

# Network motifs in biological networks: Roles and Generalizations

N. Kashtan<sup>2,3</sup>, S. Itzkovitz<sup>1,2</sup>, R. Milo<sup>1,2</sup>, U. Alon<sup>1,2</sup>

<sup>1</sup>*Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel 76100*

<sup>2</sup>*Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel 76100*

<sup>3</sup>*Department of Computer Science and Applied Mathematics,  
Weizmann Institute of Science, Rehovot, Israel 76100*

Biological and technological networks contain patterns, termed network motifs, which occur far more often than in randomized networks. Network motifs were suggested to be elementary building blocks that carry out key functions in the network. It is of interest to understand how network motifs are embedded in the network, and whether they combine to form larger structures. To address this, we present a systematic approach to detect 'motif generalizations': families of motifs of different sizes that share a common architectural theme. To define motif generalizations, we first define 'roles' in a subgraph according to structural equivalence. For example, the feedforward loop triad, a motif in transcription, neuronal and some electronic networks, has three roles, an input node, an output node and an internal node. The roles are used to define possible generalizations of the motif. The feedforward loop can have three simple generalizations, based on replicating each of the three roles and their connections. We present algorithms for efficiently detecting motif generalizations. We find that the transcription networks of bacteria and yeast display only one of the three generalizations, the multi-output feedforward generalization. In contrast, the neuronal network of *C. elegans* mainly displays the multi-input generalization. Forward-logic electronic circuits display a multi-input, multi-output hybrid. Thus, networks which share a common motif can have very different generalizations of that motif. We assign possible functions for each of the different motif generalizations in transcription, neuronal and electronic networks.

PACS numbers: 05, 89.75

## I. INTRODUCTION

Biological and engineered networks were recently found to contain network motifs: small subgraphs that occur in the network far more often than in randomized networks [25, 35]. Gene regulation networks [20], neuron networks, and some electronic circuits are all information processing networks that were found to share many of the same network motifs [25]. These network motifs were suggested to function as elementary building blocks of the networks. One of the motifs shared by biological information processing networks is the feedforward loop (FFL). The feedforward loop in transcriptional networks was suggested to act as a 'persistence detector' circuit that rejects transient activation signals yet allows rapid response to inactivation signals [22, 35]. It is important to understand the function of elementary circuits such as the FFL also in neuronal networks and other information processing networks [2, 8, 11, 14, 15, 24, 27, 29, 30, 34, 36, 37]. In particular it is important to understand whether each motif is an independent circuit or is a part of a larger structure. Thus, motifs with larger number of nodes need to be analyzed. Exploring large motifs is a hard computational task. The complexity of counting subgraphs originates from two sources: (a) the total number of  $n$ -node subgraphs in a network grows sharply with  $n$ , scaling as  $D^{n-1}$  where  $D$  is the degree of the hub (most connected node in the network) [17, 19]. (b) Classifying the subgraphs into types (isomorphic classes) becomes harder with subgraph size [26]. In addition the number of types of subgraphs increases with  $n$ : there are 13

connected directed subgraphs with 3-nodes, 199 with 4 nodes, 9364 with 5 nodes etc. [13]. For large  $n$ , there are no known efficient algorithms for exhaustively counting all  $n$ -node subgraphs in a given network. Even sampling algorithms such as [19] are limited for large subgraphs that appear only very few times in the network. Here, we present an approach for uniting related groups of motifs of different sizes into families termed motif generalizations. This allows generalizing from small motifs to the larger complexes in which they appear, using efficient algorithms. We find that networks that share the same motif can have different generalizations of that motif. We discuss the function of the FFL generalizations in transcription, neuronal and electronic networks.

## II. RESULTS

### A. Roles in a subgraph

We begin by defining roles of nodes in a subgraph. A group of nodes in a subgraph share the same role if there is a permutation of these nodes, together with their corresponding edges, that preserves the subgraph structure (See METHODS for formal definitions). For example, in the v-shaped subgraph in Fig. 1a, nodes b and c can be permuted leaving the structure intact, whereas nodes a and b can not. Thus, this subgraph has two roles, role 1 and role 2 (Fig. 1b). The FFL has three roles (Fig. 1c, triad 6), whereas the 3-loop (Fig 1c, triad 7) has only one role (because a cyclic permutation of the three nodes pre-

serves its structure). The thirteen triads have between one and three roles each (Fig. 1c), with a total of 30 different roles.

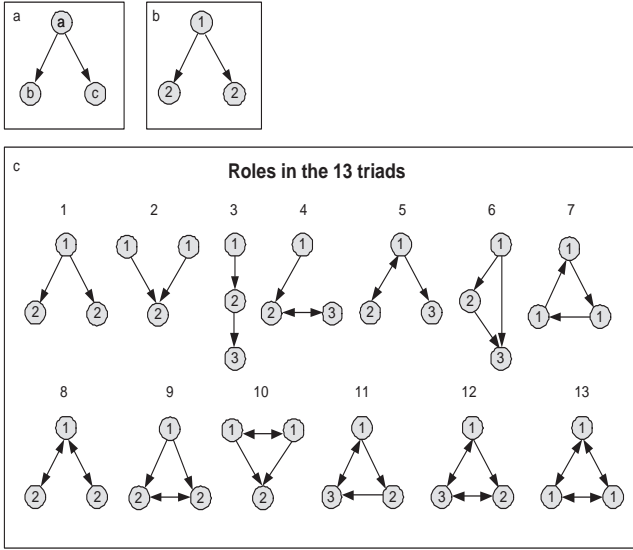


FIG. 1: **a.** A directed 3-node subgraph (triad) **b.** This triad has two roles. **c.** Roles in all the possible 13 connected triads. In each subgraph there are between one and three roles.

## B. Statistical Significance of Roles

Network motifs are detected by calculating the statistical significance of subgraphs in the network compared to suitably randomized networks [25, 35]. Here, we calculate the statistical significance of the roles in subgraphs. We counted how many times each role (for example, each of the 30 roles in 3-node subgraphs) appears in the real network ( $N_{\text{real}}$ ) and compare it to the number of times it appears in an ensemble of random networks [25] ( $N_{\text{rand}}$ ). Note that the same role is counted only once per node. In most cases the roles of network motifs are found to appear significantly more often than in random networks, whereas roles of subgraphs which are not network motifs are not significantly over-represented. We now focus on the roles of the FFL in various networks in which the FFL is a motif. The FFL has three roles, which we term input role (X), output role (Z) and internal or secondary input role (Y) (Fig. 2a). In transcription networks of *E. coli* and yeast, role X appears less often than role Y, which appears far less often than the role Z (Table 1a). The situation is reversed in the neuronal network of *C. elegans*, where X is the most abundant role and Z the rarest. In forward-logic electronic chips, Y is the least frequent role, appearing about half as often as X and Z.

The fact that the ratios of the role appearances are not 1:1:1 means that some of the FFL share nodes with each other. We calculated the mean number of FFLs

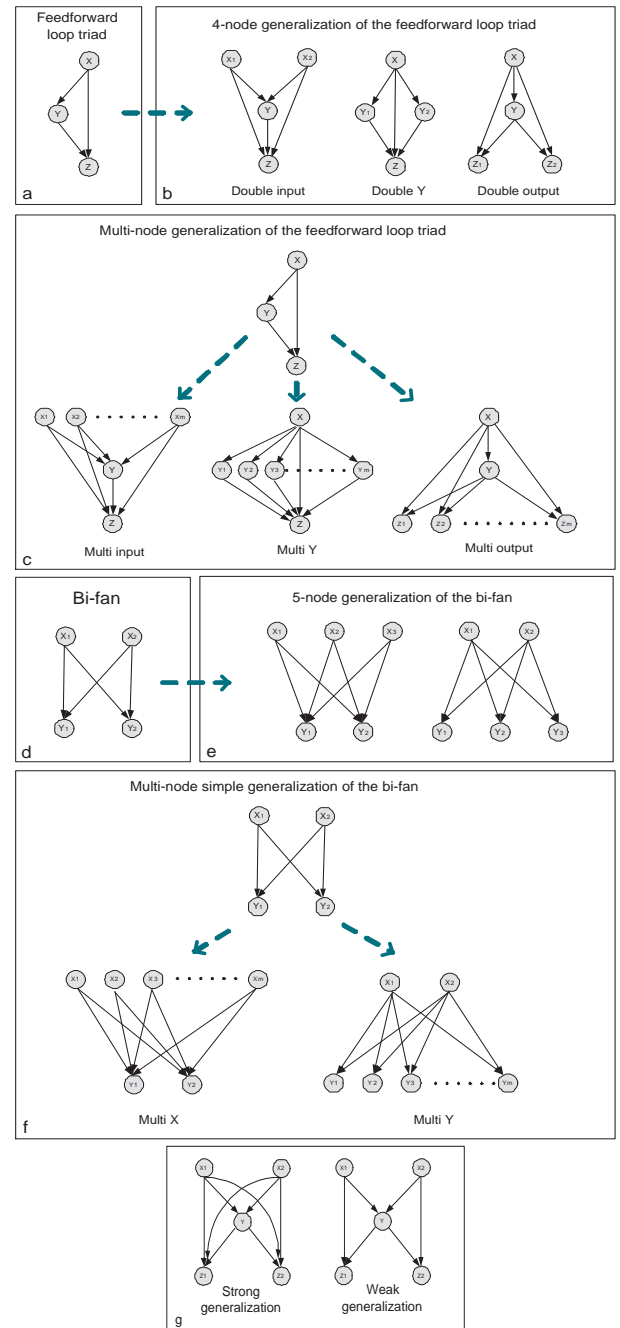


FIG. 2: **a.** The feedforward loop triad has three roles: X (input node), Y (internal - secondary input) node and Z (output node) **b.** 4-node simple generalizations of the feedforward loop. The X node is duplicated to form the double-X generalization. The Y and Z nodes are duplicated to form the double-Y and double-Z generalizations respectively. **c.** Simple multi-node generalizations of the FFL. **d.** The bi-fan, a 4-node motif with two roles X (input role) and Y (output role). **e.** 5-node simple generalizations of the bi-fan. In each of the two generalizations one of the two roles is replicated. **f.** Simple multi-node generalization of the bi-fan: an X or Y node is replicated to form the multi-input or multi-output bi-fan generalization respectively. **g.** Strong and weak generalization rules. A 5-node generalization of the FFL with two X nodes, one Y node, and two Z nodes. In the strong generalization every combination of a X,Y,Z triplet of nodes forms a FFL.

	Transcriptional networks						Neuronal networks			Electronic chips		
	E. coli			yeast			C. elegans			S15850		
$N_{FFL}$	42			62			40			424		
1a												
Role	$N_{real}$	$N_{rand}$	Z-score	$N_{real}$	$N_{rand}$	Z-score	$N_{real}$	$N_{rand}$	Z-score	$N_{real}$	$N_{rand}$	Z-score
X	<b>10</b>	4 ± 1	4	<b>15</b>	7 ± 2	4	<b>32</b>	14 ± 3	5	<b>410</b>	2 ± 2	270
Y	<b>19</b>	6 ± 2	7	<b>18</b>	8 ± 2	5	<b>24</b>	12 ± 3	4	<b>212</b>	2 ± 2	140
Z	<b>42</b>	8 ± 3	11	<b>51</b>	10 ± 3	13	<b>19</b>	13 ± 3	2	<b>424</b>	2 ± 2	270
1b												
Role	$S_{real}$	$S_{rand}$	Z-score	$S_{real}$	$S_{rand}$	Z-score	$S_{real}$	$S_{rand}$	Z-score	$S_{real}$	$S_{rand}$	Z-score
X	<b>4.2</b>	1.9 ± 1.3	2	<b>4.1</b>	1.6 ± 0.5	5	1.3	1.3 ± 0.3	0	1.0	1 ± 0	0
Y	2.2	1.4 ± 0.9	0.8	<b>3.4</b>	1.5 ± 0.5	3	1.7	1.6 ± 0.3	0	<b>2.0</b>	1 ± 0	NA
Z	1	1 ± 0.9	0	1.2	1.1 ± 0.5	0.3	<b>2.1</b>	1.4 ± 0.4	2	1.0	1 ± 0	0
1c												
Role	$N_{real}$	$N_{rand}$	Z-score	$N_{real}$	$N_{rand}$	Z-score	$N_{real}$	$N_{rand}$	Z-score	$N_{real}$	$N_{rand}$	Z-score
X	10	10 ± 2	0.1	15	11 ± 2	1.7	<b>32</b>	22 ± 3	3.8	<b>410</b>	110 ± 9	33
Y	19	17 ± 3	0.7	18	18 ± 3	0	24	18 ± 4	1.5	212	195 ± 12	1.4
Z	<b>42</b>	33 ± 2	4.1	<b>51</b>	38 ± 6	2.3	19	122 ± 3	0.9	<b>424</b>	230 ± 11	18

TABLE I: **Role statistics of the feedforward loop in different networks.** **a.** Number of appearances of each role in the real network ( $N_{real}$ ) and in randomized networks ( $N_{rand}$ : mean ± SD).  $Zscore = (N_{real} - \langle N_{rand} \rangle) / \sigma$ , where  $\sigma$  is the standard deviation in the random networks.  $N_{FFL}$  ( $= N_{FFLreal}$ ) is the number of times the feedforward loop subgraph appears in the real network **b.** Mean number of feedforward loops in which each role participates in the real network ( $S_{real} = N_{FFLreal} / N_{real}$ ) and in randomized networks ( $S_{rand} = \langle N_{FFLrand} / N_{rand} \rangle$ ). **c.** Role statistics in comparison to randomized networks constrained to have the same number of feedforward loop subgraphs as in the real network [25]. Significant roles (bold) tend to be the replicated role in the significant generalized FFL structures found in the network.

that each role participates in (Table 1b). In the *E. coli* transcription network we find that a node playing role X participates in 4.2 FFLs on average, which is significantly higher than in random networks (Table 1b). The participation of the other two roles is not highly significant. In the neuronal network, in contrast, the Z role participates in more FFLs than the Z role in randomized networks. In order to find if the tendency to share nodes is simply due to the degree distribution of the network, we compared the number of roles in the real network to an ensemble of random networks that preserves the total number of FFLs (using a simulated annealing algorithm [25, 28]). In all the networks we analyzed we find at least one role which appears more often than in FFL-preserving random networks (Table 1c). In the transcription networks of *E. coli* and yeast we find the Z role significant, in neurons the X role, and in the forward-logic chip the X and Z roles. These results suggest that in *E. coli* and yeast, FFLs tend to share X and Y nodes. In neurons the FFL tend to share Y and Z nodes, and in electronic circuits, they share Y nodes. This suggests a picture of particular kinds of complexes of FFLs in these networks. For example, in *E. coli* and yeast, the role statistics suggest structures with many Z's connected to the same X and Y nodes. In order to formalize this notion, and apply it to other significant structures, we now define subgraph generalizations based on node roles.

### C. Subgraph Generalizations

Subgraph generalizations are extensions of an n-node subgraph to a family of subgraphs with additional nodes which share its basic structure. Consider the FFL (Fig. 2a). This 3-node subgraph has three simple generalizations to the level of 4 nodes (Fig. 2b). In a simple generalization a single role and its connections are replicated. In the first simple generalization, the X role and its connections are replicated. This generalization is termed double-X FFL or double-input FFL. The other two generalizations are obtained by replicating the Y or Z roles. This replication process can be continued, leading to higher-order motif generalizations, the multi-X (multi-input), multi-Y and multi-Z (multi-output) FFL generalizations (Fig. 2c). More complex generalizations can be obtained by replicating more than one of the roles. For example, replicating both the X and Z roles yields five-node generalizations. When replicating more than one role (and in some cases replicating even a single role), one can define two kinds of generalizations: in strong generalizations, every X,Y,Z triplet forms a FFL. In weak generalizations, every node participates in at least one FFL, but not all possible FFLs are formed (Fig. 2g). This procedure of generalization can be applied to any subgraph (see formal definition in METHODS). For example simple generalizations of the 4-node bi-fan to the level of 5 nodes and above are shown in Fig. 2d-f. We now describe the statistical significance of the generalizations of the motifs found in various networks.

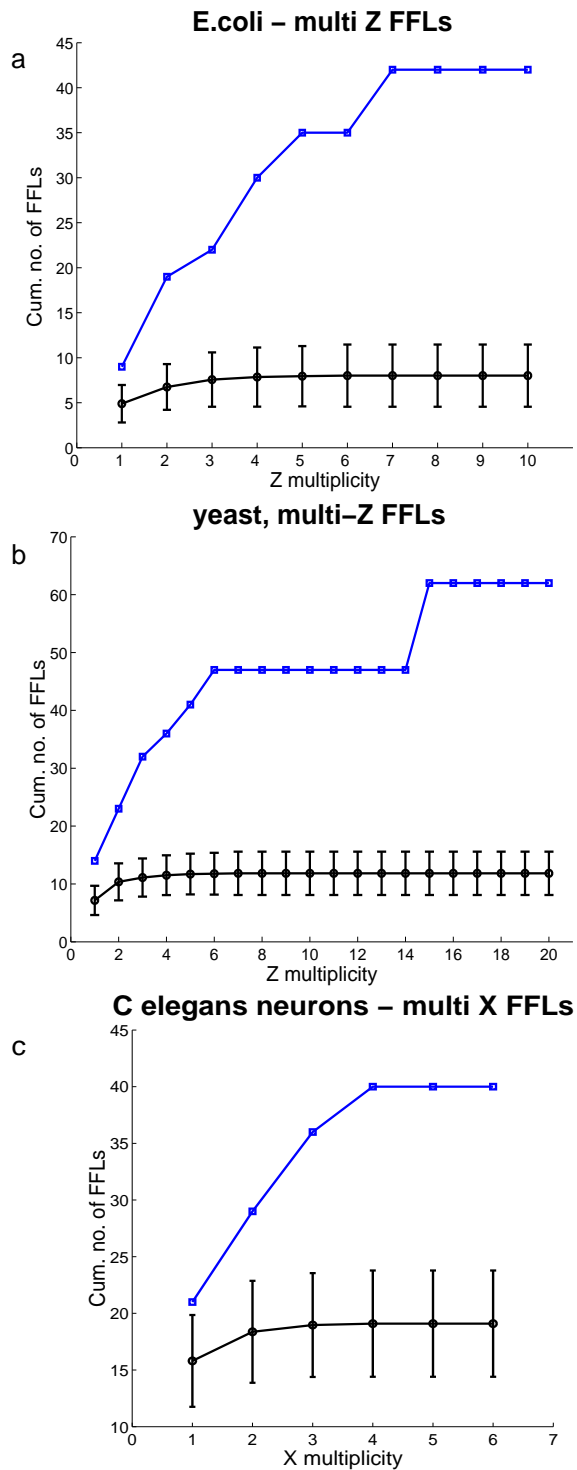


FIG. 3: Statistical significance of motif generalizations. The cumulative number of multi-Z FFLs in the real network (blue) and randomized networks - mean and SD (black) in **a.** *E. coli* transcription network. **b.** *S. cerevisiae* transcription network. **c.** The cumulative number of multi-X FFLs in the real and randomized networks (mean and SD) in the *C. elegans* neuronal network.

## D. Network Motif Generalizations

While enumerating all subgraphs of a given size is a difficult task, enumerating generalizations of a given subgraph can be performed very efficiently by an algorithm described in METHODS. The algorithm is based on using the appearances of the basic subgraph as nucleation points for a search for generalizations. As an example for motif generalizations, we applied this algorithm to networks in which the FFL and bi-fan are motifs, to ask whether any of the possible FFL or bi-fan generalizations occur in the networks. In the transcriptional networks of *E. coli* and yeast we find that the multi-Z FFL generalization is highly significant (Fig. 3a,b). The other two possible generalizations are not significant (in the *E. coli* data set, multi-X's and multi-Y's do not occur at all, in yeast both appear only twice). One example, the maltose utilization system, is illustrated in Fig. 4a. In each multi-Z FFL, the different genes (Z roles) share a common biological function (see tables 3,4 which list all multi-Z FFL complexes in the *E. coli* and yeast networks). We studied the neuronal network of *C. elegans*, in which nodes are neurons and edges are synaptic connections. We considered only edges which represent 5 or more synapses. We find that this network shows a different FFL generalization: the multi-X FFL (statistical significance - Fig. 3c). Multi-Y and multi-Z FFLs are found in far smaller numbers (double-X's and double-Y's FFL appear 3 times each). An example of a multi-X FFL in the locomotion control circuit of the worm is shown in Fig. 4b. We note that in the neuronal network where edges represent all synaptic connections (not only those with 5 or more synapses), we find numerous examples of the multi-Z and multi-Y FFLs, but the multi-X FFL is still the most common form (data not shown). In forward-logic electronic chips we find no simple generalization of the FFL. These electronic circuits do, however, show a complex FFL generalization - a structure with two Xs, a single Y and two Zs (a weak generalization, Fig. 4c). In the five forward-logic electronic chips we have analyzed, 70 percent to 100 percent of the FFLs are embedded in instances of this 5-node structure.

The most prominent 4-node network motif in these networks is the bi-fan [25] (Fig. 2d). The bi-fan has two roles and therefore two simple generalizations (Fig. 2f). We find that both simple generalizations of the bi-fan (multi-output and multi-input) are significant in transcription, neuronal and electronic networks (Table 2). The multi-output bi-fan generalizations are more significant and the maximal Y multiplicity is higher than the maximal X multiplicity in all these networks. In these networks we find structures of multi-output bi-fan with 10 Ys and more, while multi-input bi-fan do not exceed 6 input X roles.

Generalization	Subgraph size	Transcriptional Networks		Neurons	Electronic chips
		E. coli	yeast	C. elegans	S15850
basic bi-fan	4 (2X,2Y)	+ (N=209)	+ (N=1812)	+ (N=126)	+ (N=1040)
multi output	5 (2X,3Y)	+ (N=264)	+ (N=14857)	+ (N=152)	+ (N=1990)
	6 (2X,4Y)	+ (C=0.015)	+ (C=3.5)	+ (C=0.17)	+ (C=0.28)
multi input	5 (3X,2Y)	+ (N=20)	+ (N=81)	+ (N=25)	+ (N=226)
	6 (4X,2Y)	- (N=0)	+ (N=14)	+ (C=0.015)	+ (C <sub>i</sub> 0.001)
equal multi input-outputs	6 (3X,3Y)	+ (N=6)	+ (N=21)	- (N=0)	+ (N=301)

TABLE II: **Bi-fan generalizations in different networks.** (aX,bY) represents the multiplicity of each of the roles in the generalization (Fig. 2f). '+' : Statistically significant generalizations, '-' : non-significant generalizations. Number of appearances (N), or concentration ( $\times 10^{-3}$ ) (C), are listed.

Complex size	No.	X	Y	Z	Function
1	1	arcA	appY	appCBA	Anaerobic/stationary phase
	2	crp	fucPIKUR	fucAO	Fucose utilization
	3	crp	fur	cirA	Iron citrate uptake
	4	crp	galS	mglBAC	Carbon utilization
	5	crp	malI	malXY	Maltose utilization
	6	crp	melR	melAB	Melibiose utilization
	7	hns	flhDC	fliAZY	Flagella regulation
	8	metJ	metR	metA	Methionine biosynthesis
	9	ompR-envZ	csgDEFG	csgBA	Osmotic stress response
2	10	crp	caiF	caiTABCDE fixABCX	Carnitine metabolism
	11	crp	nagBACD	manXYZ nagE	Carbon utilization
	12	himA	ompR-envZ	ompC ompF	Osmotic stress response
	13	rpoN	fhIA	fdhF hycABCDEFGH	Formate hydrogen lyase system
	14	rpoN	glnALG	glnHPQ nac	Nitrogen utilization
3	15	crp	malT	malEFG malK-lamB-malM malS	Maltose utilization
4	16	crp	araC	araBAD araE araFG-araH1-H2 araJ	Arabinose utilization
	17	rob	marRAB	fumC nfo sodA zwf	Drug resistance
5	18	flhDC	fliAZY	flgBCDEFGHIJK flhBAE fliE fliFGHIJK fliLMNOPQR	Flagella system
7	19	fnr	arcA	cydAB cyoABCDE focA-pfB glpACB icdA nuoABCDEFGHIJKLMN sdhCDAB-b0725-sucABCD	Anaerobic metabolism

TABLE III: **Feedforward loops in *E. coli* transcription network classified into multi-Z complexes.** Complex size is the number of genes (Z-role node) in the FFL generalization

Complex size	No.	X	Y	Z	Function
1	1	TUP1	RME1	IME1	Meiosis
	2	RIM101	IME1	DIT1	Sporulation
	3	MIG1	HAP2-3-4-5	CYC1	Formation of apocytochromes
	4	MIG1	GAL4	GAL1	Galactokinase
	5	MIG1	CAT8	JEN1	Lactate uptake
	6	MIG2	CAT8	JEN1	(2X-FFL complex)
	7	GAT1	DAL80-GZF3	GAP1	Nitrogen utilization
	8	TUP1	ALPHA1	MFALPHA1	Mating factor alpha
	9	GAL11	ALPHA1	MFALPHA1	(2X-FFL complex)
2	10	TUP1	ROX1	ANB1 CYC7	Anaerobic metabolism
	11	GLN3	GAT1	GAP1 GLN1	Nitrogen utilization Glutamate synthetase
	12	GLN3	GAT1	DAL80 GLN1	Nitrogen utilization Glutamate synthetase
	13	GLN3	DAL80	GAP1	Nitrogen utilization
	14	PDR1	YRR1	UGA4 SNQ2 YOR1	Drug resistance
	15	GCN4	MET4	MET16 MET17	Metionine biosynthesis
3	16	HAP1	ROX1	ERG11 HEM13 CYC7	Anaerobic metabolism
	17	SPT16	SWI4-SWI6	CLN1 CLN2 HO	Cell cycle and mating type switch
4	18	GCN4	LEU3	ILV1 ILV2 ILV5 LEU4	Leucine and branched amino acid biosynthesis
	19	UME6	INO2-INO4	CHO1 CHO2 INO1 OPI3	Phospholipid biosynthesis
6	20	PDR1	PDR3	HXT11 HXT9 IPT1 PDR5 SNQ2 YOR1	Drug resistance
15	21	GLN3	DAL80	CAN1 DAL1 DAL2 DAL3 DAL4 DAL5 DAL7 DCG1 DUR1 DUR3 GDH1 PUT1 PUT2 PUT4 UGA1	Nitrogen utilization

TABLE IV: Feedforward loops in yeast transcription network classified into multi-Z complexes. Complex size is the number of genes (Z-role node) in the FFL generalization.

### III. DISCUSSION

#### A. Functions of generalized FFL motifs

The function of the FFL depends, among other things, on the signs of the interactions (positive or negative regulation). In the case of positive regulation, the 3-node FFL has been suggested to function as a persistence detector [35]: it filters out short input stimuli to X, and responds only to persistent signals. On the other hand, it responds quickly to OFF steps in the input to X [35]. With other sign combinations, the 3-node FFL can function as a pulse-generator or response accelerator [22].

What about generalizations of the FFL? The two sensory transcription networks, from a prokaryote (*E. coli*) and a eukaryote (*S. cerevisiae*), showed the same generalization of the FFL: both networks display the multi-output FFL generalization. The other two generalizations, multi-input and multi-Y, are not found significantly in these transcription networks. Multi-output FFL complexes are found throughout the transcription networks in diverse systems (Tables 3,4). The X role is usually a global transcription factor which controls many genes, the Y role is usually a 'local' transcription factor which controls specific systems, and the Z roles are the regulated genes which share a specific function. Often, multi-output FFLs in *E. coli* that respond to specific stimuli have a non-homologous multi-output FFL counterpart in yeast which responds to similar stimuli. This suggests convergent evolution to the same regulation pattern [6, 25]. Examples include systems that respond to carbon limitation, drugs, and nitrogen starvation in both organisms (Tables 3,4). Multi-output FFLs can also appear in systems that make up a protein machine, for example, a multi-output FFL in *E. coli* controls genes whose products make up the flagellar basal-body motor [18] (X=flhDC, Y=fliA, Z= class 2 flagella genes).

What is the function of the multi-output FFL? The multi-output FFL has a single input node X, a single internal node Y (secondary input) and a number of output nodes  $Z_1..Z_m$  (Fig. 2c,4a). The arrows in the FFL diagram should be assigned numbers representing the strength of the interaction of the transcription factors (TFs) with the promoters of the various Z-genes [31]. These numbers correspond to the activation or repression coefficients of each gene (the concentration of the TF required for 50 percent effect [23, 31, 34]. Here, we consider for simplicity the most common case, that of FFLs with positive regulations. The multi-Z configuration with activators is the most abundant FFL configuration in both *E. coli* and yeast [22]. We find, using simulations of this circuit (see METHODS for details), that the multi-output FFL can encode a temporal order of expression of the Z genes, by means of different

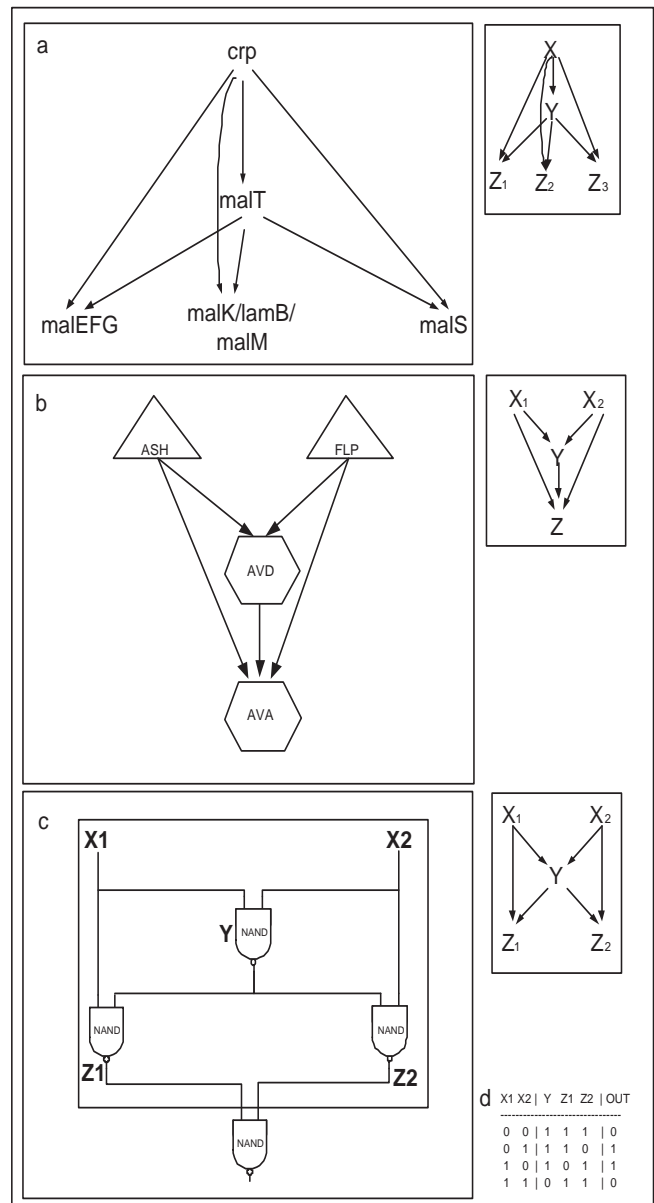


FIG. 4: Examples of the different FFL generalizations found in different networks. **a.** A three-Z FFL in the maltose utilization system of *E. coli*. Crp senses glucose starvation, malT senses maltotriose, and malEFG, malK and malS participate in maltose metabolism and transport. **b.** An example of a double-X FFL in the locomotion neural circuit of *C. elegans*. AVA and AVD are ventral cord command interneurons. AVD functions as touch modulator for backward locomotion. AVA functions as driver cell for backward locomotion. ASH and FLP are head sensory neurons sensitive to noxious chemicals and nose touch **c.** A significant generalized form of the FFL (2X,Y,2Z) found in forward-logic electronic chips. This 5-node structure appears as a part of a 6-node module, which implements XOR (Exclusive OR) using 4 NAND gates. **d.** Truth table of this circuit. There are 2 input bits X1 and X2 and a single output bit which is equal to (X1 XOR X2).

activation thresholds for each of the output genes (Fig. 5a,b). This temporal ordering feature is shared with another common network motif, the single-input module [35]. Indeed, high resolution expression measurements on the flagella multi-output FFL showed that the class 2 flagella genes are activated in a temporal order that corresponds to the functional order of the gene product in the assembly of the flagellar motor [18]. Furthermore, this multi-Z FFL can act as a persistence detector for all of the output genes (Fig. 5b): the Z genes are expressed only if the input stimulus to X is present for a long enough time. Thus the generalization preserves the functionality of the original FFL motif. The minimal stimulus duration needed to activate each gene can be tuned differentially by setting different activation thresholds for the activation of each Z gene by the Y TF (the gene with the lowest activation threshold is turned on first after the stimulation of X). The turn-off order of the Z genes can also be controlled by the activation

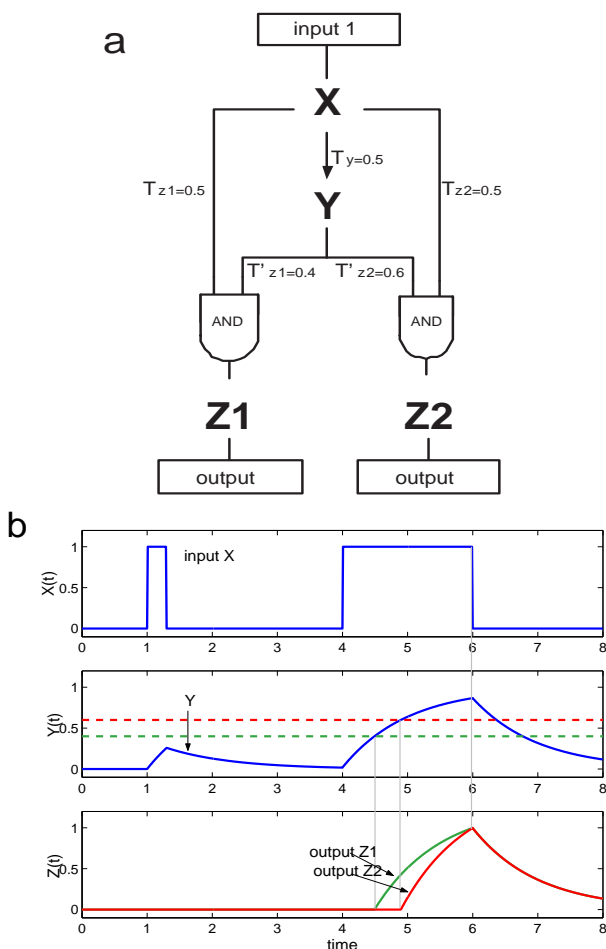


FIG. 5: Kinetics of a double-output FFL generalizations following pulses of stimuli. **a.** A double-output FFL with positive regulation and AND-logic input function for  $Z_1$  and  $Z_2$ . Numbers on the arrows are activation thresholds. **b.** Simulated kinetics of the double-output FFL.

or repression coefficients of the Y TF. In summary, the multi-output FFL preserves the functionality of the simple FFL, and in addition can encode temporal expression programs among the different Z genes.

A different FFL generalization, the multi-input FFL, is found in the neuronal synaptic connectivity of *C. elegans*. This network, which includes about 280 nodes representing neurons and about 400 connections (with 5 or more synapses), is found to chiefly display the multi-input FFL (Fig. 2c), and not the other two generalizations. The multi-input FFL has a number of input nodes  $X_1 \dots X_m$ , a single internal node Y (secondary input) and a single output node Z. This occurs 29 times in the network, with between 1 and 4 inputs. For example, the backward locomotion circuit of the worm is governed by two ventral-cord command interneurons AVD and AVA [5, 16]. These two neurons are linked in a multi-input FFL with several input neurons, such as ASH and FLP (Fig. 4b), which are head sensory neurons sensitive to nose touch and noxious chemicals [5, 16]. This circuit implements an avoidance reflex, eliciting backward motion in response to head stimulation.

What is the function of the multi-input FFL? This circuit may function as an integration unit of multiple sensory inputs, performing persistence detection on each of these inputs or on their sum. In the locomotion example, the FFL circuit would elicit backward motion only if the stimulation of one of the sensory neurons is longer than some threshold duration determined by the parameters of the circuit. A transient stimulation would not be enough to elicit backward motion. In general, the function of this circuit depends on the signs on the arrows and on two input-functions (gates): one input function integrates the multiple X inputs to Y, and the other integrates the inputs from Y and  $X_1 \dots X_m$  to Z. We simulated one possible two-input FFL, where the input-function governing the Y node is an OR gate,  $X_1 OR X_2$ , and the input-function of the Z node is  $Y AND (X_1 OR X_2)$  (Fig. 6a,b,c). This choice of input-functions ensure that Z responds to either  $X_1$  or  $X_2$ , and that Y is needed for Z to respond to  $X_1$  and  $X_2$  (as is the case, for example, in the circuit of Fig. 4b, in which ablation of AVD results in loss of sensory input to AVA [5]). These input functions could in principle be implemented by simple neurons which integrate weighted inputs. It is important to note that the simplest equations that describe transcription networks, also describe neurons with graded potential and no spiking (as *C. elegans* neurons are thought to be [10, 16] - see METHODS). The simulations show that the circuit can act as a persistence detector for both  $X_1$  and  $X_2$  (Fig. 6b). Furthermore, we find that sufficiently closely spaced short pulses of  $X_1$  and  $X_2$  can elicit a response, even if each pulse alone can not (Fig. 6c). This highlights a 'memory-like' function of Y, which can store information from recent stimulations

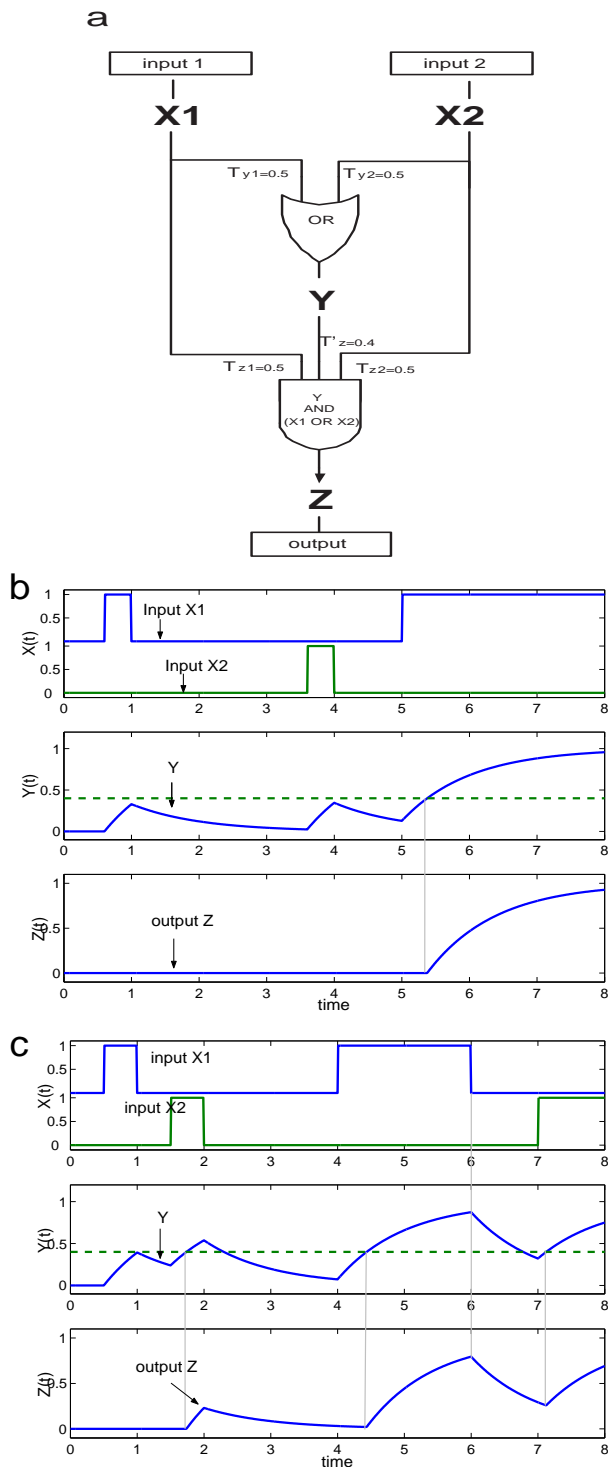


FIG. 6: Kinetics of a double-input FFL generalization following pulses of stimuli. **a.** A double-input FFL. Input functions for Y and Z, and the activation thresholds, are shown as gates and numbers on the arrows. **b.** Simulated kinetics of the two-input FFL, with short well-separated stimuli pulses of  $X_1$  and  $X_2$ , followed by a persistent  $X_1$  stimulus. **c.** Simulated kinetics of the double-input FFL, with short  $X_1$  stimulus followed rapidly by a short  $X_2$  stimulus pulse.

over its lifetime (its relaxation time). In the basic 3-node FFL, Y can store information about recurring pulses of X. In the multi-input FFL, Y can store information from multiple inputs, and increase sensitivity to one input if the other input has recently been detected.

Forward-logic electronic chips are networks in which nodes represent logic gates. These circuits are optimized to perform a hard-wired logical function between input and output nodes. Forward-logic chips, taken from an engineering database (ISCAS89), were previously found to display the FFL network motif [25]. Here we find that they display a specific generalization of the FFL, with two input and two output nodes (Fig. 4c). Analyzing the appearances of this pattern, we find that this 5-node generalized FFL motif is part of a well-known engineering module built of 4 NAND gates, which implements XOR (exclusive OR) logic on the two inputs [12] (see truth table in Fig. 4d). This demonstrates the possibility of using network motifs and their generalizations to reverse-engineer electronic and other information processing networks.

## B. Summary

This study presented a systematic approach for defining and detecting generalizations of network motifs based on replications of roles. Motif generalizations are families of subgraphs of different sizes which share a common structural theme, and which appear significantly more often in the network than in randomized networks. The generalizations are produced by replicating nodes in a basic motif structure. The generalizations preserve the roles of nodes in the motif (for example, by replicating input or output roles), and therefore can often preserve the functionality of the network motif on which the generalization is based. We find that different networks which display the same motifs can show very different generalizations of these motifs. These generalized motifs were demonstrated to carry out specific information processing functions by means of simulations. These functions can in principle be tested experimentally in transcription and neuronal systems.

Motif generalizations cover a substantial portion of high order motifs in various biological and technological networks we have studied. However motifs generalizations in the present form do not cover all possible types of families of structures that share similar architectural themes. It would be important to find additional rules for defining families of motifs beyond the current notion of motif generalization by role replication. Motifs and their generalizations can help us understand the structure, function and design principles of complex networks.

## IV. METHODS

### A. Roles in a subgraph - formal Definition:

We classify nodes in a subgraph into structurally equivalent classes. Each class represents a role [21, 38]. The measure of structural equivalence that we use here is automorphic equivalence [9, 38, 40, 41]. Let  $S(V_s, E_s)$  be a subgraph, then an automorphism is a one-to-one mapping,  $\tau$ , from  $V_s$  to  $V_s$ , such that  $(v_i, v_j) \in E_s$  if and only if  $(\tau(v_i), \tau(v_j)) \in E_s$ . Two nodes  $v_i$  and  $v_j$  are automorphically equivalent if and only if there is some automorphism,  $\tau$ , that maps one of the nodes to the other ( $\tau(v_i) = v_j$ ). For each subgraph  $S$ , we classify all its  $n$  nodes into roles by examining structural equivalence of all possible pairs of the nodes. By the transitivity of automorphic equivalence, we are promised to get a partition of the nodes into distinct classes. This concept can be readily generalized for networks with weights on the edges or with different types of nodes.

### B. Subgraph generalization - formal Definition:

Let  $S$  be the basic subgraph where  $r_1..r_L$  be the set of roles of  $S$  with multiplicity  $(d_1, \dots, d_L)$  respectively. simple generalization of  $S$  is a subgraph which are formed by replication of a single role  $r_i$  and its edges to preserve the roles connectivity of  $S$ . Note that in a simple generalization only a single role is replicated. A generalized form of a subgraph is defined by the following pair  $(M, V^L)$  where  $M$  is an  $L \times L$  image matrix, which describes the connectivity between roles ( $M[i, j] = 1$  if there is an edge between role  $i$  and  $j$ , and  $M[i, j] = 0$  otherwise).  $V^L \in N^L$  is an  $L$ -dimensional vector which defines the multiplicity of each role. The FFL which is an example of a basic subgraph, is represented by  $(M_{FFL}, (1, 1, 1))$  where

$$\mathbf{M}_{FFL} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

and the vector  $(1, 1, 1)$  describes the roles multiplicity: in the basic FFL each of the three roles X,Y,Z appears once. A FFL with two output nodes is represented by the pair  $(M_{FFL}, (1, 1, 2))$ . A FFL with  $m$  output nodes ( $m$  Z-role nodes) is represented by  $(M_{FFL}, (1, 1, m))$  (Fig 2c). Such a generalization has only one degree of freedom - the multiplicity of the Z role in the structure. There are cases, such as multiplicity of more than one role, where we need additional definition in order to distinguish between different types of structures. For this we define the generalization rule  $r$ . We define two possible generalization rules: strong generalization rule and weak generalization rule. An example of a strong and weak  $(M_{FFL}, (2, 1, 2))$  generalization is illustrated in Fig 2g. If  $S$  is the basic  $n$ -node subgraph with set of  $L$  roles represented by the multiplicity vector  $(d_1, \dots, d_L)$  then a *legal  $n$ -node set* is

every set of  $n$  nodes in the structure that consists of  $d_i$  nodes of role  $i$  (for all  $1 \leq i \leq L$ ). For example every set of three nodes in the multi output FFL, consisting of the X node, Y node and one of the Z role nodes, is a *legal  $n$ -node set*. A strong generalization rule,  $r_s$ , requires that every legal  $n$ -node set in the structure forms the basic subgraph  $S$ . A weak generalization rule,  $r_w$ , requires that every node in the structure participates in at least one *legal  $n$ -node set* (figure 2g). Thus, in such cases, the generalized form is represented by a triplet  $(M_S, V^L, r)$  instead of a pair  $(M_S, V^L)$ , where  $r$  is the generalization rule ( $r \in r_s, r_w$ ). Note that weak generalization can represent more than one unique structure of generalization of a given size.

### C. Algorithm for motif generalizations detection

We begin by finding the network motifs (significant subgraphs) of size  $n$  (usually  $n=3-5$ ) in the network as described in [25, 35]. For each motif, for each of its roles, we prepare a list of all the nodes that play that role. Based on the appearances of the motifs in the network as starting points, using the role lists, we perform a search for all its generalizations. This search reduces computation time and allows finding significant generalization forms of the basic motifs, which are beyond reach using algorithms that enumerate all subgraphs.

A network is described by a directed interaction graph  $G(V, E)$ , where  $V$  is the set of nodes and  $E$  is the set of edges. An edge  $(v_i, v_j) \in E$  represents a directed link between nodes  $v_i$  and  $v_j$ . For every  $n$ -node subgraph  $S$  which is detected as a network motif [25, 35] we search for its simple generalizations (multiplicity of one of the roles). We begin by building an induced graph  $G'(V', E')$ . The nodes in  $G'$  are only those that act as members (nodes) of  $S$  appearances in  $G$ , and the edges are only the edges in  $G$  between these nodes.  $G'$  is usually a much smaller graph than  $G$ , but it contains all the information we need for our purpose. For each simple generalization type  $j$  (multiplicity of the  $j$ -th role of the subgraph) the following is done: A non-directed graph  $\hat{G}(\hat{V}, \hat{E})$  is built where each node represents a specific basic subgraph  $S$  in  $G$  (a specific set of nodes in  $G$  that form a subgraph of type  $S$ ). The number of nodes in  $\hat{G}$  equals the number of times  $S$  appears in the original graph  $G$ . Two nodes in  $\hat{G}$  are connected if and only if they follow the generalization type,  $j$ , and the generalization rule (strong or weak). Setting the edges in  $\hat{G}$  is done efficiently by using the appearances of the basic subgraph in  $G'$  as starting points. For each specific 'starting point' subgraph  $S_1$  in  $G'$  we pass through all the 'neighboring' subgraphs  $S_2$  ('neighboring' in the sense that they share all node roles excluding  $j$ -th node roles) and check if each joint subgraph  $(S_1 \cup S_2)$  in  $G'$  forms a generalization type  $j$ . After setting all edges in  $\hat{G}$ , the next step is

to find all maximal cliques [4](a group of nodes in which every two are connected) in  $\hat{G}$ . Each maximal clique represents a maximal generalization type  $j$  of  $S$  (i.e. the generalization with maximal number of appearances of the basic subgraph). We store the size and the members (nodes in the original network) of all maximal generalizations. Complex generalizations (when more than one role is replicated) can be detected in a similar way by changing the rules we set the edges in  $\hat{G}$ .

In order to compute the statistical significance of a certain generalization form of a motif  $S$ , we first find for each appearance of  $S$  in the network the maximal size generalization in which it appears. Then we count the cumulative number of times  $S$  appears in the union of all the maximal generalizations (up to size  $k$ ). In order to verify that the generalization significance is not due to many stand-alone appearances of the basic subgraph (e.g. a single-Z FFL in the case of multi-Z FFL generalization), we subtract the number of time  $S$  appears as a stand-alone structure in the network from the cumulative results (Note that in Fig 3 we do not show the results after subtractions). We compare these numbers to the corresponding numbers in randomized networks. It is important to note that the randomized networks preserve the incoming, outgoing and mutual edge degree for each node. The networks are not constrained to have the same number of 3-node or higher subgraphs as in the real network (in [25] in contrast, 4-node motifs were detected based on randomized networks that preserved 3-node subgraph counts).

#### D. Mathematical model of FFL generalizations

In the transcriptional model (1)  $X(t)$  is the activity of the transcription factor  $X$ ,  $Y(t)$  of  $Y$ ,  $Z_j(t)$  is the expression level of gene  $Z_j$ . In Fig. 5a,b we used

$$dY/dt = F(X, T_y) - \alpha Y$$

$$dZ_j/dt = F(X, T_{Z_j})F(Y, T'_{Z_j}) - \alpha Z_j$$

where  $F(U, T) = 1$  if  $U > T$  and 0 otherwise,  $\alpha$  is the protein lifetime (30) ( $\alpha = 1$ ), and the activation thresholds are  $T_Y = 0.5$ ,  $T_{Z_1} = T_{Z_2} = 0.5$ ,  $T'_{Z_1} = 0.4$ ,  $T'_{Z_2} = 0.6$ . In the neuronal model  $X_i(t)$ ,  $Y(t)$  and  $Z(t)$  represent neuron membrane potentials. In Fig. 6a-c we used

$$dY/dt = F(X_1, T_{Y_1}) + F(X_2, T_{Y_2}) - \alpha Y$$

$$dZ/dt = F(Y, T'_Z)(F(X_1, T_{Z_1}) + F(X_2, T_{Z_2})) - \alpha Z$$

Here  $\alpha = 1$  is the relaxation time of the neurons, and the activation thresholds are  $T_{Y_1} = T_{Y_2} = 0.5$ ,  $T'_Z = 0.4$ ,  $T_{Z_1} = 0.5$ ,  $T_{Z_2} = 0.5$  are the synaptic activation thresholds. The input function of  $Z$  represents strong synapses from  $Y$  and weaker ones from  $X_1$  and

$X_2$ , such that both  $Y$  and either  $X_1$  or  $X_2$  needed for  $Z$  to be activated to a level that allows activation of its downstream (post synaptic) neurons or muscle cells.

#### E. Network databases

(N=number of nodes, E=number of edges). Self edges were excluded. Transcription network of *E.coli* [35], version 1.1 (N=423, E=519) available at <http://www.weizmann.ac.il/mcb/UriAlon/> was based on selected data from [25, 33] and literature. Transcription network of yeast (*S. cerevisiae*) [25], version 1.3 (N=685, E=1052) available at <http://www.weizmann.ac.il/mcb/UriAlon/> was based on selected data from [7, 25]. Neuronal synaptic connection network of *C. elegans*. (N=280, E=400) was based on [39] as arranged in [1]. The network was compiled with a cutoff of at least 5 synapses for connections between neurons. Target muscle cells were excluded. Electronic forward-logic chips [25]. Parsed the ISCAS89 benchmark data set [3] available at [www.cbl.ncsu.edu/CBL\\_Docs/iscas89.html](http://www.cbl.ncsu.edu/CBL_Docs/iscas89.html). Roles statistics (Table 1) and bi-fan generalizations data (Table 2) are shown for chip S15850 (N=10383, E=14240), and are representative of all logic chips in the database.

- 
- [1] Achacoso, T.B. and Yamamoto, W.S. 1992. *AY's Neuroanatomy of C. elegans for Computation*. CRC Press.
- [2] Bolouri, H. and Davidson, E.H. 2002. *Modeling transcriptional regulatory networks*. *Bioessays* **24**: 1118-1129.
- [3] Brglez, F., Bryan, D., and Kozminski, K. 1989. *Combinational Profiles of Sequential Benchmark Circuits*. *Proc. IEEE Int. Symposium on Circuits and Systems*: 1929-1934.
- [4] Bron, C. Kerbosch, J. 1973. *Finding all cliques of an undirected graph*. *Commun. ACM* **16**: 575-577.
- [5] Chalfie, M., Sulston, J.E., White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. 1985. *The Neural Circuit for Touch Sensitivity in Caenorhabditis elegans*. *The Journal of Neuroscience* **5**: 956-964.
- [6] Conant, G.C. and Wagner, A. 2003. *Convergent evolution of gene circuits*. *Nat Genet* **34**: 264-266.
- [7] Costanzo, M.C., Crawford, M.E., Hirschman, J.E., Kranz, J.E., Olsen, P., Robertson, L.S., Skrzypek, M.S., Braun, B.R., Hopkins, K.L., Kondu, P., Lengieza, C., Lew-Smith, J.E., Tillberg, M., and Garrels, J.I. 2001. *YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information*. *Nucleic Acids Res* **29**: 75-79.
- [8] Elowitz, M.B. and Leibler, S. 2000. *A synthetic oscillatory network of transcriptional regulators*. *Nature* **403**: 335-338.
- [9] Everett, M.G., Boyd, J.P., and Borgatti, S.P. 1990. *Ego-centered and local roles: A graph theoretic approach*. *Journal of Mathematical Sociology* **15**: 163-172.
- [10] Goodman, M.B., Hall, D.H., Avery, L., and Lockery, S.R. 1998. *Active currents regulate sensitivity and dynamic range in C. elegans neurons*. *Neuron* **20**: 763-772.
- [11] Guet, C.C., Elowitz, M.B., Hsing, W., and Leibler, S. 2002. *Combinatorial synthesis of genetic networks*. *Science* **296**: 1466-1470.
- [12] Hansen, M.C., Yaclin, H., and Hayes, J.P. 1999. *Unveiling the ISCAS-85 Benchmarks: A case study in reverse engineering*. *IEEE Design and Test*: 72-80.
- [13] Harary, F. and Palmer, E.M. 1973. *Graphical Enumeration*. Academic Press, NY.
- [14] Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. 1999. *From molecular to modular cell biology*. *Nature* **402**: C47-52.
- [15] Hasty, J., McMillen, D., and Collins, J.J. 2002. *Engineered gene circuits*. *Nature* **420**: 224-230.
- [16] Hope, I.A. 1999. *C. elegans A practical approach*. Exford university press.
- [17] Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G., and Alon, U. 2003. *Subgraphs in Random Networks*. *Phys Rev E* **68**: 026127.
- [18] Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M.G., and Alon, U. 2001. *Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria*. *Science* **292**: 2080-2083.
- [19] Kashtan, N., Itzkovitz, S., Milo, R., and Alon, U. 2003. *Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs*. Submitted.
- [20] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K., and Young, R.A. 2002. *Transcriptional regulatory networks in Saccharomyces cerevisiae*. *Science* **298**: 799-804.
- [21] Lorrain, F. and White, H.C. 1971. *Structural equivalence of individuals in social networks*. *Journal of Mathematical Sociology* **1**: 49-80.
- [22] Mangan, S. and Alon, U. 2003. *The structure and function of the feed-forward loop network motif*. *Proc Natl Acad Sci U S A*. **100**(21):11980-5
- [23] McAdams, H. and Arkin, A. 1998. *Simulation of prokaryotic genetic circuits*. *Annu Rev Biophys Biomol Struct*. **27**: 199-224.
- [24] McAdams, H. and Arkin, A. 2000. *Towards a circuit engineering discipline*. *Curr Biol* **10**: R318-320.
- [25] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. 2002. *Network motifs: simple building blocks of complex networks*. *Science* **298**: 824-827.
- [26] Nesetril, J. and Poljak, S. 1985. *On the complexity of the subgraph problem*. *Commen. Math. Univ. Carol.* **26**: 415-419.
- [27] Newman, M. 2003. *The structure and function of complex networks*. *SIAM Review* **45**: 167-256.
- [28] Newman, M. and Barkema, G. 1999. *Monte Carlo methods in statistical physics*. Oxford university press.
- [29] Ouzounis, C.A. and Karp, P.D. 2000. *Global properties of the metabolic map of Escherichia coli*. *Genome Res* **10**: 568-576.
- [30] Rao, C.V. and Arkin, A.P. 2001. *Control motifs for intracellular regulatory networks*. *Annu Rev Biomed Eng* **3**: 391-419.
- [31] Ronen, M., Rosenberg, R., Shraiman, B.I., and Alon, U. 2002. *Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics*. *Proc Natl Acad Sci U S A* **99**: 10555-10560.
- [32] Rosenfeld, N., Elowitz, M.B., and Alon, U. 2002. *Negative autoregulation speeds the response times of transcription networks*. *J Mol Biol* **323**: 785-793.
- [33] Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Diaz-Peredo, E., Sanchez-Solano, F., Perez-Rueda, E., Bonavides-Martinez, C., and Collado-Vides, J. 2001. *RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12*. *Nucleic Acids Res* **29**: 72-74.
- [34] Savageau, M.A. 2001. *Design principles for elementary gene circuits: Elements, methods, and examples*. *Chaos* **11**: 142-159.
- [35] Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. 2002. *Network motifs in the transcriptional regulation network of Escherichia coli*. *Nat Genet* **31**: 64-68.
- [36] Strogatz, S.H. 2001. *Exploring complex networks*. *Nature* **410**: 268-276.
- [37] Tyson, J.J., Chen, K.C., and Novak, B. 2003. *Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell*. *Curr Opin Cell Biol* **15**: 221-231.
- [38] Wasserman, S. and Faust, K. 1994. *Social Network Analysis*. Cambridge University Press.
- [39] White, J., Southgate, E., Thomson, J., and Brenner, S.

1986. *The structure of the nervous system of the nematode *Caenorhabditis elegans**. Phil. Trans. Roy. Soc. London Ser. B **314**: 1-340.
- [40] Winship. 1988. *Thoughts about roles and relations: An old document revisited*. Social Networks **10**: 209-231.
- [41] Winship, C. and Mandel, M. 1983. *Roles and Positions : A critique extension of the blockmodeling approach*. Sociological Methodology **1983-1984**: 314-344.