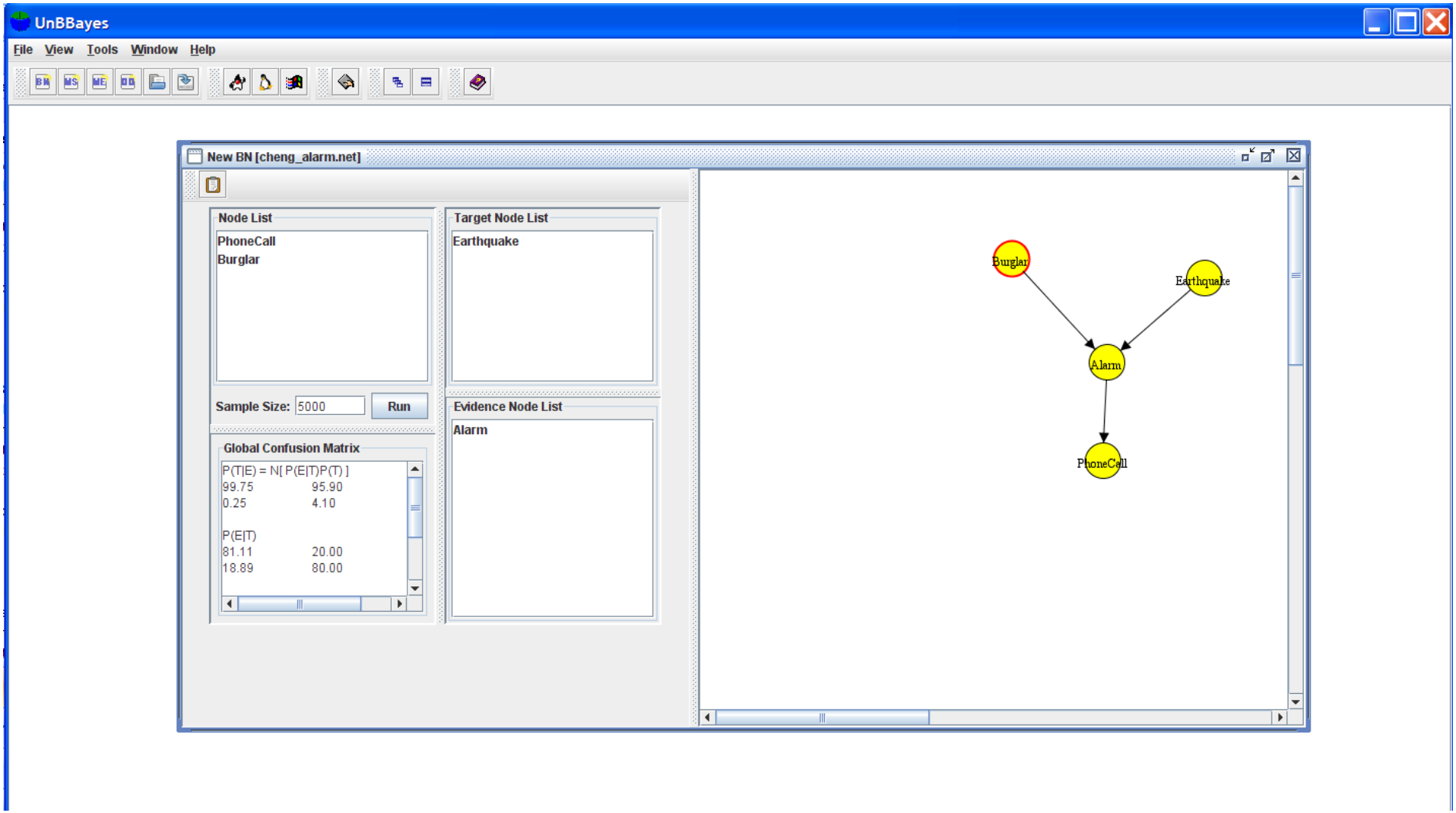
Inferring Cellular Networks Using Probabilistic Graphical Models

Jianlin Cheng, PhD
University of Missouri
2010

Bayesian Network Software

- http://www.cs.ubc.ca/~murphyk/Software/BNT/bnsoft.html

Demo


UnBBayes
File View Tools Window Help
New BN [cheng_alarm.net]
Node List
PhoneCall
Burglar
Target Node List
Earthquake
Sample Size: 5000
Run
Evidence Node List
Alarm
Global Confusion Matrix
P(T|E) = N[P(E|T)P(T)]
99.75 95.90
0.25 4.10
P(E|T)
81.11 20.00
18.89 80.00
Burglar
Earthquake
Alarm
PhoneCall

References

- E. Segal, M. Shpira, A. Regev, D. Peer, D. Botstein, D. Koller, and N. Friedman. **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. Nature Genetics. 2003.
- N. Friedman. **Inferring cellular networks using probabilistic graphical models**. Science. 2004.

Research in molecular biology is undergoing a revolution

- mRNA transcript quantities
- protein-protein
- protein-DNA interactions
- chromatin structure
- Protein quantities
- Protein localization
- Protein modification

Challenge

- Provide methodologies for transforming high-throughput heterogeneous data sets into biological insights about the underlying mechanisms
- Data is noisy
- Data integration
- Generate Hypothesis

Biological Networks – Gene Regulatory Networks

Legend: A transcription factor molecule binds to the DNA at its binding site, and thereby regulates the production of a protein from a gene.

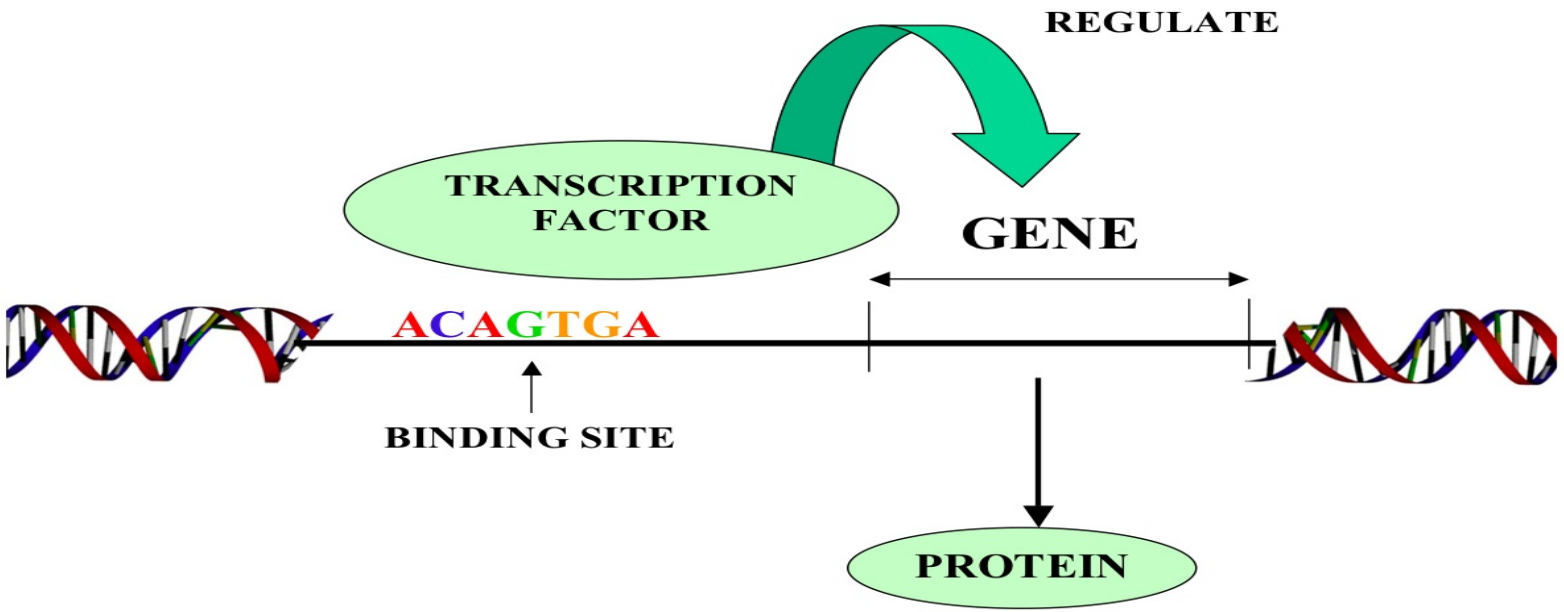
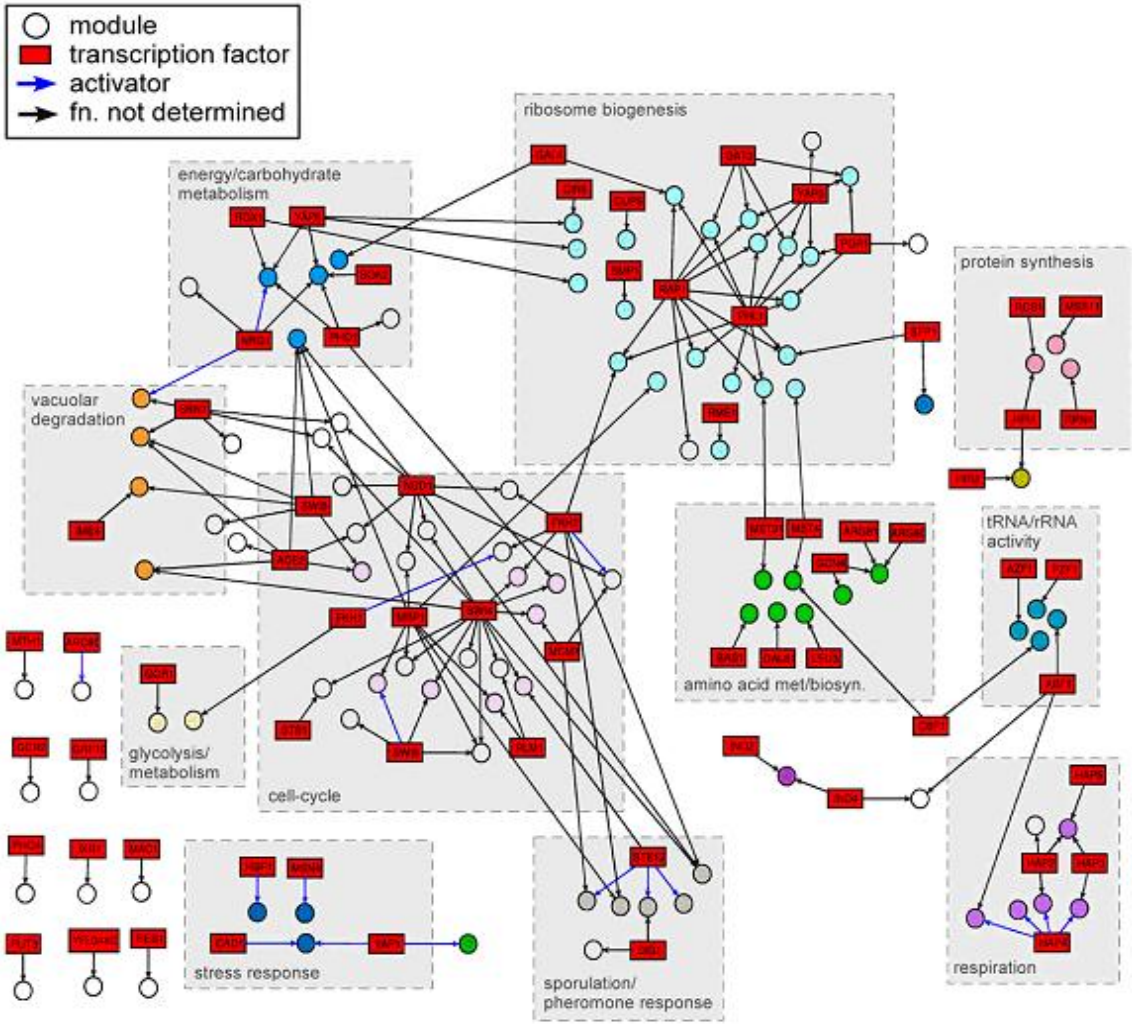
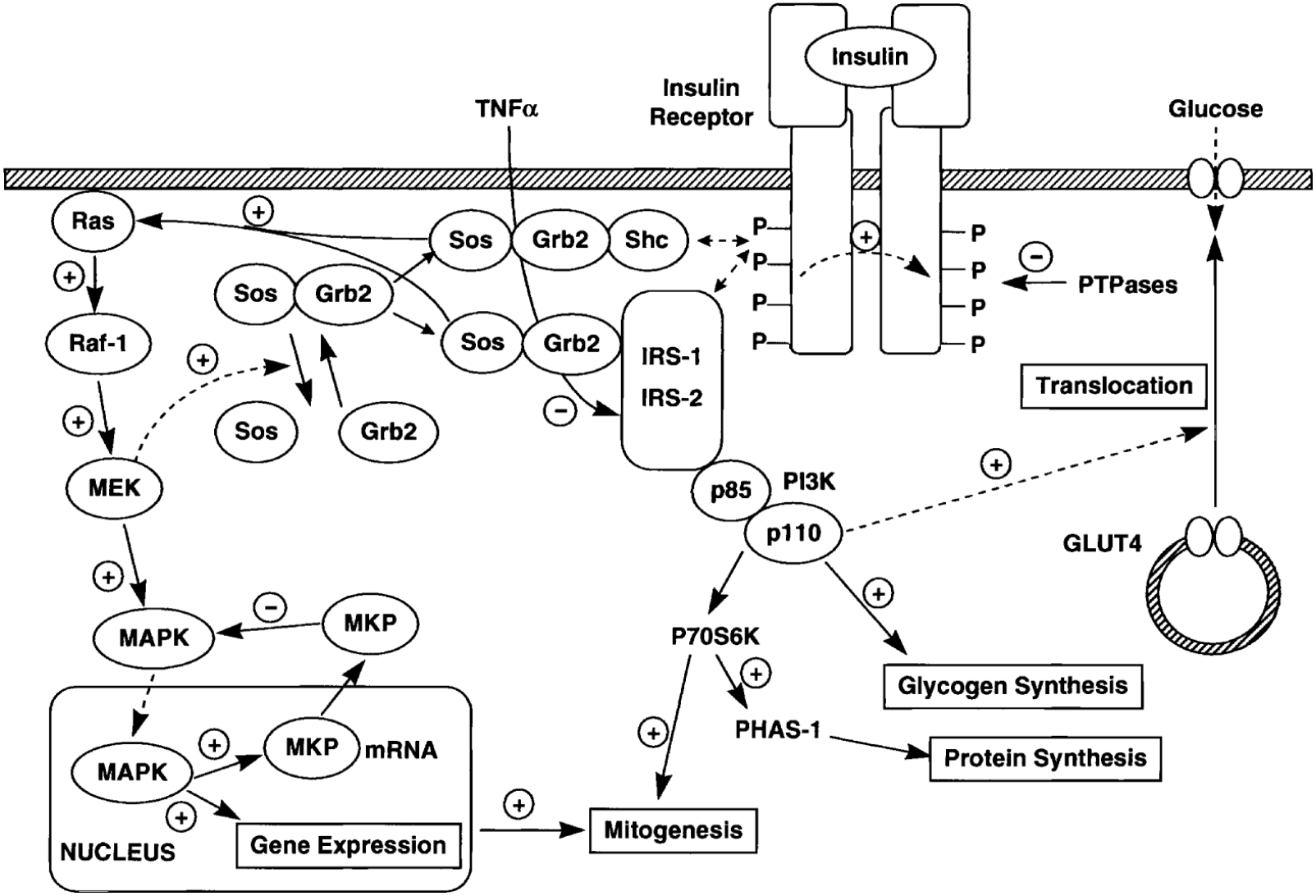


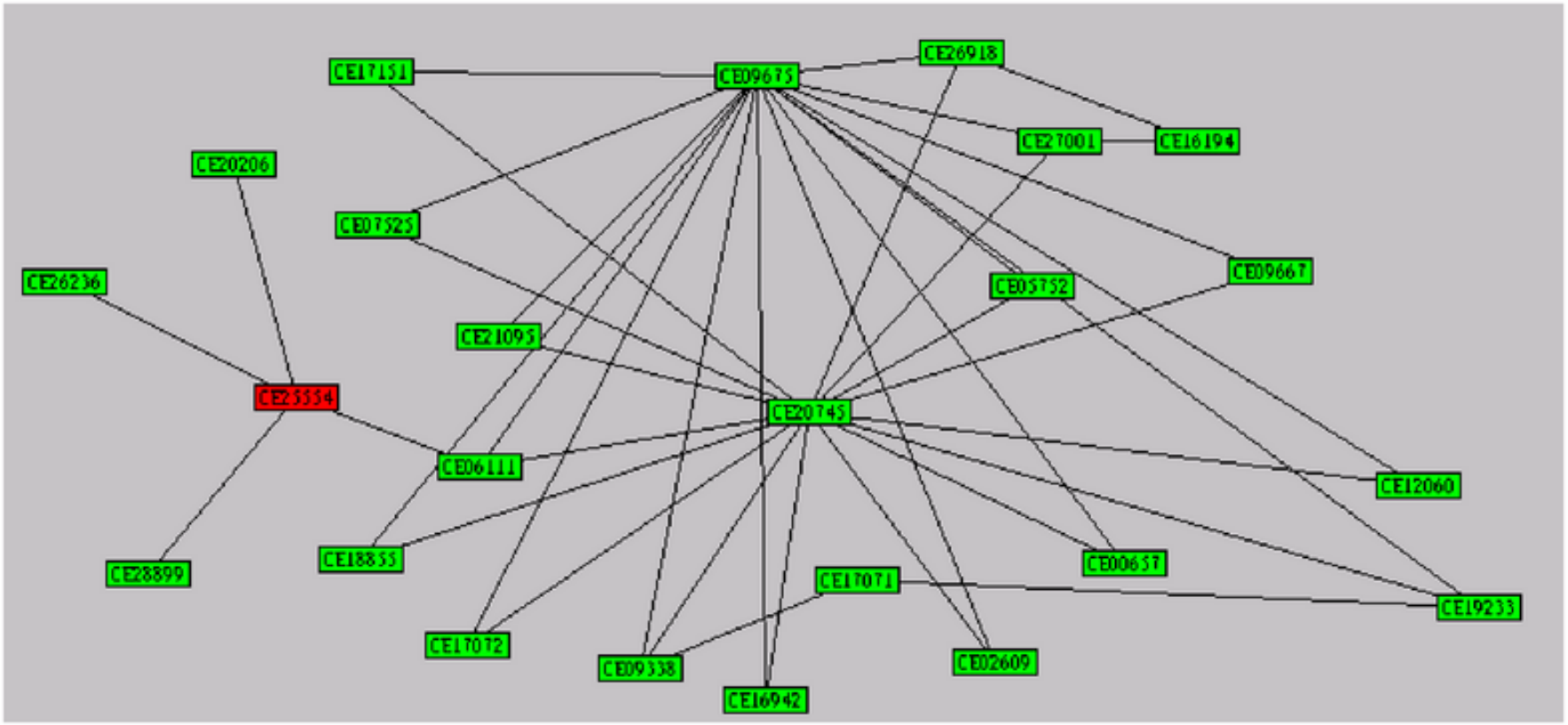
Figure 1: Rich media gene modules network



Signal Transduction Network



Protein Interaction Network



Model-Based Approaches VS Procedure Approaches

- **Procedure**: Binding sites – Gene expression. (a) cluster co-expressed genes to find common sites (b) group genes with similar binding sites and test if they are coexpressed
- **Declarative**: design a model that describes the relations between the two types of data. Learn parameter from data and make predictions

Probabilistic Models

- Stochasticity for measurement noise
- Learning Algorithms
- Select model that fits the actual observations
- Inference
- Make predictions
- Generate insights and hypothesis

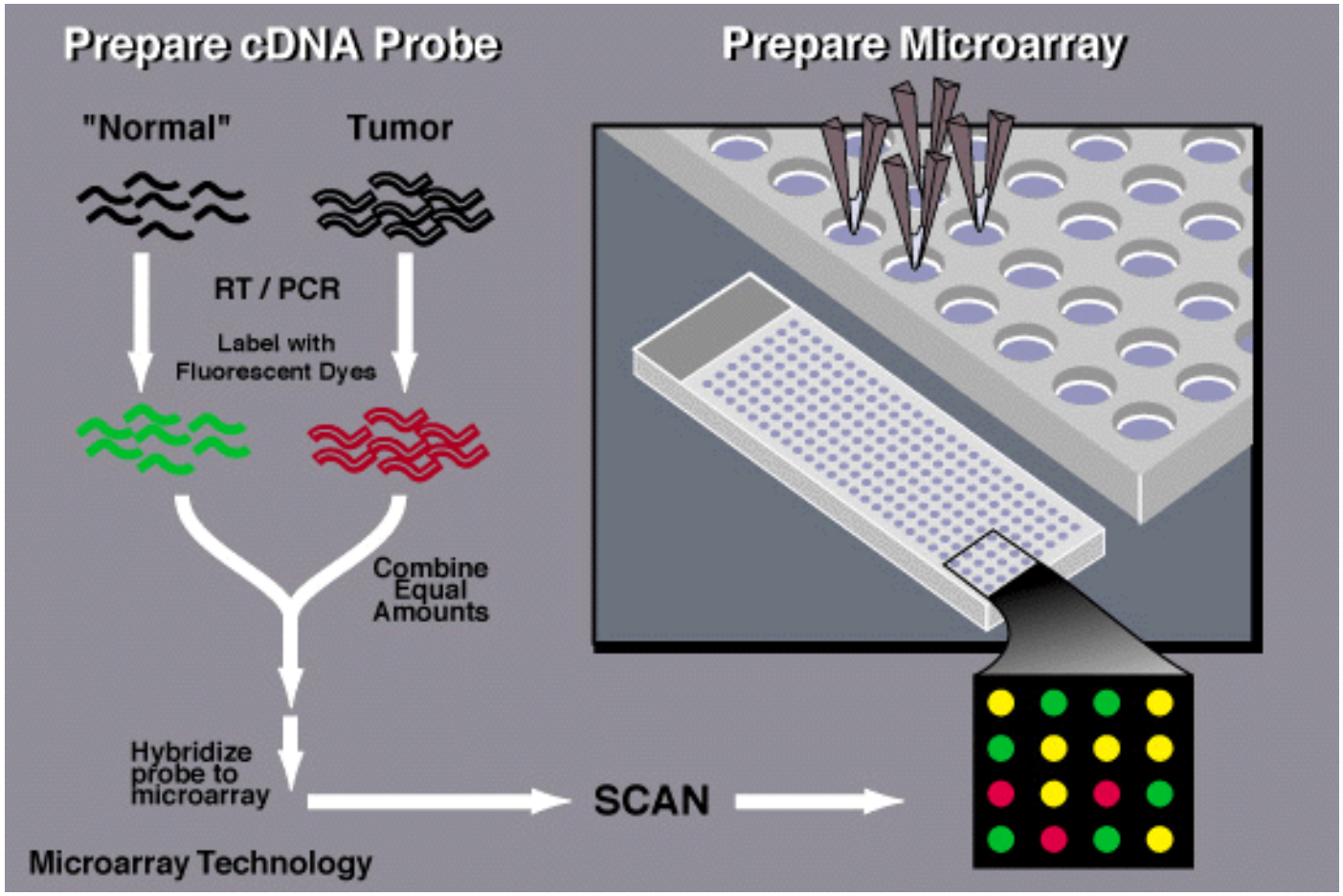
Modeling Examples

- Hidden Markov Model for sequence analysis
- Probabilistic Graphical Model for cellular networks

Advantages

- Concise language for describing probability distributions over the observations
- Approaches to learning from data that are derived from basic well-understood principles
- Use of observations to fill in model details
- Provide principles for combining multiple local models into a joint global model
- Declarative nature provides an advantage to extend model to account for additional aspects of the system

Infer Gene Regulatory Network from Gene Expression Data



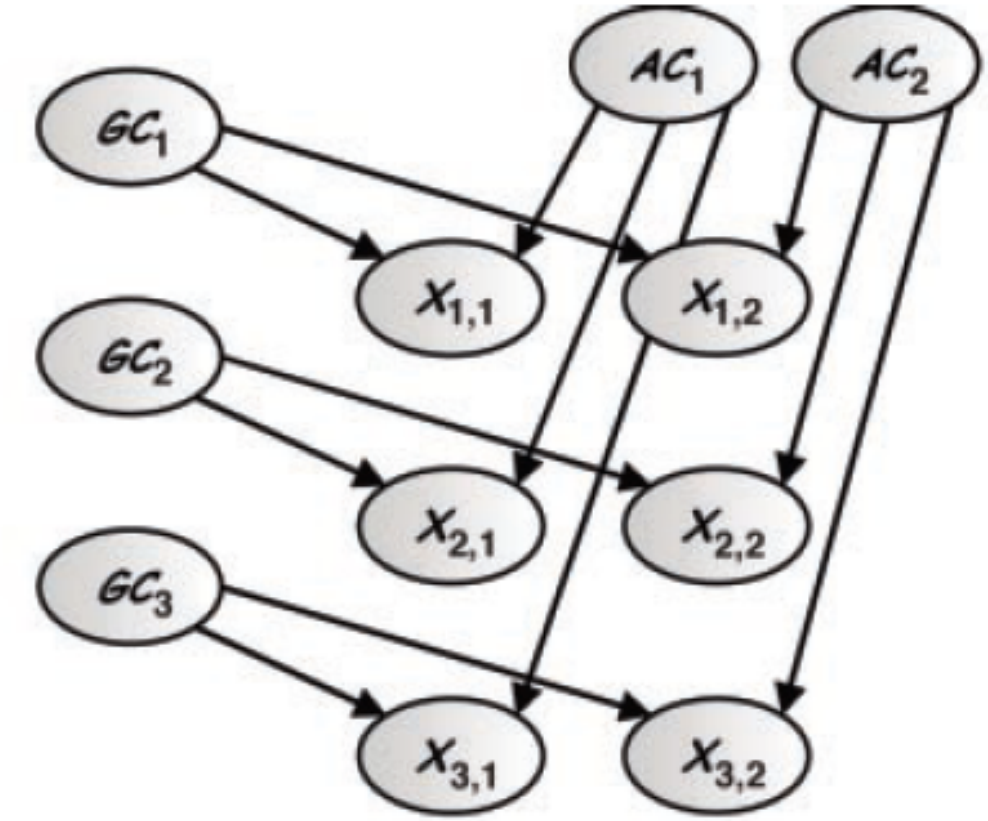
Model for gene expression and cis-regulatory elements

- **Assumptions 1**: genes can be partitioned into clusters of coexpressed genes, and the genes in each cluster have a typical expression level in each array.
- **Assumption 2**: arrays are partitioned into array clusters, which capture relevant biological context, and that the expression of a gene is roughly the same in the arrays that belong to the same array cluster

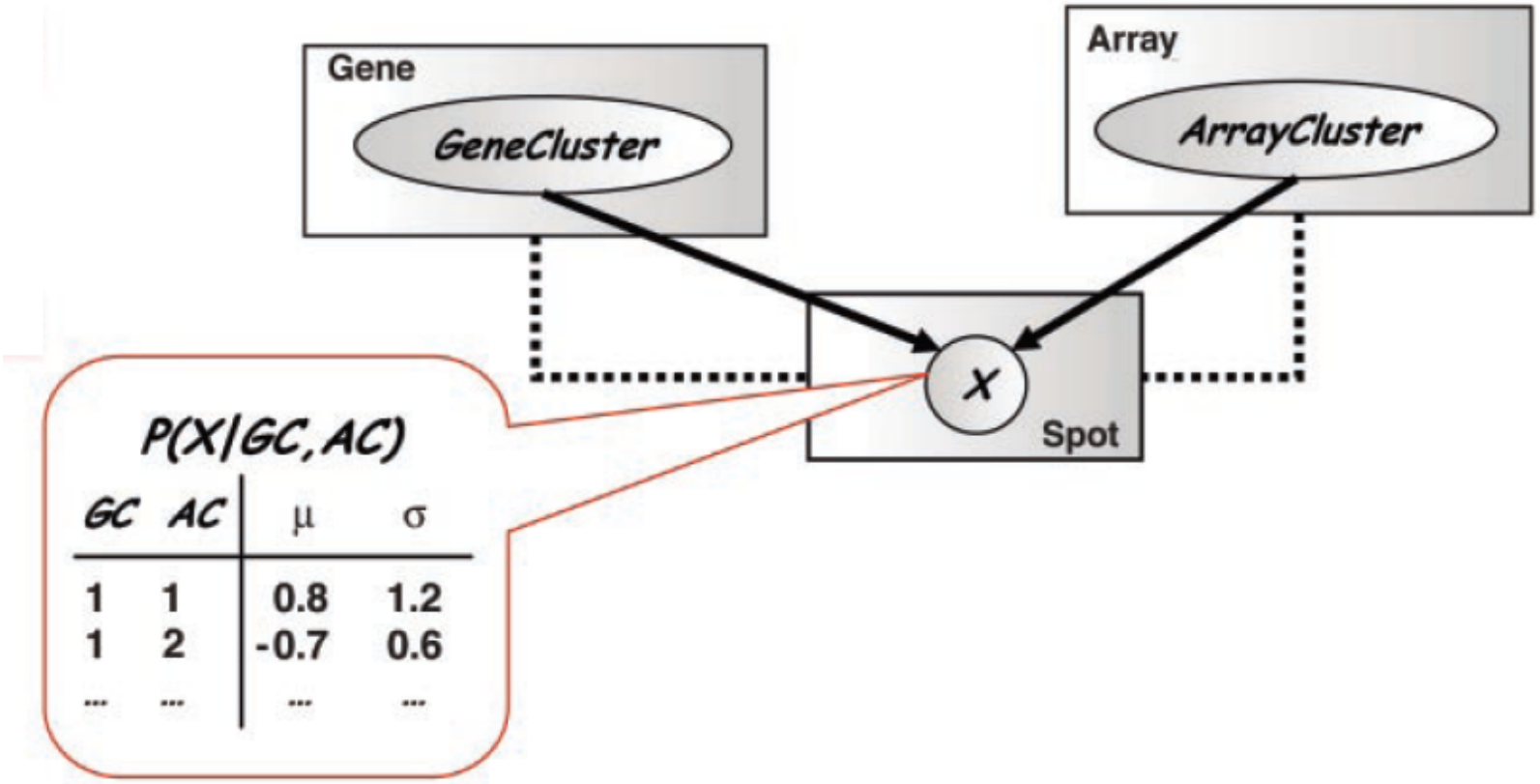
Random Variables

- $X_{g,a}$, where g is an index over gene and a is an index over arrays
- GeneCluster_g: denotes the cluster assignment of gene g
- ArrayCluster_a denotes the cluster assignment of array a.
- Assumption: the expression of gene g in array a depends on the value of GeneClusterg and ArrayClustera

Regular Bayesian Networks



Conditional Distribution



Learning Models from Data

- Parameter estimation – maximum likelihood problem (P(data| model))
- Model selection: select among different model structures to find one that best reflects the dependencies in the domain. P(model | data)

- The model just described can achieve high likelihood if the cluster and gene assignment partitions the original measurements into blocks with approximately uniform expression within each block

- Expectation Maximization procedure that iterates between an **E-step**, which uses current parameters to find the probabilistic cluster assignment of genes and arrays, and an **M-step**, which re-estimates the distribution within each gene/array cluster combination on the basis of this assignment.

Reconstruction of Regulatory Networks

- A key challenge in gene expression analysis is the reconstruction of regulatory networks.
- Distinguish correlation and regulation
- Direct and in-direct regulation

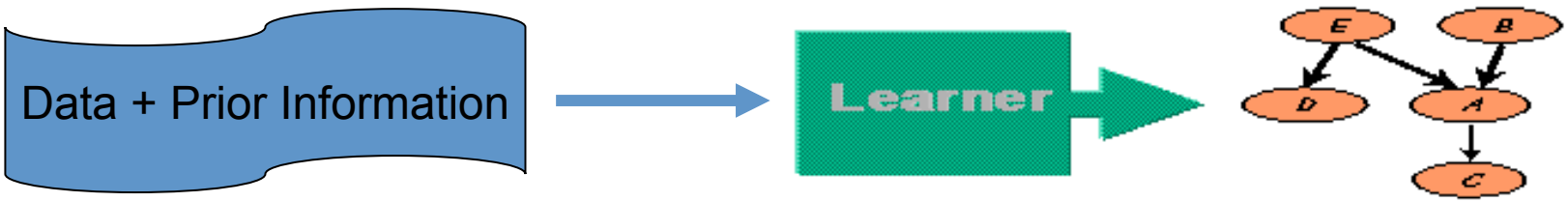
Challenges of Gene Bayesian Network

- Massive number of variables (genes)
- Small number of samples (dozens)
- Sparse networks (only a small number of genes directly affect one another)
- Two crucial aspects: computational complexity and statistical significance of relations in learned models

N. Friedman, 2005

Approach 1: Learning BN from Gene Expression Data

Measured expression level of each gene (discretized) → Random variables Affecting on another

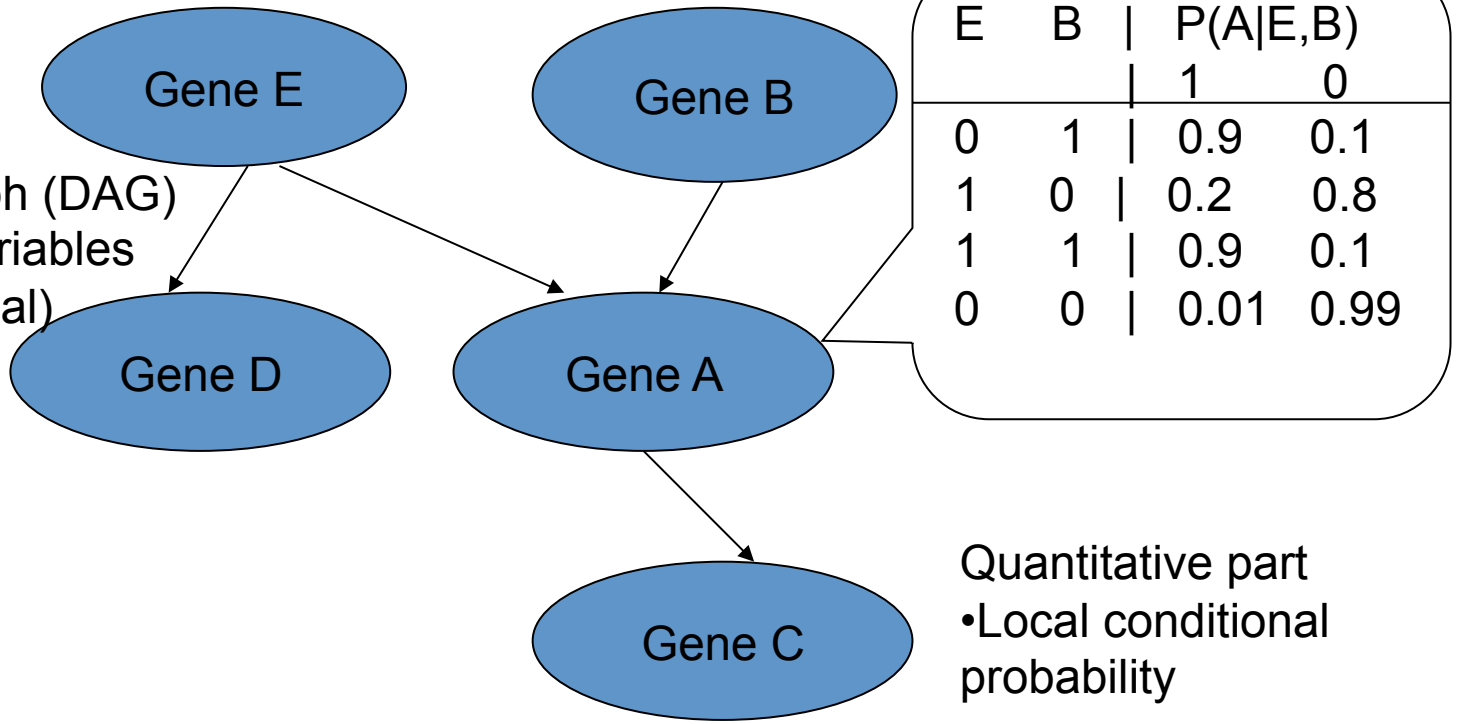


Learn parameters (conditional probabilities) from data
Learn structure (casual relation) from data
Make inference given a learned BN model

N. Friedman, 2005

Gene Bayesian Network

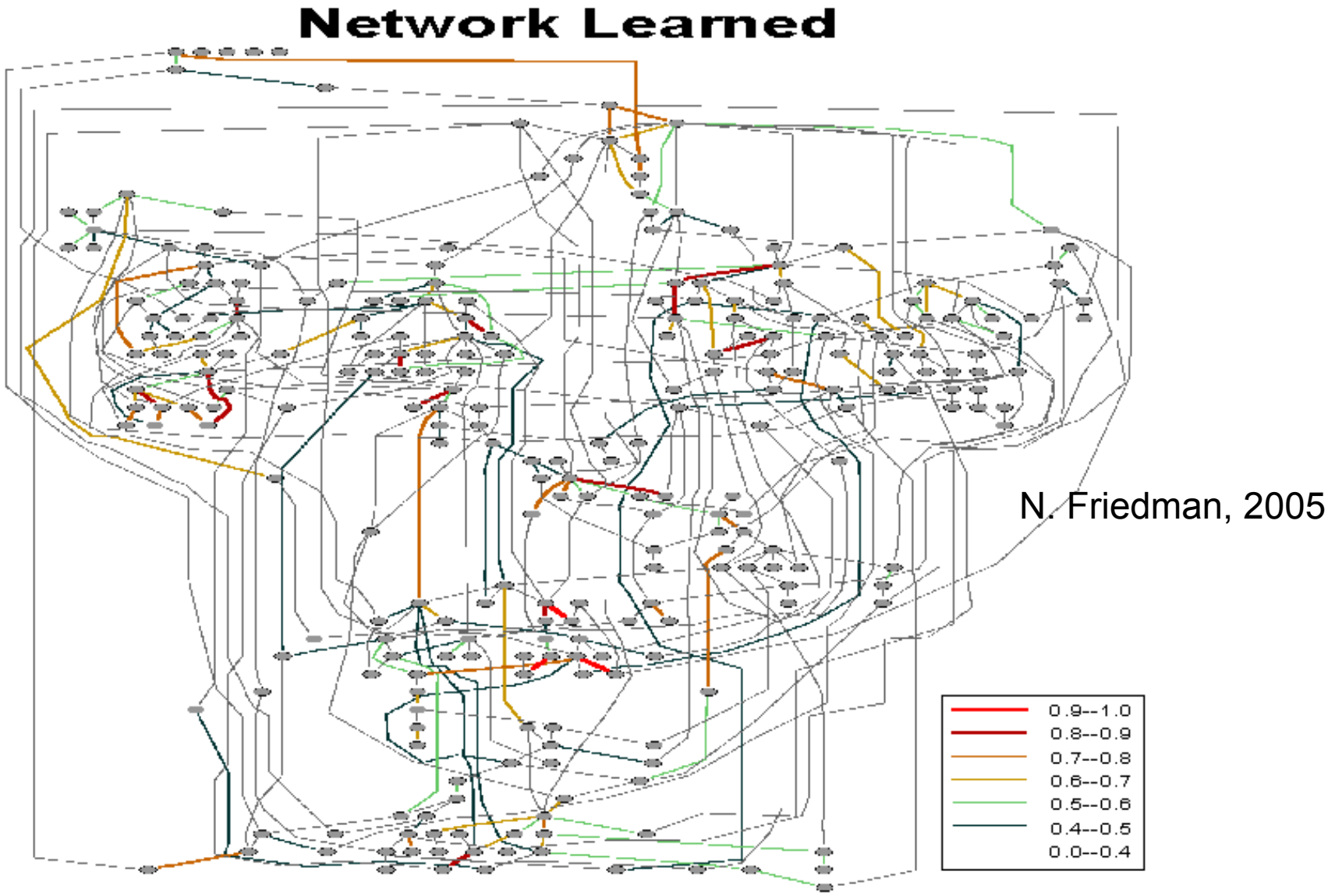
Qualitative Part:
Directed acyclic Graph (DAG)
• Nodes – random variables
•Edges – direct (causal) influence



Quantitative part
•Local conditional probability

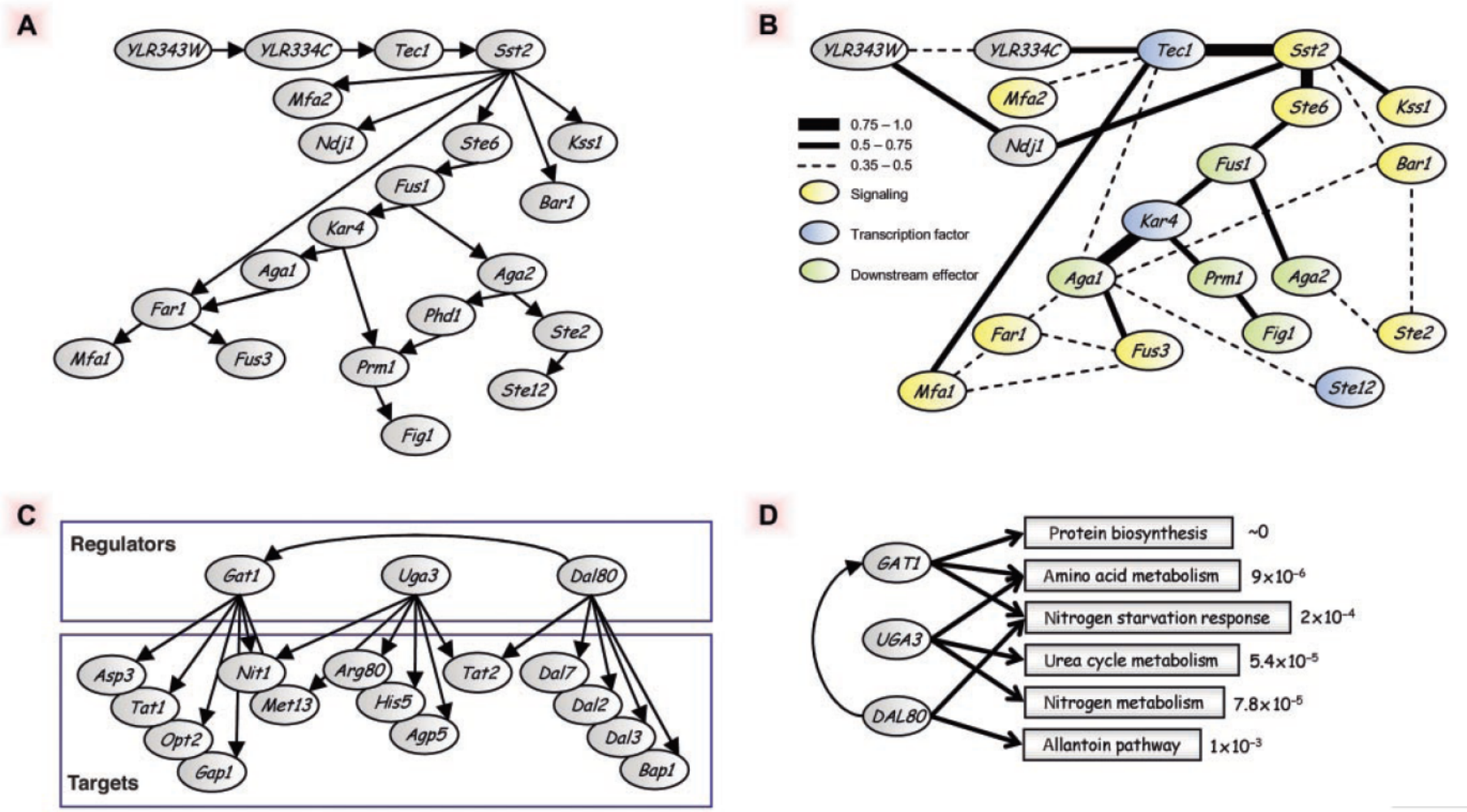
Solutions

- **Sparse candidate algorithm** (by Nir Friedman): Choose a small candidate set for direct influence for each gene. Find optimal BN constrained on candidates. Iteratively improve candidate set.
- **Bootstrap confidence estimate**: use re-sampling to generate perturbations of training data. Use the number of times a relation (or feature) is repeated among networks learned from these datasets to estimate confidence of Bayesian network features.



N. Friedman, 2005

Data: 76 samples of 250 cell-cycle related genes in yeast genome
Discretized into 3 expression levels. Run 100 bootstrap using sparse learning algorithm.
Compute the confidence of features (relations). Most high confident relations make bio-sens


A
YLR343W
YLR334C
Tec1
Sst2
Mfa2
Ndj1
Ste6
Kss1
Fus1
Bar1
Kar4
Aga1
Aga2
Far1
Phd1
Ste2
Mfa1
Fus3
Prm1
Ste12
Fig1
B
0.75 – 1.0
0.5 – 0.75
0.35 – 0.5
Signaling
Transcription factor
Downstream effector
C
Regulators
Gat1
Uga3
Dal80
Asp3
Nit1
Arg80
Tat2
Dal7
Tat1
Met13
His5
Dal2
Opt2
Agp5
Dal3
Gap1
Bap1
Targets
D
GAT1
UGA3
DAL80
Protein biosynthesis
~0
Amino acid metabolism
9×10-6
Nitrogen starvation response
2×10-4
Urea cycle metabolism
5.4×10-5
Nitrogen metabolism
7.8×10-5
Allantoin pathway
1×10-3

Co-Regulation

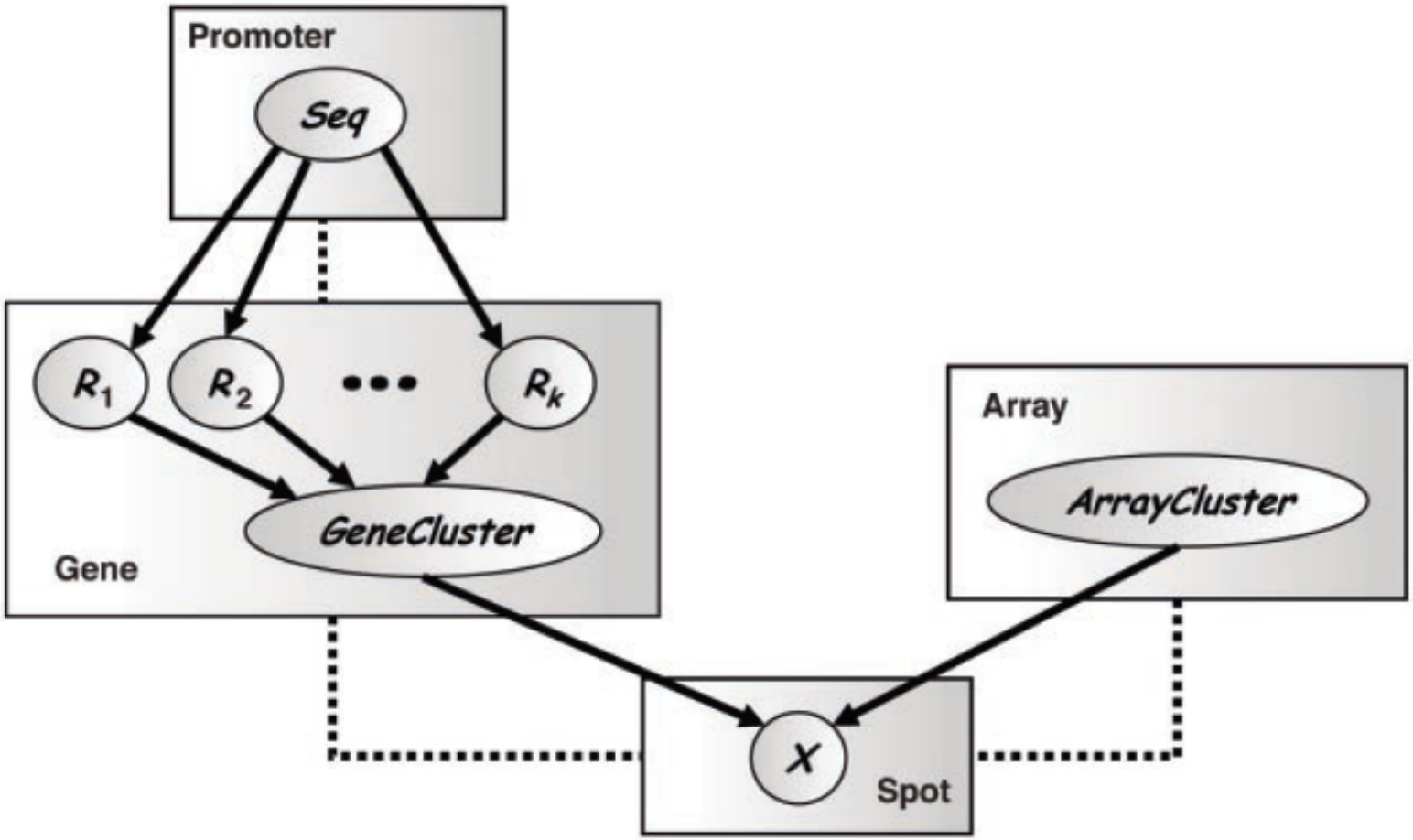
- A key regulation mechanism involves binding of transcription factors to promoter regions of genes.
- Identify the transcription factor binding sites in the promoter region of genes that can explain observed co-expression.

Module Network Approach

A regulatory module is a set of genes that are regulated in convert by a shared regulation program.

A regulation program specifies the behavior of the genes in the module as a function of the expression level of a small set of regulators

Regulatory Model



$R_{g,j}$ as depending on the promoter sequence *Seq_g*

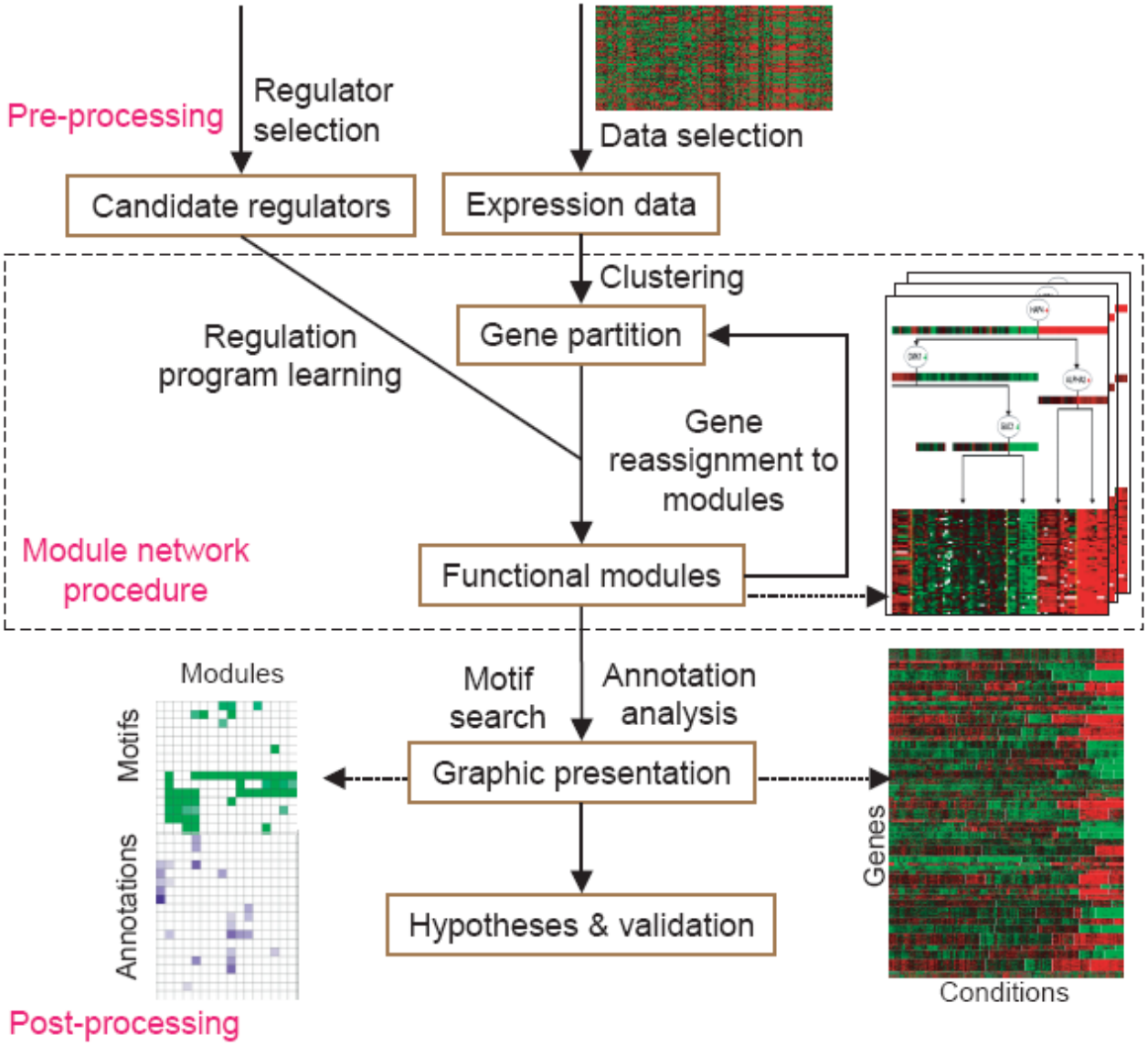
Integration of Sequence and Expression Data

- The parameters of this conditional probability characterize the specific motif recognized by the transcription factor. This **extension** allows us to learn the characterization of the binding site while learning how its presence influences gene expression.

Procedure

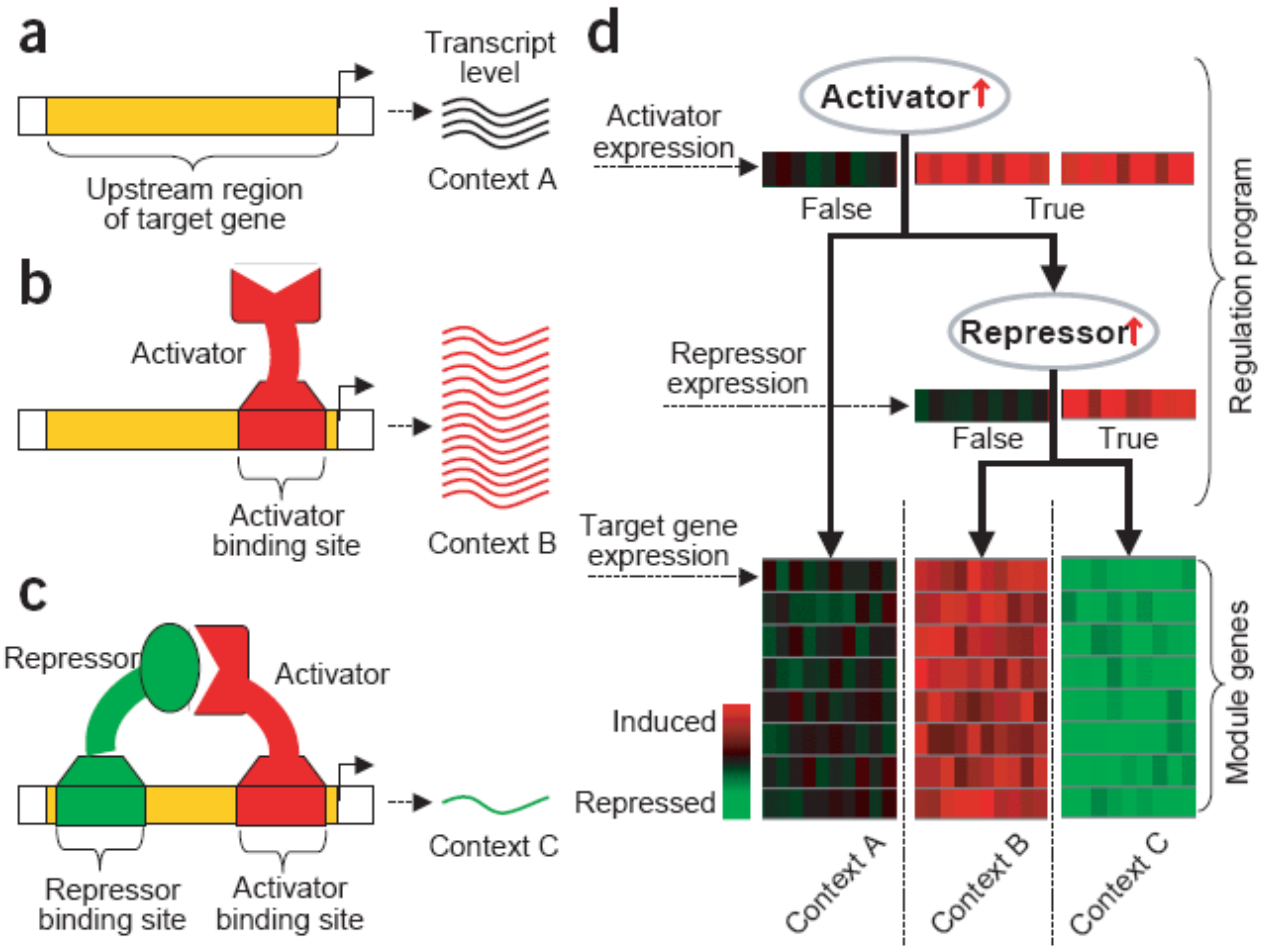
- **Inputs**: a gene expression data set and

  a large precompiled set of candidate regulatory genes for the corresponding organism (independent of data set) containing both know and putative transcription factors and signal transduction molecules
- **Goal**: search for a partition of genes into modules and for a regulation program for each module
- **Output**: a list of modules and associated regulation programs

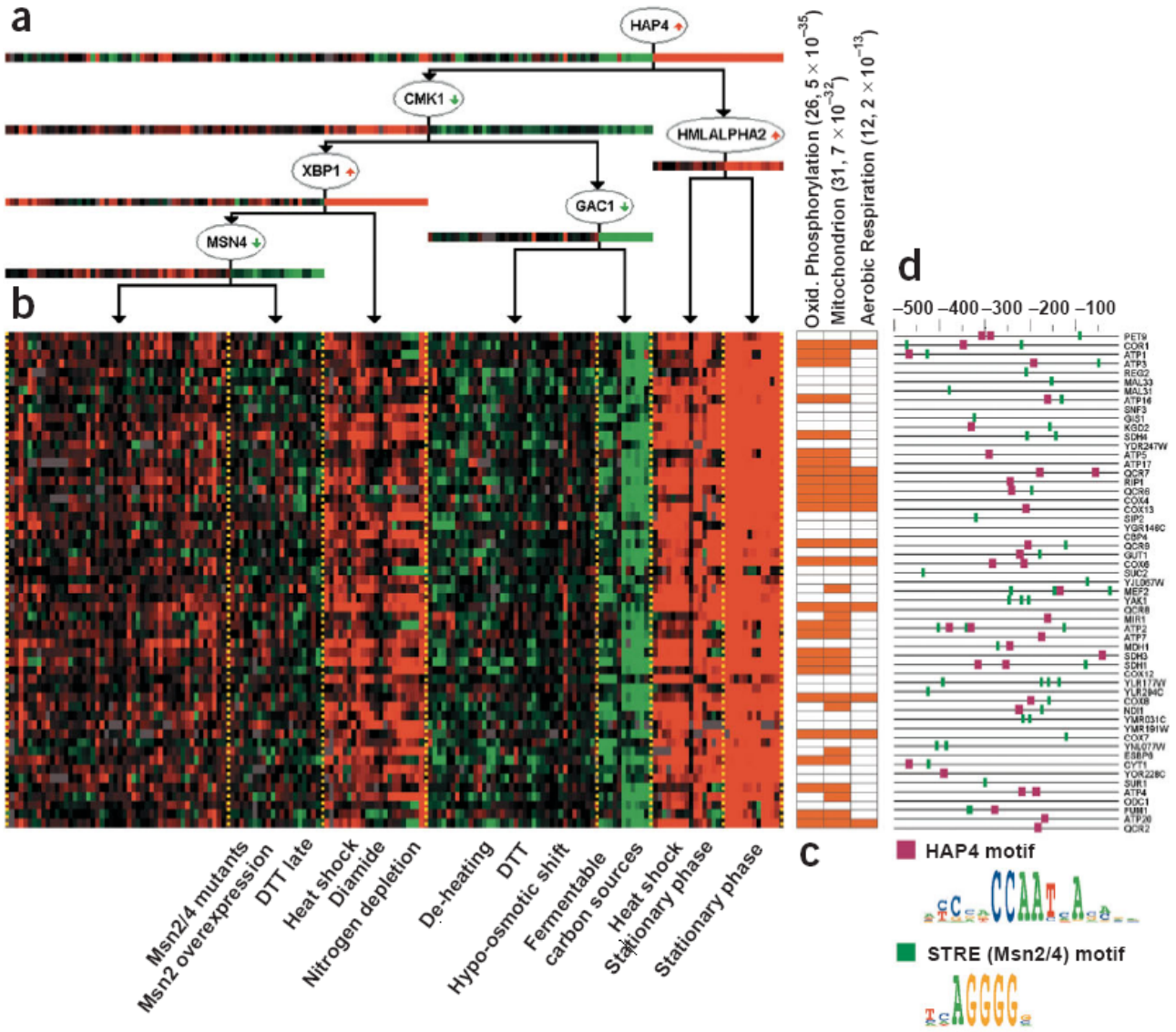

Pre-processing
Regulator selection
Data selection
Candidate regulators
Expression data
Clustering
Gene partition
Regulation program learning
Gene reassignment to modules
Module network procedure
Functional modules
Modules
Motif search
Annotation analysis
Motifs
Annotations
Graphic presentation
Genes
Conditions
Hypotheses & validation
Post-processing

- **Results**: apply the method to Yeast gene expression data set consisting of 2355 genes and 173 arrays.
- Each inferred modules contained a functionally coherent set of genes (metabolic pathways, oxidative stress, cell cycle-related processes, etc)
- Many module has a match between predicted regulator and its known cis-regulatory binding motif.

One Example


a
Transcript level
Upstream region of target gene
Context A
b
Activator
Activator binding site
Context B
c
Repressor
Activator
Repressor binding site
Activator binding site
Context C
d
Activator
Activator expression
False
True
Repressor
Repressor expression
False
True
Regulation program
Target gene expression
Induced
Repressed
Module genes
Context A
Context B
Context C

Row: genes
Column: arrays



Evaluation of Module Content and Regulation Program

- We evaluate all 50 modules to test whether the proteins encoded by genes in the same module had related functions. We scored the functional/biological coherence of each module according to percentage of its genes covered by annotations. Most of modules had a coherence level above 50%.

#	Module[a]	# G[b]	C (%)[c]	Reg.[d]	M	C	G	Reg.[d]	M	C	G	Reg.[d]	M	C	G	Reg.[d]	M	C	G	Reg.[d]	M	C	G	Reg.[d]	M	C	G
1	Respiration and carbon regulation	55	84	Hap4				HMLAlpha2				Cmk1				Gac1				Xbp1				Msn4			
2	Energy, osmolarity and cAMP signaling	64	64	Tpk1				Kin82				Yer184c				Cmk1				Ppt1				Kns1			
3	Energy and osmotic stress I	31	65	Xbp1				Kin82				Tpk1															
4	Energy and osmotic stress II	42	38	Ypl230w				Yap6				Gac1				Wsc4											
5	Glycolysis and folding	37	86	Gcn20				Ecm22				Bmh1				Bas1											
6	Galactose metabolism	4	100	Gal4				Gac1				Hir3				Ime4											
7	Snf kinase regulated processes	74	47	Ypl230w				Yap6				Tos8				Sip2											
8	Nitrogen catabolite repression	29	66	Gat1				Plp2																			
9	Amino acid metabolism I	39	95	Gat1				Ime4				Cdc20				Slt2											
10	Amino acid metabolism II	37	95	Xbp1				Hap4				Afr1				Uga3				Ppt1							
11	Amino acid and purine metabolism	53	92	Gat1				Ppz2				Rim11															
12	Nuclear	47	47	HMLAlpha2				Ino2																			
13	Mixed I	28	50	Pph3				Ras2				Tpk1															
14	Ribosomal and phosphate metabolism	32	81	Ppt1				Sip2				Cad1															
15	mRNA,rRNA and tRNA processing	43	40	Lsg1				Tpk2				Ppt1															
16	RNA processing and cell cycle	59	36	Ypl230w				Ime4				Ppt1				Tpk2				Rho2				Mcm1			
17	DNA and RNA processing	77	43	Tpk1				Gis1				Ppt1															
18	TFs and RNA processing	59	68	Gis1				Pph3				Tpk2				Lsg1											
19	TFs and nuclear transport	48	56	Ypl230w				Met18				Ppt1															
20	TFs I	53	92	Cdc14				Mcm1				Ksp1															
21	TFs II	50	54																								
22	TFs, cell wall and mating	39	59	Ptc3				Sps1																			
23	TFs and sporulation	43	60	Rcs1				Ypl133c																			
24	Sporulation and TFs	74	39	Gcn20				Gat1				Ste5															
25	Sporulation and cAMP pathway	59	37	Xbp1				Ypl230w				Sip2				Not3											
26	Sporulation and cell wall	78	40	Ypl230w				Yap6				Msn4															
27	Cell wall and transport I	23	48	Shp1				Bcy1				Gal80				Ime1				Yak1							
28	Cell wall and transport II	63	46	Ypl230w				Kin82				Msn4															
29	Cell differentiation	41	71	Ypl230w				Ypk1				Cna1															
30	Cell cycle (G2/M)	30	70	Cdc14				Clb1				Far1															
31	Cell cycle, TFs and DNA metabolism	71	85	Gis1				Ste5				Clb5															
32	Cell cycle and general TFs	64	72	Ime4				Ume1				Xbp1				Prr1				Cnb1				Arp9			
33	Mitochondrial and signalling	87	60	Tpk1				Cmk1				Yer184c				Gis1											
34	Mitochondrial and protein fate	37	78	Ypk1				Sds22				Rsc3															
35	Trafficking and mitochondrial	87	56	Tpk1				Sds22				Etr1															
36	ER and nuclear	79	86	Gcn20				Yjl103c				Not3				Tup1											
37	Proteasome and endocytosis	31	71	Ime4				Cup9				Bmh2				Hrt1											
38	Protein modification and trafficking	62	79	Ypl230w				Ptc3				Cdc42															
39	Protein folding	23	87	Bmh1				Bcy1				Ypl230w															
40	Oxidative stress I	15	80	Yap1				Sko1				Far1															
41	Oxidative stress II	15	73	Tos8				Flo8																			
42	Unkown (sub-telomeric)	82	45	Gcn20																							
43	Unknown genes I	36	42																								
44	Unknown genes II	29	14	Apg1				Pcl10																			
45	Unknown genes III	39	5	Xbp1				Kar4																			
46	Mixed II	52	42	Gcn20				Tos8				Sip2															
47	Mixed III	41	63	Gcn20				Ume1				Cnb1															
48	Mixed IV	35	29	Fkh1				Sho1																			
49	Ty ORFs	16	6																								
50	Missing values	64	39																								

Enrichment for motif known to participate in regulation by respective regulator

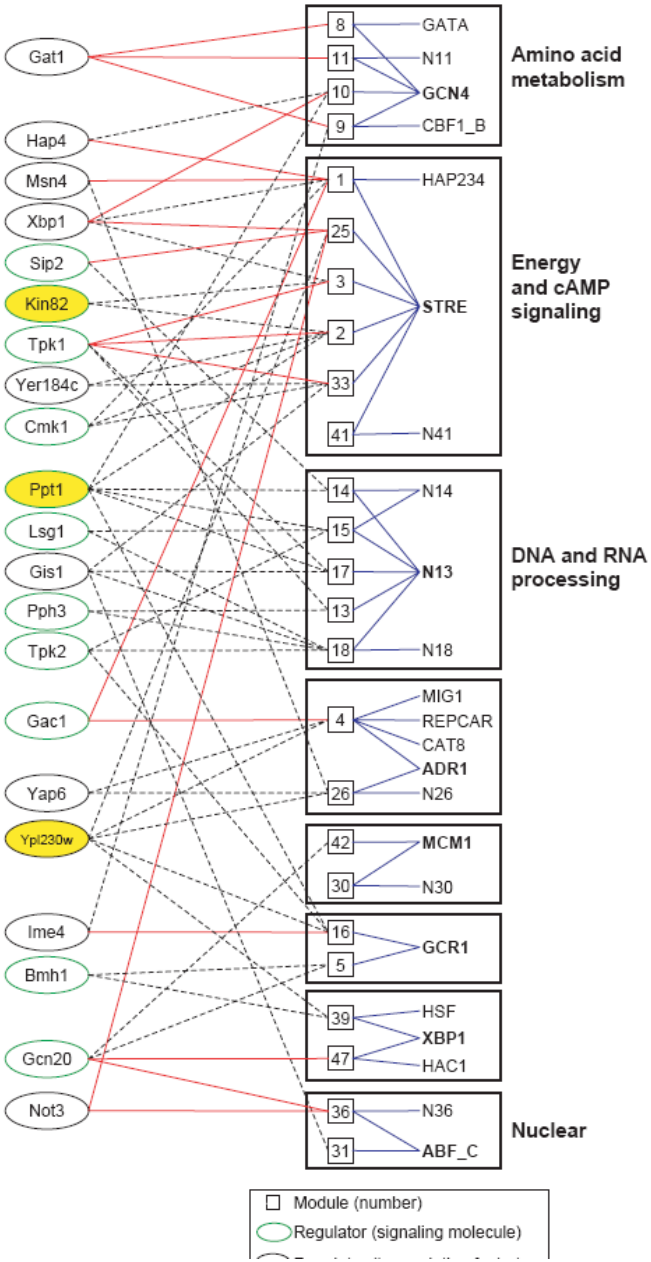
Respective regulator known to have a role under the predicted condition

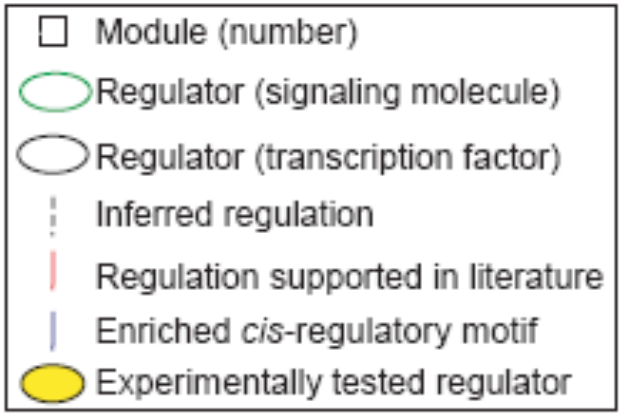
Respective regulator known to regulate module genes or their implied process

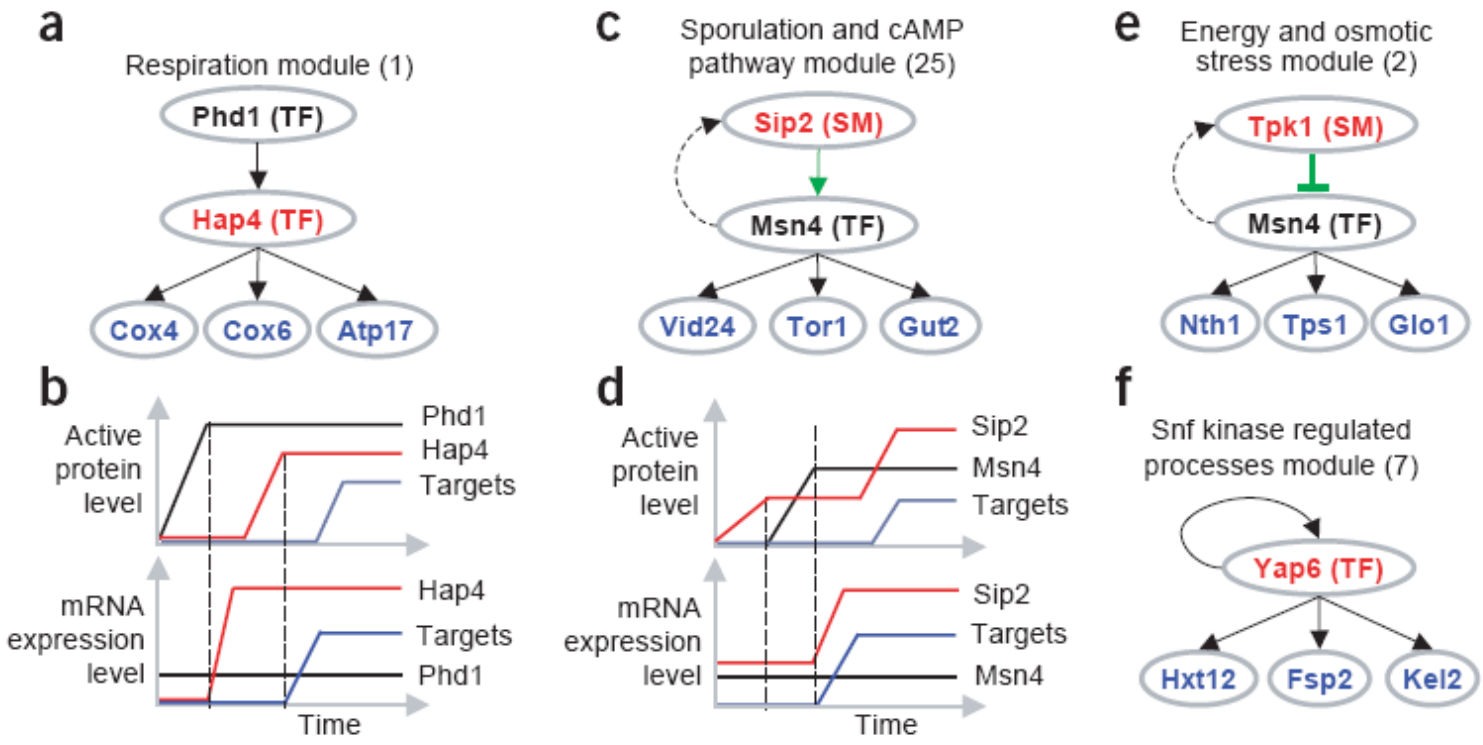
Partial evidence

Partial evidence

Partial evidence


Gat1
Hap4
Msn4
Xbp1
Sip2
Kin82
Tpk1
Yer184c
Cmk1
Ppt1
Lsg1
Gis1
Pph3
Tpk2
Gac1
Yap6
Ypl230w
Ime4
Bmh1
Gcn20
Not3
8
11
10
9
GATA
N11
GCN4
CBF1_B
Amino acid metabolism
1
25
3
2
33
41
HAP234
STRE
N41
Energy and cAMP signaling
14
15
17
13
18
N14
N13
N18
DNA and RNA processing
4
26
MIG1
REPCAR
CAT8
ADR1
N26
42
30
MCM1
N30
16
5
GCR1
39
47
HSF
XBP1
HAC1
36
31
N36
ABF_C
Nuclear
Module (number)
Regulator (signaling molecule)


Module (number)
Regulator (signaling molecule)
Regulator (transcription factor)
Inferred regulation
Regulation supported in literature
Enriched *cis*-regulatory motif
Experimentally tested regulator


a
Respiration module (1)
Phd1 (TF)
Hap4 (TF)
Cox4
Cox6
Atp17
b
Active protein level
Phd1
Hap4
Targets
mRNA expression level
Hap4
Targets
Phd1
Time
c
Sporulation and cAMP pathway module (25)
Sip2 (SM)
Msn4 (TF)
Vid24
Tor1
Gut2
d
Active protein level
Sip2
Msn4
Targets
mRNA expression level
Sip2
Targets
Msn4
Time
e
Energy and osmotic stress module (2)
Tpk1 (SM)
Msn4 (TF)
Nth1
Tps1
Glo1
f
Snf kinase regulated processes module (7)
Yap6 (TF)
Hxt12
Fsp2
Kel2

Candidate regulators

- Compiled a set of 466 candidate regulators annotated in Yeast Genome and Proteome databases
- Use Yeast gene expression data set consisting of 173 microarrays that measure responses to various stress conditions.
- We downloaded these data in log (base 2) ratio to control format from Stanford Microarry Database. Chose a subset of 2355 genes that have a significant change in gene expression under the measured stress conditions

- **Protein annotations**: downloaded Gene Ontology and Munich Information center for Protein Sequence (MIPS) function and KEGG.
- **Regulation program**: Regression tree (decision nodes and leaf nodes); the model semantics is that given a gene g in the module and an array a in a context, the probability of observing some expression value for a gene in array is governed by the normal distribution specified for the context.

Learning Module Networks

- In each iteration, the procedure searches for a regulation program for each module and then reassign each gene to the module whose program best predicts its behavior. Repeated until it converges.
- Search for the model with the highest score by using the EM algorithm.

EM Algorithm

- **M-Step**: given a partition of genes into modules and learns the best regulation program (regression tree) for each module. The regulation program is learned through a combinatorial search over the space of trees. The tree is grown from the root to its leaves. At any given node, the query that best partitions the gene expression into two distinct distribution is chosen.

- **E-step**: given the inferred regulation programs, we determine the module whose associated regulation program best predicts each gene's behavior. Select the module whose program gives the gene's expression profile the highest probability and re-assign the gene to this module.
- We initialize our modules to 50 clusters using Pcluster, a hierahical agglomerative clustering. We then applied the EM algorithm to this starting point, refining both the gene partitioin and the regulatory program.

Evaluating statistical significance of modules

- All of the statistical evaluations were done and visualized in GeneXPress. The tool can evaluate the output of any clustering program for enrichment of gene annotations and motifs

Annotation enrichment

- We associated each gene with the processes in which it participates. Resulted in 923 GO categories, 208 MIPS categories, and 87 KEGG pathways. For each module and for each annotation, we calculated the fraction of genes in the module associated with that annotation and used the hypergeometric distribution to calculate a P-value for this fraction.

Promoter Analysis

- We search for motifs (represented as Position-Specific Scoring Matrices) within 500 bp upstream of each gene. We downloaded TRANSFAC, containing 34 known function cis-regulatory motifs. We also use a motif finder to find 50 potentially novel motifs.

Motif Combination

- We searched for statistically significant occurrences of motif pairs. We constructed a motif pair attribute, which assigns a “true” value for each gene if and only if both motifs of the pair are found in the upstream region of that gene. For each module and for each motif pair attribute, we calculated the fraction of genes in the module associated with that attribute and used the hypergeometric distribution to calculate a P value for this fraction.

Regulator Annotations

- We associate regulators with annotations and binding sites in the same way we associate with these attributes to the modules. Because a regulator may regulate more than one module, its targets consist of the union of the genes in all modules predicted to be regulated by that regulator. We tested the targets of each regulator for enrichment of the same motifs and gene annotations as above using the hypergeometric P value.