Biological networks: Motifs and modules

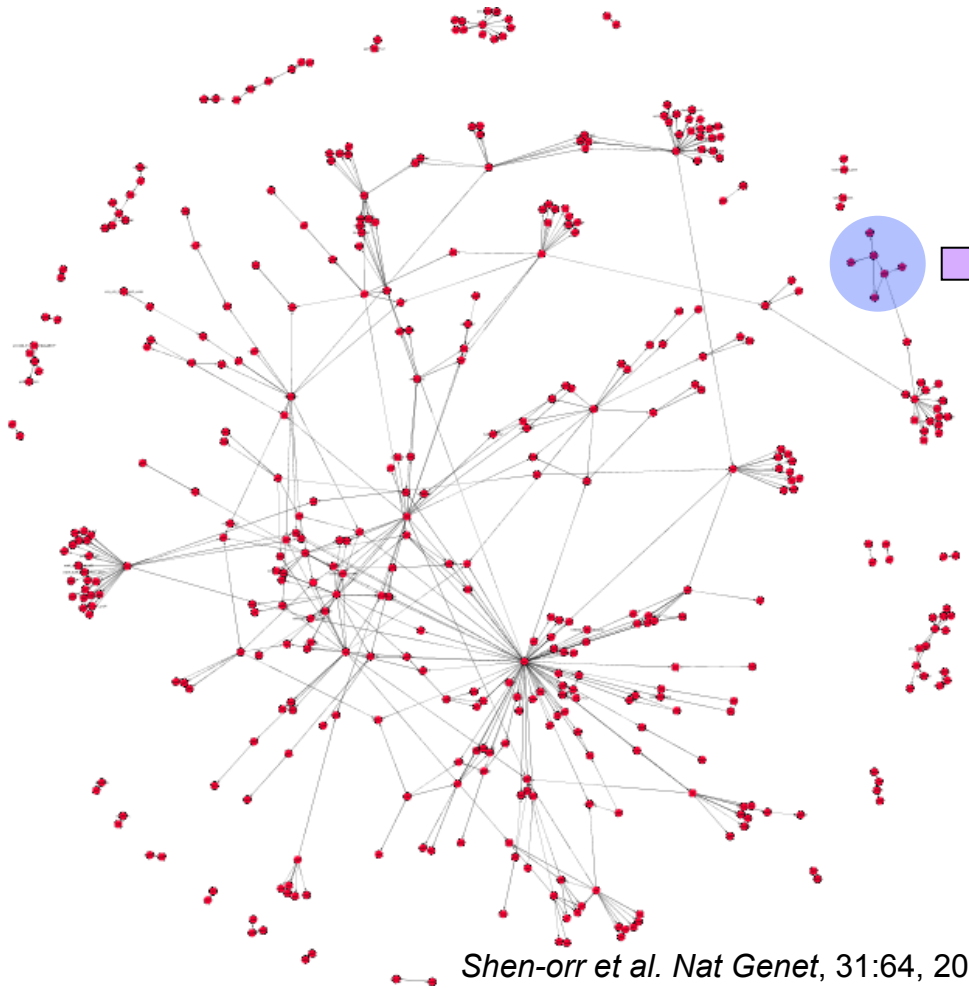
Bing Zhang

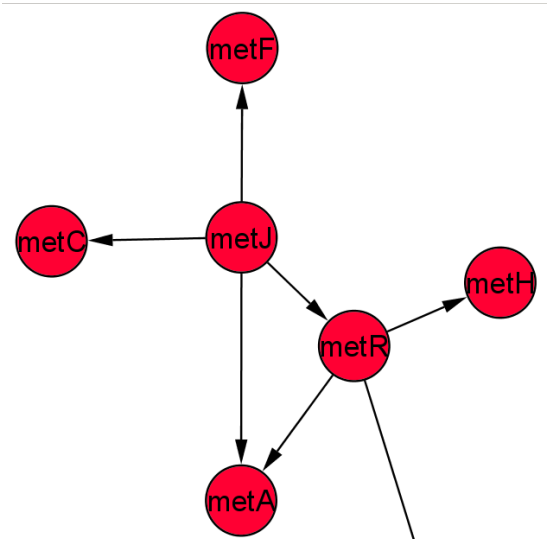
Department of Biomedical Informatics

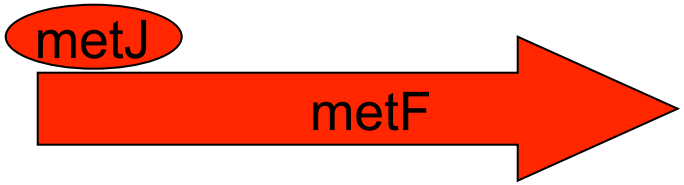
Vanderbilt University

bing.zhang@vanderbilt.edu

E.coli transcriptional regulatory network

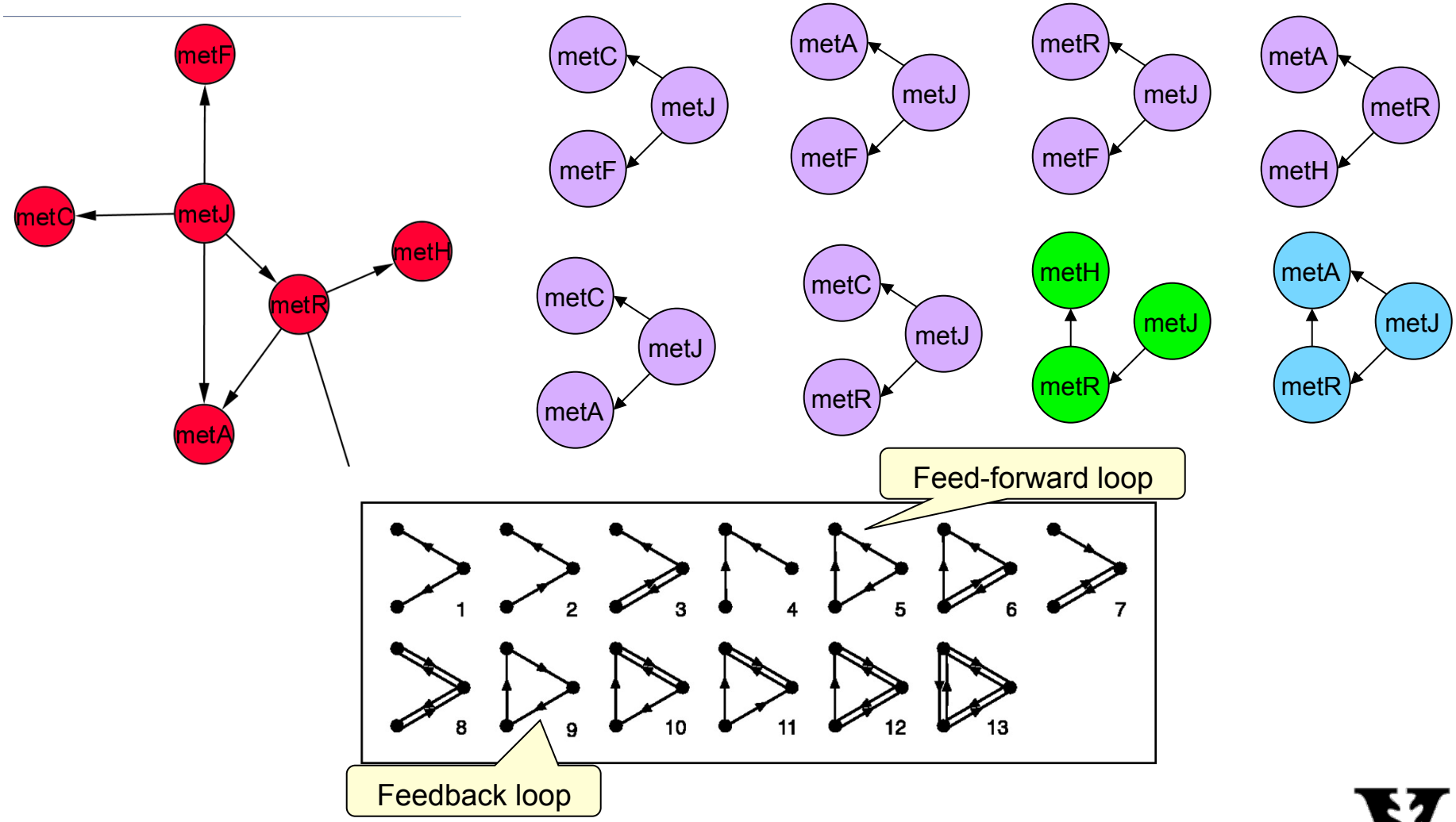






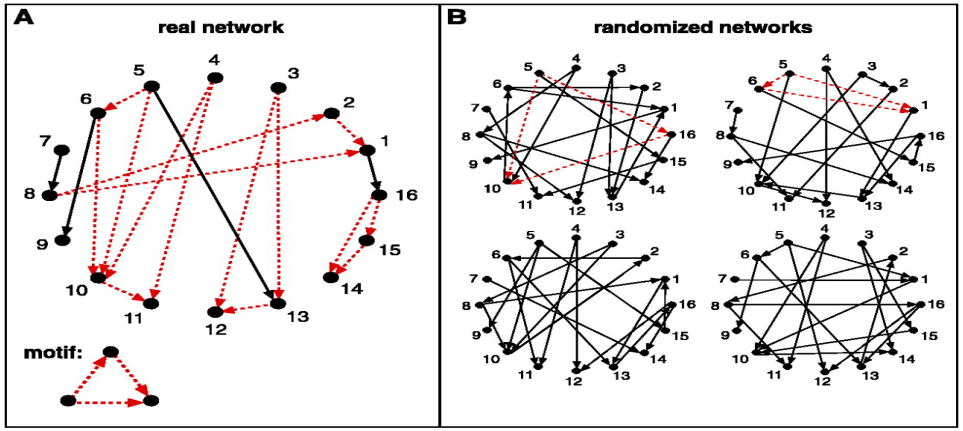
Shen-orr et al. Nat Genet, 31:64, 2002

Three-node patterns



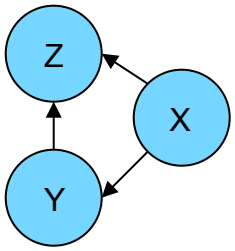
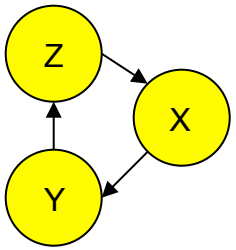
Network motifs

- **Network motifs:** Patterns that occur in the real network significantly more often than in randomized networks.
- Random network generation
  - ER model
  - Degree-preserving



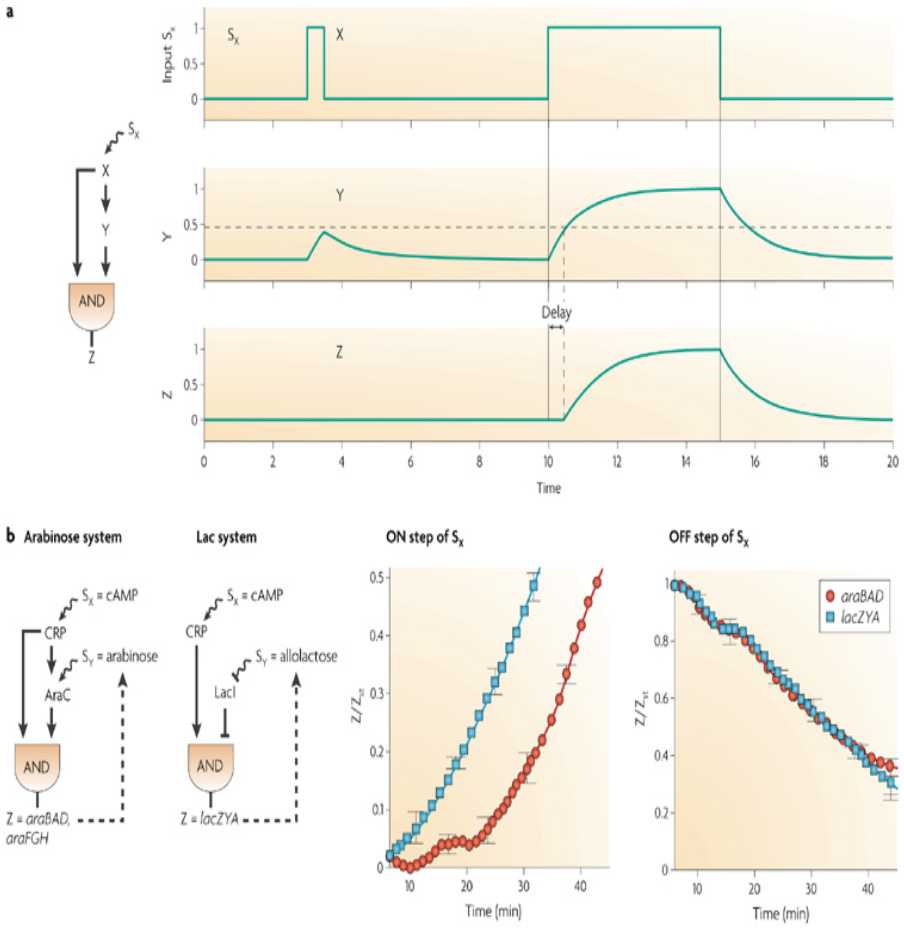
Milo et al., Science, 298:824, 2002

Feed forward loop (FFL) is a network motif in *E.coli* transcriptional regulatory network

	Feed-Forward Loop (FFL)	Feedback Loop
		
E. coli	42	0
ER random nets	1.7 ± 1.3 (Z=31)	0.6 ± 0.8
Degree-preserving random nets	7 ± 5 (Z=7)	0.2 ± 0.6

Alon, An introduction to system biology, 2007

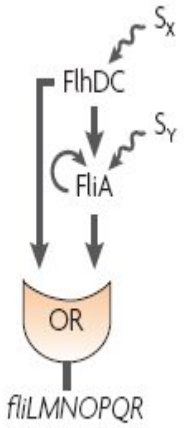
FFL with an AND input function



- Signal Sx appears
- X is activated and rapidly binds its downstream promoters
- Y begins to accumulate
- Z production starts only when Y concentration crosses the activation threshold for the Z promoter (**turn on: delay**)
- The higher the activation threshold, the longer the delay
- **The delay is useful in filtering out brief spurious pulses of signal**
- Sx is removed
- X rapidly becomes inactive and Z production stops (**turn off: no delay**)
- Arabinose system vs lactose system

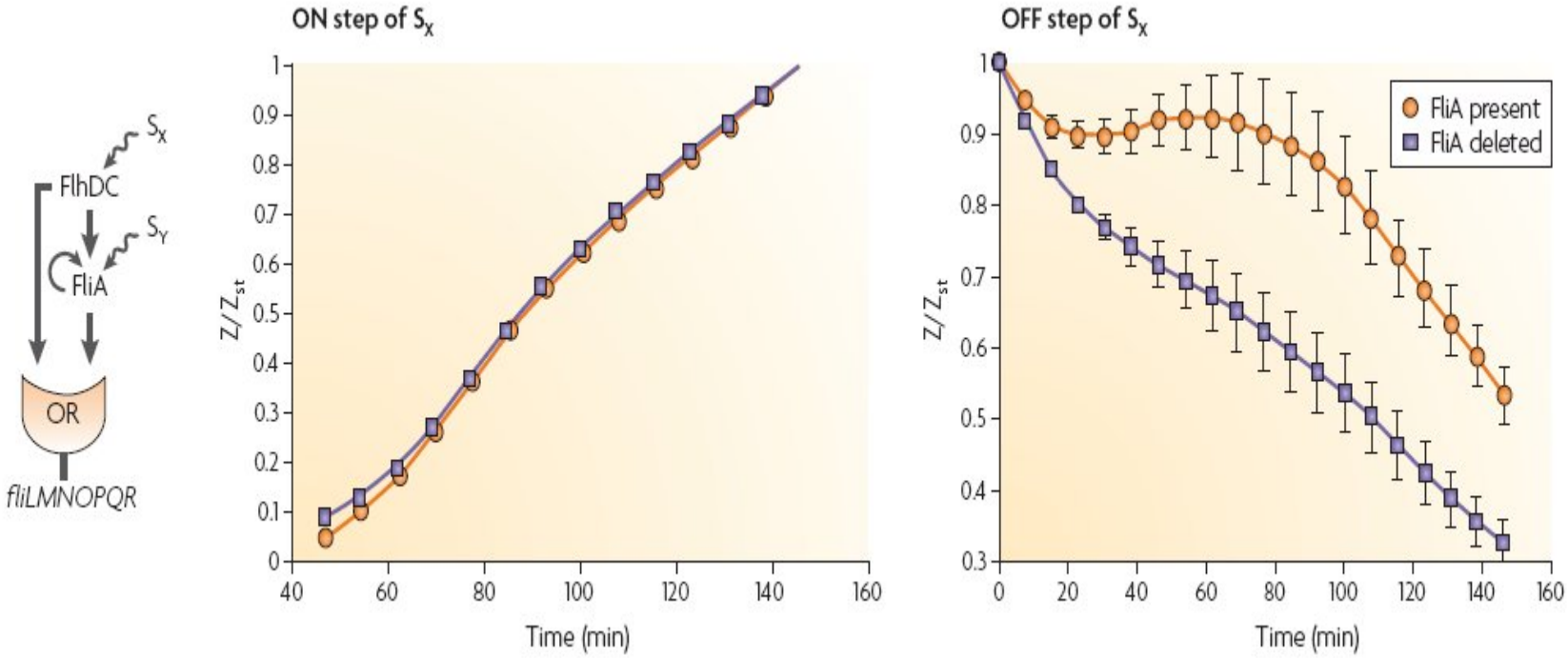
Alon, Nat Rev Genet, 8:450, 2007

FFL with an OR input function



Alon, Nat Rev Genet, 8:450, 2007

FFL with an OR input function

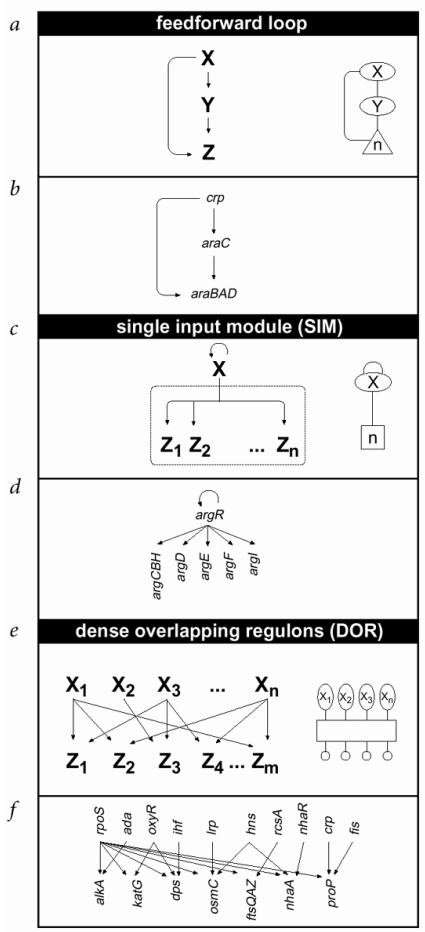


Experiments on the flagella system of *E.coli*

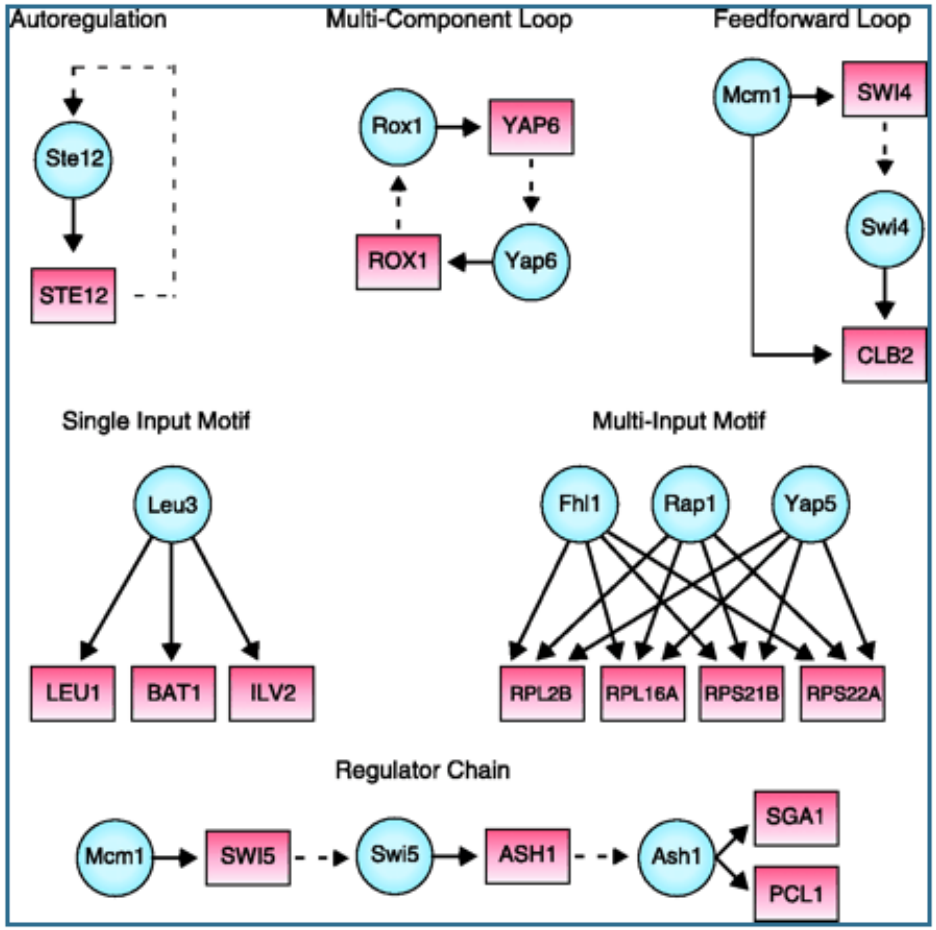
Alon, Nat Rev Genet, 8:450, 2007

Other types of network motifs in transcriptional regulatory networks

E. coli



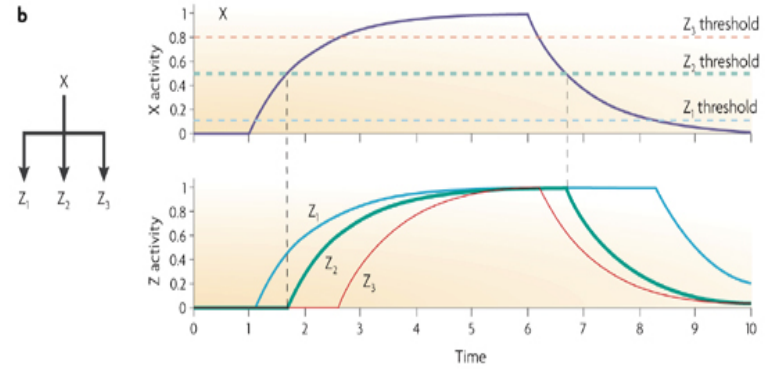
Shen-orr et al. Nat Genet, 31:64, 2002



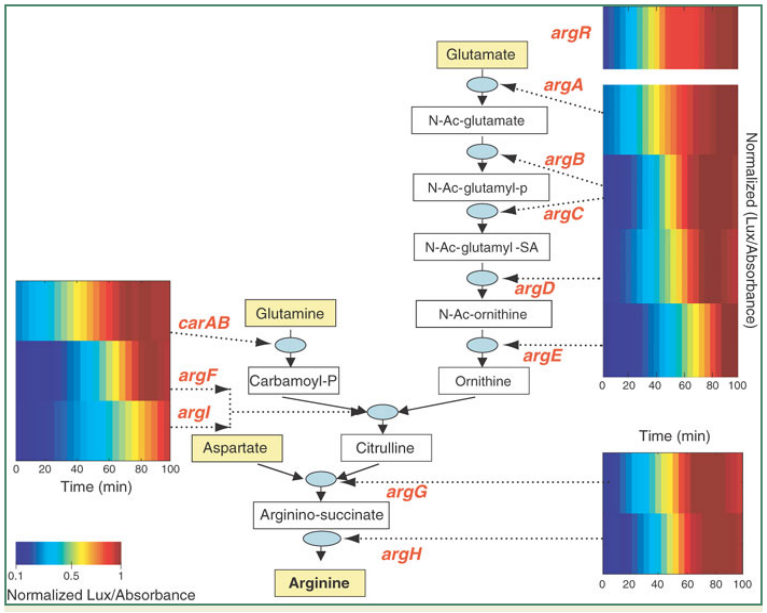
Yeast

Lee et al. Science, 298:799, 2002

Single Input Motif (SIM) generates a temporal expression program



Alon, Nat Rev Genet, 8:450, 2007



Zaslaver et al. Nat Genet, 36:486, 2004

- The activity of X gradually rises
- It crosses the different thresholds for each target promoter in a defined order to generate a temporal order
- Target genes: last-in-first-out (LIFO)
- Arginine production pathway: the earlier the protein functions in the pathway, the earlier its gene is activated
- Other biological processes
  - Cell cycle
  - Circadian clock
  - Developmental processes

Motifs in the yeast protein interaction network and their evolutionary conservation

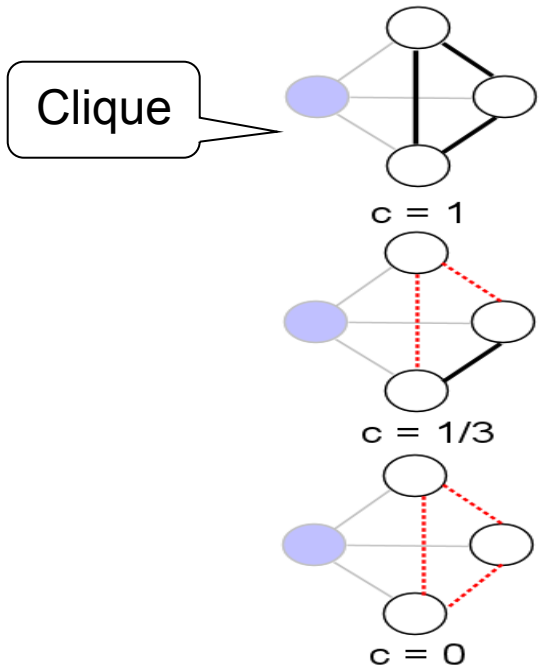
#	Motifs	Number of yeast motifs	Natural conservation rate	Random conservation rate	Conservation ratio
1		9,266	13.67%	4.63%	2.94
2		167,304	4.99%	0.81%	6.15
3		3,846	20.51%	1.01%	20.28
4		3,649,591	0.73%	0.12%	5.87
5		1,763,891	2.64%	0.18%	14.67
6		9,646	6.71%	0.17%	40.44
7		164,075	7.67%	0.17%	45.56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

Wuchty et al. Nat Genet, 35:176, 2003

Clique: a completely connected subgraph

- **Natural conservation rate:** percentage of motif instances in which all component proteins have an ortholog in all five eukaryotes investigated, indicating the evolutionary pressure to maintain the motif instances.
- **Random conservation rate:** conservation rate calculated when the same number of orthologs were randomly placed on the yeast protein interaction network, with no correlation between the network topology and the ortholog position.

Clustering coefficient: measurement of cliquishness



- Measures the local cohesiveness of a network
- If a node has *k* neighbors, at most *k* (*k*-1)/2 edges are allowed to exist between these neighbors.
- Clustering coefficient for a node is the fraction of the allowable edges that actually exist around the node.
- Clustering coefficient for a network is the average of all node clustering coefficients.

Watts and Strogatz, Nature, 393:440, 1998

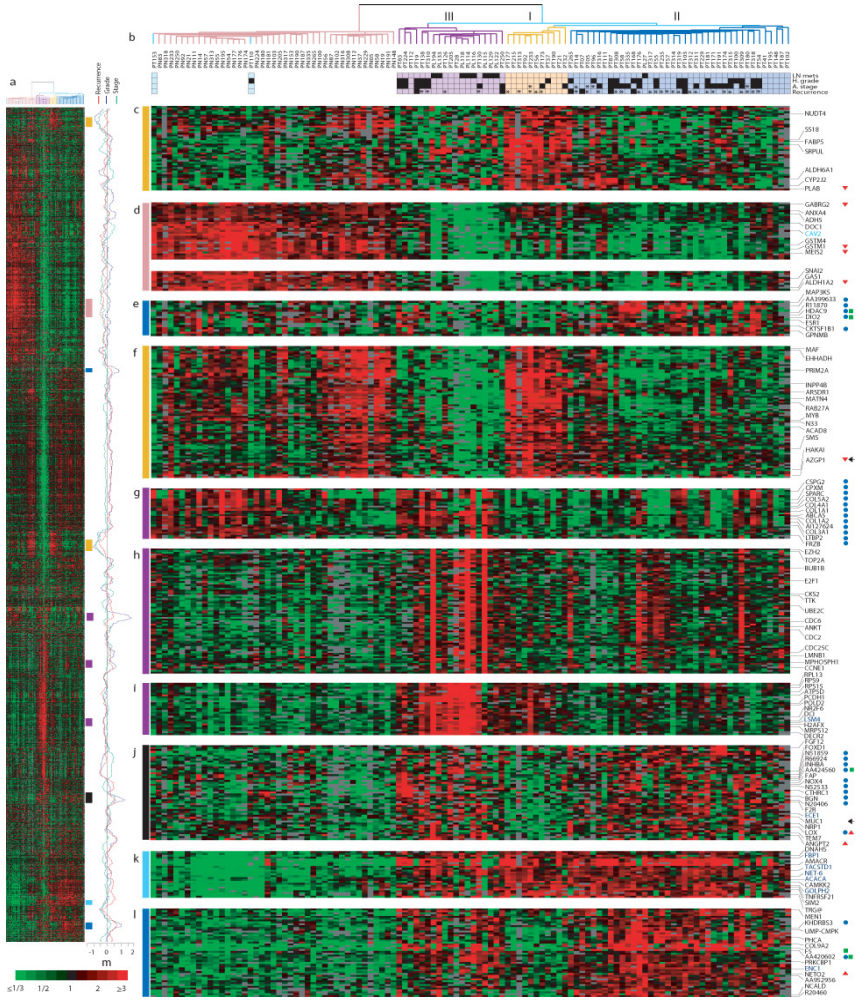
Biological networks have high clustering coefficients

	Gene co-expression network	Protein-protein interaction network
Source	**GNF**	**HPRD**
Nodes	**Genes**	**Proteins**
Edges	**Co-expression**	**Physical interaction**
Number of nodes	**6,342**	**5,881**
Number of edges	**74,830**	**23,333**
Clustering coefficient (actual)	**0.16**	**0.18**
Clustering coefficient (random)	**0.0020**	**0.0009**

Modularity of biological networks

- High clustering coefficient is the signature of a network's modularity.
- **Modularity** refers to a group of physically or functionally linked molecules (nodes) that work together to achieve a relatively distinct function.
- Examples
  - Transcriptional module: a set of co-regulated genes sharing a common function
  - Protein complex: assembly of proteins that build up some cellular machinery, commonly spans a dense sub-network of proteins in a protein interaction network
  - Signaling pathway: a chain of interacting proteins propagating a signal in the cell

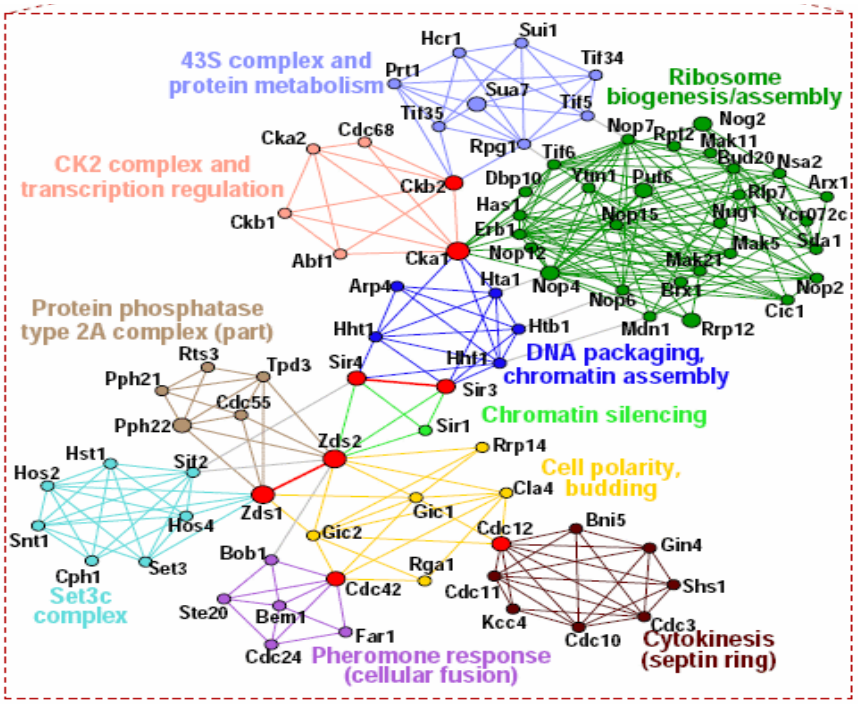
Transcriptional module identification: Hierarchical clustering



- Gene by condition matrix with expression signals
- Gene by gene matrix with pair-wise similarities
- Hierarchical clustering based on the similarities
- Biological interpretation
  - *cis*-regulatory element analysis
  - Functional enrichment

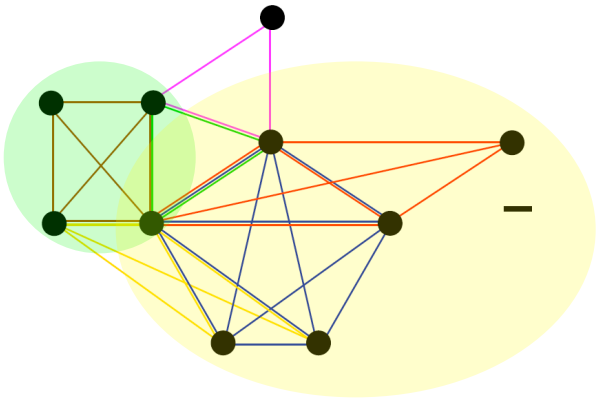
Lapointe et al, PNAS, 101:811, 2004

Protein complex identification: Clique-based approach

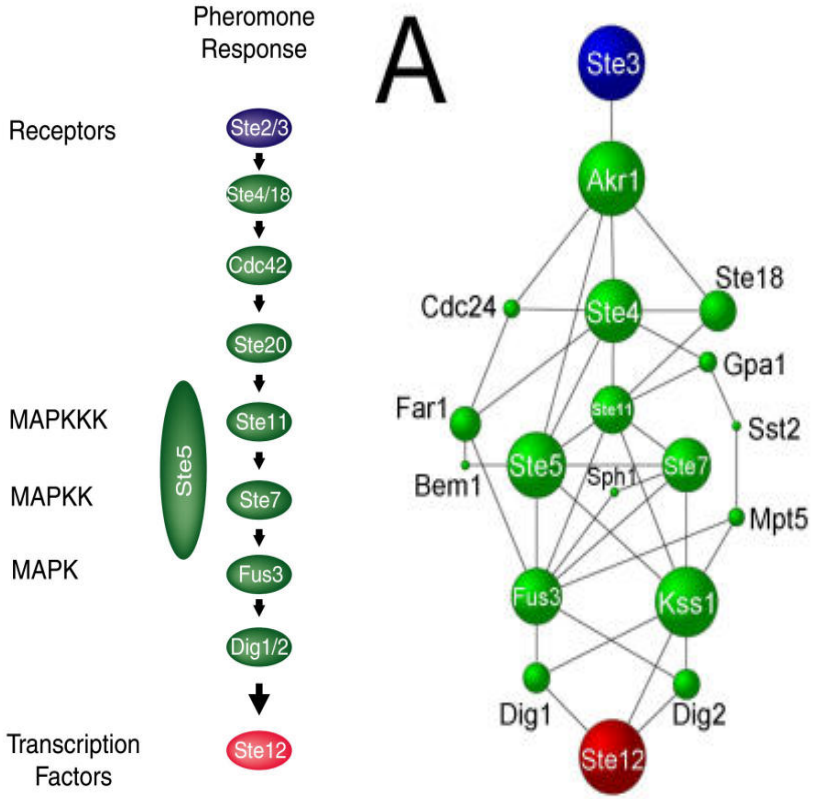


Palla et al, Nature, 435:841, 2005

- **Enumerate:** enumerate all cliques of size *k* (*k*-cliques)
- **Merge:** generate communities of all *k*-cliques that can be reached from each other through a series of adjacent *k*-cliques (where adjacency means sharing *k*-1 nodes)
- **Biological interpretation:**
  - Functional enrichment



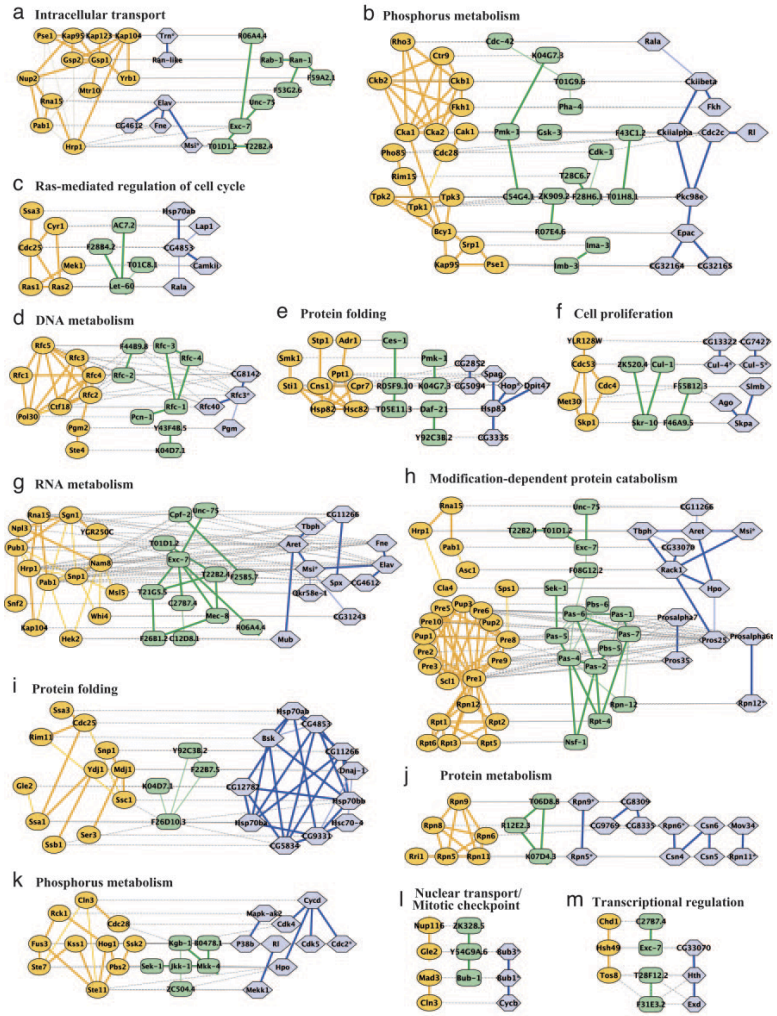
Signaling pathway identification: NetSearch

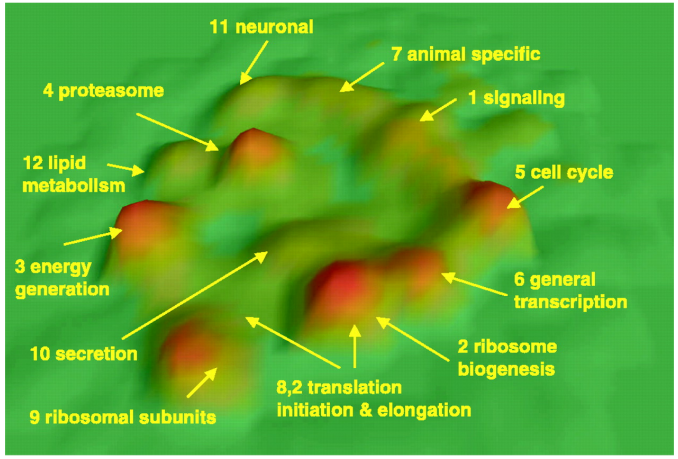


Steffen et al., BMC Bioinformatics, 3:34, 2002

- Enumerate all possible linear paths of a specified length through the interaction map starting at any membrane protein and ending on any DNA-binding protein
- Use microarray gene expression data to rank all paths according to the degree of similarity in the expression profiles of pathway members
- Highest ranking linear pathways that have common starting points and endpoints are combined into the final model of the branched networks
- Biological interpretation
  - Functional enrichment

Module conservation across species





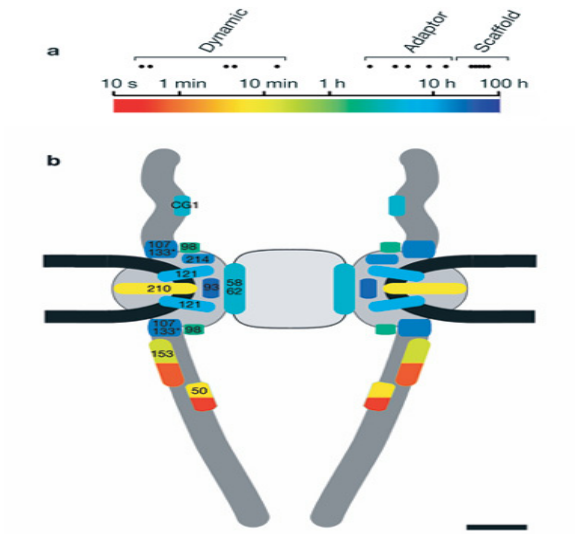
Stuart et al, Science, 302:249, 2003

Transcriptional modules: yeast, worm, fly, human

Sharan et al, PNAS, 102:1975, 2005

Protein complexes: yeast, worm, fly

Dynamic organization of the modules



Rabut et al, Nat Cell Biol, 6:1114, 2004

- Nuclear pore complex (NPC): large protein complex mediate the traffic between cytoplasm and nucleoplasm
- NPC components exhibited a wide range of residence times covering five orders of magnitude from seconds to days.
  - Scaffold
  - Adaptor
  - Dynamic

Summary

- Network motif
  - Definition: a pattern that occur in the real network significantly more often than in randomized networks
  - Identification: compare to random networks
  - Biological interpretation
    - Building blocks of biological networks
    - FFL-AND: protects against brief input fluctuations; FFL-OR: allow additional time for biological process
    - SIM: generates a temporal expression program
- Network module
  - Definition: a group of physically or functionally linked molecules that work together to achieve a relatively distinct function
  - Identification
    - Transcriptional modules: hierarchical clustering
    - Protein complexes: clique-based approaches
    - Signaling pathways: netsearch
  - Biological interpretation
    - Functional enrichment
    - Evolutionary conservation

Key references

- *Alon.* Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8:450, 2007