Computational Discovery of Gene Modules and Regulatory Networks

Georg Gerber

MIT Department of EECS and

MIT/Harvard Health Sciences and Technology (HST)

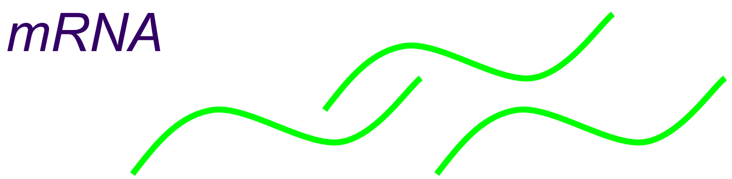
Motivation

High-level goal: Use high throughput data to discover patterns of combinatorial regulation and to understand how the activity of genes involved in related biological processes is coordinated and interconnected.

- Many previous efforts used expression data alone.
- Genome-wide binding data suggested new approaches, since this data provides *direct* evidence of physical interactions.

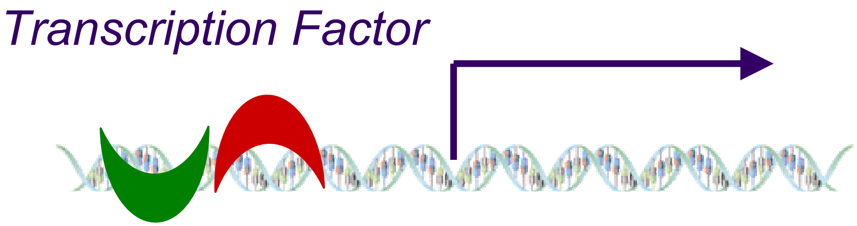
Expression and Binding Data

Gene expression data



expression - reflects ***functional*** changes in mRNA levels in different conditions

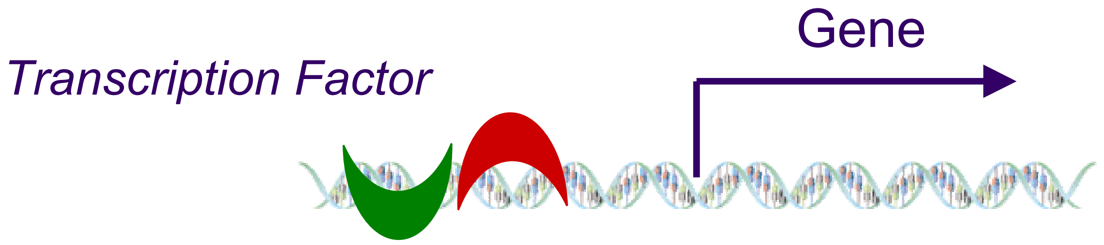
Protein-DNA binding data



binding – reflects ***physical*** interactions (connectivity)

These two data sources offer complementary information…

Protein-DNA Binding Data



Transcription Factor

Gene

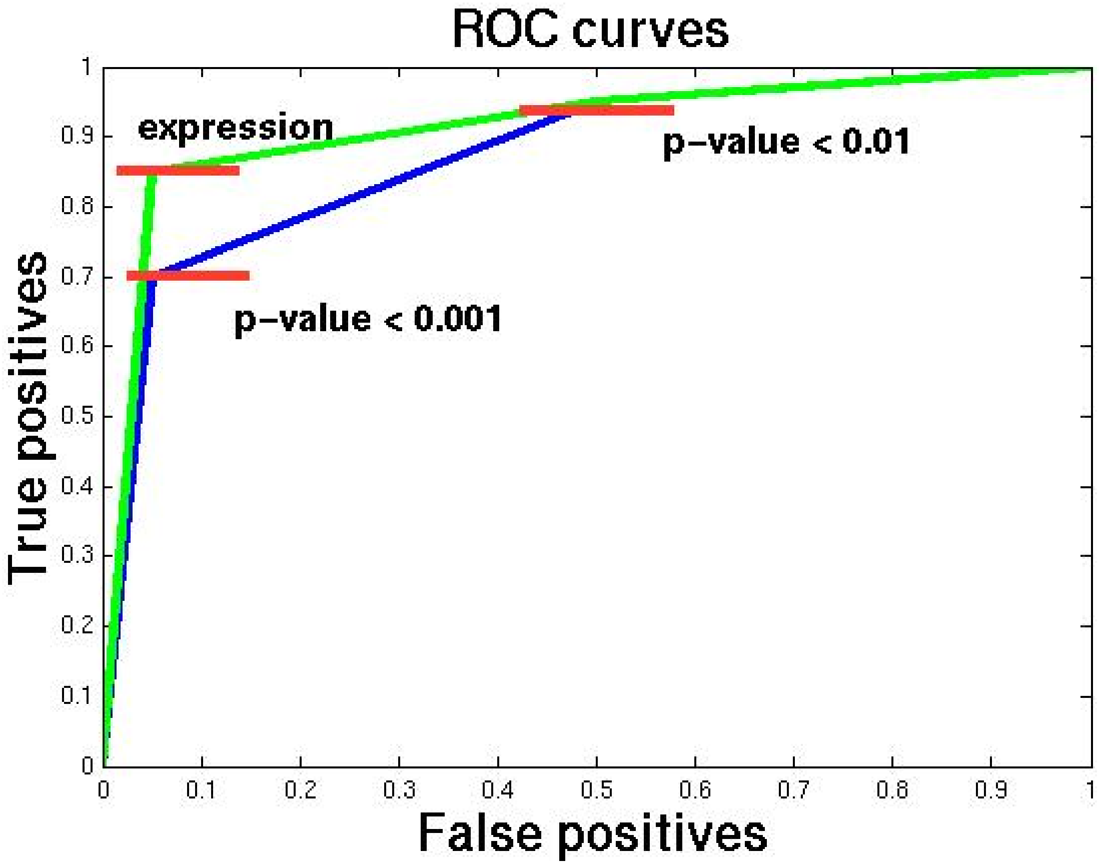
Figure removed for copyright reasons.
See Fig. 1B in Lee, T. I., et al. "Transcriptional Regulatory Networks in Saccharomyces cerevisiae."
Science 298 no. 5594 (25 October 2002): 799-804.

Previous work used an error model for binding data and a p-value cutoff to determine binary relationships.

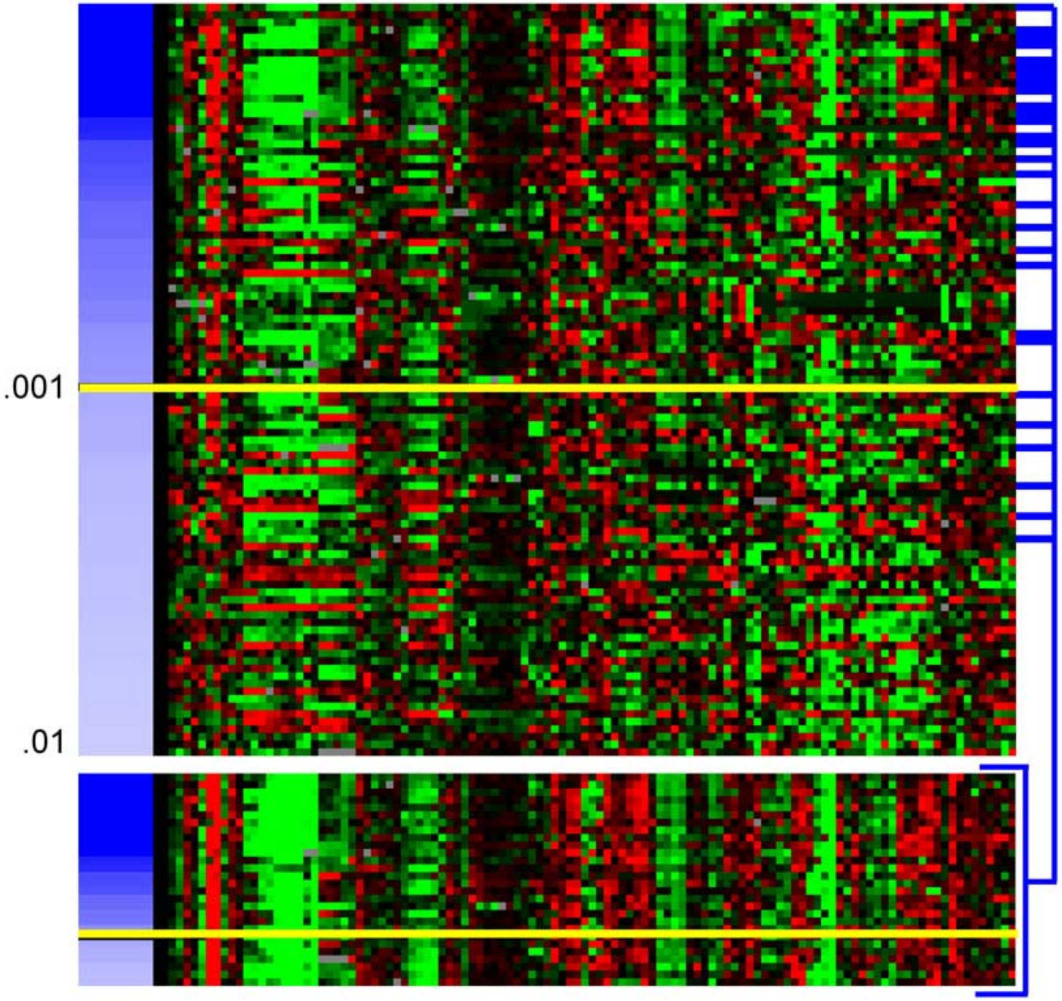
Lee *et al*, *Science*, 2002

Limitations of Binding Data Alone

The p-value cut-off for binding data alone yielded a low false positive rate (5%), but also a low true positive rate (70%).



Limitations of Binding Data Alone



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

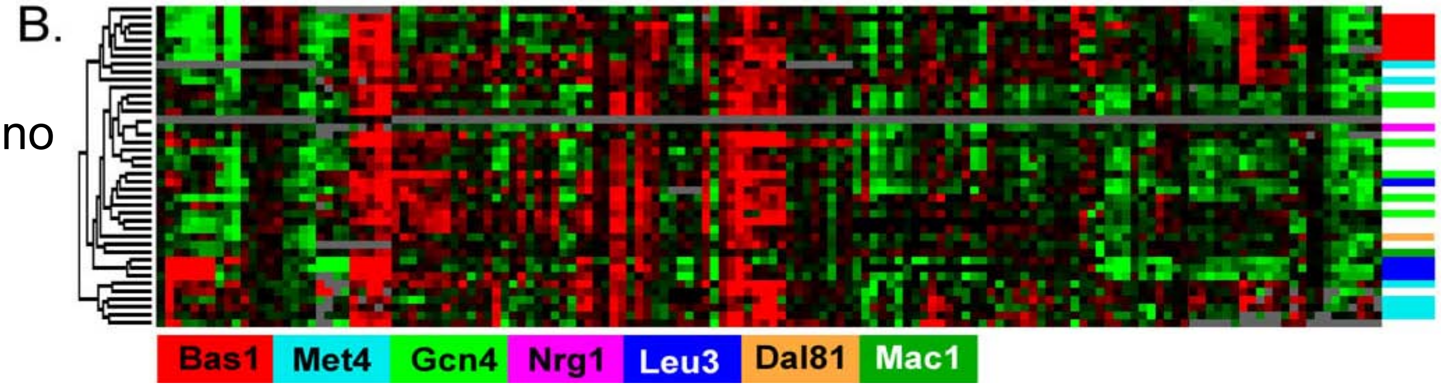
Binding p-values form a continuum – where do you draw the cut-off line?

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).

Bar-Joseph, Z., Georg Gerber et al. “Computational discovery of gene modules and regulatory networks.” Nature Biotechnology 21 (2003): 1337-1342. Used with permission.

Limitations of Expression Data Alone

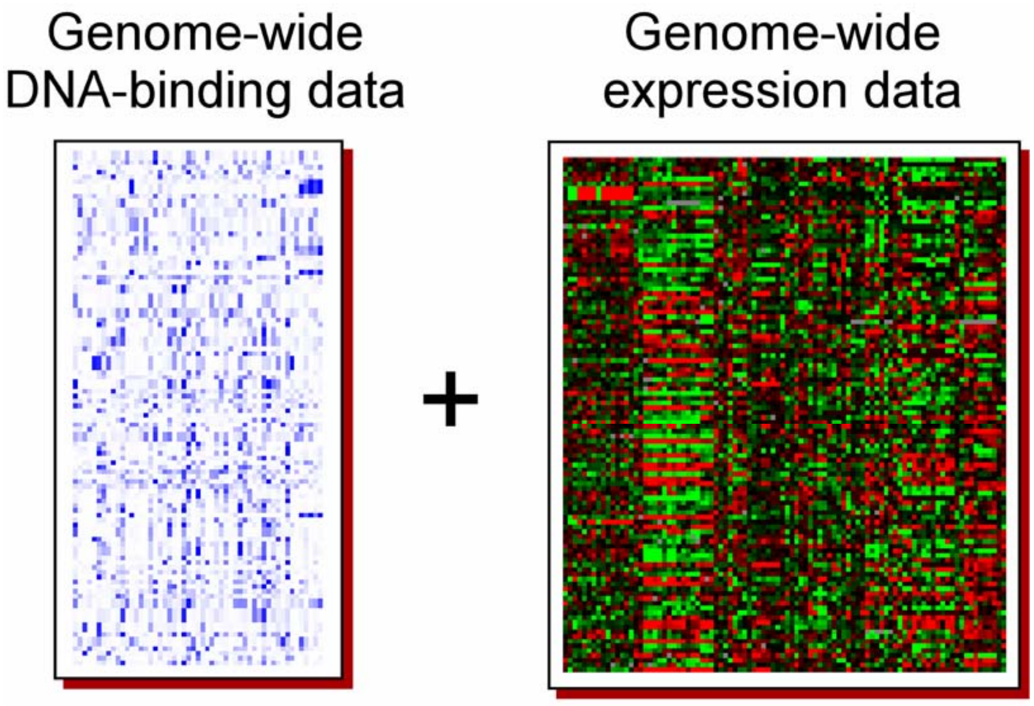
Hierarchical clustering of amino acid synthesis genes



Expression data alone can't effectively distinguish among genes that have similar expression patterns but are under the control of different regulatory networks.

Bar-Joseph, Z., Georg Gerber et al. "Computational discovery of gene modules and regulatory networks." Nature Biotechnology 21 (2003): 1337-1342. Used with permission.

The Genetic RegulAtory Modules (GRAM) Algorithm



Input data to the algorithm

Bar-Joseph, Gerber, Lee and *et al, Nature Biotech.*, 2003

Bar-Joseph, Z., Georg Gerber et al. “Computational discovery of gene modules and regulatory networks.” Nature Biotechnology 21 (2003): 1337-1342. Used with permission.

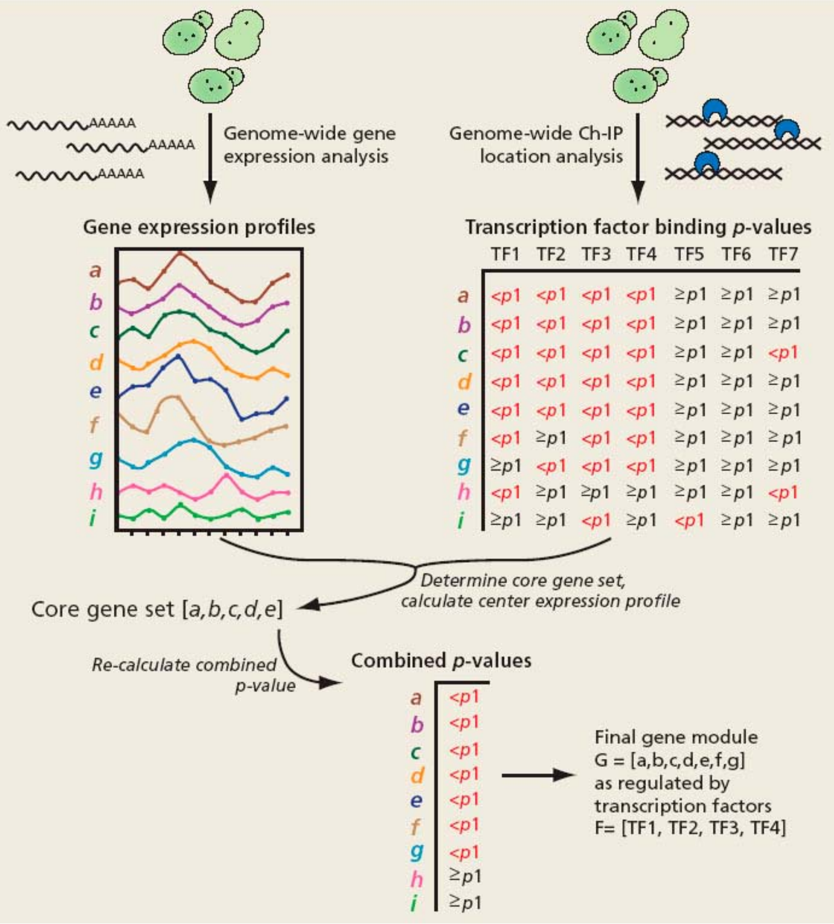
GRAM Algorithm Overview

High-level goal: to discover *gene modules.* Modules help to reduce genetic network complexity without significant loss of explanatory power.

- We define a gene module as a set of genes that is:
  1. co-bound (bound by the same set of TFs, up to limits of experimental noise) **and**
  2. co-expressed (has the same expression pattern, up to limits of experimental noise).
- We interpret this to mean that the genes in the module are co-regulated, and hence likely have a common biological function.

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).


AAAAA
AAAAA
AAAAA
Genome-wide gene expression analysis
Genome-wide Ch-IP location analysis
Gene expression profiles
a
b
c
d
e
f
g
h
i
Determine core gene set, calculate center expression profile
Core gene set [a,b,c,d,e]
Re-calculate combined p-value
Final gene module G = [a,b,c,d,e,f,g] as regulated by transcription factors F= [TF1, TF2, TF3, TF4]

Transcription factor binding *p*-values

	TF1	TF2	TF3	TF4	TF5	TF6	TF7
a	<*p*1	<*p*1	<*p*1	<*p*1	≥*p*1	≥*p*1	≥*p*1
b	<*p*1	<*p*1	<*p*1	<*p*1	≥*p*1	≥*p*1	≥*p*1
c	<*p*1	<*p*1	<*p*1	<*p*1	≥*p*1	≥*p*1	<*p*1
d	<*p*1	<*p*1	<*p*1	<*p*1	≥*p*1	≥*p*1	≥*p*1
e	<*p*1	<*p*1	<*p*1	<*p*1	≥*p*1	≥*p*1	≥*p*1
f	<*p*1	≥*p*1	<*p*1	<*p*1	≥*p*1	≥*p*1	≥*p*1
g	≥*p*1	<*p*1	<*p*1	<*p*1	≥*p*1	≥*p*1	≥*p*1
h	<*p*1	≥*p*1	≥*p*1	≥*p*1	≥*p*1	≥*p*1	<*p*1
i	≥*p*1	≥*p*1	<*p*1	≥*p*1	<*p*1	≥*p*1	≥*p*1

Combined *p*-values

a	<*p*1
b	<*p*1
c	<*p*1
d	<*p*1
e	<*p*1
f	<*p*1
g	<*p*1
h	≥*p*1
i	≥*p*1

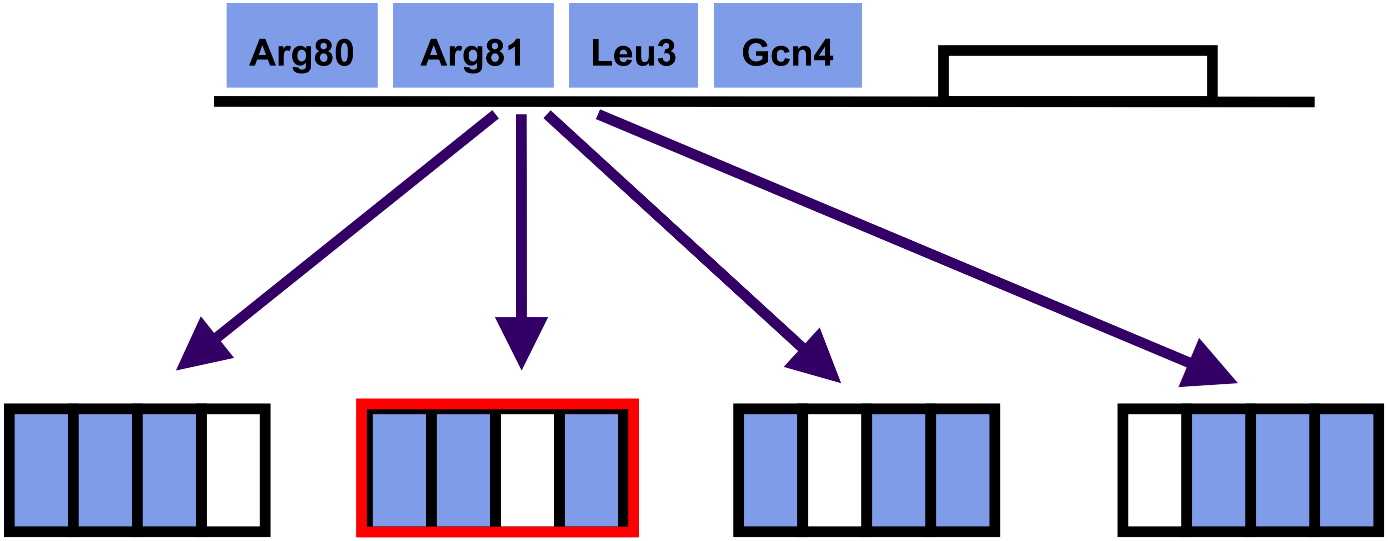
GRAM step 0:

- For each gene *i*:

  Generate all possible subsets of factors that bind to gene *i* with p-value < 0.001. Associate the gene with all the TF subsets via a hash-table.

- Result is the set of all possible binding patterns (as indicated by strict binding p-values), with the corresponding genes mapped to the patterns.

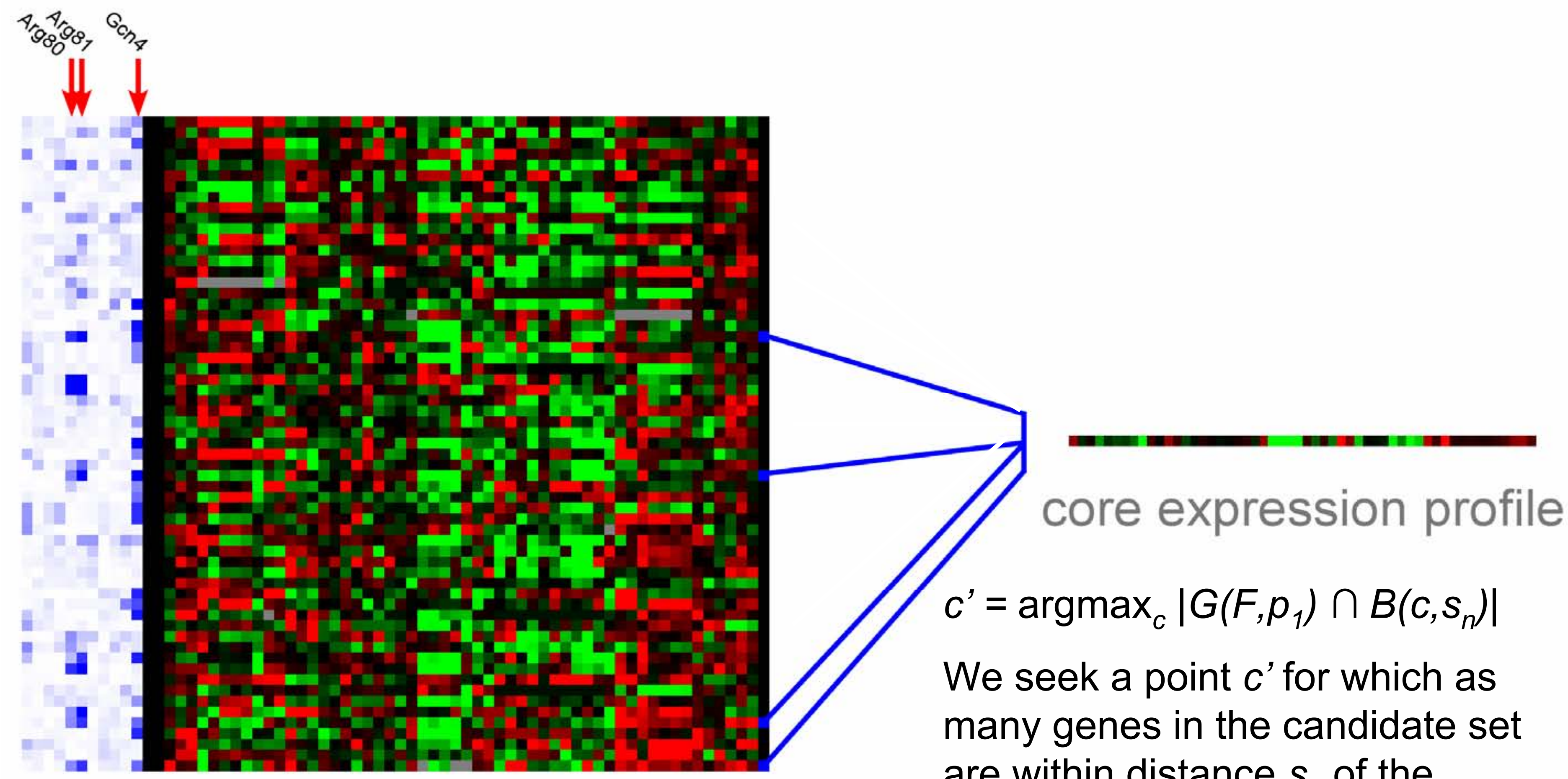
GRAM Algorithm Step 1: exhaustively search all subsets of TFs (starting w/ the largest sets)



	Arg80	Arg81	Leu3	Gcn4
g_1	1	1	0	1
g_2	1	0	1	1
g_3	1	1	1	1
g_4	0	0	0	0
g_5	1	1	1	1
g_6	1	1	0	1

For every set of transcription factors *F*, the genes in $G(F,p_1)$ serve as candidates for a module regulated by the factors in *F*.

GRAM Algorithm Step 2: find a core expression profile for the module

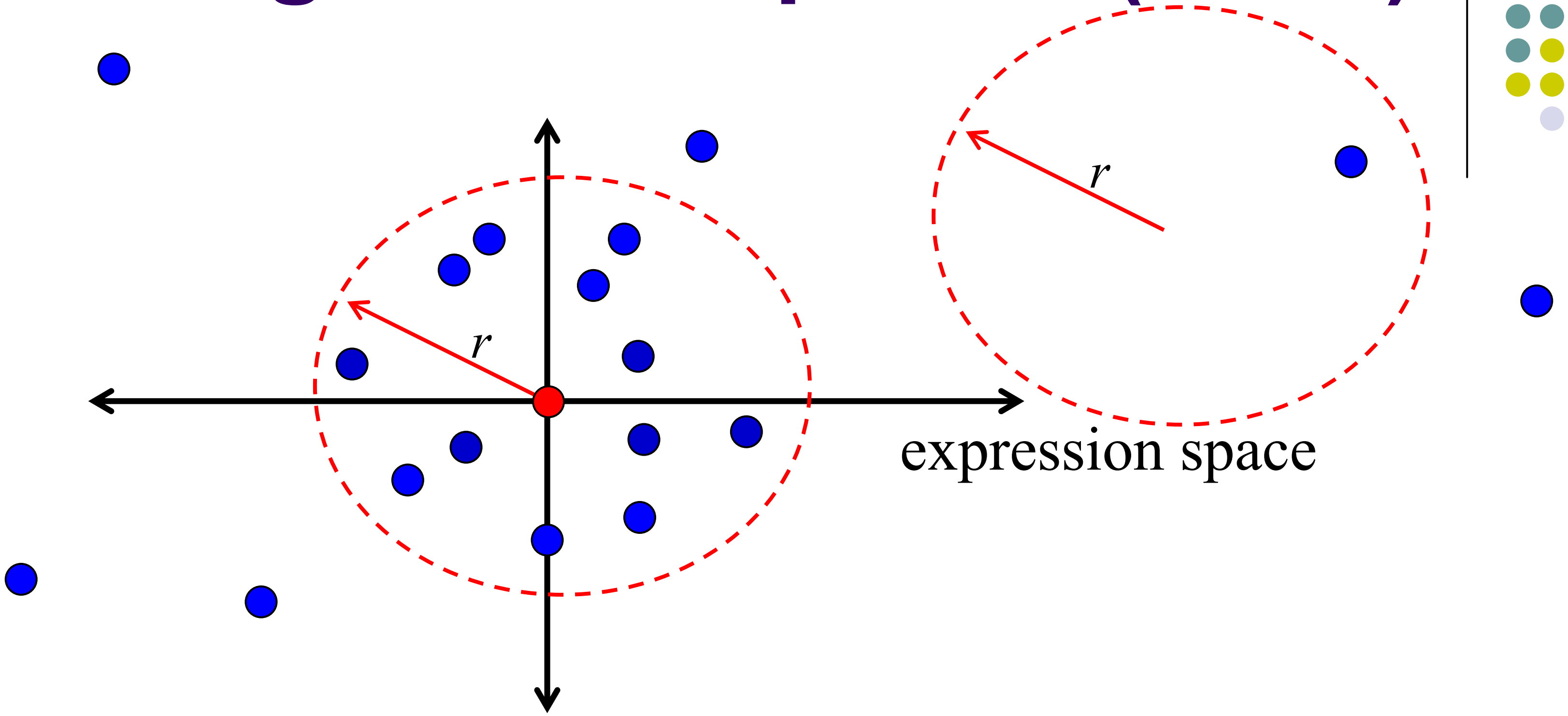


$c' = \text{argmax}_c\ |G(F,p_1) \cap B(c,s_n)|$

We seek a point *c'* for which as many genes in the candidate set are within distance s_n of the point *c'*.

Bar-Joseph, Z., Georg Gerber et al. "Computational discovery of gene modules and regulatory networks." Nature Biotechnology 21 (2003): 1337-1342. Used with permission.

Finding the core profile (cont.)

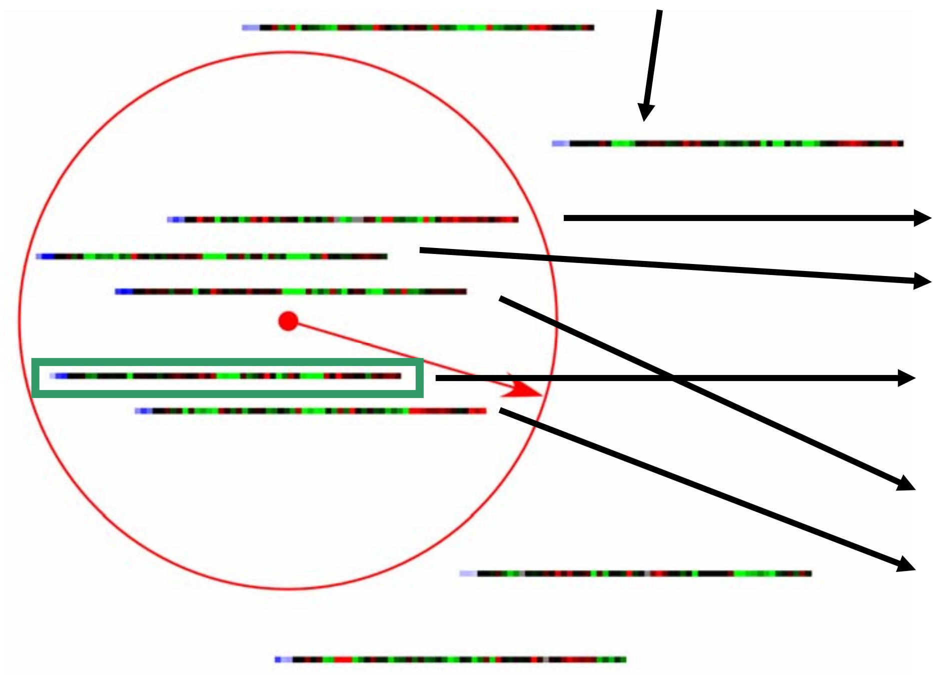


- Consider a set of genes bound by the same TFs.
- The core profile is a point in expression space that describes a ball containing the maximal number of genes within a distance *r*.
- This estimate is **robust**, in the sense that it is insensitive to outliers (think of a median versus a mean).
- To compute it exactly requires an O(2^n) algorithm (n=# of genes in set).
- Using results from computational geometry, we get an O(n^3) approximation algorithm (with provable error bounds).

GRAM Algorithm Step 3: add/remove genes

2. Remove genes with significantly far expression profiles

1. Include genes that are close and are bound by same TFs (binding p-value < 0.001)

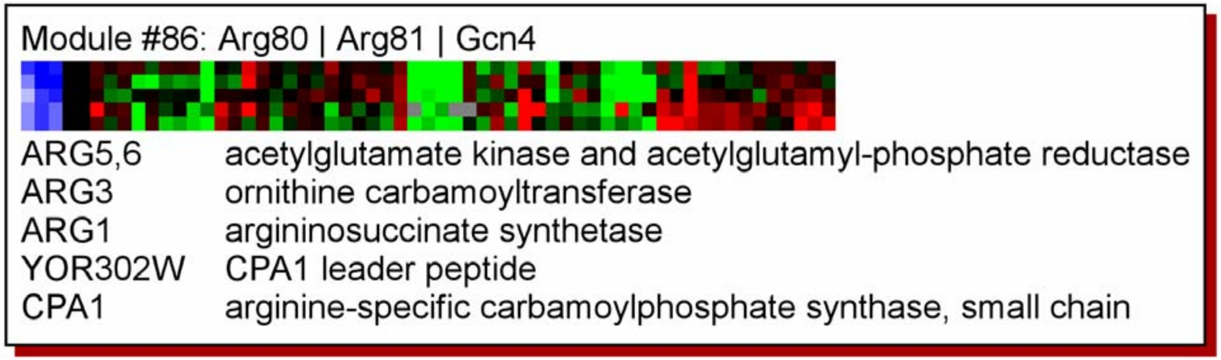


	Arg80	Arg81	Leu3	Gcn4	
g_1	.0004	.00003	.33	.0004	√
g_2	.00002	0.0006	.02	.0001	√
g_3	.0007	.002	.15	.0002	√
g_4	.007	.2	0.04	.7	x
g_5	.00001	.00001	.0001	.0002	√
g_6	.00001	.00007	.5	.0001	√

Expanded set = $G(F,p_2) \cap B(c',s_n)$, where $p_2 > p_1$.

3. Relax the binding threshold/ add genes with significantly close expression profiles

GRAM Algorithm: final module



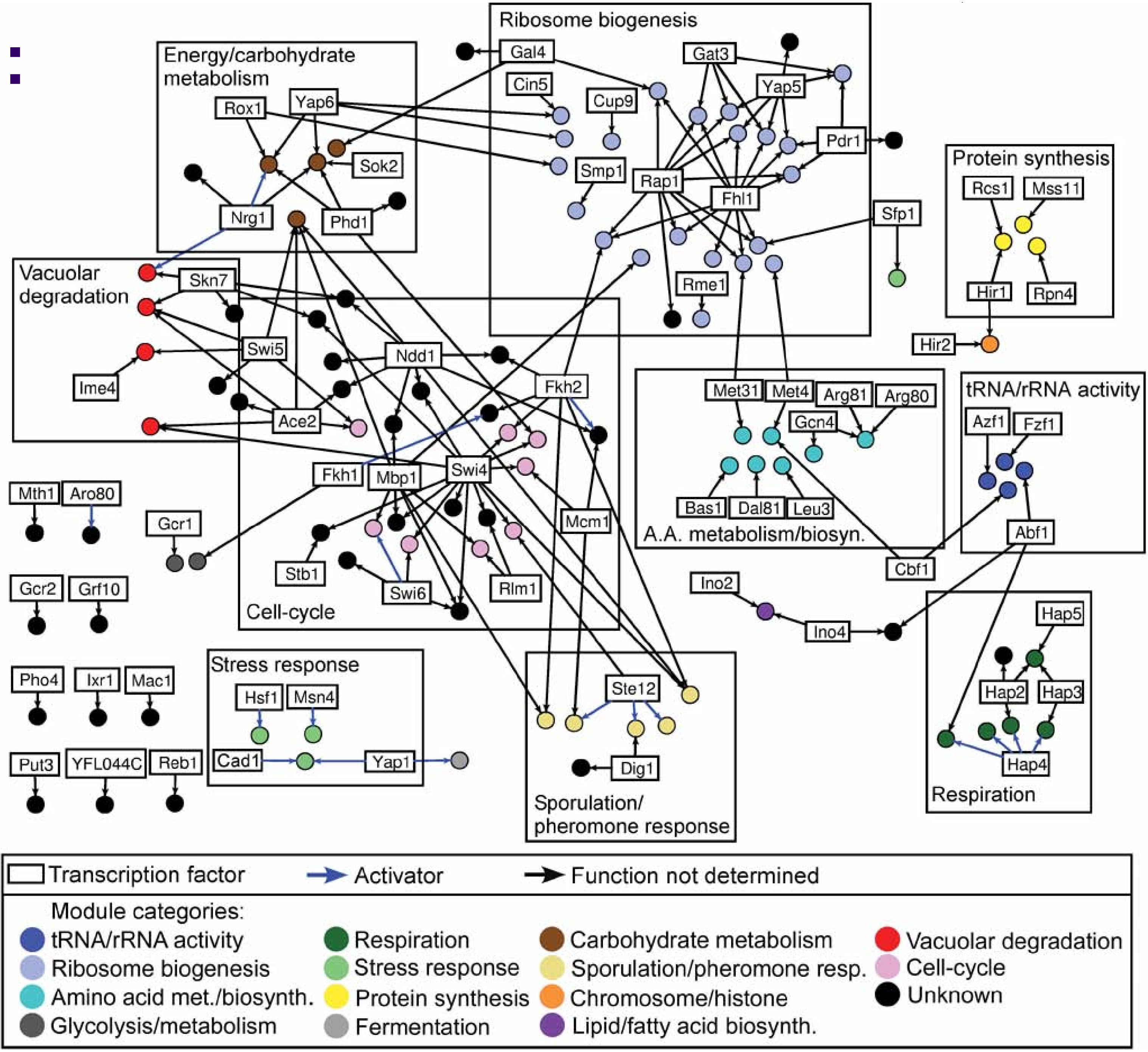
Module #86: Arg80 | Arg81 | Gcn4

ARG5,6	acetylglutamate kinase and acetylglutamyl-phosphate reductase
ARG3	ornithine carbamoyltransferase
ARG1	argininosuccinate synthetase
YOR302W	CPA1 leader peptide
CPA1	arginine-specific carbamoylphosphate synthase, small chain

- The module is:
  1. co-bound (bound by the same set of TFs, up to limits of experimental noise) **and**
  2. co-expressed (has the same expression pattern, up to limits of experimental noise).
- We interpret this to mean that the genes in the module are co-regulated, and hence likely have a common biological function.

Results:

Rich Media Modules



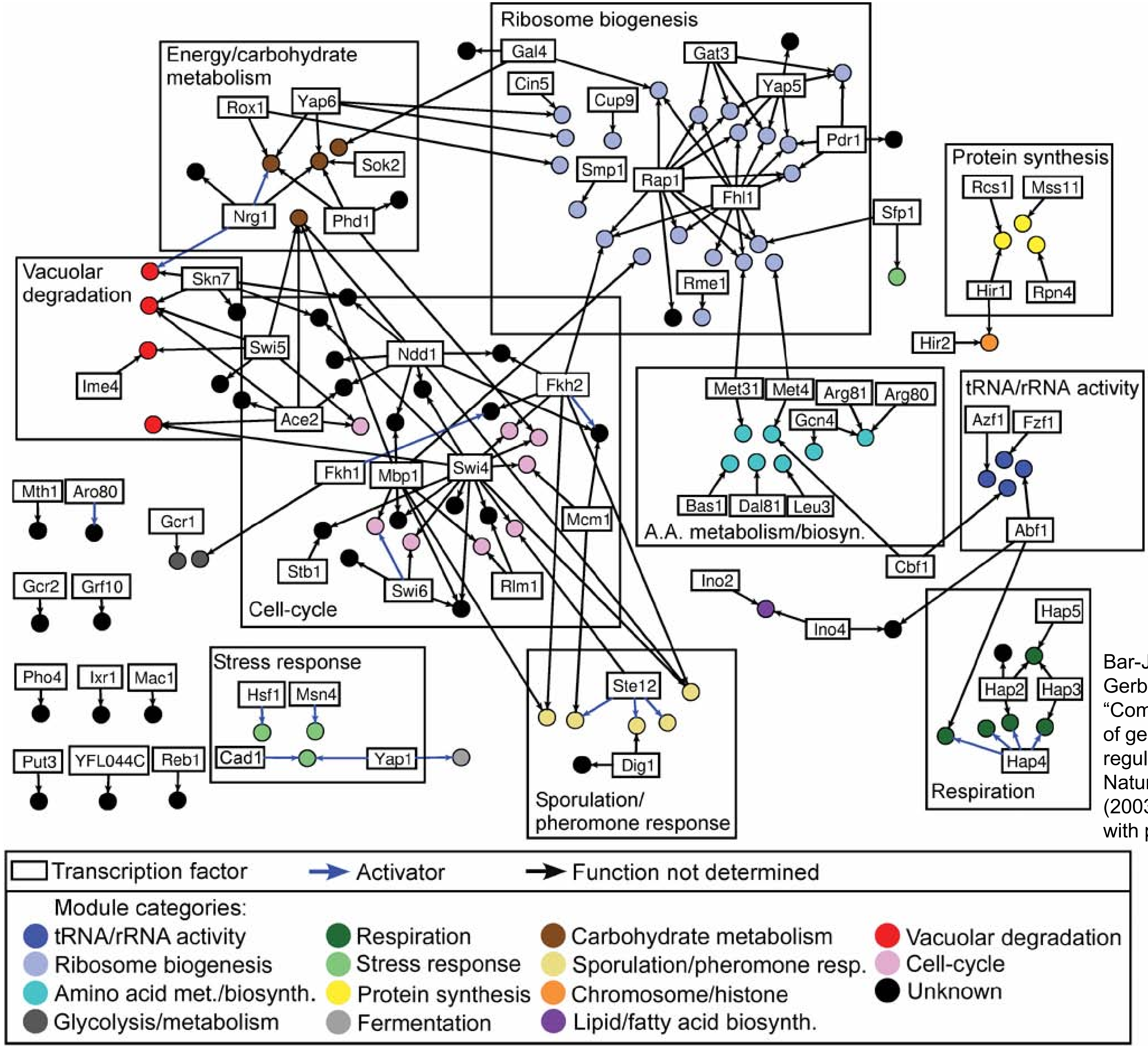
Bar-Joseph, Z., Georg Gerber et al. "Computational discovery of gene modules and regulatory networks." Nature Biotechnology 21 (2003): 1337-1342. Used with permission.

Rich Media Gene Modules Network Results

- Binding data for 106 transcription factors profiled in rich media conditions (YPD).
- Over 500 expression experiments in a variety of conditions.
- Discovered 106 modules ranging in size from 52 genes to 5; modules are controlled by 68 factors and contain 655 genes.

Rich Media Gene Modules Network: Identifying Activators

- Activator defined by:
  - TF regulates module.
  - TF expression profile is positively correlated with core profile of module.
- Statistical significance of activator relationships by computing correlation coefficients between all transcriptional regulators studied and all gene modules and taking the 5% positive tail of the distribution.

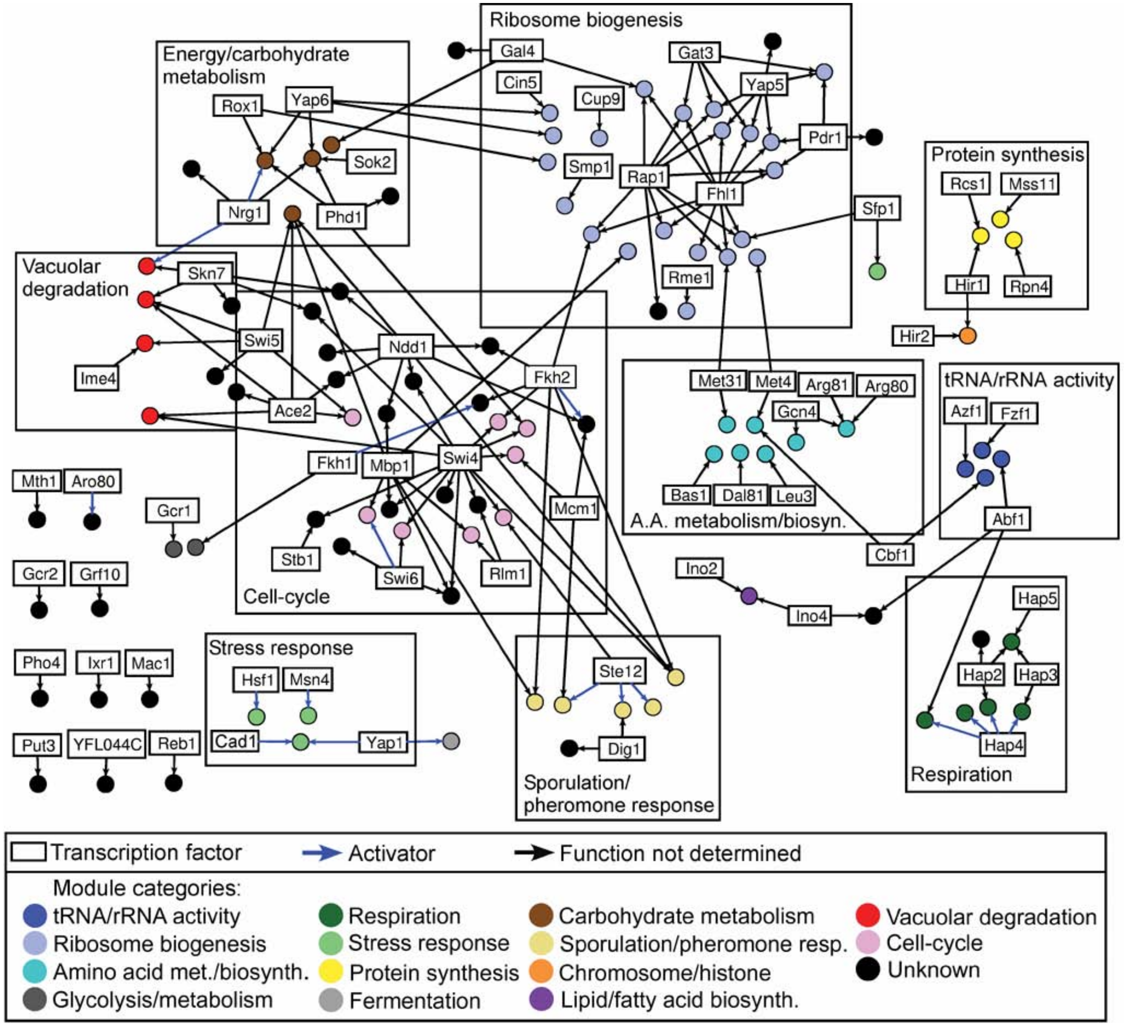


Bar-Joseph, Z., Georg Gerber et al. “Computational discovery of gene modules and regulatory networks.” Nature Biotechnology 21 (2003): 1337-1342. Used with permission.

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, involved in 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

We found a network…now what? How can we validate our results?



Bar-Joseph, Z., Georg Gerber et al. "Computational discovery of gene modules and regulatory networks." Nature Biotechnology 21 (2003): 1337-1342. Used with permission.

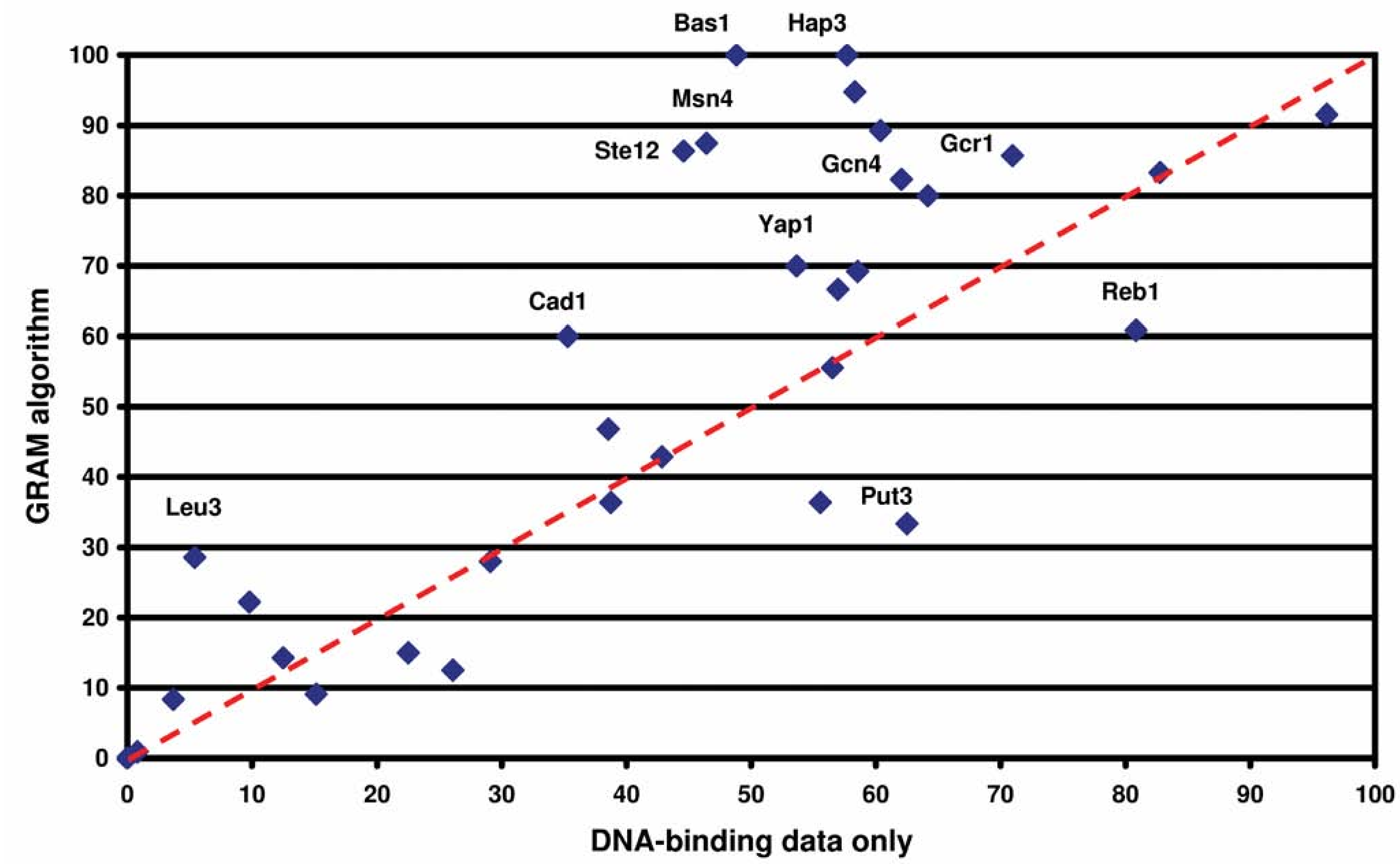
Validation Ideas

- 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.

Validation: Motifs From TRANSFAC

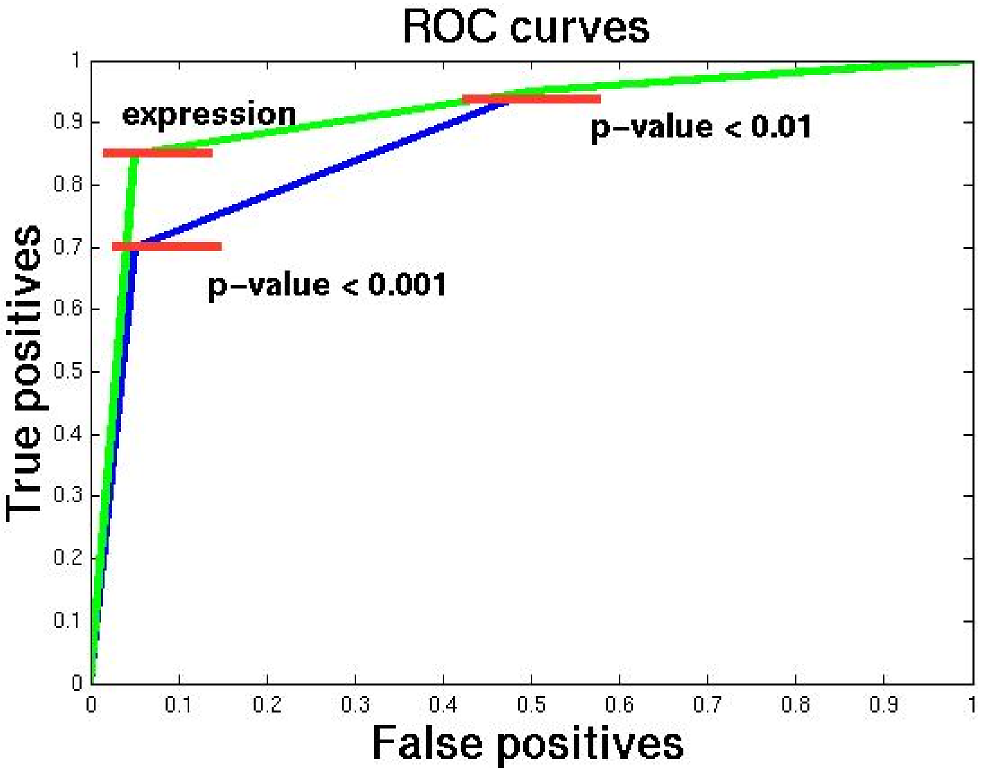


We identified 34 TFs w/ well-characterized motifs in TRANSFAC and looked at enrichment for the motifs in modules versus gene lists obtained from binding data alone.

Bar-Joseph, Z., Georg Gerber et al. “Computational discovery of gene modules and regulatory networks.” Nature Biotechnology 21 (2003): 1337-1342. Used with permission.

Validation: Biological Experiments to Verify Error Rate

- Did we improve the true positive rate without significantly affecting the false positive rate?



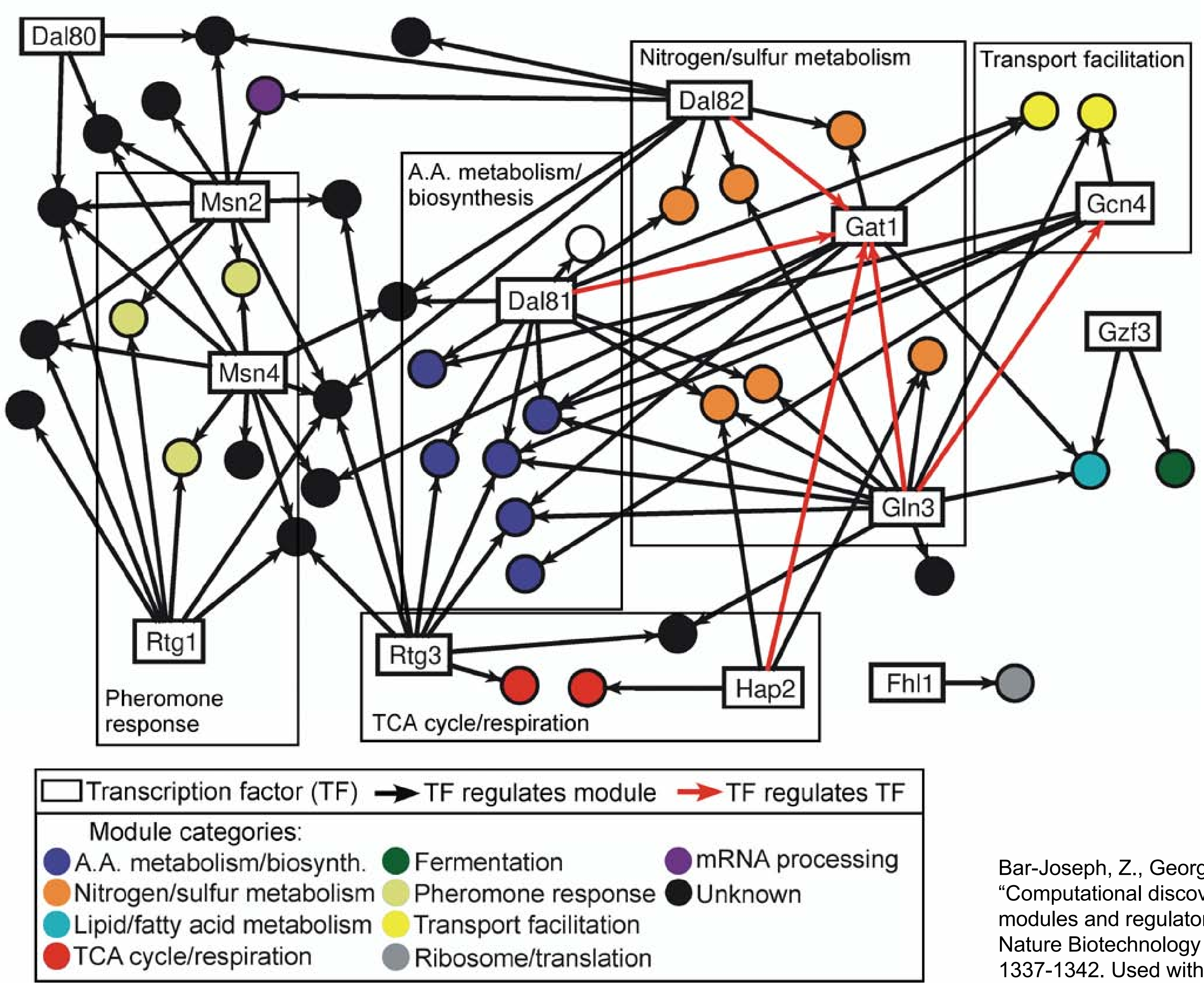
Validation: Biological Experiments to Verify Error Rate

- Added interactions not predicted by binding data alone: 627 out of 1560 unique regulator-gene interactions (40%) predicted by GRAM had binding p-values > .001.
- Performed gene-specific chromatin-IP experiments for the factor Stb1 and 36 genes.
  - Profiled genes were picked randomly from the full set of yeast genes, with representatives selected from four p-values ranges.
  - Three additional genes were determined to be bound by Stb1 that had p-values between .01 and .001.
  - GRAM identified *all* three genes as bound by Stb1 without adding any additional genes that were not detected in the gene-specific chromatin-IP experiments.

More New Experiments: Rapamycin Gene Modules Network

- How will GRAM perform on new binding data?
  - Generated new binding data for 14 transcription factors profiled in rapamycin.
  - 39 gene modules containing 317 unique genes and regulated by 13 transcription factors; added 119 genes (38%) with p-value > .001.
  - Many features of the network consistent with the literature; found modules containing genes belonging to relevant MIPS categories.
- Can we analyze this smaller second condition network in more detail to discover new biology?

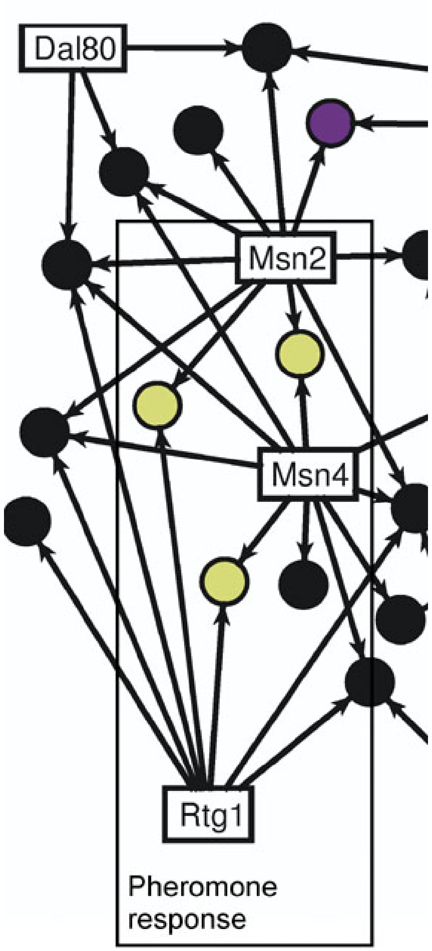
Rapamycin modules network



Bar-Joseph, Z., Georg Gerber et al. “Computational discovery of gene modules and regulatory networks.” Nature Biotechnology 21 (2003): 1337-1342. Used with permission.

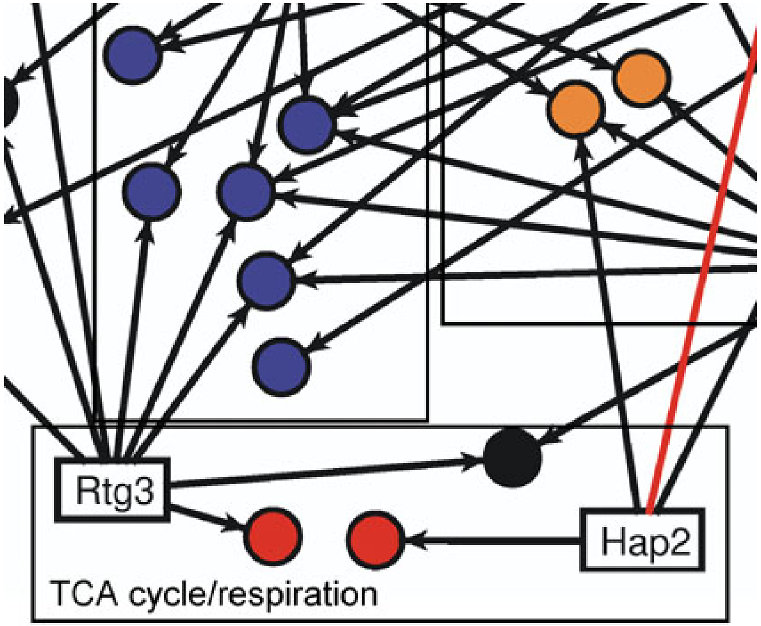
Unexpected Findings in the Rapamycin Regulatory Network: New Roles for TFs

- Msn2 and Msn4 typically characterized as general stress response TFs; found they control five modules associated with pheromone response.



New Roles for TFs (cont.)

- Rtg3 generally thought to regulate directly genes of the TCA cycle and indirectly contribute to nitrogen metabolism; results suggest Rtg3 may directly regulate genes involved in nitrogen metabolism.
- Hap2 part of well-characterized complex that regulates respiration; results suggest Hap2 also involved in regulating nitrogen metabolism (there's a small amount of support in the literature for this).

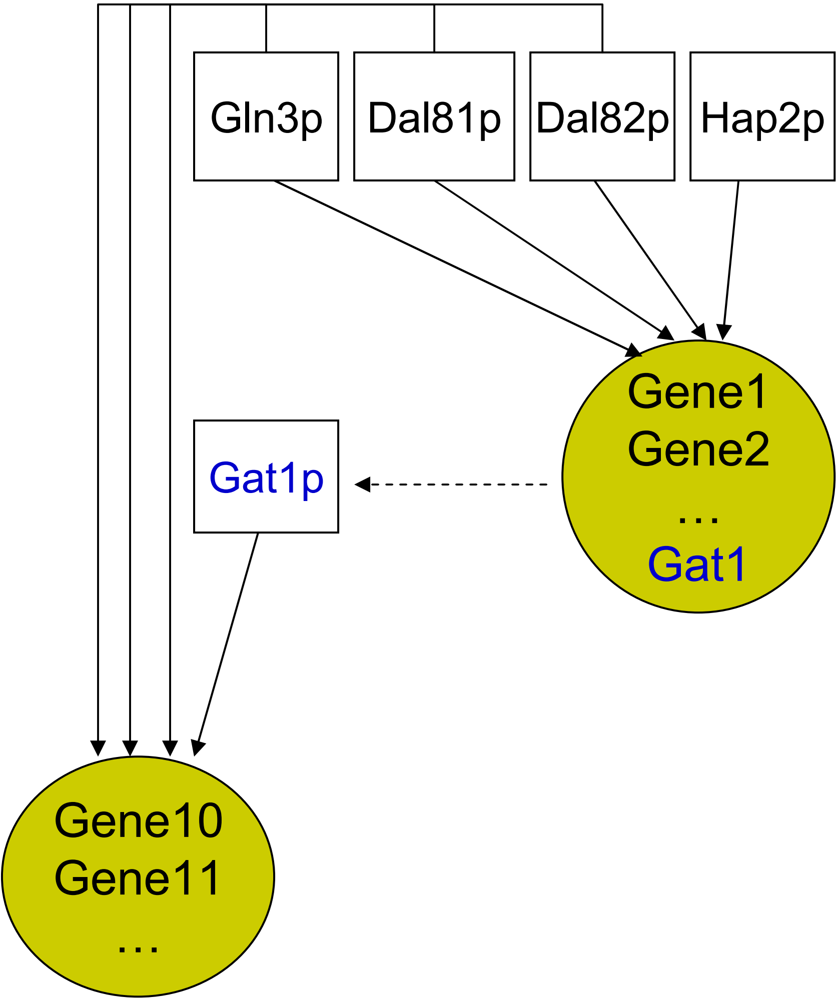


Unexpected Findings in the Rapamycin Regulatory Network: Network Complexities/Module Interactions

Figure removed for copyright reasons.
See Fig. 3 in Lee, T. I., et al. “Transcriptional Regulatory Networks in Saccharomyces cerevisiae.” Science 298 no. 5594 (25 October 2002): 799-804.

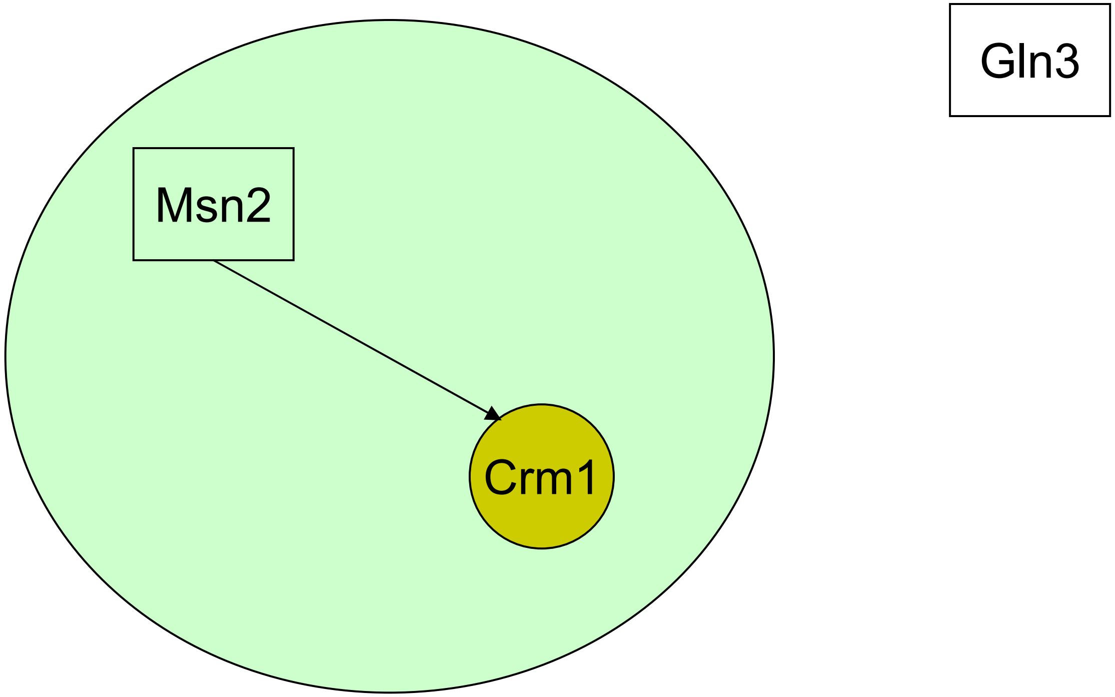
Unexpected Findings: Feed-forward Transcriptional Regulation

- Gat1 (a general activator of nitrogen responsive genes) contained in several modules along with genes involved in nitrogen metabolism.
- These modules are bound by Dal81, Dal82, Gln3 and Hap2.
- Gat1 also binds several gene modules along with Dal81, Dal82, and Gln3.
- Could be used for amplification, delay, etc.



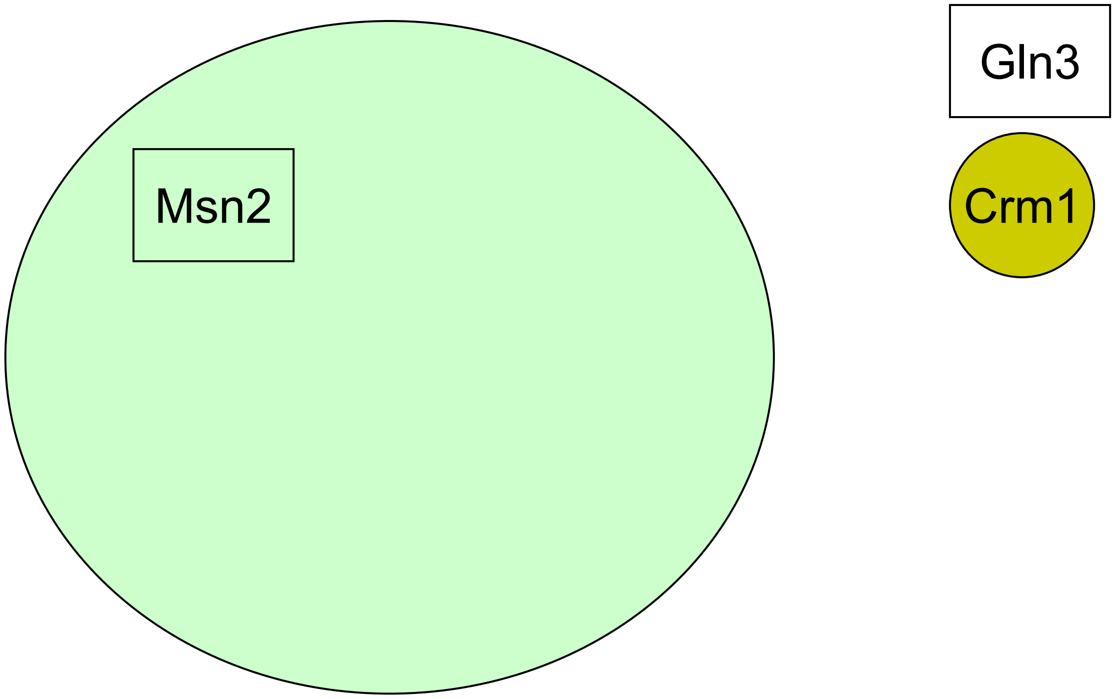
Unexpected Findings: Complex Module Interactions

- Non-transcriptional (or mixed) interactions between modules:
  - Msn2 binds to a module containing Crm1 (a nuclear export factor critical in allowing Gln3 to move from the cytoplasm to the nucleus). Suggests that Msn2 activation after rapamycin treatment may act to enhance or enable a step in Gln3 activation.



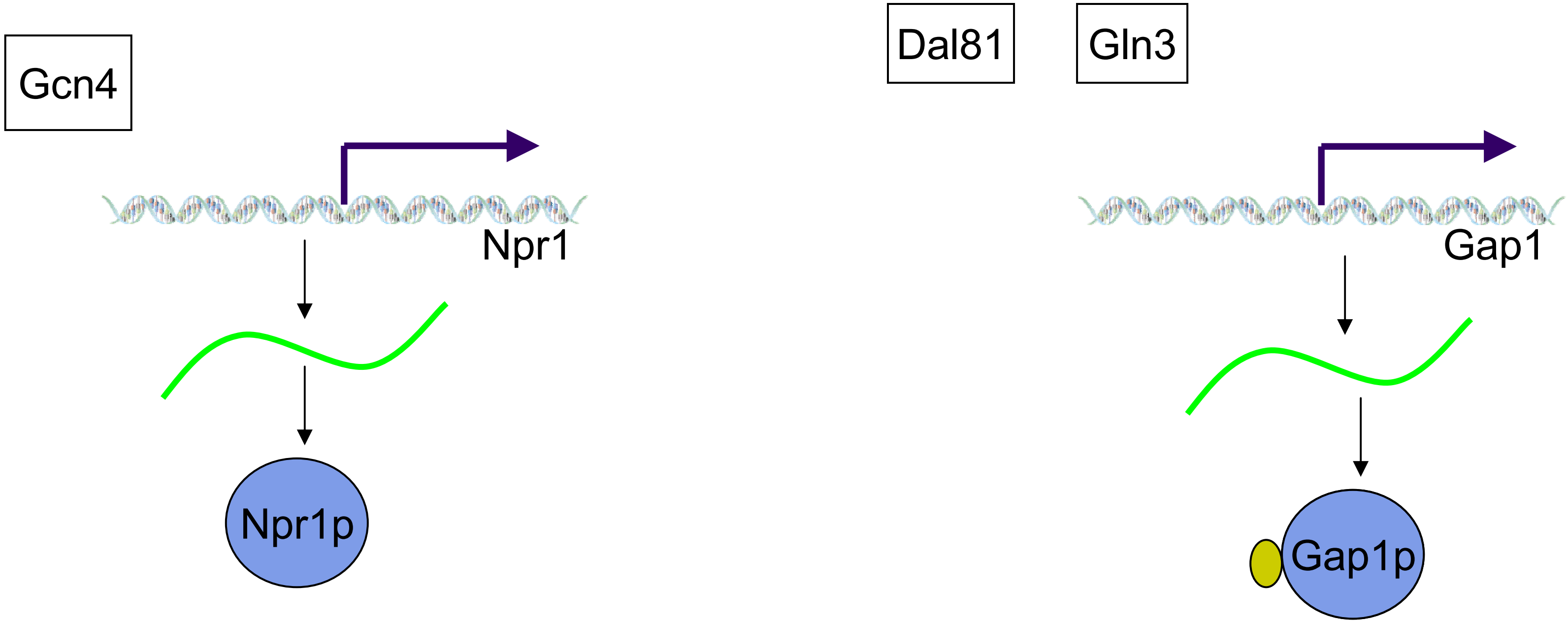
Unexpected Findings: Complex Module Interactions

- Non-transcriptional (or mixed) interactions between modules:
  - Msn2 binds to a module containing Crm1 (a nuclear export factor critical in allowing Gln3 to move from the cytoplasm to the nucleus). Suggests that Msn2 activation after rapamycin treatment may act to enhance or enable a step in Gln3 activation.



Unexpected Findings: Complex Module Interactions

- Gcn4 binds to a module containing Npr1 (serine/threonine protein kinase known to promote the function of the general permease Gap1).
- Gap1 contained in a module regulated by Dal81 and Gln3. Suggests regulatory connections in which Gap1 is transcriptionally regulated by Dal81/Gln3, Npr1 is transcriptionally regulated by Gcn4, and then Gap1 is non-transcriptionally activated by Npr1.

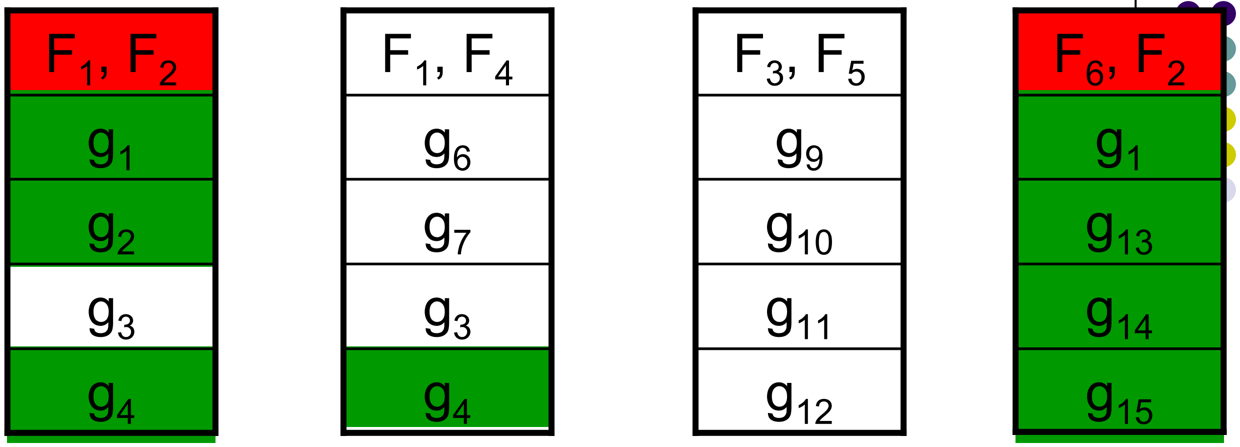


Sub-network Discovery and Dynamics: The Cell-Cycle

- We combined GRAM with our continuous representation and alignment algorithms to construct a dynamic model for the cell-cycle.
- The algorithmic steps were:
  - Identify genes relevant to the sub-system.
  - Identify factors controlling these genes and the modules involved.
  - Build a dynamic model for the activation of the modules by the identified factors.

Sub-Networks Discovery Algorithm

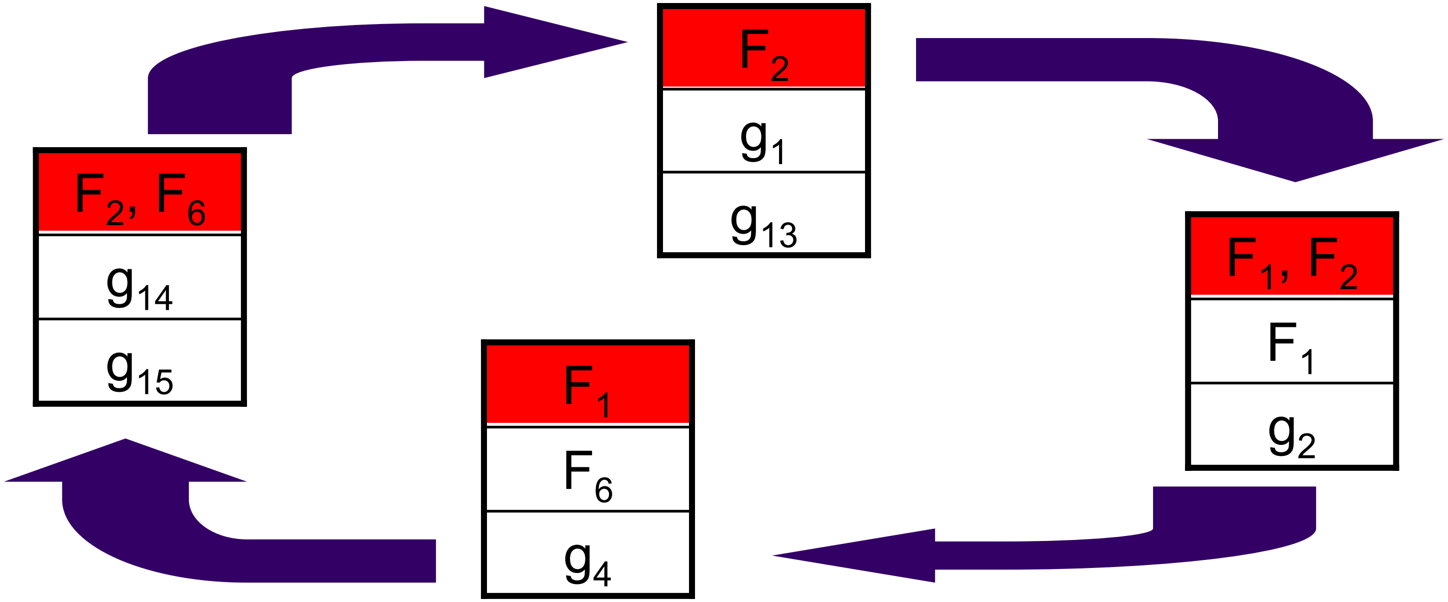
1.



F_1, F_2	F_1, F_4	F_3, F_5	F_6, F_2
g_1	g_6	g_9	g_1
g_2	g_7	g_{10}	g_{13}
g_3	g_3	g_{11}	g_{14}
g_4	g_4	g_{12}	g_{15}
p-value: 10^{-4}	p-value: 0.2	p-value: 0.7	p-value: 10^{-6}

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

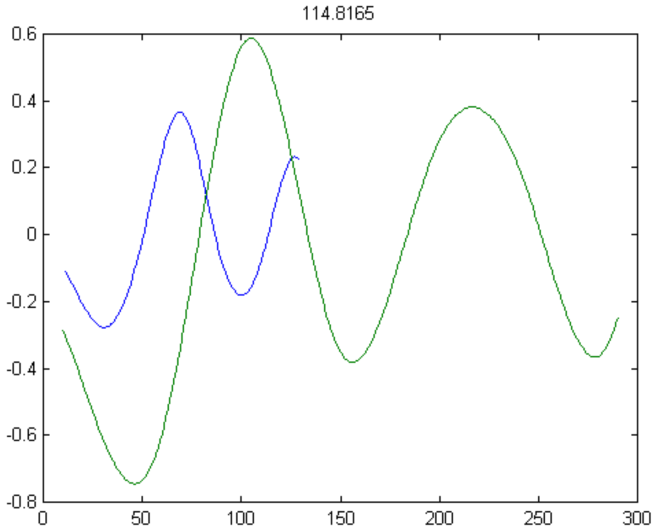
2.

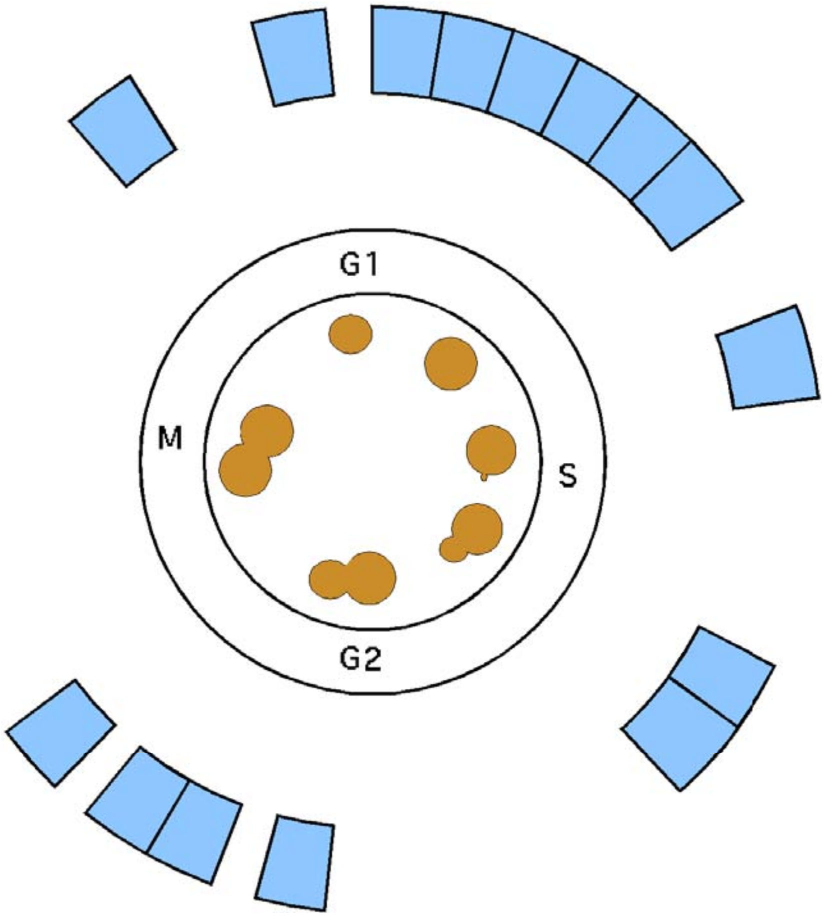


Assembly of the Cell Cycle Transcriptional Regulatory Network

Blue boxes: gene modules

Modules were fit with splines, and then aligned to a reference module at M/G1 point using our continuous alignment algorithm.





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

Figure removed for copyright reasons.
See Fig. 4 in Lee, T. I., et al. “Transcriptional Regulatory Networks in Saccharomyces cerevisiae.”
Science 298 no. 5594 (25 October 2002): 799-804.

Lee *et al, Science,* 2002

Doing Computational Biology Research: Practical Take-aways

- Focus on **biologically relevant** problems.
- Think about how you're going to **validate** your findings from day one!
  - Challenging, because "ground truth" is not always clear and new discovery is important.
- Don't neglect good/creative **visualization** – this is critical for communicating with biologists.
- **Collaborate** with biologists!
  - Can be challenging, because different language, style of thinking, knowledge-base, priorities, etc.

Acknowledgements

- Ziv Bar-Joseph
- Ernest Fraenkel
- David Gifford
- Ben Gordon
- Tommi Jaakkola
- Tony Lee
- Nicola Rinaldi
- François Robert
- Jane Yoo
- Rick Young
- Itamar Simon
- Dacheng Zhao

Algorithmic Details

Some notation…

- Let e_i denote an expression vector and b_i a vector of binding p-values for gene i
- Let $T(i,p)$ denote the set of all transcription factors that bind to gene i with p-value less than p, i.e., the list of indices j such that $b_{ij} < p$.
- Let $F \subseteq T(i,p)$ denote a subset of the transcription factors that are bound to gene i.
- Let $G(F,p)$ be the set of all genes such that for any gene $i \in G(F,p)$, $F \subseteq T(i,p)$, i.e., all genes to which all the factors in F bind with a given significance threshold.

More notation…

- Denote an open ball w/ center *c* and radius *r* by *B(c,r)*, e.g., gene *i* ε *B(c,r)* iff $d(e_i,c) < r$ (where *d* is a distance function in expression space). If we define *c* and *r* appropriately, this indicates a set of co-expressed genes.
- Consider $G(F,p_1) \cap B(c,s_n)$. This is the set of genes that are bound by the set of transcription factors *F* (with p-value threshold p_1) and for which the genes' expression vectors are within a distance s_n of the point *c*.

Details on finding the core expression profile

- Assume we can define a **co-expression threshold** s_n (more on how we do this later). This means that for genes i,j s.t. $i,j \, \varepsilon \, B(c,s_n)$, this implies that i and j are co-expressed.
- Suppose we are given a set V of arbitrary genes. We want to find c = $\mathrm{argmax}_{c'} \, |B(c',s_n) \cap V|$ (this will give us the biggest subset of V s.t. all the genes in it are enclosed in a ball with radius s_n).
- This method for finding co-expressed genes is **robust**, in that the subset found is not influenced by outliers (genes outside the co-expression threshold).

Finding the core expression profile (cont.)

- The naïve method for finding c = $\operatorname{argmax}_{c'}$ $|B(c',s_n) \cap V|$, would be to take all possible subsets of *V*, compute their centers c', find all the genes in *V* within a distance s_n of each center, and take *c'* that gives the biggest set. This is $O(2^{|V|})$, which is impractical.
- We can use a result from computational geometry to get an approximation algorithm that's $O(|V|^3)$.

Finding the core expression profile (cont.)

Theorem (adapted from Badoiu and Clarkson 2002):

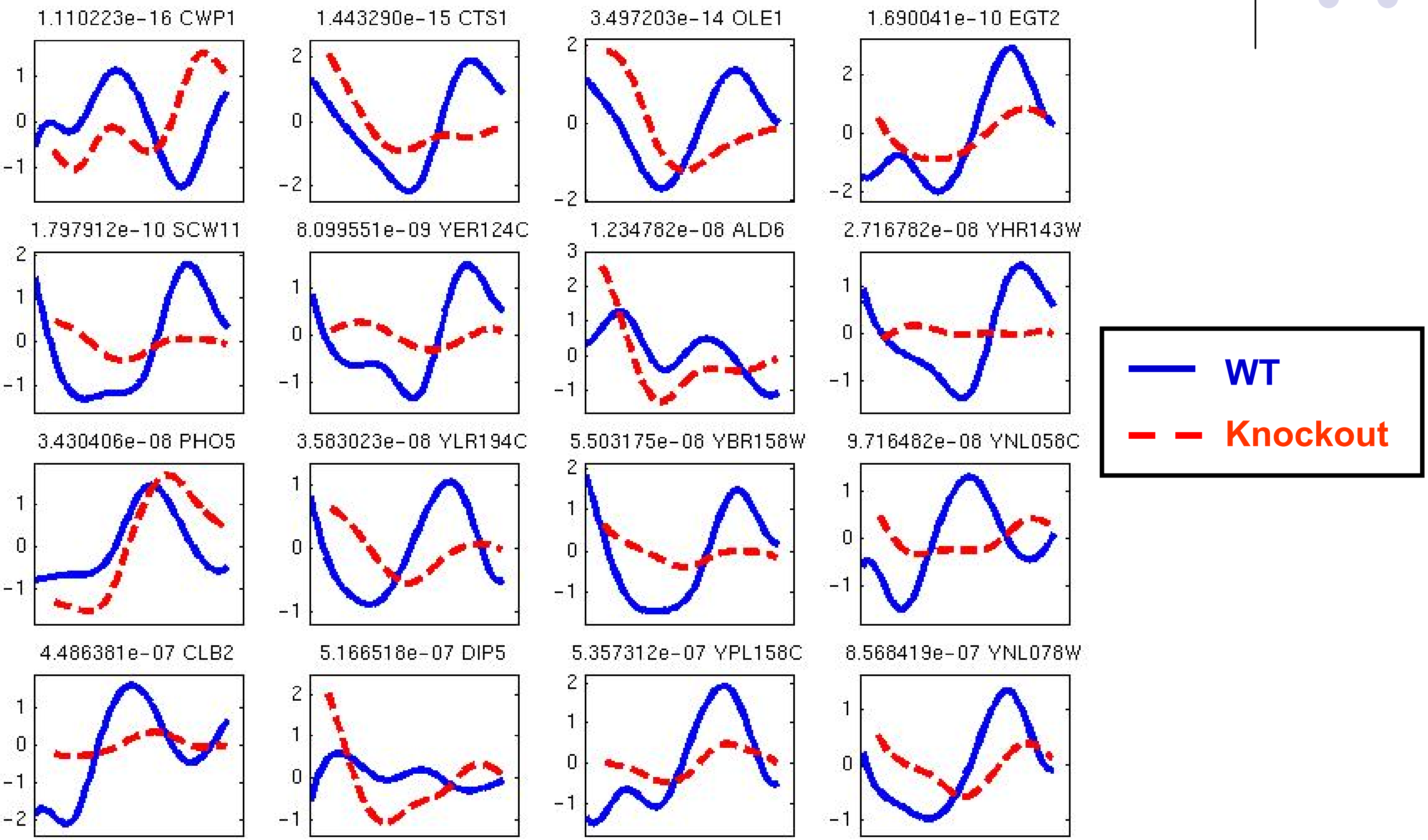
Given a set $U = B(c,r) \cap V$ and $\alpha < |U|$, there exists a set $U' \subseteq U$ with center c' and $|U'| = \alpha$, s.t. for all $i \varepsilon U$, $d(e_i,c') < (1 + 2/\alpha)\, r$.

- *U* is the maximal set we're looking for (the biggest possible set of *co-expressed genes* embedded in a bigger set of genes *V*). We can (approximately) find *U* by an $O(|V|^3)$ algorithm:
  - Let *U'* range over each triplet of genes in *V*.
  - Find the center *c'* of *U'* and find $\tilde{U} = B(c',s_n) \cap V$ (all genes *i* in *V* s.t. $d(e_i,c') < s_n$).
  - Take the set $\tilde{U}$ s.t. $|\tilde{U}|$ is maximal.
- The set $\tilde{U}$ approximates *U*, the maximal subset of co-expressed genes. That is, we' re guaranteed to find the maximal subset of co-expressed genes with radius at least $3s_n/5$.

Finding the core expression profile (cont.)

- How do we compute s_n (a co-expression threshold that depends on the number of genes in V, where $n = |V|$)?
- Let $f(V,r) = max_{B(c,r) \subseteq V} |B(c,r) \subseteq V|$ (the size of the maximal ball contained in V w/ radius r).
- Consider $P(f(V,r) \geq m \mid |V| = n, r)$ (the probability that the maximal ball contained in V w/ radius r will have m or more genes, considered over all sets of genes V with n elements).
- We can define $s_n = argmax_r\, P(f(V,r) \geq m \mid |V| = n, r) \leq \beta$, where β is some threshold (e.g., 0.05) and m is the minimum module size (e.g., m=5).
- Intuitively, if we're given a set V of n randomly selected genes, and we find the maximal subset of these genes within a radius s_n of each other, only 5% of the time will this subset consist of 5 or more genes.
- We can determine s_n by sampling random sets V of size n, going over all triplets of genes, computing their centers, and finding the distance to the fifth closest gene in V. We then take the minimum such distance. This will give us a distribution of distances. We take s_n as the 5% value.

Results for the Fkh1/2 Knockout



See Bar-Joseph, Gerber, and *et al, PNAS,* 2003.