10-810 /02-710
Computational Genomics

Physical networks active learning

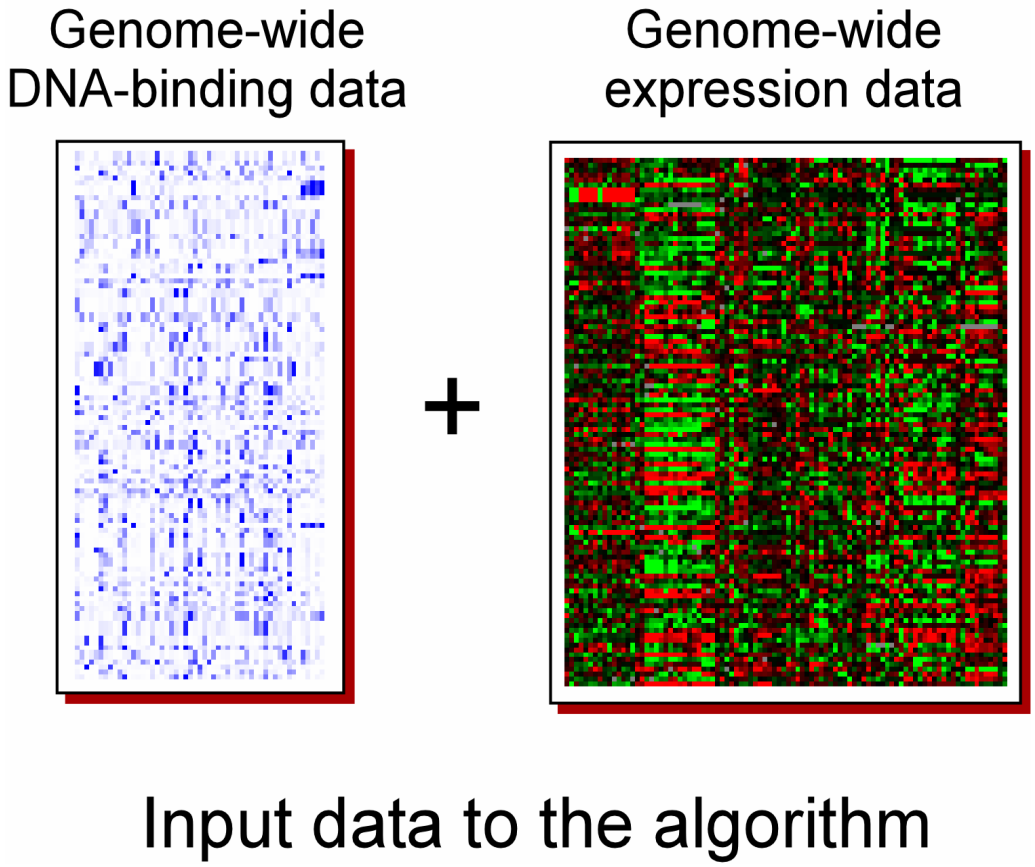
GRAM

Modules

- Gene Module
  - Set of genes that are co-regulated and co-expressed.
- Functional Module
  - Collection of gene modules with related function.

Modules provide an abstraction which reduces genetic network complexity without significant loss of explanatory power, and allows us to determine the significance of the model.

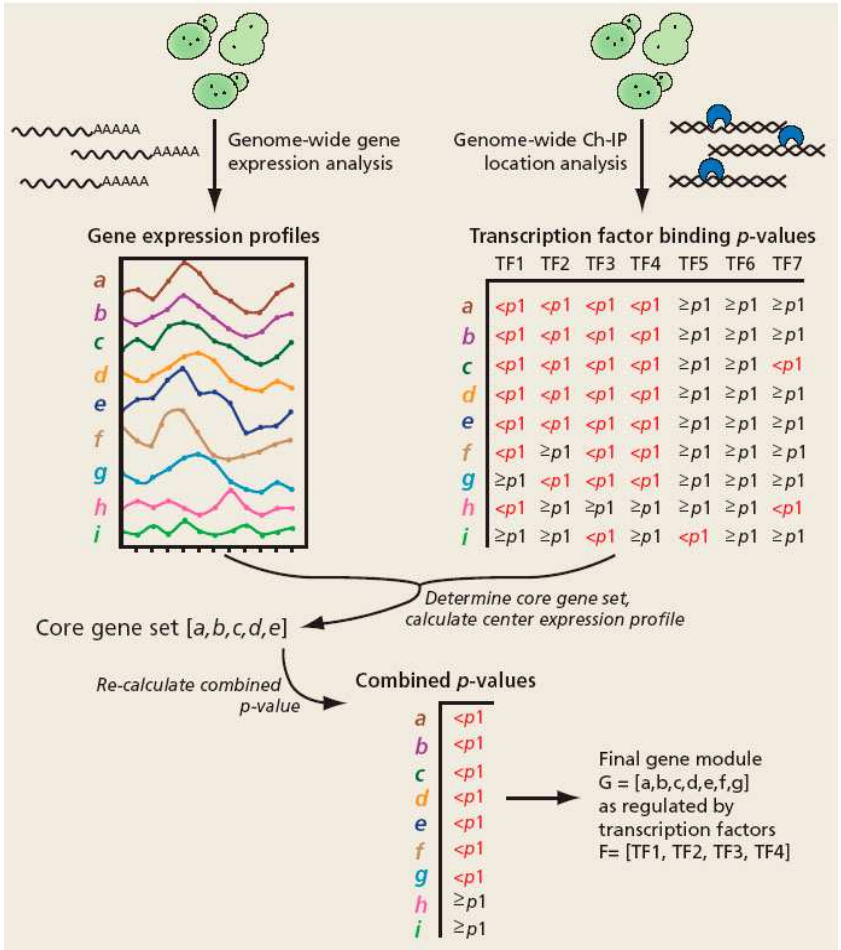
Genetic RegulAtory Modules (GRAM)



Input data to the algorithm

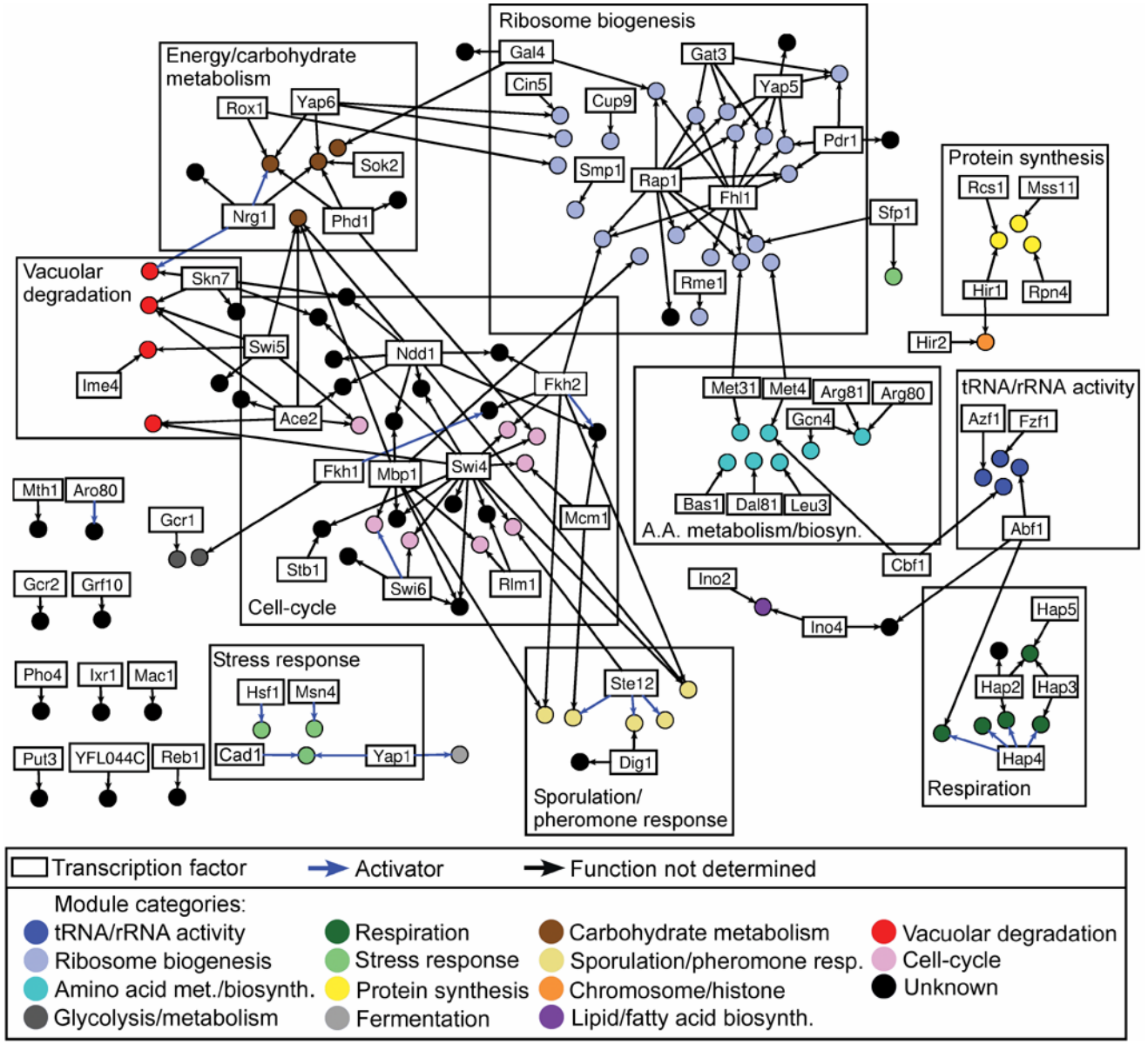
GRAM Algorithm Overview

- For each regulator combination, look at all genes bound (using a strict binding p-value).
- Find a core gene expression profile.
- Remove genes far away from core.
- Add genes close to the core (with relaxed p-value threshold).



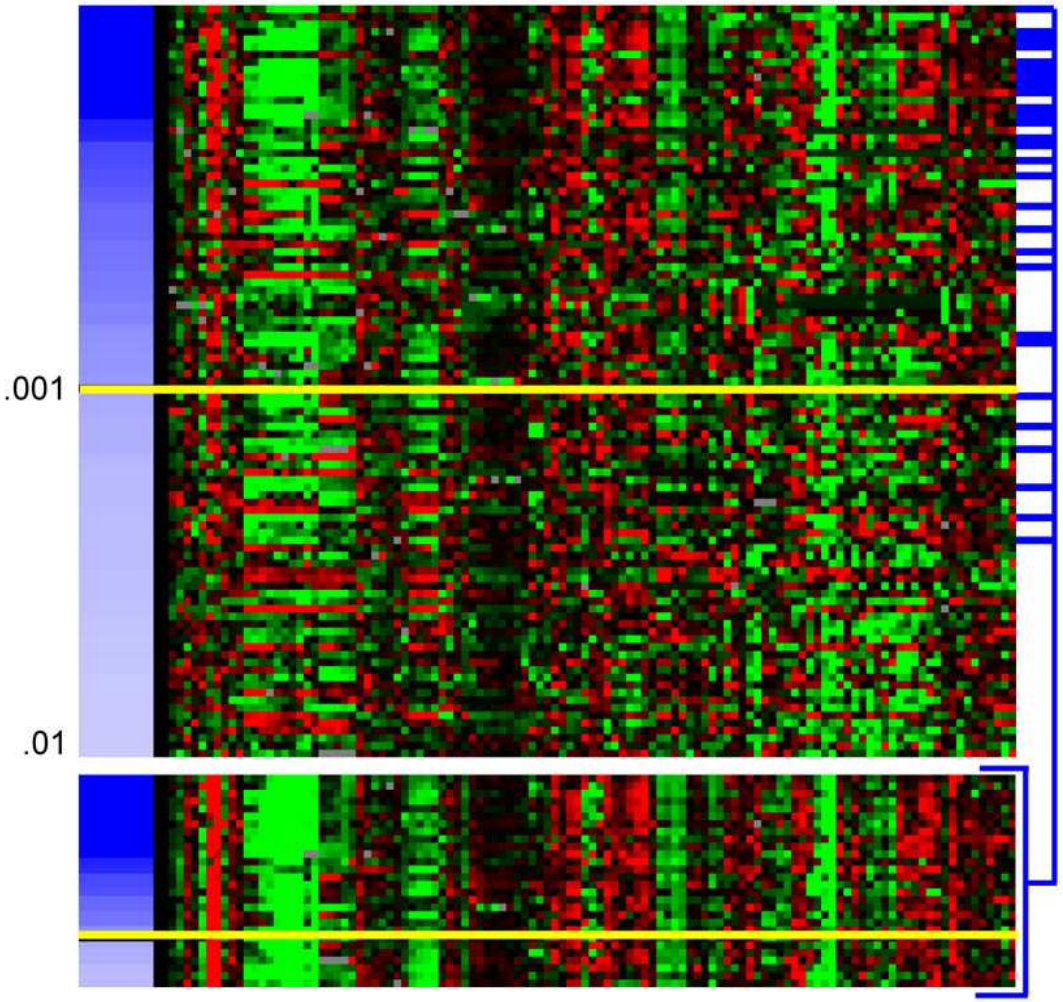
Results:

Rich Media Modules



Bar-Joseph et al Nature Biotechnology 2003

The Importance of Information Fusion: Using Only Binding



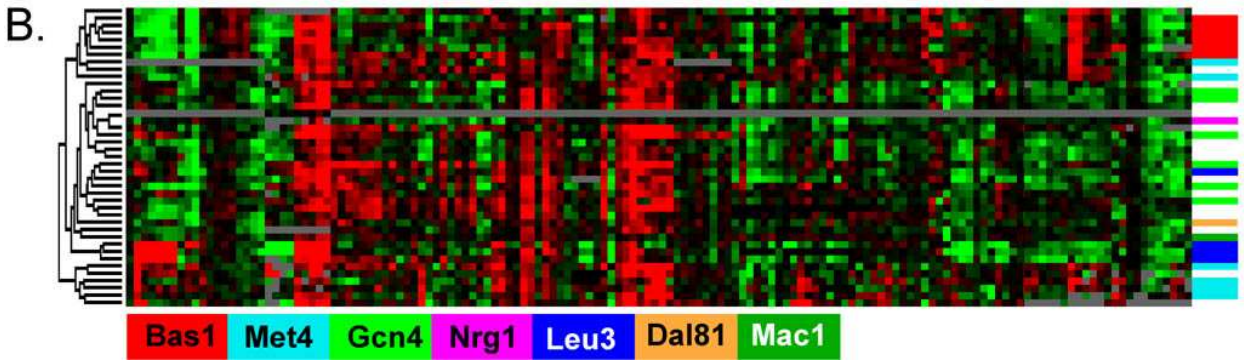
99 genes bound by Hap4 with a p-value < .01

Binding p-values form a continuum – where do you draw the cut-off line?

28 genes were selected by the GRAM algorithm; all are involved in respiration. Six of these genes (PET9, ATP16, KGD2, QCR6, SDH1, and NDI1) would not have been identified as Hap4 targets using the stringent .001 p-value threshold (p-values range from .0011 to .0036).

The Importance of Information Fusion: Using Only Expression

A cluster of amino acid synthesis genes



Eleven Significant Activators Found; Ten Previously Identified in Literature

Factor	Module function	Correlation	Comments
Ste12	Pheromone response	+0.64	Activator, required for pheromone response
Hap4	Respiration	+0.60	Activator of CCAAT box containing genes
Yap1	Detoxification	+0.53	Activator, possibly involved in oxidative stress response
Nrg1	Carbohydrate transport	+0.50	Previously identified as a repressor
Fkh1	Cell cycle	+0.49	Activator of cell cycle genes
Cad1	Detoxification	+0.47	Activator, involved in multi-drug resistance
Aro80	Energy and metabolism	+0.40	Activator, regulation of amino acid synthesis
Swi6	Cell cycle	+0.39	Activator of cell cycle genes
Msn4	Stress response	+0.38	Activator, involved in stress response
Fkh2	Cell cycle	+0.37	Activator of cell cycle genes
Hsf1	Stress response	+0.36	Activator of heat shock related genes

Validation Ideas

- Literature.
- Curated databases (e.g., GO/MIPS/TRANSFAC).
- Other high throughput data sources.
- “Randomized” versions of data.
- New experiments.

GRAM Network Validation

- Literature:
  - Many TF interactions predicted by modules corresponded well to literature (but what about ones that didn't…)
- Curated databases:
  - Computed enrichment for genes in modules for MIPS categories using the hypergeometric distribution.
  - Modules belong to diverse array of categories corresponding to cellular processes such as amino acid biosynthesis, carbohydrate and fatty acid metabolism, respiration, ribosome biogenesis, stress response, protein synthesis, fermentation, and the cell cycle.
- "Randomized" data:
  - When compared to results generated using binding data alone, there was 3-fold increase in modules significantly enriched in MIPS categories.

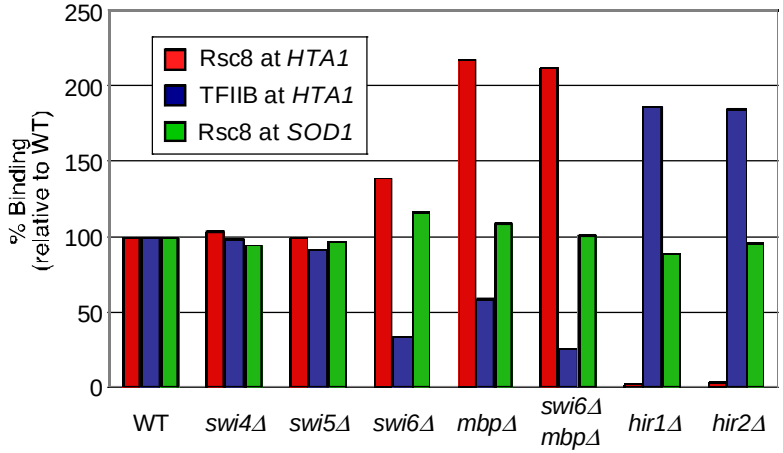
Predicting Mechanisms of Transcription Factors Regulation

Binding Predictions

Cluster 9 regs: STB1 YPD | SWI4 YPD				p-value
avg corr: 0.62927				
YCR065W	HCM1	HCM1	G1	0.0012
YDR501W	YDR501W	YDR501W	G1	0.00002
YGR109C	CLB6	CLB6	G1	0.0013
YGR221C	YGR221C	YGR221C	G1	0.0009
YIL140W	SRO4	SRO4	G1	0.008
YIL141W	YIL141W	YIL141W	G1	0.008
YMR179W	SPT21	SPT21	G1	0.007
YNL289W	PCL1	PCL1	G1	0.000005
YPL256C	CLN2	CLN2	G1	0.00007

Combinatorial regulation

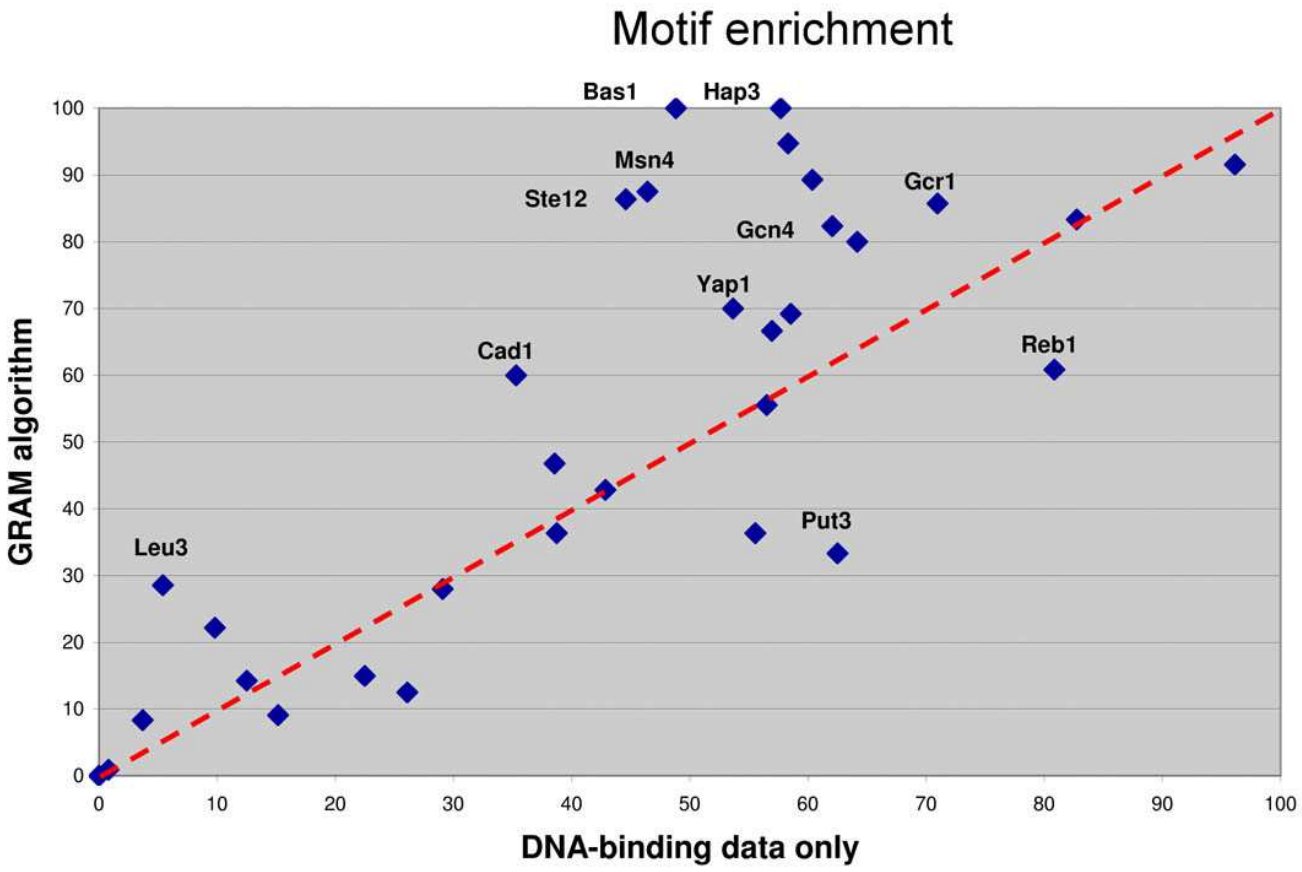
Cluster 33 regs: HIR1 YPD | HIR2 YPD | Rsc8				
avg corr: 0.81661				
YBR009C	HHF1	histone H		
YBR010W	HHT1	histone H		
YDR224C	HTB1	histone H2		
YDR225W	HTA1	histone H2		
40.10.03	chromoso	2.52E-09	44	4
04.05.01.0	transcripti	9.44E-06	334	4
04.05.01	mRNA syı	2.07E-05	406	4



Ng *et al*, *Genes Dev*. *2002*

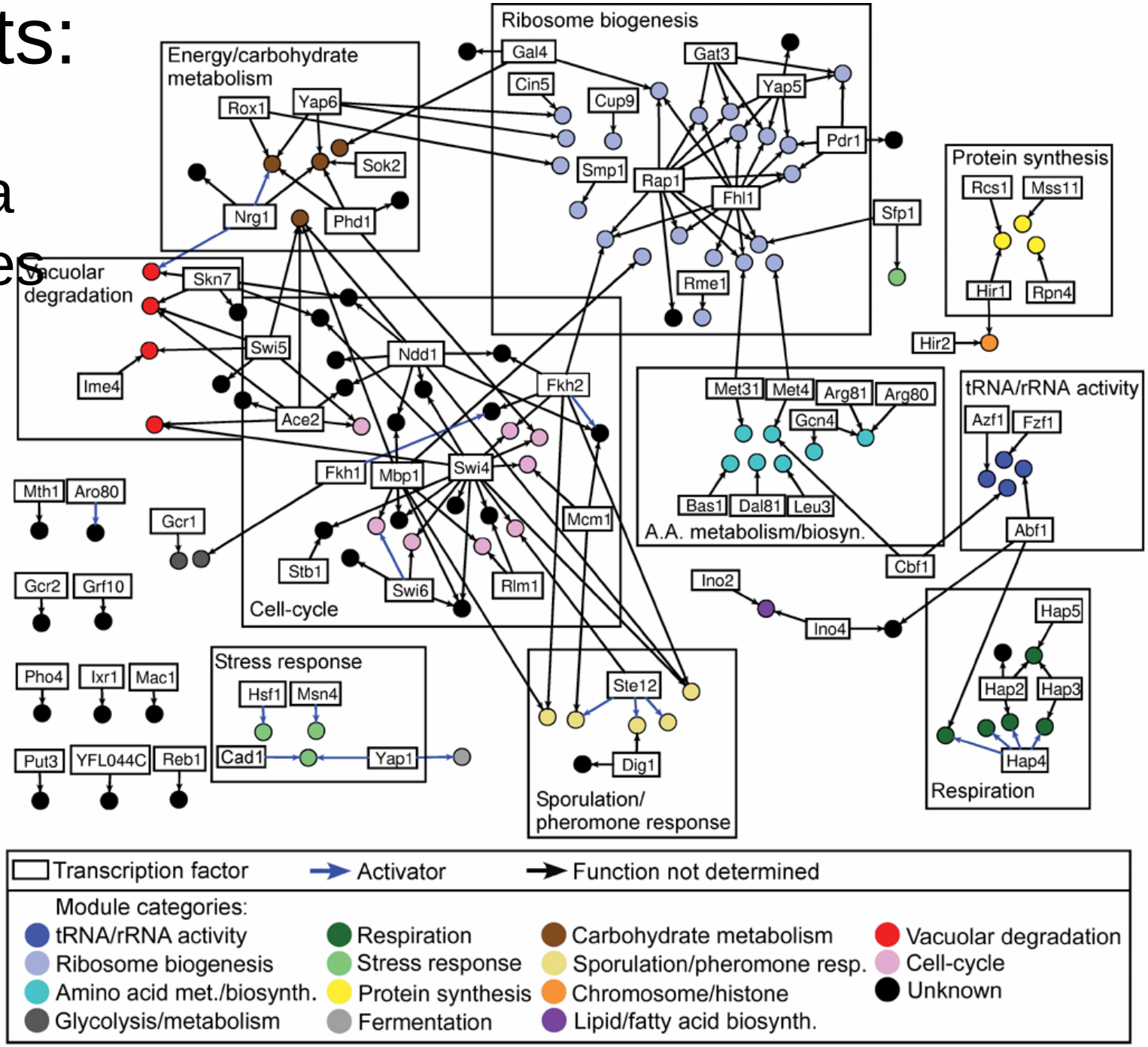
Connection to Sequence Data

What is the percentage of genes bound by factors with known motifs that contain the motif ?



Results:

Rich Media Modules



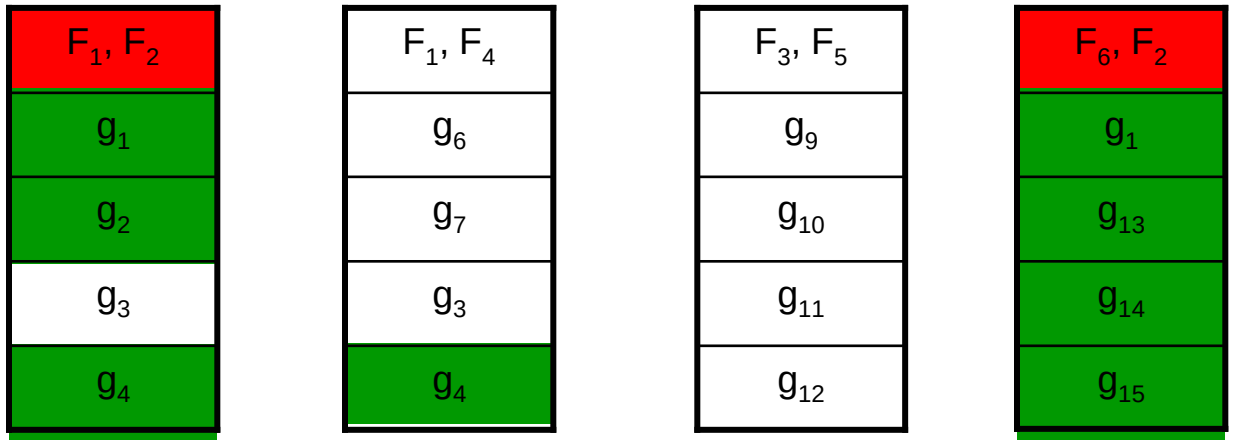
Sub-Network Discovery

- Identify genes involved in the system.
- Identify the factors controlling the system, and the modules involved.
- Determine a dynamic model for the activation of the modules by the identified factors.

We extend GRAM and combine it with our continuous representation and alignment algorithms to construct a dynamic model for a sub-network

Sub-Networks Discovery Algorithm

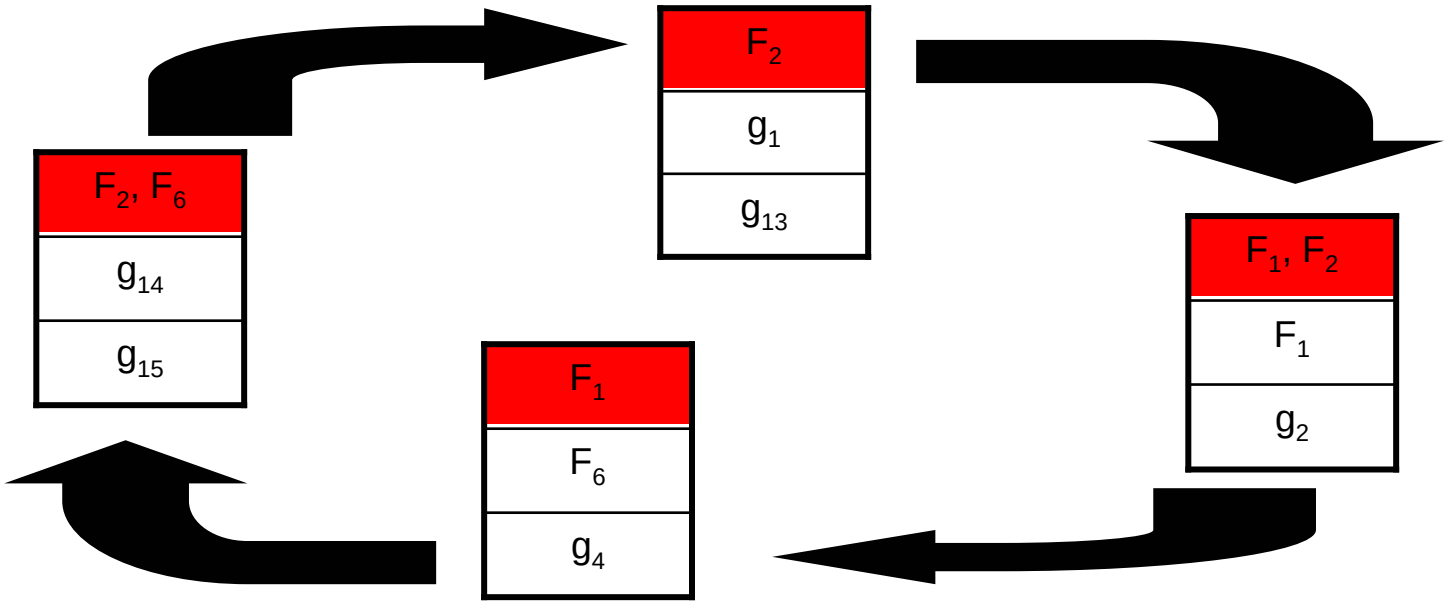
1.



p-value: 10^{-4} p-value: 0.2 p-value: 0.7 p-value: 10^{-6}

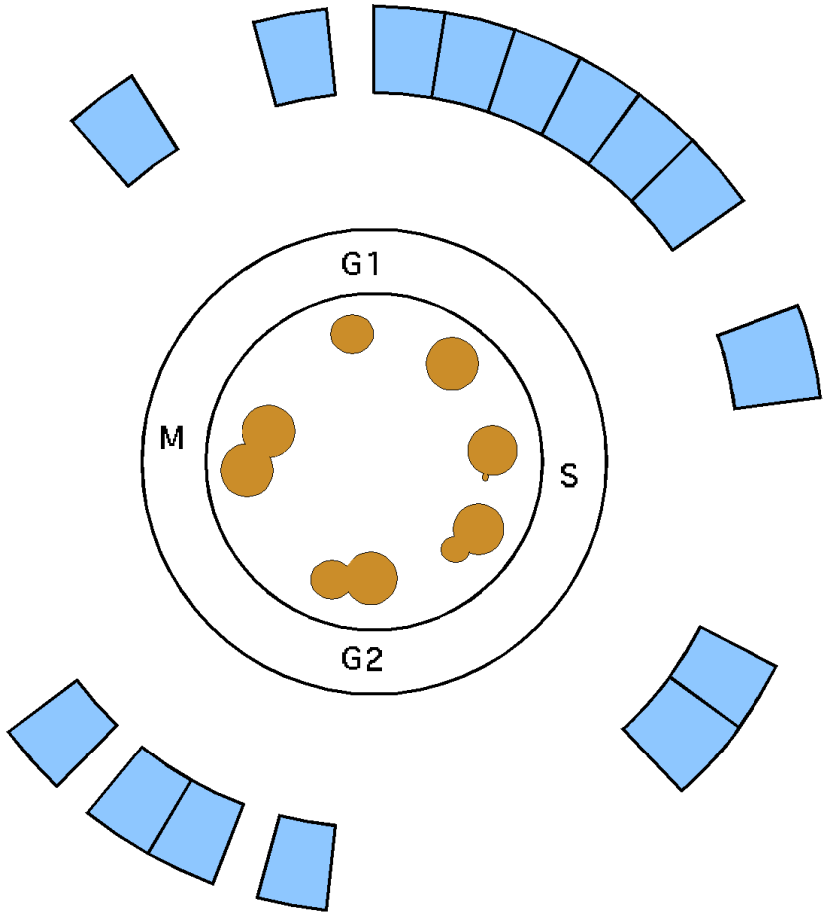
Factors = {F_1,F_2,F_6} Genes = {g_1,g_2,g_4,g_{13},g_{14},g_{15},g_{22},g_{24}}

2.



Assembly of the Cell Cycle Transcriptional Regulatory Network

Blue boxes: gene modules

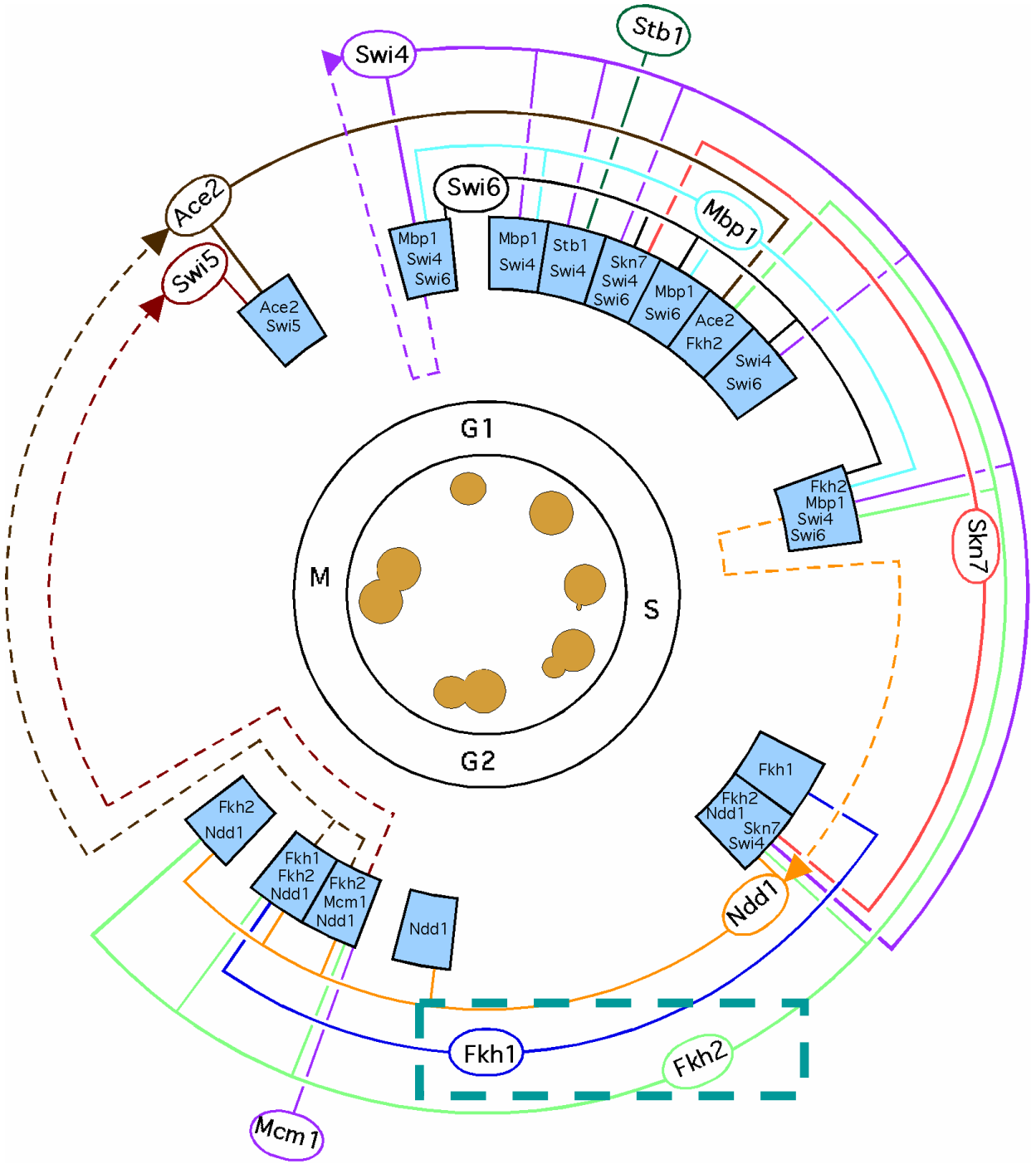


Assembly of the Cell Cycle Transcriptional Regulatory Network

Blue boxes: gene modules

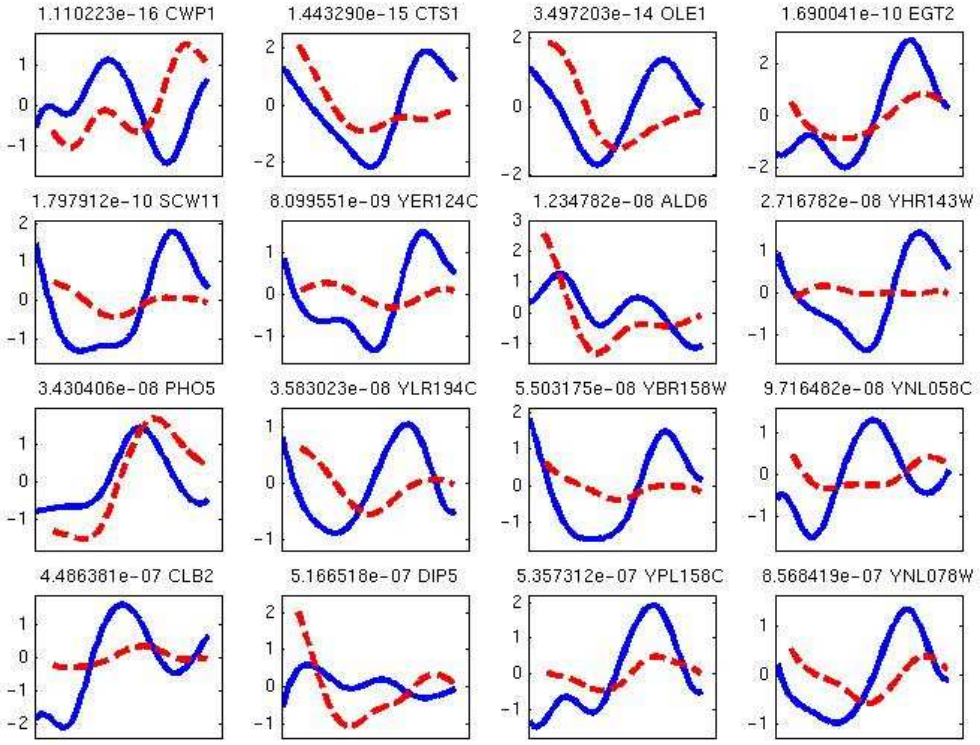
Individual regulators: ovals, connected to their modules

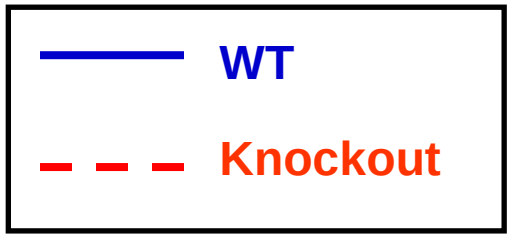
Dashed line: extends from module encoding a regulator to the regulator protein oval



Lee et al Science 2002

Results for the Fkh1/2 Knockout





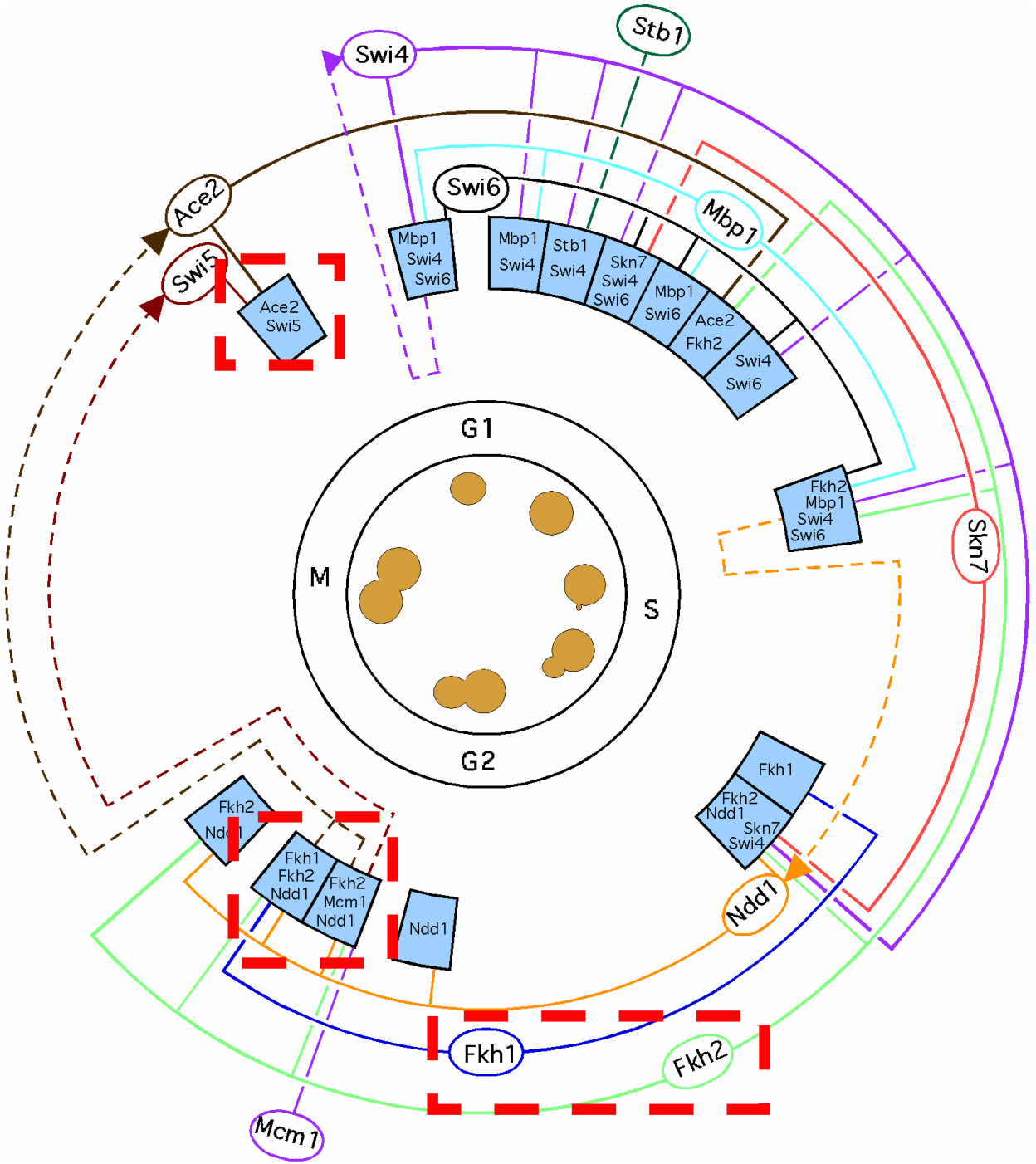
Bar-Joseph et al PNAS 2003

Assembly of the Cell Cycle Transcriptional Regulatory Network

Blue boxes: gene modules

Individual regulators: ovals, connected to their modules

Dashed line: extends from module encoding a regulator to the regulator protein oval

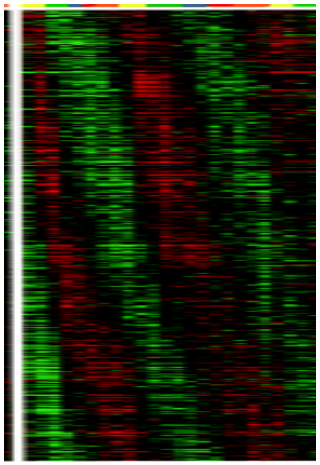


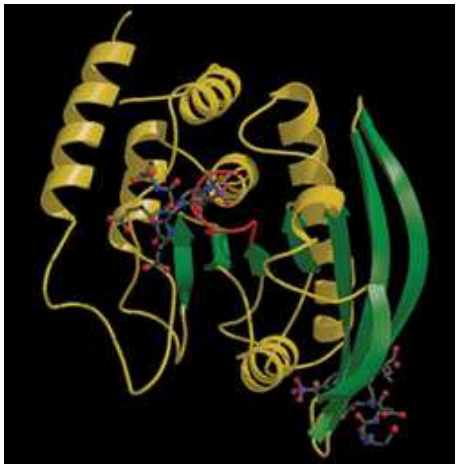
Physical networks

Data integration

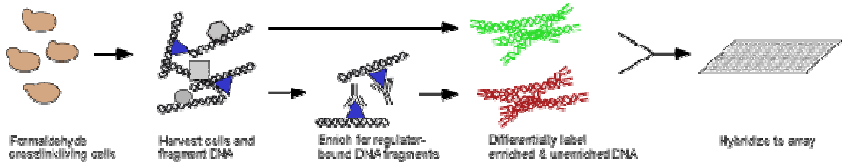
Protein interactions

Gene expression



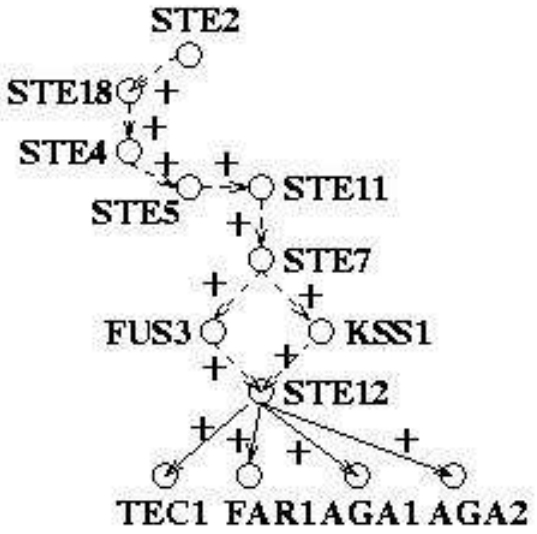


Protein-DNA binding



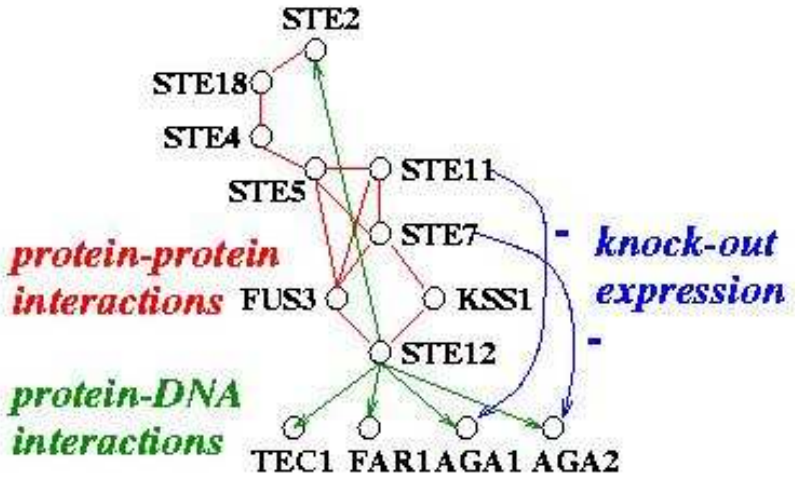
Yeast mating pathway

- A graph depicting physical interactions and functional annotations.



A mechanistic model of gene regulation

- Physical data:
  - Yeast binding data
  - DIP database (PPI)
- Functional data:
  - Rosetta compendium knockout data



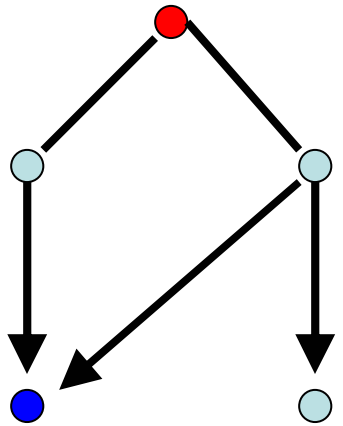
Inferring the mechanistic model from observed data

Key question: How do we construct the model from known mechanisms and constraints from observed data?

- Decompose data into pairwise items.
- Construct potential functions specifying constraints of each item.
- Combine potential functions by multiplication.

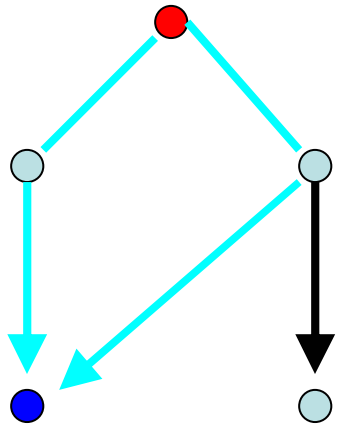
Requirements to explain knock-out data

- There is at least one connecting path.



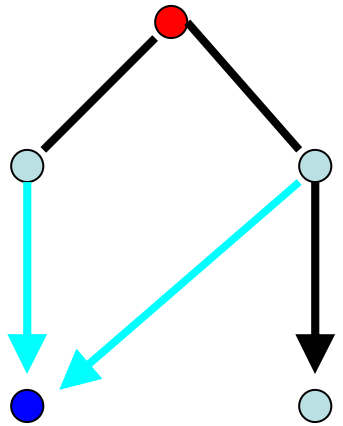
Requirements to explain knock-out data

- There is at least one connecting path.
- Edge directions along the path are consistent with the knock-out effect.



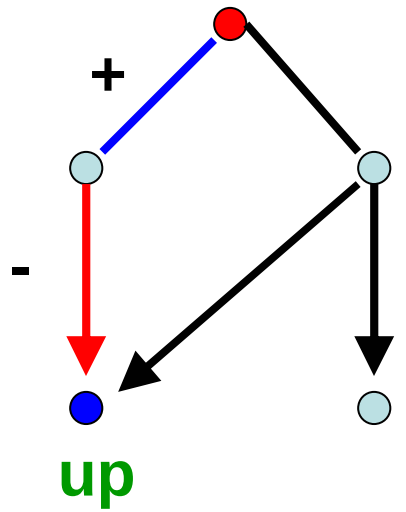
Requirements to explain knock-out data

- There is at least one connecting path.
- Edge directions along the path are consistent with the knock-out effect.
- The last edge on each path is a protein-DNA edge.



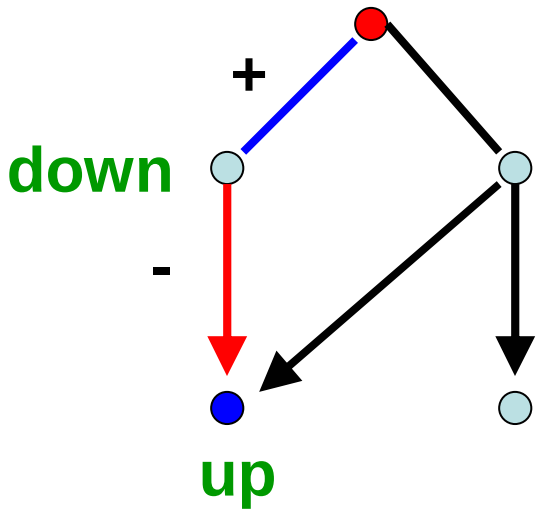
Requirements to explain knock-out data

- There is at least one connecting path.
- Edge directions along the path are consistent with the knock-out effect.
- The last edge on each path is a protein-DNA edge.
- The aggregate sign along the path is consistent with the knock-out effect.



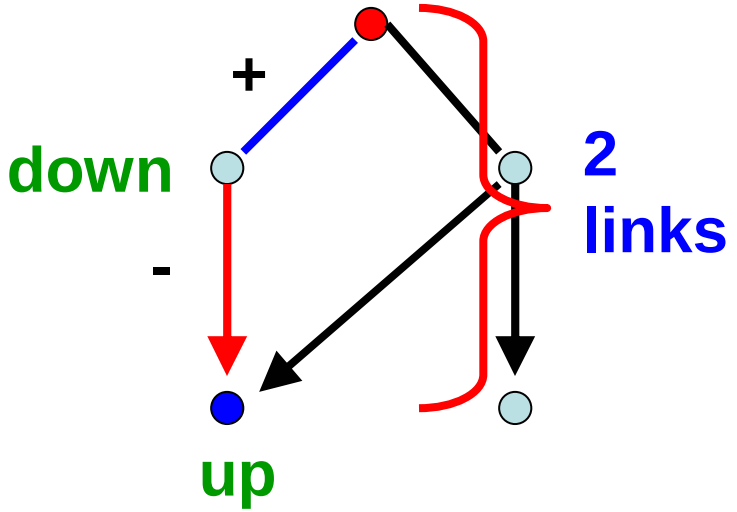
Requirements to explain knock-out data

- There is at least one connecting path.
- Edge directions along the path are consistent with the knock-out effect.
- The last edge on each path is a protein-DNA edge.
- The aggregate sign along the path is consistent with the knock-out effect.
- Intermediate genes along the path either have knock-out effects on or were not tested.



Requirements to explain knock-out data

- There is at least one connecting path.
- Edge directions along the path are consistent with the knock-out effect.
- The last edge on each path is a protein-DNA edge.
- The aggregate sign along the path is consistent with the knock-out effect.
- Intermediate genes along the path either have knock-out effects or were not tested.
- The path length is upper bounded.

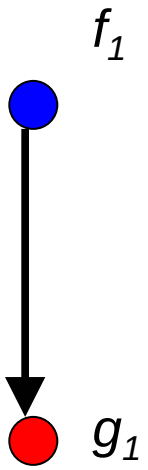


Factor graph formalism

- Factor graph is an undirected bipartite graph where edges represent dependency
- The joint likelihood is written using a set of potential functions, one for each edge in the graph and possibly for potential functions representing node probabilities
- The key challenge is to determine the set of potential functions and how to encode them

Associations with binding data

- Assume we have p-value *y* for the event *x* (binding of f_1 to g_1).
- How can we use this value in a probabilistic setting?
- Possible solution: use likelihood ratio:



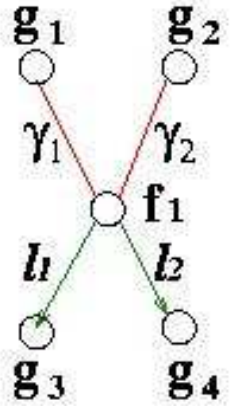
$$\frac{p(x \mid y)}{p(\sim x \mid y)}$$

x – the event of f binding to g

y – observed p-value

Associations with binding data

- Location data:
- Given a possible protein-DNA interaction e_i, the potential function $\phi_{ei}(x_{ei};y_{ei})$ is related to the direct evidence about this interaction:



$$\phi_{ei}(x_{ei};y_{ei}) = [\frac{p(y_{ei} \mid x_{ei}=1)}{p(y_{ei} \mid x_{ei}=0)}]^{x_{ei}}$$

- And similarly for protein interaction.

Determining the confidence in the observed data

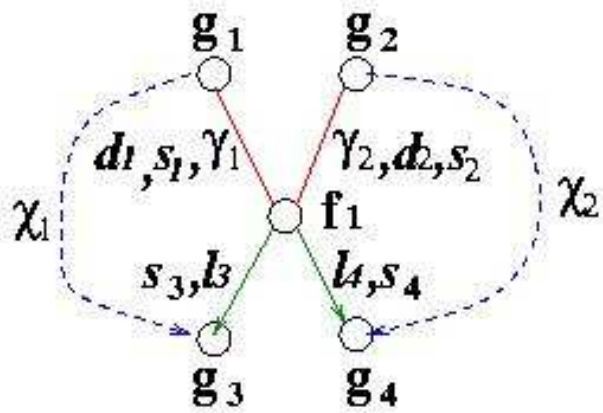
- In order to determine the probabilistic term in the potential function we use an appropriate error model.
- As a crude approximation, $p(y_{ei} \mid x_{ei})$ can be obtained from the binding p-value
- First, set p(measurement | interaction does not exist) = p-value
- The other side p(measurement | interaction exists) is set to a fixed value.
- The potential term for the protein interaction case is defined analogously.

Associations with knock-out expression data

- Given knockout expression data, we need to determine whether or not the knockout of gene *i* influenced gene *j*
- The interaction effect is associated with the observed data o by:

$$\phi_{i,j}(k_{i,j}, o_{k,i,j}) = [\frac{p(o_{k,i,j} \mid k_{i,j} = 1)}{p(o_{k,i,j} \mid k_{i,j} = 0)}]$$

- *k* can be explained by cascades of molecular interactions, i.e., paths in the physical model.



Knockout (cont.)

- Explanation conditions can be expressed as a logic clause of variables along the paths connecting a knock-out pair:
  - the knock-out interaction effect (χ_k)
  - edge presences (E_k),
  - edge directions (D_k), and sign (S_k),
  - and path selections (Σ_k).

- The potential term can also incorporates the situations of multiple paths and uncertainties of explanation.

Inference

- Potential functions are combined by multiplication.
- Goal: find the optimal configuration of the variables.
- This is done using a maximum likelihood approach using a variant of belief propagation.
- Using a graph known as a factor graph, the max-product algorithm is applied to obtain a MAP configuration.
- If the network is small, we can apply the max-product recursively to obtain all MAP configurations.

Datasets

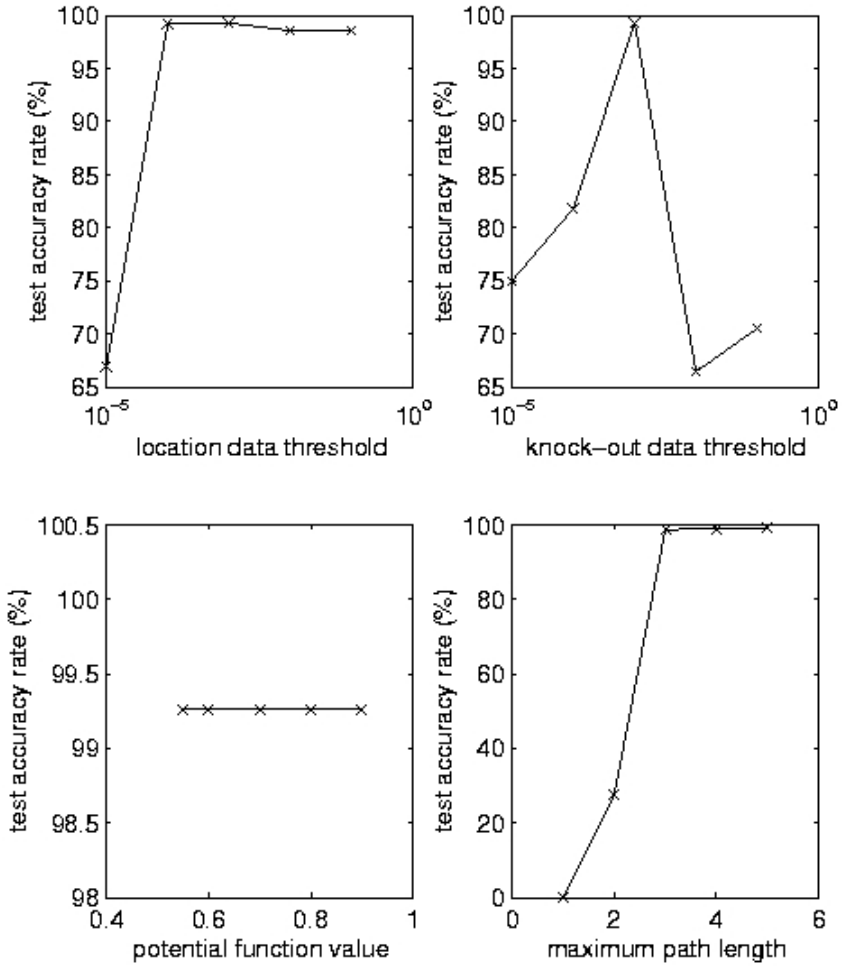
- 46 genes including 2 transcription factors (STE12 and MCM1).
- Binding p-value threshold 0.001 result in 34 protein-DNA edges (Lee et al., 2003).
- 30 protein-protein edges (DIP).
- 164 knock-out pairs from 10 experiments (Hughes et al., 2000).
- Maximal path length set to 5.

Results: yeast mating pathway

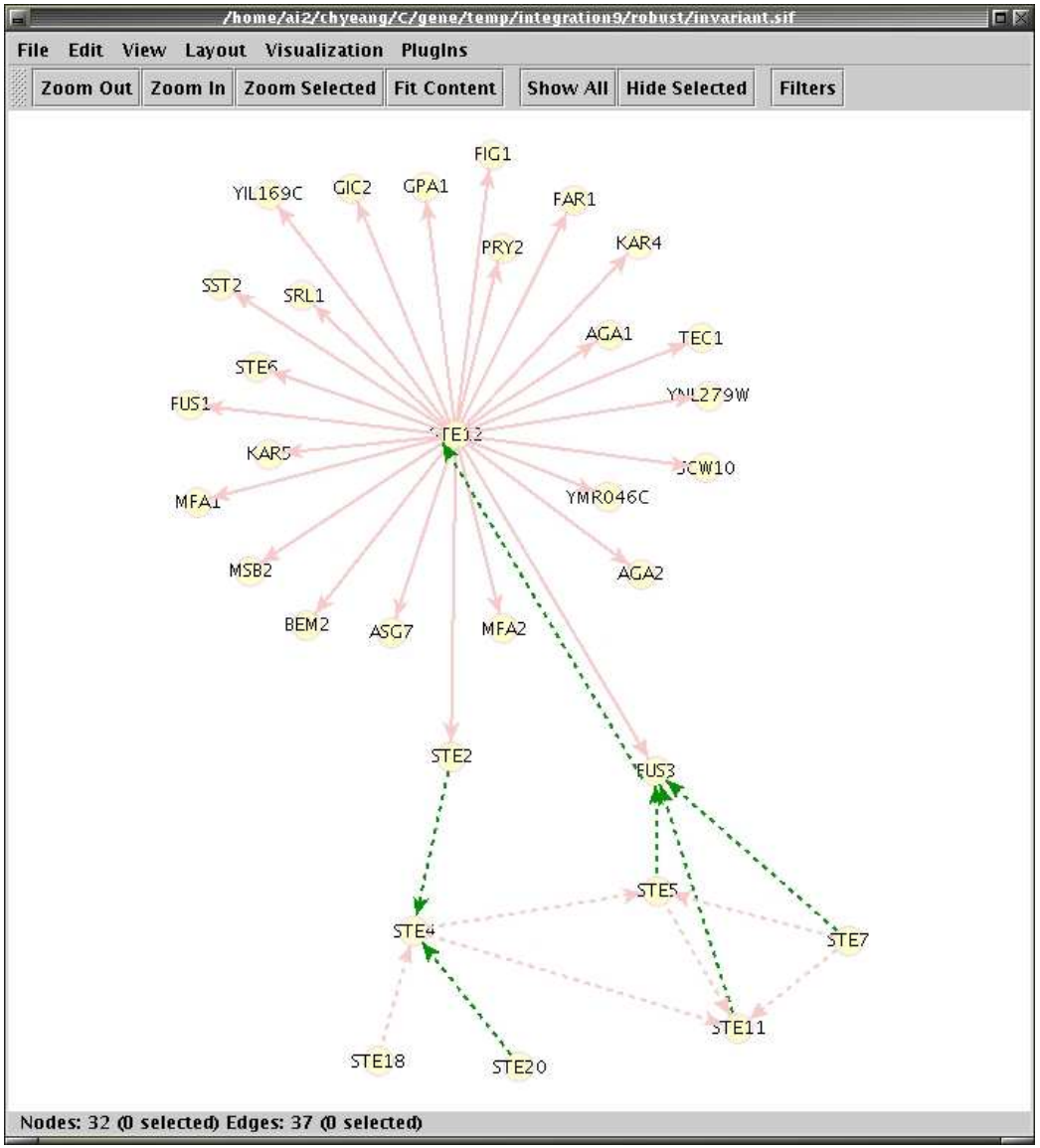
- 129 knock-out pairs are connected via valid paths.
- 8 MAP configurations.
- 129 knock-out pairs are explained by all MAP models.
- 106 knock-out pairs are explained by non-trivial inference.
- 2 knock-out pairs whose explanatory paths are not constrained by other knock-out pairs

Robustness of the model

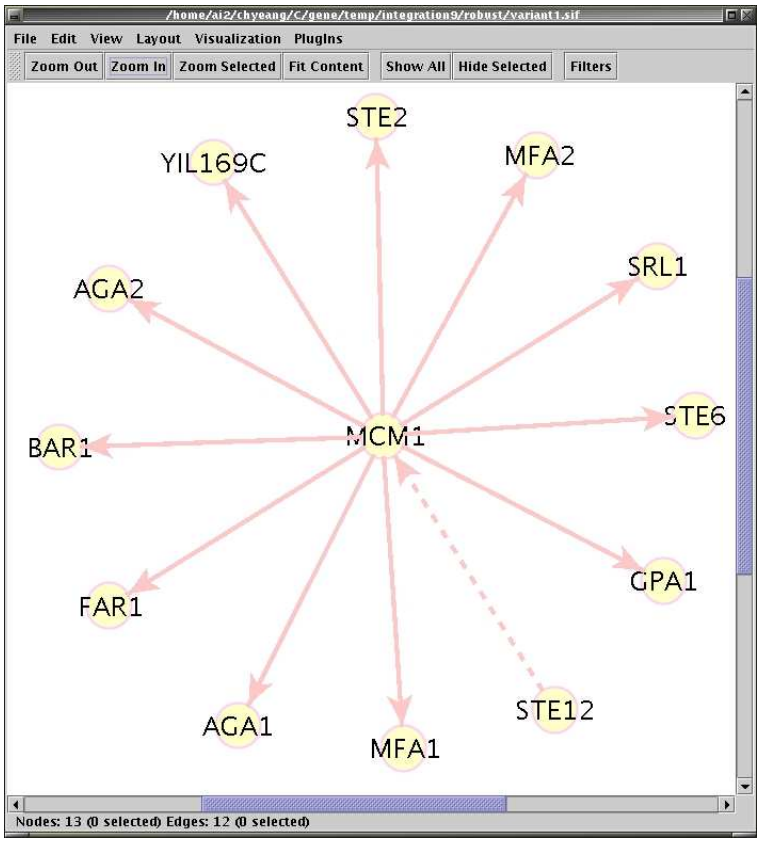
- Are prediction outcomes sensitive to parameter settings?
- Robustness tests on location and knock-out p-value cutoffs, potential values and path length

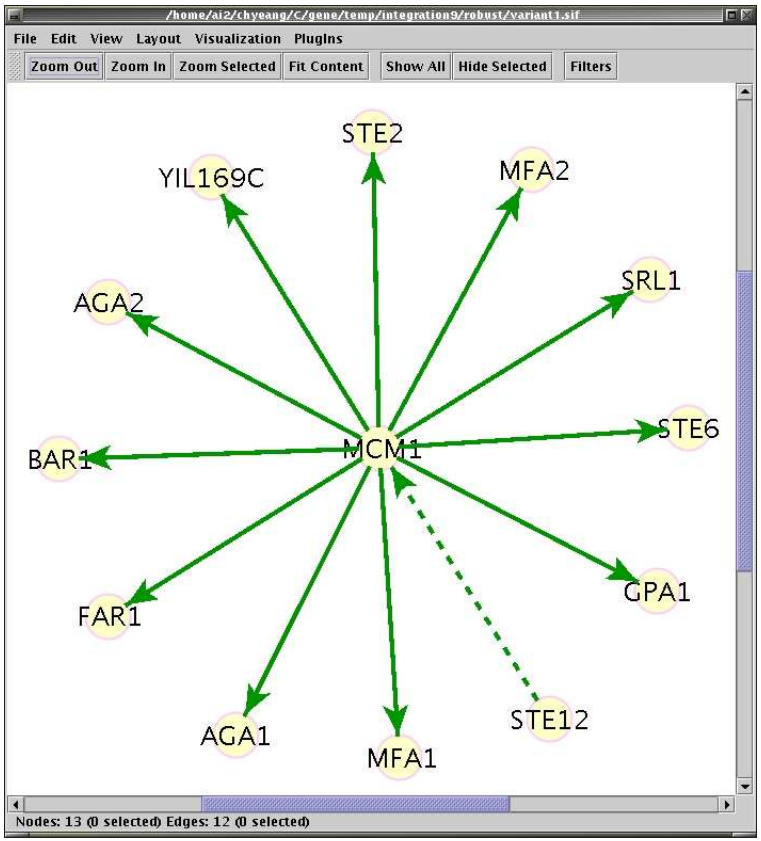


Common features for all MAP models

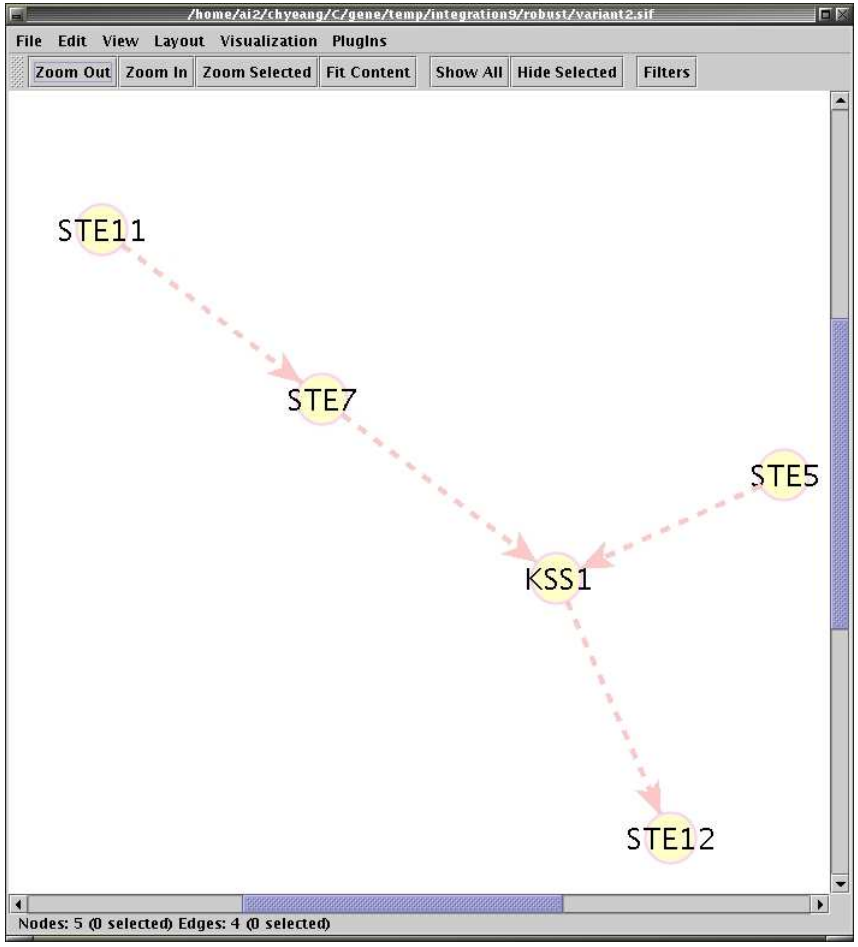


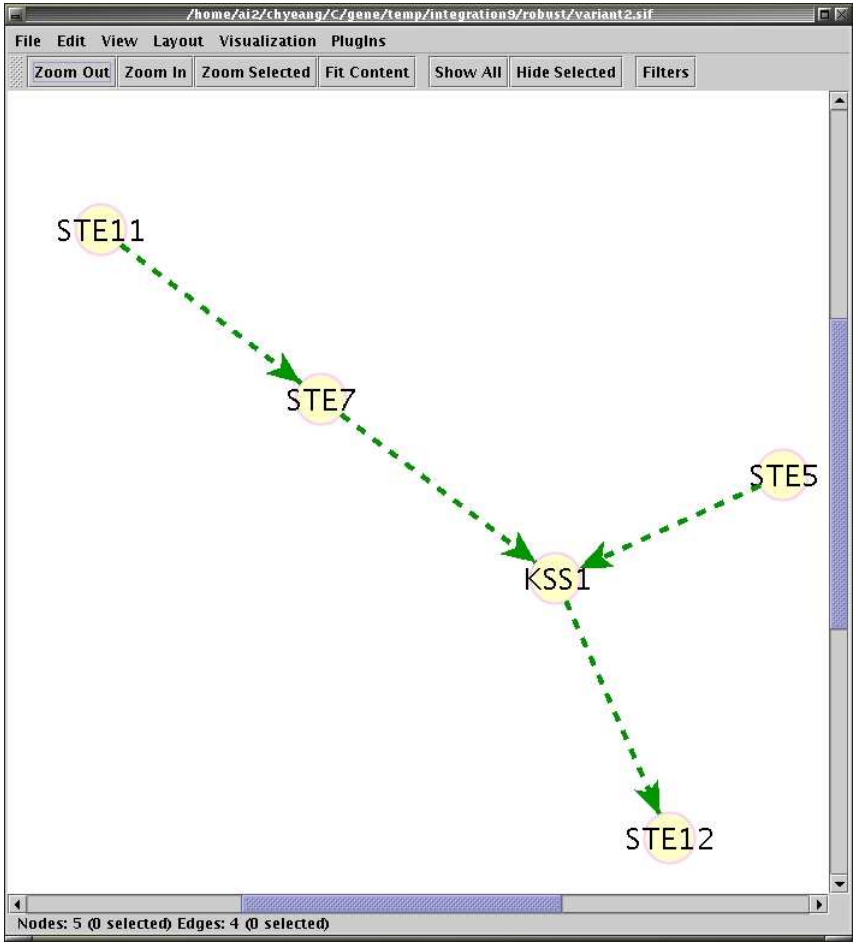
Variant features





Variant features





Resolving ambiguities in the model

- Resolving ambiguities in the model requires new experiments
- There are many possible experiments (knockout of every gene)
- How can we chose which one to perform?

Active learning

- Assume we want to teach a computer to distinguish between cats and dogs …








Can you give me some outdoor dog and indoor cat pictures?

Sure!

Active Learning for sampling time series data

- On the basis of current model, *M*, the learner
  - predicts the answers O_x to various possible queries q_x
  - computes which query's answer will be most beneficial in improving model quality (or minimizing the loss)
  - Perform the experiment, updates model with the answer

$$\min\langle Loss(q_x)\rangle = \min E[Loss(M^{O_x})]$$

Targeting specific network

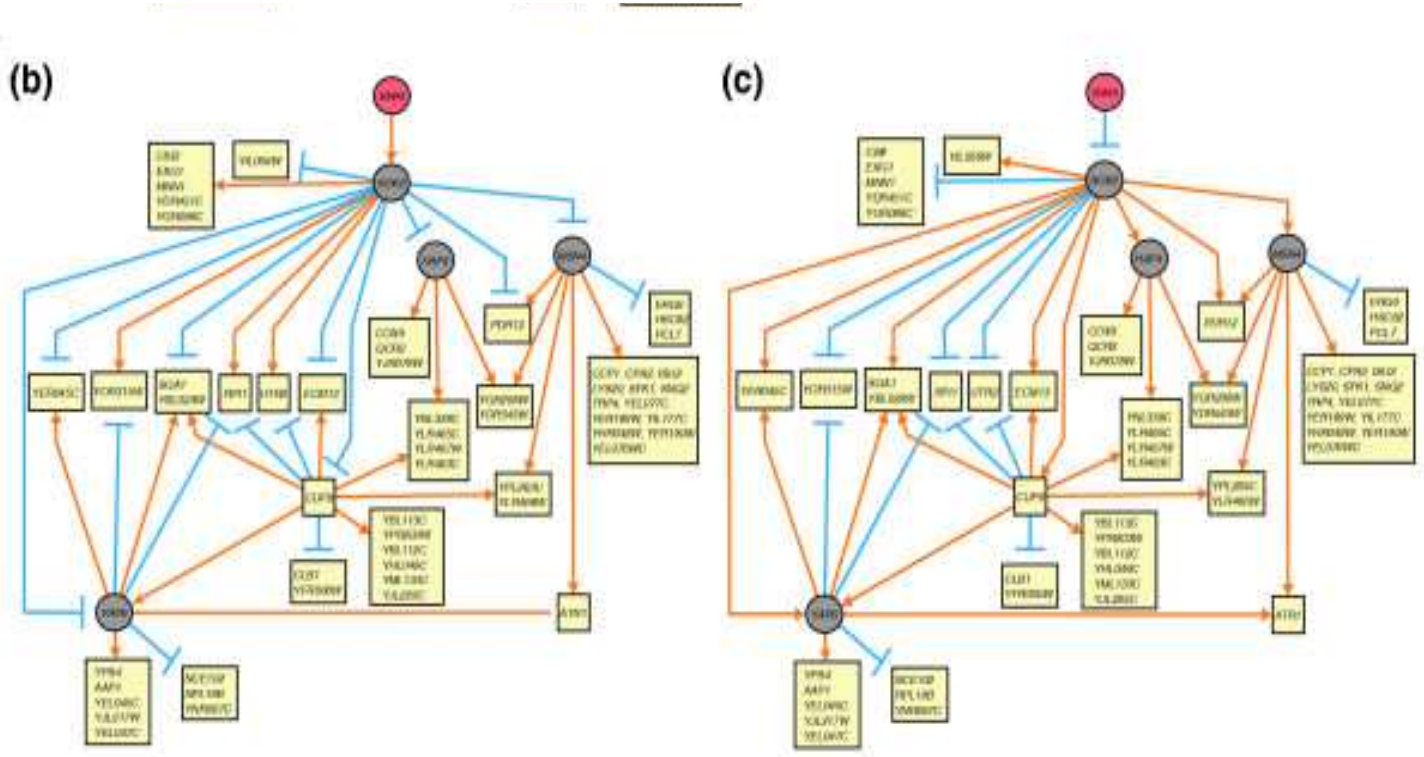


Table 2

Table 2

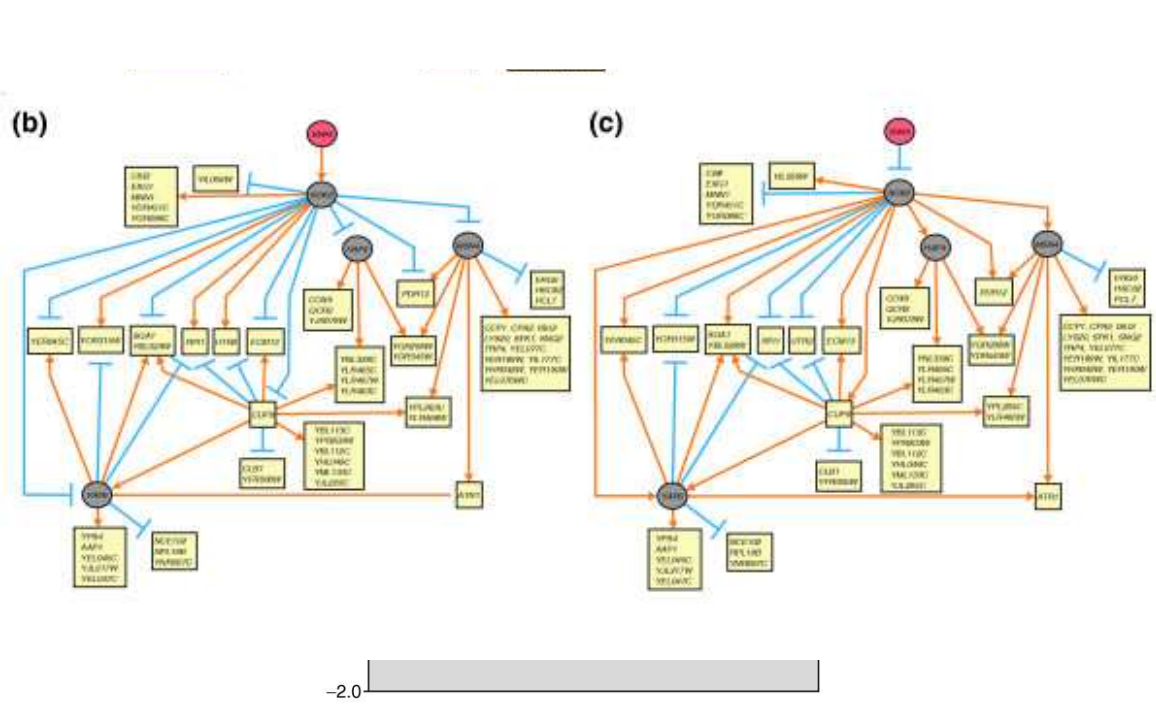
Top-ranking knock-out experiments proposed for model discrimination

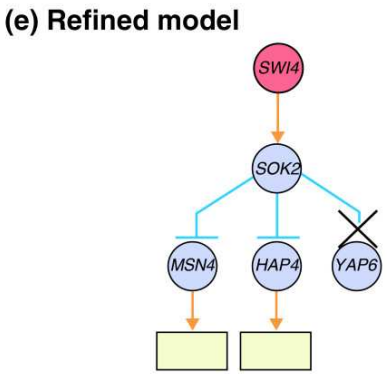
Gene	Function	Score	Downstream genes	Rank	Model
HHF1	Histone	52.1429	74	1	2
*SOK2**	Regulator for meiosis and PKA pathway	45.0279	64	2	1
CKA1	Protein kinase of cell cycle	45.0075	64	3	5
A2	Mating response	40.9023	58	4	4
*YAP6**	Stress response regulator	35.1652	50	5	1, 3
NRG1	Regulator of glucose dependent genes	31.6501	45	6	3
FKH1	Regulator of cell cycle	29.1194	41	7	2
FKH2	Regulator of cell cycle	26.7131	38	8	7
SLT2	Protein kinase of cell wall integrity pathway	23.4727	31	9	8
*MSN4**	Regulator of stress response	21.8224	31	10	1
*HAP4**	Regulator of cellular respiration	6.3310	9	34	1

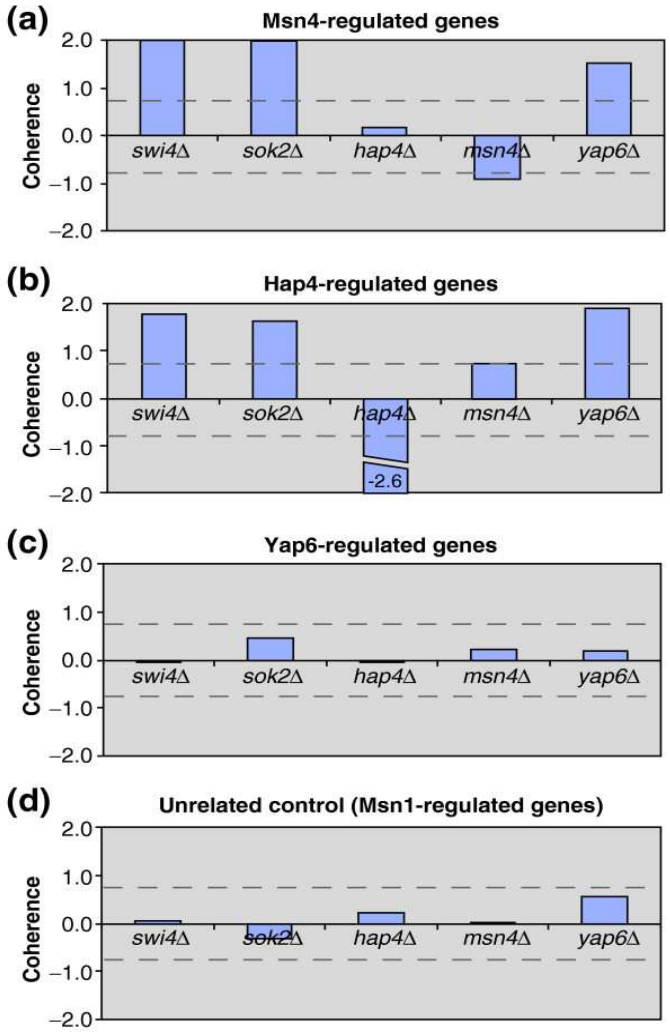
Using the ranked list

- How should we use the list in the previous table?
- Performing all the experiments at once ignores the dependency between these experiments
- Its much better to carry them one at a time
- However, that may cause other problems that are less desirable.

Experiments carried out







Projects

- Poster session: May 8th 1:00p until 2:30p in NSH 1507
- Each group should be ready to present their poster at that time
- Writeups (6-8 pages, no more than 8) are due at the poster session