Structure and dynamics of feedback in the large scale genetic regulatory network of *E. coli*

Sanjay Jain

Department of Physics and Astrophysics, University of Delhi, Delhi.

Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore.

Santa Fe Institute, Santa Fe, New Mexico, USA.

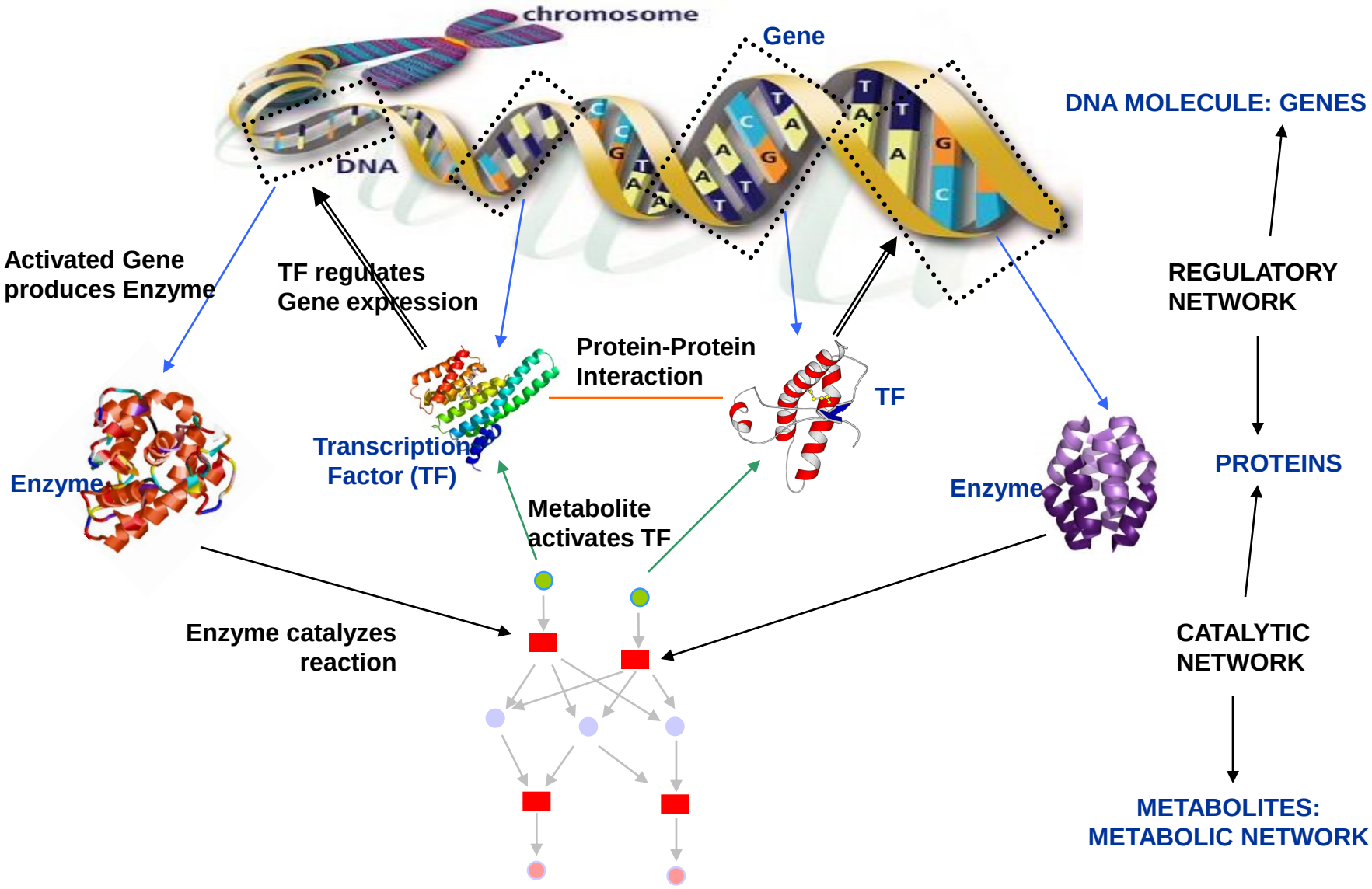
Homepage: http://people.du.ac.in/~jain/

Email: jain@physics.du.ac.in, jain_physics@yahoo.co.in

Work done at the Physics Department, University of Delhi
in collaboration with

Areejit Samal
Saurabh Mahajan
Megha Mittal
Manoj Raghav

Mutually Interacting Networks inside a cell



Outline

Regulon database: http://regulondb.ccg.unam.mx/ (Salgado et al, Nucl. Acids Res. 2006).

Freyre-Gonzales et al, Genome Biol. 2008.

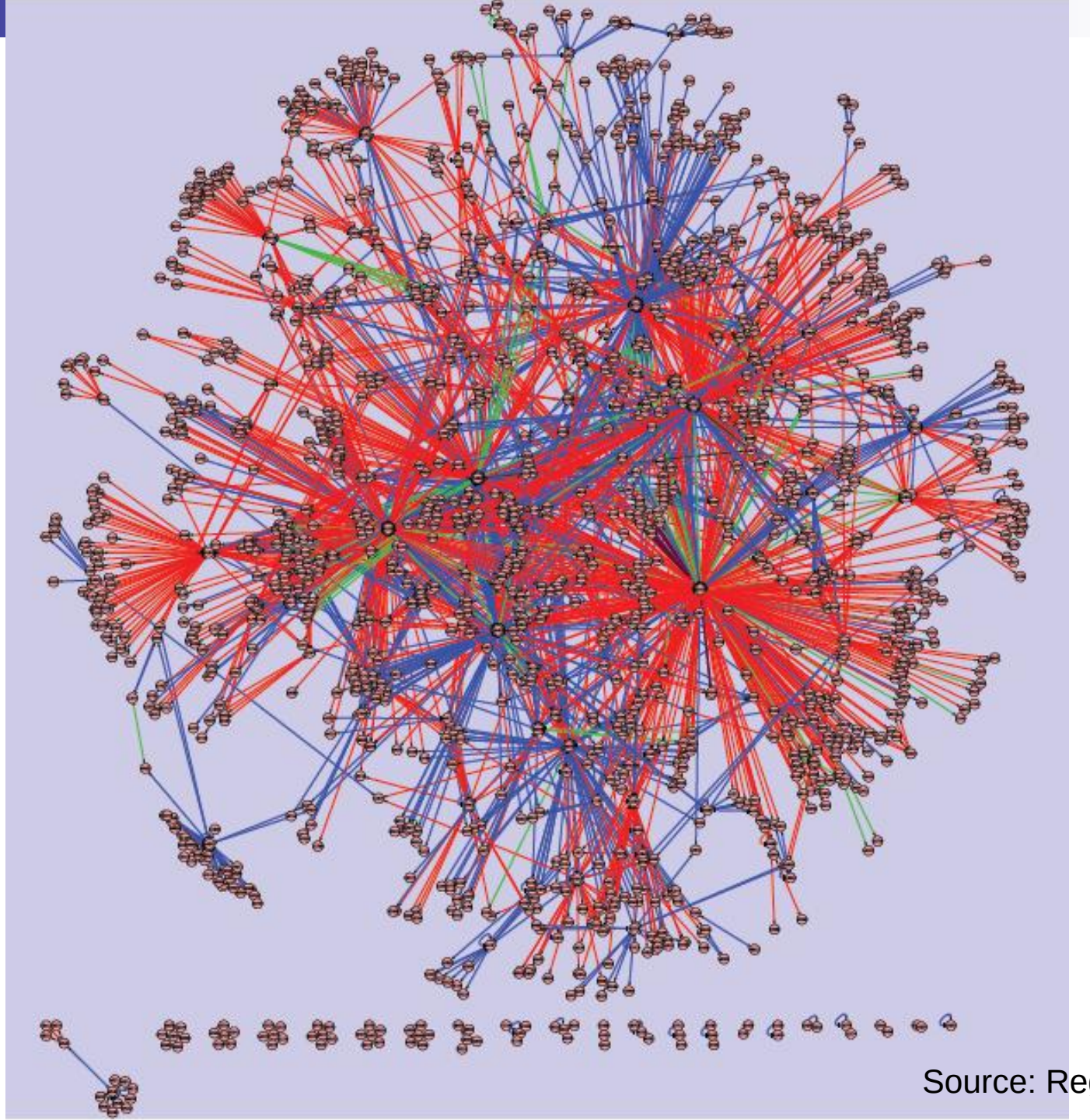
Database that lists more than 5000 interactions among about 2000 genes of E. coli (currently)

Description of network structure, including nature of feedbacks

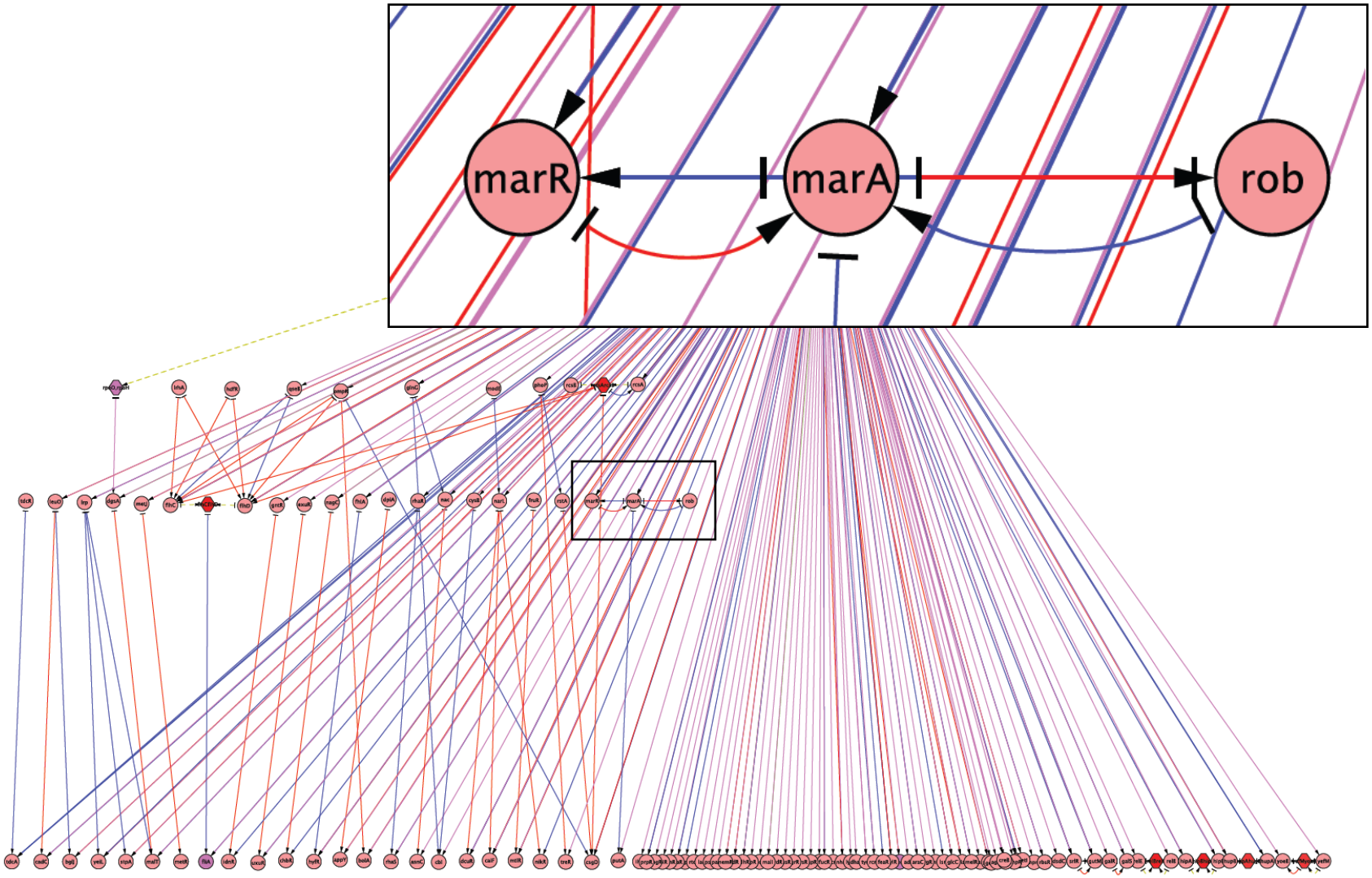
Dynamics: Boolean dynamics of the regulation of metabolism

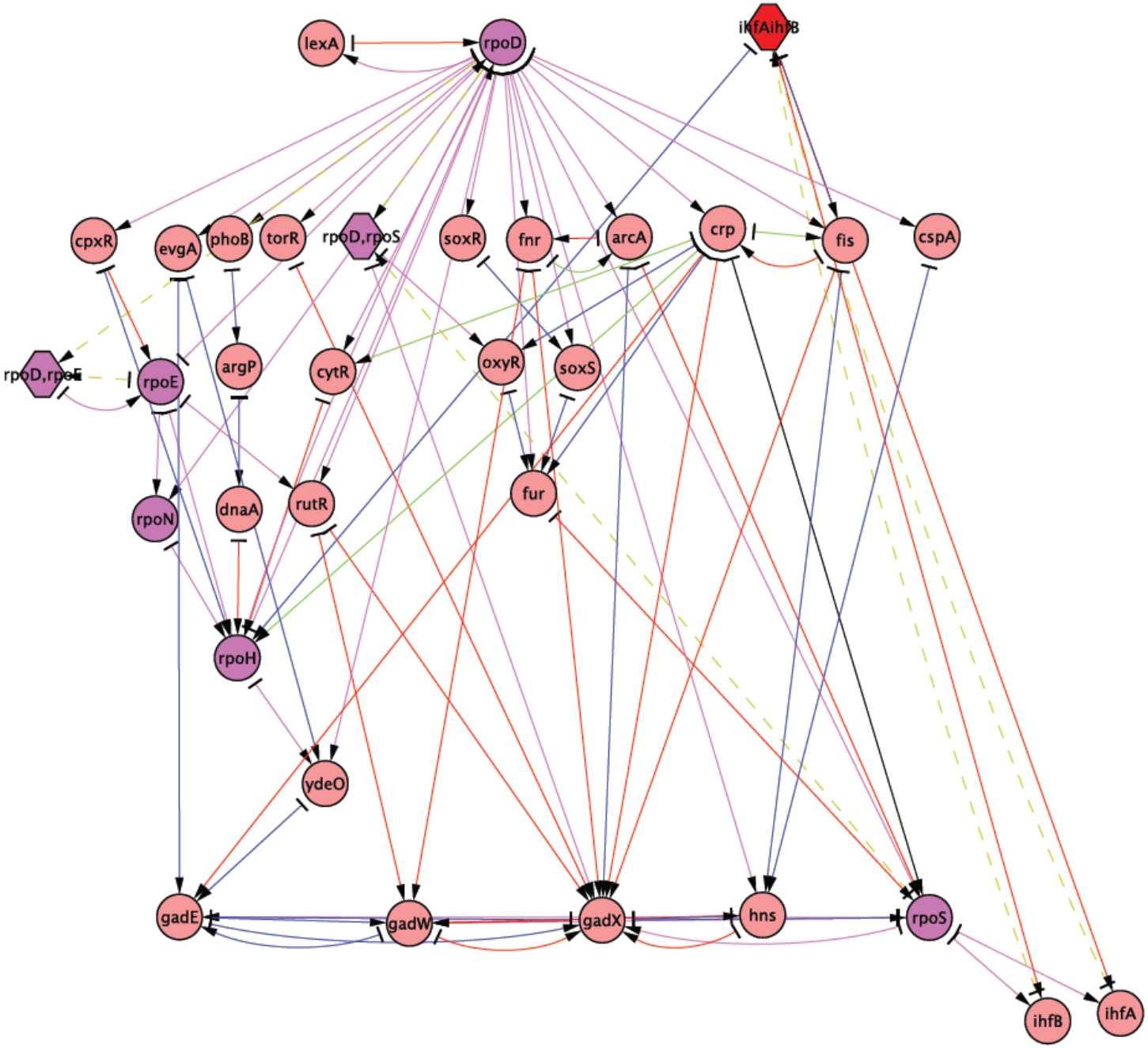
Dynamics of the core

Conclusions and unanswered questions

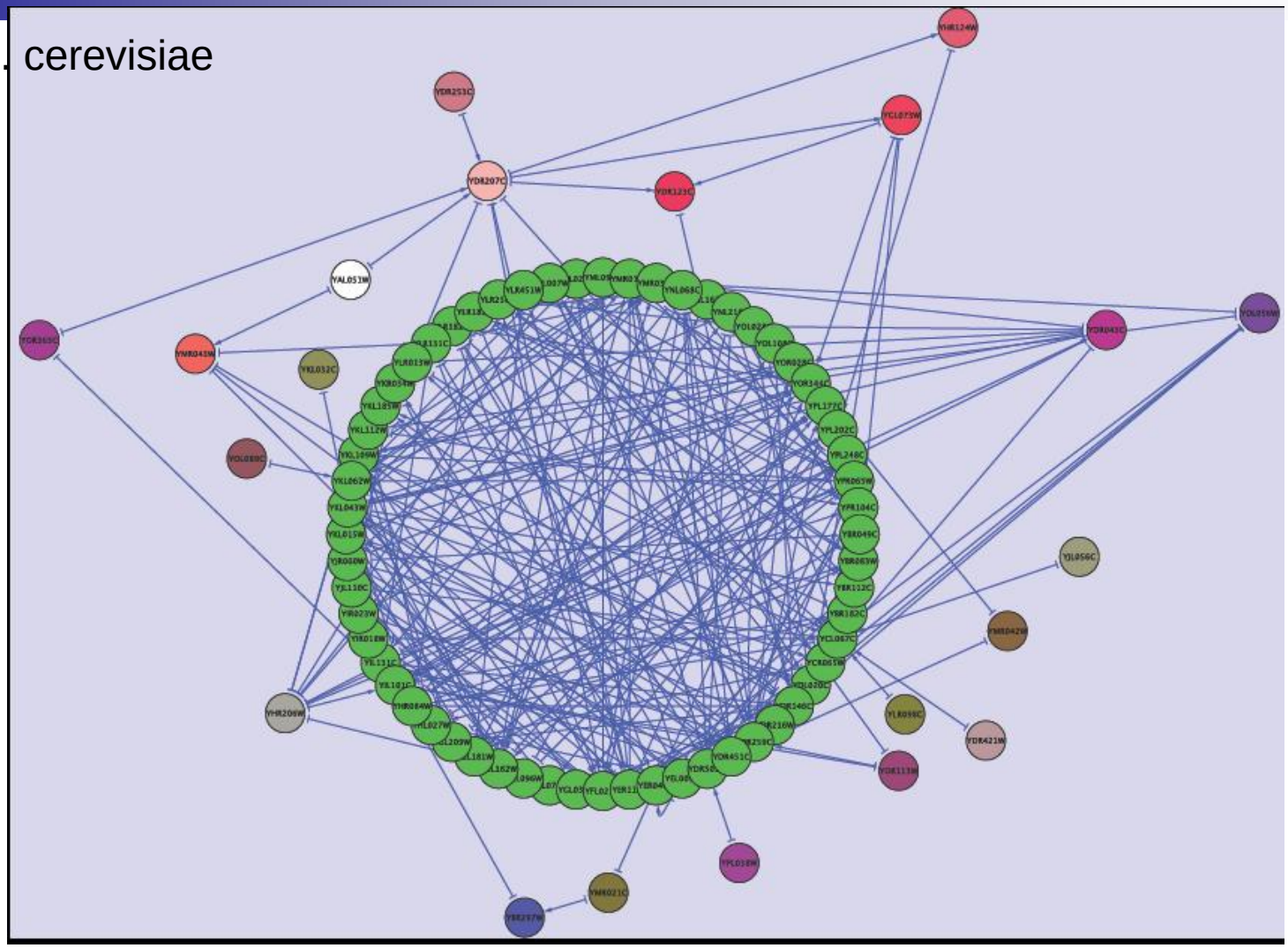


Source: Regulon DB


marR
marA
rob

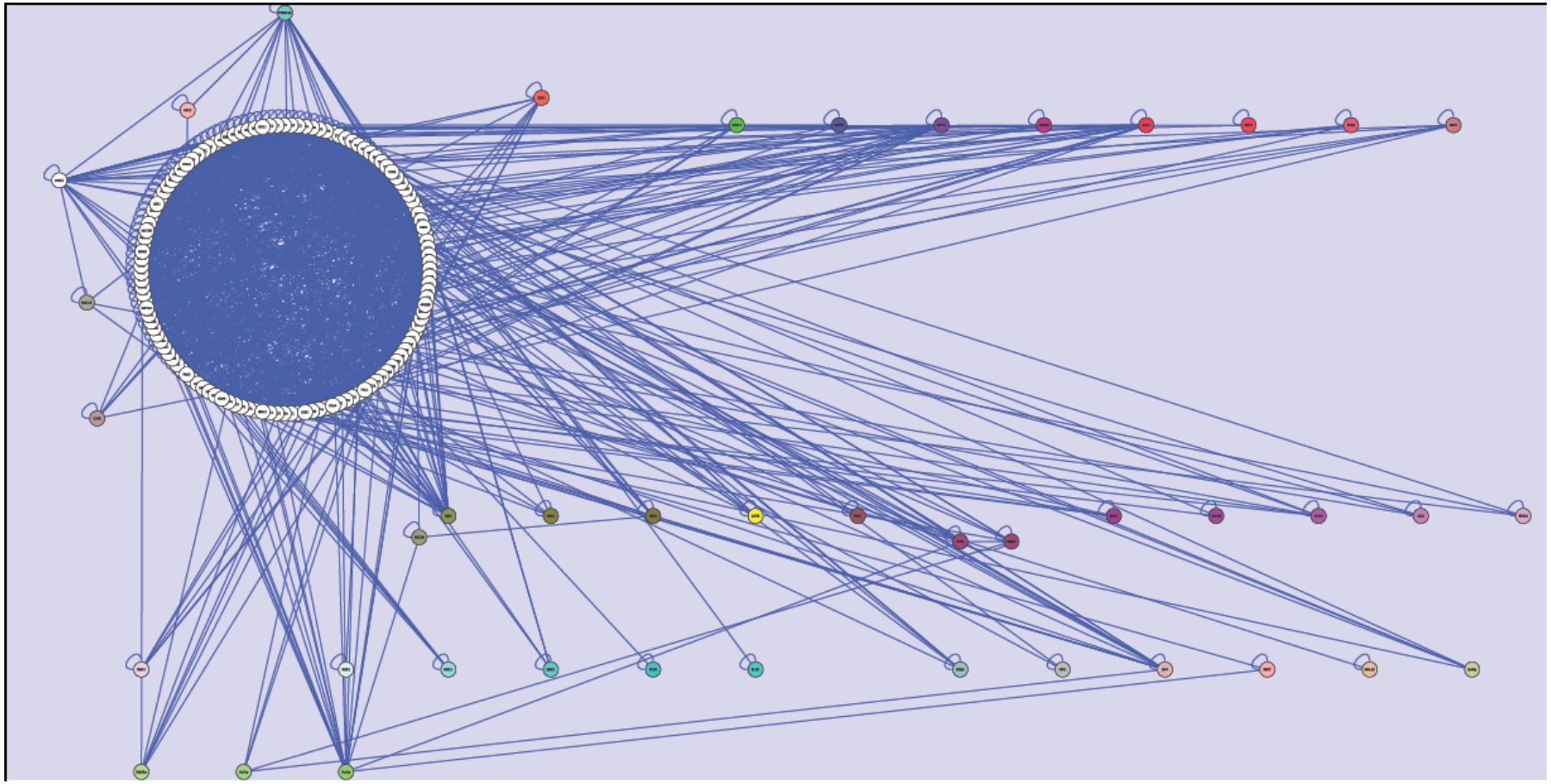

lexA
rpoD
ihfAihfB
cpxR
evgA
phoB
torR
rpoD,rpoS
soxR
fnr
arcA
crp
fis
cspA
rpoD,rpoE
rpoE
argP
cytR
oxyR
soxS
rpoN
dnaA
rutR
fur
rpoH
ydeO
gadE
gadW
gadX
hns
rpoS
ihfB
ihfA

For S. cerevisiae



Tnet Database: 4410 Genes, 12873 Links, 157 TFs
Balaji, Madan Babu, Iyer, Luscombe, and Aravind, J. Mol. Bio. (2006) 360(1): 213-27.

22 Irreducible components
136 Nodes in the largest irreducible component
Hannes Beushausen, Areejit Samal, S.J.



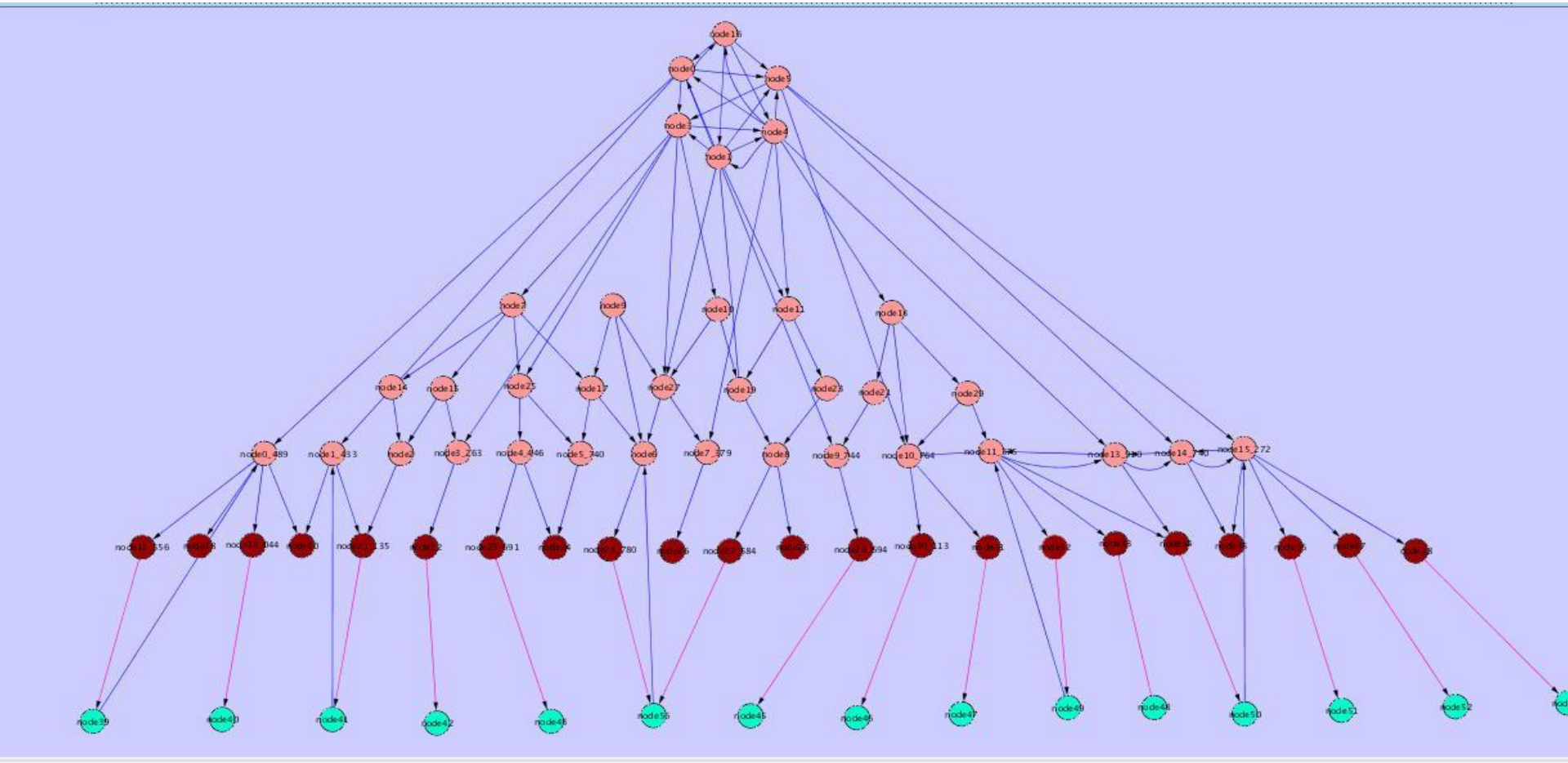
Yeastract Database: 5784 Genes, 31880 Links, 169 TFs

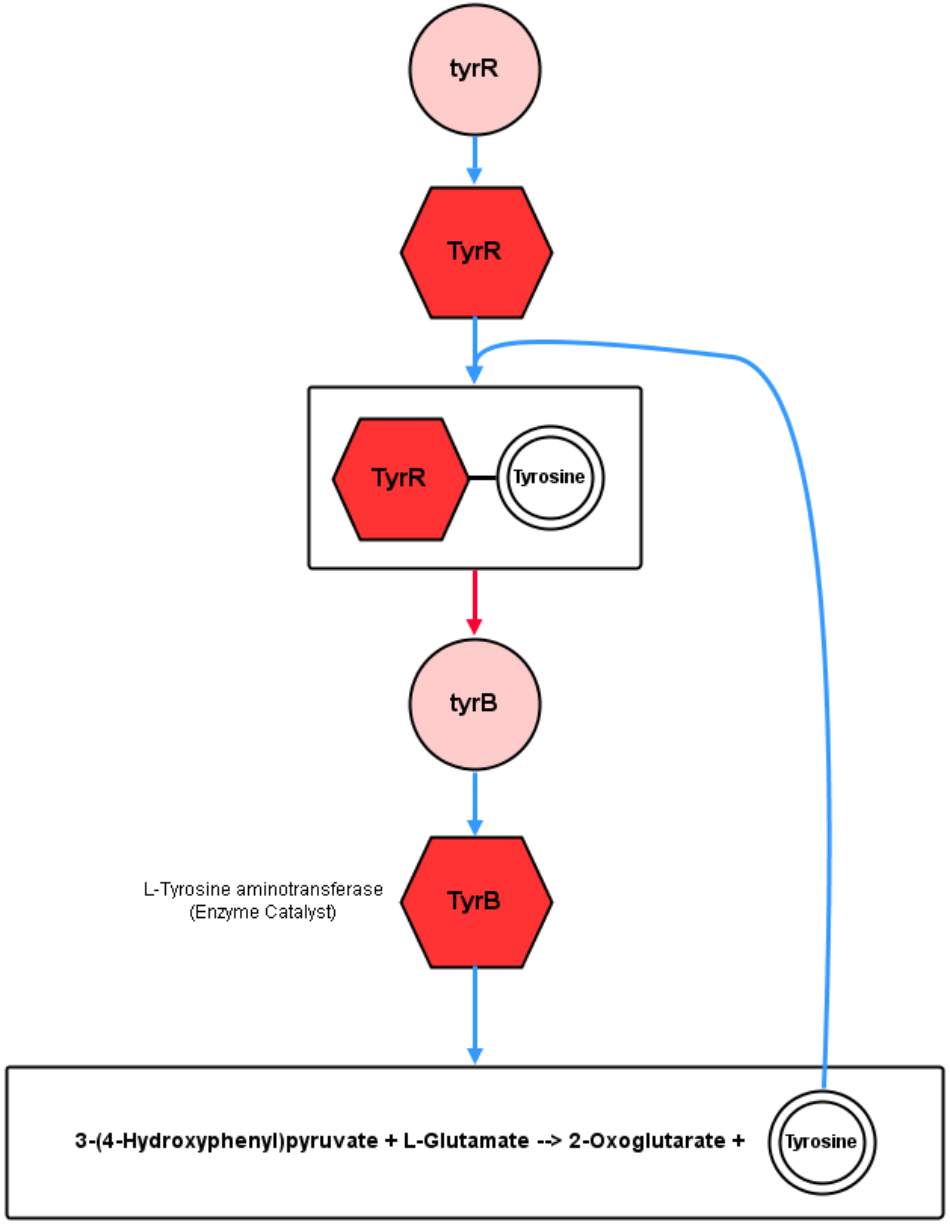
Ref: Teixeira, et. al., Nucl. Acids Res. (2006) and Monteiro, et. al. Nucl. Acids Res. (2008)

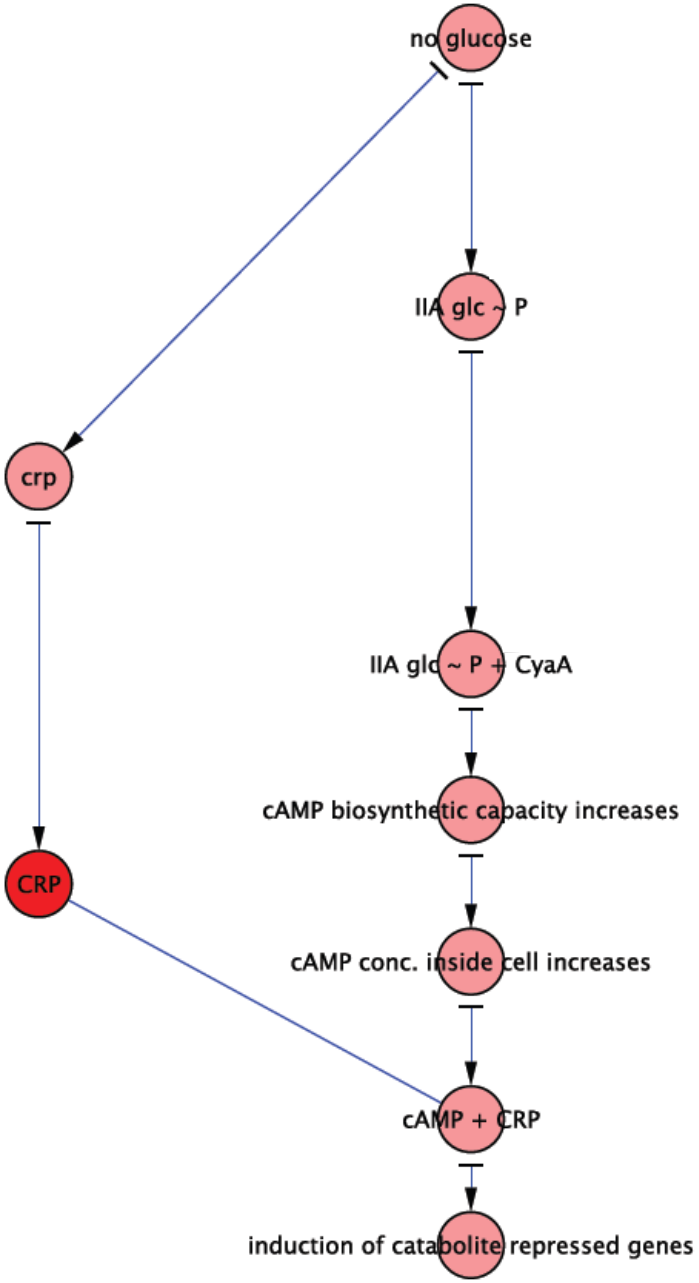
42 Irreducible components

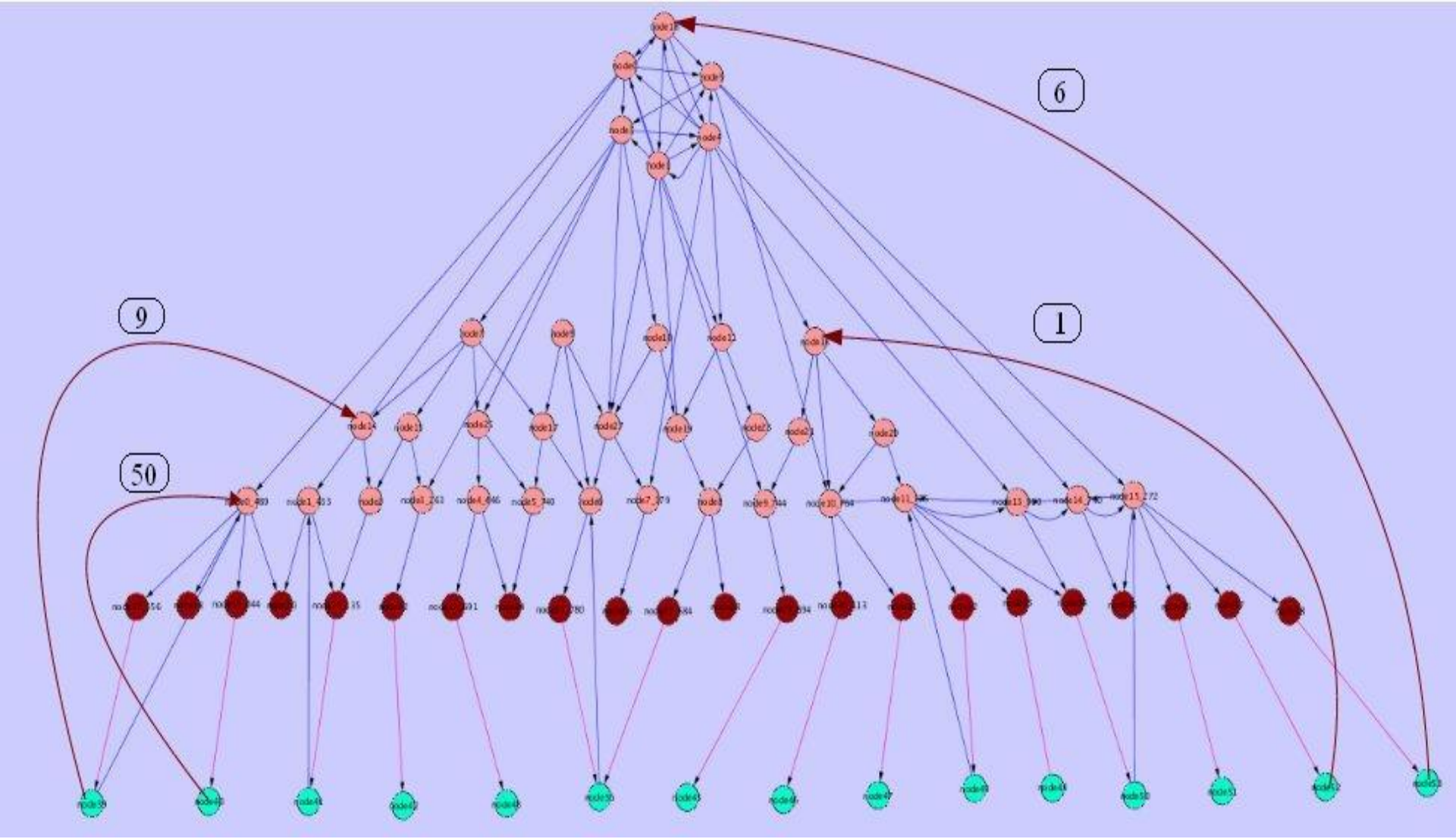
128 Nodes in the largest irreducible component

Hannes Beushausen, Areejit Samal, S.J.


node16
node0
node1
node3
node4
node12
node14
node15
node25
node17
node27
node19
node23
node21
node20
node0_489
node1_433
node2
node3_263
node4_946
node5_740
node6
node7_379
node8
node9_744
node10_964
node11_876
node13_949
node14_780
node15_272
node36
node40
node41
node42
node43
node56
node45
node46
node47
node49
node48
node50
node51
node52


tyrR
TyrR
TyrR
Tyrosine
tyrB
L-Tyrosine aminotransferase
(Enzyme Catalyst)
TyrB
3-(4-Hydroxyphenyl)pyruvate + L-Glutamate --> 2-Oxoglutarate +
Tyrosine


no glucose
IIA glc ~ P
crp
IIA glc ~ P + CyaA
cAMP biosynthetic capacity increases
CRP
cAMP conc. inside cell increases
cAMP + CRP
induction of catabolite repressed genes


6
9
1
50

Dynamics: Some system level questions

- **Does the cell have homeostasis under perturbations of the genes' configuration?**

  The internal state of the cell is subject of lots of fluctuations, e.g., perturbations in the genes' on-off configuration. After the perturbation the cell needs to revert back to its old tasks, and not get derailed into a completely different trajectory.

- **Does the cell have flexibility and adaptability under environmental changes?**

  When the environment changes in a sustained way, the cell needs to modify its internal state and processes in such a way so as to be able to function in and take advantage of the changed circumstances.

Main result: The cell seems to have both the above dynamical properties, and that is a consequence of the architecture of the regulatory network.

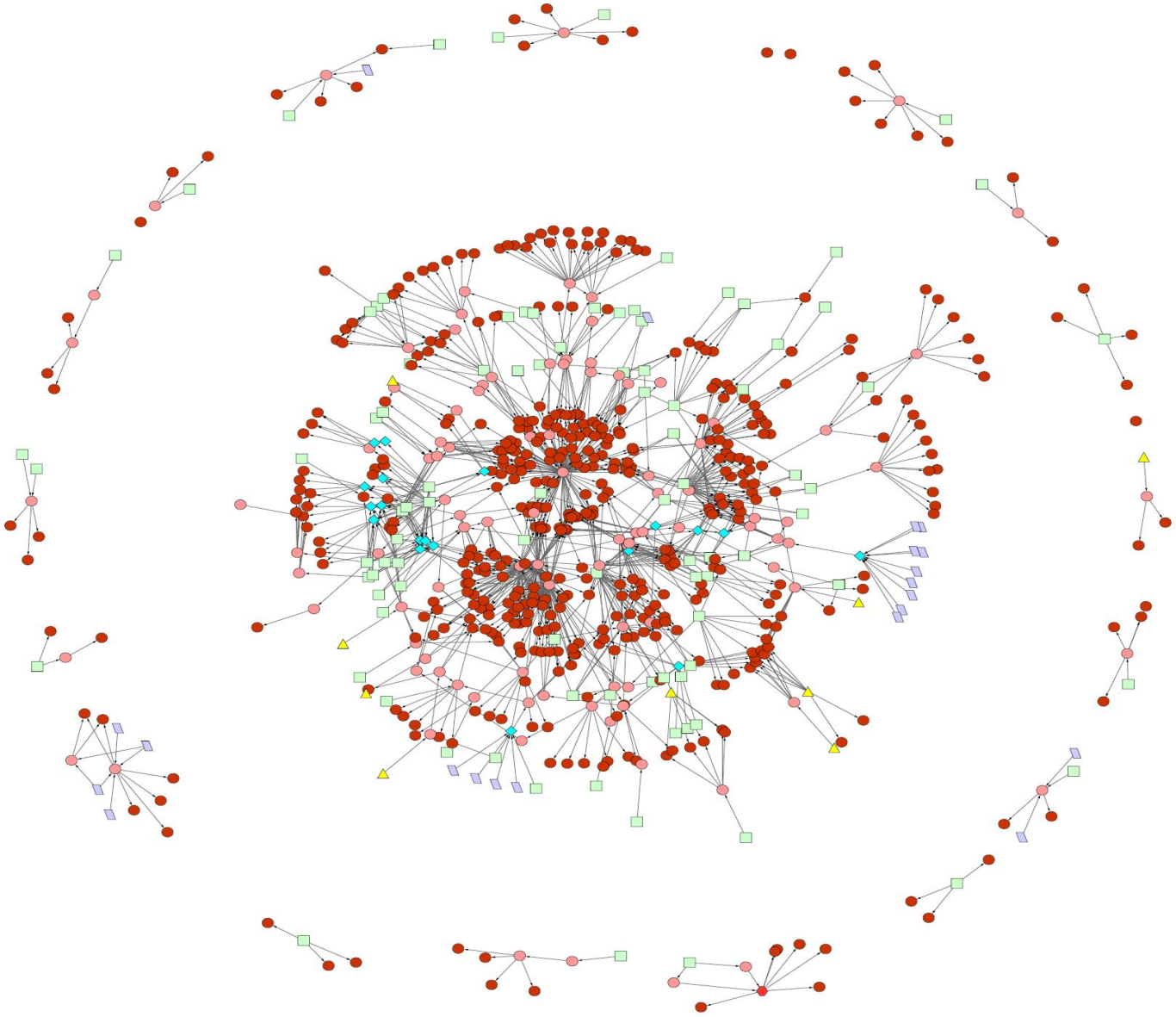
Description of the *E. coli* Genetic network controlling metabolism

- The network has 583 genes including
  - 479 genes coding for enzymes
  - 104 genes coding for 103 transcription factors
- 96 external metabolites (whose presence or absence switches certain genes on or off)
- 19 conditions (e.g., pH)
- 21 internal fluxes of metabolic reactions (representing the presence or absence of internal metabolites)
- 9 stimuli (stress, heat shock, etc.).

In the network, a directed link from one node to another means that the former regulates the latter

Network database available at: http://gcrg.ucsd.edu (Bernhard Palsson's Group Webpage)

Reference: M.W. Covert, Knight, Reed, Herrgard and Palsson, **Nature** 2004;
Barrett, Herring, Reed and Palsson, **PNAS** 2005

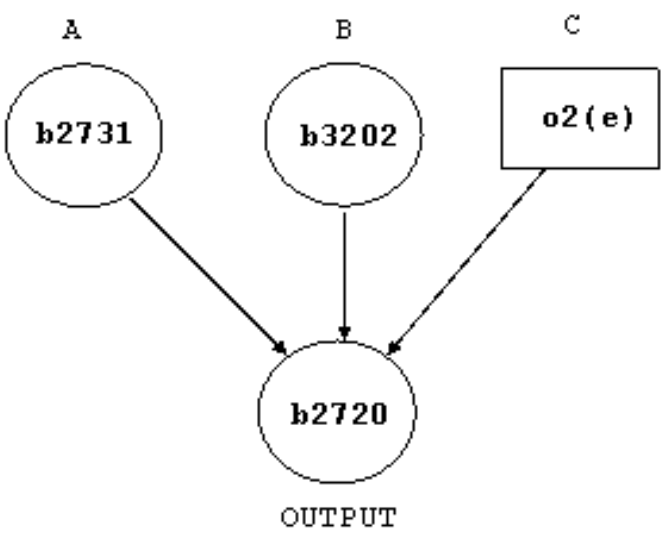


The graph shows the Genetic network controlling metabolism in *Escherichia coli*. The pink nodes represent genes coding for transcription factors, brown nodes represent genes that code for enzymes and the green nodes represent external metabolites.

Areejit Samal and S. J., arxiv:q-bio.MN/0703059, BMC Systems Biology 2:21 (2008)

The database also gives the regulatory logic of every gene as a function of its input nodes' configuration, i.e., given the on-off status of all its input nodes, whether the gene will be turned on or off.

Example of an input function in the form of a boolean rule



Truth Table

A	B	C	OUTPUT
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	0

b2720 = IF (b2731 AND b3202 AND NOT (o2(e)))

The dynamical system

Using the information in the database, we construct the following discrete dynamical system:

$$g_i(t+1) = G_i(\vec{g}(t), \vec{m})$$

where i = 1 …583

$\vec{g}(t)$

$$\begin{bmatrix} g_1(t) \\ g_2(t) \\ g_3(t) \\ . \\ . \\ . \\ g_{583}(t) \end{bmatrix}$$

g_i = 1 if the ith gene is on,
g_i = 0 if it is off.

State of the genetic network at time *t*

$\vec{m}$

$$\begin{bmatrix} m_1 \\ m_2 \\ . \\ . \\ m_{96} \end{bmatrix}$$

m_i = 1 if metabolite i is present in the external environment, m_i = 0 if it is absent.

In general, the concentrations of external metabolites change in time. In the present study, we have considered buffered minimal media (i.e., vector ***m*** constant in time).

Configuration of the external environment

In this approach, time is taken as discrete, i.e., t = 1,2,…

G_i: boolean functions containing all the information of the internal wiring of the network as well as the logic of each gene's regulation.

Stuart Kauffman (1969,1993) studied dynamical systems of the form: $g_i(t+1) = G_i(\vec{g}(t))$ with random network and random Boolean functions

Treatment of the external environment

The external environment: **m** = (m_1,m_2,…,m_{96}).

96 = 86 organic molecules + 10 inorganic ones that can enter the cell.

The other stimuli like stress, heat shock, etc., are assumed to be absent.

We considered two libraries of **minimal** media:

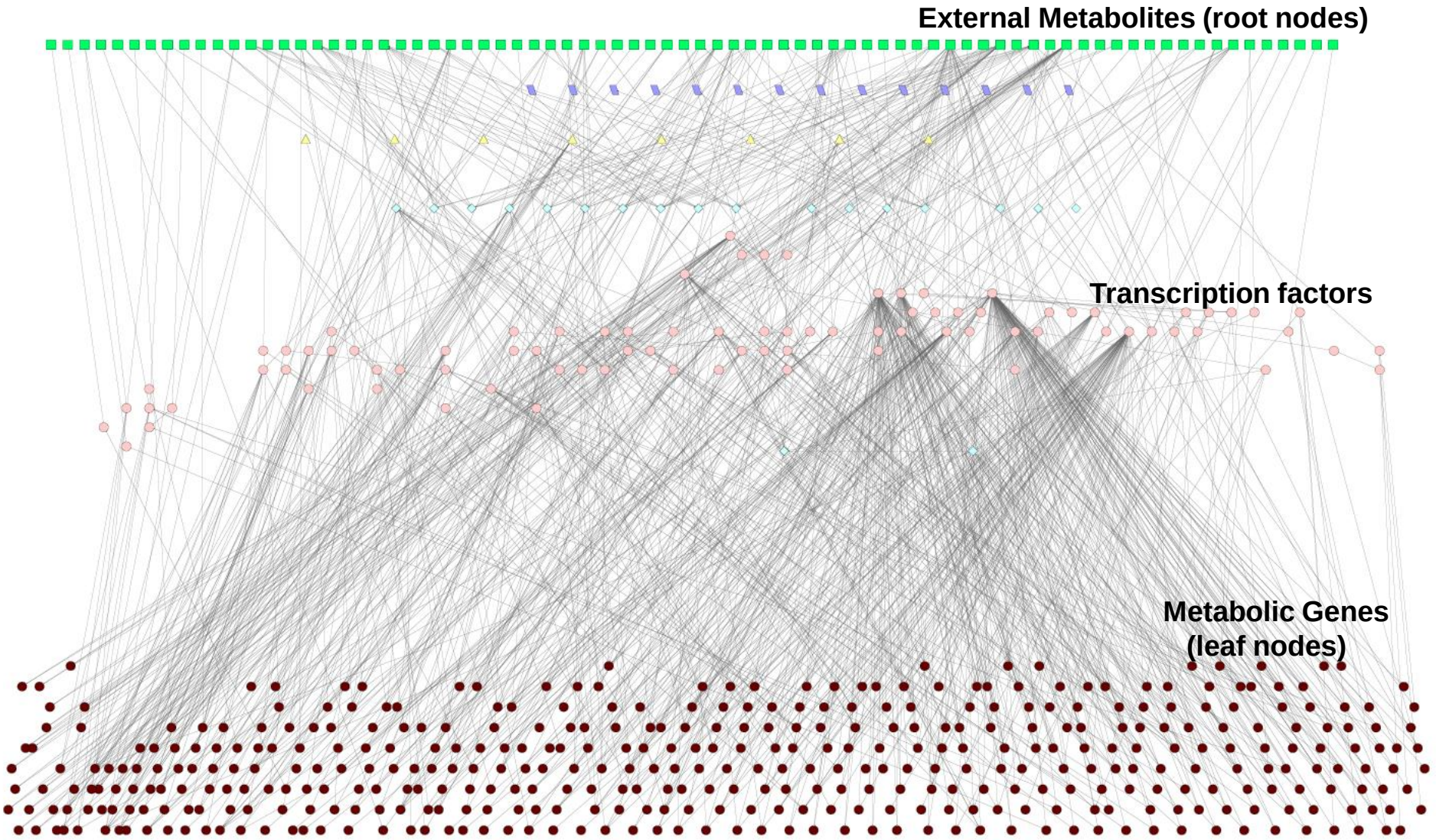
(a) **93** minimal media (61 aerobic and 32 anaerobic).

- Containing only one of the 86 organic sources of carbon.
- Inorganic ions present: ammonium, sulphate, phosphate, hydrogen, iron, potassium and sodium, (and oxygen for aerobic media).
- Only those media were chosen in which the metabolic network supports a nonzero growth rate as determined by Flux Balance Analysis (FBA).

(b) A much larger library of **15427** minimal media

- Containing all possible combinations of single sources of carbon, nitrogen, sulphur, phosphorus, etc., from among the various metabolites ingested by E. coli and which also support a nonzero growth rate of the cell as determined by FBA.

Structure of the largest connected component



This is an acyclic graph, with maximal depth 4, and external metabolites as root nodes

Cellular Homeostasis

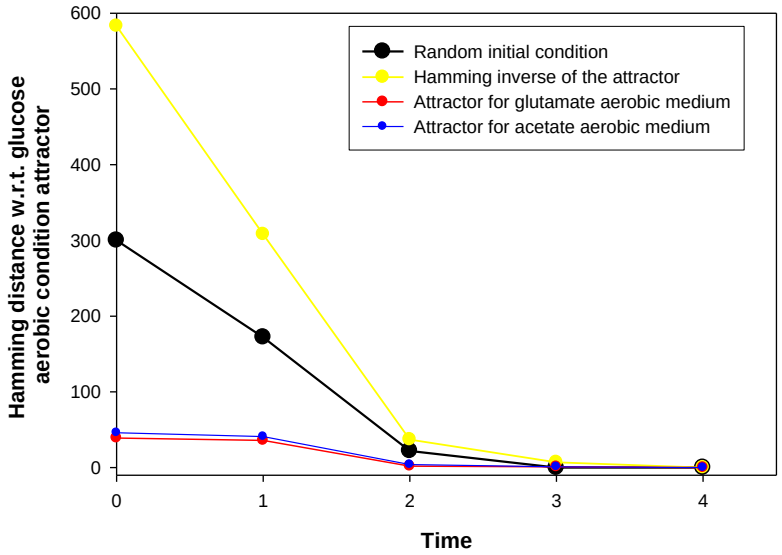
System A

For a fixed environment (e.g., glucose aerobic minimal medium) a fixed point is the global attractor. Sampled 10000 initial conditions for each of 93 media.

$$g_i(t+1) = G_i(\vec{g}(t), \vec{m})$$

Start with different $\boldsymbol{g}(t)$ as initial configuration of genes, and determine the attractor for the system for each initial configuration of genes.

Fix $\boldsymbol{m}$ to some buffered minimal media e.g. Glucose aerobic condition



A fixed point global attractor also found for each medium in the larger library of 15427 minimal media.

Maximum time to reach attractor was 4 time steps.

Thus any perturbation of the genes' configuration will be washed out in a few time steps. The system is robust to such perturbations.

System B:

Sampled 1000 randomly chosen initial conditions for each of 93 media.

For 89 media: 36 distinct attractors (fixed points: 8, and 2-cycles: 28).

For 4 media: 10 distinct attractors (fixed points: 4, and 2-cycles: 6).

Again maximum time taken to reach the attractor: 4 time steps.

For cycles, no. of genes locked in a fixed configuration = 562 to 567.

no. of oscillating genes = 21 to 16.

Thus we have a **large "**frozen core**"** and **small "**twinkling islands**"** in each attractor (Kauffman's nomenclature). *E. coli*'s genetic network seems to be in the "ordered" regime.

Each of the frozen genes has the same configuration across all the attractors (36 or 10). Thus for any given medium most genes (562 or more out of 583) will end up in same configuration, independent of initial conditions.

Conclusion: High degree of **homeostasis**. Genetic perturbations die out fast, restoring an overwhelming majority of genes to a configuration that is independent of the perturbation.

The *E. coli* TRN exhibits flexibility of response to changed environmental conditions

$$g_i(t+1) = G_i(\vec{g}(t), \vec{m})$$

Determine the attractors of the genetic system for different environments ***m***

Vary ***m*** across a set of 15427 buffered minimal media

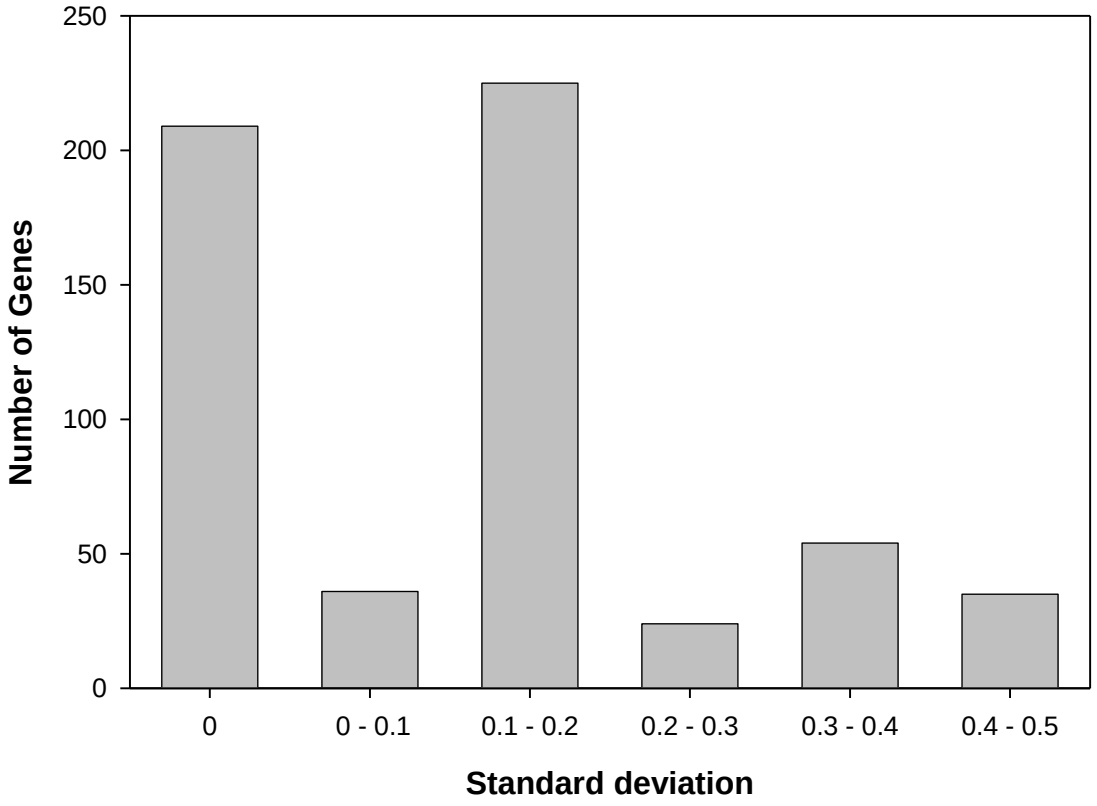
- We asked the question:
  - How *different* are the global attractors from each other for different environmental conditions?
  - We compared the various attractors of the system A for the library of 15,427 environmental conditions.
  - Result: The *largest hamming distance* obtained between two attractors corresponding to different environmental conditions was 145.

- Therefore, while the system is insensitive to fluctuations in gene configurations in a fixed external environment, it can move to quite a different attractor when it encounters a change in environment.

No. of genes whose configuration was not the same in the 15427 attractors = 374.
No. of genes whose configuration was the same in all the 15427 attractors = 209.
(Of the 209 genes, 75 genes were always off and 134 always on.)

Each gene takes a value 0 or 1. The standard deviation of a gene's value across 15427 attractors is a measure of its variability across environmental conditions.

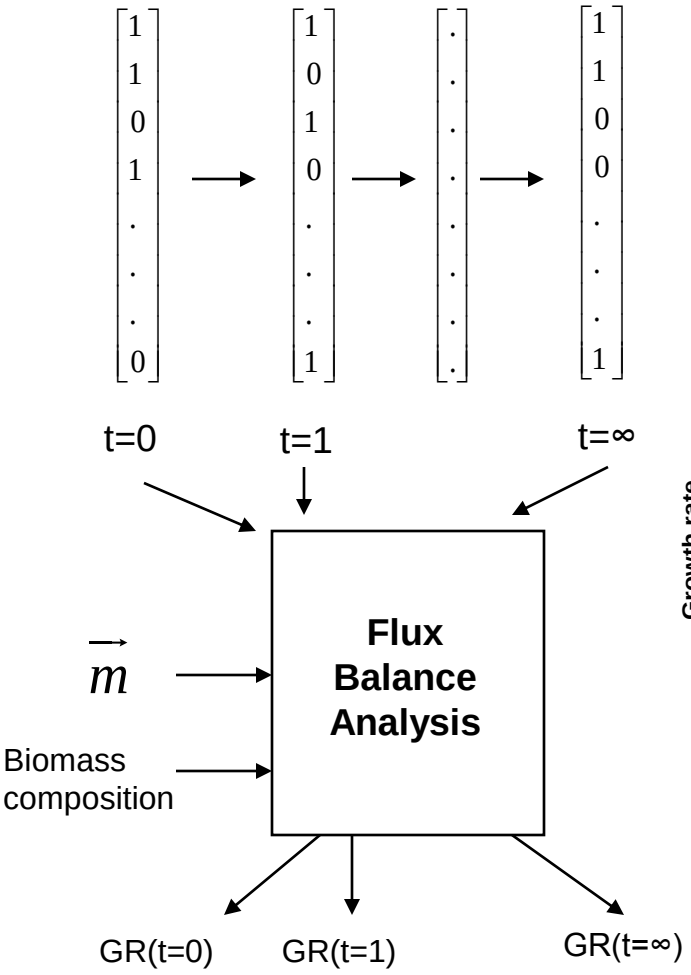
The histogram of standard deviations of a gene:



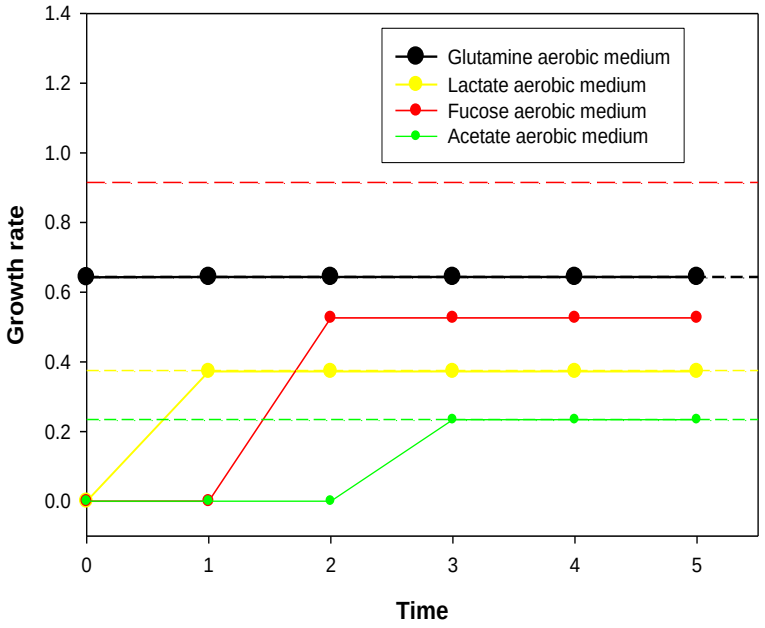
Note: maximum possible value of the standard deviation is 0.5.

This quantifies the considerable variety in a gene's variability across environmental conditions

Adaptability



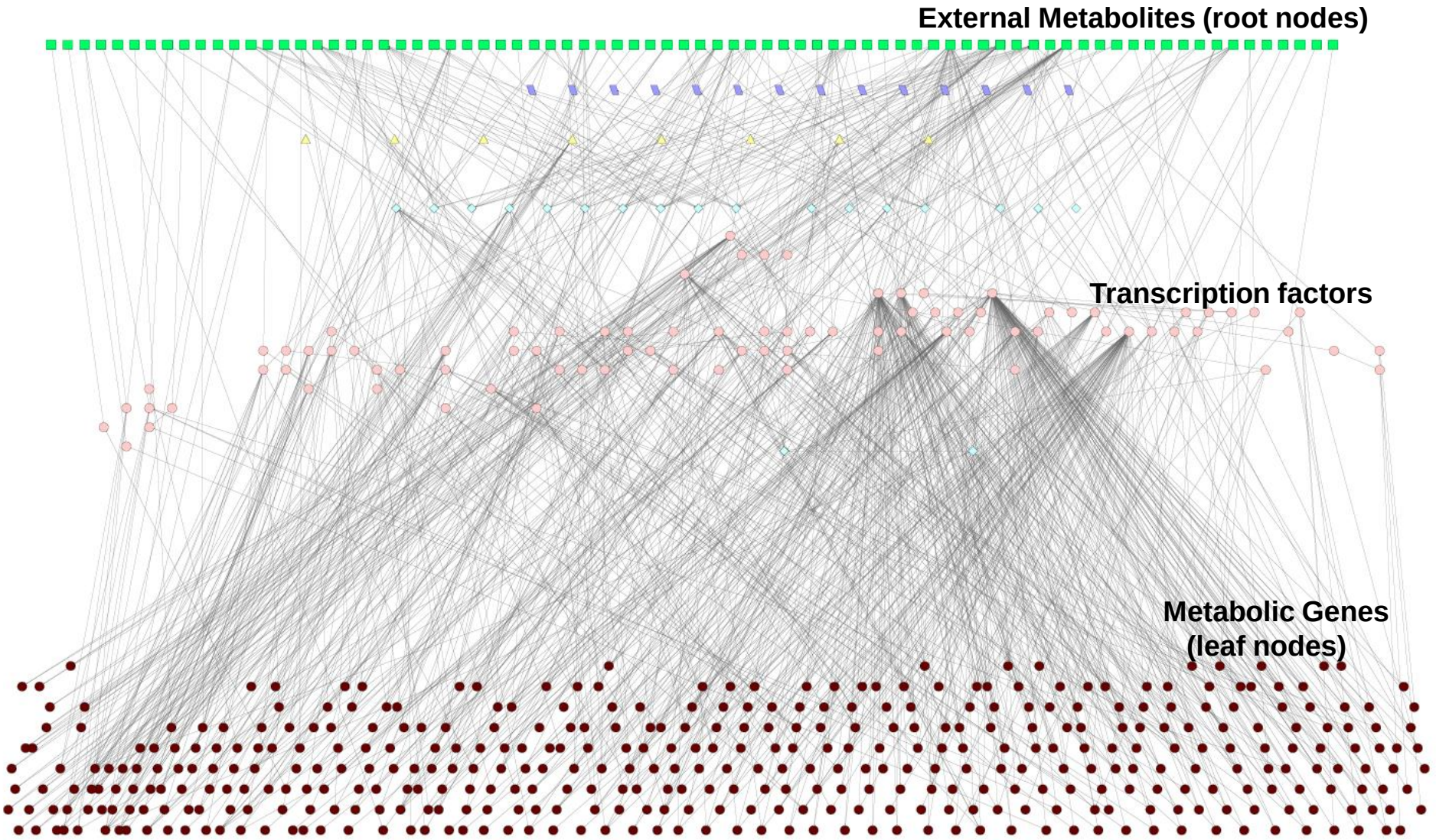
Question: How well is the attractor of any particular medium "adapted" to that medium? Does the movement to the attractor "improve" the cell's "metabolic functioning" in the medium?



Answer:
Growth rate increases by a factor of 3.5, averaged over pairs of minimal media
From one minimal medium to another the average time taken to reach the attractor is only 2.6 steps

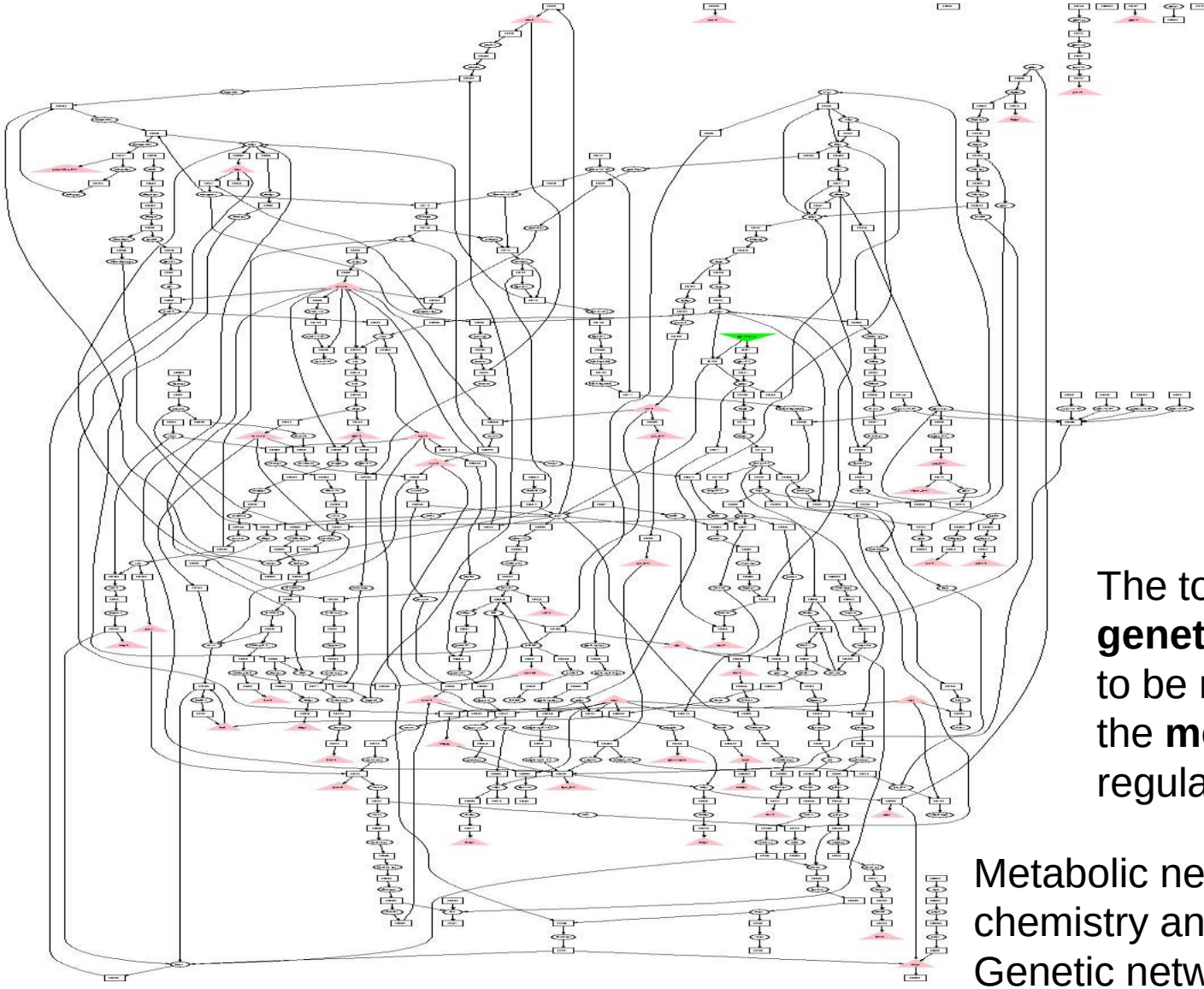
Thus the regulatory dynamics enables the cell to adapt to its environment to improve its metabolic efficiency very substantially, fairly quickly.

Graph structure explains homeostasis and flexibility



This is an acyclic graph, with maximal depth 4, and external metabolites as root nodes

The metabolic network: set of all active reactions in glucose aerobic minimal medium (obtained using Flux Balance Analysis)

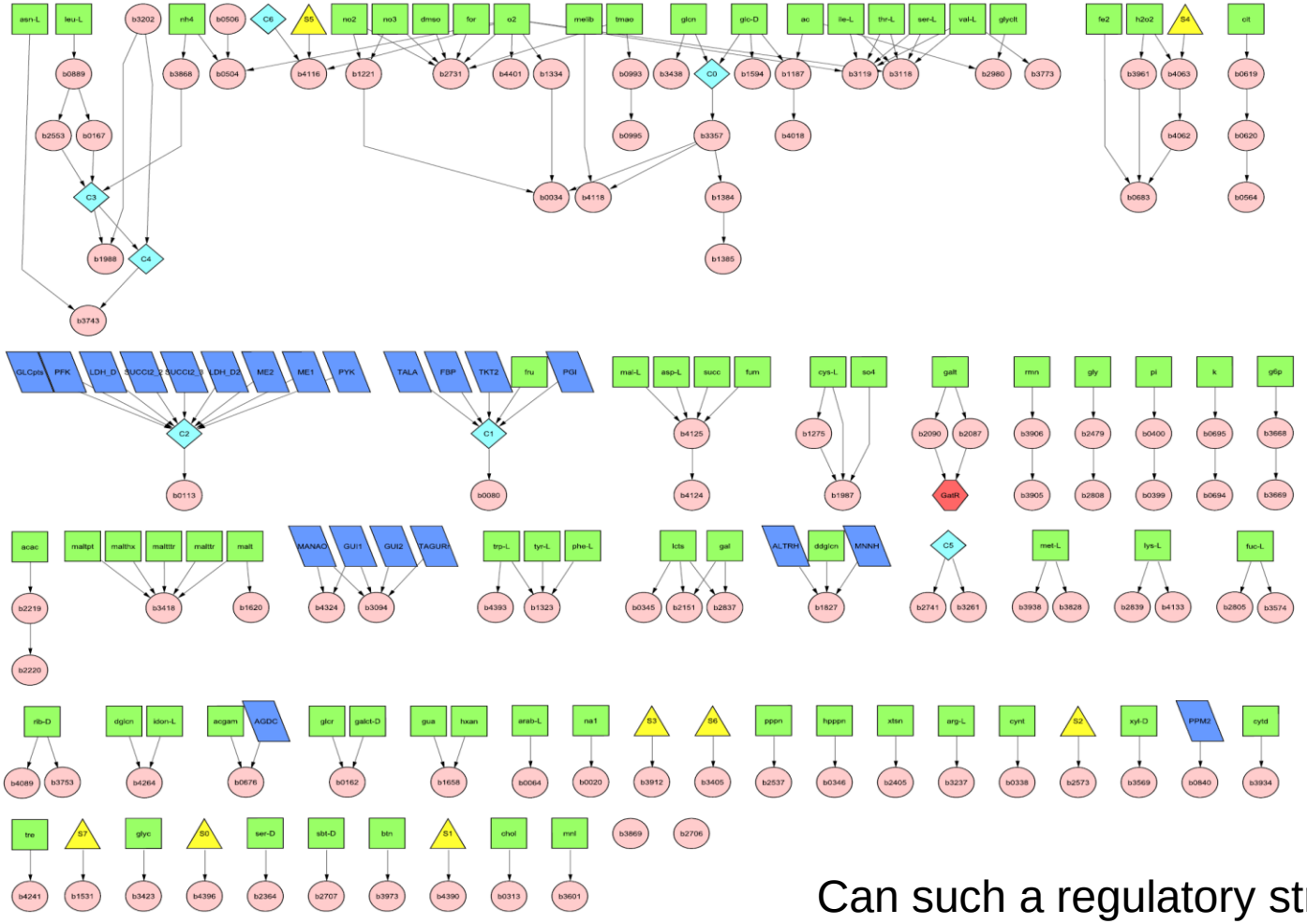


The topology of the **genetic** network seems to be much **simpler than** the **metabolic** network it regulates

Metabolic network: constrained by chemistry and physics
Genetic network: shaped by evolution

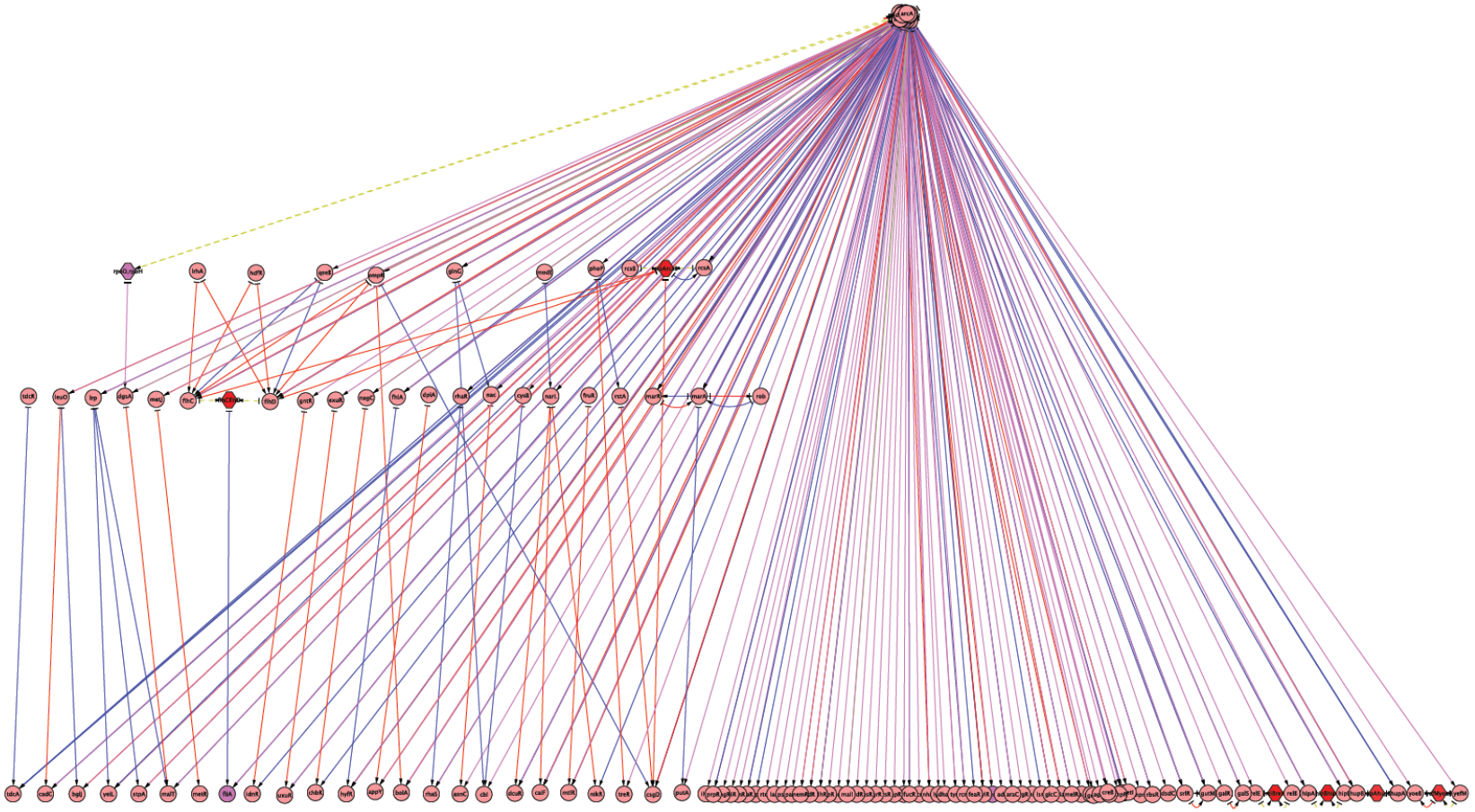
Modularity, flexibility and evolvability

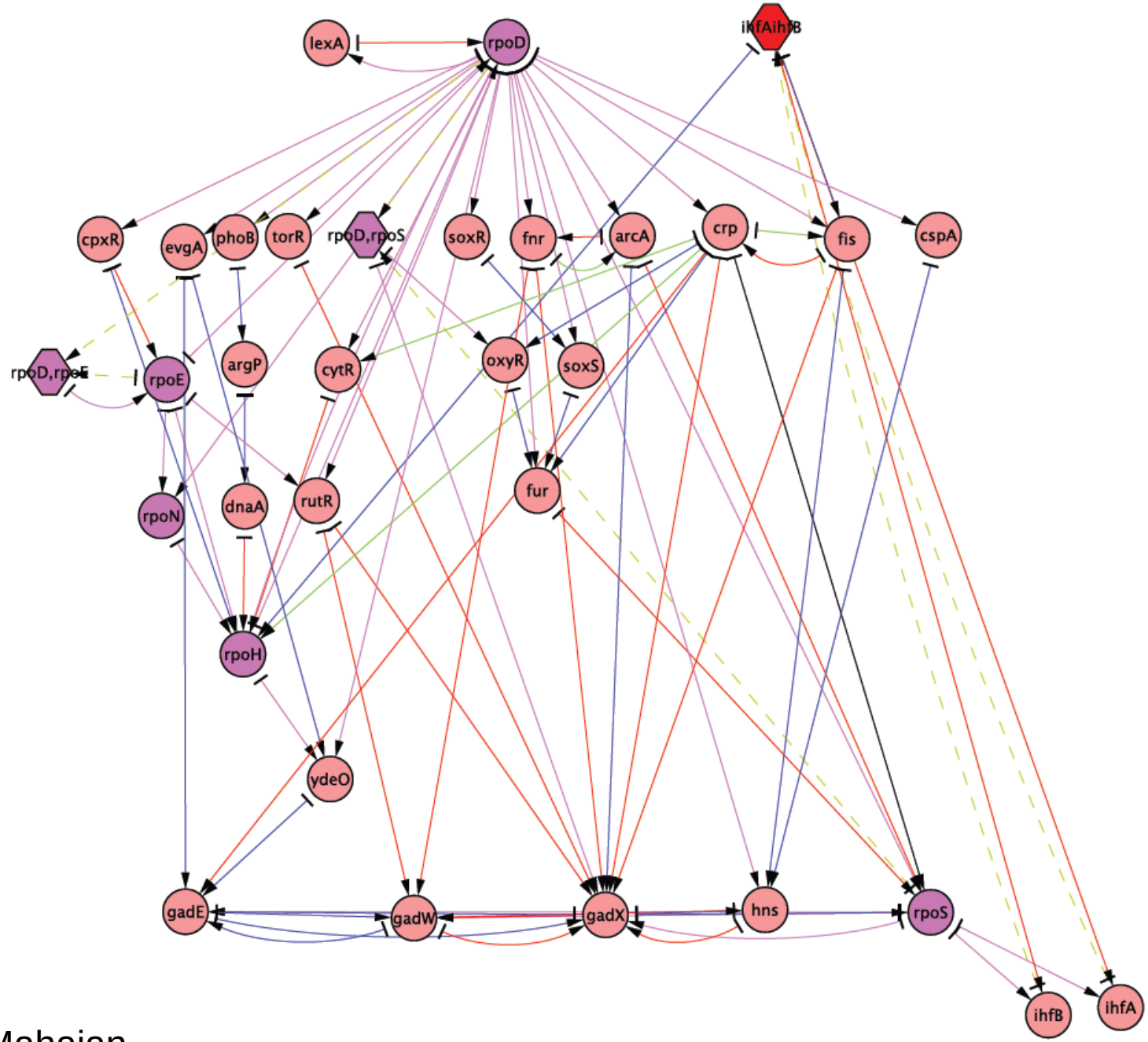
Consider the graph left after deleting the leaf nodes along with all their links. This is a highly disconnected graph. The large connected cluster has broken up into 38 components. Disconnected components are dynamically independent, hence modules.



While exploring new pathways, modularity ensures that existing ones remain intact

Can such a regulatory structure be useful for multicellular organisms?





Saurabh Mahajan

Results

- The attractors of this system for all environmental conditions considered are again fixed points.

- For a given environmental condition, the fixed point attractor is independent of the initial conditions of the genes and the corresponding proteins.

	Attractor 1		Attractor 2		Attractor 3		Attractor 4	
	Number of genes that agree	% agreement	Number of genes	% agreement	Number of genes	% agreement	Number of genes	% agreement
All (61)	49	80	32	52	49	80	35	57
Clean (34)	30	88	18	53	33	97	22	64
Low (27)	19	70	14	51	16	60	13	50

Table 4. The agreement between model output and experimental data for individual attractors. Attractors 1 to 4 are respectively from ECs 1 to 4 - aerobic exponential phase, aerobic stationary phase, anaerobic exponential phase and anaerobic stationary phase.

Caveats

- The results depend upon existing databases which are incomplete. Hence all the results are provisional. (Also it is possible that inherent biases in collecting the biochemical information result in a simple picture.)
- The discrete time, discrete state approach is an approximation.
- Model is deterministic whereas the real behaviour is stochastic.