	134	(44.5%)	145	(19.0%)	141	(17.0%)	80	(20.7%)
	Total	4,654		406		301		763		829		386	

Figure 8 (From Luscombe 2003) : Standard topological statistics generated by SANDY for the transcriptional network of *S. cerevisiae.*