Bioinformatics: Network Analysis

Network Motifs

COMP 572 (BIOS 572 / BIOE 564) - Fall 2013
Luay Nakhleh, Rice University

Recall

- Not all subgraphs occur with equal frequency
- Motifs are subgraphs that are over-represented compared to a randomized version of the same network
- To identify motifs:
  - Identify all subgraphs of n nodes in the network
  - Randomize the network, while keeping the number of nodes, edges, and degree distribution unchanged
  - Identify all subgraphs of n nodes in the randomized version
  - Subgraphs that occur significantly more frequently in the real network, as compared to the randomized one, are designated to be the motifs

Outline

- Motifs in cellular networks: case studies
- Efficient sampling in networks
- Comparing the local structure of networks
- Motif evolution

Motifs in Cellular Networks: Case Studies

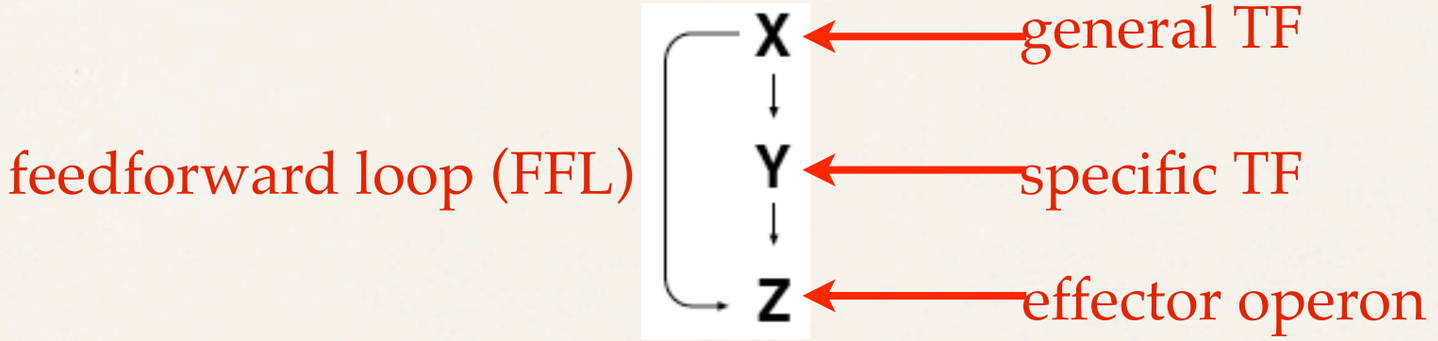
Motifs in Transcription Regulation Networks:
The Data

- Research group: Uri Alon and co-workers
- Organism: E. coli
- Nodes of the network: 424 operons, 116 of which encode transcription factors
- (Directed) Edges of the network: 577 interactions (from an operon that encodes a TF to an operon that is regulated by that TF)
- Source: mainly RegulonDB database, but enriched with other sources

Motifs in Transcription Regulation Networks: Findings

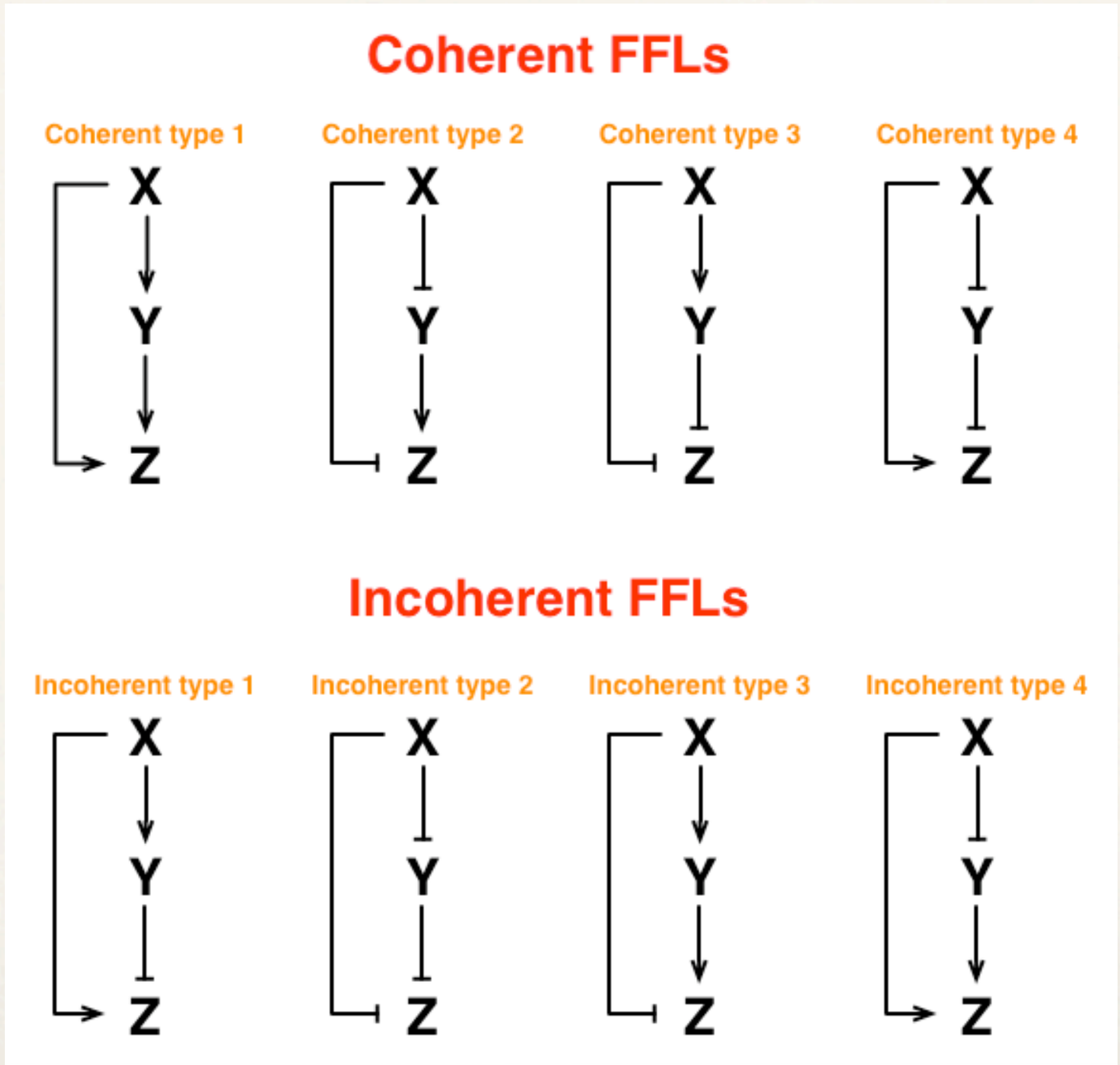
- Alon and colleagues found that much of the network is composed of repeated appearances of three highly significant motifs
  - feedforward loop (FFL)
  - single input module (SIM)
  - dense overlapping regulons (DOR)
- Each network motif has a specific function in determining gene expression, such as generating "temporal expression programs" and governing the responses to fluctuating external signals
- The motif structure also allows an easily interpretable view of the entire known transcriptional network of the organism

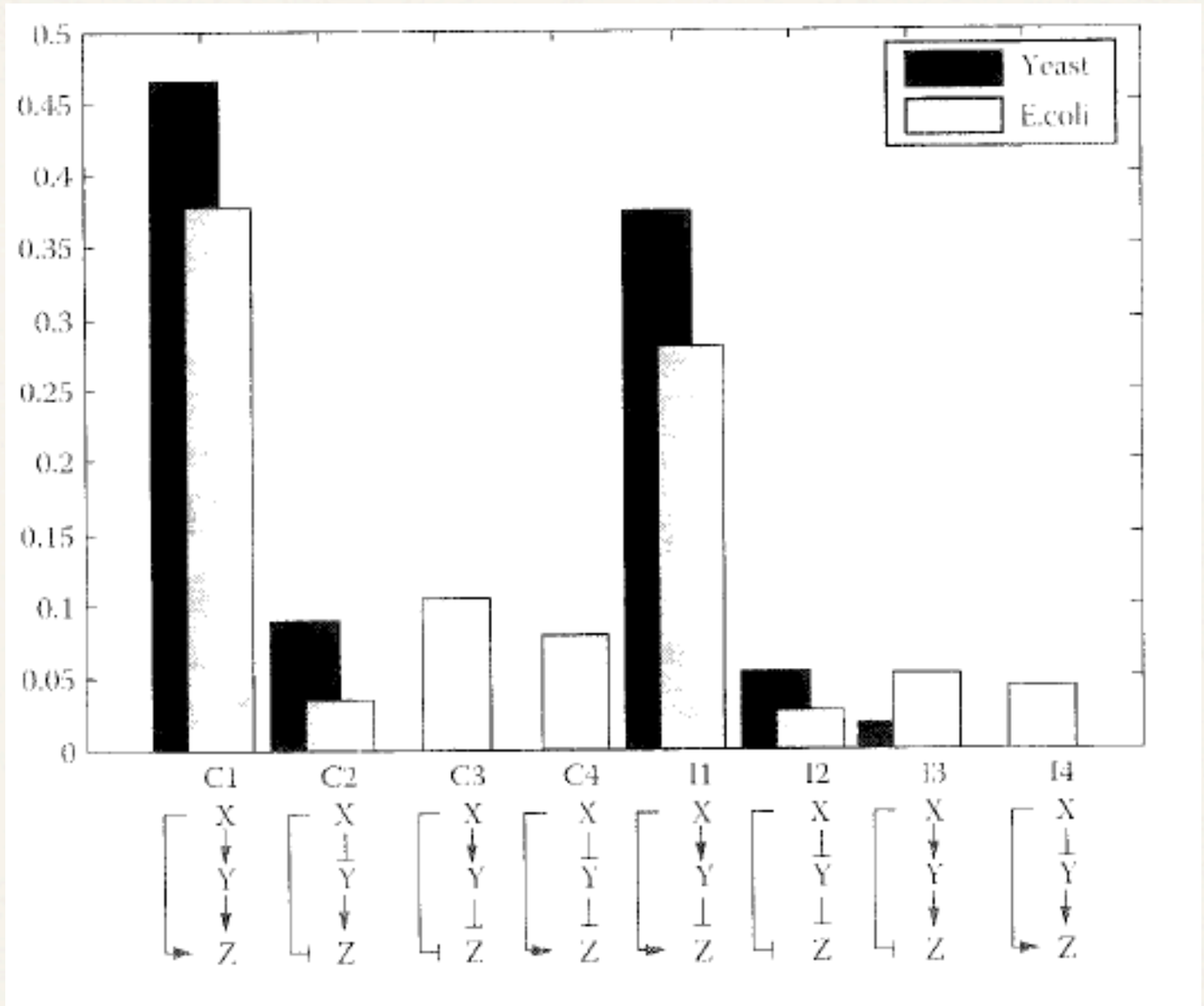
Motifs in Transcription Regulation Networks:
Motif Type (1): Feedforward loops



FFL is { coherent if the direct effect of X on Z has the same indirect effect of X on Z through Y
incoherent otherwise

FFL Types


Coherent FFLs
Coherent type 1
Coherent type 2
Coherent type 3
Coherent type 4
X
Y
Z
Incoherent FFLs
Incoherent type 1
Incoherent type 2
Incoherent type 3
Incoherent type 4

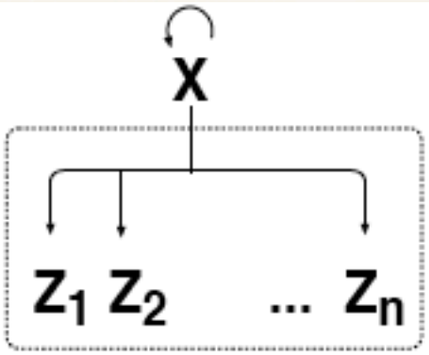


Relative abundance of the eight FFL types in the transcription networks of yeast and E. coli.

FFL types are marked C and I for coherent and incoherent, respectively.

Motifs in Transcription Regulation Networks:
Motif Types (2) and (3): Variable-size motifs

Single input module (SIM)



* All operons $Z_1,...,Z_n$ are regulated with the same sign
* None is regulated by a TF other than X
* X is usually autoregulatory

Dense overlapping regulon (DOR)

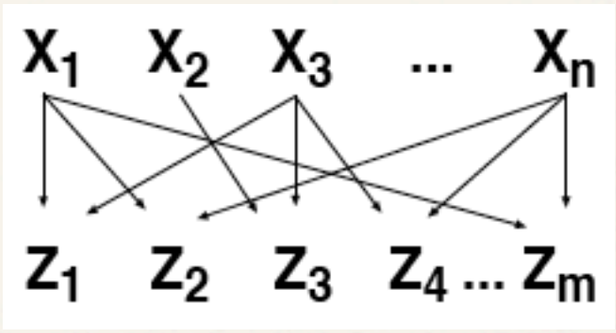


Table 1 • Statistics of occurrence of various structures in the real and randomized networks

Structure	Appearances in real network	Appearances in randomized network (mean ± s.d.)	*P* value
Coherent feedforward loop	34	4.4 ± 3	$P < 0.001$
Incoherent feedforward loop	6	2.5 ± 2	$P \sim 0.03$
Operons controlled by SIM (>13 operons)	68	28 ± 7	$P < 0.01$
Pairs of operons regulated by same two transcription factors	203	57 ± 14	$P < 0.001$
Nodes that participate in cycles*	0	0.18 ± 0.6	$P \sim 0.8$

*Cycles include all loops greater than size 1 (autoregulation). *P* value for cycles is the probability of networks with no loops.

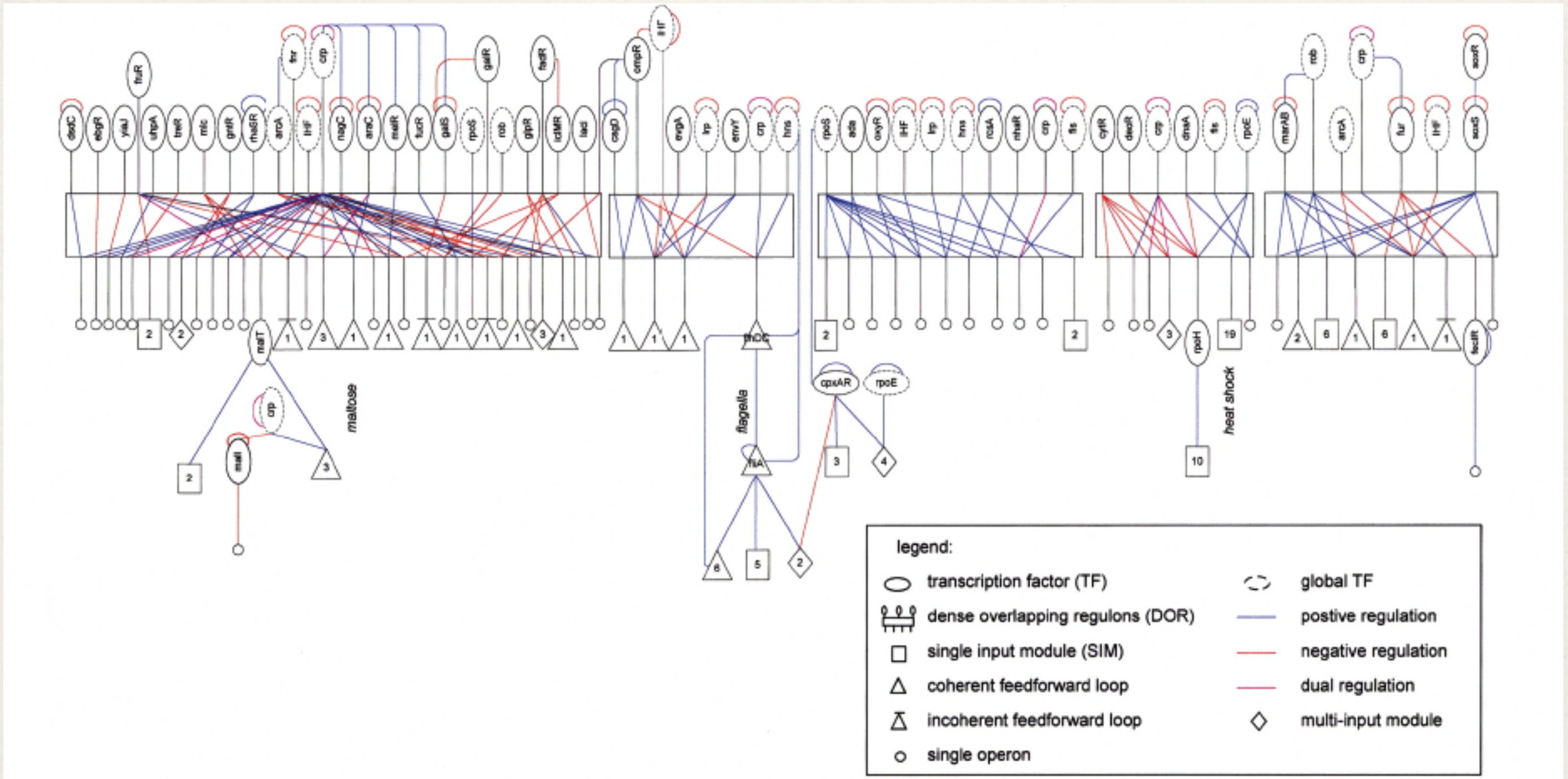


Fig. 3 Part of the network of direct transcriptional interactions in the *E. coli* data set, represented using network motifs. Nodes represent operons, and lines represent transcriptional regulation, directed so that the regulating transcription factor is above the regulated operons. Network motifs are represented by their corresponding symbols (Fig. 1). The DORs are named according to the common function of their output operons. Each transcription factor appears in only a single subgraph, except for transcription factors regulating more than ten operons ('global transcription factors'), which can appear in several subgraphs. For an image of the entire network, see Web Fig. A online.

Motifs in Transcription Regulation Networks:
Functional Roles of Motifs

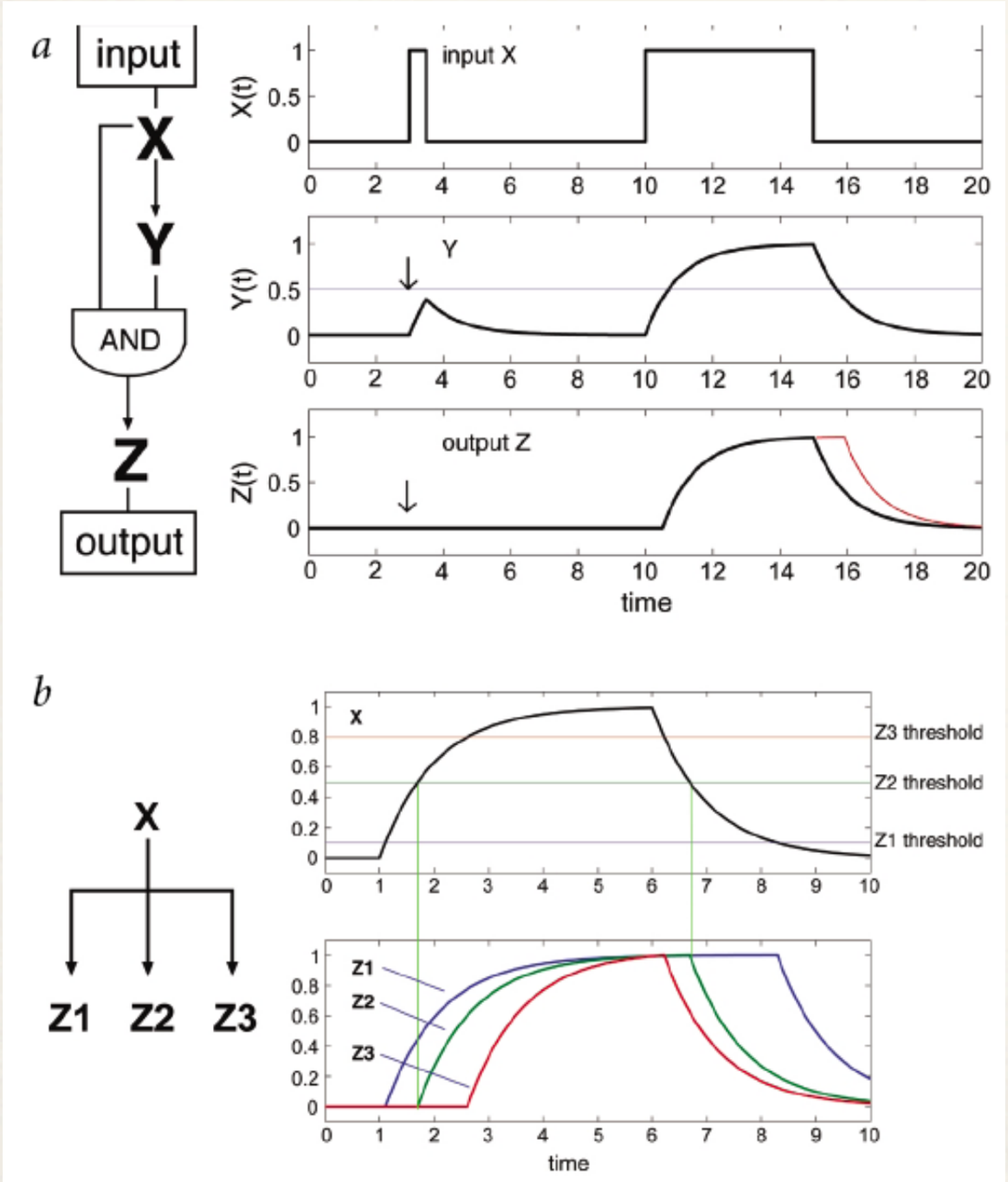


Fig. 2 Dynamic features of the coherent feedforward loop and SIM motifs. *a*, Consider a coherent feedforward loop circuit with an 'AND-gate'–like control of the output operon Z. This circuit can reject rapid variations in the activity of the input X, and respond only to persistent activation profiles. This is because Y needs to integrate the input X over time to pass the activation threshold for Z (thin line). A similar rejection of rapid fluctuations can be achieved by a cascade, X→Y→Z; however, the cascade has a slower shut-down than the feedforward loop (thin red line in the Z dynamics panel). ***b***, Dynamics of the SIM motif. This motif can show a temporal program of expression according to a hierarchy of activation thresholds of the genes. When the activity of X, the master activator, rises and falls with time, the genes with the lowest threshold are activated earliest and deactivated latest. Time is in units of protein lifetimes, or of cell cycles in the case of long-lived proteins.

Motifs in Other Networks

- Following their success at identifying motifs in transcription regulation network in E. coli, Alon and co-workers analyzed other types of networks: gene regulation (in E. coli and S. cerevisiae), neurons (in C. elegans), food webs (in 7 ecological systems), electronic circuits (forward logic chips and digital fractional multipliers), and WWW

Motifs in Other Networks
Motif Types

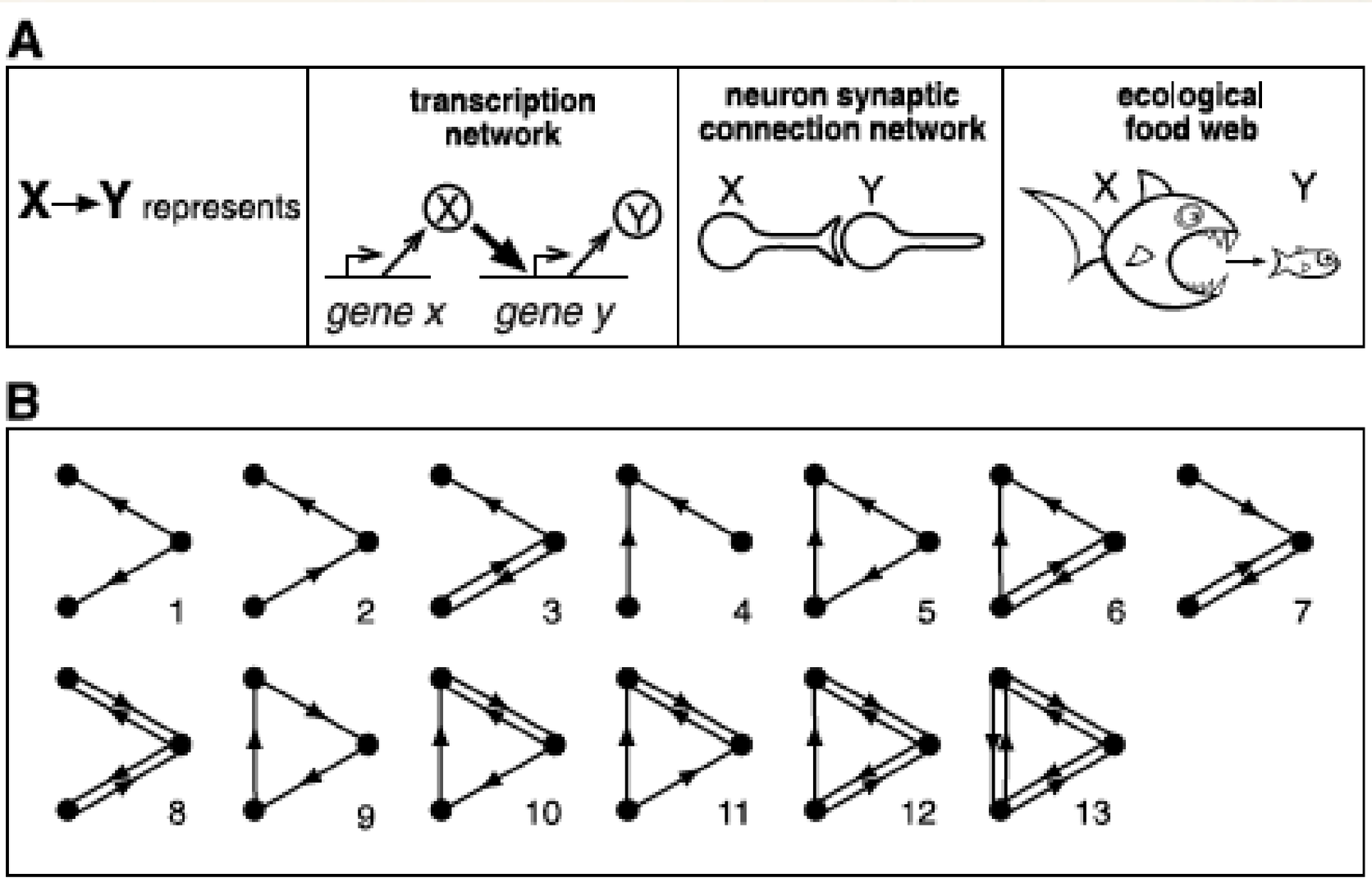


Fig. 1. (**A**) Examples of interactions represented by directed edges between nodes in some of the networks used for the present study. These networks go from the scale of biomolecules (transcription factor protein X binds regulatory DNA regions of a gene to regulate the production rate of protein Y), through cells (neuron X is synaptically connected to neuron Y), to organisms (X feeds on Y). (**B**) All 13 types of three-node connected subgraphs.

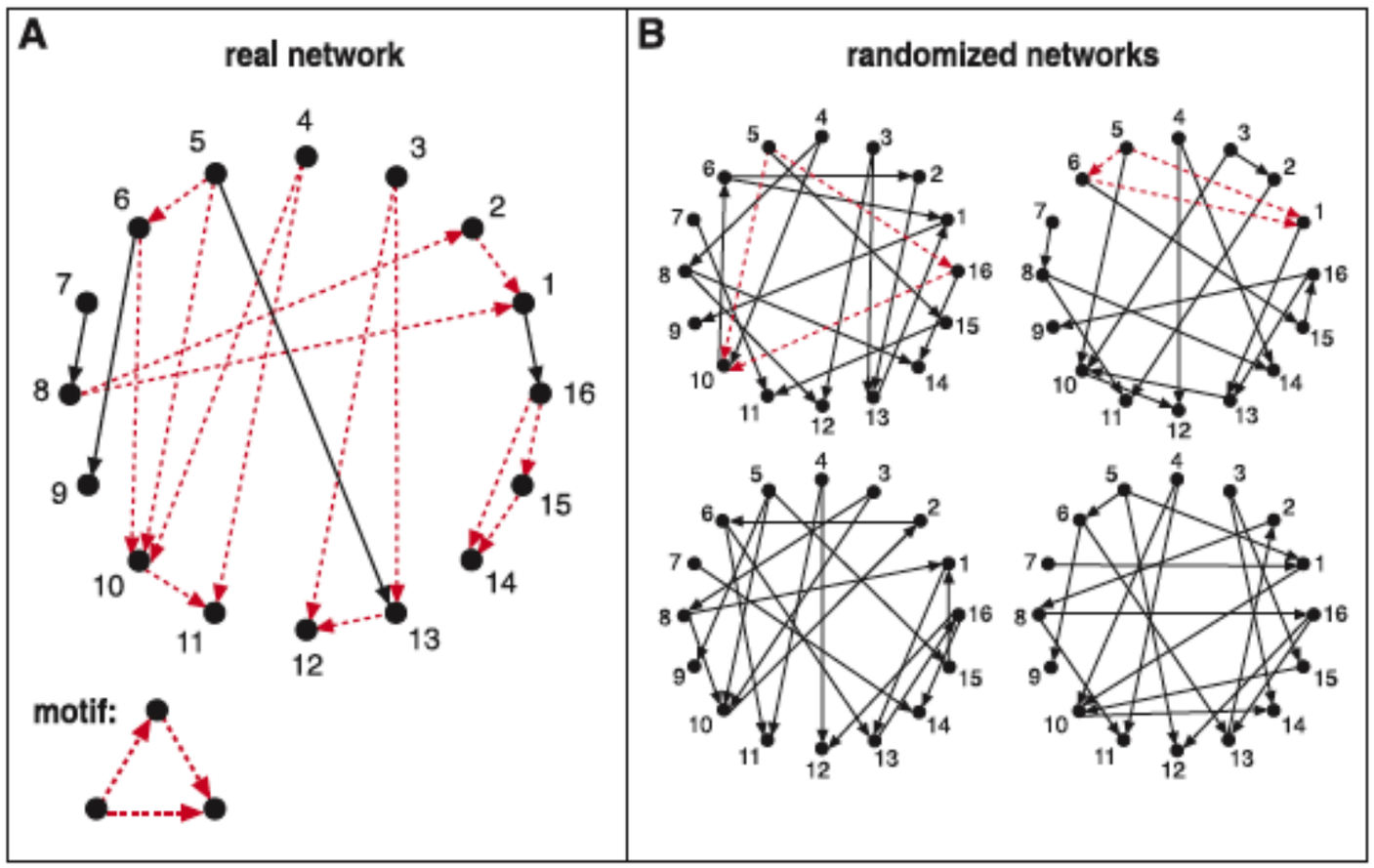


Fig. 2. Schematic view of network motif detection. Network motifs are patterns that recur much more frequently (**A**) in the real network than (**B**) in an ensemble of randomized networks. Each node in the randomized networks has the same number of incoming and outgoing edges as does the corresponding node in the real network. Red dashed lines indicate edges that participate in the feedforward loop motif, which occurs five times in the real network.

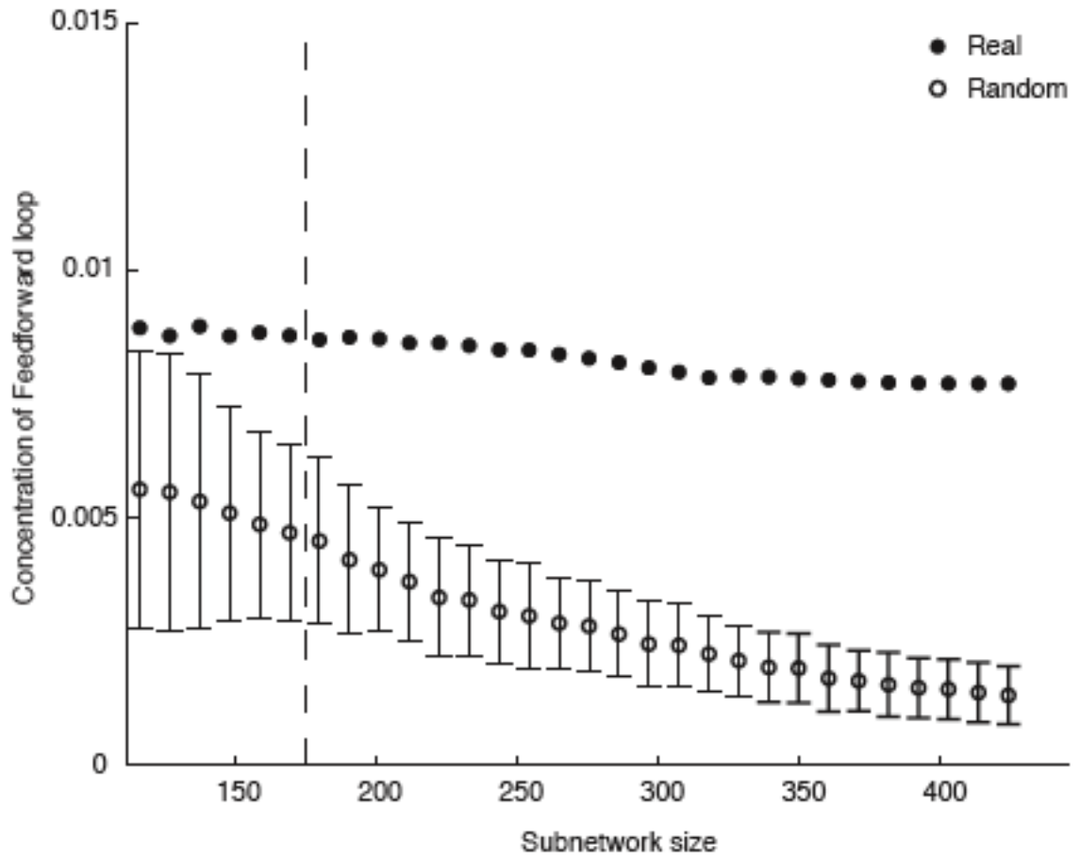


Fig. 3. Concentration C of the feedforward loop motif in real and randomized subnetworks of the *E. coli* transcription network (*11*). C is the number of appearances of the motif divided by the total number of appearances of all connected three-node subgraphs (Fig. 1B). Subnetworks of size S were generated by choosing a node at random and adding to it nodes connected by an incoming or outgoing edge, until S nodes were obtained, and then including all of the edges between these S nodes present in the full network. Each of the subnetworks was randomized (*17*, *18*) (shown are mean and SD of 400 subnetworks of each size).

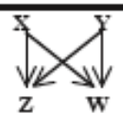
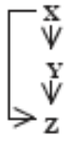
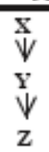
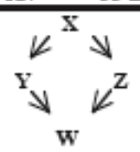
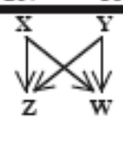
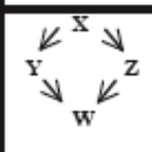
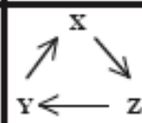
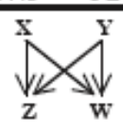
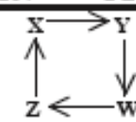
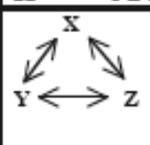
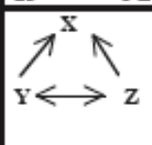
Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Gene regulation (transcription)					**Feed-forward loop**			**Bi-fan**			
E. coli	424	519	40	7 ± 3	10	203	47 ± 12	13			
*S. cerevisiae**	685	1,052	70	11 ± 4	14	1812	300 ± 40	41			
Neurons					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
C. elegans†	252	509	125	90 ± 10	3.7	127	55 ± 13	5.3	227	35 ± 10	20
Food webs					**Three chain**			**Bi-parallel**			
Little Rock	92	984	3219	3120 ± 50	2.1	7295	2220 ± 210	25			
Ythan	83	391	1182	1020 ± 20	7.2	1357	230 ± 50	23			
St. Martin	42	205	469	450 ± 10	NS	382	130 ± 20	12			
Chesapeake	31	67	80	82 ± 4	NS	26	5 ± 2	8			
Coachella	29	243	279	235 ± 12	3.6	181	80 ± 20	5			
Skipwith	25	189	184	150 ± 7	5.5	397	80 ± 25	13			
B. Brook	25	104	181	130 ± 7	7.4	267	30 ± 7	32			
Electronic circuits (forward logic chips)					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

Table 1. Network motifs found in biological and technological networks. The numbers of nodes and edges for each network are shown. For each motif, the numbers of appearances in the real network (N_{real}) and in the randomized networks ($N_{rand} \pm$ SD, all values rounded) (*17, 18*) are shown. The P value of all motifs is $P < 0.01$, as determined by comparison to 1000 randomized networks (100 in the case of the World Wide Web). As a qualitative measure of statistical significance, the Z score $= (N_{real} - N_{rand})/SD$ is shown. NS, not significant. Shown are motifs that occur at least $U = 4$ times with completely different sets of nodes. The networks are as follows (*18*): transcription interactions between regulatory proteins and genes in the bacterium *E. coli* (*11*) and the yeast *S. cerevisiae* (*20*); synaptic connections between neurons in *C. elegans*, including neurons connected by at least five synapses (*24*); trophic interactions in ecological food webs (*22*), representing pelagic and benthic species (Little Rock Lake), birds, fishes, invertebrates (Ythan Estuary), primarily larger fishes (Chesapeake Bay), lizards (St. Martin Island), primarily invertebrates (Skipwith Pond), pelagic lake species (Bridge Brook Lake), and diverse desert taxa (Coachella Valley); electronic sequential logic circuits parsed from the ISCAS89 benchmark set (*7, 25*), where nodes represent logic gates and flip-flops (presented are all five partial scans of forward-logic chips and three digital fractional multipliers in the benchmark set); and World Wide Web hyperlinks between Web pages in a single domain (*4*) (only three-node motifs are shown). e, multiplied by the power of 10 (e.g., 1.46e6 $= 1.46 \times 10^6$).

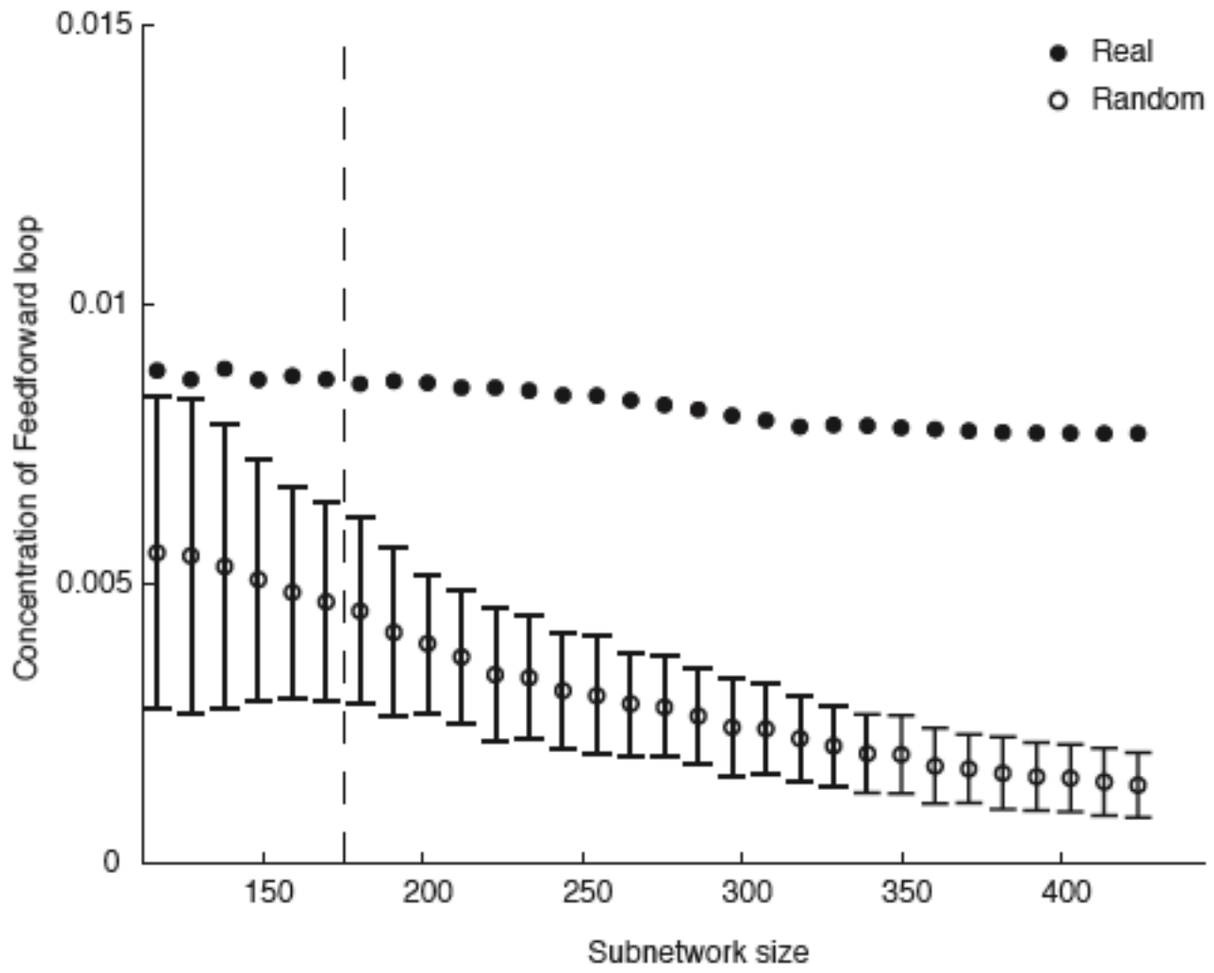


Fig. 3. Concentration *C* of the feedforward loop motif in real and randomized subnetworks of the *E. coli* transcription network (*11*). *C* is the number of appearances of the motif divided by the total number of appearances of all connected three-node subgraphs (Fig. 1B). Subnetworks of size *S* were generated by choosing a node at random and adding to it nodes connected by an incoming or outgoing edge, until *S* nodes were obtained, and then including all of the edges between these *S* nodes present in the full network. Each of the subnetworks was randomized (*17*, *18*) (shown are mean and SD of 400 subnetworks of each size).

Issues with the Null Hypothesis

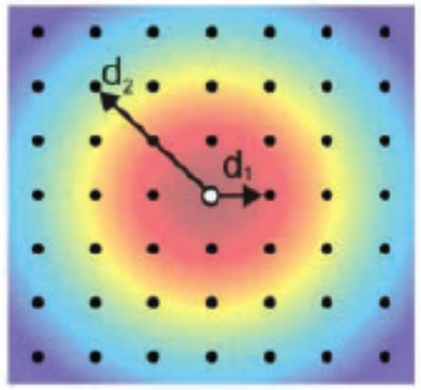
- In analyzing the neural-connectivity map of C. elegans, Alon and co-workers generated randomized networks in which the probability of two neurons connecting is completely independent of their relative positions in the network

- However, in reality, two neighboring neurons have a greater chance of forming a connection than two distant neurons at opposite ends of the network

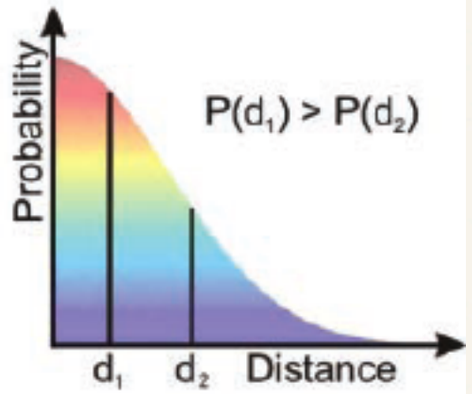
- Therefore, the test performed by Alon and co-worker was not null to this form of localized aggregation and would misclassify a completely random but spatially clustered network as one that is nonrandom and that has significant network motifs

- In this case, a random geometric graph is more appropriate

A



B



C

Motif	Feed-forward	Bi-fan	Bi-parallel
Observed	42	347	65
Mean	27.1	18.83	43.32
Std.	5.33	8.63	6.4
Z score	2.79	38.02	3.38

Fig. 1. (**A**) Construction of Gaussian "toy network." We used a 30 by 30 grid of 900 nodes. Edges were added on the basis that the probability P of two nodes being connected reduces with the distance d between them. Thus, $P(d_1) > P(d_2)$ when $d_1 < d_2$. This feature will be present to some degree in neural networks such as that of *C. elegans* (*14*). (**B**) Color-coded probability $P(d)$ of connecting to a node as a function of distance for the Gaussian toy network. (**C**) Overrepresentation of motif patterns in the Gaussian toy network. We focused on three motif patterns (feedforward, bi-fan, and bi-parallel) found in (*1*) to be significantly overrepresented in the *C. elegans* neural map. The observed number of each motif, as counted in the Gaussian toy network of (A), was compared with the mean number of motifs counted in 2000 randomized networks (*14*). For all three cases, the Z scores $\left(\frac{\text{Observed} - \text{Mean}}{\text{Std.}}\right)$ were larger than 2, signifying that the null hypothesis can be rejected and all motifs are significantly overrepresented.

- The issue of null models hold also for regulatory networks...

The evolution of genetic networks by non-adaptive processes

Michael Lynch

Abstract | Although numerous investigators assume that the global features of genetic networks are moulded by natural selection, there has been no formal demonstration of the adaptive origin of any genetic network. This Analysis shows that many of the qualitative features of known transcriptional networks can arise readily through the non-adaptive processes of genetic drift, mutation and recombination, raising questions about whether natural selection is necessary or even sufficient for the origin of many aspects of gene-network topologies. The widespread reliance on computational procedures that are devoid of population-genetic details to generate hypotheses for the evolution of network configurations seems to be unjustified.

Neutral forces acting on intragenomic variability shape the *Escherichia coli* regulatory network topology

Troy Ruths[1] and Luay Nakhleh[1]

Department of Computer Science, Rice University, Houston, TX 77251

Edited by Sean B. Carroll, University of Wisconsin, Madison, WI, and approved March 27, 2013 (received for review October 9, 2012)

Cis-regulatory networks (CRNs) play a central role in cellular decision making. Like every other biological system, CRNs undergo evolution, which shapes their properties by a combination of adaptive and nonadaptive evolutionary forces. Teasing apart these forces is an important step toward functional analyses of the different components of CRNs, designing regulatory perturbation experiments, and constructing synthetic networks. Although tests of neutrality and selection based on molecular sequence data exist, no such tests are currently available based on CRNs. In this work, we present a unique genotype model of CRNs that is grounded in a genomic context and demonstrate its use in identifying portions of the CRN with properties explainable by neutral evolutionary forces at the system, subsystem, and operon levels. We leverage our model against experimentally derived data from *Escherichia coli*. The results of this analysis show statistically significant and substantial neutral trends in properties previously identified as adaptive in origin—degree distribution, clustering coefficient, and motifs—within the *E. coli* CRN. Our model captures the tightly coupled genome–interactome of an organism and enables analyses of how evolutionary events acting at the genome level, such as mutation, and at the population level, such as genetic drift, give rise to neutral patterns that we can quantify in CRNs.

Efficient Sampling in Networks

The Issue

- Identifying network motifs requires computing subgraph concentrations
- The number of subgraphs grows exponentially with their number of nodes
- Hence, exhaustive enumeration of all subgraphs and computing their concentrations are infeasible for large networks
- In this part, we describe *mfinder*, an efficient method for estimating subgraph concentrations and detecting network motifs

Subgraph Concentrations

- Let N_i be the number of appearances of subgraphs of type i

- The concentration of n-node subgraphs of type i is the ratio between their number of appearances and the total number of n-node connected subgraphs in the network:

$$C_i = \frac{N_i}{\sum_i N_i}$$

Subgraphs Sampling

- The algorithm samples n-node subgraphs by picking random connected edges until a set of n nodes is reached

Definitions: E_S is the set of picked edges.

V_S is the set of all nodes that are touched by the edges in E_S.

Init V_S and E_S to be empty sets.

1. Pick a random edge $e_1 = (v_i, v_j)$. Update $E_S = \{e_1\}, V_S = \{v_i, v_j\}$
2. Make a list L of all neighboring edges of E_S.

   Omit from L all edges between members of V_S. If L is empty return to 1.
3. Pick a random edge $e = (v_k, v_l)$ from L.

   Update $E_S = E_S \cup \{e\}, V_S = V_S \cup \{v_k, v_l\}$
4. Repeat steps 2–3 until completing *n*-node subgraph *S*.
5. Calculate the probability *P* to sample *S*.

Sampling Probability

To sample an n-node subgraph, an ordered set of n-1 edges is iteratively randomly picked. In order to compute the probability, P, of sampling the subgraph, we need to check all such possible ordered sets of n-1 edges [denoted as (n-1)-permutations] that could lead to sampling of the subgraph

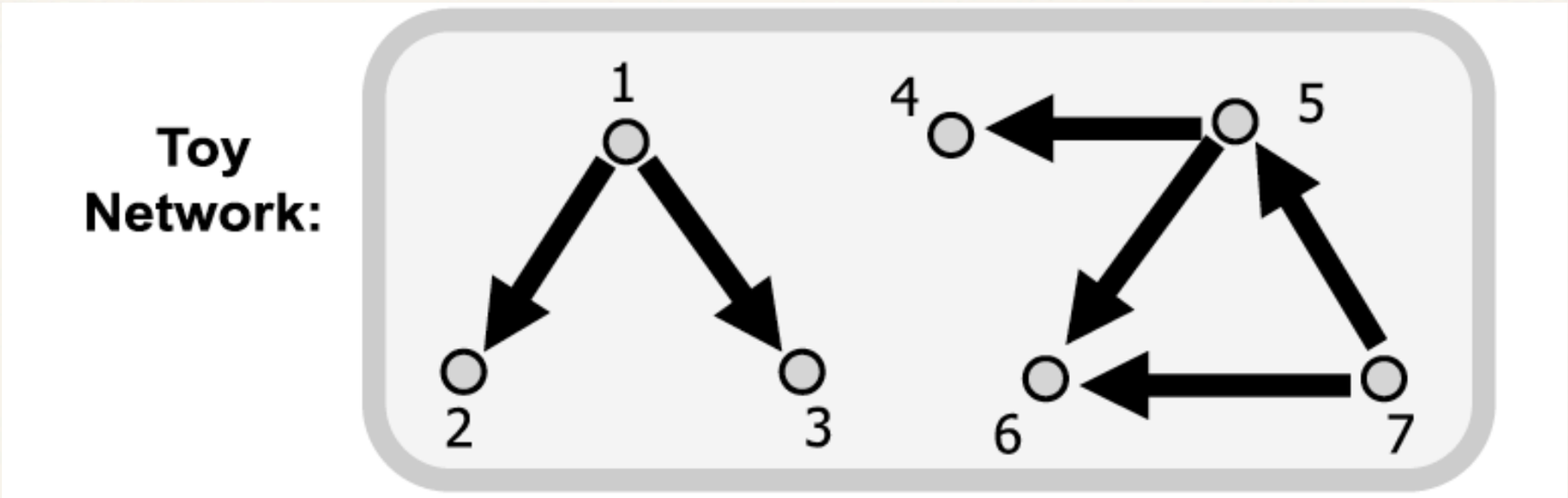
The probability of sampling the subgraph is the sum of the probabilities of all such possible ordered sets of n-1 edges:

$$P = \sum_{\sigma \in S_m} \prod_{E_j \in \sigma} \mathbf{Pr}[E_j = e_j | (E_1, \ldots, E_{j-1}) = (e_1, \ldots, e_{j-1})]$$

where S_m is the set of all (n-1)-permutations of the edges from the specific subgraph edges that could lead to a sample of the subgraph. E_j is the *j*-th edges in a specific (n-1)-permutation (σ)

Correction for Non-uniform Sampling

- Different probabilities of sampling different subgraphs



Probability to sample {1,2,3}:
There are 2 possibilities to sample {1,2,3}:
1. Pick first (1,2): Pr=1/E=1/6.
   then pick (1,3): Pr=1.
   Pr[(1,2) then (1,3)]=1/6*1=1/6.
2. Pick first (1,3):Pr=1/E=1/6.
   then pick (1,2): Pr=1.
   Pr[(1,3) then (1,2)]=1/6*1=1/6.
In Total: Pr[{1,2,3}] = 1/6 + 1/6 = 1/3=12/36

Probability to sample {4,5,6}:
There are 2 possibilities to sample {4,5,6}:
1. Pick first (5,4): Pr=1/E=1/6.
   then pick (5,6): Pr=1/2.
   Pr[(5,4) then (5,6)]=1/6*1/2=1/12
2. Pick first (5,6): Pr=1/E=1/6.
   then pick (5,4): Pr=1/3.
   Pr[(5,6) then (5,4)]=1/6*1/3=1/18.
In Total: Pr[{4,5,6}] = 1/12 + 1/18 = 5/36

After each sample, a weighted score of W=1/P is added to the score of the relevant subgraph type

Calculating the Concentrations of n-node Subgraphs

- Define score S_i for each subgraph of type i
- Initialize S_i to 0 for all i
- For every sample, add the weighted score W=1/P to the accumulated score S_i of the relevant type i
- After S_T samples, assuming we sampled L different subgraph types, calculate the estimated subgraph concentrations:

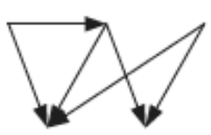
$$C_i = \frac{S_i}{\sum_{k=1}^{L} S_k}$$

Table 1. Sampling method versus exhaustive enumeration on a WWW network

Subgraph		Exhaustive enumeration Total no. of subgraphs 287M (runtime: 2.9 h)		Sampling method No. of samples 5K (runtime: 15 s)	No. of samples 50K (runtime: 37 s)	No. of samples 2.5M (runtime: 28 min)
ID		Appearances	Concentration ($\times 10^{-3}$)	Concentration ($\times 10^{-3}$)	Concentration ($\times 10^{-3}$)	Concentration ($\times 10^{-3}$)
6		47 015 127	163.8	181.2	168.4	162.7
12		2 319 911	8.1	10.3	6.7	8.2
14		1 363 964	4.8	6.0	4.9	4.8
36		218 449 147	761.0	732.2	754.8	762.2
38*		499 763	1.74	1.97	1.75	1.73
46*		1 164 456	4.1	4.9	4.1	4.1
74		4 049 373	14.1	17.4	15.7	13.9
78		4 954 123	17.3	18.5	17.7	17.2
98		9474	0.030	0.006	0.048	0.030
102		40 607	0.14	0.08	0.16	0.14
108*		309 167	1.08	1.08	1.08	1.08
110*		106 614	0.37	0.51	0.37	0.37
238*		6 779 926	23.6	25.9	24.2	23.5

Accuracy

Table 2. Subgraphs of size 3–5 in the transcriptional regulation network of *E.coli*

Subgraph size	Subgraph ID	Shape	Full enumeration Appearances (Z-score)	Concentration ($\times 10^{-3}$)	Sampling method Concentration ($\times 10^{-3}$) (Z-score)	No. of samples
3	S1		4777	917.60	916.60	1K (~5K total three-node subgraphs)
	S2		160	30.73	31.13	
	S3		227	43.60	43.64	
	M4		42 ($z = 10$)	8.07	8.69 ($z = 10$)	
4	M5		209 ($z = 9$)	2.49	2.69 ($z = 8$)	10K (~85K total four-node subgraphs)
	M6		51 ($z = 15$)	0.61	0.65 ($z = 15$)	
5	M7		54 ($z = 120$)	0.038	0.035 ($z = 30$)	50K (~1.4M total five-node subgraphs)
	M8		271 ($z = 16$)	0.189	0.196 ($z = 11$)	
	M9		20 ($z = 18$)	0.014	0.013 ($z = 8$)	
	M10		18 ($z = 12$)	0.013	0.014 ($z = 8$)	

Running Time

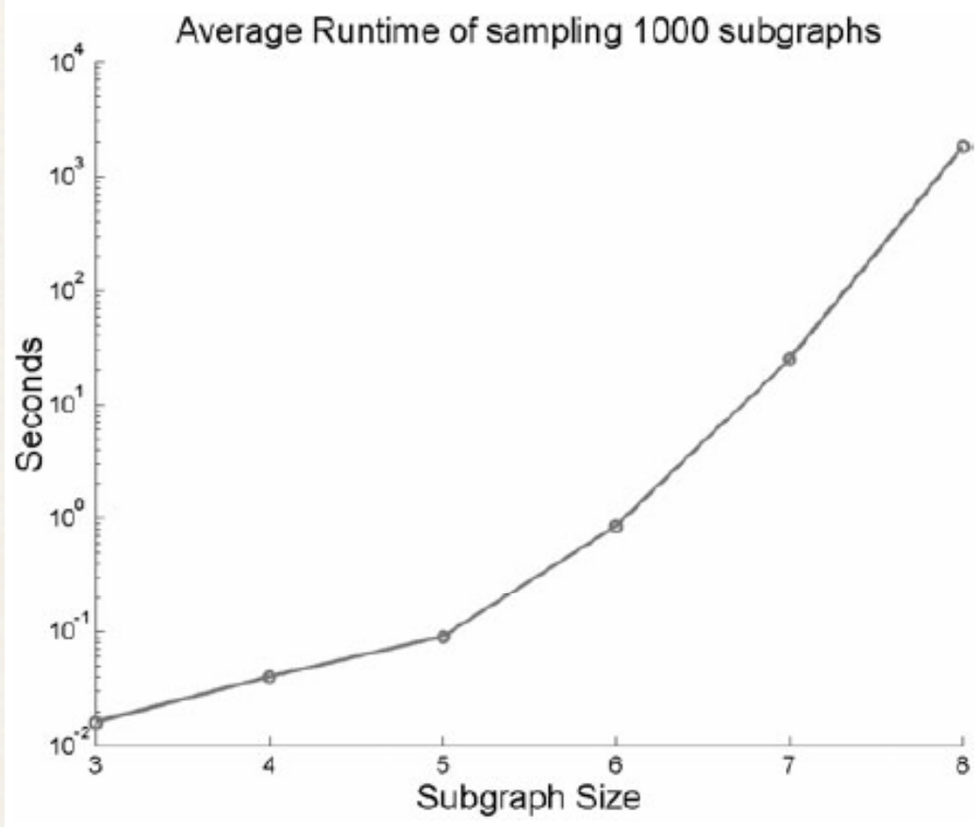


Fig. 3. Runtime per 1000 samples for different subgraph sizes: three-node up to eight-node subgraphs (semi-log scale). The network analyzed is the transcriptional regulation of *E.coli* (Shen-Orr *et al.*, 2002). The scaling of the runtime of the sampling method qualitatively agrees with the theoretical analysis of $O(K^{n-1}n^{n+1})$, where n is the subgraph size.

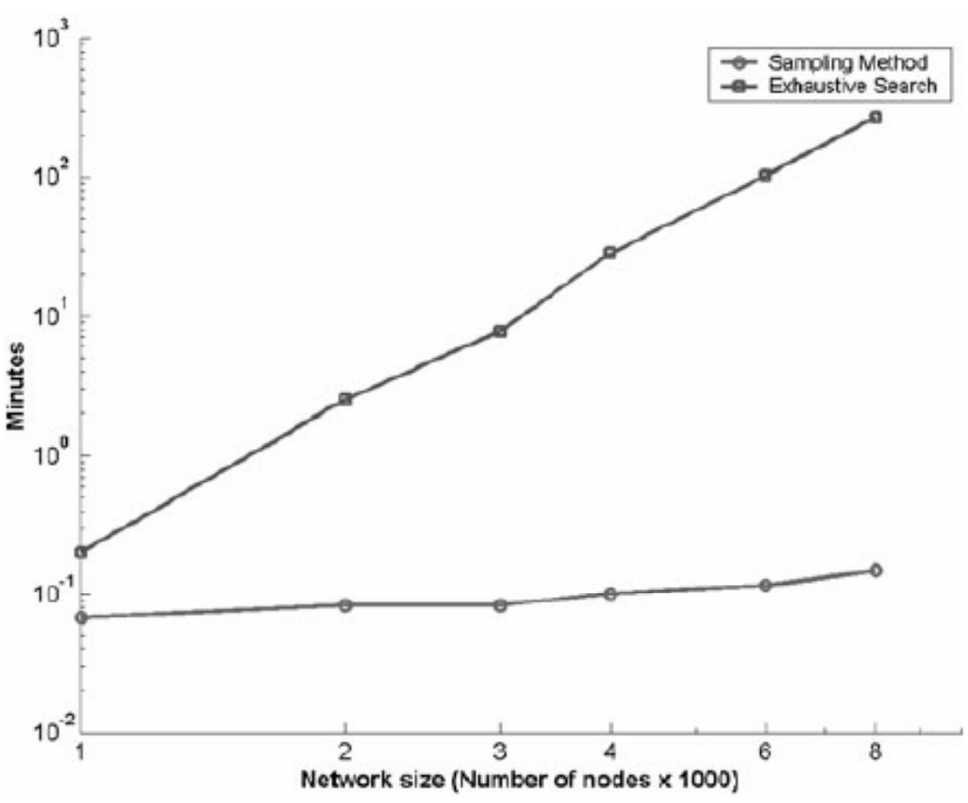


Fig. 4. Runtime of the sampling method versus an exhaustive enumeration as a function of network size (log–log scale). The networks are synthetic scale-free networks ($\gamma = 2$) with equal average connectivity ($\langle d \rangle = 2.4$). The hub degree is 10% of the total number of nodes. The sampling method was run with 100 000 samples for all the networks. The runtime of the exhaustive enumeration scales as the total number of subgraphs, while the runtime of the sampling method is almost constant.

Convergence

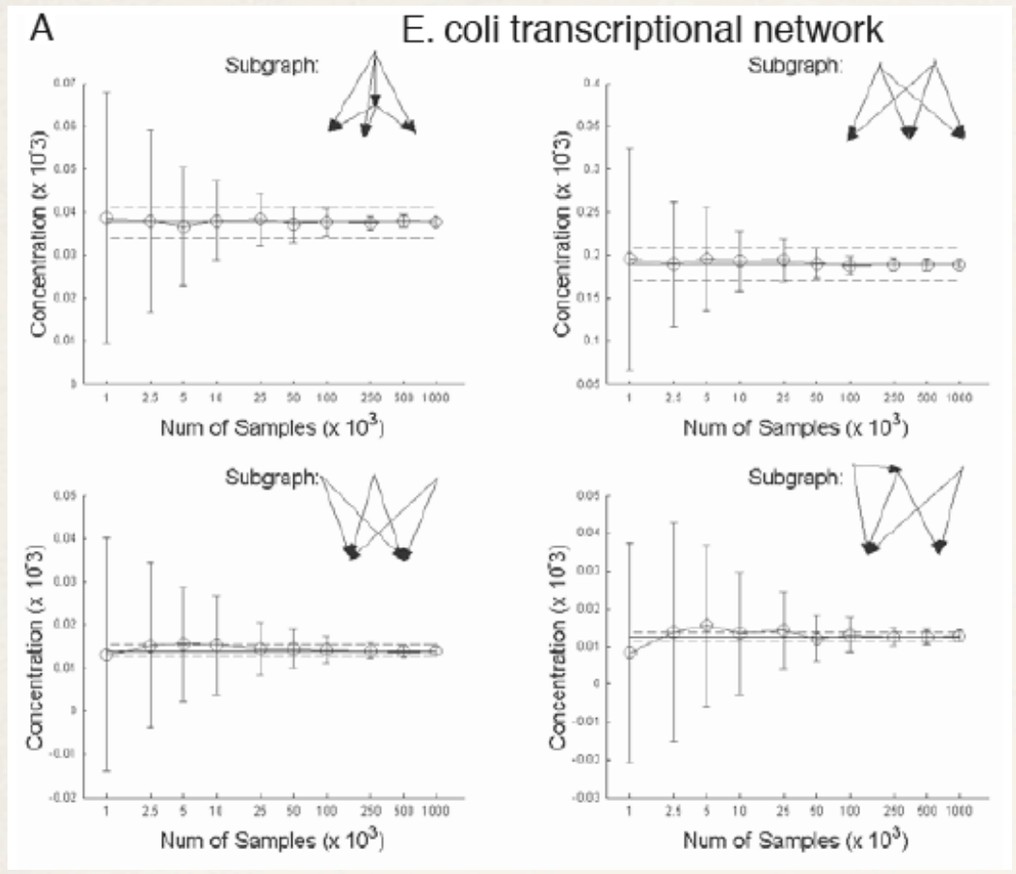


Fig. 5. Convergence of the sampling method results on different networks. Concentrations calculated by the sampling method for different subgraphs on different networks as a function of number of samples. The true concentration was found by exhaustive enumeration (horizontal full line). We ran the algorithm 100 times for each number of samples (S_T) on each of the networks. The average concentration (circles) and standard deviation are shown. Real concentrations $\pm 10\%$ are shown by dashed lines. It can be seen that the algorithm results on all four networks, for all subgraphs, converge to the true concentrations. (**A**) Transcription network of *E.coli*. All the four five-node subgraphs were found as network motifs. Despite the low concentration of the subgraphs, they are estimated accurately with a small error ratio even with relatively few samples (10^5). The total number of connected five-node subgraphs in the network is 1.43×10^6. (**B**) Transcription network of yeast (*Saccharomyces cerevisiae*). Three of the subgraphs (all but the bottom right subgraph) are found to be network motifs. Results of a high concentration subgraph (bottom right) also converge rapidly to the real concentration. The total number of connected five-nodes subgraphs in the network is 2.5×10^6. (**C**) Neuronal network of *C.elegans*. All the four four-node subgraphs were found as network motifs. This network is characterized by relative high density (average degree $= 15.5$). The total number of connected four-node subgraphs is 8.75×10^5. (**D**) Ythan Estuary food web. All the four five-node subgraphs were detected as network motifs. Total number of connected five-node subgraphs is 9.4×10^5.

How Many Samples Are Enough?

- It is a hard problem
- Further, the number of samples required for good estimation with a high probability is hard to approximate when the concentration distribution is not known a priori
- Alon and co-workers used an approach similar to adaptive sampling
- Let $V_i = (\hat{c}_1^i, \hat{c}_2^i, \ldots, \hat{c}_k^i)$ and $V_{i-1} = (\hat{c}_1^{i-1}, \hat{c}_2^{i-1}, \ldots, \hat{c}_k^{i-1})$ be the vectors of estimated subgraphs concentration after the iterations i and i-1, respectively. The average instantaneous convergence rate is

$$\mathrm{CG_{avg}} = \frac{1}{k}\sum_{j=1}^{k}\frac{\left|\hat{c}_j^i - \hat{c}_j^{i-1}\right|}{0.5(\hat{c}_j^i + \hat{c}_j^{i-1})}(\forall \hat{c}_j^i > C_{\min})$$

and the maximal instantaneous convergence rate is

$$\mathrm{CG_{max}} = \max_j \left\{ \frac{\left|\hat{c}_j^i - \hat{c}_j^{i-1}\right|}{0.5(\hat{c}_j^i + \hat{c}_j^{i-1})} \middle| \forall \hat{c}_j^i > C_{\min} \right\}$$

By setting the thresholds CG_{avg}, CG_{max} and the value of C_{min}, the required accuracy of the results and the minimum concentration of subgraphs can be adjusted

Comparing the Local Structure of Networks

- To understand the design principles of complex networks, it is important to compare the local structure of networks from different fields

- The main difficulty is that these networks can be of vastly different sizes

- In this part, we introduce an approach for comparing network local structure based on the significance profile (SP)

Significance Profile

- For each subgraph i, the statistical significance is described by the Z score:

$$Z_i = \frac{Nreal_i - \langle Nrand_i \rangle}{\mathrm{std}(Nrand_i)}$$

where

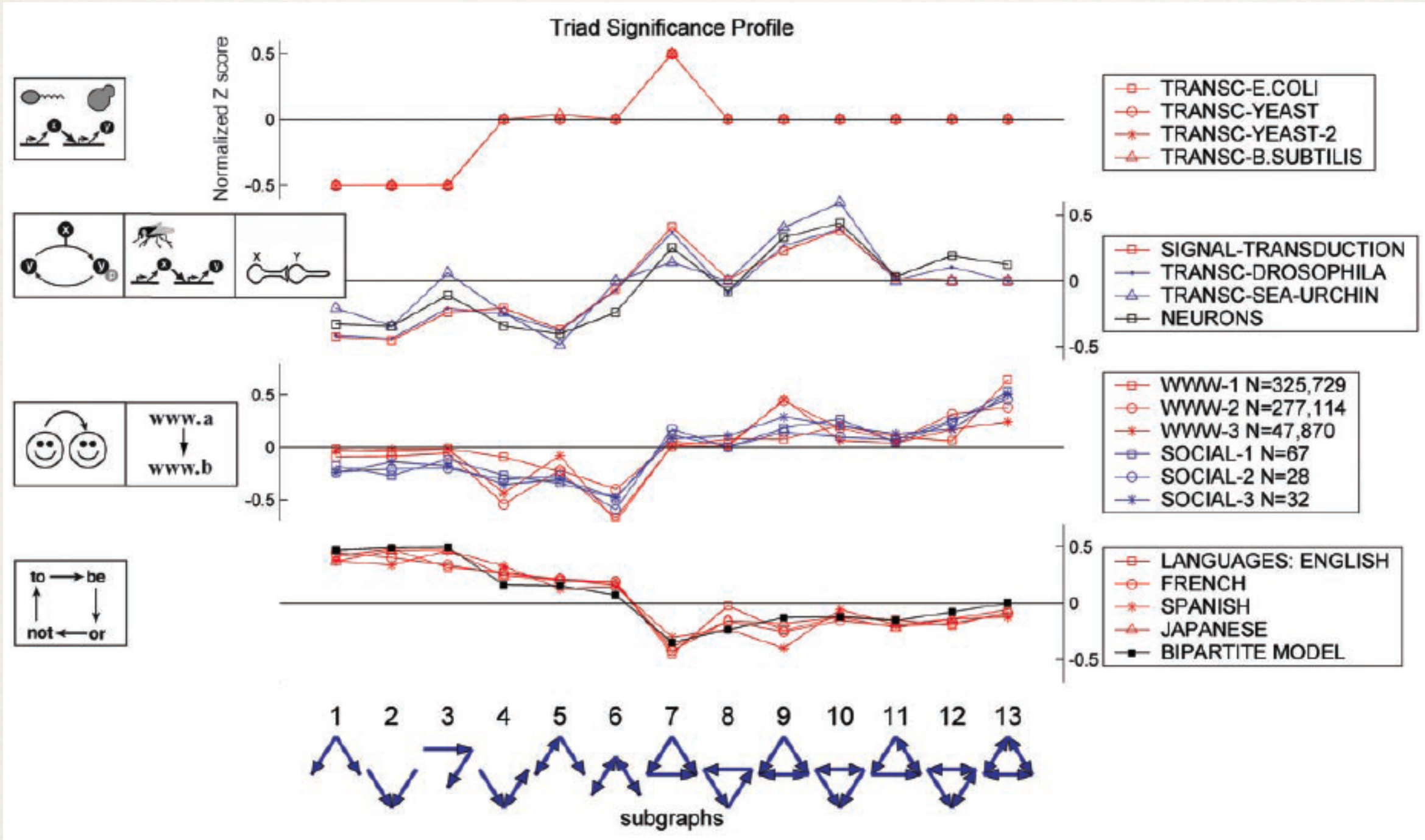
$Nreal_i$ is the number of times subgraph type i appears in the network

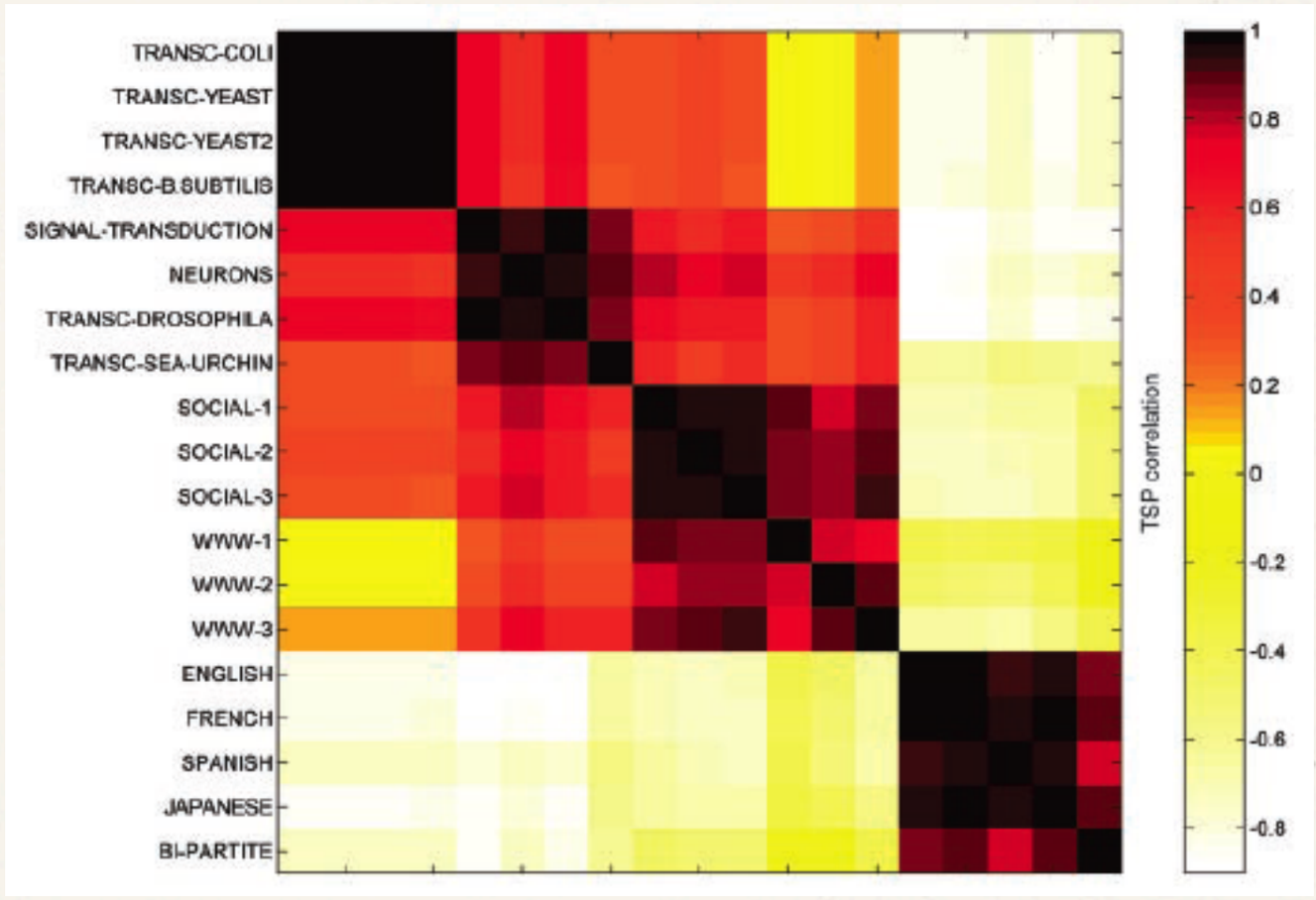
$\langle Nrand_i \rangle$ is the mean of its appearances in the randomized network ensemble

$\mathrm{std}(Nrand_i)$ is the standard deviation of its appearances in the randomized network ensemble

- The SP is the vector of Z scores normalized to length 1:

$$SP_i = \frac{Z_i}{\sqrt{\sum_i Z_i^2}}$$


Triad Significance Profile
Normalized Z score
0.5
0
-0.5
TRANSC-E.COLI
TRANSC-YEAST
TRANSC-YEAST-2
TRANSC-B.SUBTILIS
SIGNAL-TRANSDUCTION
TRANSC-DROSOPHILA
TRANSC-SEA-URCHIN
NEURONS
www.a
www.b
WWW-1 N=325,729
WWW-2 N=277,114
WWW-3 N=47,870
SOCIAL-1 N=67
SOCIAL-2 N=28
SOCIAL-3 N=32
to be
not or
LANGUAGES: ENGLISH
FRENCH
SPANISH
JAPANESE
BIPARTITE MODEL
1 2 3 4 5 6 7 8 9 10 11 12 13
subgraphs



The correlation coefficient matrix of the triad significance profiles for the directed networks on the previous slide

The Subgraph Ratio Profile (SRP)

- When analyzing subgraphs (particularly 4-node subgraphs) in undirected graphs, the normalized Z scores of the subgraphs showed a significant dependence on the network size

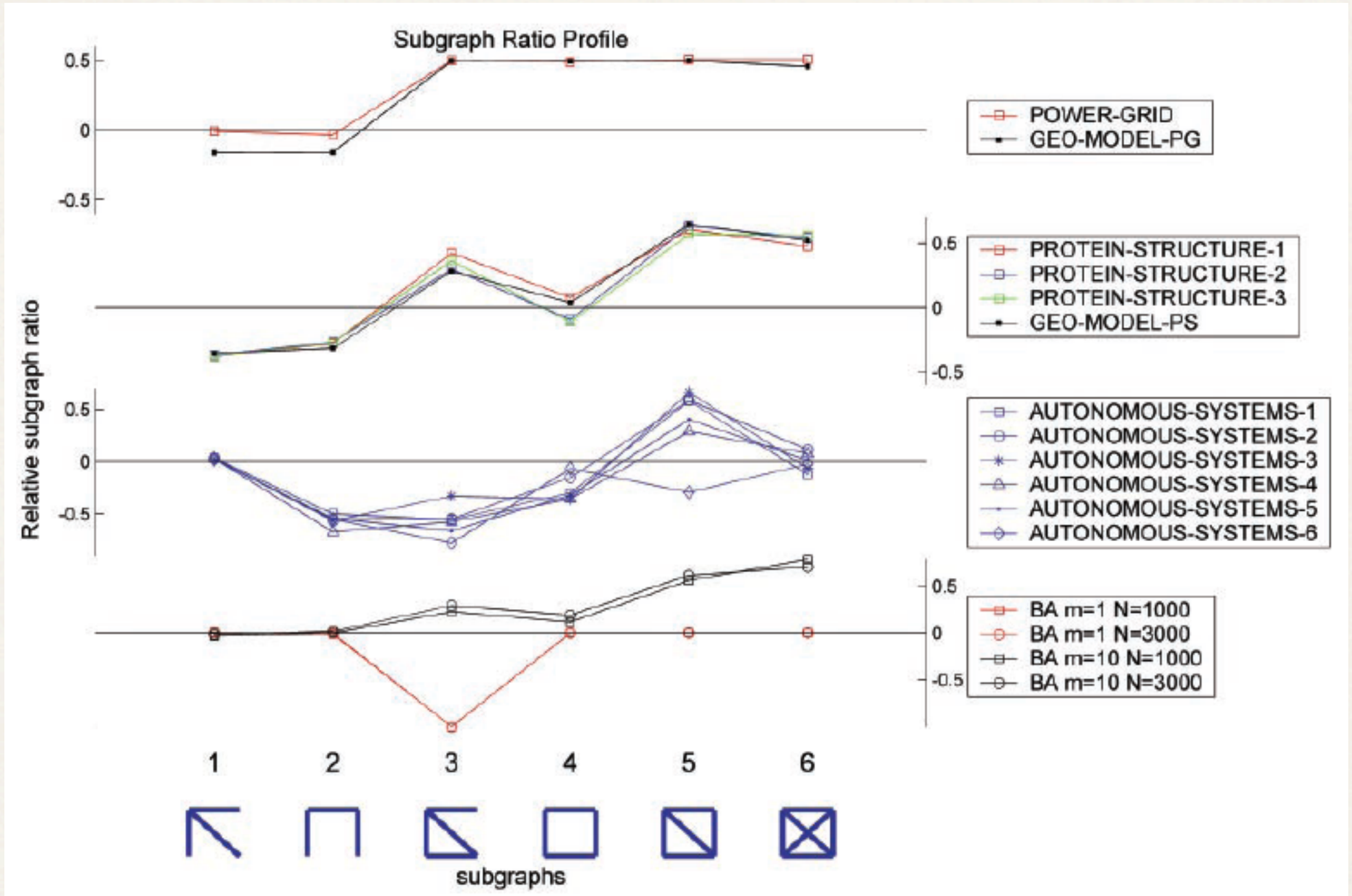
- In undirected networks, an alternative measure is the SRP

$$\Delta_i = \frac{Nreal_i - \langle Nrand_i \rangle}{Nreal_i + \langle Nrand_i \rangle + \varepsilon}$$

where ε ensures that $|\Delta|$ is not misleadingly large when the subgraph appears very few times in both the real and random networks

- The SRP is the vector of Δ_i scores normalized to length 1:

$$SRP_i = \frac{\Delta_i}{\sqrt{\sum_i \Delta_i^2}}$$


Subgraph Ratio Profile
Relative subgraph ratio
0.5
0
-0.5
POWER-GRID
GEO-MODEL-PG
PROTEIN-STRUCTURE-1
PROTEIN-STRUCTURE-2
PROTEIN-STRUCTURE-3
GEO-MODEL-PS
AUTONOMOUS-SYSTEMS-1
AUTONOMOUS-SYSTEMS-2
AUTONOMOUS-SYSTEMS-3
AUTONOMOUS-SYSTEMS-4
AUTONOMOUS-SYSTEMS-5
AUTONOMOUS-SYSTEMS-6
BA m=1 N=1000
BA m=1 N=3000
BA m=10 N=1000
BA m=10 N=3000
1
2
3
4
5
6
subgraphs

Motif Evolution

Motif Conservation

- Wuchty et al. recently showed that in S. cerevisiae, proteins organized in cohesive patterns of interactions are conserved to a substantially higher degree than those that do not participate in such motifs.

- They found that the conservation of proteins in distinct topological motifs correlates with the interconnectedness and function of that motif and also depends on the structure of the overall interactome topology.

- These findings indicate that motifs may represent evolutionary conserved topological units of cellular networks molded in accordance with the specific biological function in which they participate.

Experimental Setup

- Test the correlation between a protein's evolutionary rate and the structure of the motif it is embedded in

- Hypothesis: if there is evolutionary pressure to maintain specific motifs, their components should be evolutionarily conserved and have identifiable orthologs in other organisms

- They studied the conservation of 678 S. cerevisiae proteins with an ortholog in each of five higher eukaryotes (Arabidopsis thaliana, C. elegans, D. melanogaster, Mus musculus, and Homo sapiens) from the InParanoid database

Table 1 Evolutionary conservation of motif constituents

#	Motifs	Number of yeast motifs	Natural conservation rate	Random conservation rate	Conservation ratio
1		9,266	13.67%	4.63%	2.94
2		167,304	4.99%	0.81%	6.15
3		3,846	20.51%	1.01%	20.28
4		3,649,591	0.73%	0.12%	5.87
5		1,763,891	2.64%	0.18%	14.67
6		9,646	6.71%	0.17%	40.44
7		164,075	7.67%	0.17%	45.56
8		12,423	18.68%	0.12%	157.89
9		2,339	32.53%	0.08%	422.78
10		25,749	14.77%	0.05%	279.71
11		1,433	47.24%	0.02%	2,256.67

The third column gives the number of motifs of a given kind found in the yeast protein interaction network of 3,183 proteins, which we obtained by counting all subgraphs of two-node to five-node motifs (from the set of 28 five-node motifs, we show only two, #10 and #11). We identified 678 proteins that have an ortholog in each of the five higher eukaryotes that we studied and identified all motifs for which each component belongs to this evolutionary conserved protein subset. The natural conservation rate indicates the fraction of the original yeast motifs that is evolutionarily fully conserved, meaning that each of their protein components belongs to the 678 orthologs of the list. For example, we find that 47% of the 1,433 fully connected pentagons (#11) found in yeast have each of their five proteins conserved in each of the five higher eukaryotes. If the topology of motifs does not interfere with the conservation rate of its constituting proteins, a random ortholog distribution should give the same conservation rate for specific motifs as seen in the natural sample. The random conservation rate therefore represents the fraction of motifs that is fully conserved for the random ortholog distribution. The last column gives the ratio between the natural and the random conservation ratios, indicating that all motifs are highly conserved, some (for example, #11) having a natural conservation rate 2,256 times higher than expected in the absence of correlations between protein conservation rate and the topology of a given motif.

Table 2 Overrepresentation of human orthologous motifs in various functional classes of yeast proteins

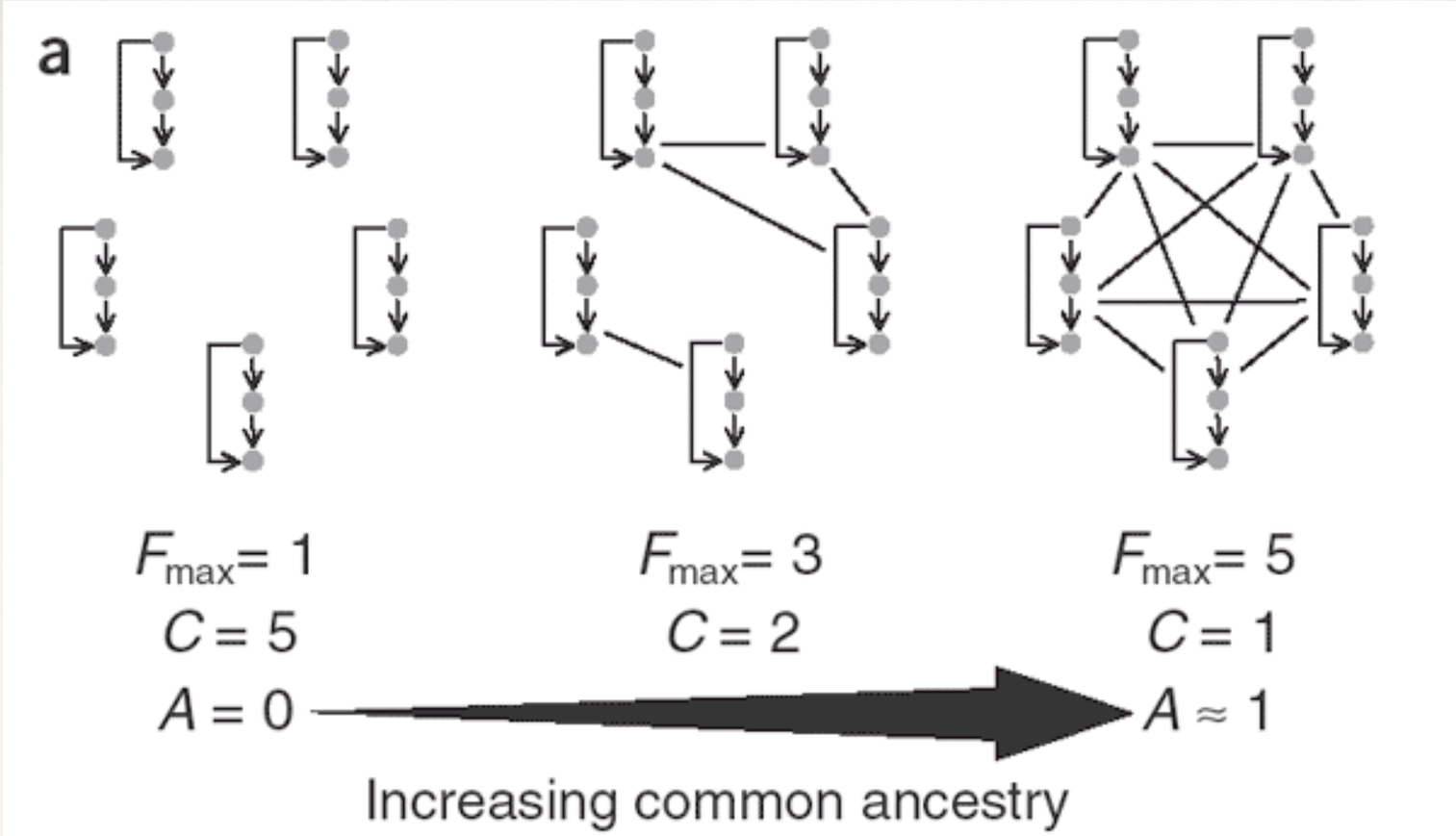
Functional class	Overrepresented motifs
Transport facilitation	 (10)
Subcellular localization	 (21)  (21)  (26)  (15)  (27)  (23)  (29)  (20)  (63)  (45)
Regulation	 (10)
Protein fate	 (14)  (16)  (13)  (33)  (27)  (20)  (26)  (24)  (16)  (60)  (41)
Cell cycle	 (11)  (14)  (13)  (11)  (14)
Cellular transport	 (11)  (12)
Transcription	 (12)  (16)  (17)  (13)  (16)  (19)  (17)  (15)  (14)  (21) (23)
Protein synthesis	(12) (11) (17) (11) (24)

We determined the number of motifs for the subnetworks defined by proteins belonging to a specific functional class, as well as the number of these motifs (μ_h) that are fully conserved in humans. Finally, for 100 randomized human orthologous sets we determined the average number of motifs (μ_r) in the random ortholog samples and the standard deviation (σ_r) for each motif. The table lists all motifs that are overrepresented by a factor of at least ten compared with a random configuration ($Z > 10$), with the specific Z values shown next to the motifs. We did not find overrepresented motifs for the classes of transposable elements, energy, cellular fate, cellular communication, cellular rescue, cellular organization, metabolism, protein activity, protein binding and proteins that are not yet classified or that are classified unclearly. If all proteins of a given motif simultaneously belong to more than one functional class, the motif will also appear in multiple functional classes.

Convergent Evolution

- Convergent evolution is a potent indicator of optimal design
- Conant and Wagner recently showed that multiple types of trascriptional regulation circuitry in E. coli and S. cerevisiae have evolved independently and not by duplication of one or a few ancestral circuits

(**a**) Two indicators of common ancestry for gene circuits. Each of $n = 5$ circuits of a given type (a feed-forward loop for illustration) is represented as a node in a circuit graph. Nodes are connected if they are derived from a common ancestor, that is, if all k pairs of genes in the two circuits are pairs of duplicate genes. $A = 0$ if no circuits share a common ancestor (the graph has n isolated vertices); A 1 if all circuits share one common ancestor (the graph is fully connected). The number C of connected components indicates the number of common ancestors (two in the middle panel) from which the n circuits derive. F_{max} is the size of the largest family of circuits with a single common ancestor (the graph's largest component). (**b**) Little common ancestry in six circuit types. We considered two circuits to be related by common ancestry if each pair of genes at corresponding positions in the circuit had significant sequence similarity. Each row of the table shows values of C, A and F_{max} for a given circuit type, followed in parentheses by their average values standard deviations and P values



b

	Circuit type	Number of circuits	Number of families (C)	Index of common ancestry (A)	Largest circuit family (F_{max})
Yeast	Feed-forward	48	44 (46.8 ± 1.9; $P = 0.08$)	0.082 (0.023 ± 0.035; $P = 0.08$)	5 (1.9 ± 1.4; $P = 0.05$)
	Bi-fan	542	435 (469.0 ± 37.7; $P = 0.18$)	0.197 (0.135 ± 0.070; $P = 0.18$)	49 (41.0 ± 31.1; $P = 0.33$)
	MIM-2	176	168 (164.5 ± 8.8; $P = 0.60$)	0.045 (0.065 ± 0.050; P =0.60)	5 (7.4 ± 6.2; $P = 0.59$)
	Reg. chain (3)	33	33	0	1
E. coli	Feed-forward	11	11	0	1
	Bi-fan	27	27	0	1

$$A = 1-(C/n)$$

n: number of circuits (nodes in the graph)

C: number of components in the graph

F_{max} is size of largest family

Table 1 Gene families are not over-represented in circuit types

Organism	Circuit type	P_{motif}[a]	$P_{motif|duplicate}$[b]	P[c]
S. cerevisiae	Bi-fan	0.82	0.80	NA
	Feed-forward	0.38	0.42	0.21
	Multi-input motif	0.77	0.76	NA
	Regulator chains	0.64	0.67	0.30
E. coli	Bi-fan	0.50	0.67	0.11
	Feed-forward	0.82	0.67	NA

[a]Probability of a randomly chosen regulatory gene occurring in a given circuit type. [b]Probability of a regulatory gene occurring in a circuit type given that one of its duplicates occurs in that circuit type (see **Supplementary Methods** online). [c]P value for one-sided exact binomial test of the null hypothesis $P_{motif} = P_{motif\ |\ duplicate}$. NA indicates that a test has not been carried because $P_{motif} > P_{motif\ |\ duplicate}$. The number of transcriptional regulators was $n = 112$ and $n = 22$ for the yeast and *E. coli* analyses, respectively.

A Textbook Focused on Network Motifs

- "An Introduction to Systems Biology: Design Principles of Biological Circuits", by Uri Alon, Chapman & Hall/CRC, 2007.

Acknowledgments

- Materials in this lecture are mostly based on:
  - "Superfamilies of evolved and designed networks", by Milo et al.
  - "Network motifs: simple building blocks of complex networks", by Milo et al.
  - A comment on the above two by Artzy-Randrup et al.
  - "Network motifs in the transcriptional regulation network of Escherichia coli", by Shen-Orr et al.
  - "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs", by Kashtan et al.
  - "Convergent evolution of gene circuits", by Conant and Wagner.
  - "Evolutionary conservation of motif constituents in the yeast protein interaction network", by Wuchty et al.