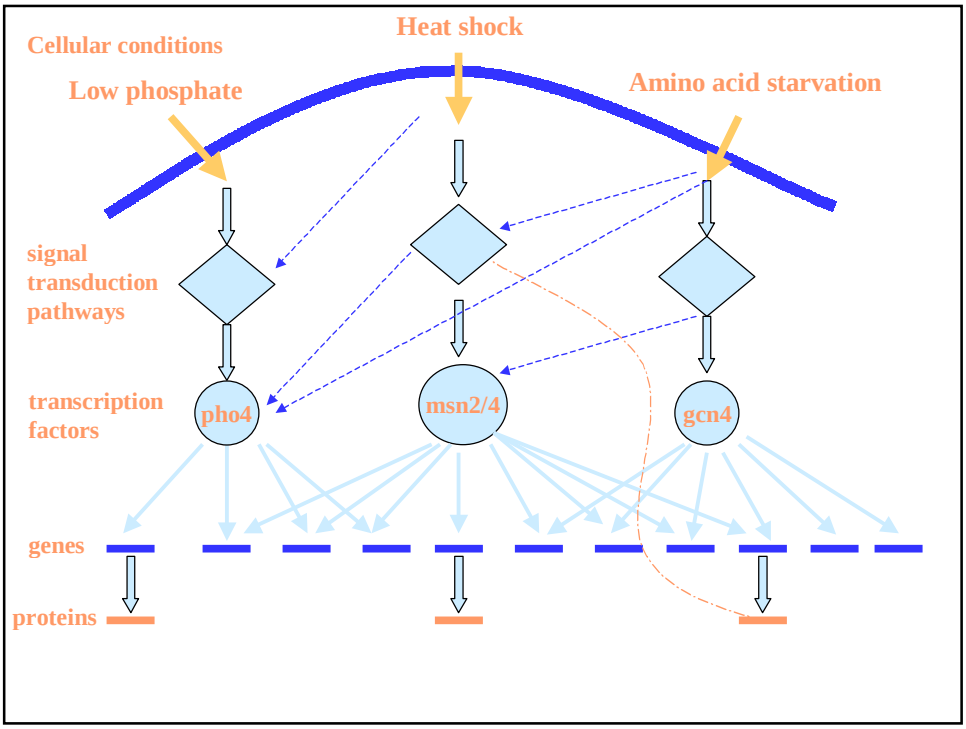
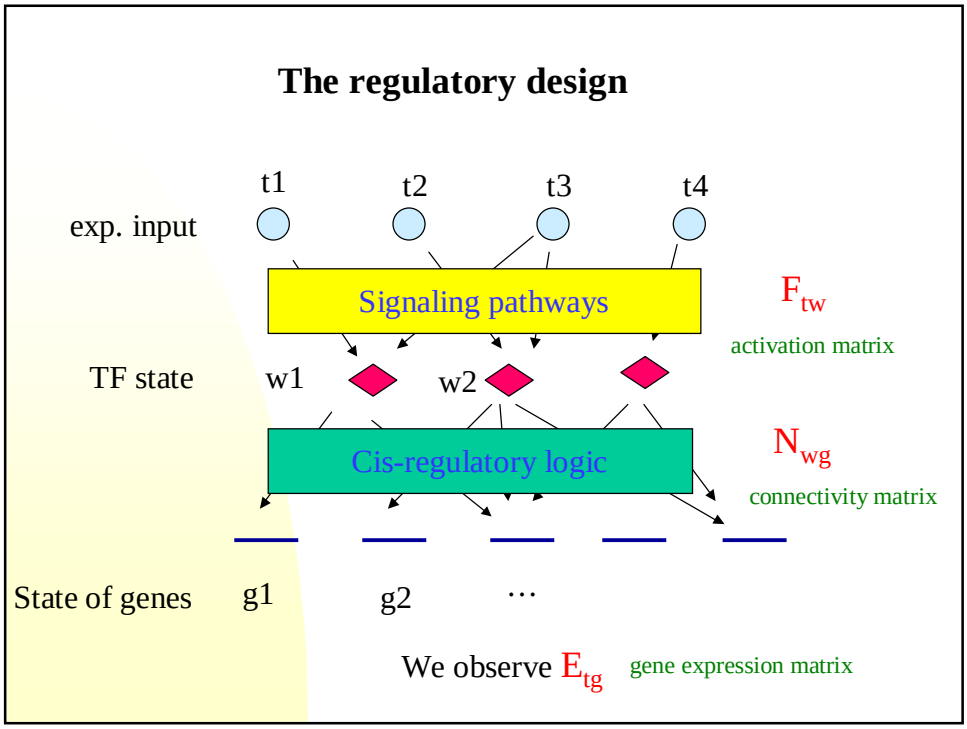
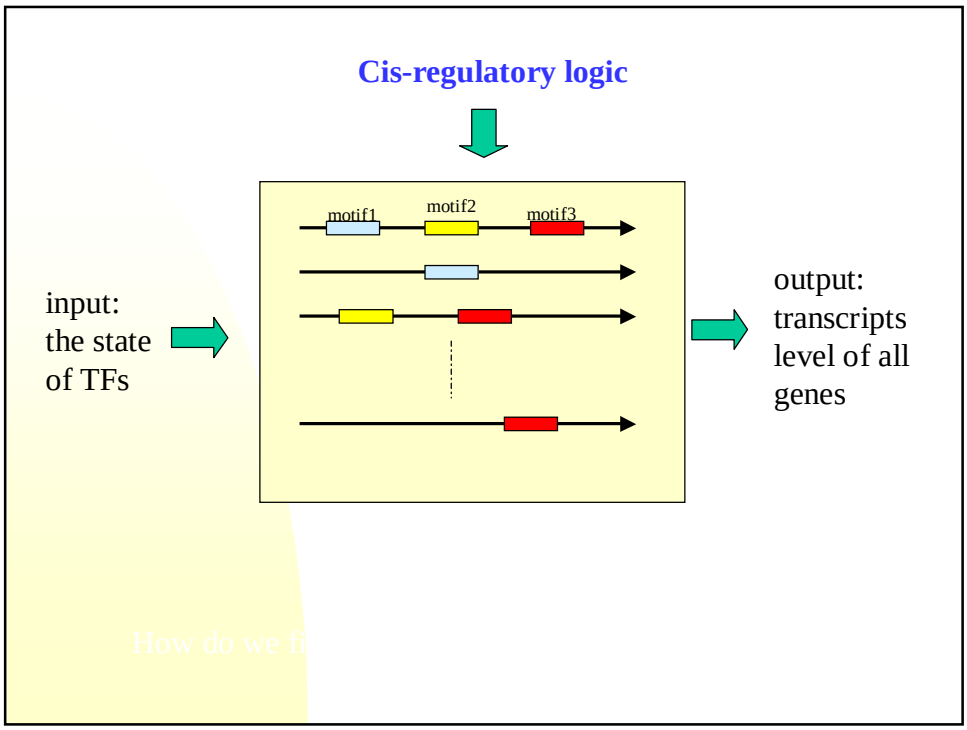
Reconstructing the Transcription Networks of A Cell Using Computational Genomics

Hao Li, UCSF

KITP, Santa Barbara
Feb 27, 2003




The regulatory design
exp. input
t1
t2
t3
t4
Signaling pathways
F_{tw}
activation matrix
TF state
w1
w2
Cis-regulatory logic
N_{wg}
connectivity matrix
State of genes
g1
g2
…
We observe E_{tg}
gene expression matrix


Cis-regulatory logic
input:
the state
of TFs
motif1
motif2
motif3
output:
transcripts
level of all
genes

grand goals

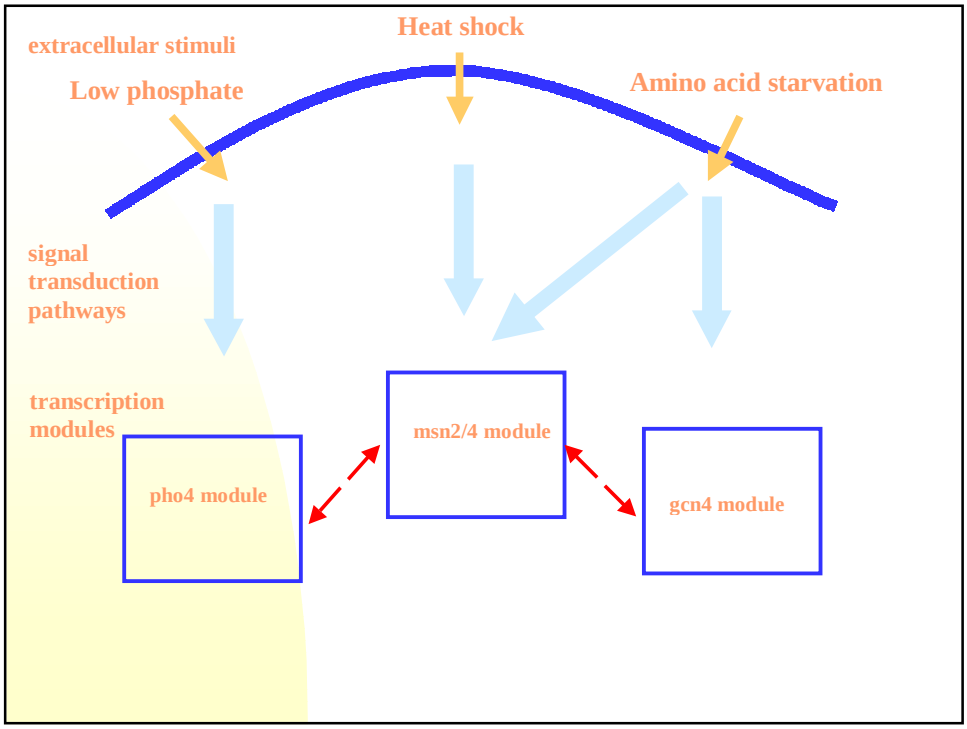
Dissecting all signaling pathways→ predict TF state from conditions

Deciphering cis-regulatory logic → how the logic function is defined by sequence motifs

Understanding the fundamental constraints and design principles of signaling pathways and cis-regulatory logic

Define a transcription module:

1. **A transcription factor;**
2. **The binding sites of the transcription factor**
3. **Genes that are directly regulated by this transcription factor.**



Modest Goals:

1. **Determine the structures of transcription modules (TF, binding site and target genes);**
2. **Identify conditions under which a module is activated;**
3. **Infer module interactions (overlaps, combinatorial control etc).**

data

Genome sequence + gene annotation

Gene expression data (array) under different conditions

Expression profile of mutants (deletion or over expression)

CHIP-on-CHIP

Step 1. Construct transcription module:
A. Identify the DNA regulatory motif recognized by a particular transcription factor using REDUCE;

↓

Step 1. Construct transcription module:
B. Identify genes that are regulated by this transcription factor using the MODEM algorithm

↓

Step2. Identify activation of transcription modules using correlation between motif weighted expression

↓

Step 3. Infer interactions between co-activated transcription modules by examining genes shared by these modules

Step 1. Construct transcription module:
A. Identify the DNA regulatory motif recognized by a particular transcription factor using REDUCE;

↓

Step 1. Construct transcription module:
B. Identify genes that are regulated by this transcription factor using the weighted profile method.

↓

Step2. Identify activation of transcription modules using its X value pattern

↓

Step 3. Infer interactions between co-activated transcription modules by examining genes shared by these modules

$$\log(E_{tg}) \approx \sum_{w} F_{tw} N_{wg}$$

REDUCER

Regulatory Elements Detection Using Correlation with gene Expression

focus on binding site W which correlate with genome-wide gene expression data

N_{wg} Number of occurrences of w in g's promoter

F_{tw} How strong site w correlate with expression in experiment t

$$\chi = \sum_{t,g} \left[\log(E_{tg}) - \sum_{w} F_{tw} N_{wg} \right]^2$$

How does REDUCER reduce?
An iterative scheme to identify independent motifs

Regression + model selection

Iteration:

suppose N-1 motifs found
fit expression data with the N-1 motifs → residual of the fit
find motif with strongest correlation with the residual
based on p value of Pearson correlation
take it as Nth motif
iterate

Features of REDUCE:

A. It does not depend on clustering of genes;
B. combinatorial motifs can be identified in each experiments;
C. The sign of the fitting coefficient can tell whether the factor induce or repress expression
D. Identify the conserved core of the binding sites

Analyzing microarray data systematically

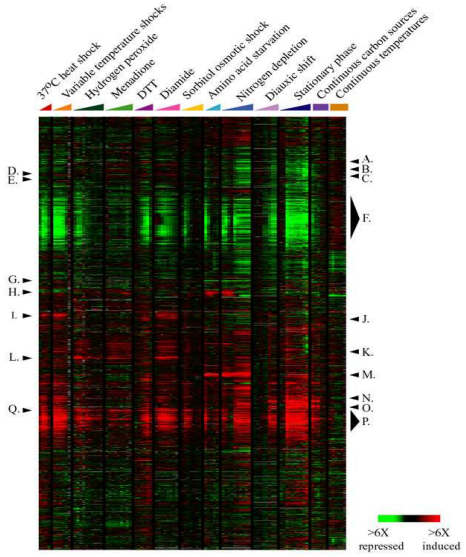
Data:

- **~300 gene deletion experiments;**
- **174 environmental stress response time course;**
- **9 sporulation time course;**
- **8 phosphate metabolism;**
- **22 cell cycle time course.**

stress response

Gasch et al.

Figure 1.



Many known motifs are identified as very significant (p-value < 10^{-20}) under the right conditions;

DNA motif	Experiment[b] (Site name/TF name)
aaatttt/aaaattt	ES
agggg/cccct	ES (STRE)
acccc	HS
acgcgt	CC (MCB)
cacaaaa/tttgtg	SPO (MSE)
cacgtgg	PHO (Pho4p site)
cgatgag	ES
taagggg	HS; DS
cgcgaaa	CC (SCB)
tgaaaaa	HS; Diamide; YPD

Combinatorial motifs are identified;

One example: in amino acid starvation time course

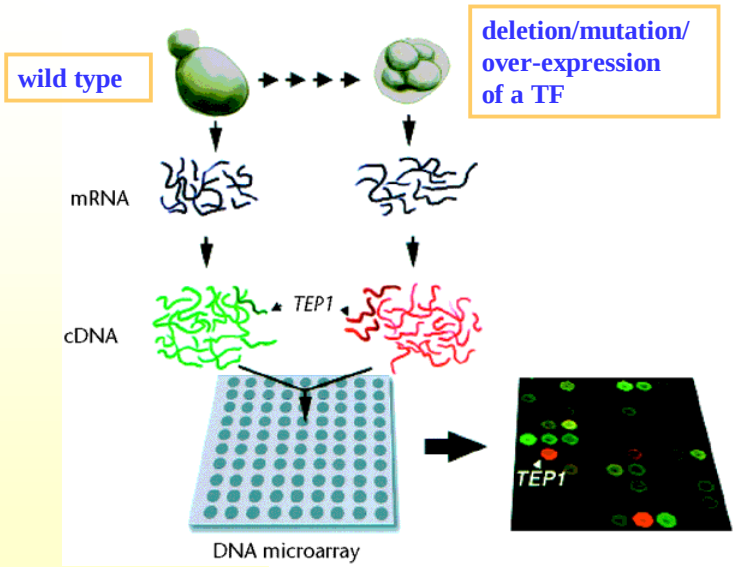
msn2/4

met4/met31

gcn4

DNA motif	p-value (-log10)
aaaattt/aaatttt	64
aaggg	23
acggtgt	3
agggg/cccct	27
ataag	6
atataaa	6
atgac	9
atgagc	17
cacgtga	7
ccacagt	6
ccgtaca	6
cgatgag	38
ctcatc	14
cttatc	8
gagtca	4
gataa	5
gataag	8
gtggc	6
tatataa	8
tcatc	6
tccgtac	7
tgaaaaa	19
tgactc	12
tgagatg	21

Transcription Factor Perturbation Experiment (TFPE):



The most significant motif identified in TFPE is typically the regulatory motif recognized by the TF or its cofactor(s)

TF	Motif (p-value)	Known binding site[a]	Biological process[b]	
GCN4	tgactca (10^{-80}) tgagtca (10^{-26})	tga[c/g]tca	transcriptional activator of amino acid biosynthetic genes	
MBP1	acgcgt (10^{-27})	MCB site acgcg[t/a]	DNA replication; cell cycle control	
MSN2	agggg (10^{-26})	agggg	Stress response	
MSN4	agggg (10^{-33})	agggg	Stress response	
PHO4	cacgtgg (10^{-30})	cacgtg	Phosphate metabolism	
RTG1	ggtcacg (10^{-5})	ggtcac	interorganelle communication	
STE12	tgaaac (10^{-14})	PRE site tgaaac[g/a]	Invasive growth; pheromone induction; pseudohyphal growth	
YAP1	tgactca (10^{-8})	tgactca	Regulation of certain oxygen detoxification enzymes	
MAC1	tgcaccc (10^{-80})	N/A	Cu/Fe utilization, stress resistance	
SIN3	cgcgcgc (10^{-24})	N/A	transcription	
TUP1	aggcac (10^{-25})	N/A	Glucose repression	

Step 1. Construct transcription module:
A. Identify the DNA regulatory motif recognized by a particular transcription factor using REDUCER;

↓

Step 1. Construct transcription module:
B. Identify genes that are regulated by this transcription factor using the MODEM algorithm.

↓

Step2. Identify activation of transcription modules using motif weighted expression

↓

Step 3. Infer interactions between co-activated transcription modules by examining genes shared by these modules

Identifying the target genes of a factor: Why MODEM?

1. **Using core motif alone can not accurately identify target genes;**
   137 genes containing CACGTGG in yeast
   only ~15 genes are true targets of Pho4p.

2. **Target genes may have different versions (mismatches) of the core motif.**
   Pho4p target VTC1 only has CACGTGC
   ~2000 genes have CACGTGG with 1 mismatch

3. **Additional information can be obtained from**

   A. **flanking sequences of the core motif;**

   B. **gene expression.**

Inputs to MODEM:

1. The core and flanking sequences S={S(1),S(2),…S(N)};
2. Gene expression (log ratio) E={E(1),E(2)…E(N)}.

```
agggcacgtggcgtt   2.5
cttcacgtggctga   4.6
cgttcacgtgggtga   1.2
cgttcacgtggcgga   2.0
```

Outputs of MODEM:

1. **Calculate the probability of each gene being a target;**

2. **Classify each gene into the target or non-target category.**

MODEM

Module construction using gene Expression and sequence Motif

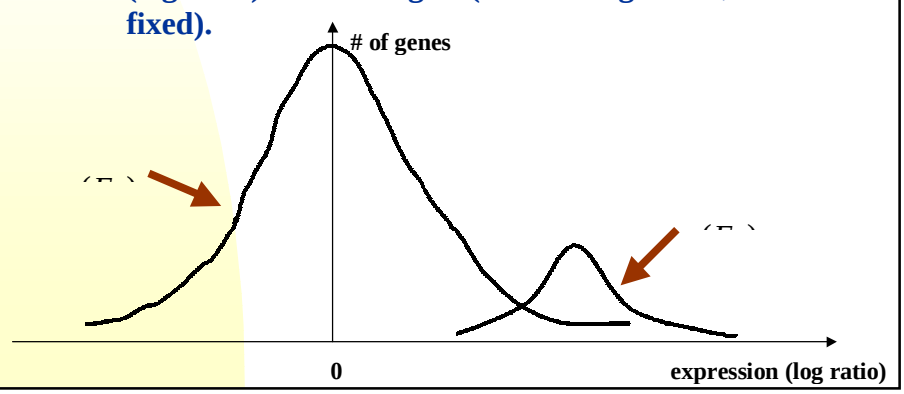
is

A joint probability model for gene expression and sequence motif

The 5 components in MODEM:

$\rho(E_k)$ **A Gaussian distribution for the gene expression (log ratio) of targets;**

$\rho_0(E_k)$ **A Gaussian distribution for the gene expression (log ratio) of non-targets (the entire genome, fixed).**



The 5 components in MODEM:

$f_{i\sigma}$ **A position specific scoring matrix (PSSM) for sequences belonging to targets;**

$f^{0}_{i\sigma}$ **A PSSM for sequences belonging to non-targets;**

A PSSM:

- **reflects how often a specific nucleotide appears at a position of the motif;**
- **if given a sequence, can be used to calculate the probability of observing such a sequence and thus score it.**

Position	1	2	3	4	5
A	0.80	0.10	0.01	0.05	0.02
C	0.02	0.85	0.20	0.90	0.01
G	0.10	0.02	0.78	0.03	0.88
T	0.08	0.03	0.01	0.02	0.09

P(ACGCG) >P(CCGCG)

The 5 components in MODEM:

λ **A prior percentage of targets among the genes containing the core motif;**

Without any sequence or expression information, λ is an estimation of target percentage among the genes containing the core motif.

The values of the parameters θ in the MODEM are

determined iteratively

to maximize the posterior probability $P(\boldsymbol{\theta} \mid S, E)$

Sum over all possible values of the hidden variable:

The value of the hidden variable reflects:
The extended motif belongs to target or non-target

$$P(\boldsymbol{\theta} \mid S, E) = \sum_{hidden} P(\boldsymbol{\theta}, hidden \mid S, E)$$

Assumptions:

1. Sequences (extended motifs) S(k), k=1, … N, are independent from each other;

2. Given the value of the hidden variable and the parameters, the distributions of sequence and expression are independent;

Posterior probability:

Target term

$$P(\boldsymbol{\theta} \mid S, E) = \prod_{k=1}^{N} \left\{ \left[\left(\prod_{i=1}^{M} f_{i\sigma_i(k)} \right) \rho(E_k) \lambda \right] + \left[\left(\prod_{i=1}^{M} f^{0}_{i\sigma_i(k)} \right) \rho_0(E_k)(1-\lambda) \right] \right\}$$

Product of all sequences

Non-target term

Posterior probability:

target PSSM parameters

Expression distribution of targets

Prior percentage of targets

$$P(\boldsymbol{\theta} \mid S, E) = \prod^{N} \left\{ \left[\left(\prod^{M} f_{i\sigma^i(k)} \right) \rho(E_k) \lambda \right] + \left[\left(\prod^{M} f^{0}_{i\sigma^i(k)} \right) \rho^0(E_k)(1-\lambda) \right] \right\}$$

non-target PSSM parameters

Expression distribution of non-targets (fixed)

MODEM Outputs:

1. Probability of being a target.

$$P(\text{target}) = \frac{\left[\left(\prod_{i=1}^{M} f_{i\sigma_i(k)}\right)\rho(E_k)\lambda\right]}{\left\{\left[\left(\prod_{i=1}^{M} f_{i\sigma_i(k)}\right)\rho(E_k)\lambda\right]+\left[\left(\prod_{i=1}^{M} f^{0}_{i\sigma_i(k)}\right)\rho_0(E_k)(1-\lambda)\right]\right\}}$$

MODEM Outputs:

2. Classify genes using a optimal Bayesian classifier:

$$\frac{P(\text{target})}{P(non-\text{target})} \geq 1$$ **kth sequence is a target**

otherwise **kth sequence is a non-target**

Target genes of Pho4p identified using the core motif CACGTGG and allowing 1 mismatch.

Rank	Gene/ORF	Probability	Extended motif	Expression	Ogawa[a]	Carroll[a]
1	PHO89	1.000	ATGGCACAACGTGGGGATGAC	5.262	√	√
1	SPL2	1.000	TGTCGGTCACGTGAGCAAAAA	4.605	√	
1	PHO84	1.000	TCCGCCCCACGTGCTGGAAAT	5.491	√	√
1	PHO11	1.000	ATGCGAAAACGTGGTAATTTA	4.287	√	√
1	PHO12	1.000	TTAAACCCACGTGTGAACGCC	4.159	√	√
1	VTC3	1.000	AGGCAGAAACGTGGAAACATA	4.297	√	√
1	VTC4	1.000	GTGCAGCCACGTGCGGATGAA	3.296	√	√
1	PHM6	1.000	CACCTCCCACGTGTCAGCGAA	2.998	√	√
1	PHO5	1.000	GCACTCACACGTGGGACTAGC	2.816	√	√
1	VTC1	1.000	TCCGAGACACGTGCTAATATC	2.485	√	√
1	YAL011W	1.000	AGGCAGAGACGTGGCACTGGC	2.233		
12	CTF4	0.999	AGAATCTCACCTGGAGAATGG	2.625		
13	YLR402W	0.998	GAGTTTGCAGGTGGGACTAAT	2.223		
14	CTF19	0.998	GAGGGCCCACGTGGCTTAATA	1.864	√	√
15	PHM5	0.998	GGCCGCACAGGTGGGCAGATC	1.757	√	√
16	REC107	0.995	CTAATCTTACGTGGTTCTTAT	2.310		
17	PHO8	0.991	GTCGGGCCACGTGCAGCGATC	1.546	√	√
18	NUP85	0.987	AAGAGGGCACTTGGTCACAAC	1.926		
19	YJR039W	0.980	GTCTTGACACGTAGGCGTTGC	1.876		
20	CDA1	0.974	TCTCATGCACTTGGAAGCAGC	1.852		
21	YML089C	0.965	GCAATTATACGTGGCAAGGAA	1.937		
22	KRE2	0.927	GTCGGGCCACGTGCAGCGATC	1.233		
23	VTC2	0.902	AAAAACCCACGTGCTGCTTGG	1.599	√	√
24	YAR069C	0.887	GTTCACACTCGTGGGGCCCAC	1.438		

a. Genes identified as targets of Pho4p experimentally by Ogawa et. al.[9] or Carroll et. al.[10] are marked with checks.

53 genes in Gcn4 module (from deletion experiment)

1. **Genes involved in biosynthesis and metabolism are greatly enriched.**

amino acid metabolism: p-value 6.28E-39
31 out of 53 genes, 58.4%
138 out of 6906 annotated genes, 1.9%

amino acid biosynthesis: p-value 6.63E-37
27 out of 53 genes, 50.9%
89 out of 6906 annotated genes, 1.2%

Completely consistent with the function of Gcn4p: a master regulator of biosynthesis.

In the 53 genes predicted to be targets of Gcn4p:

2. **39 (74%) were shown by experiments that their expressions were induced/repressed by Gcn4p;**
   10 are ORFs whose functions are unknown.
   4 genes with annotated functions:

Gene	Molecular function	Biological process
TMT1	Trans-aconitate 3-methyltransferase	unknown
ADH5	Alcohol dehydrogenase	Alcohol metabolism
STR3	Cystathionine beta-lyase	Methionine biosynthesis
ALD5	Aldehyde dehydrogenase	metabolism

Predictions based on the Gcn4 module:

1. **14 new target genes;**

2. **The 10 ORFs whose function are unknown may play roles in biosynthesis or metabolism.**

28 transcription modules have been constructed
Based on TFPE experiments

Predicted new binding sites and target genes.

Apply REDUCE+MODEM

To analyzing CHIP-on-CHIP data

From R. Young Lab

35 factors → significant core motif
module contains target genes
consistent with the regulatory
function of the factor

TF	Motif	Sig	Known motif	Sensible targets	Function of TF
ABF1	agtgat	12.2			
ACE2	ccagcaa	2.5	accagc	Yes	G1-specific trans.
ARO80	gcggagc	2.5		Yes	In mitotic cell cycle reg. due to
BAS1	tgactc	10.8		Aro9, Aro10 Yes	arpmatic aa adenine synthesis
CBF1	cacgtg(a)	97.9	cacRtg	Met2-8 Yes	methionine synthesis
CIN5	ttacata	16.8		Met.syn. genes Yes	drug resistance
DAL81	tgccgt	3.7		Transporters Yes	Cisplatin resistance nitrogen utilization
FHL1	ccgtaca	58.0		(Agp1, Bap3..) Yes	rRNA processing
FKH1	gtaaaca	7.8	RWaaaYaW	(ribosomal genes ?	chromatin silencing
FKH2	gtaaaca	19.3	RWaaaYaW	Yes	cell cycle
GCN4	tgactca	45.0	tga[c,g]tca	Cell cycle,spo genes Yes	AA response
GLN3	tttgaa	14.4	match	AA syn. genes ?	Nitrogen utilization
HAP4	ccaatca	10	Ste12 sites Match ccaat Hap2/3/4 site	Ste12 targets Yes	Carbohydrate
HSF1	ttctaga	2.8		Known genes Known genes	metabolism
INO2	cacgtga	5.3		Yes	phospholipid
INO4	gcatgtg	15		(ITR1,PSD1,..) Yes	biosyn. Phos.
LEU3	gggaccg	3.1		Phosphate genes Yes	biosyn
MAC1	gctcgtt	2.7		(leu1,bap2,leu4) Yes	
MBP1	acgcgt	107	acgcgt	Known target Yes	Cell cycle
MCM1	acaata	2.2	cccaaWWagg		
MET31	tcacgtg	2.4	Match met4/31	Yes	Met. Syn.
Met4	cacgtga	11.3	Complex site	Met. Syn. genes Yes	Met.syn.
NDD1	gtaaaca	6.7		Yes	G2/M specific
NRG1	aggcaca	6.6		Yes	Glucose metab.
REB1	cgggtaa	128	cgggtRR	Hexose-glucose trans., permease ?	
SKN7	cggccg	7.9		Yes	Osmotic stress
STB1	acgcgt	3.8	Some complex as	Known target	G1/S transition
STE12	tgaaaca	10	MBP1? atgaaa	Yes	Pheromone
SUM1	gtgtcac	8.9		Known genes Yes	response sporulation
SWI4	cgcgaaa	20.1		Spo. genes	
SWI5	gctggct	2.3	KgcTgR	Yes	
SWI6	acgcgt	18	Some complex as	Known targets Yes	
YAP1	ttactaa	2.8	MBP1? ttaNtaa	Known targets Yes	
YAP6	ccgcgga	8	Match pdr1/3 site		Drug resistance

Some statistics

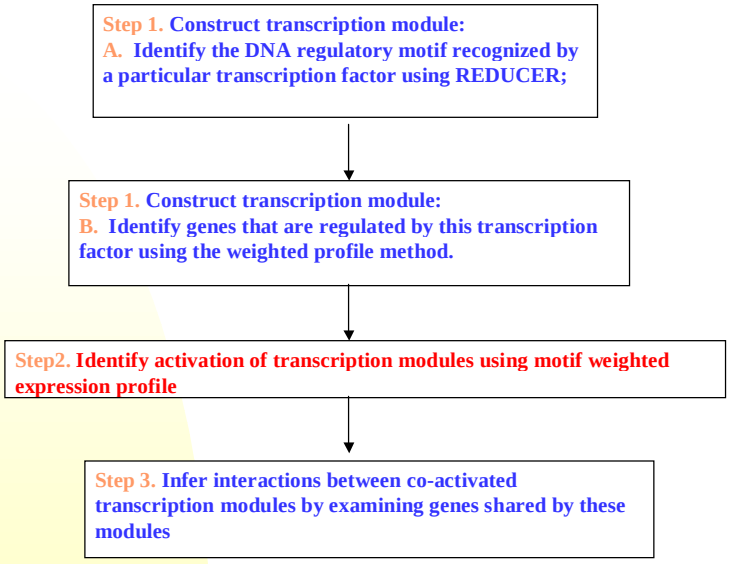
Out of 106 factors → 70 overlap with YPD
total ~800 known sites

Target Overlap of Young with YPD: 154 → 39 factors
picked at least one known target

39 factors and 35 factors picked by our analysis
→ 25 overlap

The MODEM algorithm

- **Identify target genes of a particular transcription factor with high specificity and sensitivity;**
- **Predict new target genes;**
- **Help annotate gene functions.**



Available data (Inputs):

1. **TFPE;**
2. **TF location (CHIP-on-CHIP) data**
3. **Microarray experiments under different cellular conditions**

Goal (Output):

infer the state of TFs → Identify conditions under which a particular module is activated

Approach:

Compare the expression profile of the TFPE or CHIP-on-CHIP and that of the interesting condition.

"Global" comparison between expression profiles is noisy.

Gene	1	2	3 ……											
PHO4 mutation	0.2	-0.1	0.3 ……			1.5	2.6	1.8	3.5	2.2	……	-0.2	–1.1	0.6
Hypo-osmotic shock	-0.9	1.1	-2.0 ……			1.5	2.6	1.8	3.5	2.2	……	1.0	0.1	2.8

Pho4p target genes

We need to search "local" similarity between the two profiles.

Method: motif weighted profile comparison

"local similarity" between global expression profiles

Is the same TF module activated in both experiments?
focus on a subset of genes with the potential binding site

define X value as: X(g,t,m) = E(g,t) * N(g,m)

E(g,t): logarithm base 2 of the expression ratio for gene g in the experiment t;
N(g,m): the number of occurrence of motif m in the gene g.

Compare x values between a TF perturbation experiment and a specific condition

Predict activation conditions of TFs/modules

TF	Biological process	Activation conditions
GCN4	transcriptional activator of amino acid biosynthetic genes	amino acid starvation, nitrogen depletion
MBP1	DNA replication; cell cycle control	cell cycle, diauxic shift, nitrogen depletion, heat shock
MSN2	Stress response	environmental stresses, phosphate metabolism
MSN4	Stress response	environmental stresses

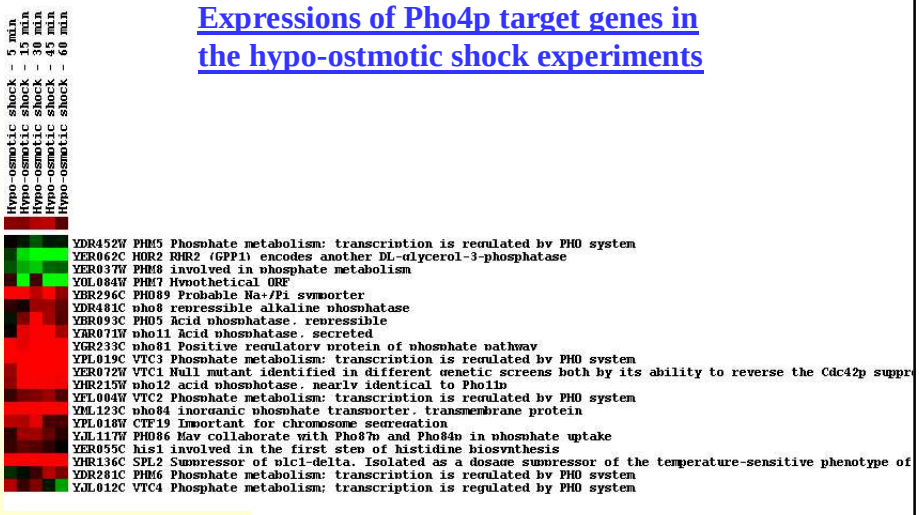
Predict activation conditions of TFs/modules

TF	Biological process	Activation conditions
PHO4	Phosphate metabolism	phosphate metabolism, hypo-osmotic shock
RTG1	interorganelle communication	amino acid starvation, nitrogen depletion
STE12	Invasive growth; pheromone induction; pseudohyphal growth	nitrogen depletion
YAP1	Regulation of certain oxygen detoxification enzymes	amino acid starvation, nitrogen depletion

Unexpected activations suggest:

- Those modules may play roles in cell's adaptation to these cellular conditions;
- There probably exist unknown links in the networks.

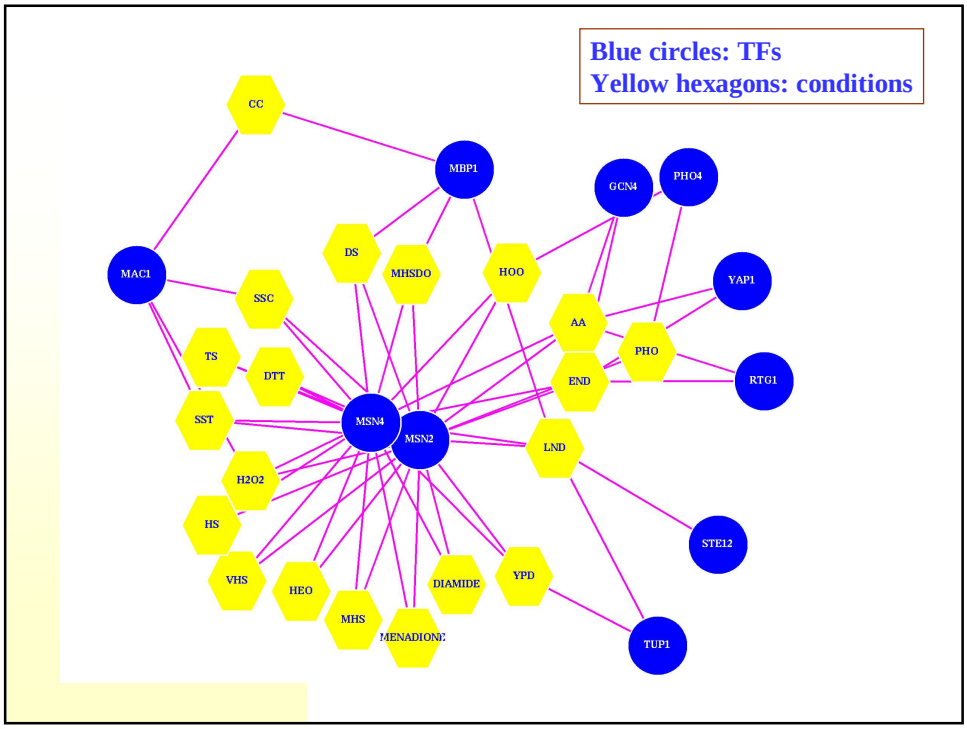
Expressions of Pho4p target genes in the hypo-ostmotic shock experiments



Pho4's activation under hypo-osmotic shock can not be identified by "global" comparison of expression profiles.

Target genes of Pho4p identified from the 45 minute time point of the hypo-osmotic shock experiments using core motif CACGTGG.

Rank	Gene/ORF	Probability	Extended motif	Expression (log2 ratio)	Ogawa[a] et. al.[9]	Carroll[a] et. al.[10]
1	SPL2	1.000	ATGTACGCACGTGGGCGAAAG	4.980	√	
1	VTC3	1.000	ATTAAGCCACGTGGGCCCTCG	1.940	√	√
1	VPS8	1.000	ATACAAGCACGTGGGCCCTCC	1.680		
4	PHO12	0.999	GCGTTCACACGTGGGTTTAAA	1.500	√	√
5	PHO84	0.994	TTTCCAGCACGTGGGGCGGAA	1.490	√	√
6	PHO89	0.993	AATGCAGCACGTGGGAGACAA	1.220	√	√
7	PHO5	0.991	GCACTCACACGTGGGACTAGC	0.640	√	√
8	PHO11	0.988	GCGTTCACACGTGGGTTTAAA	1.030	√	√
9	MNN1	0.976	TTAAAAGCACGTGGCACGAGA	1.210		
10	PHM6	0.970	TCGCTGACACGTGGGAGGTGG	0.700	√	√
11	NAB3	0.852	ACTCAATCACGTGGGATACCA	0.700		

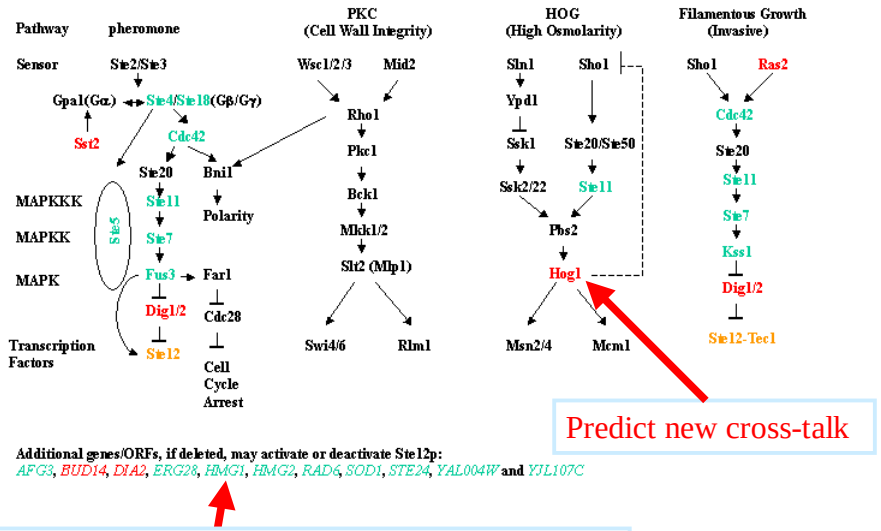


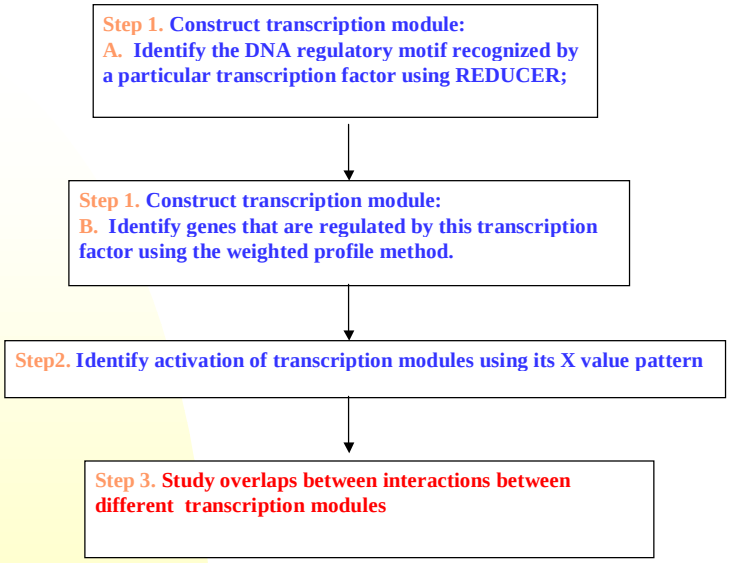
Consider genetic perturbations:

~300 single-gene deletion microarray experiments.

Mapping the "activation" genes onto known signaling pathways can provide useful insight.

Identify "activation" genetic perturbations





Identify genes co-regulated by more than one transcription factor and find combinatorial control between different transcription factors.

One example

Overlap between SUM1 and NDT80

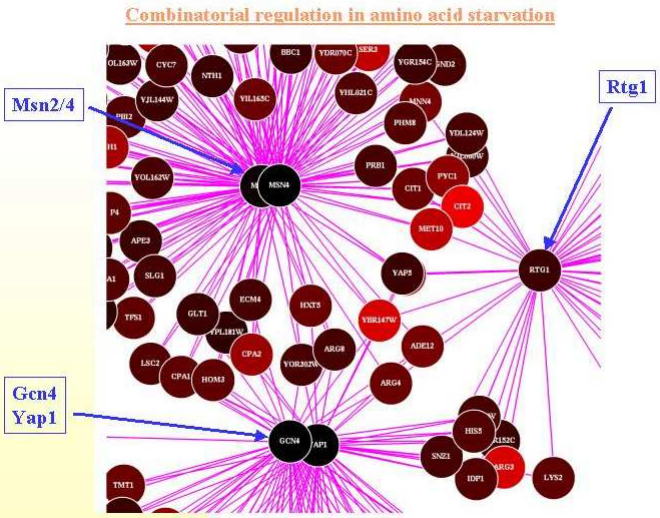
SUM1 module
- Core motif gtgtcac
- 67 genes
- from CHIP-on-CHIP

NDT80 module
- core motif cacaaaa
- 88 genes
- from NDT80 deletion

Overlap between module SUM1 and NDT80

YIR028W	dal4	CGCTTTGCTGTCACGTCGATA	allantoin transport
YOR313C	SPS4	TACATTGGTTTCACATAACAT	sporulation-specific protein
YNL318C	HXT14	TGAATTTGTGTCATCATTAGA	High-affinity hexose transporter
YBR148W	YSW1	CAGACGCGTGTCAGCAAAGGG	Spore-specific protein
YDR523C	SPS1	TTTTTATGTGTCATTTTTTTT	involved in middle/late stage of meiosis, required for spore wall for.
YDR042C	YDR042C	GGATTTTGTGTCATTAGCAAA	Hypothetical ORF
YGL170C	SPO74	ATTTCTTGTGACACAAAAGAG	Protein involved in sporulation
YBR180W	DTR1	AGACATTCTGTCACCTGGTGA	dityrosine transporter MFS-MDR
YLR343W	YLR343W	AATCAGAGTGACACAAATTTT	Hypothetical ORF
YLR308W	CDA2	TTGCGTTGCGTCACAAAATCA	Required for proper formation of the ascospore wall
YFR023W	PES4	AGAATCAGTATCACAAAAAAA	Suppressor of DNA polymerase epsilon mutation
YGR059W	SPR3	CTCTTTTGTGTCGCTAACAAA	sporulation-specific
YOR255W	YOR255W	AGCGATTGTGTCAGTAATGAA	Hypothetical ORF
YFR032C	YFR032C	AATGGAAGCGTCACAAATTAA	Hypothetical ORF
YJL037W	YJL037W	CGATTTAGTGTCATTTTTTTT	Hypothetical ORF
YJL038C	YJL038C	CGATTTAGTGTCATTTTTTTT	Hypothetical ORF
YHR184W	SSP1	TGATTTTGTGTCGCCTGTTTG	Involved in the control of meiotic nuclear divisions & spore formation;
YHR124W	NDT80	TAAATAGGTGACACAAAATGG	Meiosis-specific gene transcription

SUM1 site and NDT80 site overlap



constructing transcription network by combining sequence information, gene expression data, gene function, and pathway information.

1. **Core regulatory motifs (known and predicted) are identified systematically by REDUCE**
2. **Targets genes of a transcription factor are identified with high sensitivity and specificity using MODEM algorithm**
3. **Conditions that can activate a particular transcription module can be identified by comparing motif weighted profiles**
4. **examining co-activated TFs and their targets → suggest combinatorial control by the TFs**

Acknowledgement

Harman Busssemaker, Columbia

Eric Siggia, Rockefeller

Wei Wang, Stanford

David Botstein, Stanford

Mike Cherry, Stanford

Yigal Nochomovitz, UCSF biophysics

Randy Wu, UCSF biophysics

Reference: **http://mobydick.ucsf.edu/~haoli**