Supporting Appendix 1:
Network Topology

Table of Contents

In- and Out-Degree Distribution in the *E. coli* Transcription-Regulatory Network

The large-scale topological properties of the *E. coli* transcriptional-regulatory (TR) network have been studied before (1), but without considering link directionality. However, link directionality is an important attribute of TR networks. For example, the in-degree (k_{in}) distribution of connectivity (described by an exponential decay) and the out-degree (k_{out}) distribution of connectivity (described by power-law decay) are drastically different in the TR network of *S. cerevisiae* (2). As we show in Fig. 4*A* and *B*, the same is true for the *E. coli* TR network. In analogy to the definition of origons, we define terminons as the set of nodes directly or indirectly regulating an output node. Interestingly, the in- and out-degree distribution is reflected in the distributions of origon sizes (N_O) and terminon sizes (N_T) (see Fig. 4*C* and *D*).

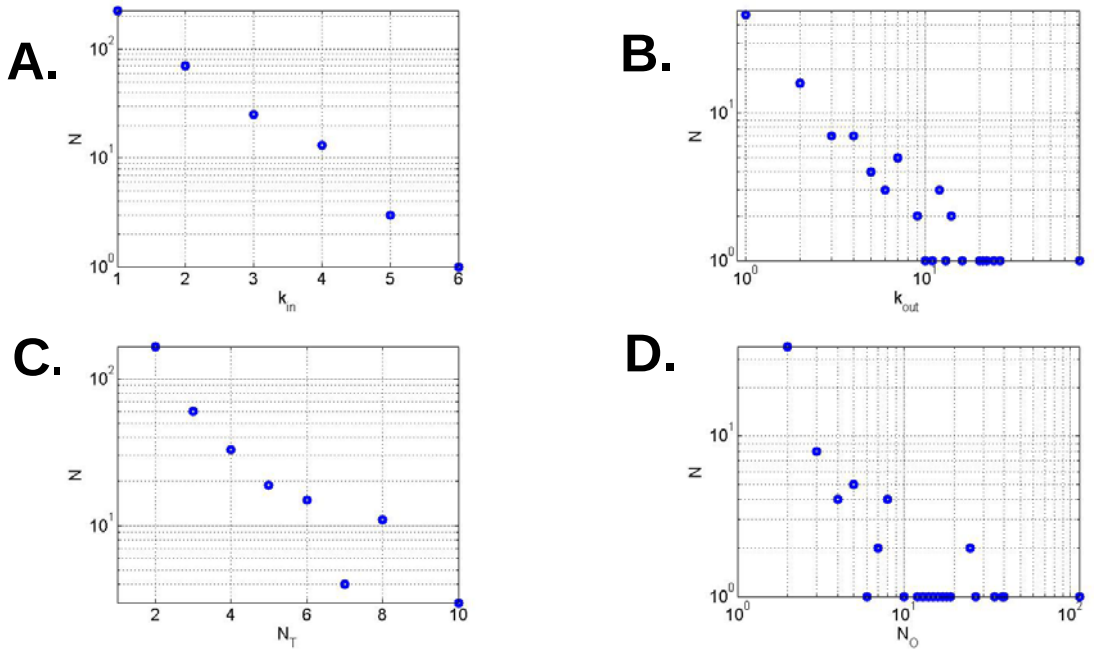

Fig. 4. The in-degree (k_{in}) distribution is best fitted with an exponential **(A)**, while the out-degree (k_{out}) distribution appears to be scale-free **(B)**. This is reflected by the histograms showing the number of nodes in terminons, N_T **(C)** and origons, N_O **(D)**.

Abundance of Three-Node Subgraphs in Randomized Versions of the *E. coli* TR Network

We studied the abundance of three-node subgraphs in randomized versions of the *E. coli* TR network to evaluate if they are over-represented in the real compared to randomized networks. To keep the layers unchanged and to keep the network acyclic, we proceeded with the randomization of the *E. coli* TR network, repeating a randomization step as follows. First, we selected four nodes: nodes $N_{1,L}$ and $N_{2,L}$ from the same layer L, and $N_{1,L'}$ and $N_{2,L'}$ from another layer $L'>L$, such that originally $N_{1,L}$ was connected to $N_{1,L'}$ and $N_{2,L}$ was connected to $N_{2,L'}$ (see Fig. 5). Then, we switched the links between these node pairs such that $N_{1,L}$ became connected to $N_{2,L'}$ and $N_{2,L}$ became connected to $N_{1,L'}$.

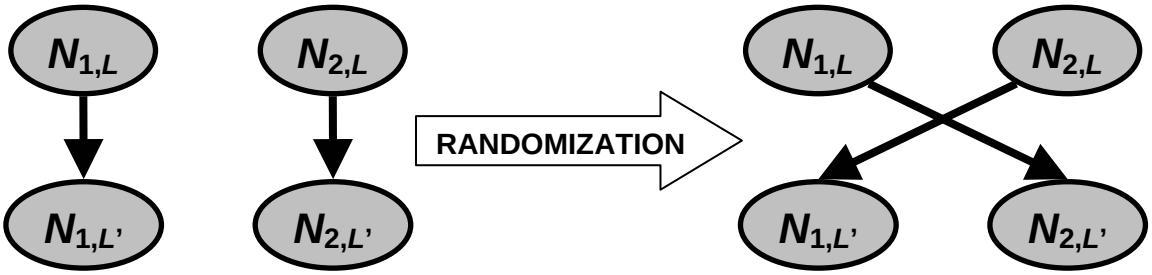

Fig. 5. The randomization protocol maintains the layers and keeps the network acyclic.

We counted all three-node subgraphs after performing 1000 randomization steps, 1000 times. The number of randomization steps we used was sufficient, as after 5000 randomization steps we obtained identical results. The histograms of cascade (CAS), convergence (CNV), divergence (DIV), and feed-forward loop (FFL) subgraph abundances are shown in Fig. 6.

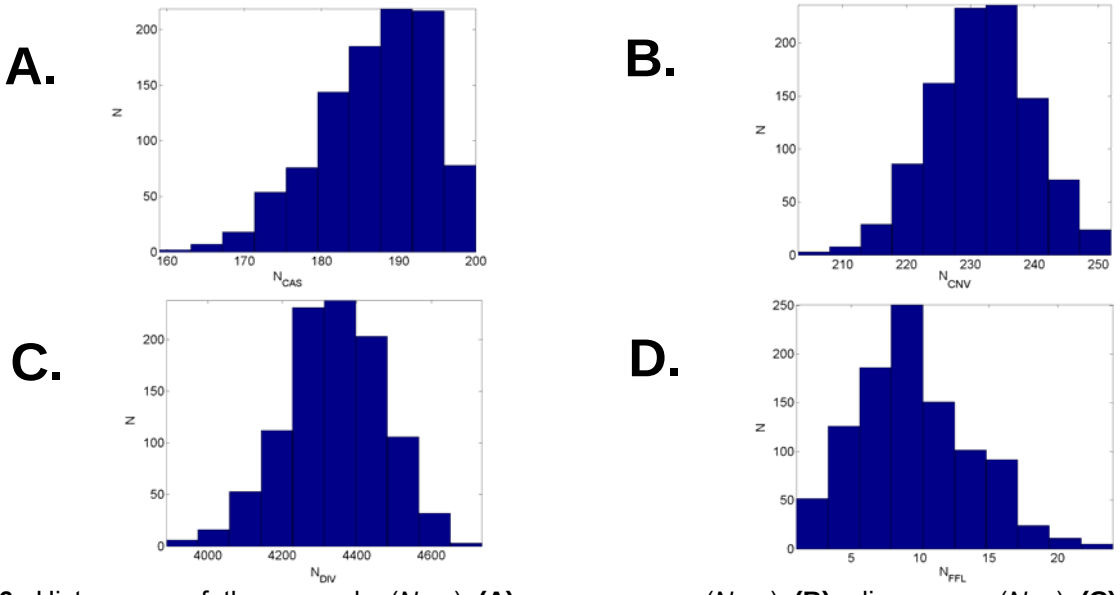

Fig. 6. Histograms of the cascade (N_{CAS}) **(A)**, convergence (N_{CNV}) **(B)**, divergence (N_{DIV}) **(C)**, and feed-forward loop (N_{FFL}) **(D)** subgraph abundances in randomized versions of the *E. coli* TR network. The number of randomized networks (N; vertical axis) containing a given number of subgraphs (N_{CAS}, N_{CNV}, N_{DIV}, N_{FFL}, horizontal axis) is shown.

Topological Analysis of Directed Tree and FFL-Tree Structures

Comparison of Regular Directed Trees with *E. coli* Origons. A tree is a connected graph with no circuits (loops). Most E. coli origons are random directed trees, defined as connected digraphs containing only DIV and CAS three-node subgraphs. In a directed tree, if a link exists between two nodes, the one from which the link originates is the parent node, while the other one is the child node. Every directed tree has a root node (which is the node with no parents). Trees have a hierarchical layered structure, where layers are defined based on the distance from the root node (nodes belong to the same layer if their path lengths from the root node are identical). We address the topological characteristics of regular directed trees in the following sections, aiming to gain insight into the topology of random trees as well.

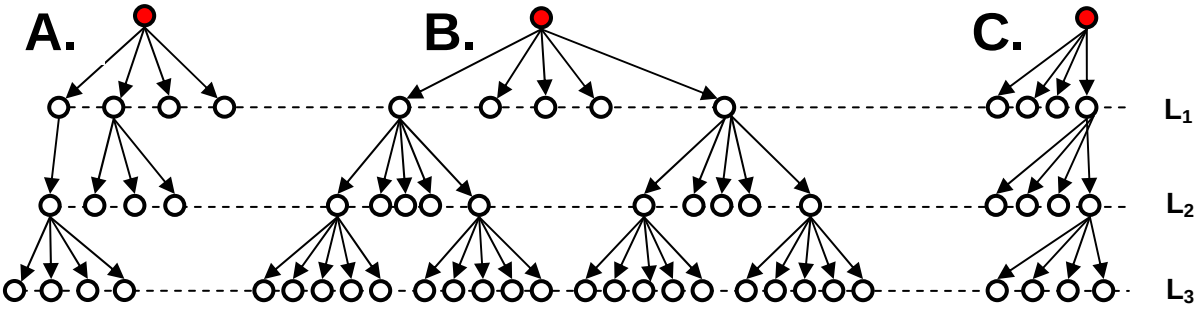

Fig. 7. Directed trees with three layers (*L*=3): random tree **(A)**, regular tree with *b*=5 and *c*=2/5 **(B)**; regular tree with *b*=4 and *c*=0.25 **(C)**. Root nodes are colored red.

A tree is regular if the branching rules at every layer are identical (i.e., the same fraction of nodes have the same number of links originating from them). Examples of regular trees can be seen in Figs. 7*B* and *C*. Regular trees can be completely characterized by three numbers: the number of layers *L*, the continuation ratio *c* and the branching ratio *b*. The number of layers L is the diameter (the maximum shortest path length) of the tree. We define the continuation ratio *c* as the fraction of nodes with children in every layer. The branching ratio is the number of links originating from nodes with children.

Theorem: The number of nodes in a regular tree is given by the expression:

$$N = bS(bc, L) + 1, \tag{S1}$$

where the function S is given by

$$S(x, y) = \begin{cases} \dfrac{(x)^y - 1}{x - 1}, & x > 1 \\ y, & x = 1 \end{cases}. \tag{S2}$$

Proof:

First, we consider the situation $bc>1$ (as in Fig. 7*B*).

Corollary: The number of nodes in layer L_{i+1} is equal to

$N_i = b(bc)^{i-1}$.

Proof of corollary: The number of nodes in layer L_1 is always equal to the branching ratio:

$N_1 = b$.

From the definition of c, only bc out of the b nodes in layer L_1 will have children. From the definition of b, each of these bc nodes has exactly b children, and therefore the number of nodes in layer L_2 is

$N_2 = (bc)b$

We continue the proof by using the method of induction. Let us suppose that the number of nodes in layer L_i is

$N_i = b(bc)^{i-1}$.

We will calculate the number of nodes in layer L_{i+1}. By the definition of the continuation ratio, the number of nodes with children from layer i is:

$N_i = bc(bc)^{i-1} = (bc)^i$,

and, each of these nodes having exactly b children, we obtain for the number of nodes in layer i+1:

$N_{i+1} = b(bc)^i$.

The total number of nodes in the tree is the sum of the number of nodes in all layers:

$$N = \sum_{i=0}^{L} N_i = 1 + \sum_{i=0}^{L} b(bc)^{i-1} = 1 + b\sum_{i=0}^{L} (bc)^{i-1}.$$

Using the formula for the sum of a geometric series,

$$\sum_{i=0}^{L} x^{i-1} = \frac{x^L - 1}{x - 1}, \qquad \text{(S3)}$$

we obtain:

$$N = 1 + b\frac{(bc)^L - 1}{bc - 1} = 1 + bS(bc, L).$$

Next, we consider the situation bc=1 (as in Fig. 7*C*). In this case, only one node from each layer has b children. Therefore, each layer will contain exactly b nodes. The total number of nodes in the network is therefore given by:

$N = 1 + bL = 1 + bS(bc, L)$,

which completes the proof.

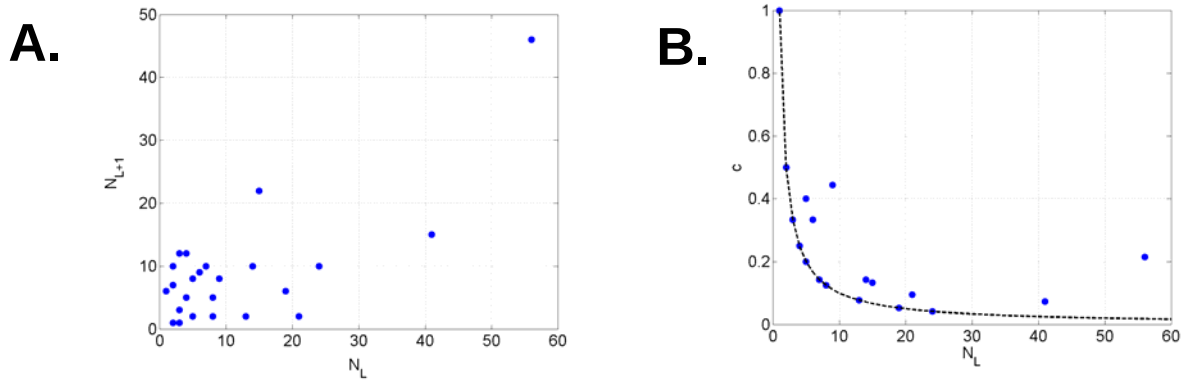

Fig. 8. Plots characterizing the topology of *E. coli* origons. **(A)** The number of nodes in layer L+1 (N_{L+1}, vertical axis) is approximately identical with the number of nodes in layer L (N_L, horizontal axis). **(B)** For all origons, the ratio of nodes with children (c, vertical axis) is shown versus the number of nodes in the layer (N_L, horizontal axis). Typically, only one node per layer has children (the black line corresponds to bc=1).

To compare the topological structure of *E. coli* origons with regular trees, in Fig. 8*A* we plot the number of nodes in layer L+1 (N_{L+1}) as a function of the number of nodes in layer L (N_L), and in Fig. 8*B* we plot the continuation ratio versus the number of nodes in layer L within each origon. As these plots indicate, the number of nodes in consecutive layers in *E. coli* origons are approximately identical, and the continuation ratio is low (the black line in Fig. 8*B* corresponds to bc=1). Therefore, the topology of *E. coli* origons is reminiscent of regular trees with low continuation ratio c, for which only one node per layer has children and the number of nodes in consecutive layers is identical (Fig. 7*C*).

The Abundance of DIVs and CASs Within Individual Origons. By their definition, directed trees contain only two types of three-node subgraphs: CASs and DIVs. For regular trees, the abundance of these two subgraphs can be determined from the parameters L, b and c as follows.

DIVs appear at branching points. Given a parent and its b children, the corresponding number of DIVs is $\binom{b}{2} = \frac{b(b-1)}{2}$. This quantity multiplied by the number of parent nodes in the network, $S(bc,L)$, gives the total number of DIVs in a directed regular tree:

$$N_{DIV} = \frac{b(b-1)}{2} S(bc, L) = \frac{(b-1)(N-1)}{2}. \qquad \text{(S4)}$$

By similar reasoning, the number of CASs in a regular directed tree is given by

$$N_{CAS} = b[S(bc, L) - 1] = N - b - 1. \qquad \text{(S5)}$$

For bc=1, these expressions become of second and first order in N:

$$N_{DIV} = N\frac{(N-L-1)}{2}, \qquad \text{(S4')}$$

and

$$N_{CAS} = \frac{L-1}{L}(N-1). \qquad \text{(S5')}$$

These relationships are confirmed by Fig. 9*A* (showing the number of DIVs as an approximately quadratic function of the number of nodes in an origon), and in Fig. 9*B* (showing the number of CASs as an approximately linear function of the number of nodes in an origon).

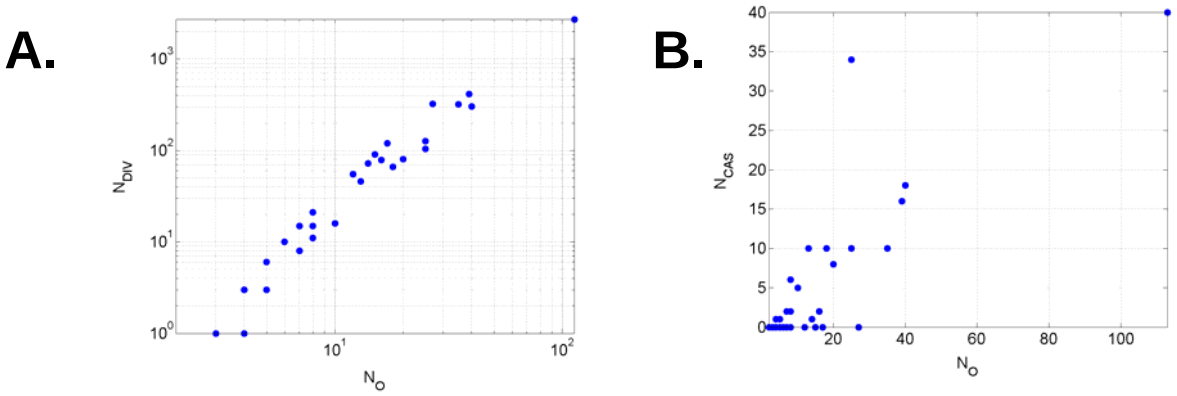

Fig. 9. The number of DIVs **(A)** and CASs **(B)** versus the number of nodes in each individual origon.

The Relationship Between the Number of Nodes, Links and FFLs in Origons. In a tree, every link (SRI) connects two consecutive layers L and L+1, and the number of SRIs, N_{SRI} depends on the number of nodes (operons) within the origon as $N_{SRI} = N_O - 1$. If extra links are added to this simple backbone structure without increasing the number of nodes, the new links must connect non-consecutive layers L and L' (L'>L+1), and N_{SRI} and N_N must satisfy the inequality $N_{SRI} - N_O + 1 > 0$ (see Fig. 10*A*). All extra links in the *E. coli* TR network create FFLs, skipping *exactly* one layer, and thus the number of FFLs, SRIs and nodes satisfy the equation $N_{FFL} = N_{SRI} - N_O + 1$. Interestingly, the number of FFLs within the 7 FFL-containing individual origons depends linearly on the number of nodes within the origon (Fig. 10*B*).

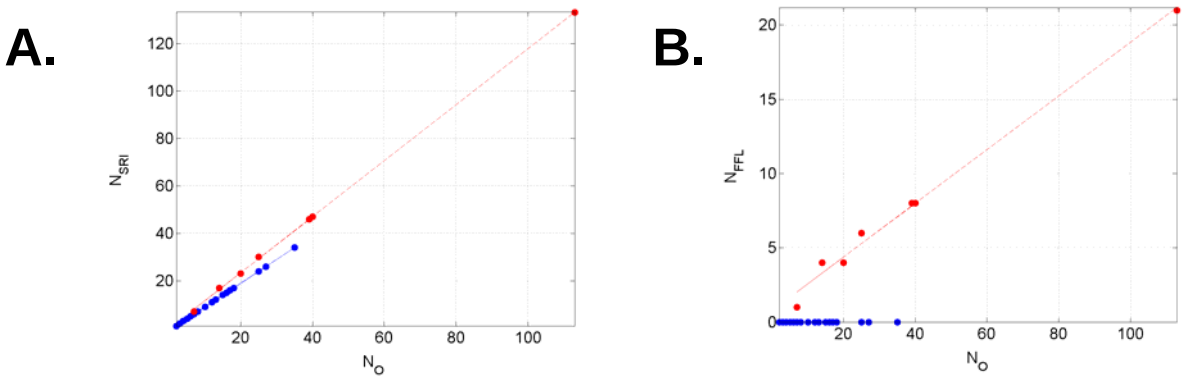

Fig. 10. The number of SRIs **(A)** and FFLs **(B)** (N_{SRI}, N_{FFL}, vertical axes) are shown versus the number of nodes in each individual origon (N_O, horizontal axis). FFLs are due to extra links added to the tree structure.

Of the 7 FFL-containing origons, only the *metJ* origon (Fig. 16) contains a single FFL. Of the other 6 FFL-containing origons, the *fnr* (Fig. 14), *rpoN* (Fig. 18) and *rob* (Fig. 17) origons contain FFLs typically starting at the root node. On the other hand, the *crp* (Fig. 12), *himA* (Fig. 15) and *cspA* (Fig. 13) origons contain FFLs starting both near and far from the root node, sharing the subnetwork responsible for flagellum synthesis as the distal FFL cluster, with *flhDC* and *fliAZY* as X and Y-nodes, respectively. Many FFLs originating at root nodes share their X- and Y-nodes, which are both sensors of related stimuli, such as the presence of carbon sources (*crp* origon), anoxia (*fnr* origon), antibiotics (*rob* origon), etc. The FFL cluster responsible for flagellum synthesis is different from the rest of the FFLs, as none of its nodes is a sensor. The shared X- and Y-nodes in this cluster encode sigma factors necessary for the synthesis of flagellar proteins, and integrate the information from three input nodes (*crp*, *cspA* and *himA*) and five sensors (*crp*, *cspA*, *himA*, *hns*, *ompR_envZ*).

The Relationship Between the Number of CNVs and BFMs

The over-represented four-node bi-fan motif (BFM) (3,4) is the union of two CNVs (or two DIVs). If we define convergence classes based on the operon pairs co-regulating a common target operon, and the number of CNVs in a given class is N_{CNV}^{CLS}, the number of BFMs that arise from these CNVs in this class is given by

$$N_{BFM} = \binom{2}{N_{CNV}^{CLS}} = \frac{N_{CNV}^{CLS}\left(N_{CNV}^{CLS}-1\right)}{2}. \qquad \text{(S6)}$$

If several classes of CNVs are considered, each containing a different number of CNVs, then the total number of BFMs is given by

$$N_{BFM} = \sum_{CLS}\binom{2}{N_{CNV}^{CLS}} = \sum_{CLS}\frac{\left(N_{CNV}^{CLS}\right)^2 - N_{CNV}^{CLS}}{2} = \frac{1}{2}\left[\sum_{CLS}\left(N_{CNV}^{CLS}\right)^2 - \sum_{CLS}N_{CNV}^{CLS}\right] =$$

$$= \frac{N_{CNV}}{2}\sum_{CLS}N_{CNV}^{CLS}\frac{N_{CNV}^{CLS}}{N_{CNV}} - \frac{N_{CNV}}{2} = \frac{N_{CNV}}{2}\left[\sum_{CLS}N_{CNV}^{CLS}P\left(N_{CNV}^{CLS}\right)-1\right] = \frac{N_{CNV}}{2}\left(< N_{CNV}^{CLS} > -1\right).$$

Therefore, the total number of BFMs depends on the total number of CNVs and on the expected number of CNVs in a given class $< N_{CNV}^{CLS} >$ as

$$N_{BFM} = \frac{N_{CNV}}{2}\left(< N_{CNV}^{CLS} > -1\right), \qquad \text{(S7)}$$

where CNVs are classified according to their pair of input nodes, and the expected number of CNVs in a given class is defined as $< N_{CNV}^{CLS} >= \sum_{CLS}N_{CNV}^{CLS}P\left(N_{CNV}^{CLS}\right)$.

If the number of CNVs in a given class is N_{CNV}^{CLS}, then the number of all BFMs (N_{BFM}) that arise from all CNVs (N_{CNV} = 227) in the network can be calculated, using the expected number of CNVs in a given class $< N_{CNV}^{CLS} >$ =2.8414 as $N_{BFM} = \frac{N_{CNV}}{2}\left(< N_{CNV}^{CLS} > -1\right)$ = 209. Based on this relationship, the high number of BFMs can be explained by the high number of CNVs in the same class. From a total of 134 CNV classes, 45 contain 2 or more CNVs, giving rise to a high number of BFMs.

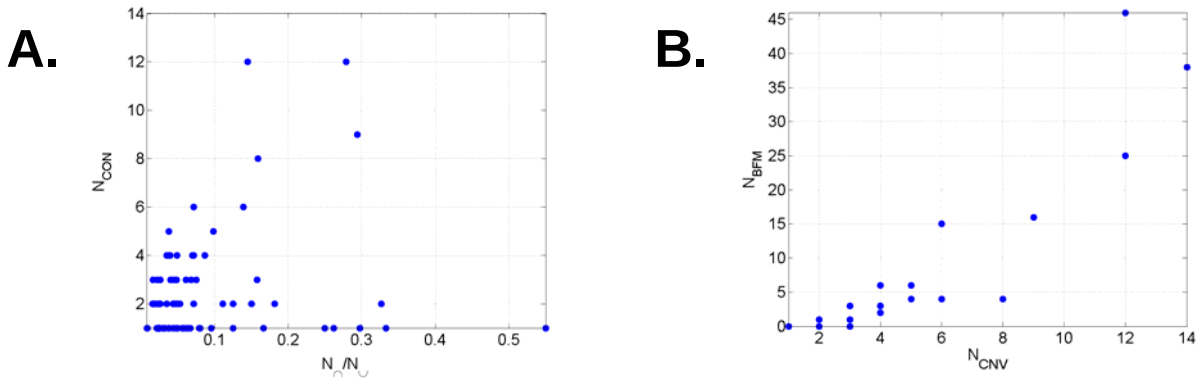

Fig. 11. The number of CNVs (N_{CNV}, vertical axis) versus the overlap ($N_{\cap}/N_{U}$, horizontal axis) of two origons **(A)** and the number of bi-fan motifs (N_{BFM}, vertical axis) versus the number of CNVs (N_{CNV}, horizontal axis) in the union of two origons **(B)**.

Since there are no CNV subgraphs within any individual origon, we studied their abundance as a function of the degree of overlap ($N_{\cap}/N_{\mathrm{U}}$) of two origons, defined as the number of nodes in their intersection ($N_{\cap}$) divided by the number of nodes in their union (N_{U}). As shown in Fig. 11*A*, there are two types of origon unions: some contain a number of CNVs proportional to the degree of overlap, while other origon unions contain a low number of CNVs, independent of their degree of overlap. Therefore, some two-origon unions tend to contain a number of CNVs proportional to their overlap, while others contain a low number of CNVs, independent of their overlap. Also, as discussed above, the number of BFMs (N_{BFM}) increases with the number of CNVs (N_{CNV}) for the union of any two origons (Fig. 11*B*).

Position of Subgraphs Within Layers and Origons

Besides the abundance of subgraphs, their position within layers and origons could also be indicative of their possible functional (signal-processing) role. Table 2 shows the distance from the input or output layers of the CAS, DIV, CNV and FFL three-node subgraphs, while Table 3 shows the distance from the input or output layers of autoregulatory loops (ARL), bi-fan motifs (BFM) and all nodes. Adding the ARL to the list is justified, as out of the 578 links listed between the 423 operons in *E. coli*, 59 have autoregulatory loops. This gives a self-link abundance of N_{ARL}/N_L = 59/578 = 0.1021, a 22 times higher value than $\frac{N}{\frac{N(N+1)}{2}} = \frac{2}{N+1}$ = 2/(423+1)=0.0047 that one expects if links are re-distributed randomly. Autoregulatory loops preferentially appear near the input layer indicating that they play a role in signal pre-processing, as previously described (5,6).

The results of this analysis indicate that FFLs tend to appear where the regulatory network is shallow (they often span the network). The small network depth near FFLs indicates that the functions fulfilled by them have to be carried out rapidly, minimizing the response delay caused by long regulatory chains (7). Also, ARLs tend to appear near the input layer, supporting their role as stabilizers of noisy sensor TF levels (5) or inducers of bistability (6,8) in face of graded external signal fluctuations.

References

1. Dobrin, R., Beg, Q. K., Barabási, A.-L. & Oltvai, Z. N. (2004) *BMC Bioinformatics* **5,** 10.

2. Guelzim, N., Bottani, S., Bourgine, P. & Kepès, F. (2002) *Nat. Genet.* **31,** 60-63.

3. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002) *Science* **298,** 824-827.

4. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. & Alon, U. (2004) *Science* **303,** 1538-1542.

5. Becskei, A. & Serrano, L. (2000) *Nature* **405,** 590-593.

6. Becskei, A., Seraphin, B. & Serrano, L. (2001) *EMBO J.* **20,** 2528-2535.

7. Rosenfeld, N. & Alon, U. (2003) *J. Mol. Biol.* **329,** 645-654.

8. Isaacs, F. J., Hasty, J., Cantor, C. R. & Collins, J. J. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 7714-7719.

CNV (227)	Type	Total	Dist=0	Dist=1	Dist=2	Dist=3	Dist=4
Input	**X**	62	47	13	1	1	0
	Y	86	0	48	35	1	2
	X,Y	146	47	60	35	2	2
Output	**X**	62	0	44	12	5	1
	Y	86	79	5	2	0	0
	X,Y	146	79	48	13	5	1

DIV (4777)	Type	Total	Dist=0	Dist=1	Dist=2	Dist=3	Dist=4
Input	**X**	59	39	16	3	1	0
	Y	316	0	211	85	8	12
	X,Y	356	39	212	85	8	12
Output	**X**	59	0	37	15	5	2
	Y	316	287	23	4	2	0
	X,Y	356	287	47	15	5	2

CAS (160)	Type	Total	Dist=0	Dist=1	Dist=2	Dist=3	Dist=4
Input	**X**	23	17	5	1	0	0
	Y	25	0	18	6	1	0
	Z	87	0	0	65	10	12
	X,Y,Z	122	17	18	65	10	12
Output	**X**	23	0	0	16	5	2
	Y	25	0	19	4	2	0
	Z	87	79	7	1	0	0
	X,Y,Z	122	79	20	16	5	2

FFL (42)	Type	Total	Dist=0	Dist=1	Dist=2	Dist=3	Dist=4
Input	**X**	10	6	3	1	0	0
	Y	19	0	14	4	1	0
	Z	42	0	0	32	5	5
	X,Y,Z	67	6	15	36	5	5
Output	**X**	10	0	0	4	5	1
	Y	19	0	14	4	1	0
	Z	42	40	2	0	0	0
	X,Y,Z	67	40	15	6	5	1

Table 2. Position of three-node subgraphs (CNV, DIV, CAS, FFL) within the *E. coli* transcriptional regulatory network. The number of X-, Y-, Z-type nodes located at distances of 0, 1, 2, 3 and 4 links (columns) from the input layer (first rows for each subgraph) and from the output layer (last rows for each subgraph) is listed.

BFM (209)	Type	Total	Dist=0	Dist=1	Dist=2	Dist=3	Dist=4
Input	**X**	31	23	7	0	1	0
	Y	69	0	35	31	1	2
	X,Y	100	23	42	31	2	2
Output	**X**	31	0	17	8	5	1
	Y	69	63	4	2	0	0
	X,Y	100	63	21	10	5	1

ARL (59)	Type	Total	Dist=0	Dist=1	Dist=2	Dist=3	Dist=4
Input	**"+"**	14	8	3	2	1	0
	"-"	42	25	13	3	1	0
	"+/-"	3	2	1	0	0	0
	Total	59	35	17	5	2	0
Output	**"+"**	14	4	9	1	0	0
	"-"	42	6	29	4	2	1
	"+/-"	3	0	1	1	1	0
	Total	59	10	39	6	3	1

ALL	Type	Total	Dist=0	Dist=1	Dist=2	Dist=3	Dist=4
Input		418	76	233	87	10	12
Output		418	312	83	16	5	2

Table 3. Position of bi-fan motifs (BFM) and autoregulatory loops (ARL) within the *E. coli* transcriptional regulatory network. The meaning of rows and columns is the same as in Table S1.