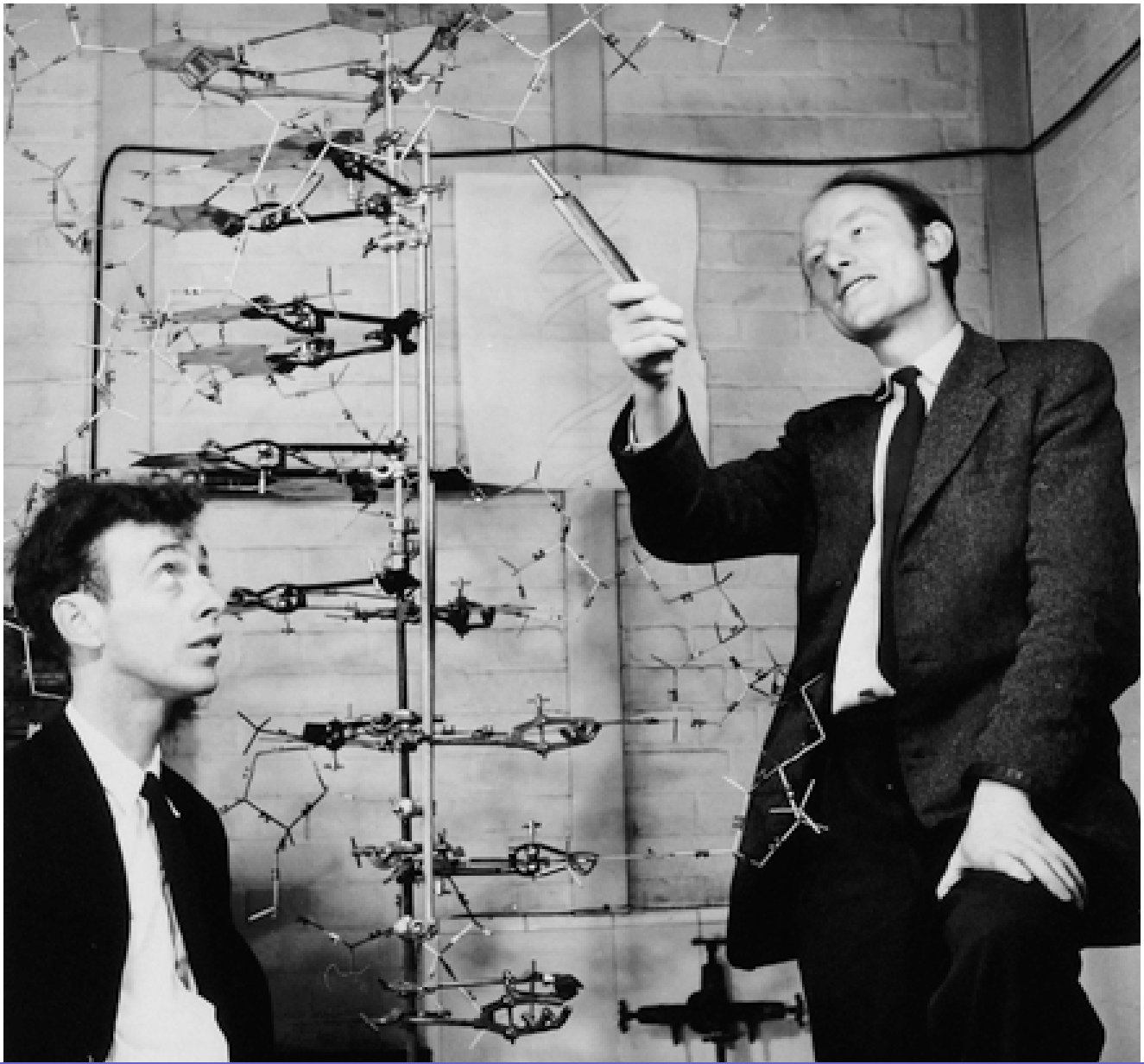
What is the Difference between a Human and a Chimp?

T. M. Murali

murali@cs.vt.edu
http://bioinformatics.cs.vt.edu/~murali

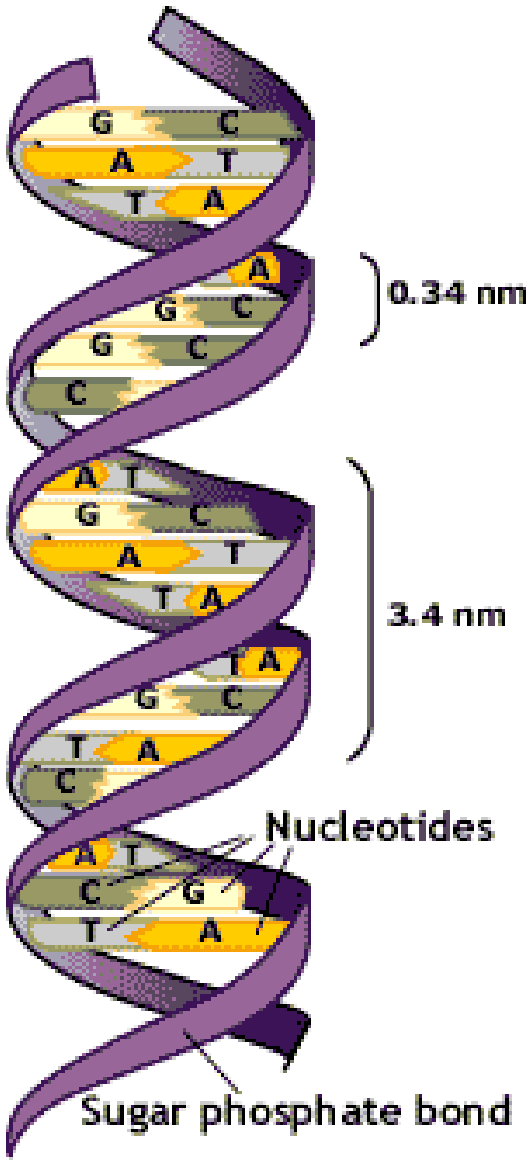
Introduction to Computational Biology and Bioinformatics (CS 3824)
March 25 and April 1, 2010

The Story Began in 1953



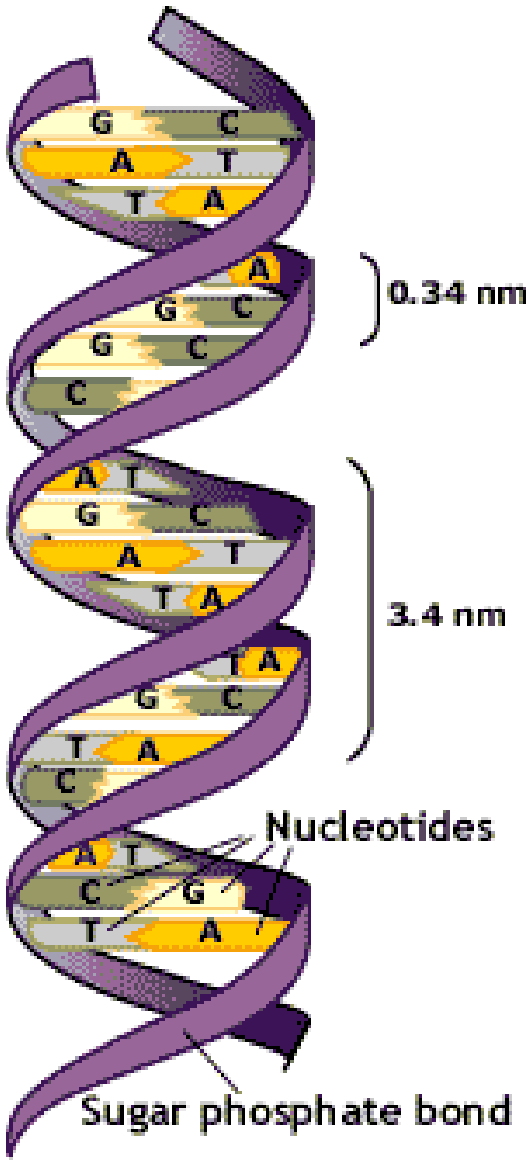
The Human Genome

- DNA is a (very long) string containing letters A, T, C, and G.
- Length of human genome is 3 billion base pairs



The Human Genome

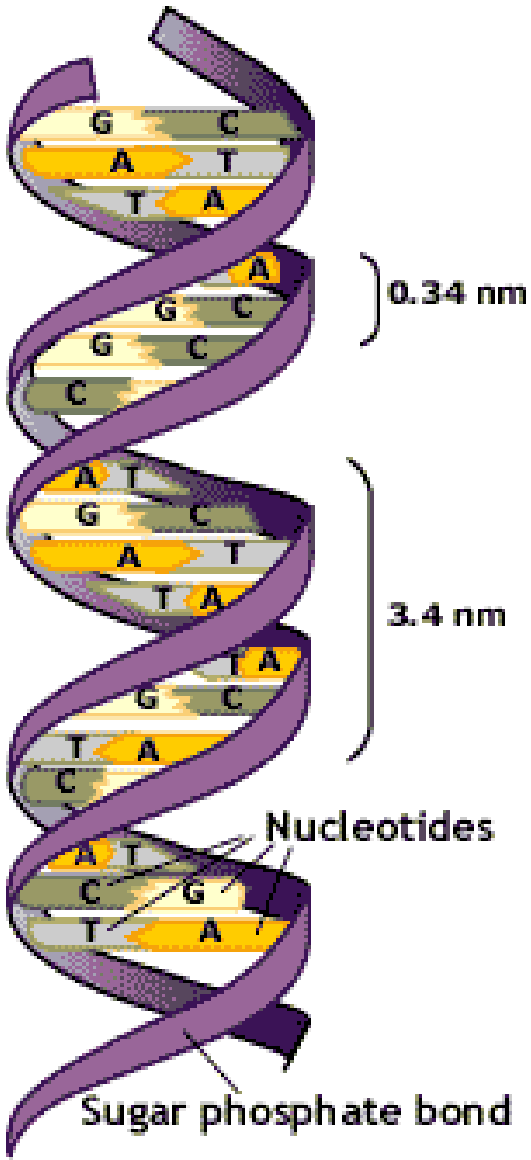
- DNA is a (very long) string containing letters A, T, C, and G.
- Length of human genome is 3 billion base pairs (1.8 metres).
- The Human Genome Project determined the spelling of the genome.



The Human Genome

- DNA is a (very long) string containing letters A, T, C, and G.
- Length of human genome is 3 billion base pairs (1.8 metres).
- The Human Genome Project determined the spelling of the genome.
- Eric Lander (Nano-Lecture, 2003 Ig Nobel Prize Ceremony):

  Genome. Bought the book, hard to read.



The Human Genome Project

The Human Genome Project

Before: human genome has about 100,000 genes.

The Human Genome Project

Before: human genome has about 100,000 genes.

After: human genome has about 30,000 genes.

Shock and Dismay

- The New York Times: **Genome Analysis Shows Humans Survive on Low Number of Genes** The two teams report that there are far fewer human genes than thought—probably a mere 30,000 or so—only a third more than those found in the roundworm. . . . The impact on human pride is another matter.
- Washington Post: It also raises new and difficult questions, such as how human beings—with all their passions and fears, their capacity for art, music, culture and war—can be all that they are with just 30,000 or so genes, only five times as many as in baker's yeast.

Genome size comparison

	Species	Chromosomes	Genes	Base pairs
	**Human** *(Homo sapiens)*	46 (23 pairs)	28-35,000	3.1 billion
	**Mouse** *(Mus musculus)*	40	22.5-30,000	2.7 billion
	**Puffer fish** *(Fugu rubripes)*	44	31,000	365 million
	**Malaria mosquito** *(Anopheles gambiae)*	6	14,000	289 million
	**Fruit fly** *(Drosophila melanogaster)*	8	14,000	137 million
	**Roundworm** *(C. elegans)*	12	19,000	97 million
	**Bacterium** * *(E. coli)*	1	5,000	4.1 million

*Bacterial chromosomes are chromonemes, not true chromosomes

JOHN BLANCHARD / *The Chronicle*

What Makes a Human Different from a Chimp?

What Makes a Human Different from a Chimp?

What Makes a Human Different from a Chimp?

What Makes a Human Different from a Chimp?

Chimp and chump genomes are only about 1.2% different!

Molecular Biology

- Genomes provide the parts lists (e.g., genes and proteins) but do not directly tell us how these parts fit.

Molecular Biology → Systems Biology

- Genomes provide the parts lists (e.g., genes and proteins) but do not directly tell us how these parts fit.
- We need to understand how genes, proteins, and other molecules interact with other in different cell states, different tissues, and under different external conditions.
- Study only of individual elements is unlikely to reveal higher-order organisation of cellular interaction networks.

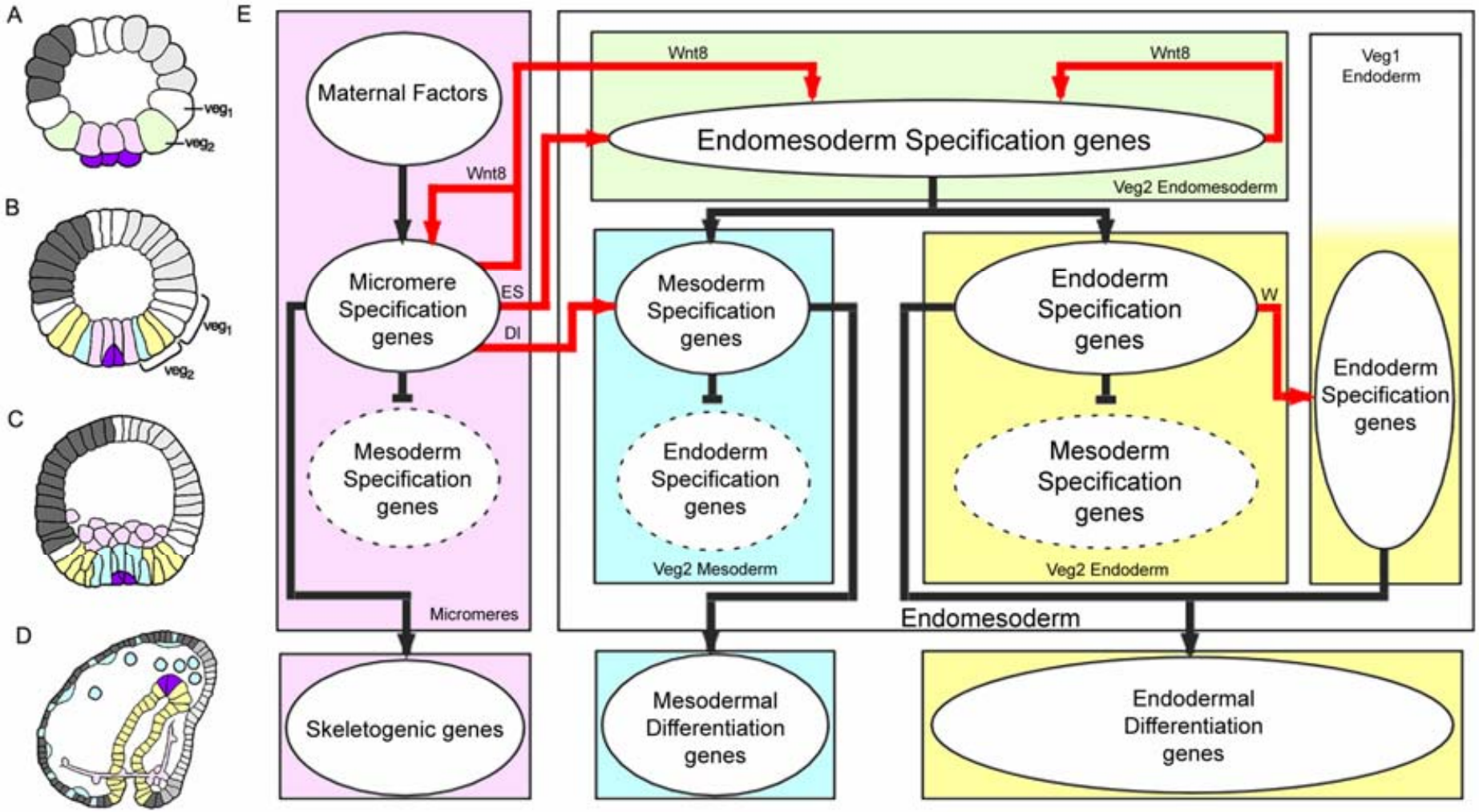
Sea Urchin (Strongylocentrotus purpuratus)

Sea Urchin (Strongylocentrotus purpuratus)

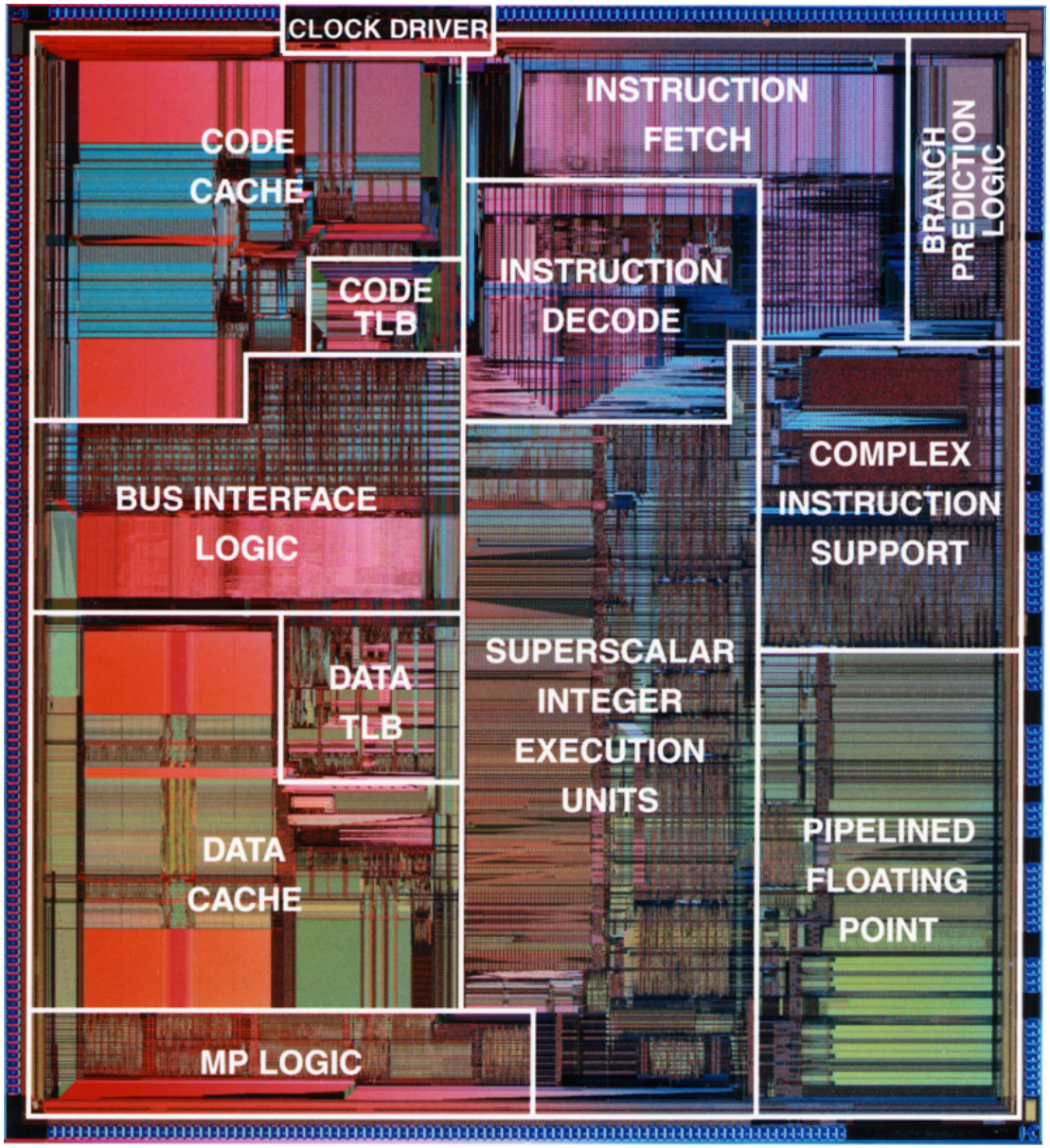
- Very important in developmental biology.
- Many principles of embryo development were discovered in the sea urchin.

A Cell

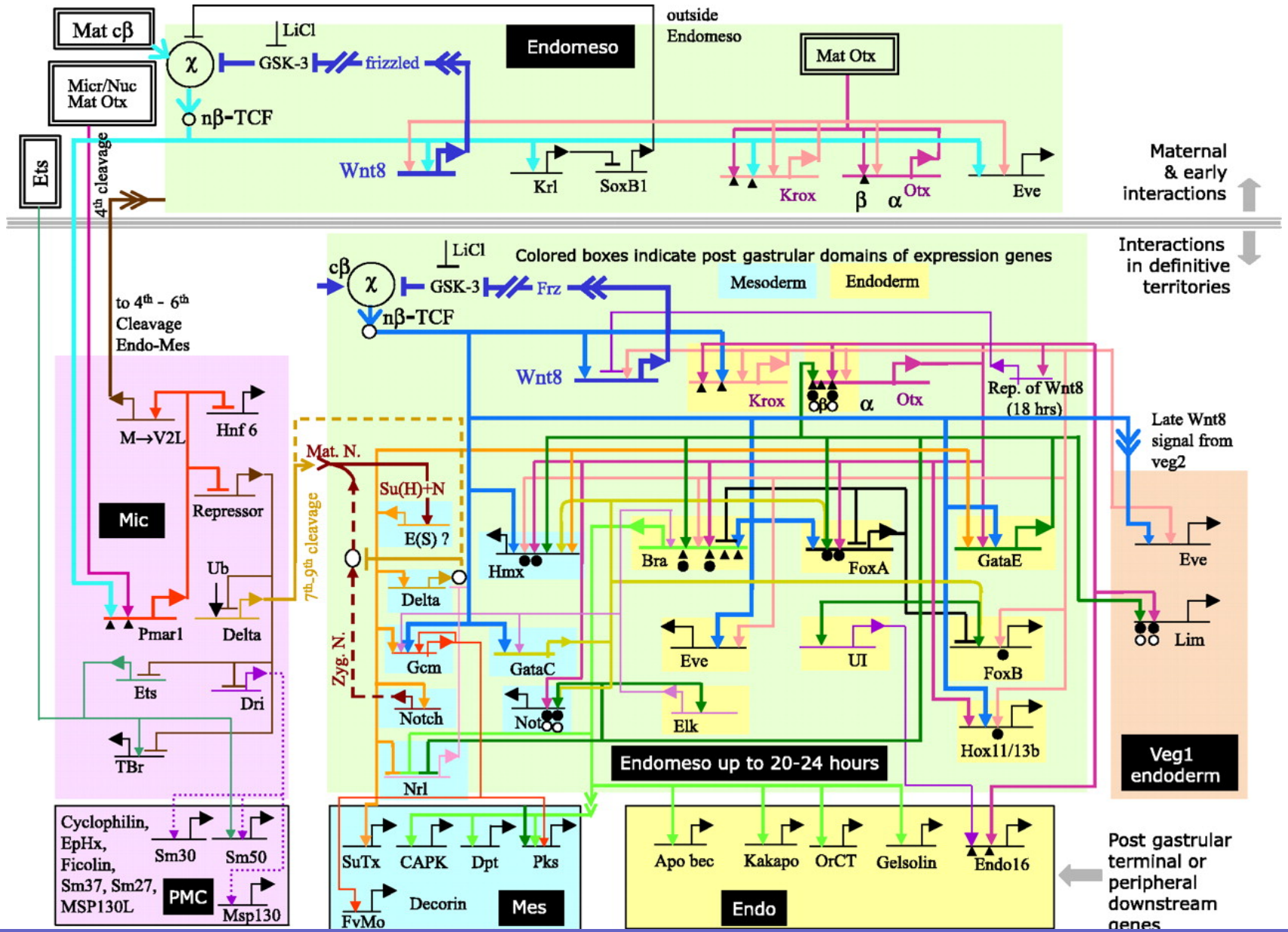
A Cell is a Modular



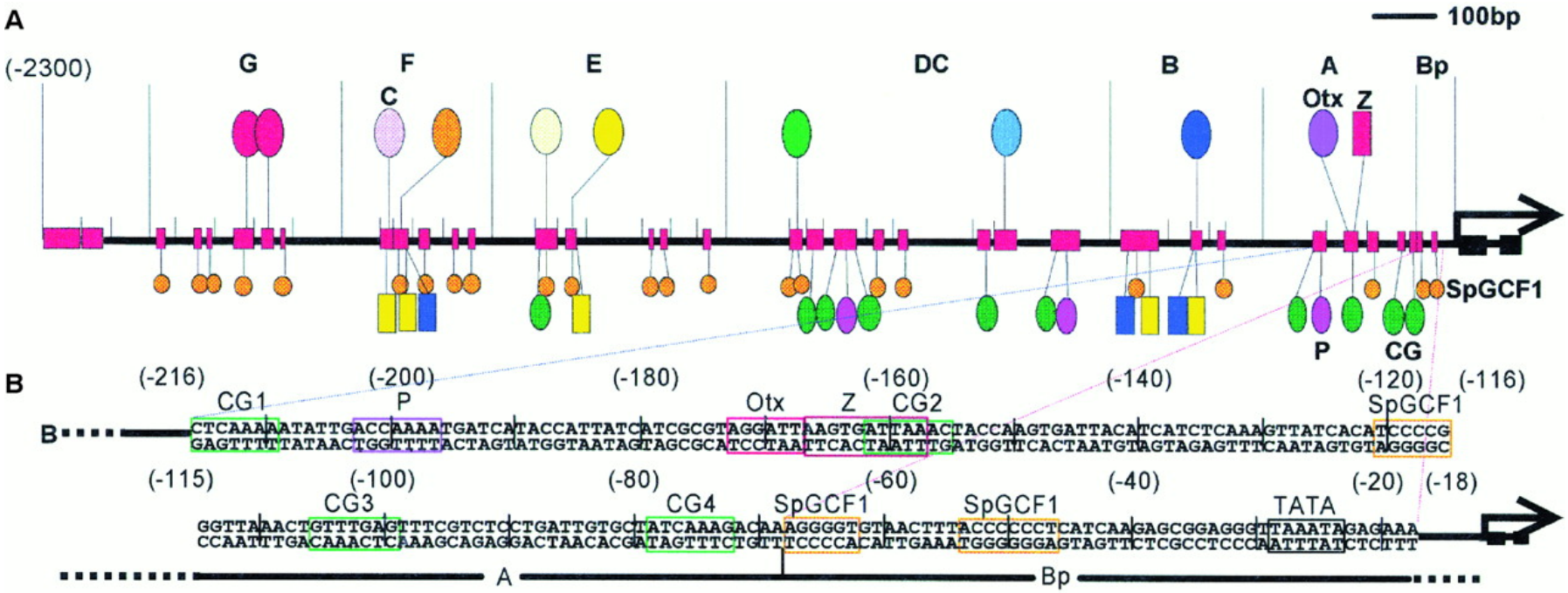
A Cell is a Modular



A Cell is a Modular Network



A Cell is a Modular Network



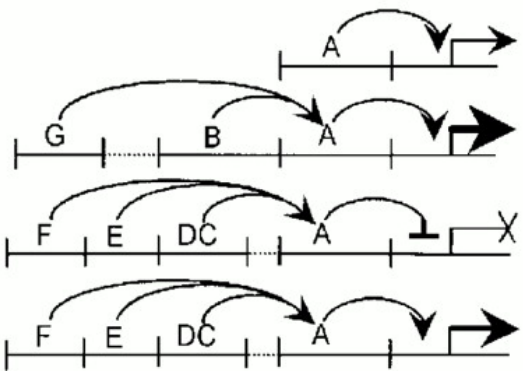
C Module A functions:

Vegetal plate expression in early development:

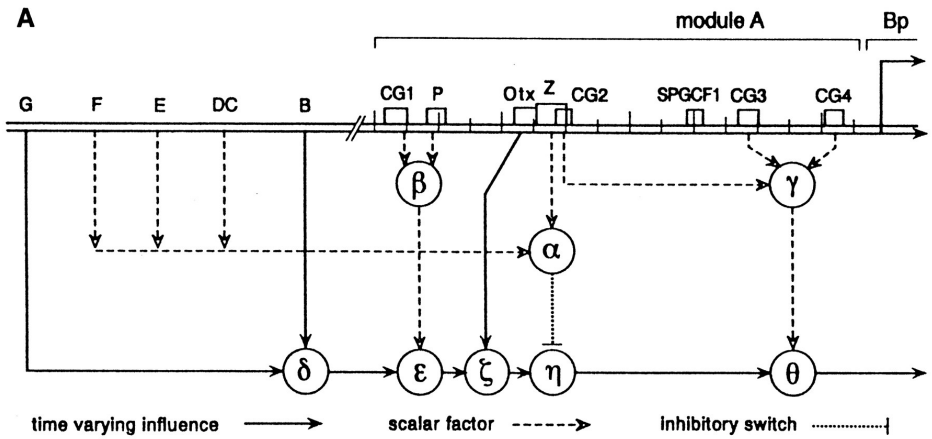
Synergism with modules B and G enhancing endoderm expression in later development:

Repression in ectoderm (modules E and F) and skeletogenic mesenchyme (module DC):

Modules E, F and DC with LiCl treatment:



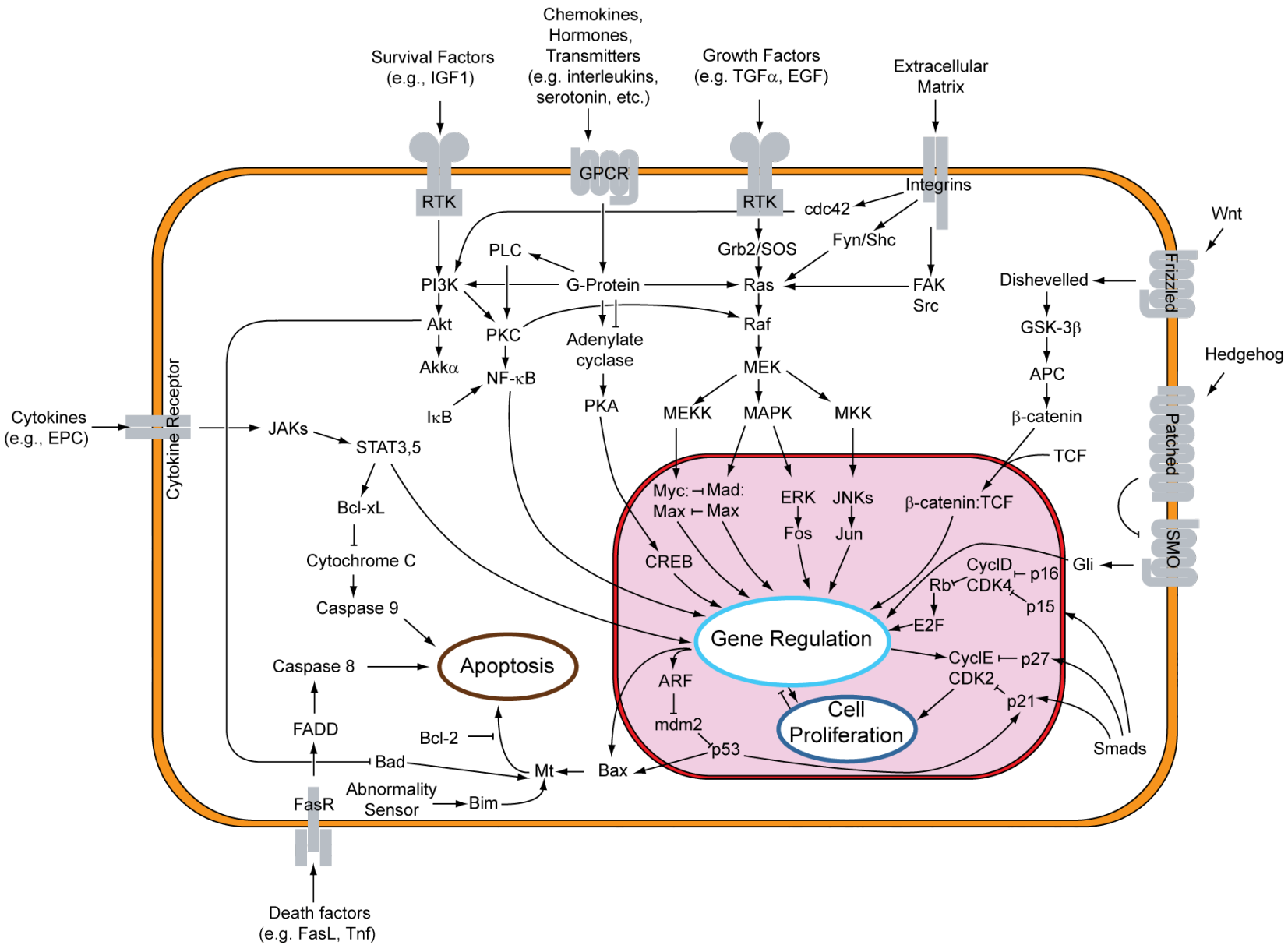
A Cell is a Modular Network that Computes



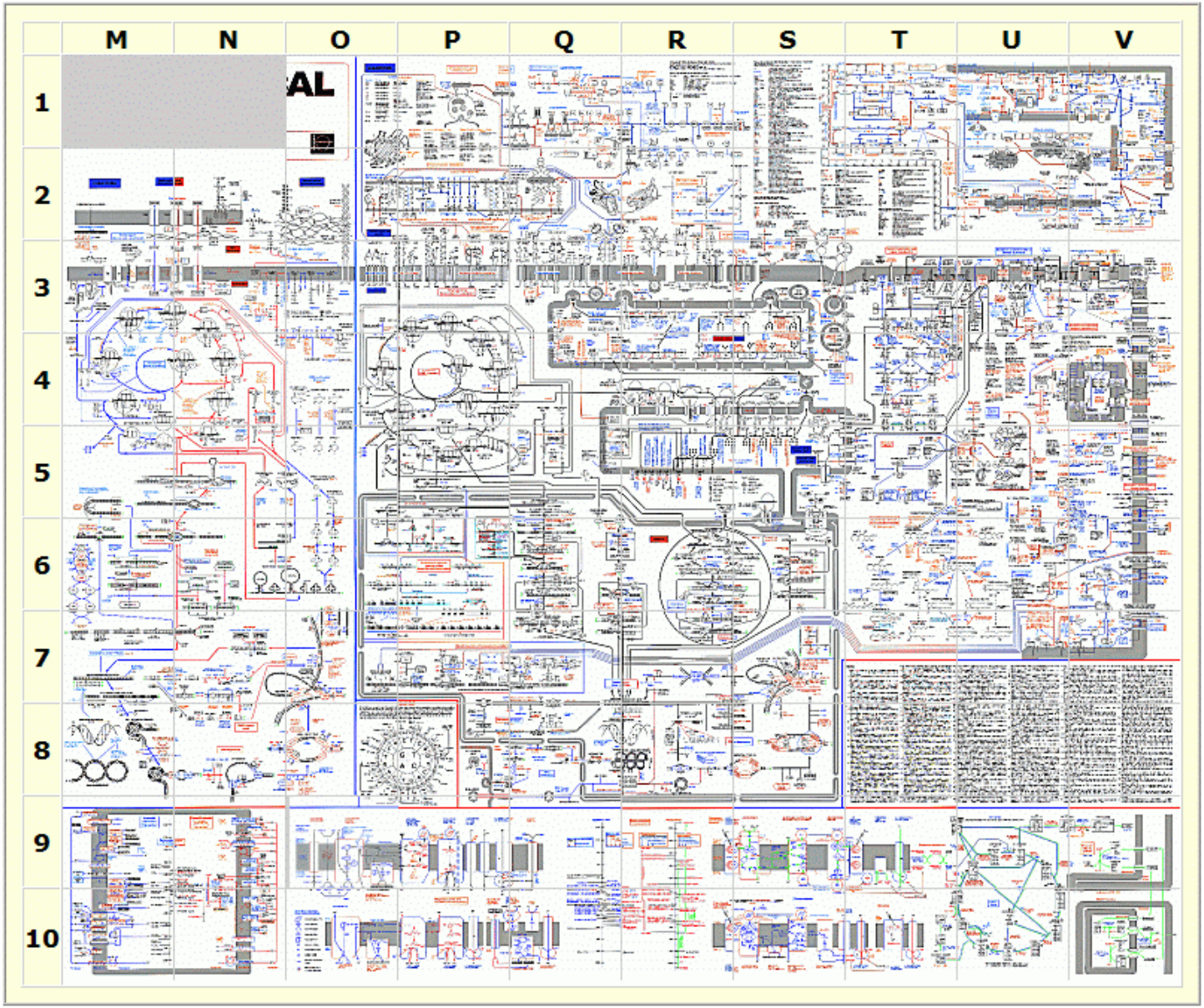
B

if (F = 1 or E = 1 or CD = 1) and (Z = 1)	Repression functions of modules F, E, and DC mediated by Z site
$\alpha = 1$	
else $\alpha = 0$	
if (P = 1 and CG_1 = 1)	Both P and CG_1 needed for synergistic link with module B
$\beta = 2$	
else $\beta = 0$	
if (CG_2 = 1 and CG_3 = 1 and CG_4 = 1)	Final step up of system output
$\gamma = 2$	
else $\gamma = 1$	
$\delta(t) = B(t) + G(t)$	Positive input from modules B and G
$\varepsilon(t) = \beta^*\delta(t)$	Synergistic amplification of module B output by CG_1-P subsystem
if ($\varepsilon(t) = 0$)	Switch determining whether Otx site in module A, or upstream modules (i.e., mainly module B), will control level of activity
$\xi(t) = Otx(t)$	
else $\xi(t) = \varepsilon(t)$	
if ($\alpha = 1$)	Repression function inoperative in endoderm but blocks activity elsewhere
$\eta(t) = 0$	
else $\eta(t) = \xi(t)$	
$\Theta(t) = \gamma^*\eta(t)$	Final output communicated to BTA

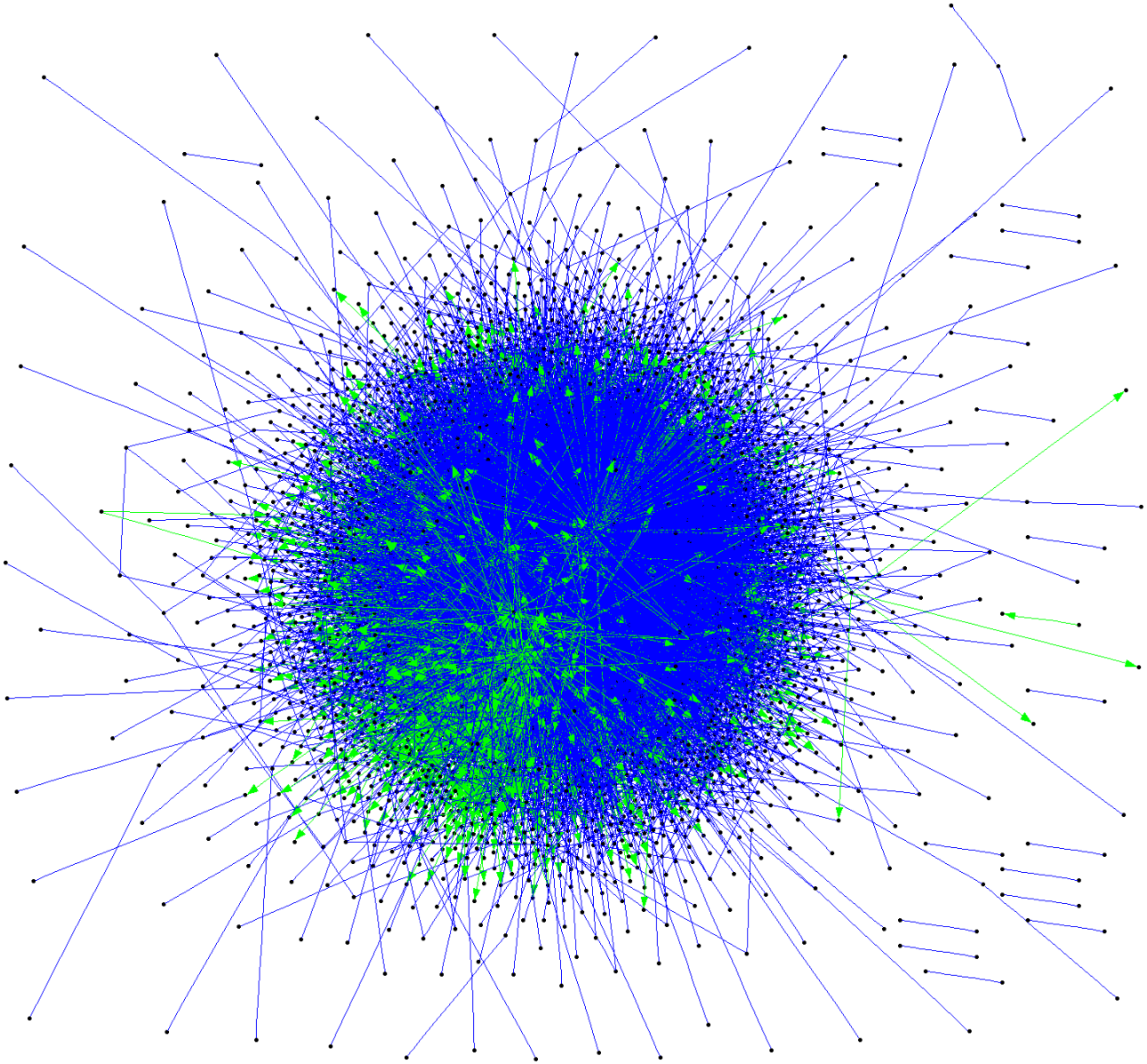
Network is Complex



Network is Complex



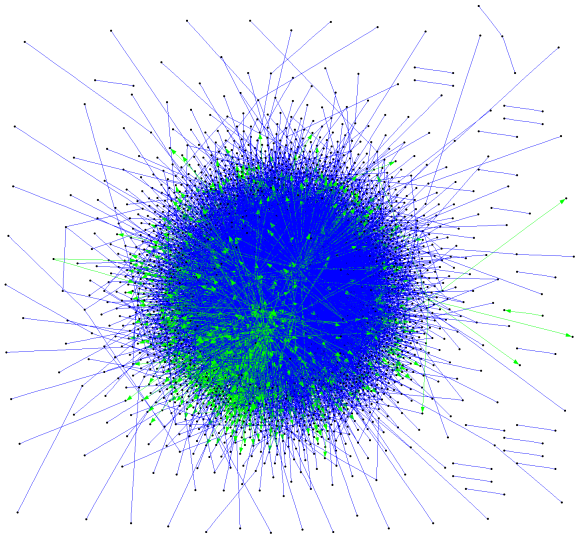
Network is Complex but Very Poorly Understood

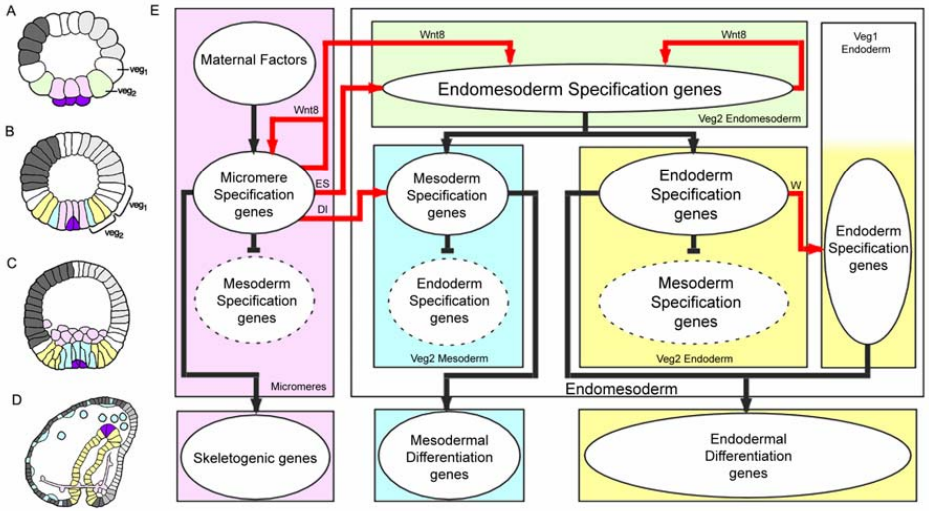


Challenges with Molecular Interaction Networks

- Biological data sets and networks are large.
- They are intricate and of very diverse types.
- They are noisy: experiments are error-prone.
- They are highly incomplete. We barely know which genes interact, let alone the detailed kinetics of each interaction.

My Research: Understanding Interaction Networks





- Automatically find modules of coherently acting molecules from data.
- Build predictive module-level models of the cell.
- Emphasise a phenomenological approach to systems biology.
- Develop techniques in graph and discrete algorithms, data mining, and machine learning and apply them to solve specific biological questions.

Research Applications

- Predict the functions that genes and proteins perform in the cell.
- Develop drugs that may be effective against multiple pathogens.
- Build models of how cells communicate with each other.
- Zero in on the molecules and interactions that are active in cancer.
- Develop biologically-relevant representations of molecular interaction networks.

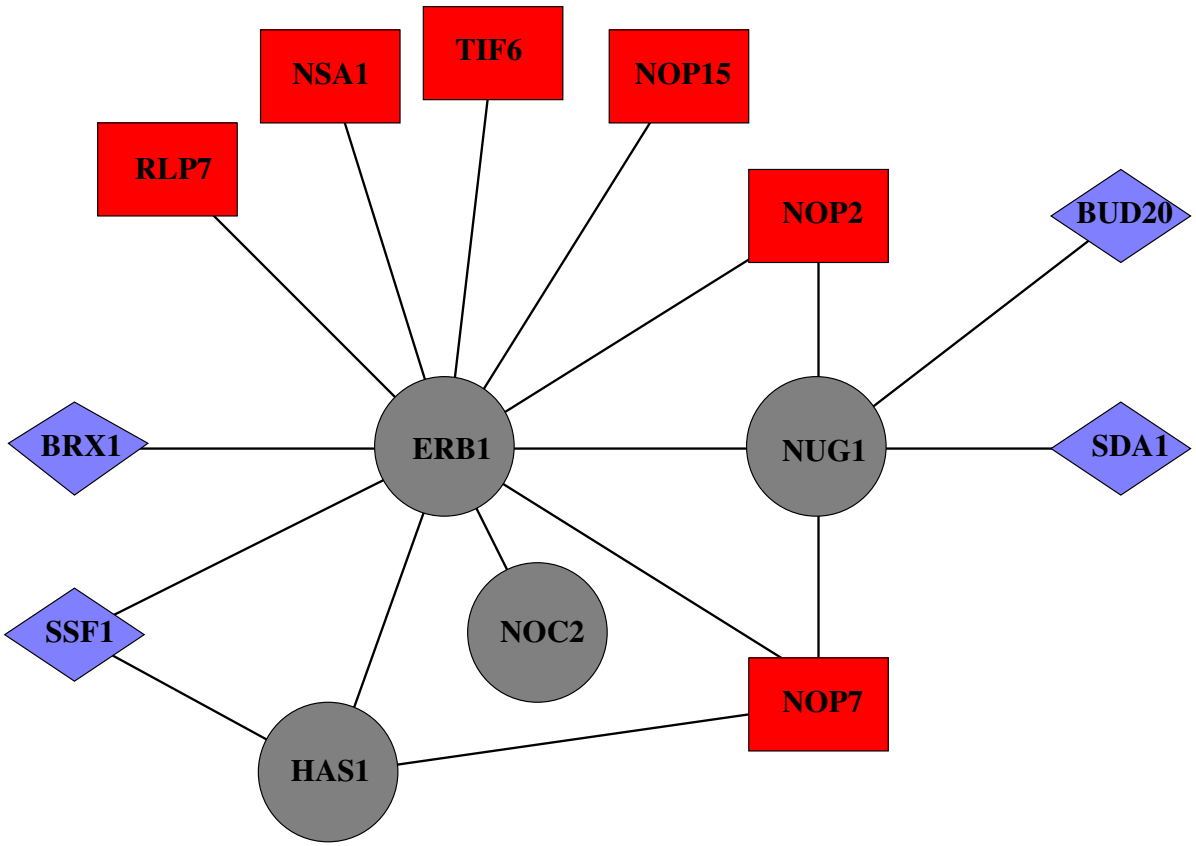
Functions of Many Genes are Unknown

- We have the sequences of 100s of genomes.
- We know the locations of genes in these genomes.
- Functions of over 50% of the genes are unknown!

Functions of Many Genes are Unknown

- We have the sequences of 100s of genomes.
- We know the locations of genes in these genomes.
- Functions of over 50% of the genes are unknown!
- Genes with similar sequences in different organisms are likely to have the same function.
- We need techniques for function prediction that go beyond sequence similarity.

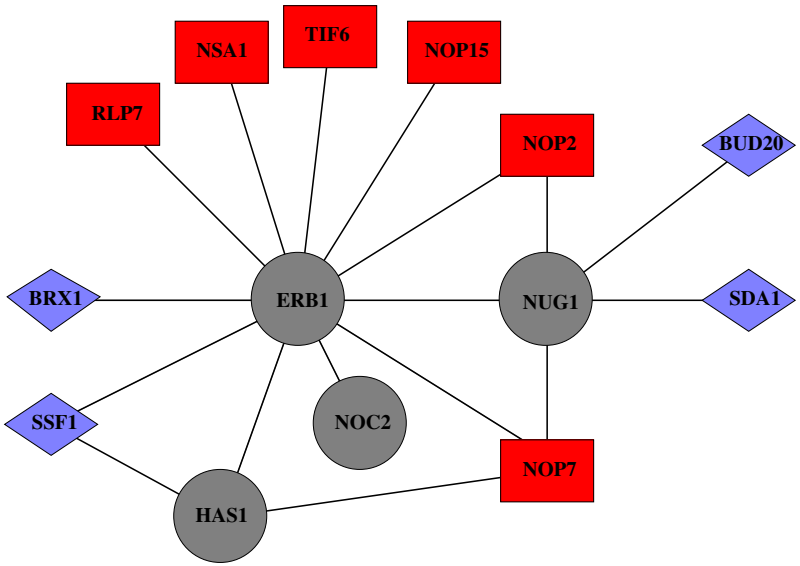
Exploit Social Network of Genes



- Exploit network structure to determine whether gray genes have the same function as the red genes or the blue genes.

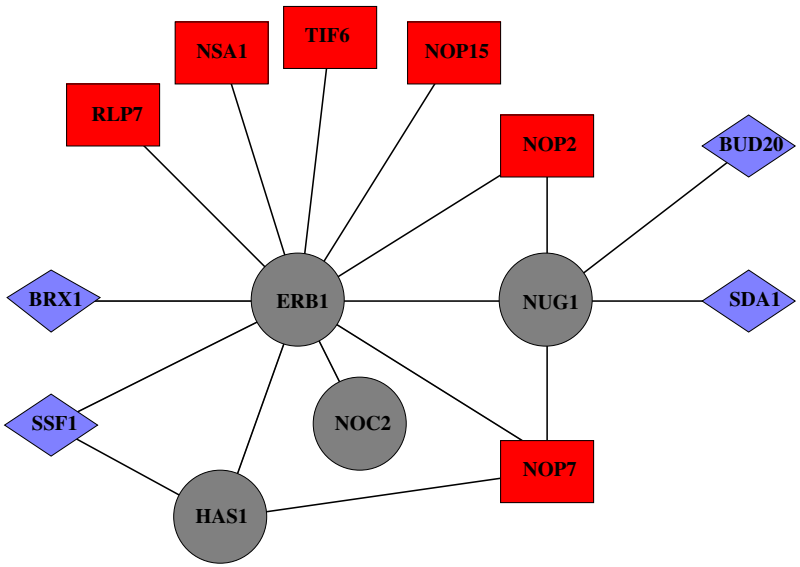
(Karaoz, Murali, Letovsky, Zheng, Ding, Cantor and Kasif, *PNAS*, 2004; Murali, Wu, and Kasif, *Nature Biotech.*, 2006.)

Node States



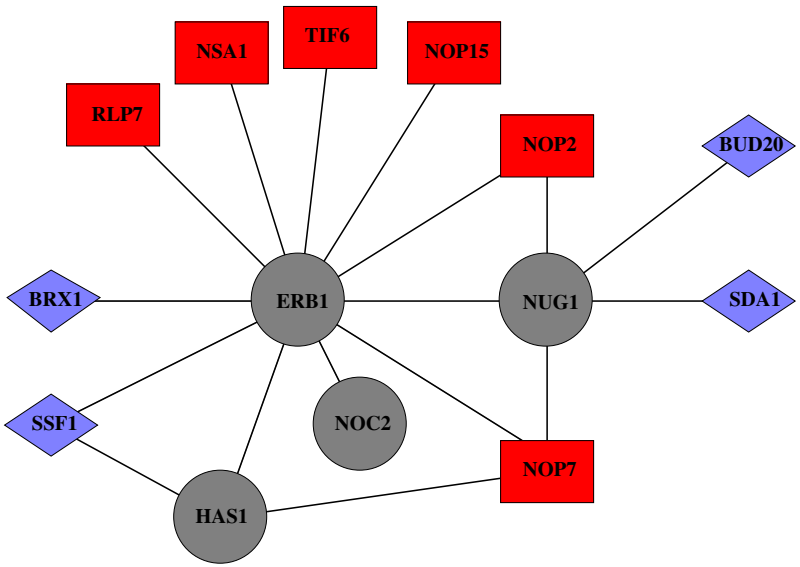
- Interaction network is a graph $G = (V, E)$.
- Each node i has an associated state s_i:
  - $s_i = 1$: gene i is annotated with current function.
  - $s_i = -1$: gene i is annotated with another function.
  - $s_i = 0$: otherwise.
- An edge between nodes i and j has a weight w_{ij}.

Maximally-Consistent Assignments



- An edge is *consistent* if it is incident on nodes with the same state.
- *Maximally-consistent assignment*: number of consistent edges is maximised.

Maximally-Consistent Assignments



- An edge is *consistent* if it is incident on nodes with the same state.
- *Maximally-consistent assignment*: number of consistent edges is maximised.

Computational goal: Assign state of -1 or $+1$ to nodes with initial state 0 to achieve maximal consistency by maximising

$$\sum_{(u,v) \text{ is an edge}} w_{uv} s_u s_v$$

Local Update Rule

- Activation rule is

$$s_i = \mathrm{sgn}\left(\sum_{j \in N_i} w_{ij} s_j\right),$$

where N_i = neighbours of node i.

Local Update Rule

- Activation rule is

$$s_i = \mathrm{sgn}\left(\sum_{j \in N_i} w_{ij} s_j\right),$$

  where N_i = neighbours of node i.
- Applying this rule:

Local Update Rule

- Activation rule is

$$s_i = \mathrm{sgn}\left(\sum_{j \in N_i} w_{ij} s_j\right),$$

  where N_i = neighbours of node i.
- Applying this rule:
  - Parallel update: each node updates itself in parallel with the other nodes.
  - Serial update: go through each node in sequence.

Local Update Rule

- Activation rule is

$$s_i = \operatorname{sgn}\left(\sum_{j \in N_i} w_{ij} s_j\right),$$

  where N_i = neighbours of node i.
- Applying this rule:
  - Parallel update: each node updates itself in parallel with the other nodes.
  - Serial update: go through each node in sequence.
- Stopping criterion: converge when no node's state changes.

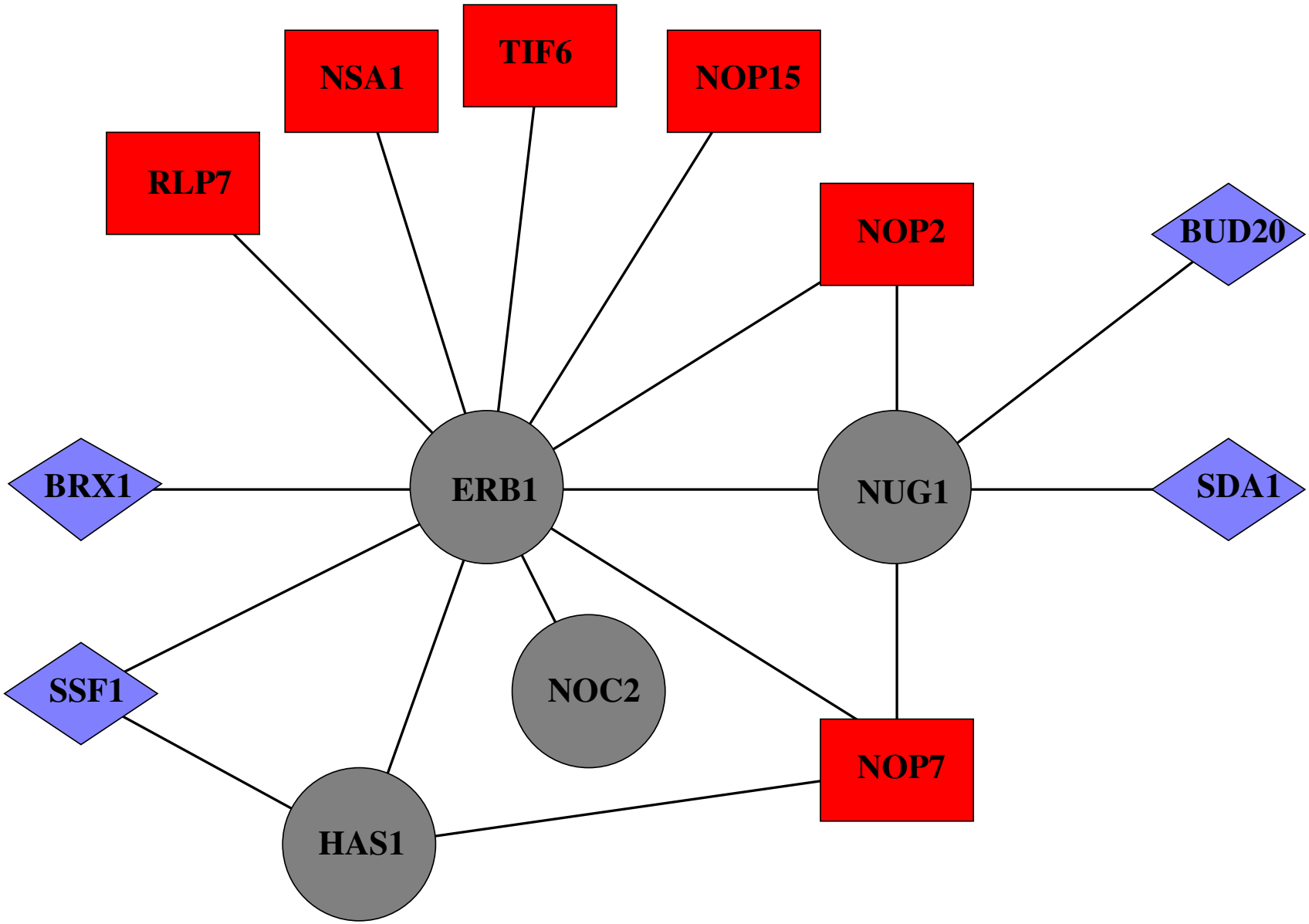
Local Update Rule

- Activation rule is

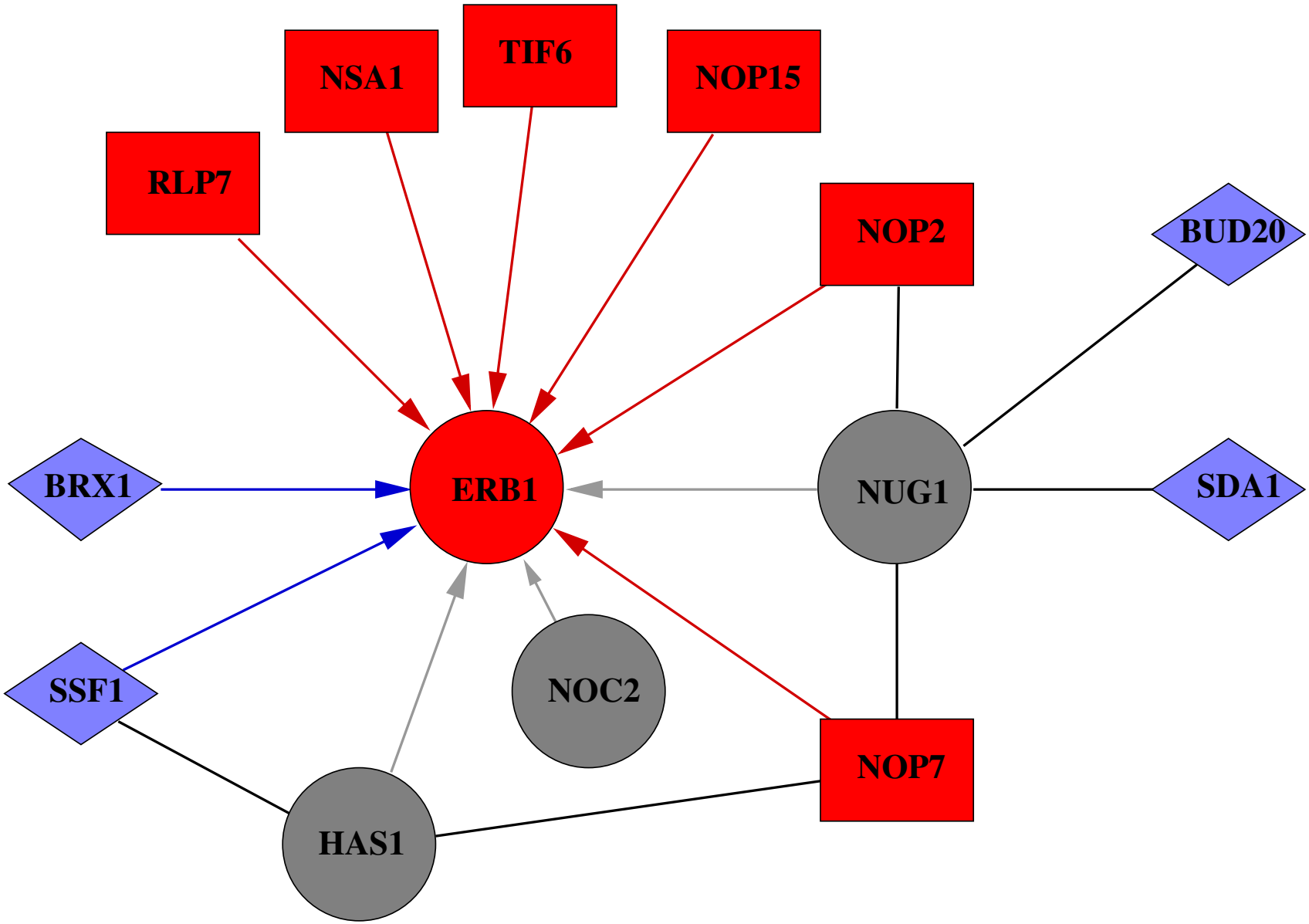
$$s_i = \operatorname{sgn}\left(\sum_{j \in N_i} w_{ij} s_j\right),$$

  where N_i = neighbours of node i.
- Applying this rule:
  - Parallel update: each node updates itself in parallel with the other nodes.
  - Serial update: go through each node in sequence.
- Stopping criterion: converge when no node's state changes.
- Proof of convergence: show $(\sum_{(u,v) \text{ is an edge}} w_{uv} s_u s_v)$ always increases when you change a node's state.

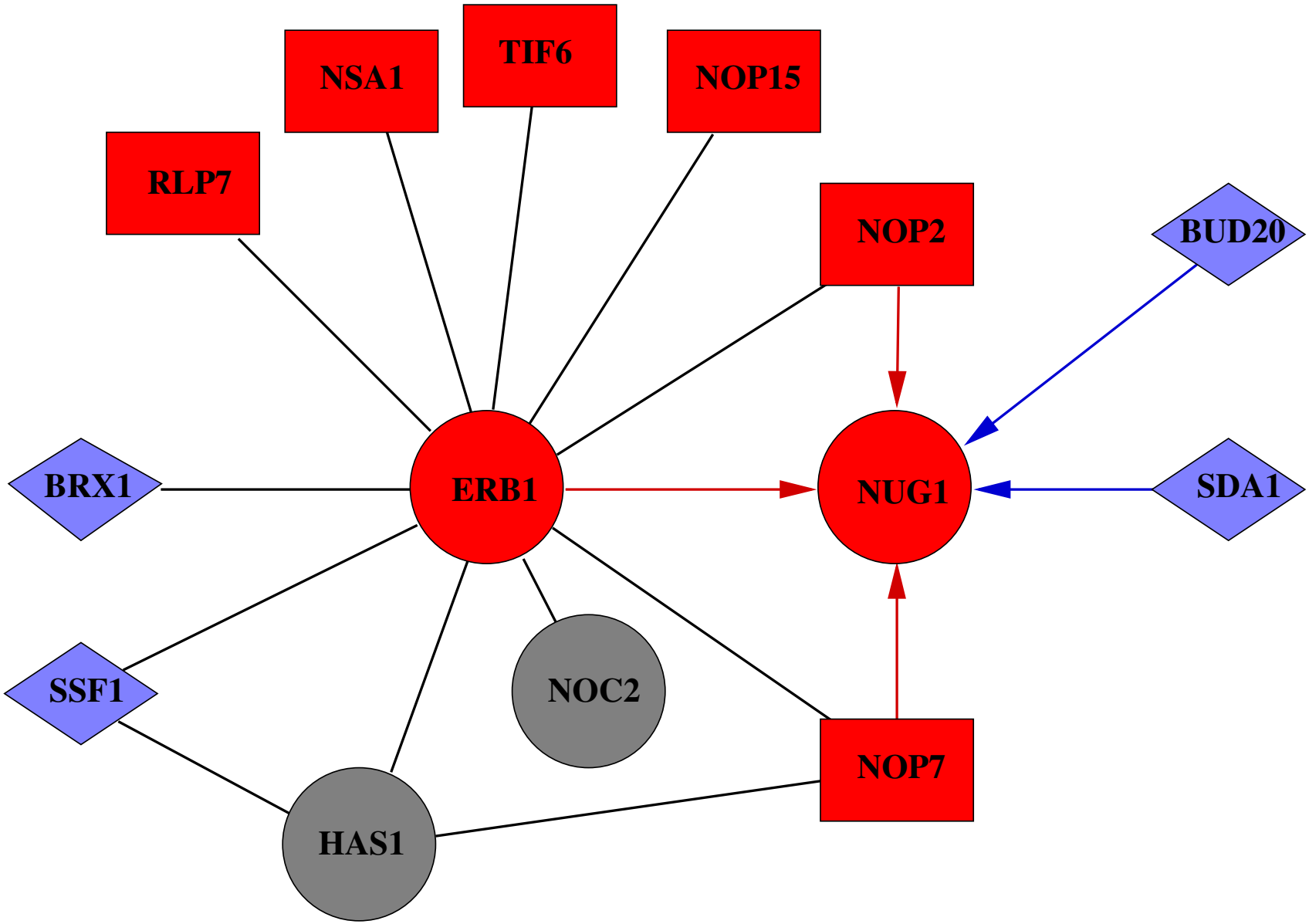
Example of Local Updates



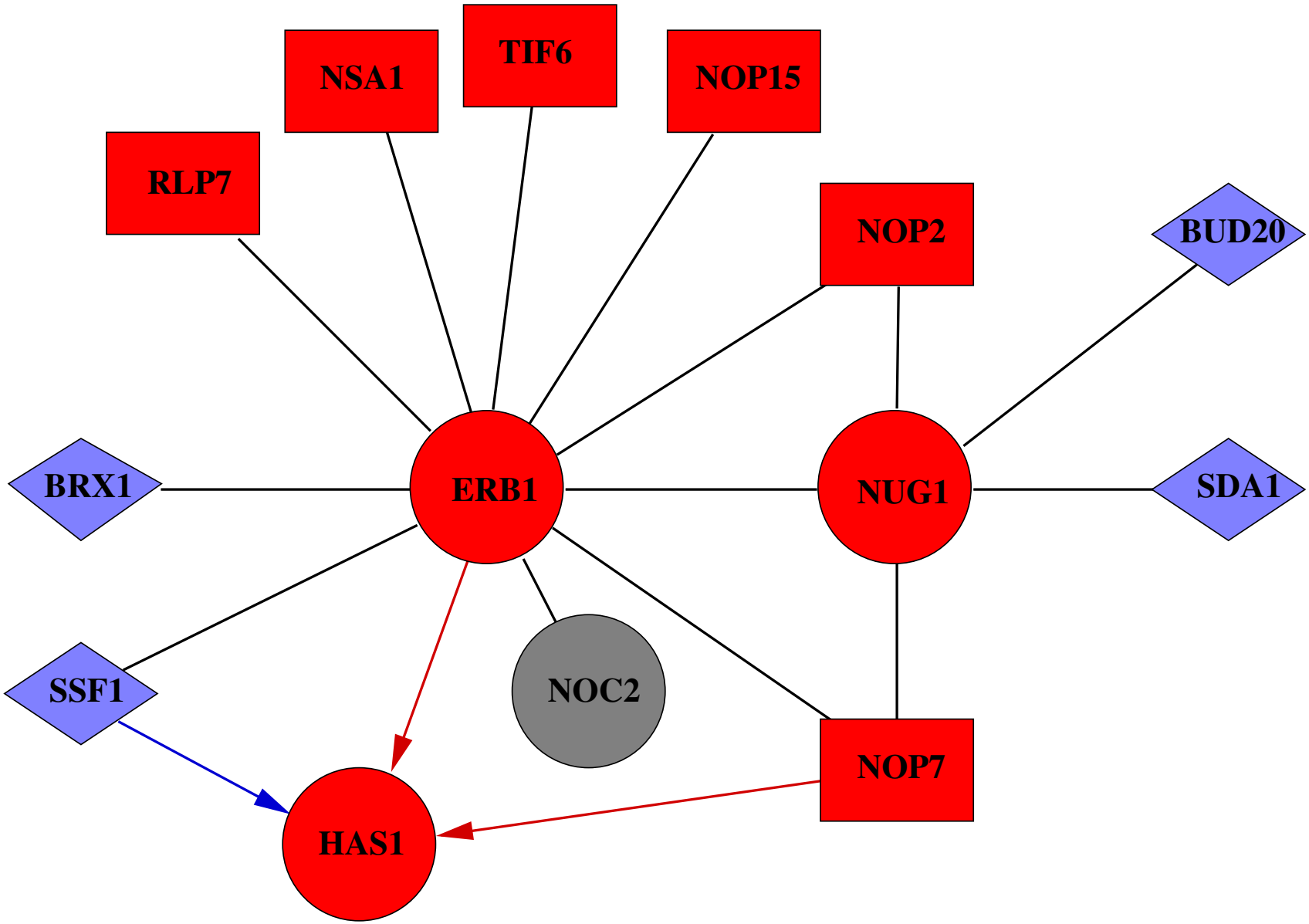
Example of Local Updates



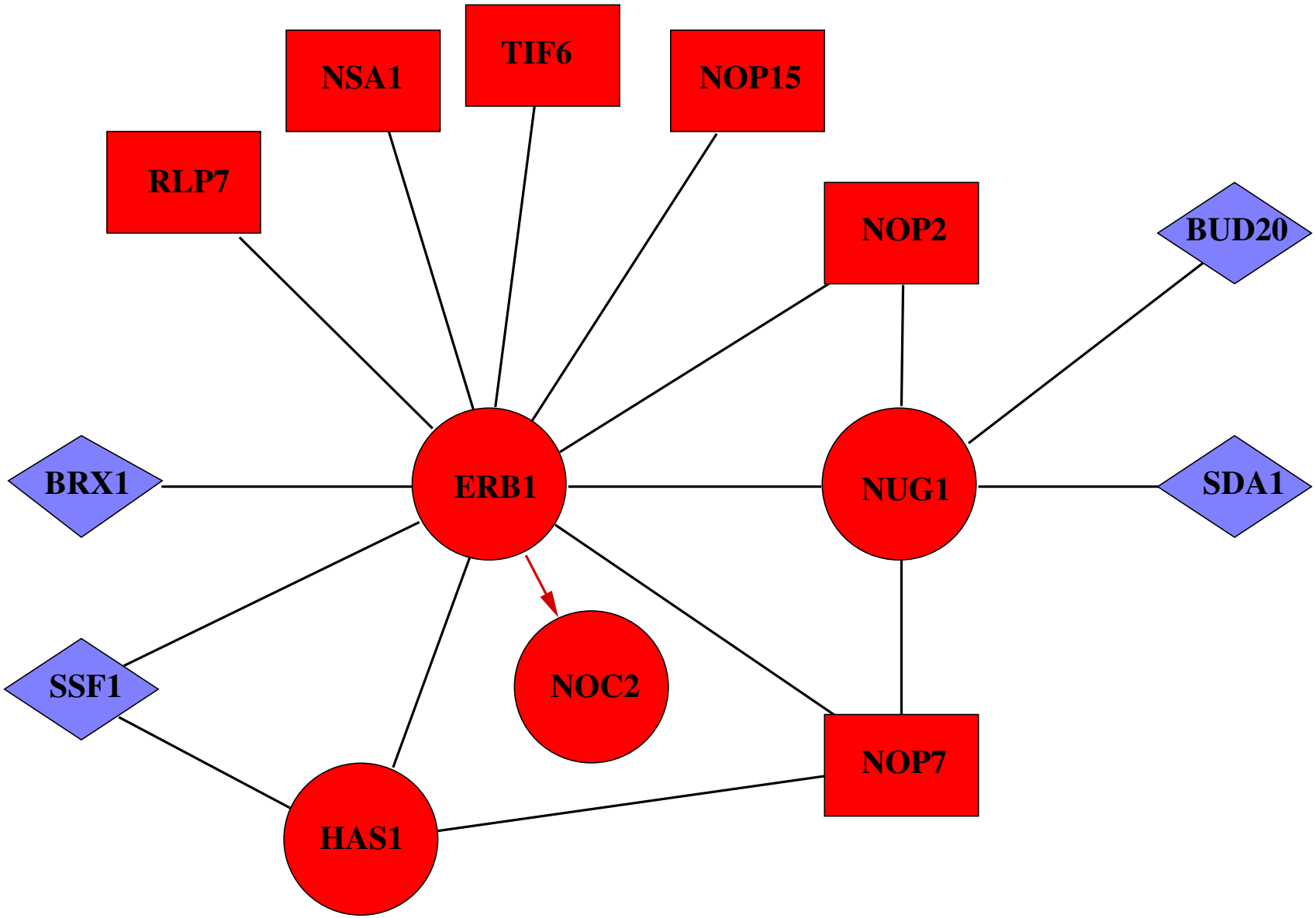
Example of Local Updates



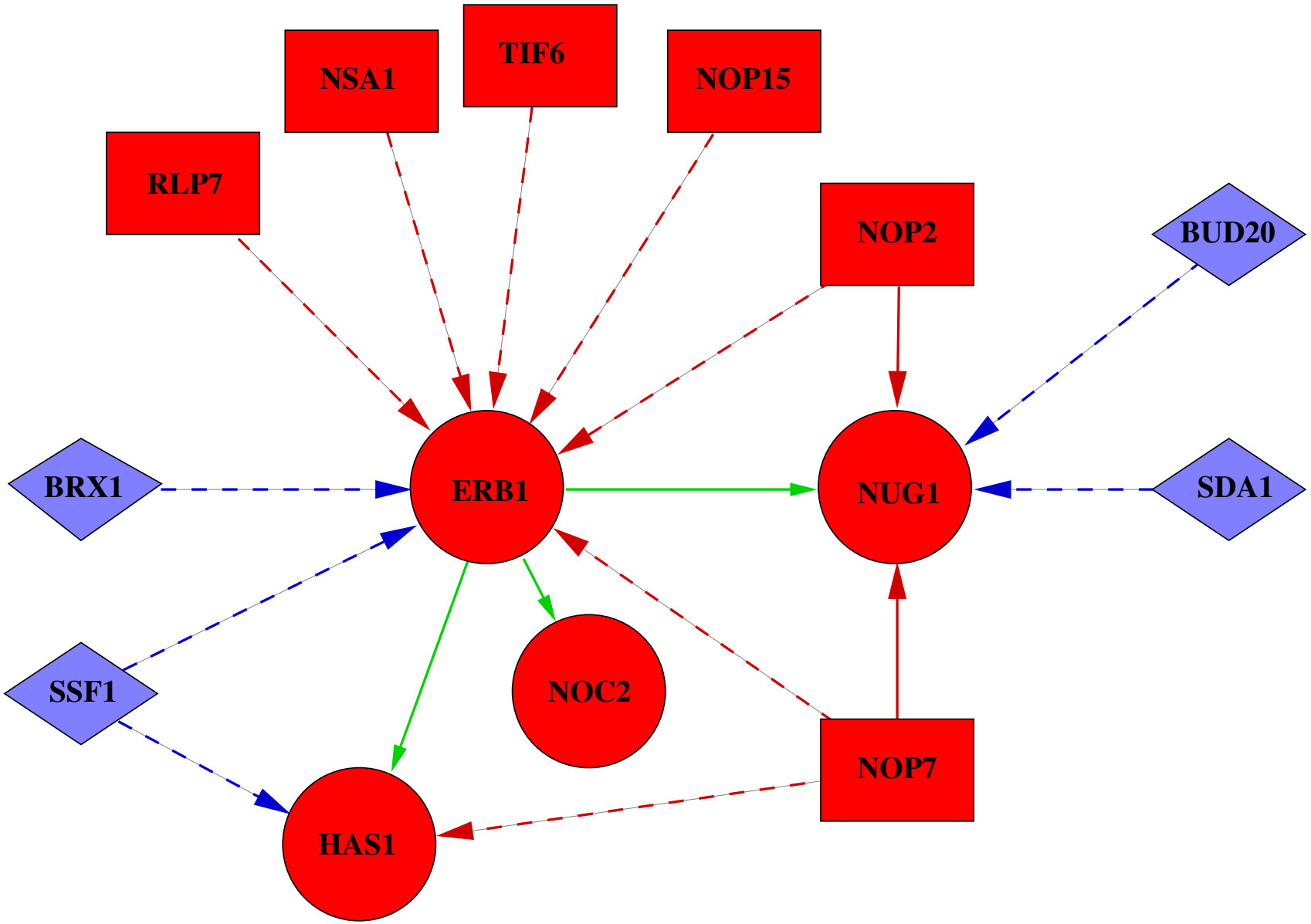
Example of Local Updates



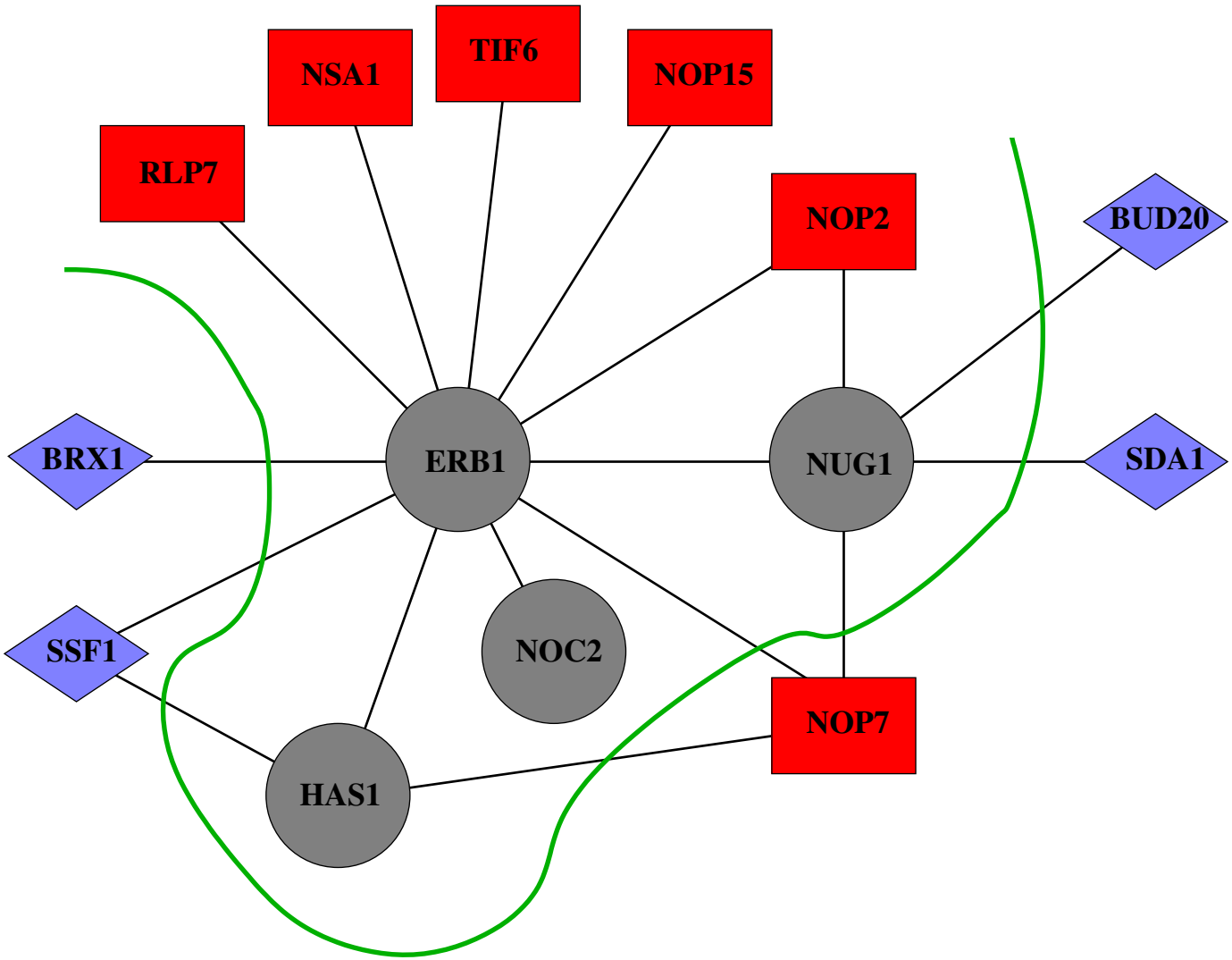
Example of Local Updates



Example of Local Updates

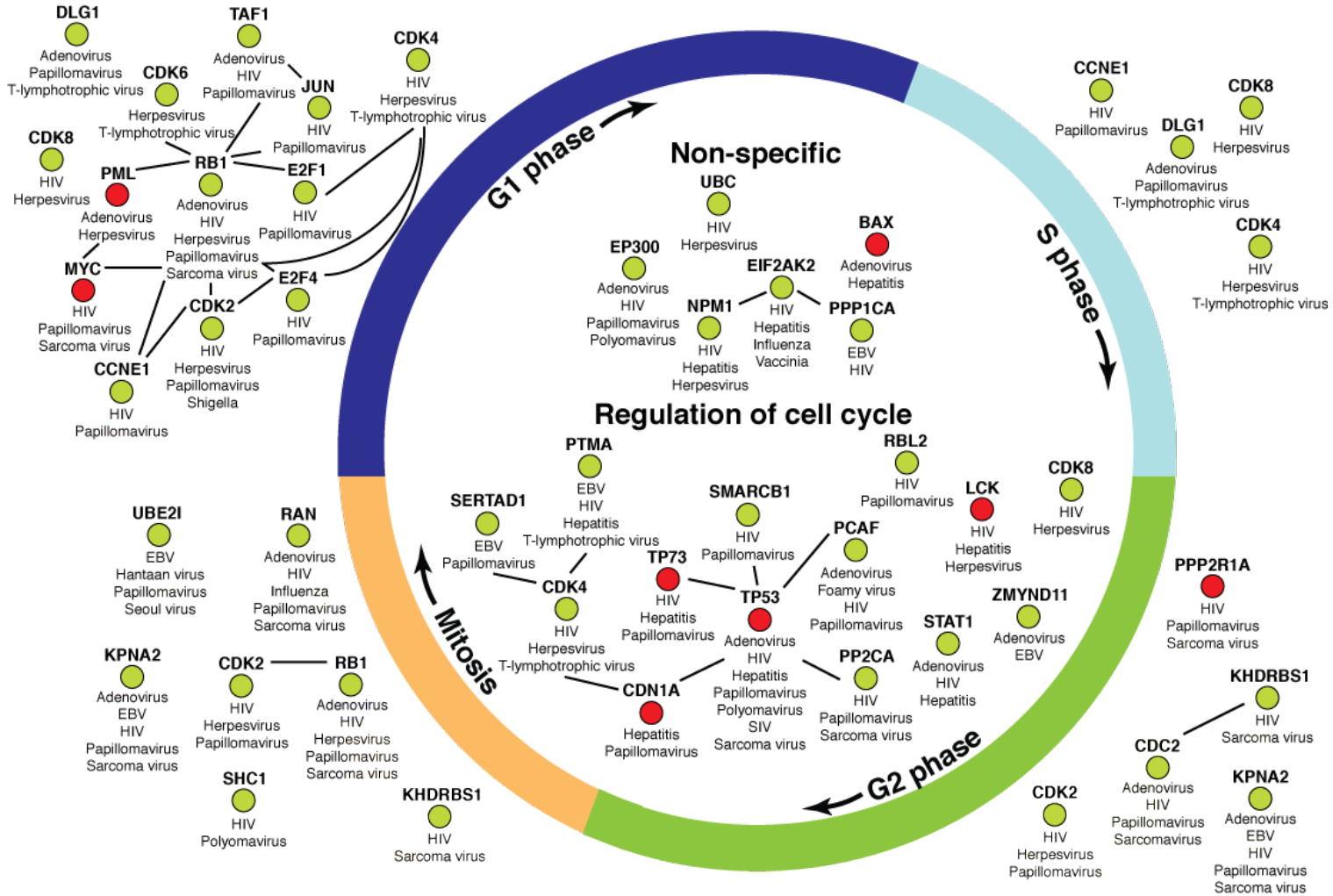


Graph cuts



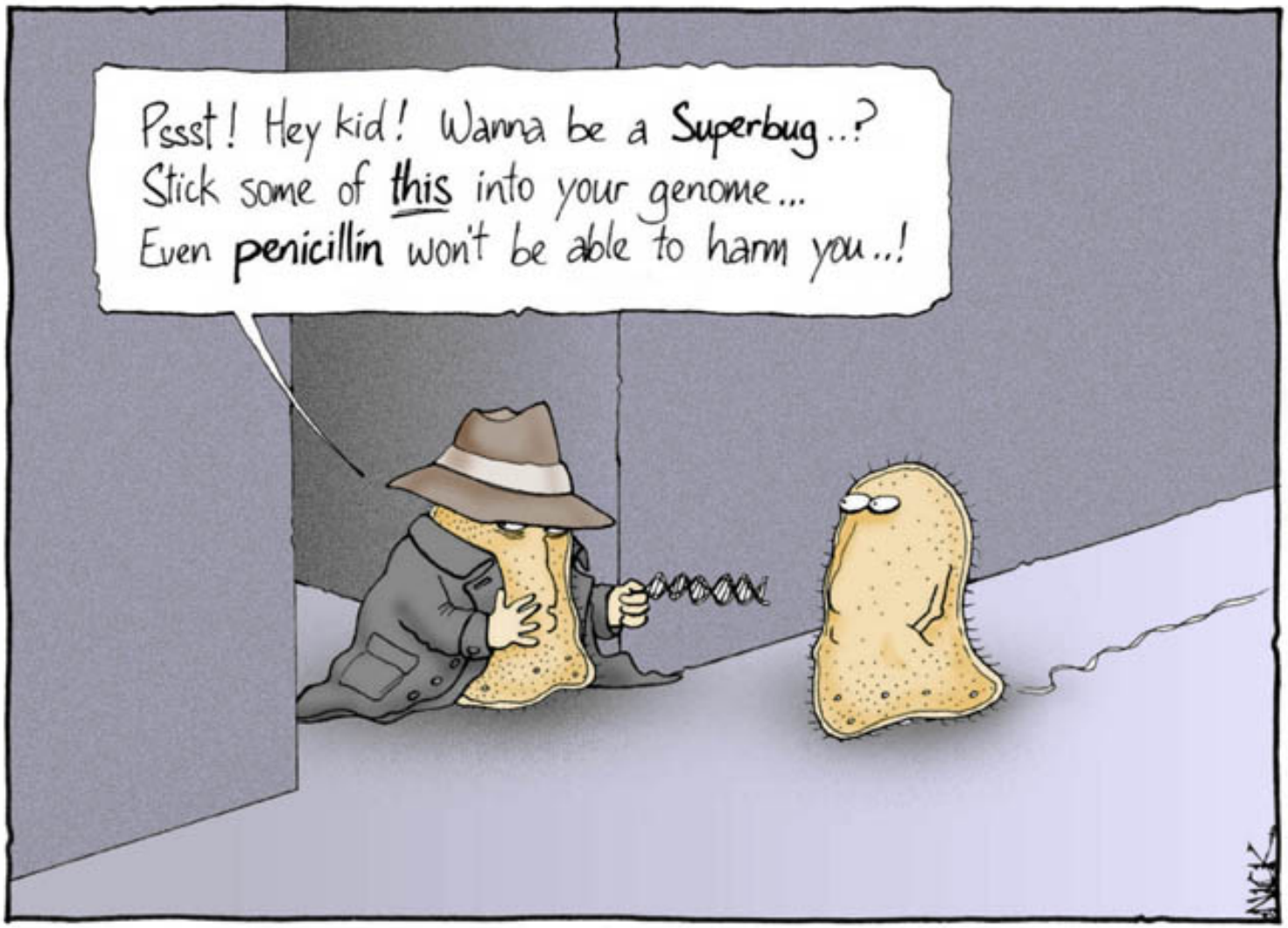
- Can also use graph cuts to solve the problem.

Human Proteins Interacting with Pathogens



Dyer, Murali, and Sobral, *PLoS Pathog*, 2008.

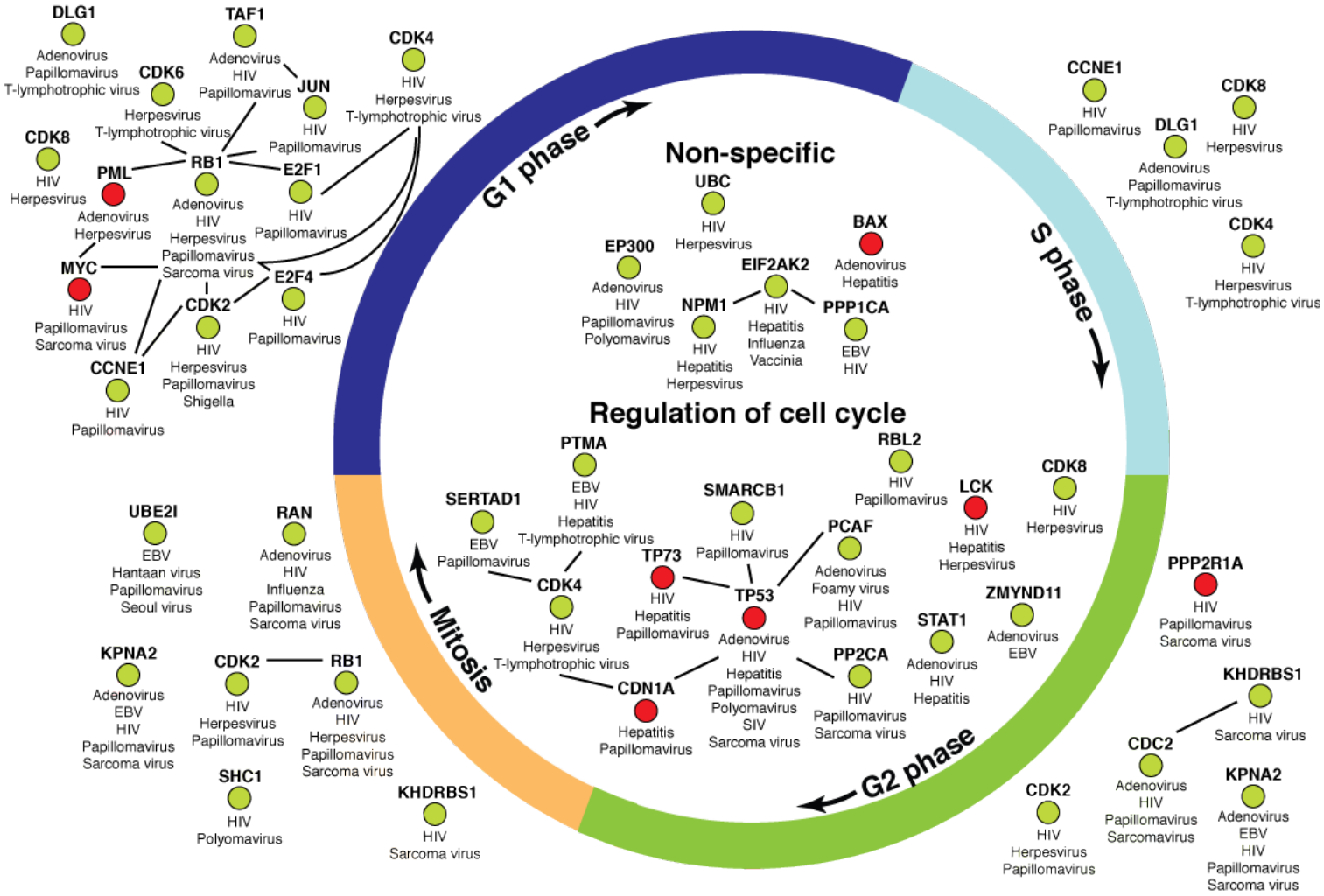
Pathogens are Becoming Drug-Resistant



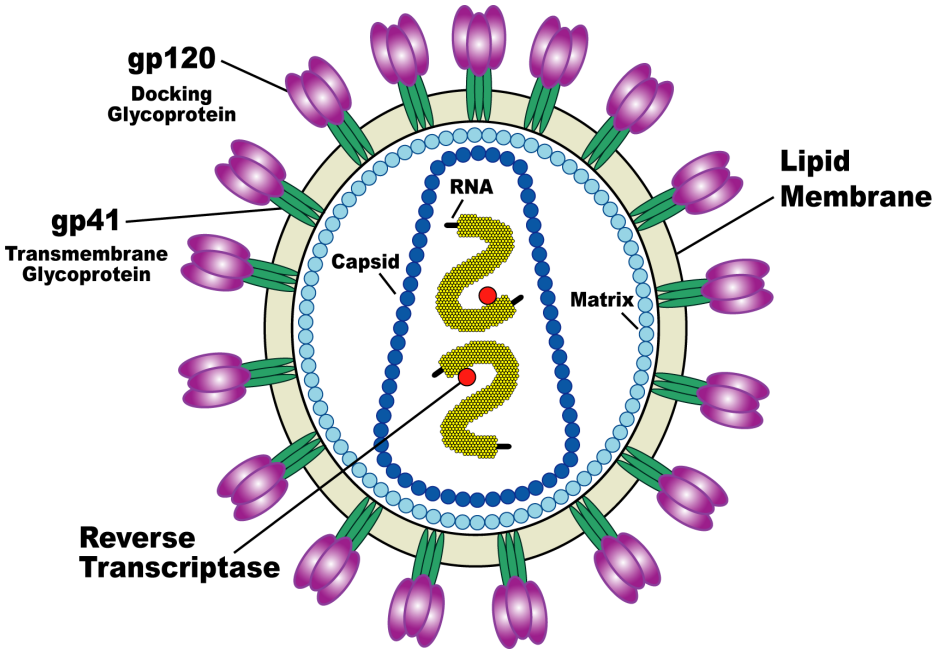
It was on a short-cut through the hospital kitchens that Albert was first approached by a member of the Antibiotic Resistance.

One-Bug-One-Drug ⇒ Many-Bugs-One-Drug

- Develop drugs that target *human* proteins.
- Prioritize human proteins interacting with *multiple* pathogens.

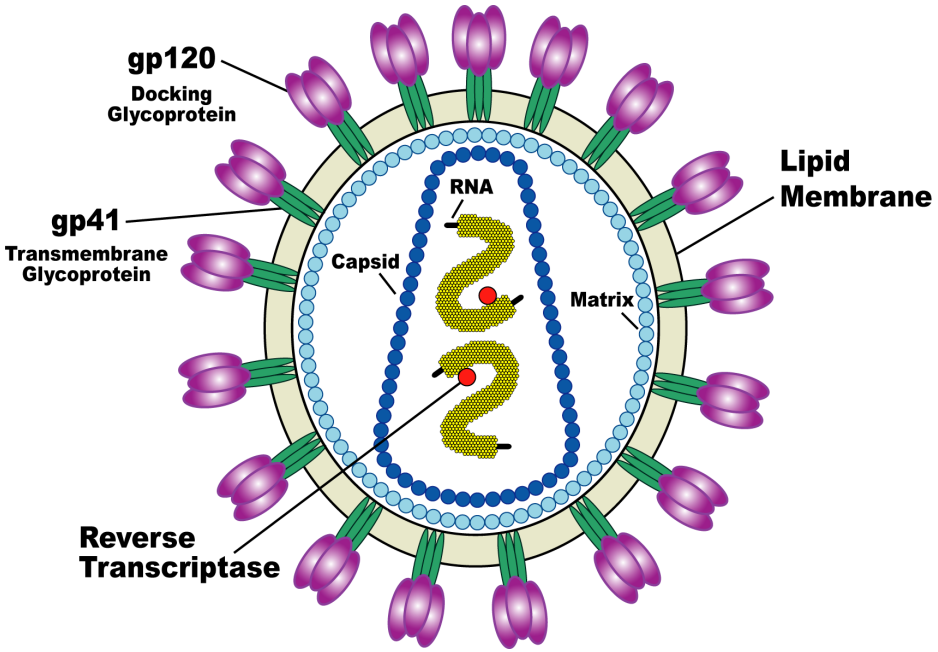


Viral Dependency Factors



- RNA viruses like HIV have very few genes.
- Viral dependency factor (VDF): *human* protein that virus needs to replicate and propagate.

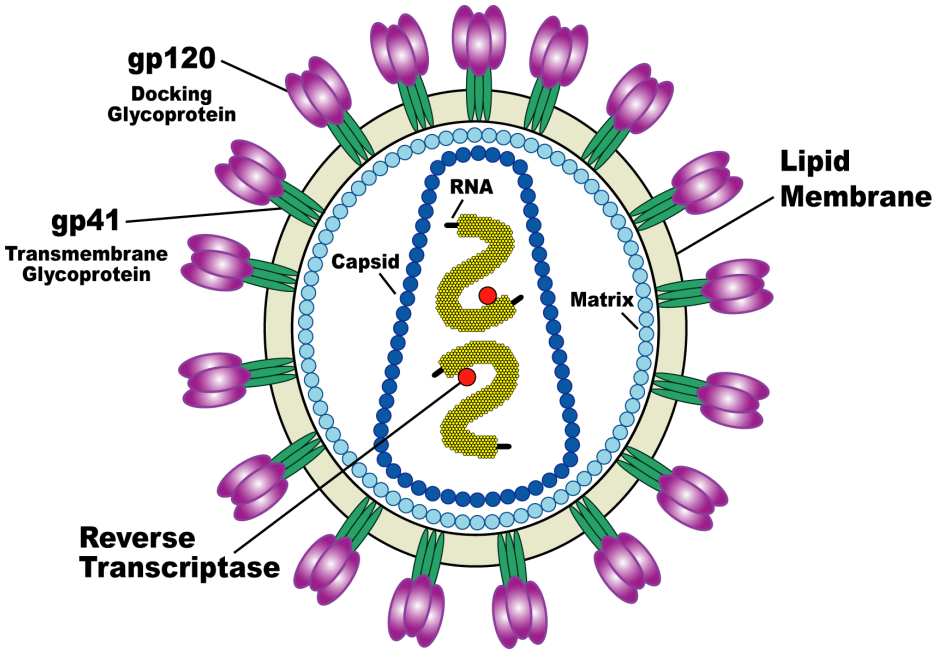
Viral Dependency Factors



- RNA viruses like HIV have very few genes.
- Viral dependency factor (VDF): *human* protein that virus needs to replicate and propagate.

- Recent genome-wide experiments have discovered dozens of VDFs for HIV, flu virus, West Nile virus, and Hepatitis C virus.
- Different experiments for HIV show very little overlap.

Viral Dependency Factors



- RNA viruses like HIV have very few genes.
- Viral dependency factor (VDF): *human* protein that virus needs to replicate and propagate.

- Recent genome-wide experiments have discovered dozens of VDFs for HIV, flu virus, West Nile virus, and Hepatitis C virus.
- Different experiments for HIV show very little overlap.
- Proteins in one experiment interact with proteins detected in other experiments.

Predicting VDFs

Collaboration with Michael Katze (Dept of Microbiology, Univ. of Washington) and Brett Tyler (VBI)

- Treat the problem as one of predicting gene function: which human genes have the function of "being used by viruses to propagate?"
- Modify previous approach: for every node v compute a function $0 \leq f(v) \leq 1$

Predicting VDFs

Collaboration with Michael Katze (Dept of Microbiology, Univ. of Washington) and Brett Tyler (VBI)

- Treat the problem as one of predicting gene function: which human genes have the function of "being used by viruses to propagate?"
- Modify previous approach: for every node v compute a function $0 \leq f(v) \leq 1$ to minimise

$$\sum_{(u,v)} w_{uv}(f(u) - f(v))^2$$

Predicting VDFs

Collaboration with Michael Katze (Dept of Microbiology, Univ. of Washington) and Brett Tyler (VBI)

- Treat the problem as one of predicting gene function: which human genes have the function of "being used by viruses to propagate?"
- Modify previous approach: for every node v compute a function $0 \leq f(v) \leq 1$ to minimise

$$\sum_{(u,v)} w_{uv}(f(u) - f(v))^2 + \lambda \sum_{v} f^2(v)$$

Predicting VDFs

Collaboration with Michael Katze (Dept of Microbiology, Univ. of Washington) and Brett Tyler (VBI)

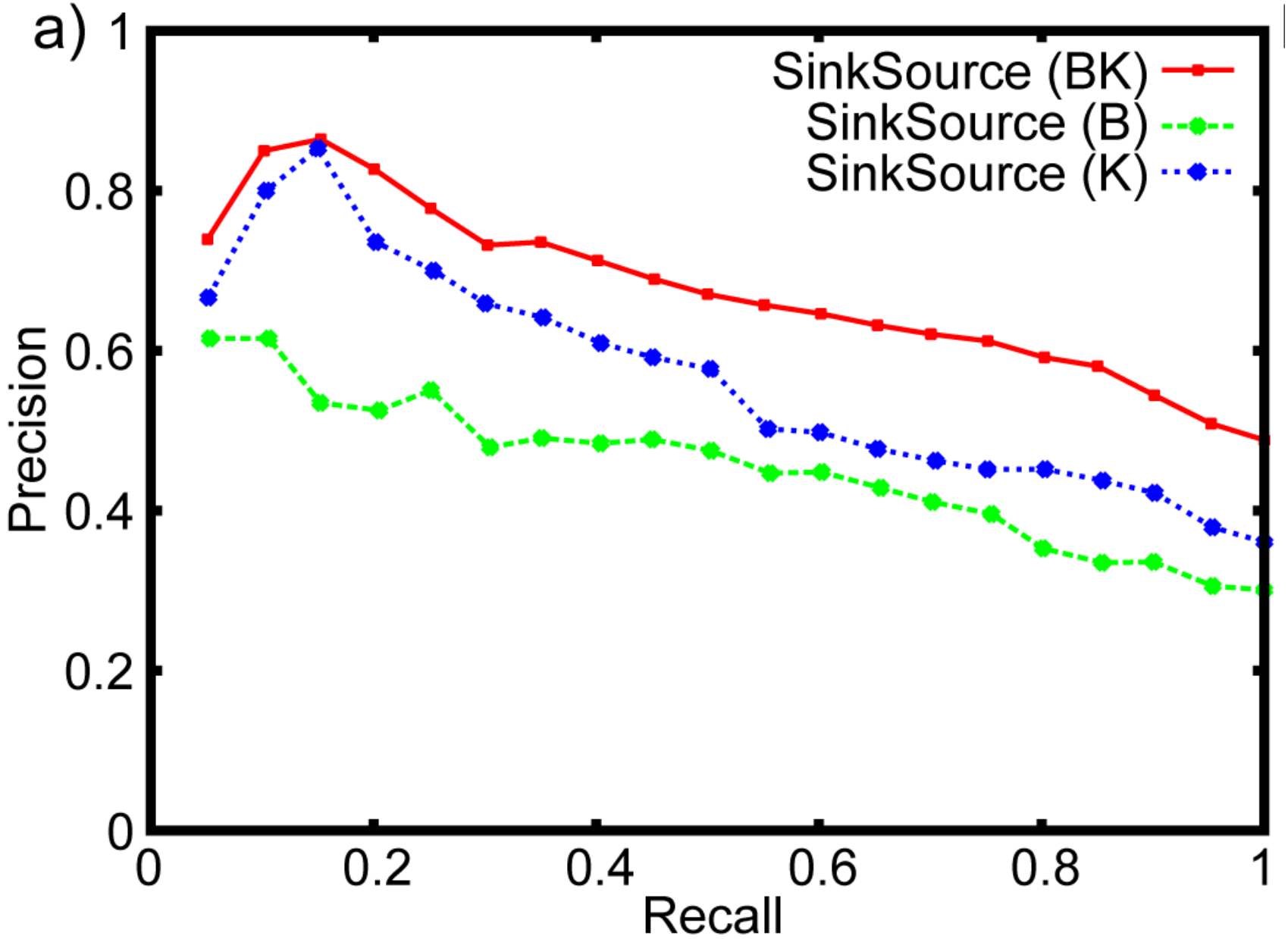
- Treat the problem as one of predicting gene function: which human genes have the function of "being used by viruses to propagate?"
- Modify previous approach: for every node v compute a function $0 \leq f(v) \leq 1$ to minimise

$$\sum_{(u,v)} w_{uv}(f(u) - f(v))^2 + \lambda \sum_{v} f^2(v)$$

- Solve linear system of equations:

$$f(v) = \frac{\sum_u w_{uv} f(u)}{\lambda + \sum_u w_{uv}}$$

Cross-Validation Results



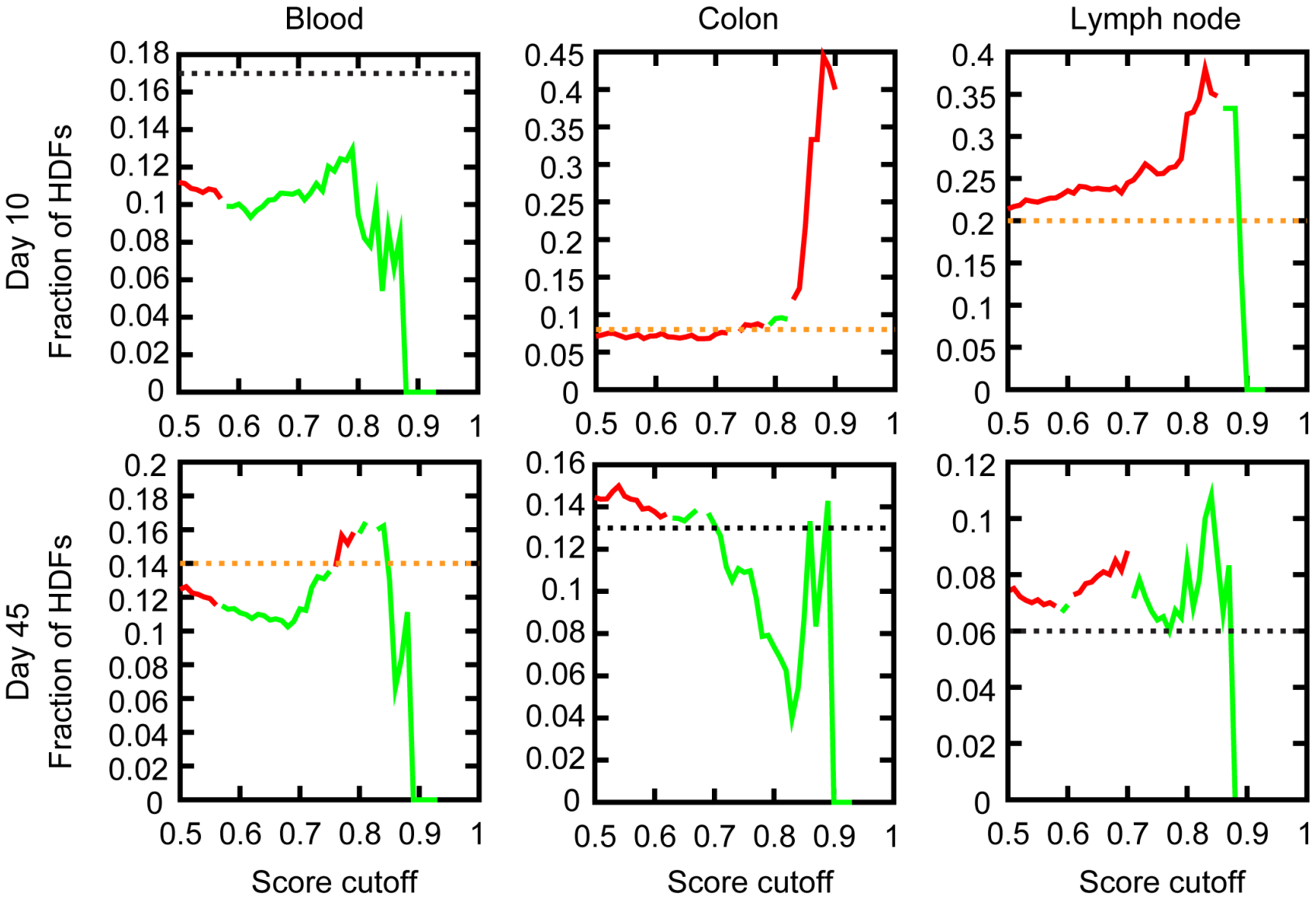
Independent Evaluation of Predictions

- VDFs for HIV were discovered in cell lines.
- Protein-protein interactions discovered in a wide variety of experiments.

Independent Evaluation of Predictions

- VDFs for HIV were discovered in cell lines.
- Protein-protein interactions discovered in a wide variety of experiments.
- Used gene expression data from SIV-infected non-human primates.
  - African Green Monkeys (AGMs) are natural hosts for SIV.
  - Pig-tailed Macaques (PMs) are susceptible to SIV.
- Computed which genes were differentially expressed between AGMs and PMs.
- Computed the statistical significance of the overlap between differentially-expressed genes and known/predicted HDFs.

Results of the Evaluation



Other Projects

- Building blocks of molecular interaction networks.
- Models to enable explicit comparisons between cell's response to different conditions.

What Makes Systems Biology Exciting?

- Use principles in computer science to solve problems that impact human life.
- Research is inter-disciplinary: collaborate closely with biochemists, bioengineers, geneticists, doctors, and plant biologists.
- Seek synergistic collaborations between computer science and biology.
- Train students in both computer science and biology.

Michael Katze	RNA viruses (HIV, flu, etc.)
Padma Rajagopalan	Liver tissue engineering
Bruno Sobral	Host-pathogen interactions
Brett Tyler	Predicting gene function, plant pathogens

How Can You Contribute?

How Can You Contribute?

- Curiosity about how the cell works, how it becomes ill, how does it survive attacks.
- Curious to learn how computer science helps to discover new things about the cell.

How Can You Contribute?

- Curiosity about how the cell works, how it becomes ill, how does it survive attacks.
- Curious to learn how computer science helps to discover new things about the cell.
- We need students who will learn both computer science and biology.
- They will form the next generation of scientists in systems biology.

Nothing in Biology Makes Sense Except in the Light of Evolution