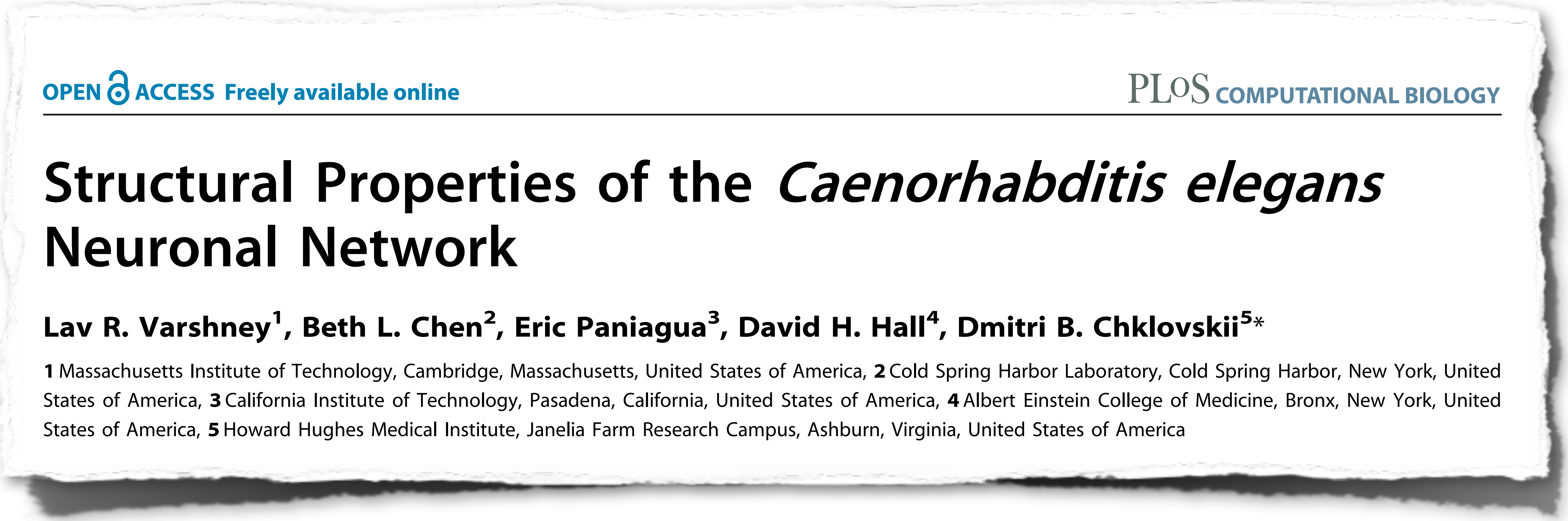
Bioinformatics 3

V8 – Gene Regulation

Tue, Nov 15, 2011

Recently in PLoS Comp. Biol.



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Structural Properties of the *Caenorhabditis elegans* Neuronal Network

Lav R. Varshney[1], Beth L. Chen[2], Eric Paniagua[3], David H. Hall[4], Dmitri B. Chklovskii[5]*

1 Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **2** Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **3** California Institute of Technology, Pasadena, California, United States of America, **4** Albert Einstein College of Medicine, Bronx, New York, United States of America, **5** Howard Hughes Medical Institute, Janelia Farm Research Campus, Ashburn, Virginia, United States of America

Reconstruction and classification of the worm's neuronal network

PLoS Comput. Biol. **7** (2011) e1001066

"Network" => What can we apply???

Excursion: *C. elegans*

Small worm: L = 1 mm, Ø ≈ 65 µm
lives in the soil, eats bacteria

Consists of 959 cells, 302 nerve cells,
all worms are "identical"

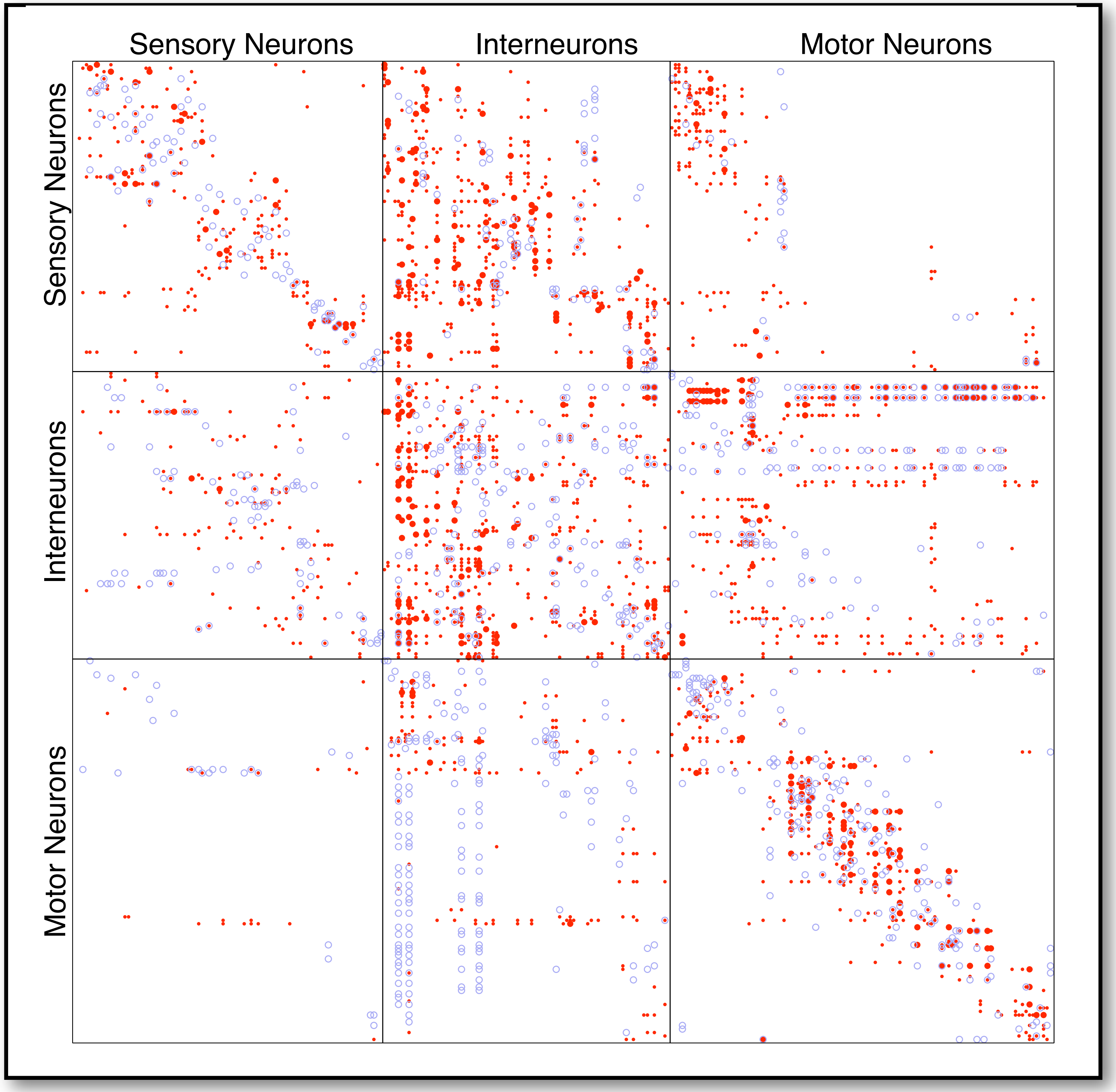
Completely sequenced in 1998 (first multicellular organism)

Very simple handling, transparent

=> One of the **prototype organisms**

Database "everything" about the worm:
www.wormbase.org

Adjacency Matrix



Two types of connections between neurons:

- **gap junctions**
  => electric contacts
  => undirected
- **chemical synapses**
  => neurotransmitters
  => directed

Observations:

- three groups of neurons (clustering)
- gap junction entries are symmetric, chemical synapses not (directionality)

PLoS Comput. Biol. **7** (2011) e1001066

Some Statistics

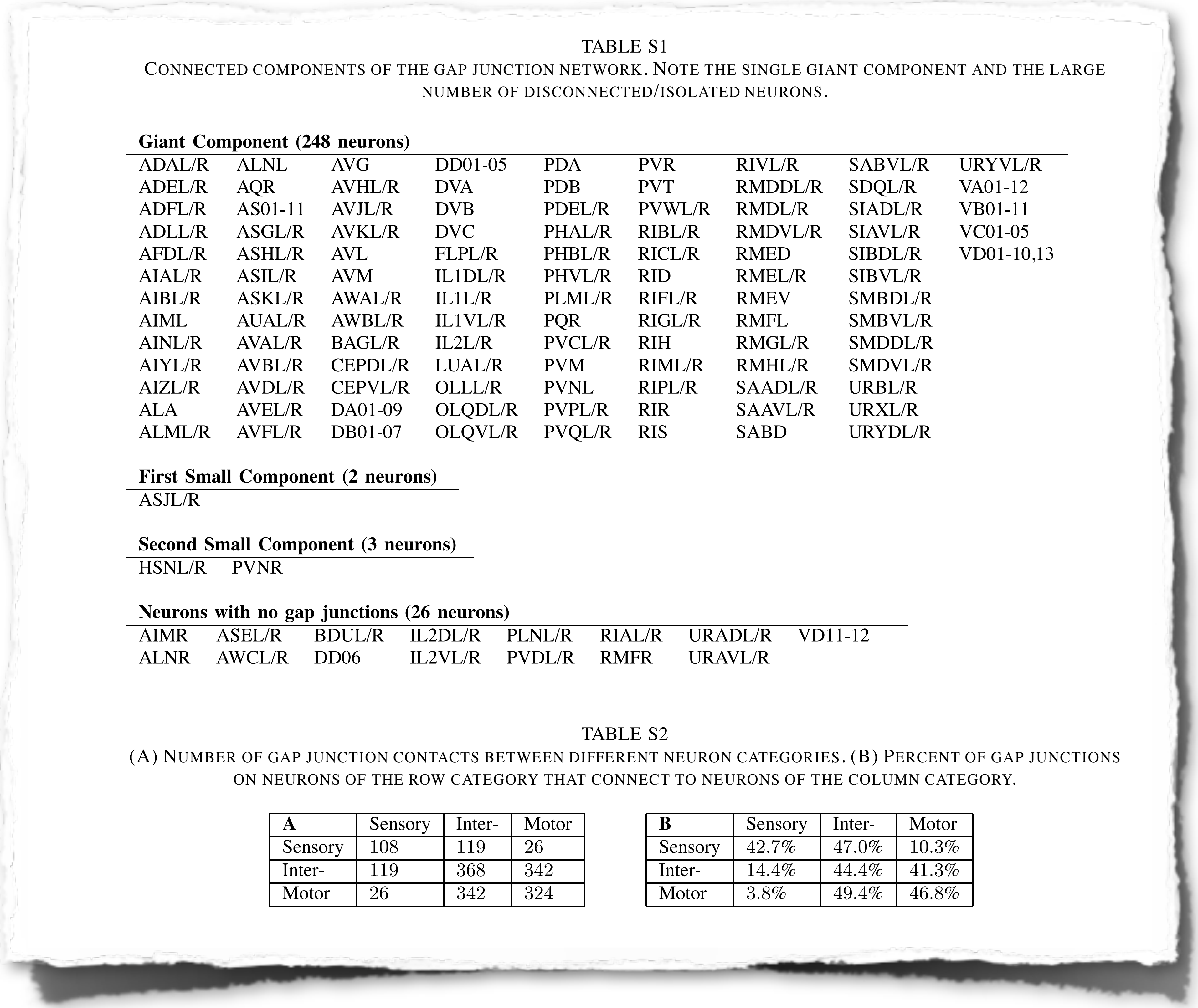


TABLE S1

CONNECTED COMPONENTS OF THE GAP JUNCTION NETWORK. NOTE THE SINGLE GIANT COMPONENT AND THE LARGE NUMBER OF DISCONNECTED/ISOLATED NEURONS.

Giant Component (248 neurons)

ADAL/R
ADEL/R
ADFL/R
ADLL/R
AFDL/R
AIAL/R
AIBL/R
AIML
AINL/R
AIYL/R
AIZL/R
ALA
ALML/R
ALNL
AQR
AS01-11
ASGL/R
ASHL/R
ASIL/R
ASKL/R
AUAL/R
AVAL/R
AVBL/R
AVDL/R
AVEL/R
AVFL/R
AVG
AVHL/R
AVJL/R
AVKL/R
AVL
AVM
AWAL/R
AWBL/R
BAGL/R
CEPDL/R
CEPVL/R
DA01-09
DB01-07
DD01-05
DVA
DVB
DVC
FLPL/R
IL1DL/R
IL1L/R
IL1VL/R
IL2L/R
LUAL/R
OLLL/R
OLQDL/R
OLQVL/R
PDA
PDB
PDEL/R
PHAL/R
PHBL/R
PHVL/R
PLML/R
PQR
PVCL/R
PVM
PVNL
PVPL/R
PVQL/R
PVR
PVT
PVWL/R
RIBL/R
RICL/R
RID
RIFL/R
RIGL/R
RIH
RIML/R
RIPL/R
RIR
RIS
RIVL/R
RMDDL/R
RMDL/R
RMDVL/R
RMED
RMEL/R
RMEV
RMFL
RMGL/R
RMHL/R
SAADL/R
SAAVL/R
SABD
SABVL/R
SDQL/R
SIADL/R
SIAVL/R
SIBDL/R
SIBVL/R
SMBDL/R
SMBVL/R
SMDDL/R
SMDVL/R
URBL/R
URXL/R
URYDL/R
URYVL/R
VA01-12
VB01-11
VC01-05
VD01-10,13

First Small Component (2 neurons)

ASJL/R

Second Small Component (3 neurons)

HSNL/R
PVNR

Neurons with no gap junctions (26 neurons)

AIMR
ALNR
ASEL/R
AWCL/R
BDUL/R
DD06
IL2DL/R
IL2VL/R
PLNL/R
PVDL/R
RIAL/R
RMFR
URADL/R
URAVL/R
VD11-12

TABLE S2

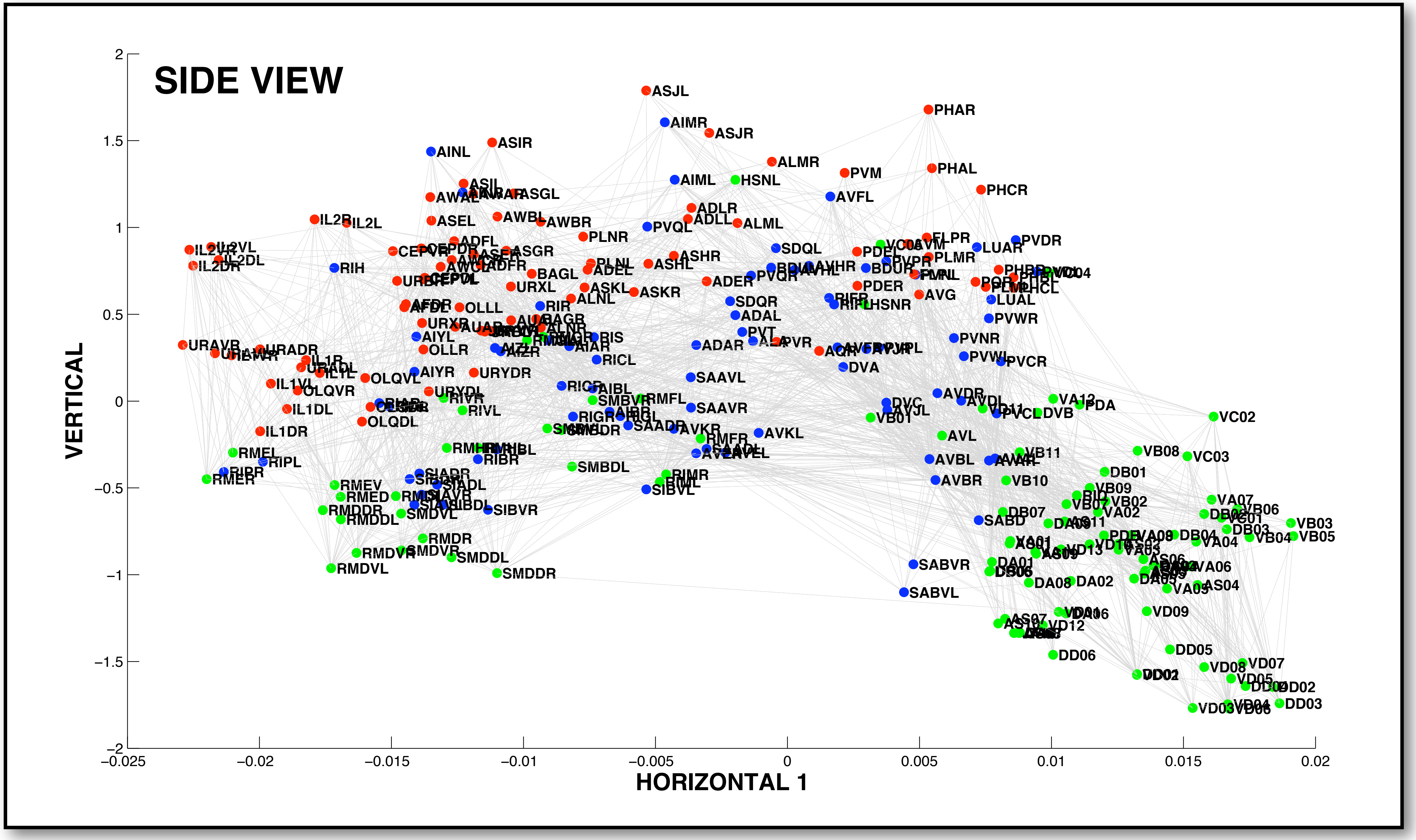
(A) NUMBER OF GAP JUNCTION CONTACTS BETWEEN DIFFERENT NEURON CATEGORIES. (B) PERCENT OF GAP JUNCTIONS ON NEURONS OF THE ROW CATEGORY THAT CONNECT TO NEURONS OF THE COLUMN CATEGORY.

A	Sensory	Inter-	Motor
Sensory	108	119	26
Inter-	119	368	342
Motor	26	342	324

B	Sensory	Inter-	Motor
Sensory	42.7%	47.0%	10.3%
Inter-	14.4%	44.4%	41.3%
Motor	3.8%	49.4%	46.8%

PLoS Comput. Biol. **7** (2011) e1001066

Information Flow



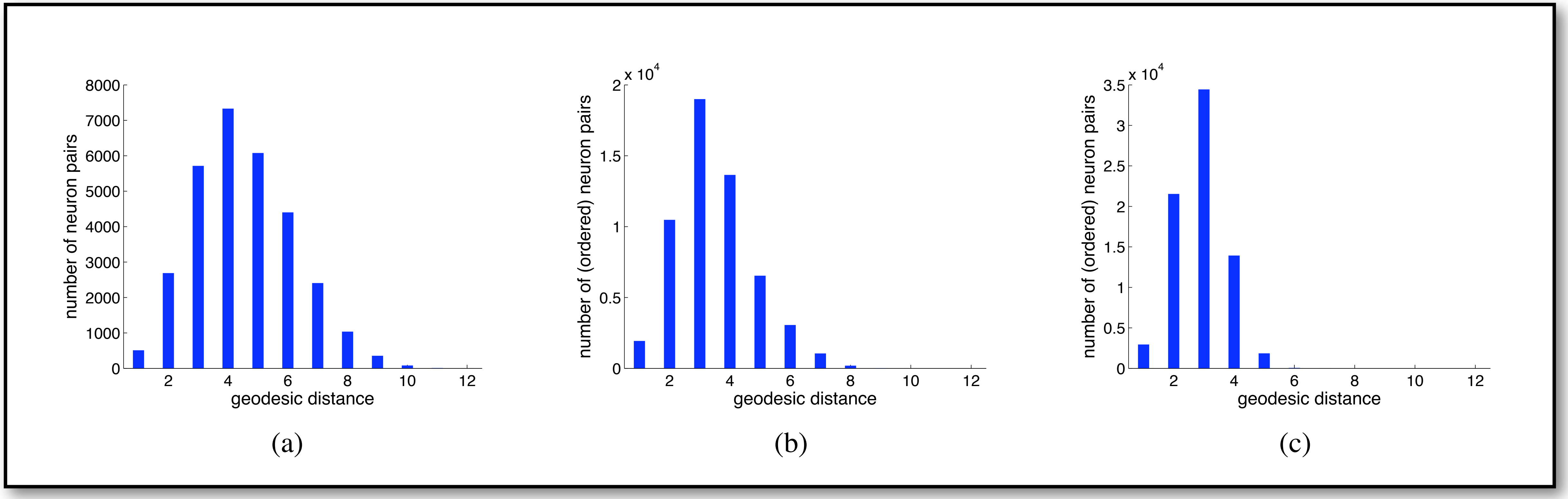
Network arranged so that information flow is (mostly) top => bottom

sensory neurons interneurons motorneurons

PLoS Comput. Biol. **7** (2011) e1001066

Network Size

Geodesic distance (shortest path) distributions of giant component of…



(a) (b) (c)

(electric)
gap junctions

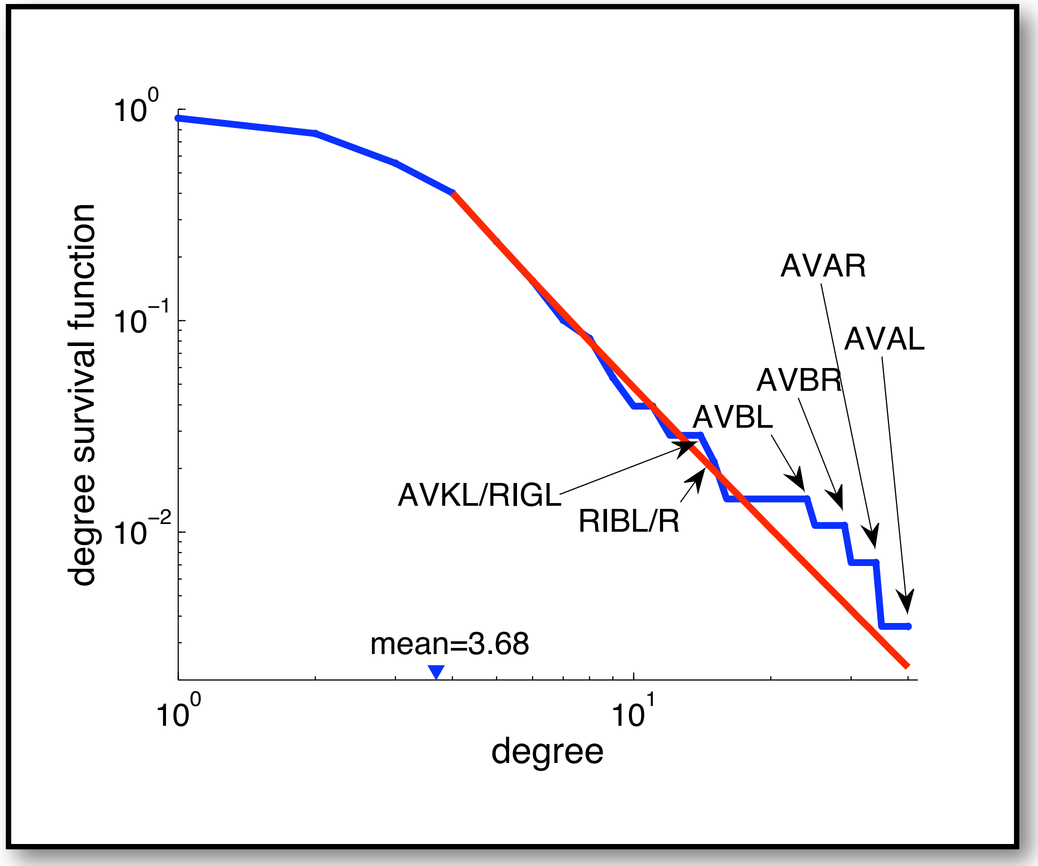
(chemical)
synapses

combined
network

=> a worm is a small animal :-)

PLoS Comput. Biol. **7** (2011) e1001066

Degree Distribution

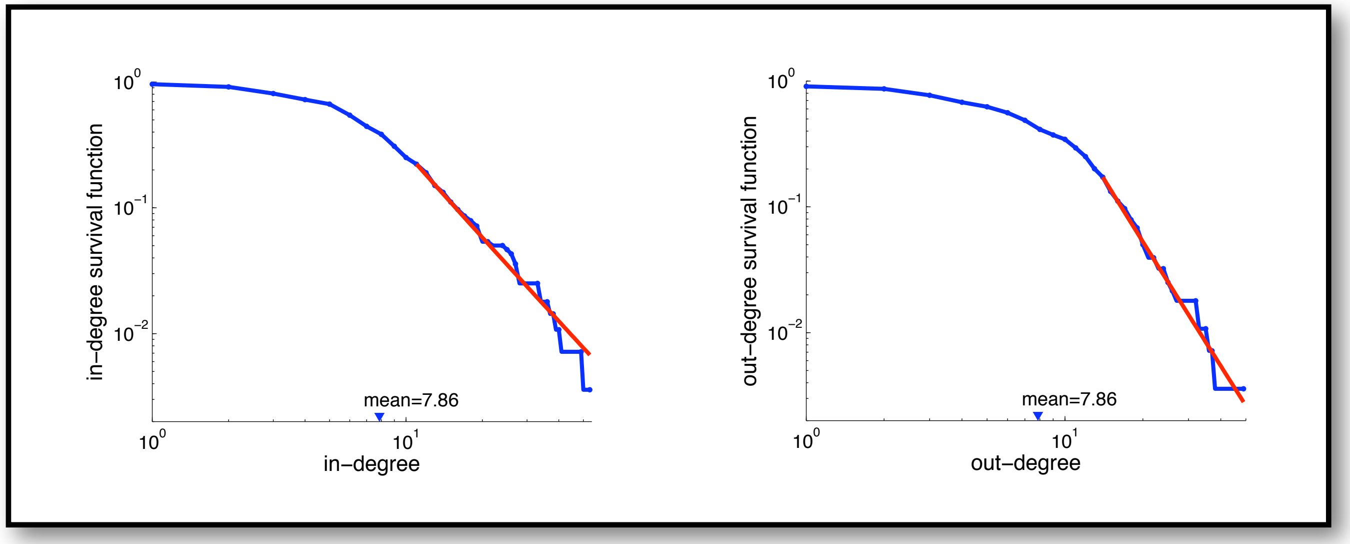


Plot of the "survival function" of P(k)
(1 – cumulative P(k))
for the (electric) gap junctions

$$P(d) = \sum_{k=d}^{\infty} p(k)$$

Power law for P(k) with γ = 3.14 (≈ π?)

In/out degrees of the
chemical synapses
=> fit with γ = 3.17 / 4.22
(but clearly not SF!)



PLoS Comput. Biol. **7** (2011) e1001066

Some More Statistics

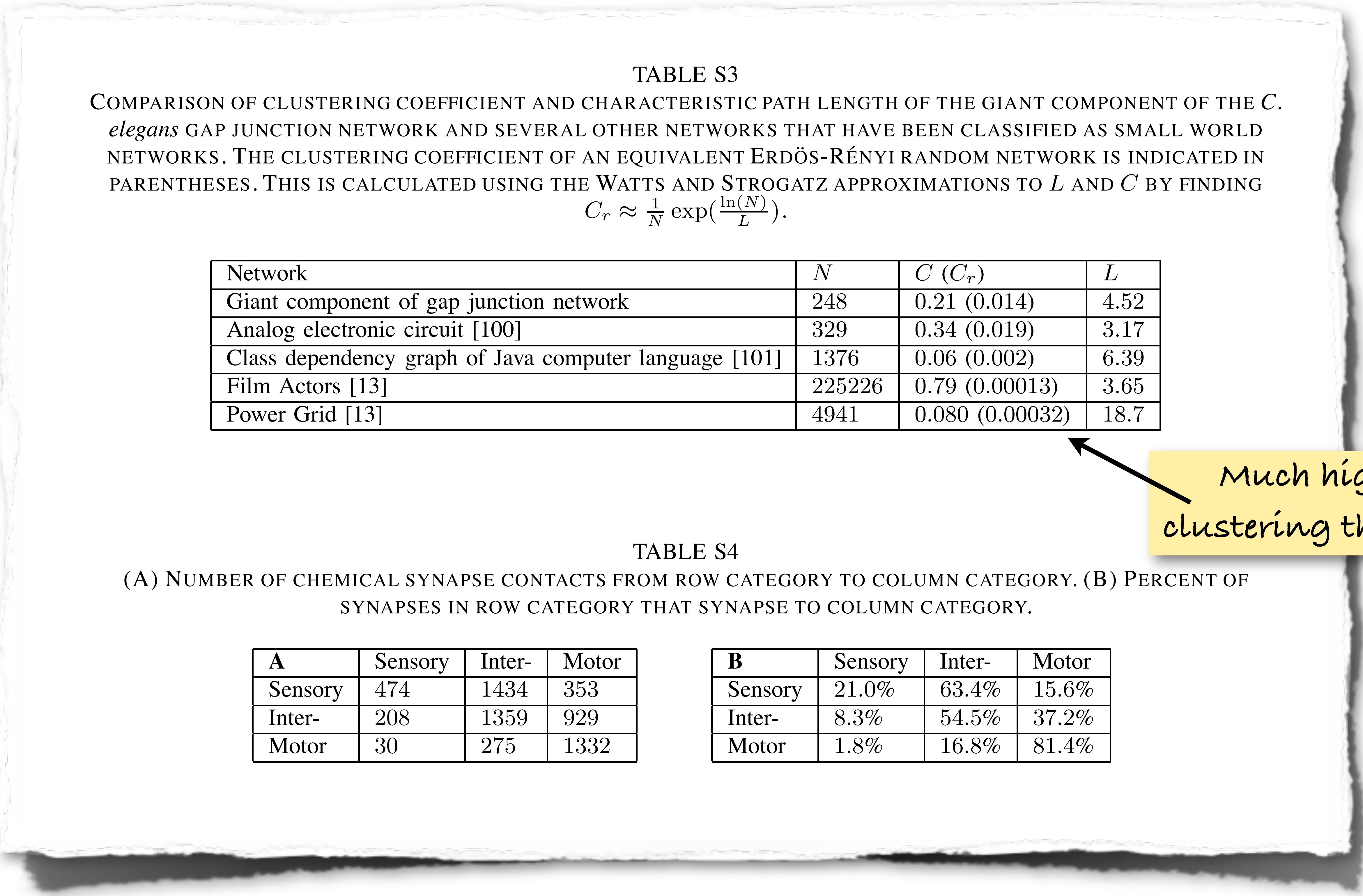


TABLE S3

Comparison of clustering coefficient and characteristic path length of the giant component of the *C. elegans* gap junction network and several other networks that have been classified as small world networks. The clustering coefficient of an equivalent Erdös-Rényi random network is indicated in parentheses. This is calculated using the Watts and Strogatz approximations to L and C by finding $C_r \approx \frac{1}{N}\exp(\frac{\ln(N)}{L})$.

Network	N	C (C_r)	L
Giant component of gap junction network	248	0.21 (0.014)	4.52
Analog electronic circuit [100]	329	0.34 (0.019)	3.17
Class dependency graph of Java computer language [101]	1376	0.06 (0.002)	6.39
Film Actors [13]	225226	0.79 (0.00013)	3.65
Power Grid [13]	4941	0.080 (0.00032)	18.7

Much higher clustering than ER

TABLE S4

(A) Number of chemical synapse contacts from row category to column category. (B) Percent of synapses in row category that synapse to column category.

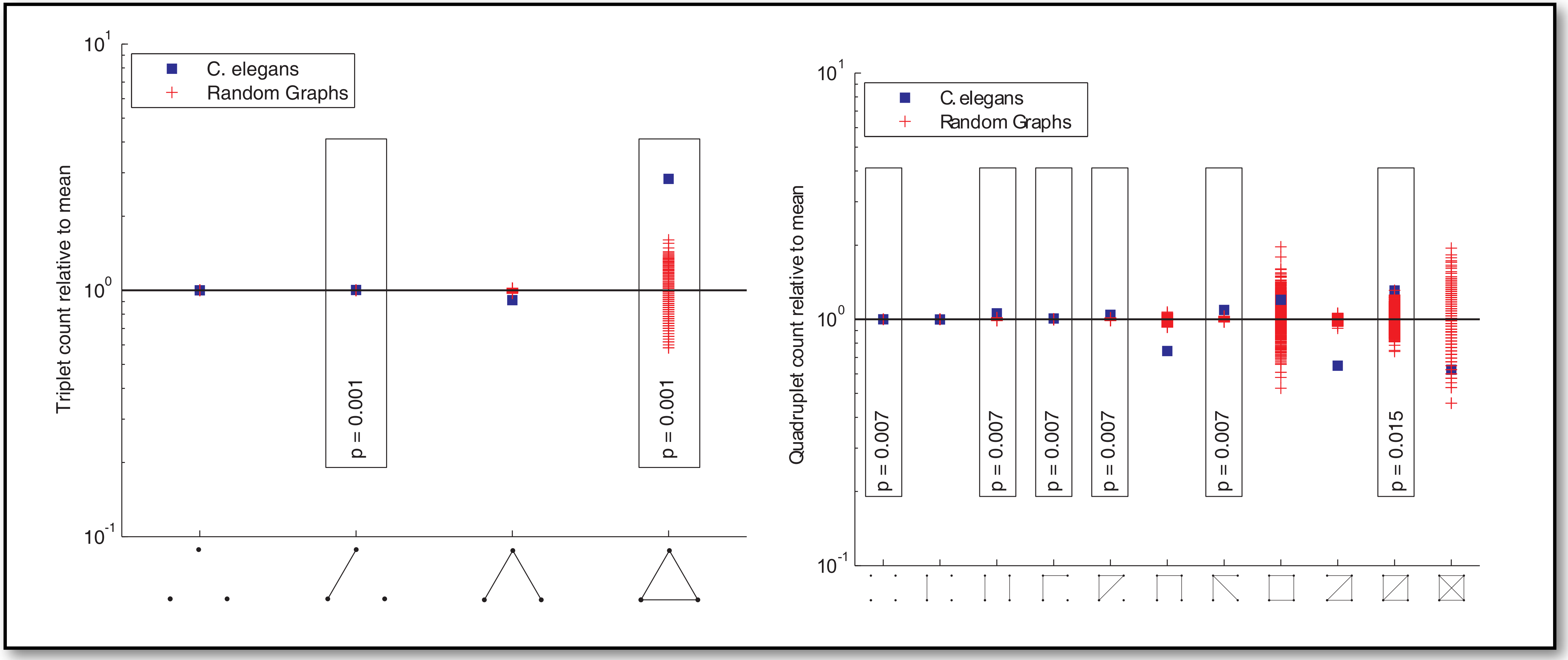
A	Sensory	Inter-	Motor
Sensory	474	1434	353
Inter-	208	1359	929
Motor	30	275	1332

B	Sensory	Inter-	Motor
Sensory	21.0%	63.4%	15.6%
Inter-	8.3%	54.5%	37.2%
Motor	1.8%	16.8%	81.4%

PLoS Comput. Biol. **7** (2011) e1001066

Network Motifs

Motif counts of the electric gap junction network relative to random network

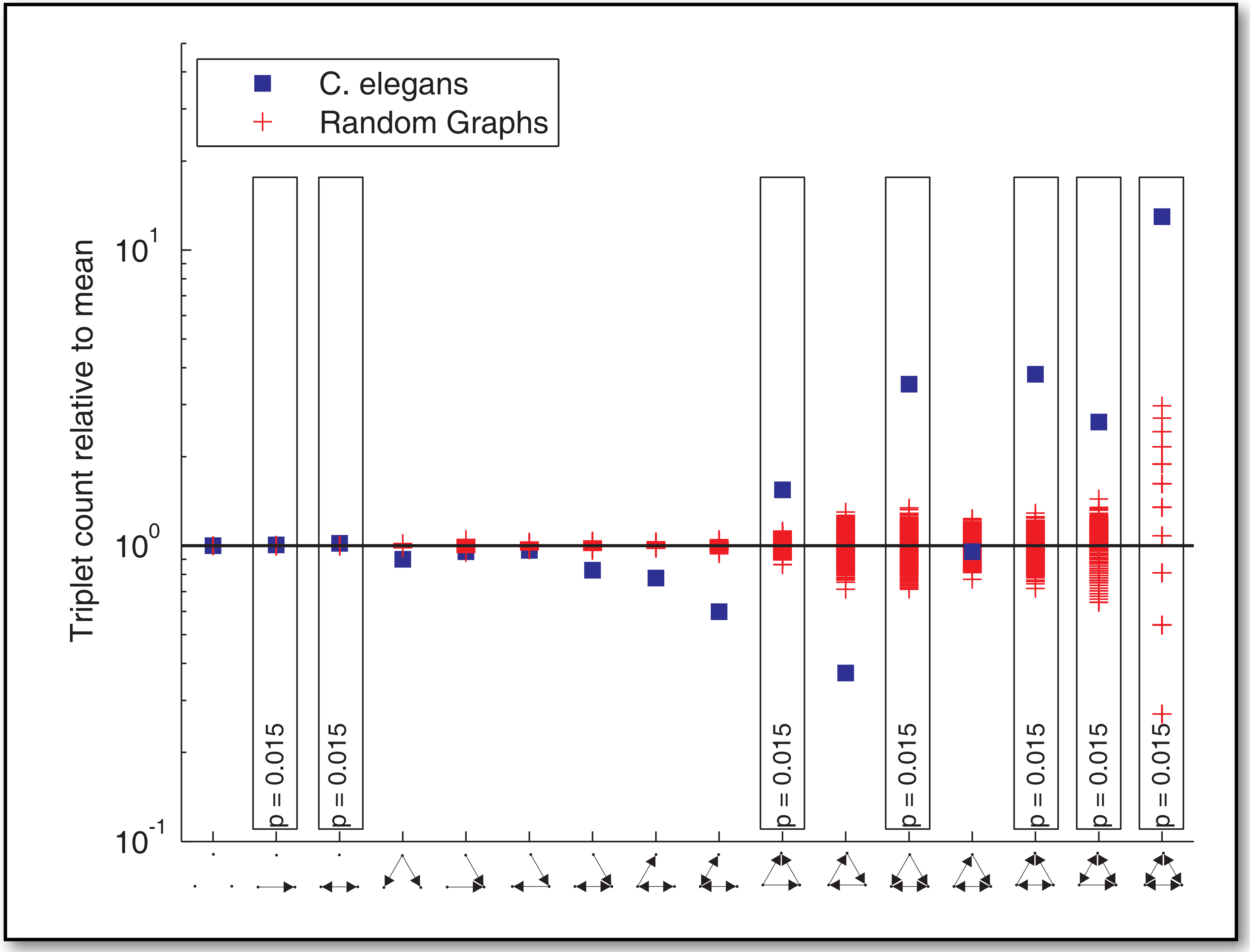


=> symmetric structures are overrepresented

=> clearly not a random network

PLoS Comput. Biol. **7** (2011) e1001066

Motifs II



Similar picture for the chemical synapses: not random

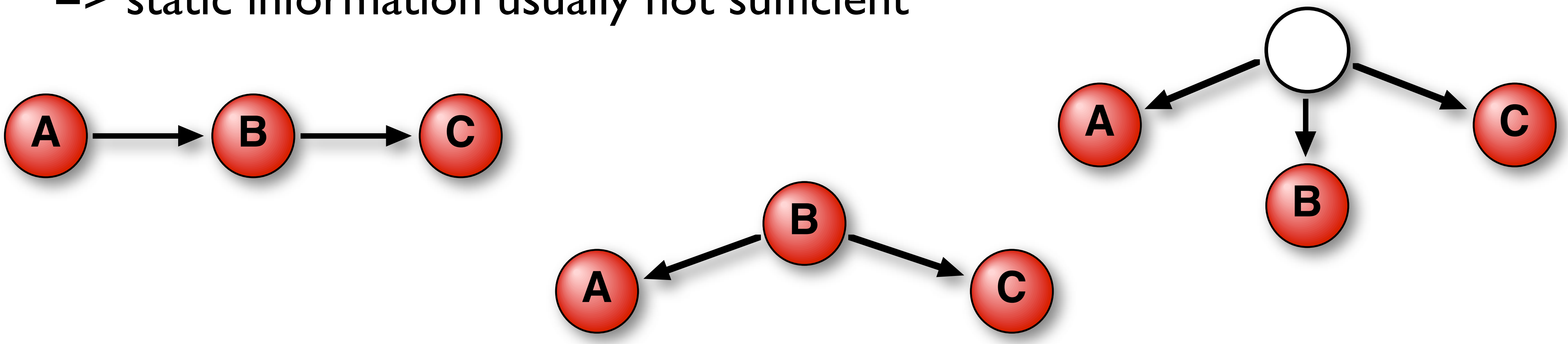
PLoS Comput. Biol. **7** (2011) e1001066

Network Reconstruction

Experimental data: DNA microarray => expression profiles

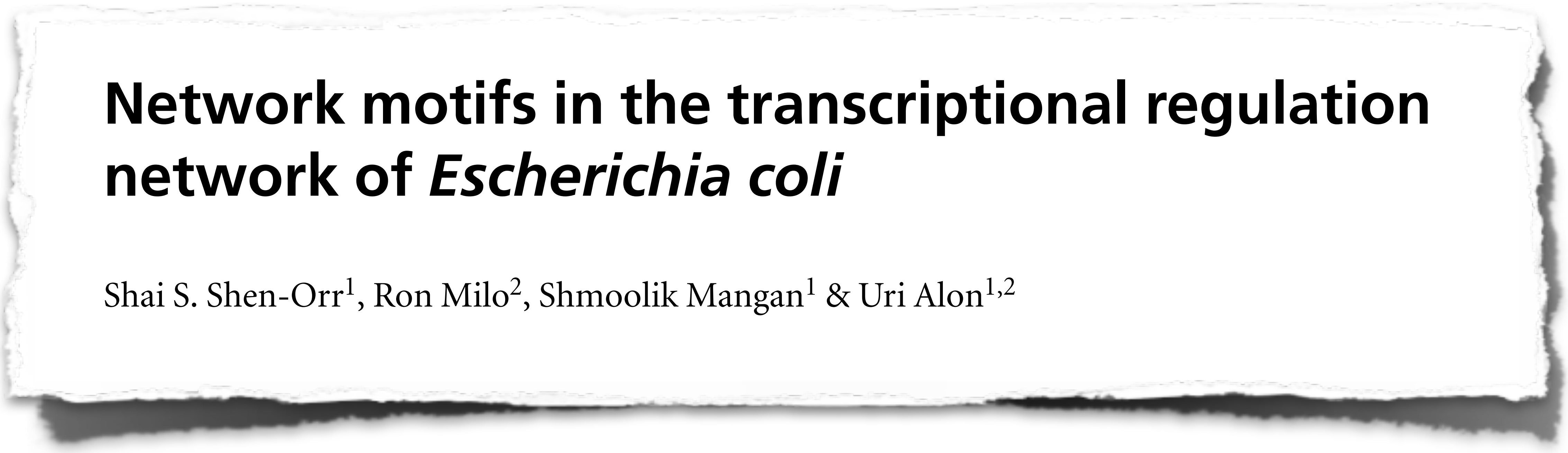
Clustering => genes that are **regulated simultaneously**
=> Cause and action??? Are all genes known???

Three different networks that lead to the same expression profiles
=> **combinatorial explosion** of number of compatible networks
=> static information usually not sufficient



Some formalism may help
=> **Bayesian networks** (formalized conditional probabilities)
but usually too many candidates…

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

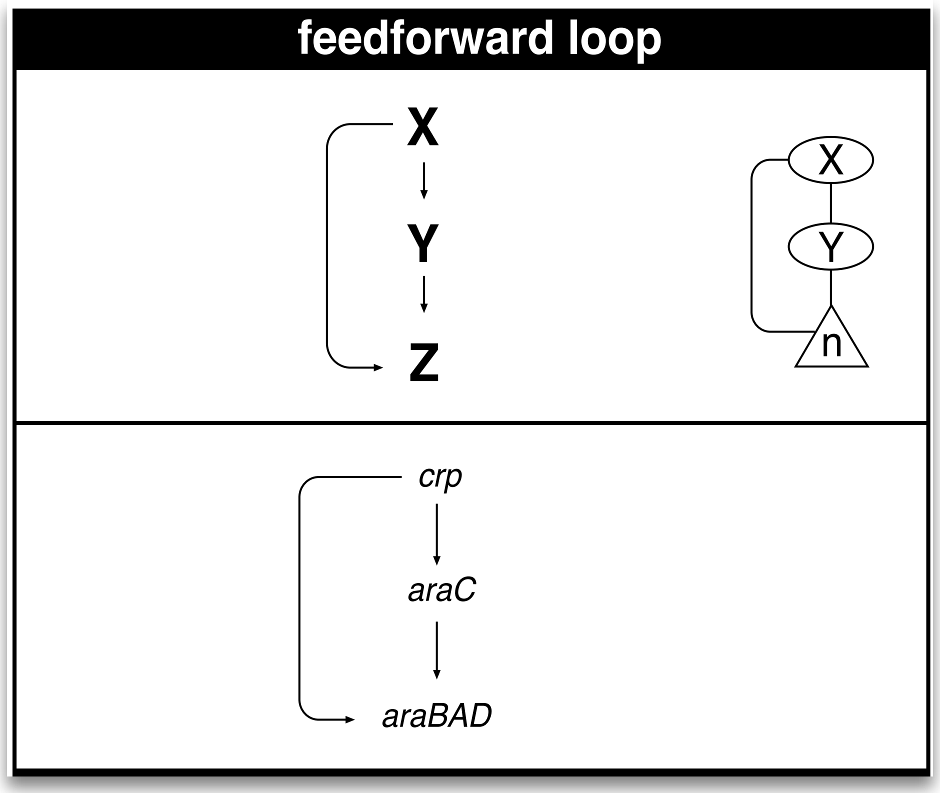
RegulonDB + their own hand-curated findings

=> break down network into motifs

  => statistical significance of the motifs?

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

Motif 1: Feed-Forward-Loop



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

Why not direct regulation without Y?

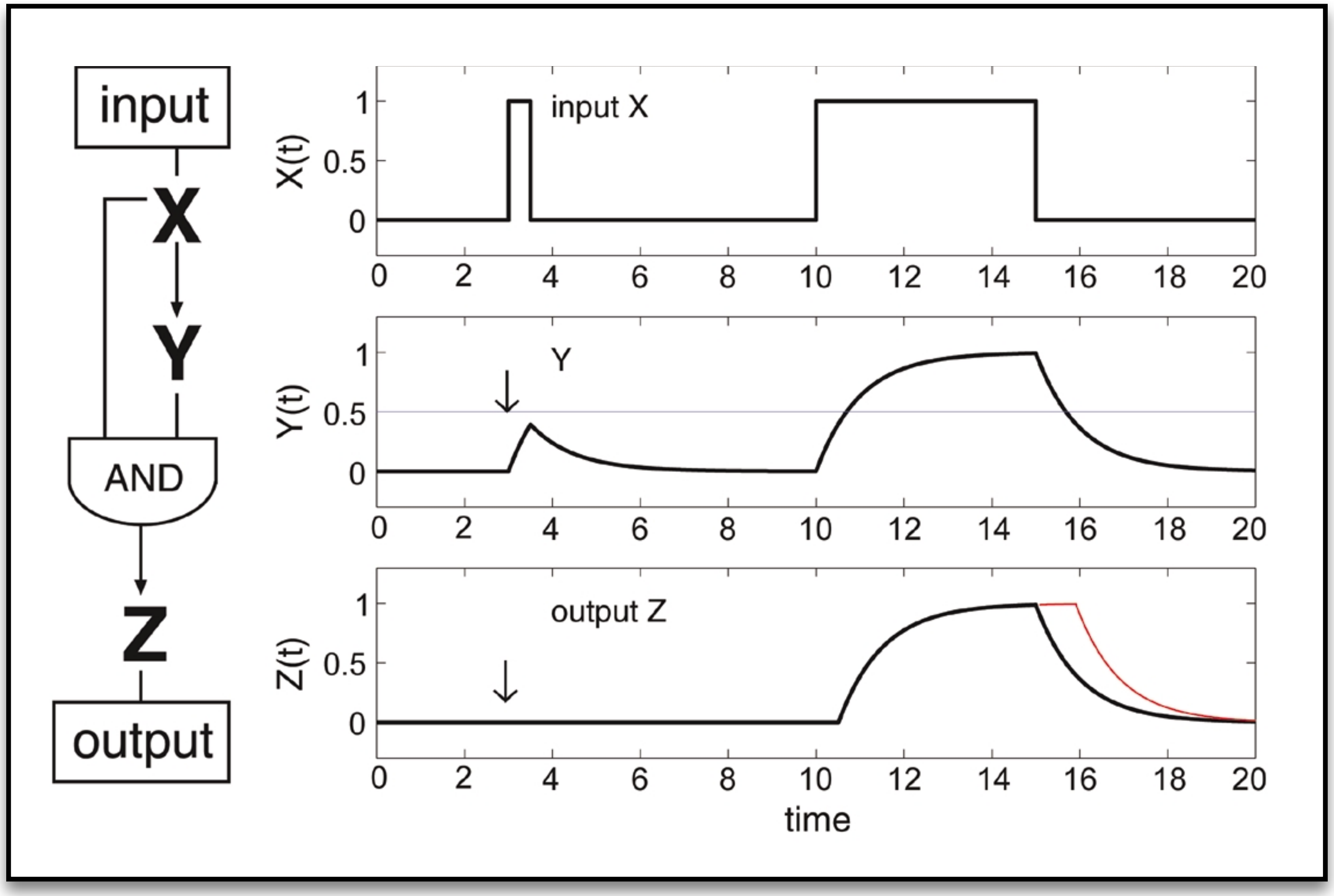
X and Y **together** regulate Z:

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

85% of the FFL 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

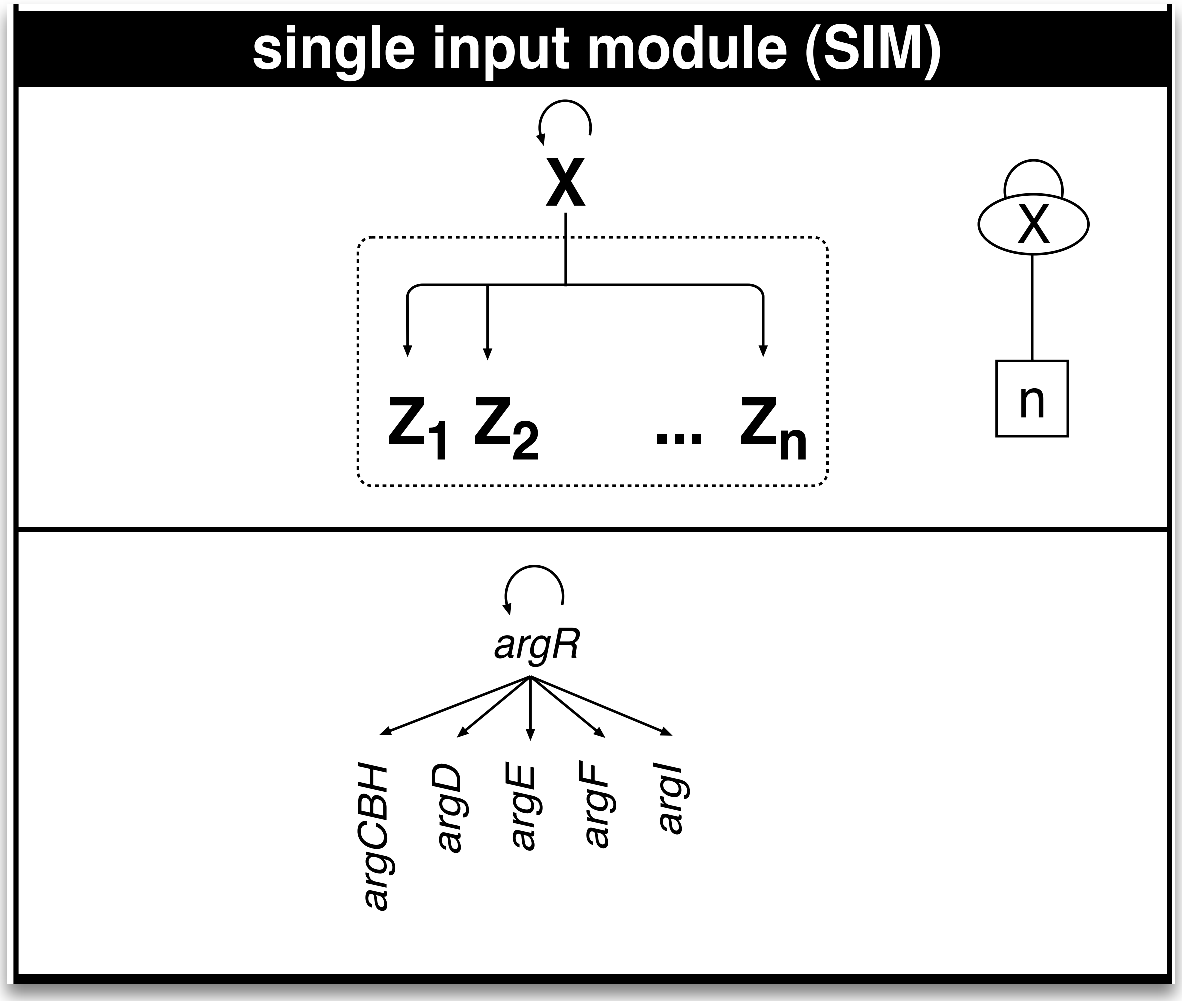
Dynamics:
- input activates X
- X activates Y (delay)
- (X && Y) activates Z

Delay between X and Y => signal must persist longer than delay
=> reject transient signal, react only to **persistent** signals
=> 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

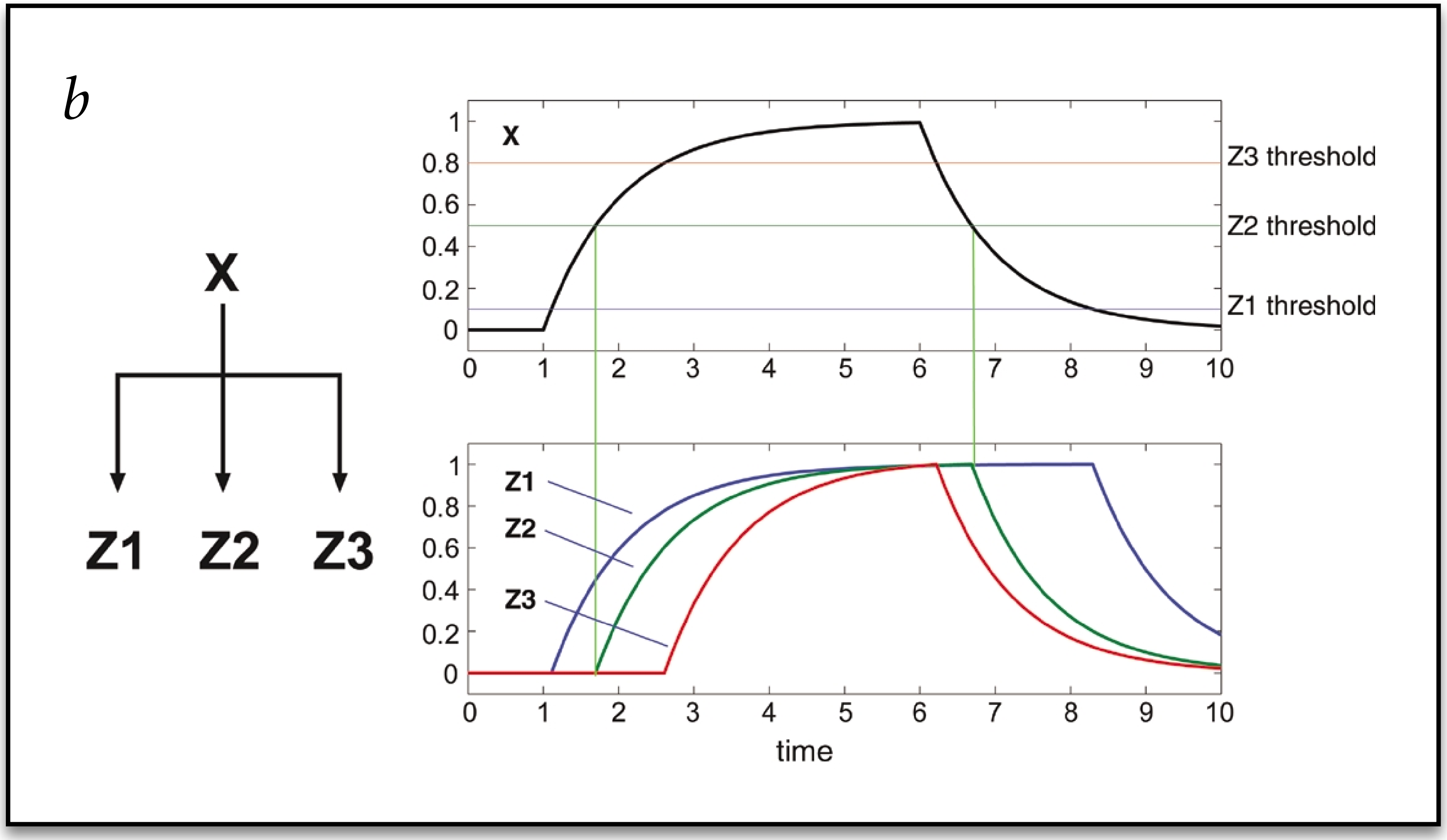
- same sign
- no additional regulation
- control usually autoregulatory (70% vs. 50% overall)

Mainly found in genes that code for **parts** of a protein **complex** or metabolic **pathway**

=> relative stoichiometries

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

SIM-Dynamics



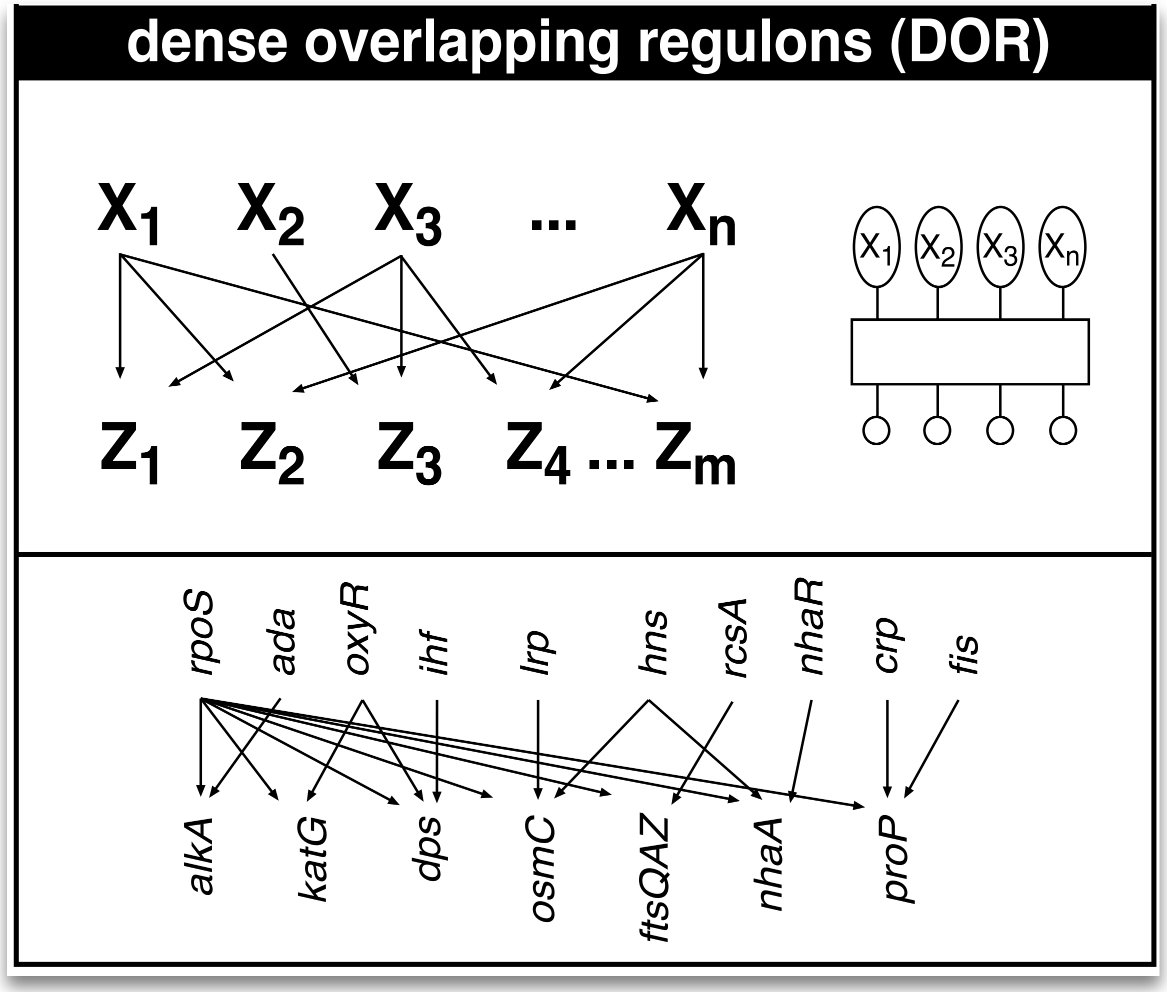
With different thresholds for each regulated operon:

=> 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: Dense 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

Motif Statistics

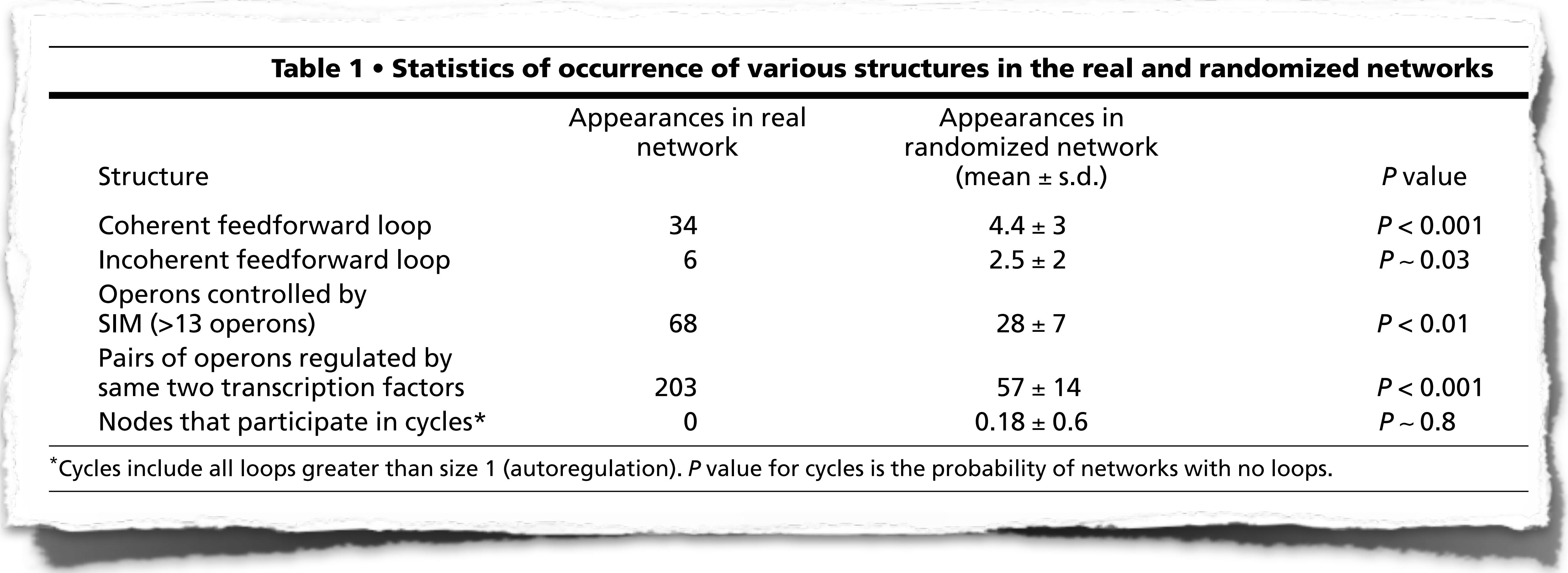


Table 1 • Statistics of occurrence of various structures in the real and randomized networks

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	$P \sim 0.8$

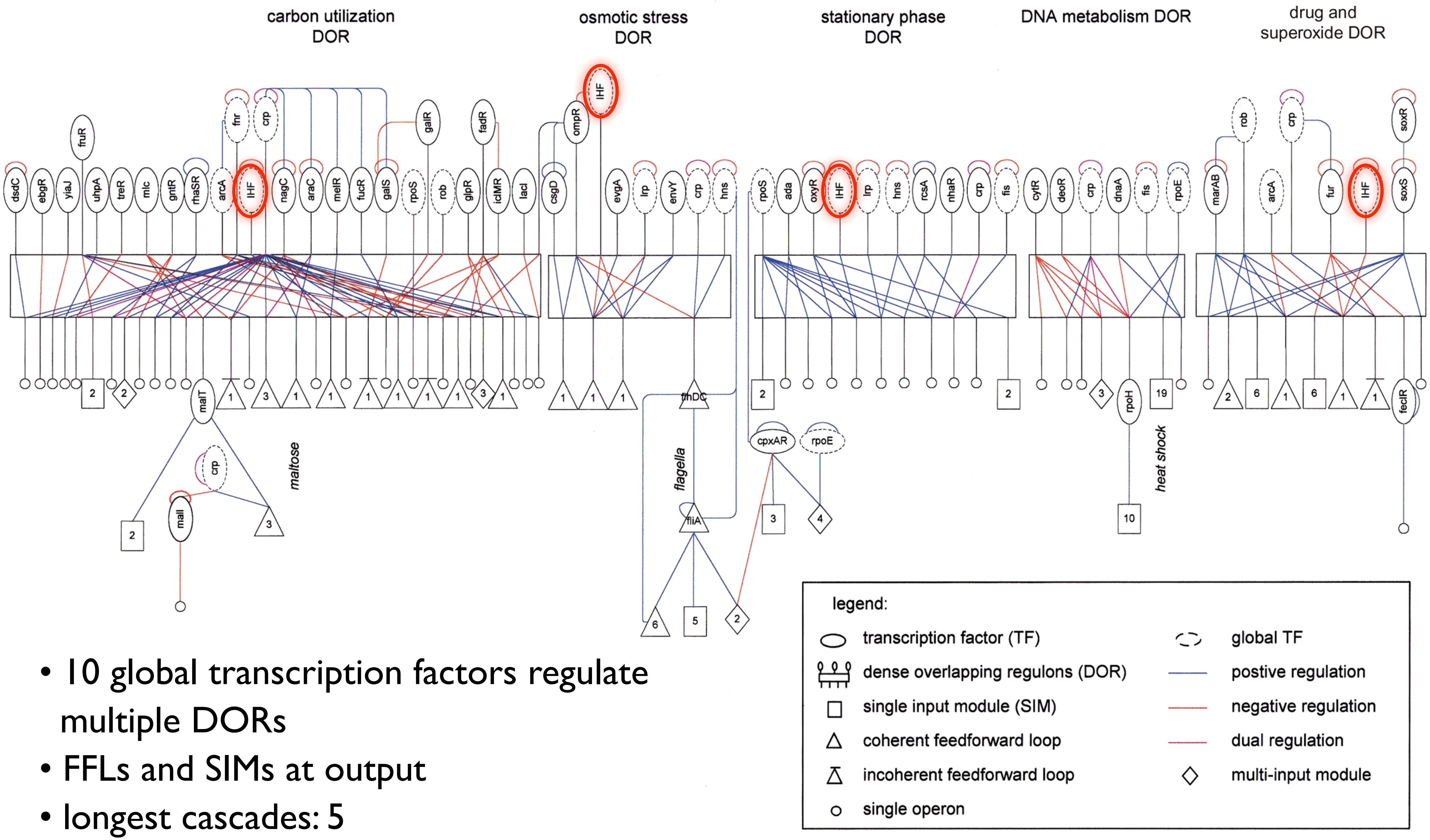
*Cycles include all loops greater than size 1 (autoregulation). *P* value for cycles is the probability of networks with no loops.

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

No cycles (X => Y => Z => X), but this is not statistically significant

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

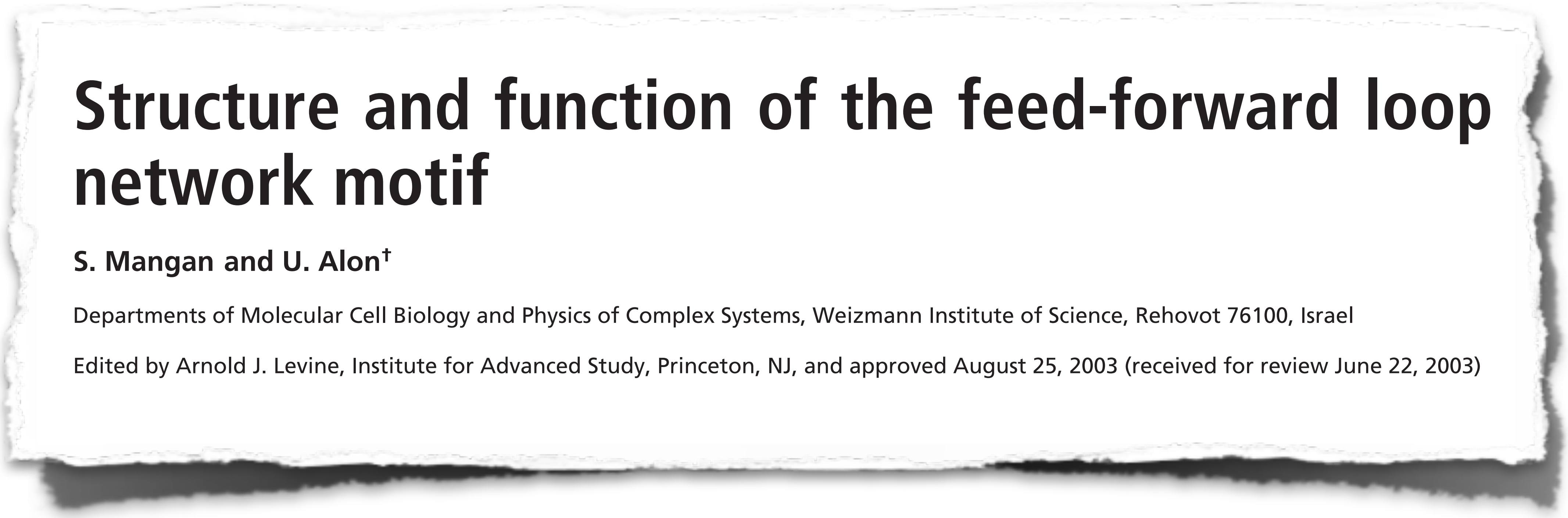
Network with Motifs



- 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

Motif-Dynamics



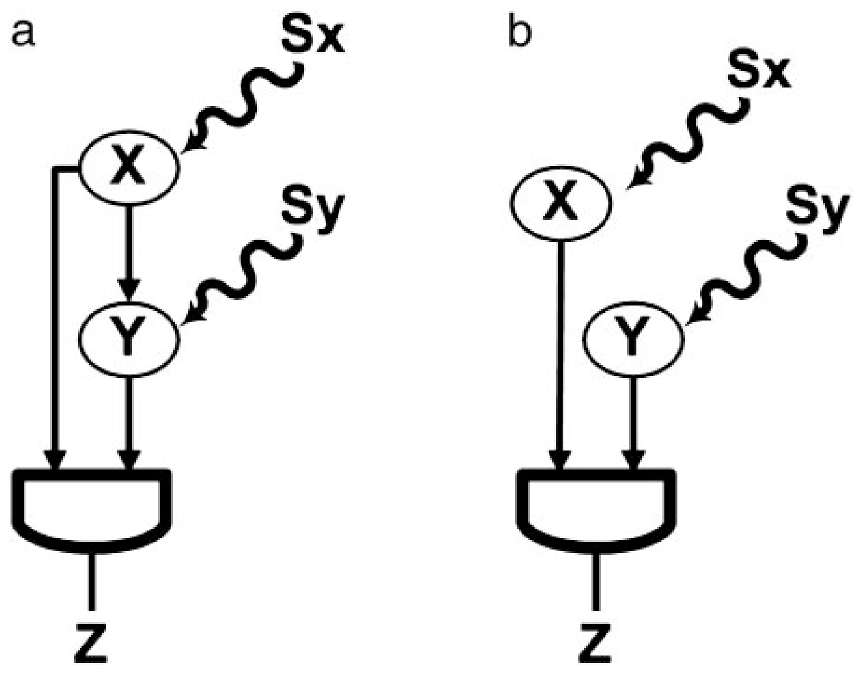
Structure and function of the feed-forward loop network motif

S. Mangan and U. Alon†

Departments of Molecular Cell Biology and Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel

Edited by Arnold J. Levine, Institute for Advanced Study, Princeton, NJ, and approved August 25, 2003 (received for review June 22, 2003)

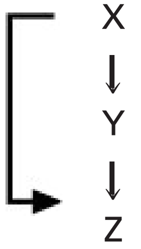
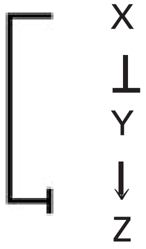
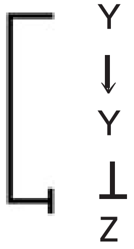
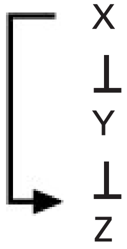
PNAS **100** (2003) 11980

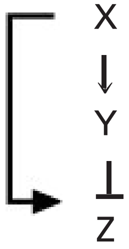
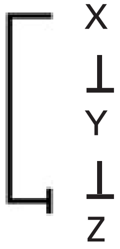
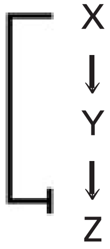
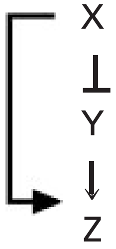


Compare dynamics of response Z to stimuli Sx and Sy for FFL (a) vs simple system (b).

Coherent and Incoherent FFLs

(in)coherent: X => Z has (opposite)same sign as X => Y => Z

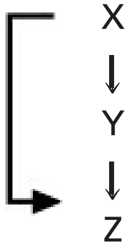
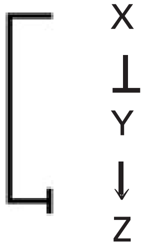
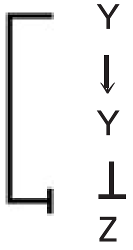
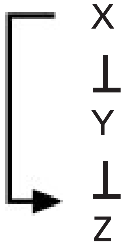
	Coherent type 1		Coherent type 2		Coherent type 3		Coherent type 4	
Species	Structure	Abundance	Structure	Abundance	Structure	Abundance	Structure	Abundance
E. coli		28		2		4		1
S. cerevisiae		26		5		0		0

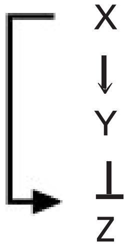
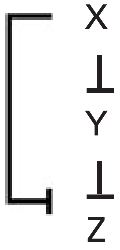
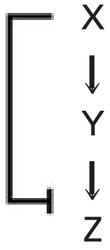
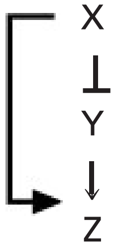
	Incoherent type 1		Incoherent type 2		Incoherent type 3		Incoherent type 4	
Species	Structure	Abundance	Structure	Abundance	Structure	Abundance	Structure	Abundance
E. coli		5		0		1		1
S. cerevisiae		21		3		1		0

Mangan, Alon, *PNAS* **100** (2003) 11980

Coherent and Incoherent FFLs

(in)coherent: X => Z has (opposite)same sign as X => Y => Z

	Coherent type 1		Coherent type 2		Coherent type 3		Coherent type 4	
Species	Structure	Abundance	Structure	Abundance	Structure	Abundance	Structure	Abundance
E. coli		28		2		4		1
S. cerevisiae		26		5		0		0

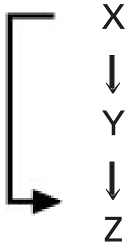
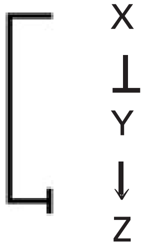
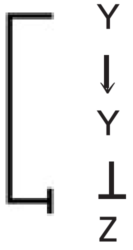
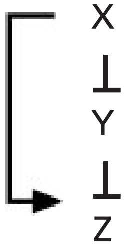
	Incoherent type 1		Incoherent type 2		Incoherent type 3		Incoherent type 4	
Species	Structure	Abundance	Structure	Abundance	Structure	Abundance	Structure	Abundance
E. coli		5		0		1		1
S. cerevisiae		21		3		1		0

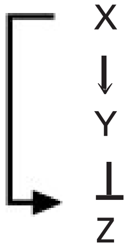
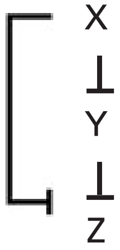
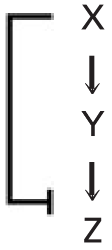
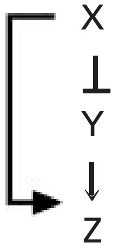
In *E. coli*: 2/3 are activator, 1/3 repressor interactions
=> relative abundances **not explained** by interaction occurences

Mangan, Alon, *PNAS* **100** (2003) 11980

Coherent and Incoherent FFLs

(in)coherent: X => Z has (opposite)same sign as X => Y => Z

Species	Coherent type 1 Structure	Coherent type 1 Abundance	Coherent type 2 Structure	Coherent type 2 Abundance	Coherent type 3 Structure	Coherent type 3 Abundance	Coherent type 4 Structure	Coherent type 4 Abundance
E. coli		28		2		4		1
S. cerevisiae		26		5		0		0
from interaction occurances:		8		2		2		2

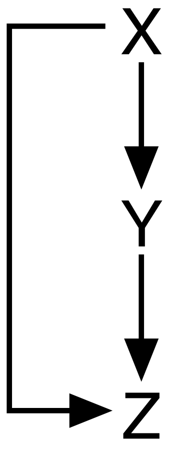
Species	Incoherent type 1 Structure	Incoherent type 1 Abundance	Incoherent type 2 Structure	Incoherent type 2 Abundance	Incoherent type 3 Structure	Incoherent type 3 Abundance	Incoherent type 4 Structure	Incoherent type 4 Abundance
E. coli		5		0		1		1
S. cerevisiae		21		3		1		0
from interaction occurances:		4		1		4		4

In *E. coli:* 2/3 are activator, 1/3 repressor interactions

=> relative abundances **not explained** by interaction occurences

Mangan, Alon, *PNAS* **100** (2003) 11980

Logic Response



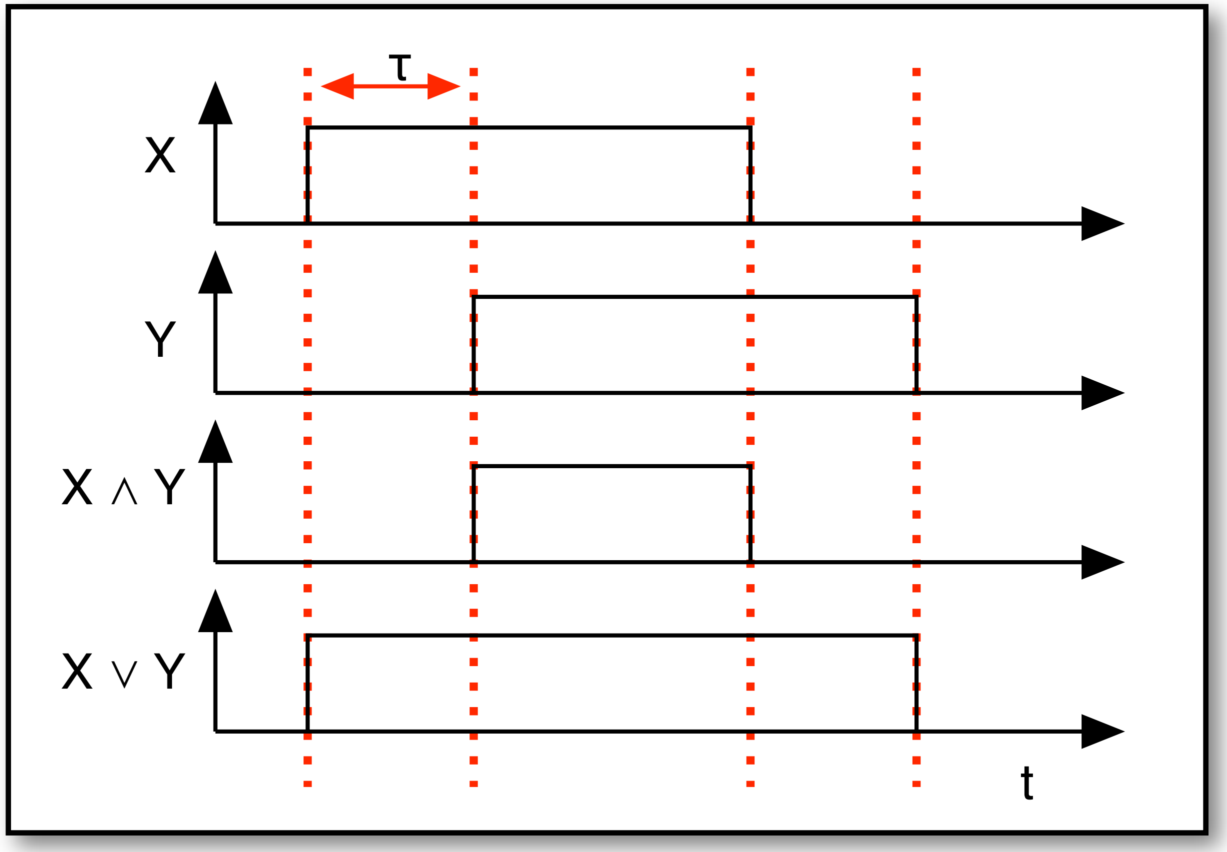
Z goes on when …

…X and Y are on
=> AND type

"Complex of TFx and TFy"

…X or Y is on
=> OR type

"TFx or TFy alone suffices"



=> same steady state response
Z is on when X is on

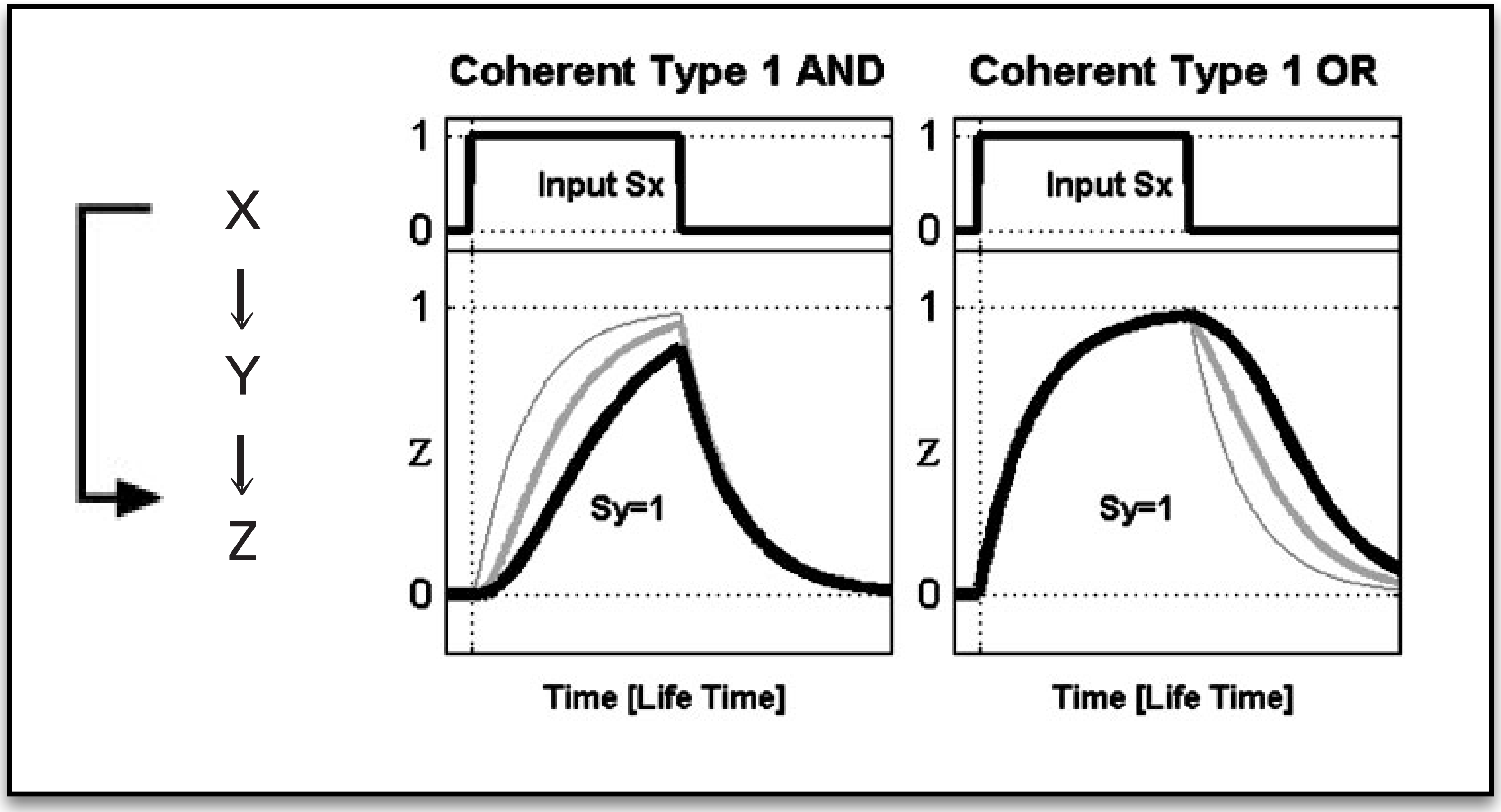
=> different dynamic responses
due to delay X => Y

Dynamics

Model with differential equations:

$$dY/dt = B_y + \beta_y f(X^*, K_{xy}) - \alpha_y Y$$

$$dZ/dt = B_z + \beta_z G(X^*, K_{xz}, Y^*, K_{yz}) - \alpha_z Z\,.$$



thick and medium lines:
coherent FFL type I
(different strengths Y=>Z)

thin line: simple system

AND: **delayed**
response to Sx-on

OR: **delayed**
response to Sx-off

=> Handle **fluctuating signals** (on- or off-fluctuations)

Mangan, Alon, *PNAS* **100** (2003) 11980

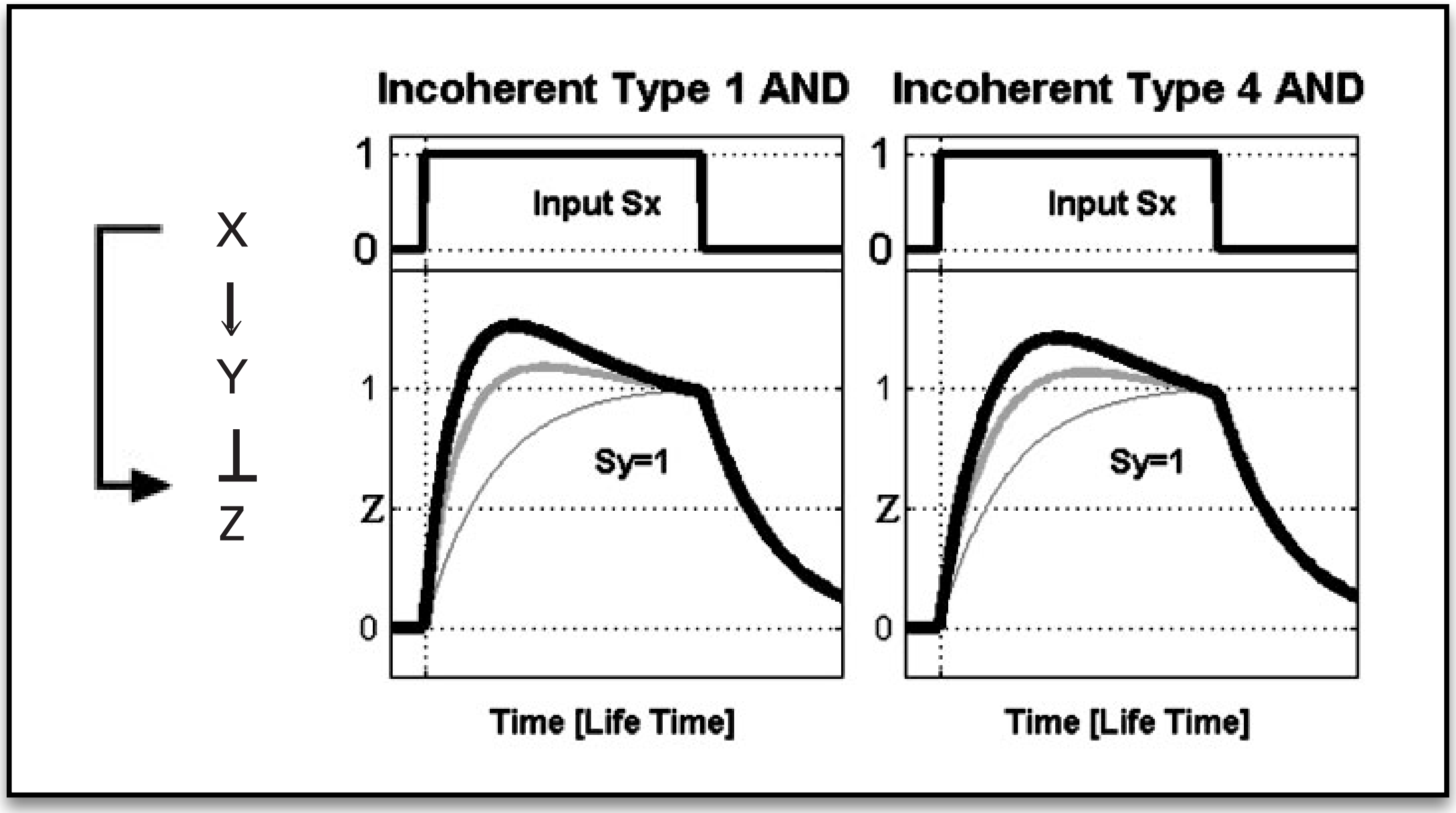
Fast Responses

Scenario: we want a fast response of the protein level
- gene regulation on the minutes scale
- protein lifetimes O(h)

At **steady state**: protein production = protein degradation

=> degradation determines $T_{1/2}$ for given stationary protein level

=> for fast response: faster degradation or negative regulation of production

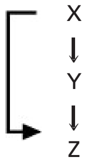
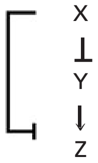
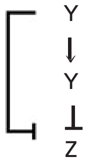
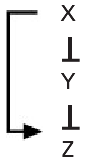


On the genes:
no autoregulation for
protein-coding genes
=> incoherent FFL for
upstream regulation

Mangan, Alon, *PNAS* **100** (2003) 11980

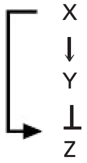
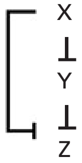
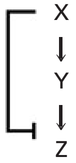
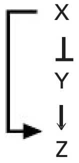
All Behavioral Patterns

Table 1. Structure and function of the coherent FFL types, with AND- and OR- gates at the Z promoter

	Coherent type 1		Coherent type 2		Coherent type 3		Coherent type 4	
Species	Structure	Abundance	Structure	Abundance	Structure	Abundance	Structure	Abundance
E. coli	 X ↓ Y ↓ Z	28	 X ⊥ Y ↓ Z	2	 Y ↓ Y ⊥ Z	4	 X ⊥ Y ⊥ Z	1
S. cerevisiae		26		5		0		0
Z Logic→	AND	OR	AND	OR	AND	OR	AND	OR
Steady-state Z(Sx,Sy)	$S_x \wedge S_y$	S_x	$\bar{S}_x \wedge S_y$	$\bar{S}_x$	$\bar{S}_x$	$\bar{S}_x \wedge \bar{S}_y$	S_x	$S_x \vee \bar{S}_y$
Response delay								
Sx on step	Delay	—	—	Delay	—	—	Delay	Delay
Sx off step	—	Delay	Delay	—	Delay	Delay	—	—
Inverted out	No	No	Yes	Yes	Yes	Yes	No	No

Coherent FFL types and their abundance in transcription databases of *E. coli* and *S. cerevisiae* (6, 11). Z(Sx,Sy): Steady-state Z expression of coherent FFLs for the four combinations of Sx and Sy on and off levels (∧,∨,¯ represent AND, OR, NOT). Response: Response delay of coherent FFLs to on and off S_x steps in the presence of Sy. —, not delayed. Inverted out means that Z goes off in response to Sx on step.

Table 2. Structure and function of the incoherent FFL types, with AND-gates at the Z promoter

	Incoherent type 1		Incoherent type 2		Incoherent type 3		Incoherent type 4	
Species	Structure	Abundance	Structure	Abundance	Structure	Abundance	Structure	Abundance
E. coli	 X ↓ Y ⊥ Z	5	 X ⊥ Y ⊥ Z	0	 X ↓ Y ↓ Z	1	 X ⊥ Y ↓ Z	1
S. cerevisiae		21		3		1		0
Z logic →	AND		AND		AND		AND	
Steady-state Z(Sx,Sy)	$S_x \wedge \bar{S}_y$		$\bar{S}_x \wedge S_y$		0		0	
Pulse								
Sx on step	Weak		—		—		Strong	
Sx off step	—		Weak		Strong		—	
Sy effect	Destroy		Destroy		Enable		Enable	
Response acceleration								
Sx on step	Accelerate		—		—		Accelerate	
Sx off step	—		Accelerate		Accelerate		—	

Incoherent FFL types and their abundance in transcription databases (6, 11). Z(Sx,Sy): Steady-state Z expression of incoherent FFL with no basal level of Y (∨, ¯ represent AND, NOT). Pulse: Response to steps of Sx, in the presence of Sy, in FFLs with no basal activity, Sy effect on pulse: Enable, no pulse is created when Sy is off; Destroy, Z output is a low pulse when Sy is on, but is high and steady when Sy is off (Fig. 3). Response acceleration: Acceleration of response of and steady-state values of incoherent FFL with basal activity to on and off steps in the presence of Sy. —, not accelerated.

Mangan, Alon, *PNAS* **100** (2003) 11980

Summary

Today:

- Gene regulation networks have **hierarchies**:
  => global "cell states" with specific expression levels

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

Next lecture:

- Simple dynamic modelling of transcription networks
  => Boolean networks, Petri nets