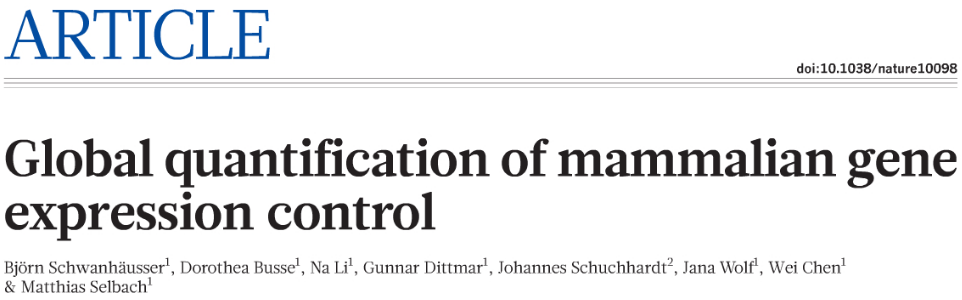
Bioinformatics 3
V8 – Gene Regulation

Mon, Nov 21, 2016

- Measuring transcription + translation rates
- Motifs in GRNs
- Master Regulatory Genes in GRNs

Rates of mRNA transcription and protein translation



ARTICLE

doi:10.1038/nature10098

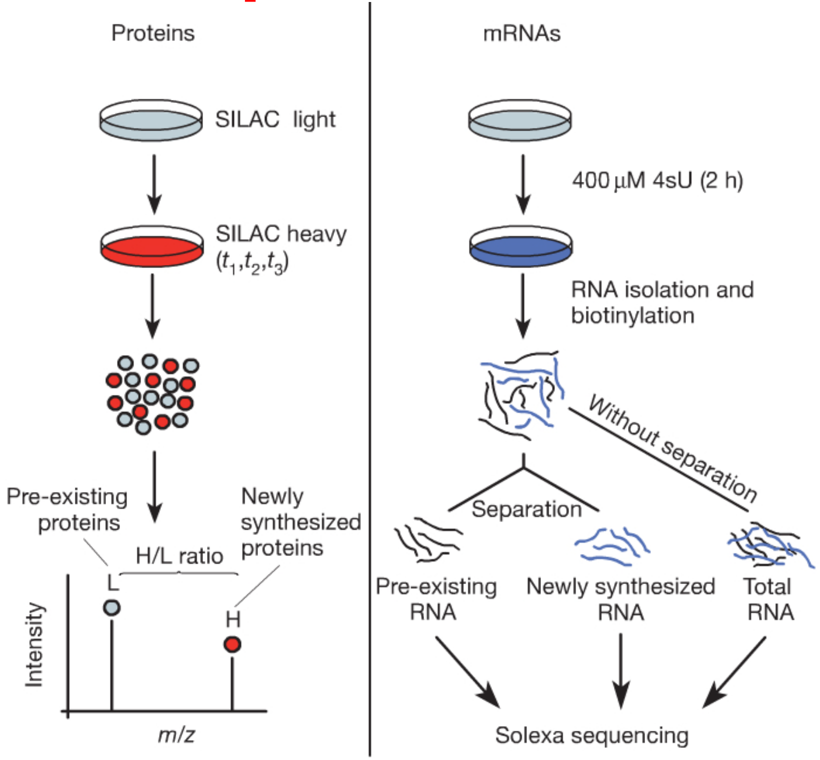
Global quantification of mammalian gene expression control

Björn Schwanhäusser[1], Dorothea Busse[1], Na Li[1], Gunnar Dittmar[1], Johannes Schuchhardt[2], Jana Wolf[1], Wei Chen[1] & Matthias Selbach[1]

SILAC: „stable isotope labelling by amino acids in cell culture" means that cells are cultivated in a medium containing **heavy** stable-isotope versions of **essential amino acids**.

When non-labelled (i.e. light) cells are transferred to heavy SILAC growth medium, newly synthesized proteins incorporate the heavy label while pre-existing proteins remain in the light form.

Schwanhäuser et al. Nature 473, 337 (2011)



Parallel quantification of mRNA and protein turnover and levels. Mouse fibroblasts were pulse-labelled with heavy amino acids (SILAC, left) and the nucleoside **4-thiouridine** (4sU, right).
Protein and mRNA turnover is quantified by mass spectrometry and next-generation sequencing, respectively.

Rates of mRNA transcription and protein translation

84,676 peptide sequences were identified by MS and assigned to 6,445 unique proteins.

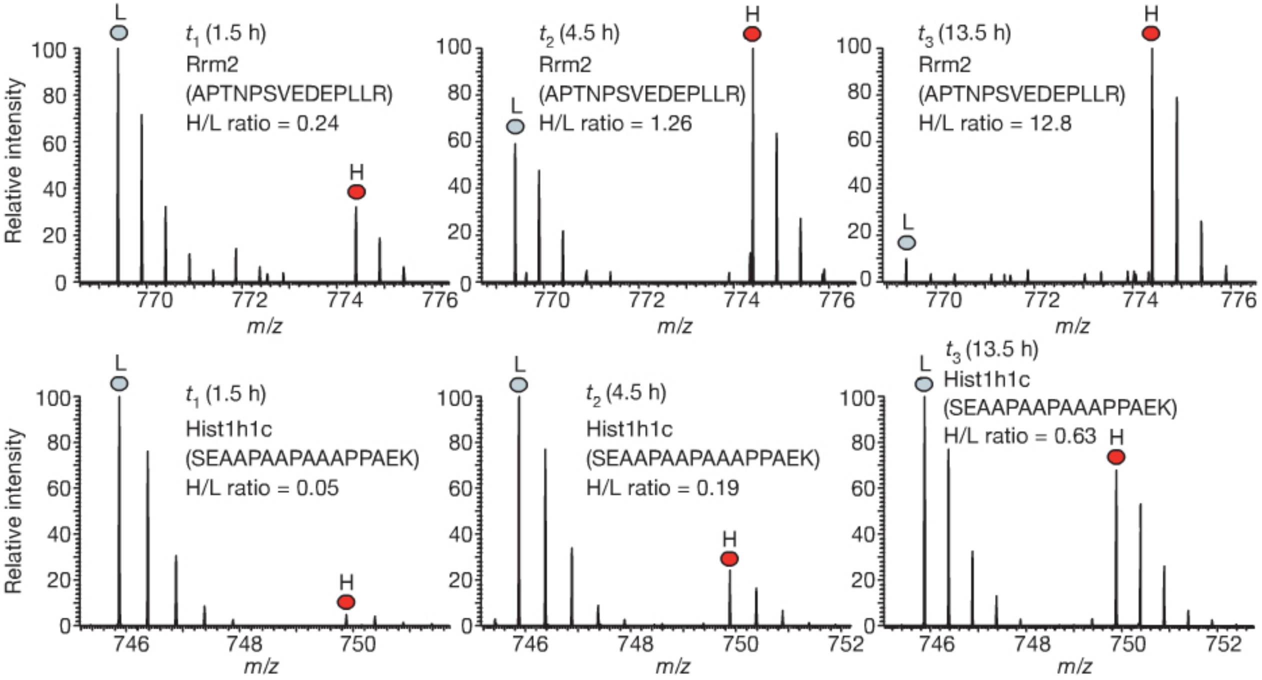
5,279 of these proteins were quantified by at least three heavy to light (H/L) peptide ratios belonging to these proteins.

Mass spectra of peptides for two proteins (x-axis: mass over charge ratio).

Over time, the heavy to light (H/L) ratios increase.

You should understand these spectra!

Top: high-turnover protein



Bottom: low-turnover protein, slow synthesis, long half-life

Schwanhäuser et al. Nature 473, 337 (2011)

Protein half-lifes and decay rates

Consider ratio *r* of protein with heavy amino acids (P_H) and light amino acids (P_L):

$$r = \frac{P_H}{P_L}$$

Assume that proteins labelled with light amino acids decay exponentially with degradation rate constant k_{dp} :

$$P_L = P_0 e^{-k_{dp} t} .$$

Express (P_H) as difference between total number of a specific protein P_{total} and P_L:

$$P_H(t) = P_{total}(t) - P_L(t)$$

Assume that P_{total} doubles during duration of one cell cycle (which lasts t_{∞}):

$$P_H(t) = P_{total}(t) - P_L(t) = P_0 2^{t/t_{cc}} - P_L(t) ,$$

$$r = \frac{P_H}{P_L} = \frac{P_0}{P_L} 2^{\frac{t}{t_{cc}}} - 1$$

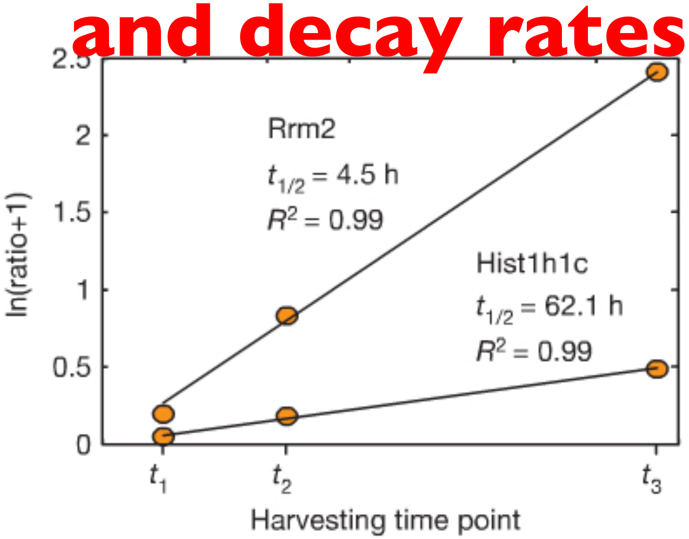
$$\frac{P_H}{P_L} + 1 = \frac{P_0}{P_L} 2^{\frac{t}{t_{cc}}}$$

take *ln* on both sides

$$\ln(ratio + 1) = \ln \frac{P_0}{P_L} 2^{\frac{t}{t_{cc}}} = \ln e^{k_{dp} t} + \ln 2^{\frac{t}{t_{cc}}} = k_{dp} t + \ln 2^{\frac{t}{t_{cc}}}$$

$$ln(ratio + 1) = k_{dp} t + \frac{t}{t_{cc}} ln2 = t \times \left(k_{dp} + \frac{ln2}{t_{cc}} \right) \; ln(ratio + 1)\, t = t^2 \times \left(k_{dp} + \frac{ln2}{t_{cc}} \right)$$

$$ln(ratio + 1)\, t = t^2 \times \left(k_{dp} + \frac{ln2}{t_{cc}} \right)$$



Consider *m* intermediate time points:

$$k_{dp} = \frac{\sum_{i=1}^{m} \log_e (r_{t_i} + 1) t_i}{\sum_{i=1}^{m} t_i^2} - \frac{\log_e 2}{t_{cc}} ,$$

From k_{dp} we get the desired half-life:

$$T_{1/2} = \frac{\log_e 2}{k_{dp}} .$$

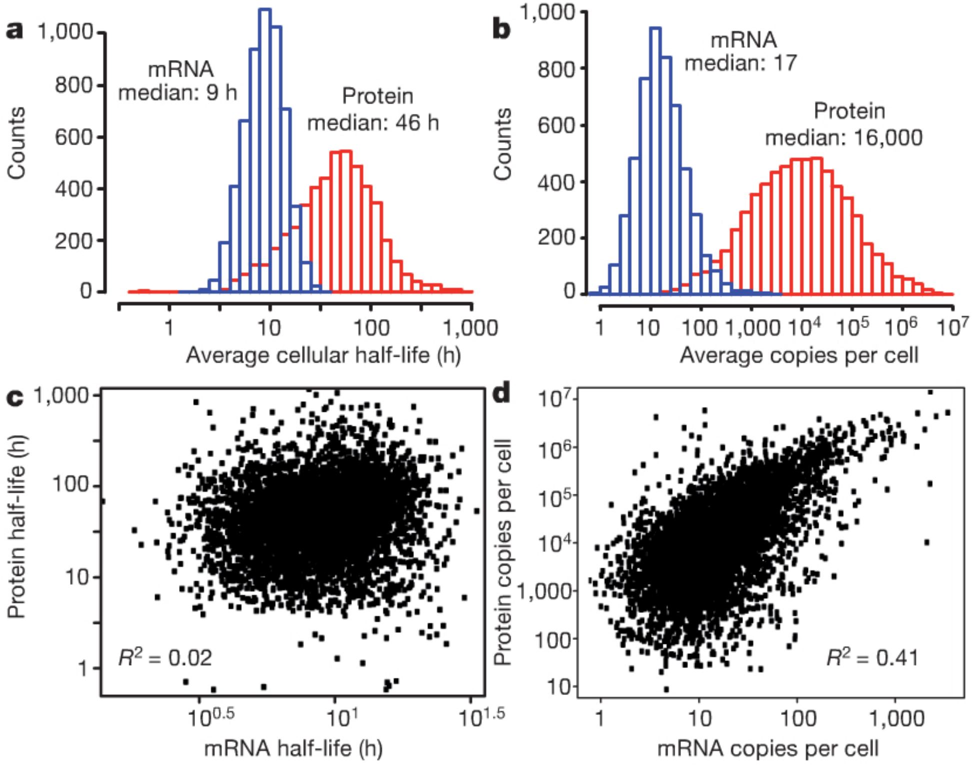
because this gives

$$P_L = P_0 e^{-k_{dp} t} = P_0 e^{-k_{dp} \frac{\log_e 2}{k_{dp}}} = P_0 e^{\log_e \frac{1}{2}} = \frac{1}{2} P_0$$

The same is done to compute mRNA half-lives (not shown).

Schwanhäuser et al. Nature 473, 337 (2011)

mRNA and protein levels and half-lives



a, b, Histograms of mRNA (blue) and protein (red) half-lives (a) and levels (b).

Proteins were on average 5 times more stable (9h vs. 46h) and 900 times more abundant than mRNAs and showed more variation.

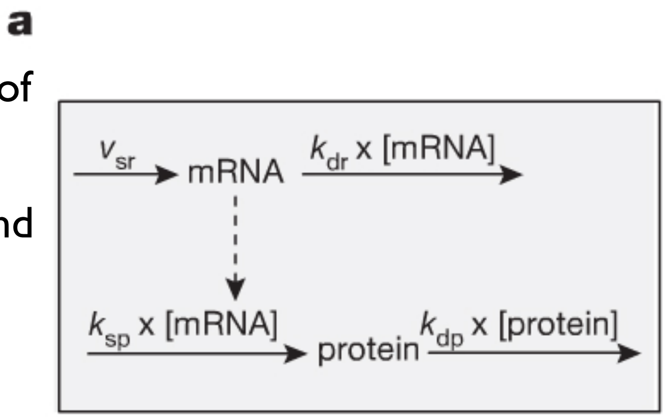
(right) mRNA and protein levels showed reasonable correlation (R^2 = 0.41)

(left) However, there was practically no correlation of protein and mRNA half-lives.

Schwanhäuser et al. Nature 473, 337 (2011)

Mathematical model of transcription and translation

A widely used minimal description of the dynamics of transcription and translation includes the synthesis and degradation of mRNA and protein, respectively



$$\frac{dR}{dt} = v_{sr} - k_{dr}R$$

$$\frac{dP}{dt} = k_{sp}R - k_{dp}P$$

The mRNA (*R*) is synthesized with a constant rate v_{sr} and degraded proportional to their numbers with rate constant k_{dr}.

The protein level (*P*) depends on the number of mRNAs, which are translated with rate constant k_{sp}.

Protein degradation is characterized by the rate constant k_{dp}.

The synthesis rates of mRNA and protein are calculated from their measured half lives and levels.

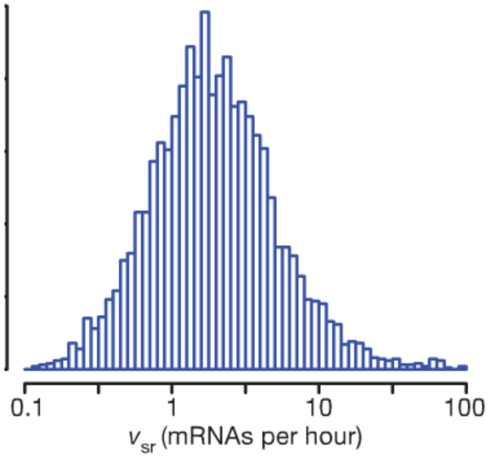
Schwanhäuser et al. Nature 473, 337 (2011)

Computed transcription and translation rates

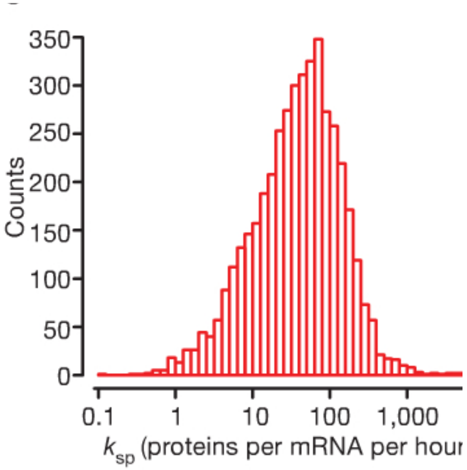
Average cellular transcription rates predicted by the model span two orders of magnitude.

The median is about 2 mRNA molecule per hour (**very slow**!).

An extreme example is the protein Mdn of which more than 500 mRNAs per hour are transcribed.



The median translation rate constant is about 40 proteins per mRNA per hour



Calculated translation rate constants are not uniform

Schwanhäuser et al. Nature 473, 337 (2011)

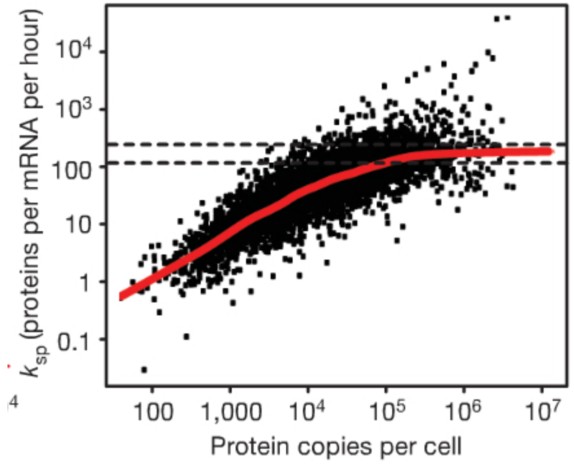
Maximal translation constant

Abundant proteins are translated about 100 times more efficiently than those of low abundance

Translation rate constants of abundant proteins saturate between approximately 120 and 240 proteins per mRNA per hour.

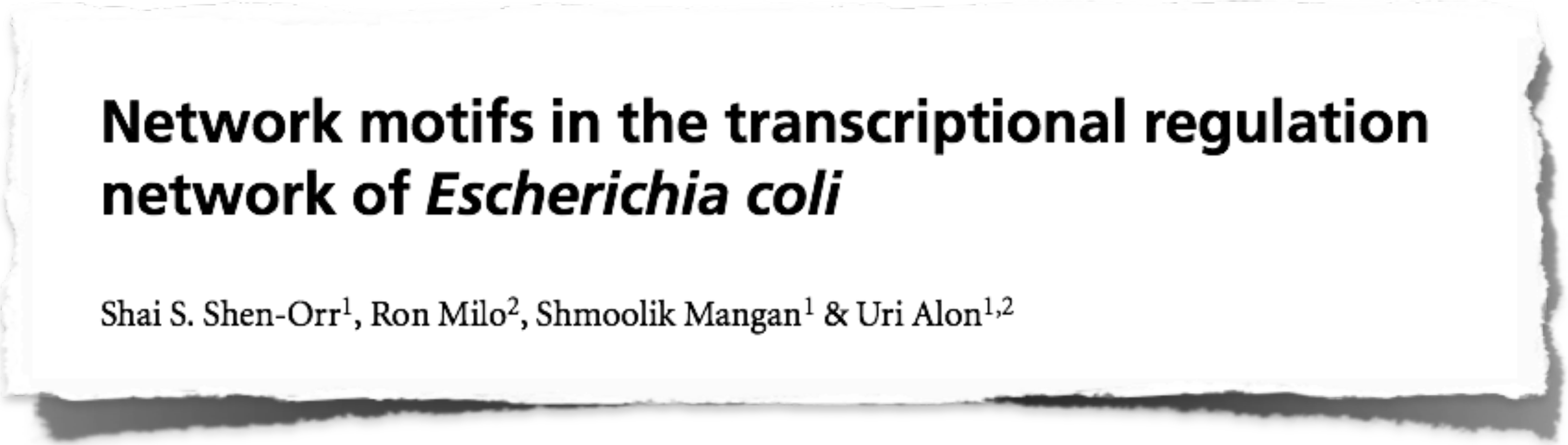
The maximal translation rate constant in mammals is not known.

The estimated maximal translation rate constant in sea urchin embryos is 140 copies per mRNA per hour, which is surprisingly close to the prediction of this model.



Schwanhäuser et al. Nature 473, 337 (2011)

Network Motifs



Network motifs in the transcriptional regulation network of *Escherichia coli*

Shai S. Shen-Orr[1], Ron Milo[2], Shmoolik Mangan[1] & Uri Alon[1,2]

Nature Genetics **31** (2002) 64

RegulonDB + their own hand-curated findings

→ break down network into motifs

→ statistical significance of the motifs?

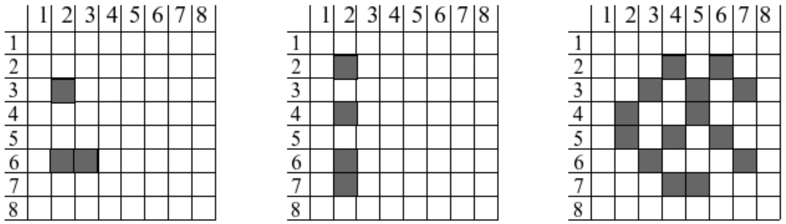
→ behavior of the motifs <=> location in the network?

Detection of motifs

Represent transcriptional network as a connectivity matrix M such that $M_{ij} = 1$ if operon j encodes a TF that transcriptionally regulates operon i and $M_{ij} = 0$ otherwise.

Scan all $n \times n$ submatrices of M generated by choosing n nodes that lie in a connected graph, for $n = 3$ and $n = 4$.

Submatrices were enumerated efficiently by recursively searching for nonzero elements.



	1	2	3	4	5	6	7	8
1								
2								
3		■						
4								
5								
6		■	■					
7								
8								

	1	2	3	4	5	6	7	8
1								
2		■						
3								
4		■						
5								
6		■						
7		■						
8								

	1	2	3	4	5	6	7	8
1								
2				■		■		
3			■		■		■	
4		■			■			
5		■		■		■		
6			■				■	
7				■	■			
8								

Connectivity matrix for causal regulation of transcription factor j (row) by transcription factor i (column). Dark fields indicate regulation.
(Left) Feed-forward loop motif. TF 2 regulates TFs 3 and 6, and TF 3 again regulates TF 6.
(Middle) Single-input multiple-output motif.
(Right) Densely-overlapping region.

For $n = 3$, the only significant motif is the feedforward loop.
For $n = 4$, only the overlapping regulation motif is significant.
SIMs and multi-input modules were identified by searching for identical rows of M.

Shen-Orr et al. Nature Gen. 31, 64 (2002)

Motif Statistics

Compute a p-value for submatrices representing each type of connected subgraph by comparing # of times they appear in real network vs. in random network.

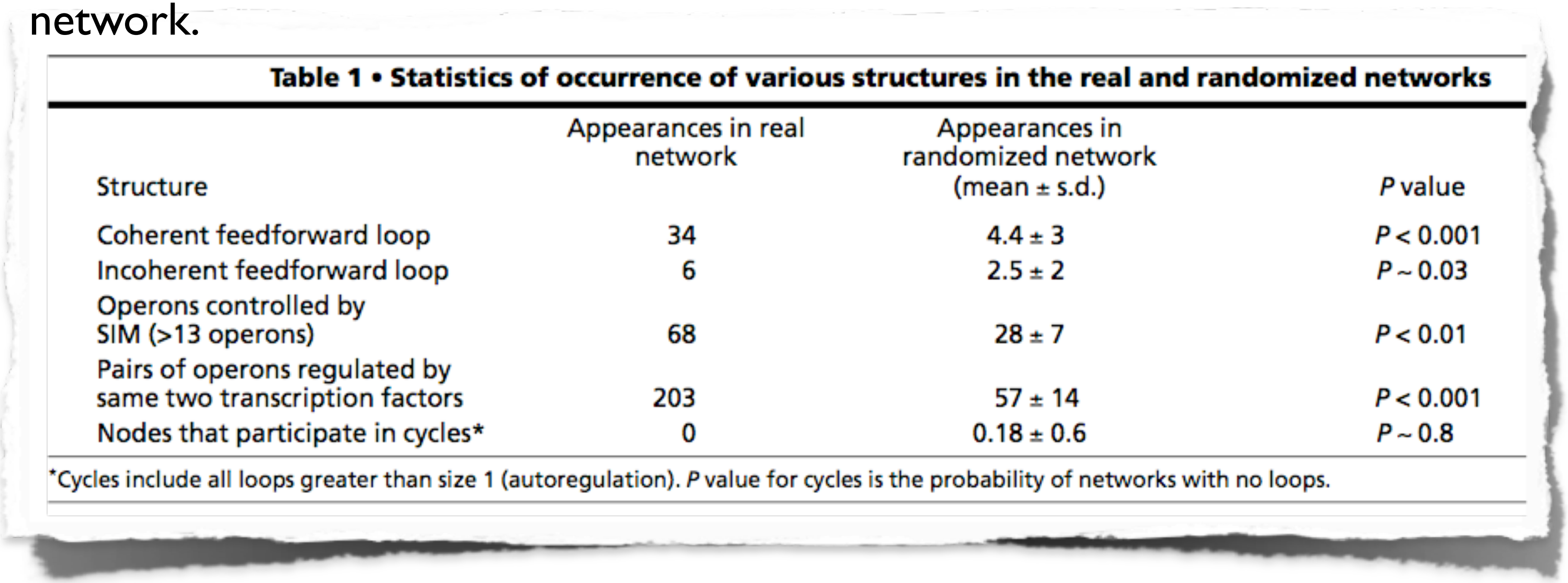


Table 1 • Statistics of occurrence of various structures in the real and randomized networks

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.

Listed motifs are highly **overrepresented** compared to randomized networks

No cycles (X → Y → Z → X) were identified,
but this was not statistically significant in
comparison to to random networks

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

Generate Random Networks

For a stringent comparison to randomized networks, one generates networks with precisely the same number of operons, interactions, transcription factors and number of incoming and outgoing edges for each node as in the real network (here the one from *E. coli*).

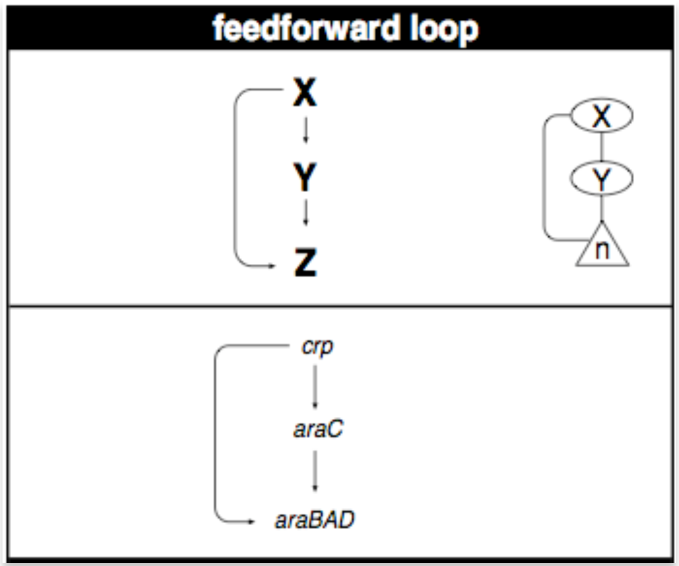
One starts with the real network and repeatedly swaps randomly chosen pairs of connections (*X1* → *Y1*, *X2* → *Y2* is replaced by *X1* → *Y2*, *X2* → *Y1*) until the network is well randomized.
This yields networks with precisely the same *n*umber of nodes with
p incoming and *q* outgoing nodes, as the real network.

The corresponding randomized connectivity matrices, *Mrand*, have the same number of nonzero elements in each row and column as the corresponding row and column of the real connectivity matrix *M*:

$$\sum_i Mrand_{ij} = \sum_i M_{ij} \quad \text{and} \quad \sum_j Mrand_{ij} = \sum_j M_{ij}$$

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

Motif 1: Feed-Forward-Loop



X = general transcription factor
Y = specific transcription factor
Z = effector operon(s)

Example for this in *E. coli*:
araBAD operon, encodes enzymes needed for the catabolism of arabinose

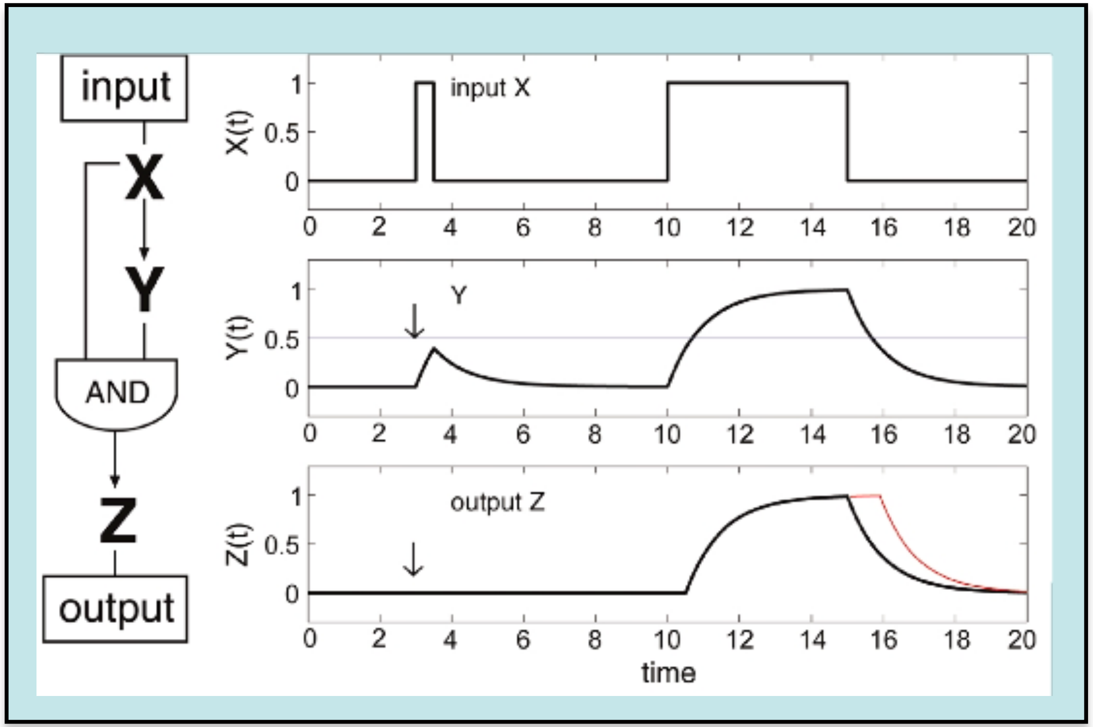
X and Y **together** regulate Z:

"**coherent**", if X and Y have the **same** effect on Z (activation vs. repression), otherwise "incoherent"

85% of the FFLs in *E. coli* are coherent

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

FFL dynamics



In a coherent FFL:
X **and** Y activate Z

Dynamics:

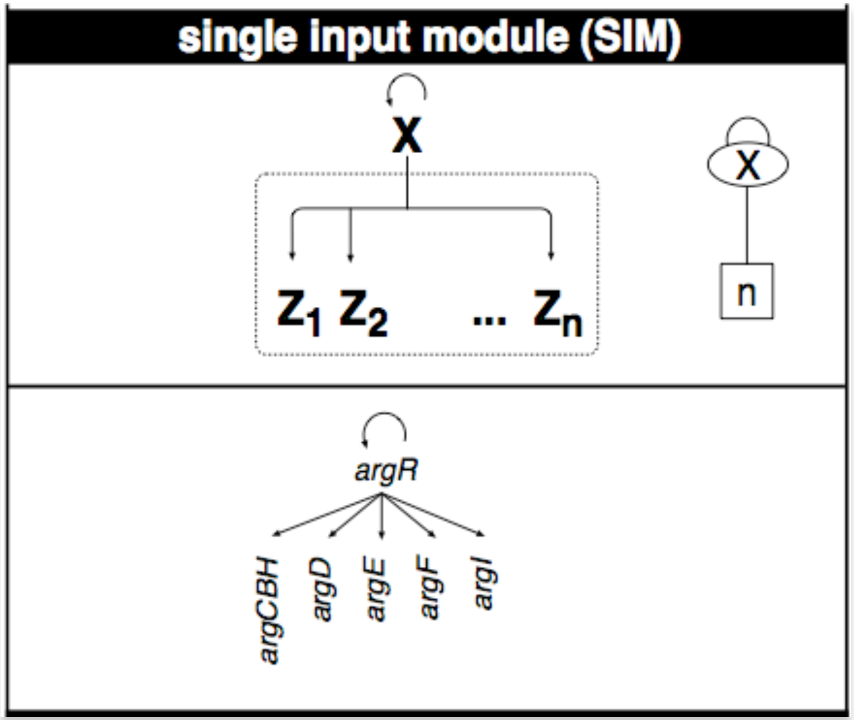
- input activates X
- X activates Y (delay)
- (X && Y) activates Z

Delay between X and Y → signal must persist longer than delay
→ reject transient signal, react only to **persistent** signals
→ enables fast shutdown

Helps with **decisions** based on **fluctuating signals**

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

Motif 2: Single-Input-Module



Set of operons controlled by a single transcription factor

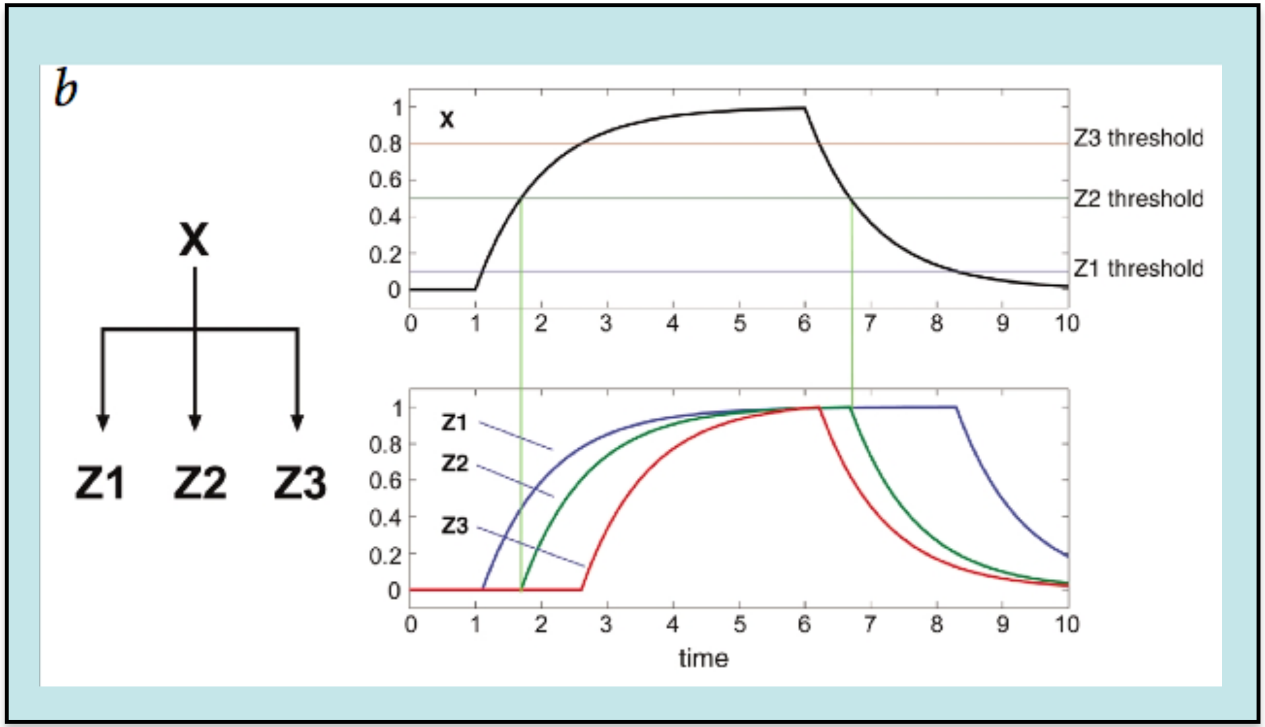
- same sign
- no additional regulation
- control is usually autoregulatory (70% vs. 50% overall)

<u>Example for this in *E. coli*:</u>
arginine biosynthetic operon *argCBH* plus other enzymes of arginine biosynthesis pathway

Mainly found in genes that code for **parts** of a protein **complex** or metabolic **pathway**
→ produces components in comparable amounts (stoichiometries)

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

SIM-Dynamics



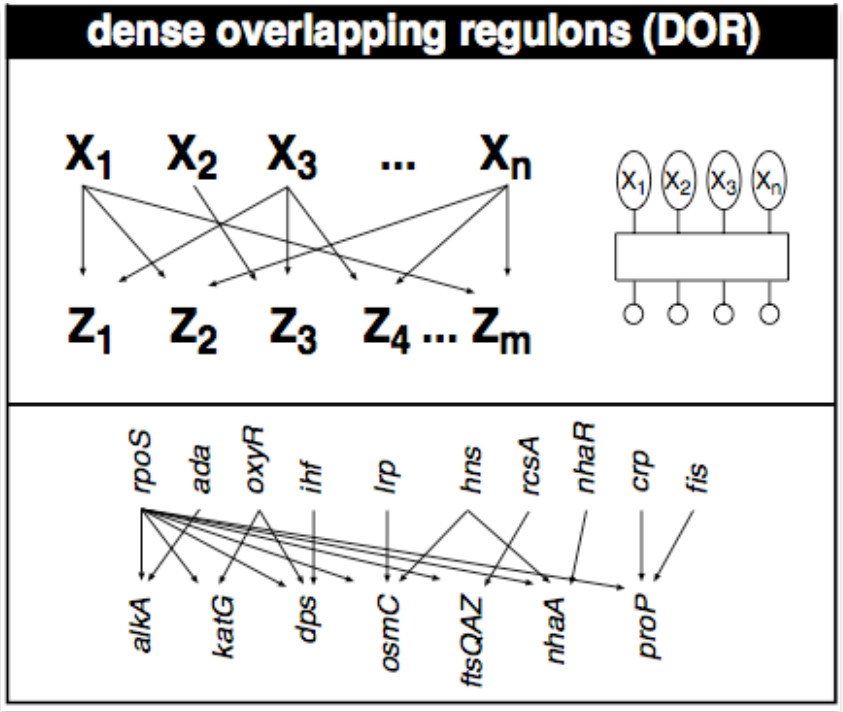
If different thresholds exist for each regulated operon:

→ first gene that is activated is the last that is deactivated

→ well defined temporal ordering (e.g. flagella synthesis) + stoichiometries

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

Motif 3: Densely Overlapping Regulon



Dense layer between groups of transcription factors and operons
→ much denser than network average (≈ community)

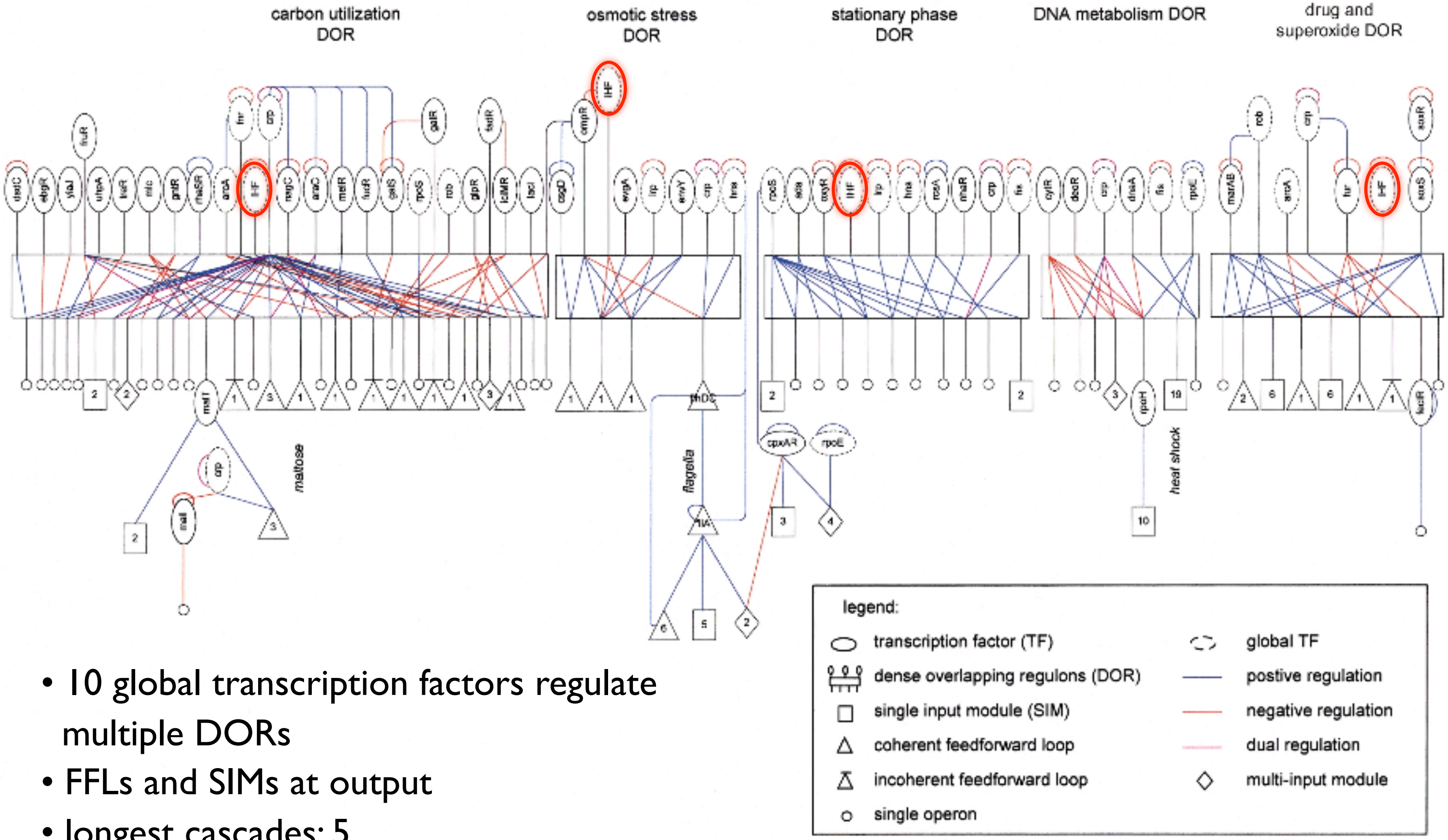
Usually each operon is regulated by a different combination of TFs.

Main "**computational**" **units** of the regulation system

Sometimes: same set of TFs for group of operons → "multiple input module"

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

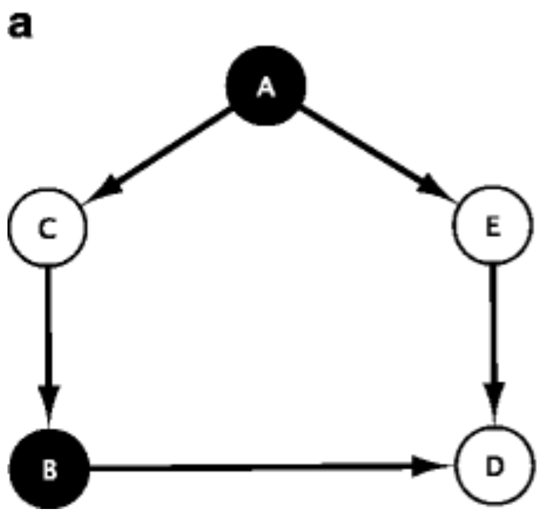
Network with Motifs



- 10 global transcription factors regulate multiple DORs
- FFLs and SIMs at output
- longest cascades: 5 (flagella and nitrogen systems)

Shen-Orr et al., *Nature Genetics* **31** (2002) 64

Identification of Master regulatory genes



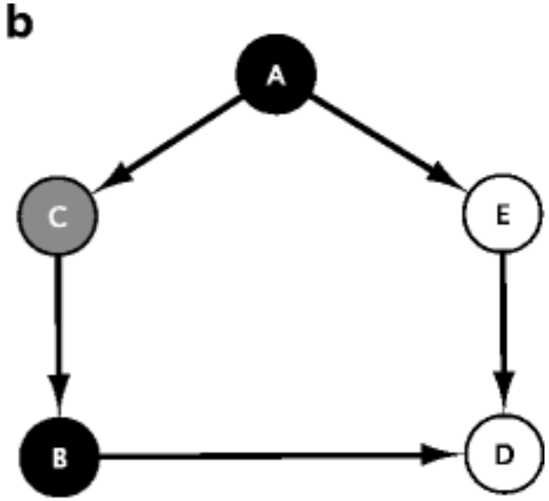
A vertex *u* **dominates** another vertex *v* if there exists a directed arc *(u,v)*.

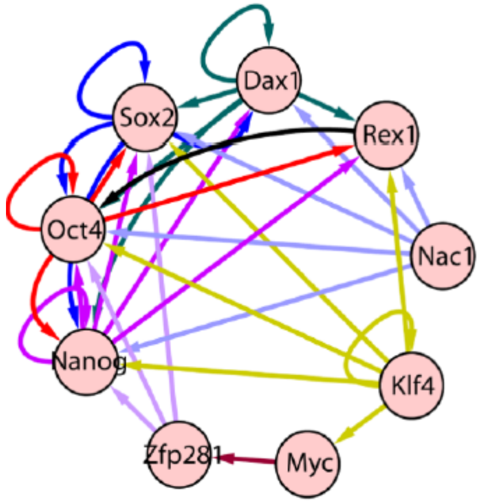
Idea: find a set of dominator nodes of minimum size that controls all other vertices. In the case of a GRN, a directed arc symbolizes that a transcription factor regulates a target gene.

In the figure, the MDS nodes {*A*,*B*} are the dominators of the network. Together, they regulate all other nodes of the network (*C*, *E*, *D*).

Nazarieh et al. BMC Syst Biol 10:88 (2016)

Identification of Master regulatory genes





Core pluripotency network,
Kim et al. Cell (2008)

The nodes of a MDS can be spread as isolates nodes over the entire graph. However, the set of core pluripotency factors is tightly connected (right).

Idea: find a connected dominating set of minimum size (MCDS).

(Left) the respective set of MCDS nodes (*black and gray*).
Here, node *C* is added in order to preserve the connection between the two dominators *A* and *B* to form an MCDS

ILP for minimum dominating set

Aim: we want to determine a set *D* of minimum cardinality such that for each *v* ∈*V*, we have that *v* ∈ *D* or that there is a node *u* ∈ *D* and an arc *(u,v)* ∈ *E*.

Let $\delta^-(v)$ be the set of incoming nodes of *v* such that *(u,v)* ∈ *E*, x_u and x_v are binary variables associated with *u* and *v*.

We select a node *v* as dominator if its binary variable x_v has value 1, otherwise we do not select it.

$$\begin{array}{lll} \text{minimize} & \sum_{v \in V} x_v & \\ \text{subject to} & x_u + \sum_{v \in \delta^-(u)} x_v \geq 1 & \forall u \in V \\ & x_v \in \{0,1\} & \forall v \in V \end{array}$$

With the GLPK solver, the runtime was less than 1 min for all considered networks.

Nazarieh et al. BMC Syst Biol 10:88 (2016)

ILP for minimum connected dominating set

A minimum connected dominating set (MCDS) for a directed graph G = (V,E) is a set of nodes $D \subseteq V$ of minimum cardinality that is a dominating set and additionally has the property that the graph *G[D]* induced by *D* is weakly connected, i.e. such that in the underlying undirected graph there exists a path between any two nodes of *D* that only uses vertices in *D*.

This time we will use two binary valued variables y_v and x_e .
y_v indicates whether node *v* is selected to belong to the MCDS.
x_e for the edges then yields a tree that contains all selected vertices and no vertex that was not selected.

$$\begin{aligned} &\text{minimize} && \sum_{v \in V} y_v \\ &\text{subject to} && \sum_{e \in E} x_e = \sum_{i \in V} y_i - 1 \end{aligned}$$

This guarantees that the number of edges is one less than the number of vertices. This is necessary (but not sufficient) to form a (spanning) tree.

Nazarieh et al. BMC Syst Biol 10:88 (2016)

ILP for minimum connected dominating set

$$\text{minimize} \quad \sum_{v \in V} y_v$$

$$\text{subject to} \quad \sum_{e \in E} x_e = \sum_{i \in V} y_i - 1$$

$$\sum_{e \in E(S)} x_e \leq \sum_{i \in S \setminus \{j\}} y_i \qquad \forall S \subset V, \forall j \in S$$

The second constraint implies that the selected edges imply a tree.

(Note that this defines an exponential number of constraints for all subgraphs of V!)

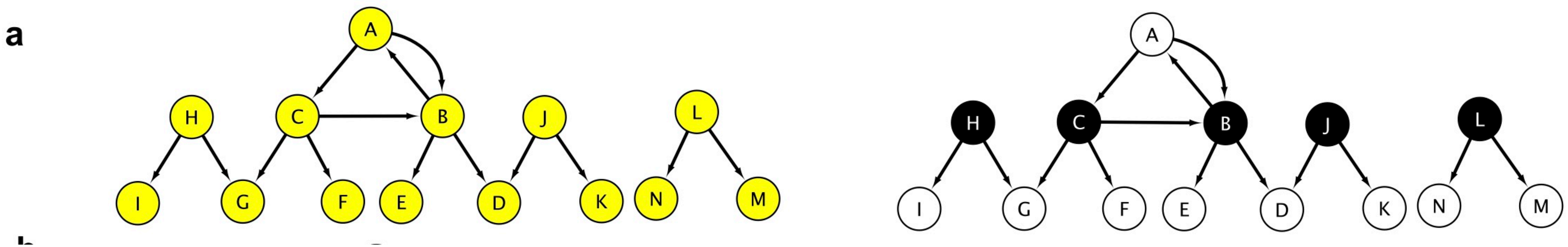
$$\begin{aligned} & y_u + \sum_{v \in \delta^-(u)} y_v \geq 1 && \forall u \in V \\ & y_v \in \{0,1\} && \forall v \in V \\ & x_e \in \{0,1\} && \forall e \in E \end{aligned}$$

The third constraint guarantees that the set of selected nodes in the solution forms a dominating set of the graph.

For dense graphs, this yields a quick solution. However, for sparse graphs, the running time may be considerable. Here we used an iterative approach (not presented).

Nazarieh et al. BMC Syst Biol 10:88 (2016)

Example MDS

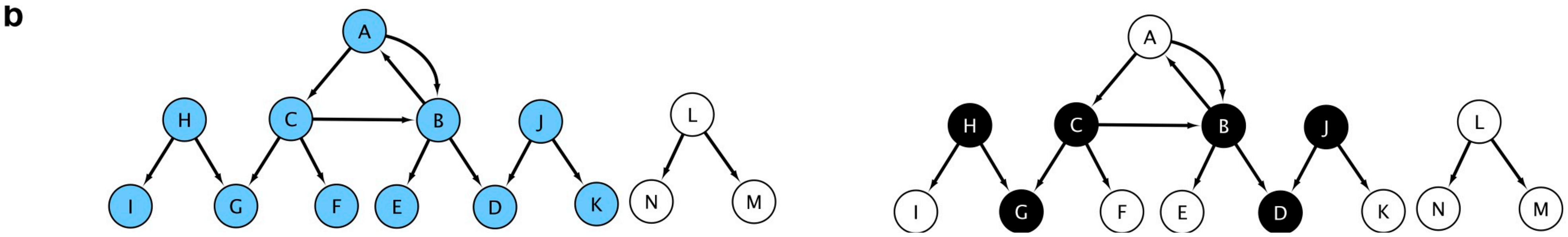


(Left) this toy network includes 14 nodes and 14 edges.

(Right) The dark colored nodes {*J*, *B*, *C*, *H*, *L*} are the dominators of the network obtained by computing a MDS.

Nazarieh et al. BMC Syst Biol 10:88 (2016)

Example MCDS

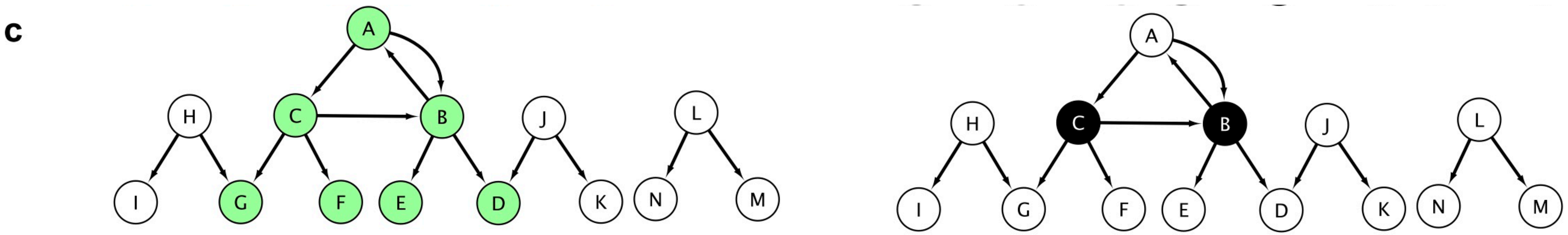


(Left) The nodes colored blue make up the **largest connected component** (LCC) of the underlying **undirected** graph.

(Right) MCDS nodes for this component are {*J, D, B, C, G, H*}.

Nazarieh et al. BMC Syst Biol 10:88 (2016)

Example MCDS

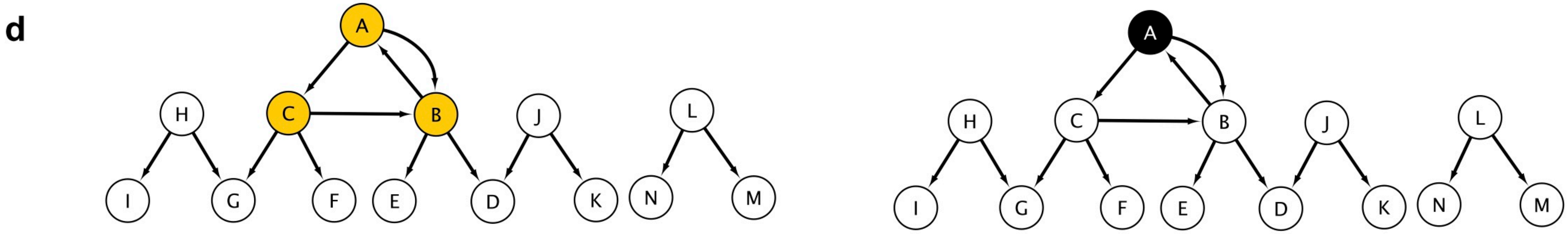


(Left) The green colored nodes are elements of **the largest connected component** underlying the **directed** graph.

(Right) The two nodes {*B*, *C*} form the MCDS for this component.

Nazarieh et al. BMC Syst Biol 10:88 (2016)

MCDS of the strongly connected component



(Left) The nodes colored orange show the LSCC in the network.

(Right) The node *A* is the only element of the MCDS

Nazarieh et al. BMC Syst Biol 10:88 (2016)

Studied networks: RegulonDB (E.coli)

This GRN contains 1807 genes, including 202 TFs and 4061 regulatory interactions. It forms a general network which controls all sorts of responses which are needed in different conditions.

Due to the sparsity of the network, its MDS contains 199 TFs.

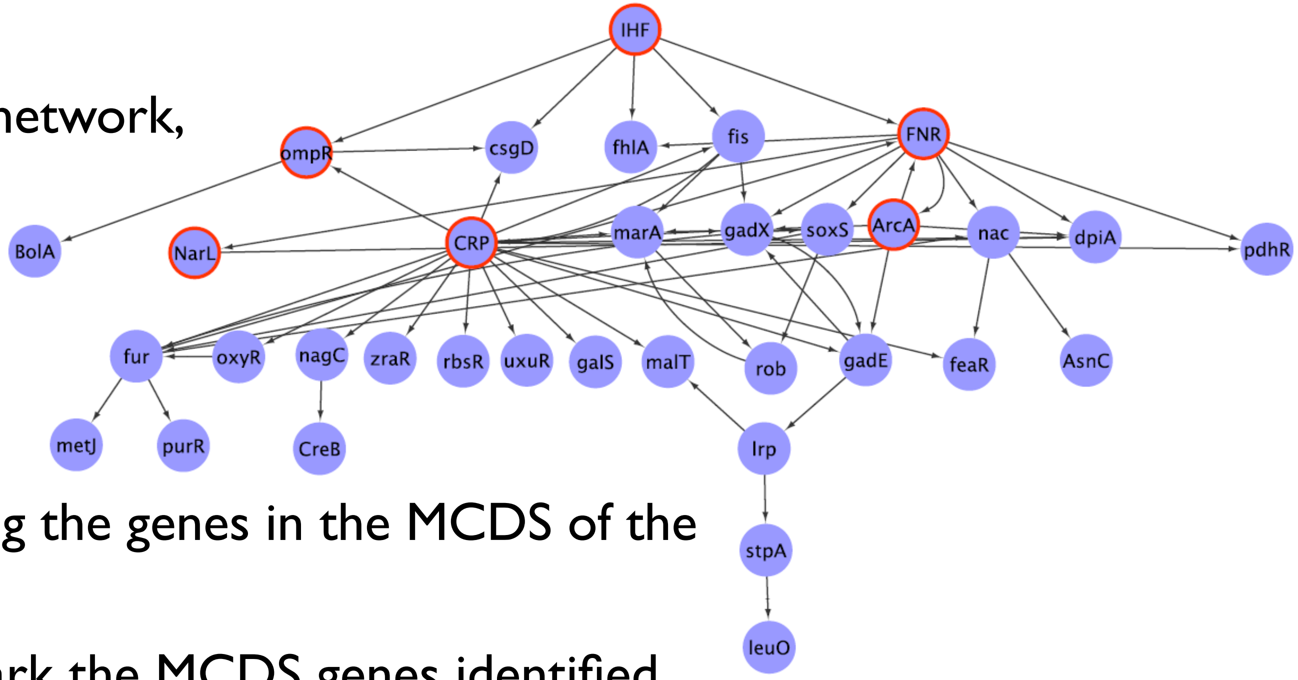


Figure: Connectivity among the genes in the MCDS of the LCC of the *E.coli* GRN.
The red circle borders mark the MCDS genes identified as global regulators by Ma *et al.* (see lecture V7).

Periodic genes in cell cycle network of yeast

Take regulatory data from Yeast Promoter Atlas (YPA).
It contains 5026 genes including 122 TFs.

From this set of regulatory interactions, we extracted a cell-cycle specific subnetwork of 302 genes that were differentially expressed along the cell cycle of yeast (MA study by Spellman et al. Mol Biol Cell (1998)).

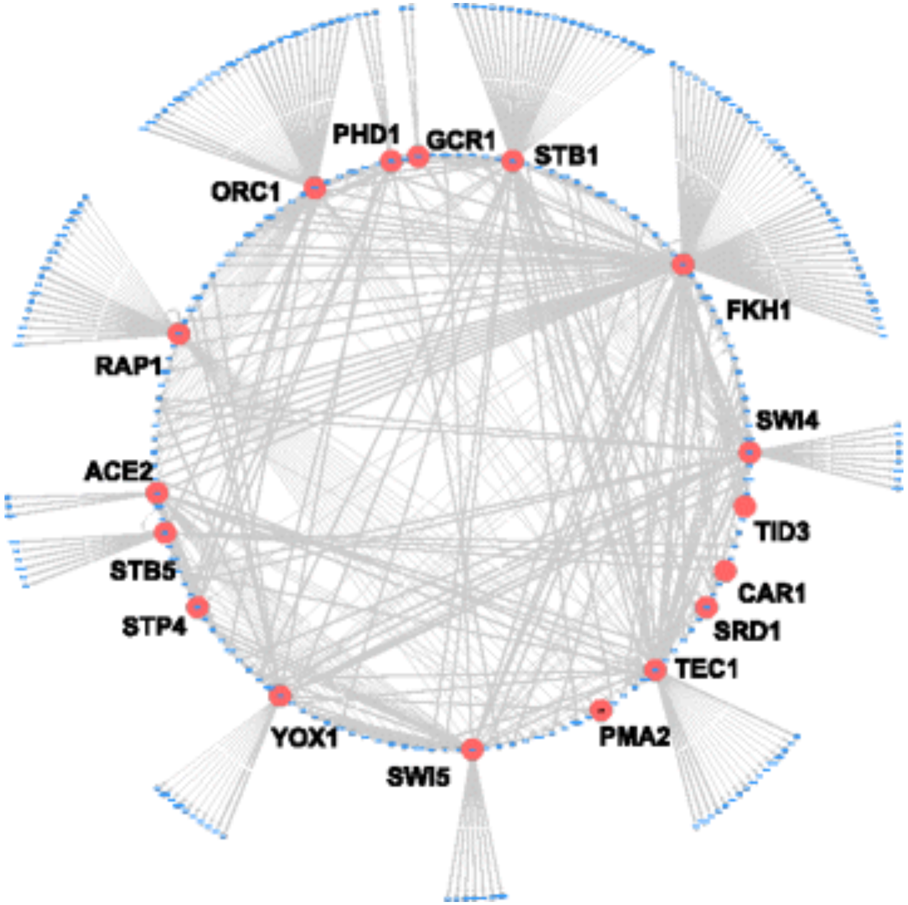
Nazarieh et al. BMC Syst Biol 10:88 (2016)

MCDS of cell cycle network of yeast

Tightly interwoven network of 17 TFs and target genes that organize the cell cycle of *S. cerevisiae*.

Shown on the circumference of the outer circle are 164 target genes that are differentially expressed during the cell cycle and are regulated by a TF in the MCDS (shown in the inner circle).

The inner circle consists of the 14 TFs from the heuristic MCDS and of 123 other target genes that are regulated by at least two of these TFs

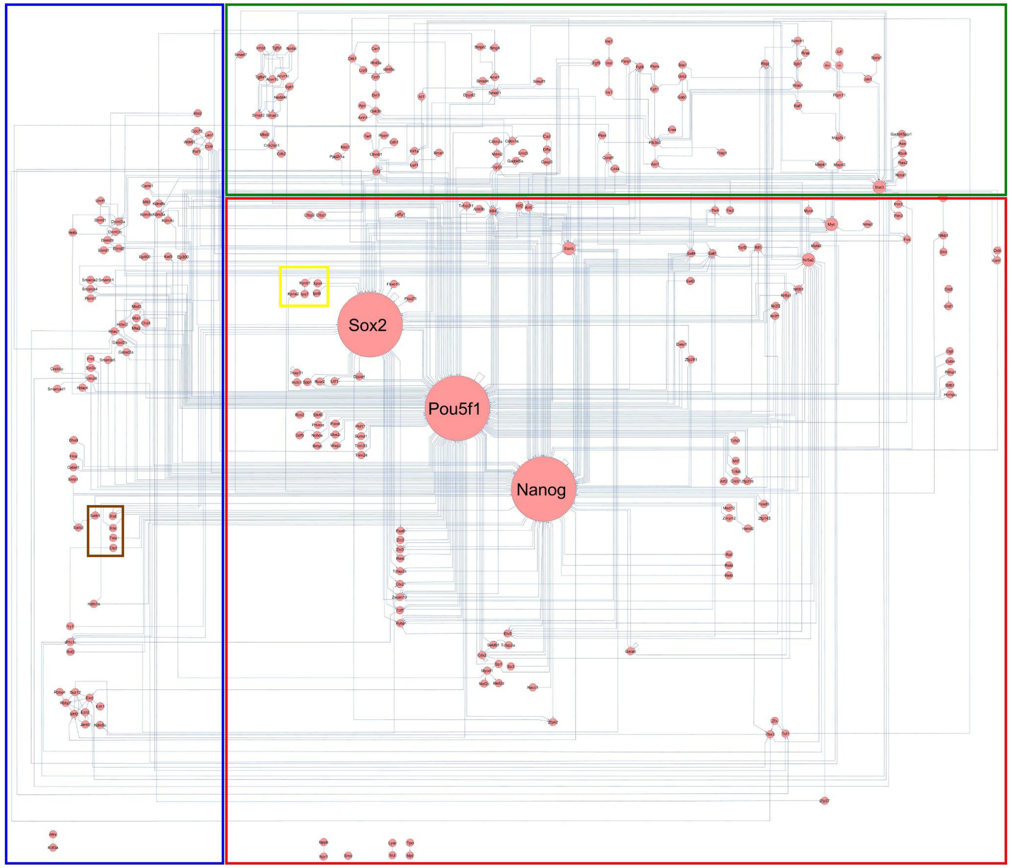


Nazarieh et al. BMC Syst Biol 10:88 (2016)

Studied networks: PluriNetwork

PluriNetWork was manually assembled as an interaction/ regulation network describing the molecular mechanisms underlying pluripotency.

It contains 574 molecular interactions, stimulations and inhibitions, based on a collection of research data from 177 publications until June 2010, involving 274 mouse genes/proteins.



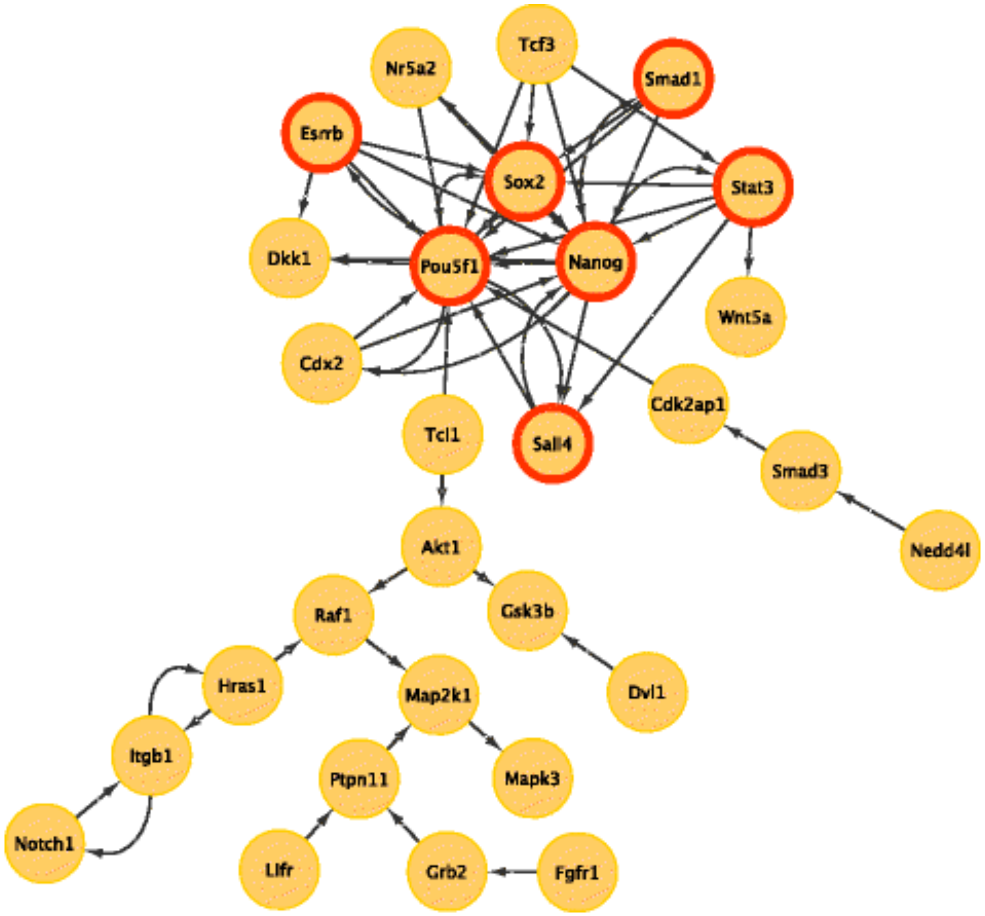
Som A, et al. (2010) PLoS ONE 5: e15165.

MCDS of mouse pluripotency network

Connectivity among TFs in the heuristic MCDS of the largest strongly connected component of a GRN for mouse ESCs.

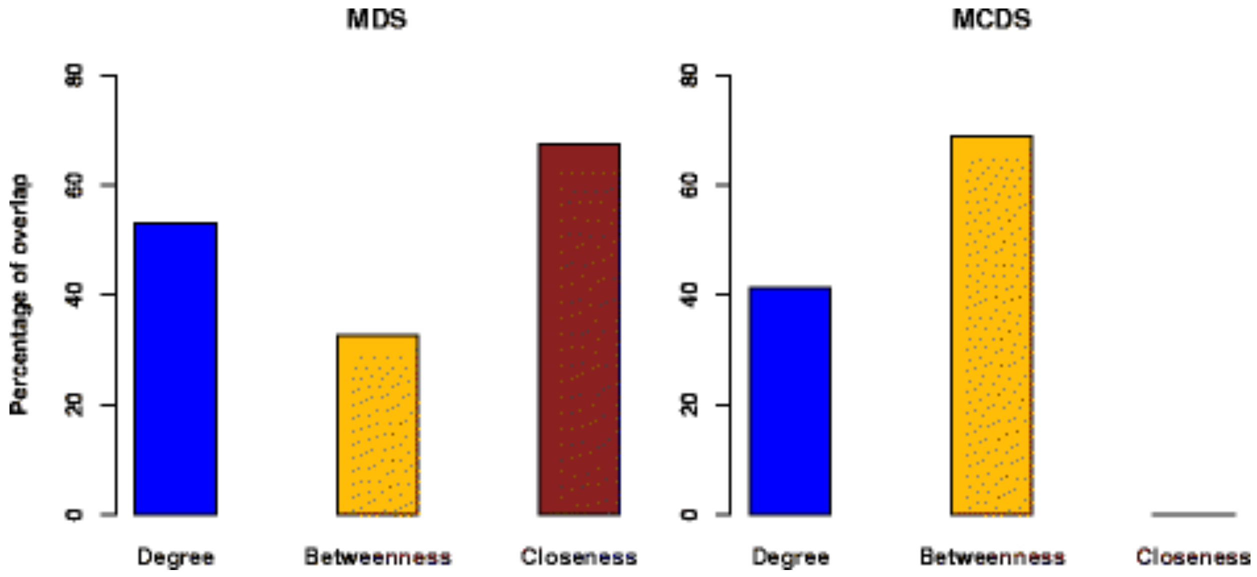
The red circle borders mark the 7 TFs belonging to the set of master regulatory genes identified experimentally.

The MCDS genes were functionally significantly more homogeneous than randomly selected gene pairs of the whole network (p = 6.41e-05, Kolmogorov-Smirnow test).



Nazarieh et al. BMC Syst Biol 10:88 (2016)

Overlap with most central nodes



Percentage overlap of the genes of the MDS and MCDS with the list of top genes (same size as MCDS) according to 3 centrality measures. Shown is the percentage of genes in the MDS or MCDS that also belong to the list of top genes with respect to degree, betweenness and closeness centrality

MDS nodes tend to be central in the network (high closeness) and belong to the most connected notes (highest degree).

When considering only outdegree nodes in the directed network, most of the top nodes of the MCDS have the highest overlap with the top nodes of the degree centrality and the betweenness centrality
(→ connector nodes).

Nazarieh et al. BMC Syst Biol 10:88 (2016)

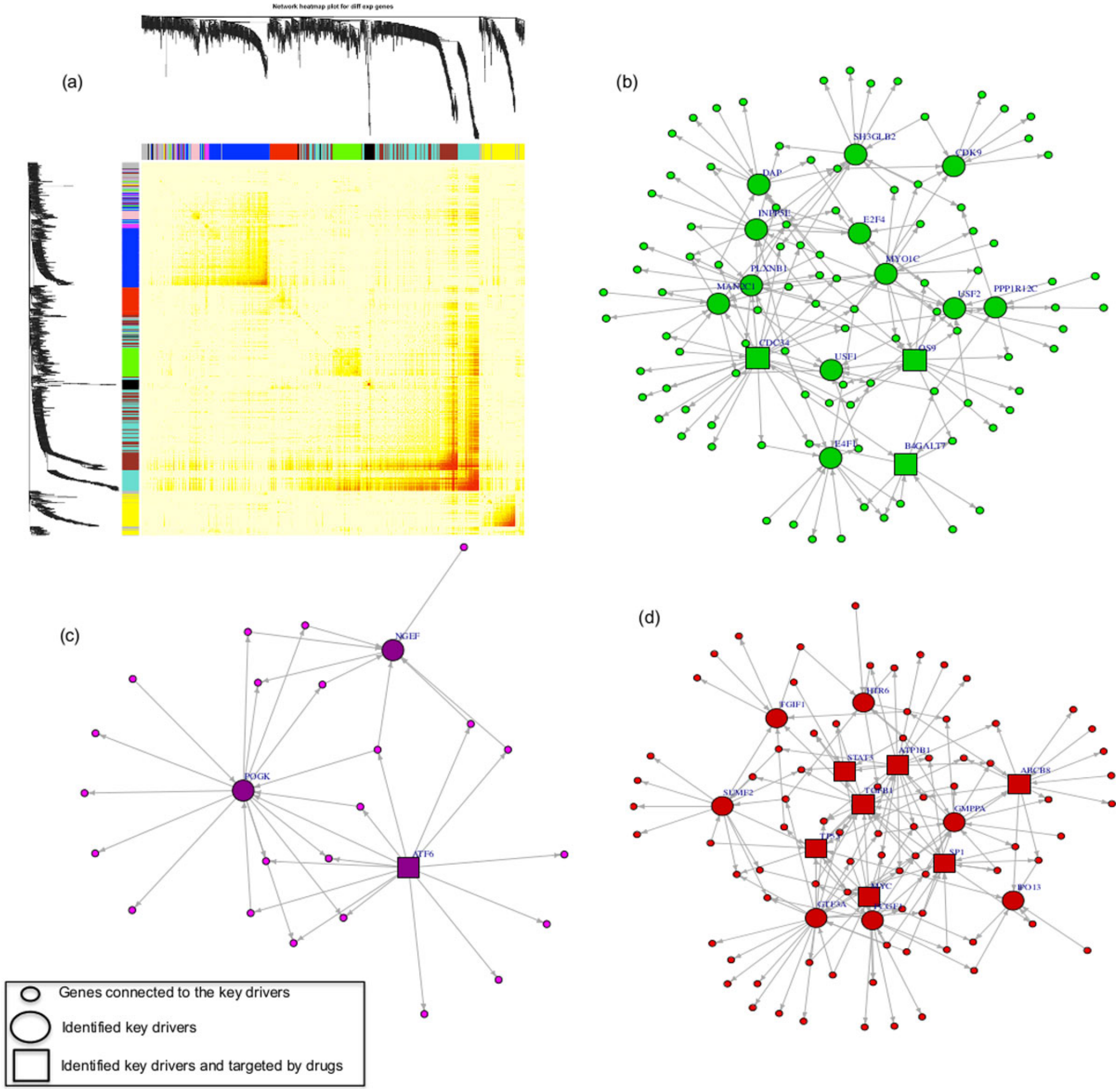
Breast cancer network

Analyze breast cancer data from TCGA → ca. 1300 differentially expressed genes.

Hierarchical clustering of co-expression network yielded 10 segregated network modules that contain between 26 and 295 gene members.

Add regulatory info from databases Jaspar, Tred, MSigDB.

(b) – (d) are 3 modules.



Hamed et al. BMC Genomics 16 (Suppl5):S2 (2015)

Breast cancer network

The MDS and MCDS sets of the nine modules
contain 68 and 70 genes, respectively.

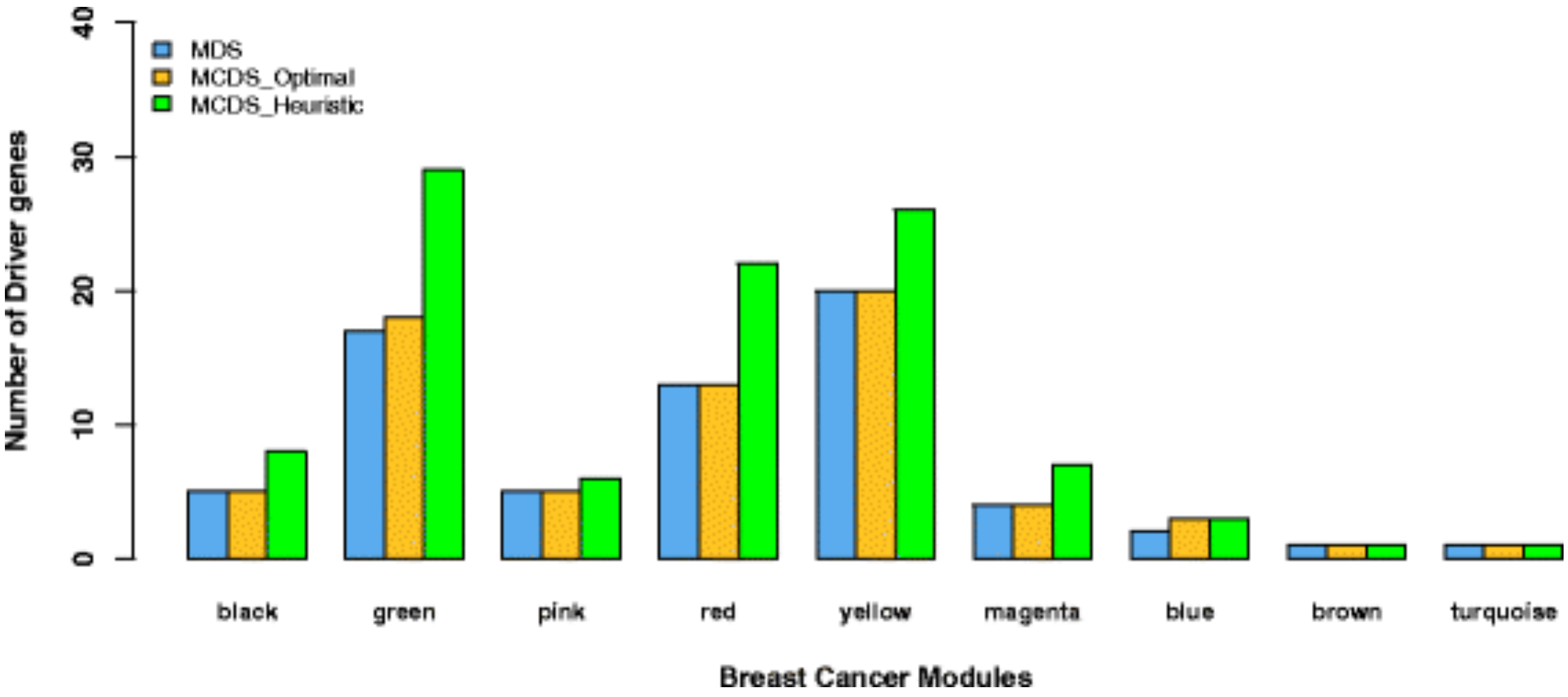
Intersect the proteins encoded by these genes with the targets of anti-cancer drugs.

20 of the 70 proteins in the MCDS are known drug targets (p = 0.03, hypergeometric test against the network with 1169 genes including 228 drug target genes).

Also, 16 out of the 68 proteins belonging to the MDS genes are binding targets of at least one anti-breast cancer drug.

Nazarieh et al. BMC Syst Biol 10:88 (2016)

|MDS| ≤ |MCDS|



Number of MCDS genes determined by the heuristic approach or by the ILP formulation and in the MDS.

Shown are the results for 9 modules of the breast cancer network

Nazarieh et al. BMC Syst Biol 10:88 (2016)

Summary

Today:

- mRNA and protein half-lifes and synthesis rates can be measured experimentally with SILAC MS

- Network **motifs**: FFLs, SIMs, DORs are overrepresented
→ different functions, different temporal behavior

- MDS and MCDS identify candidate master regulatory genes
→ who reliable are they when applied to noisy and incomplete data?

Next lecture:

- overview of methods to construct GRNs from experimental data

- benchmarking of GRN methods based on synthetic data