02-710
Computational Genomics

Systems biology

Putting it together: Data integration using graphical models

High throughput data

- So far in this class we discussed several different types of high throughput datasets, each providing an important, but limited, view of cellular activity. These include:

- Coding sequences: genes, exons, miRNAs
- Non coding sequences: enhancers, DNA motifs
- Gene and microRNA expression: microrarrays, RNA-Seq
- Protein-DNA interactions: ChIP-CHIP, CHIP-Seq, PBM
- Epigenetic data
- Etc.

Systems Biology: Motivation

High-level goal: Integrate different types of 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.

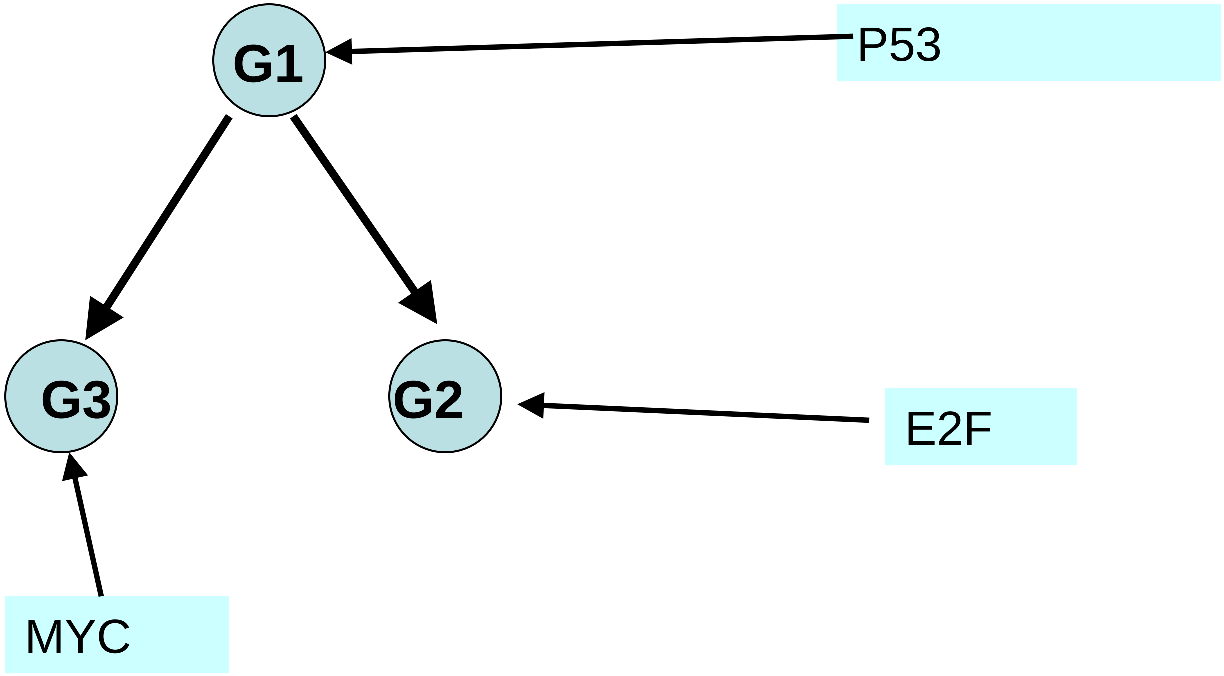
Graphical models

Independence

- Independence allows for easier models, learning and inference (for example, when using a Naïve Bayes classifier)
- For example, with 3 binary variables we only need 3 parameters rather than 7.
- The saving is even greater if we have many more variables …
- In many cases it would be useful to assume independence, even if its not the case
- Is there any middle ground?

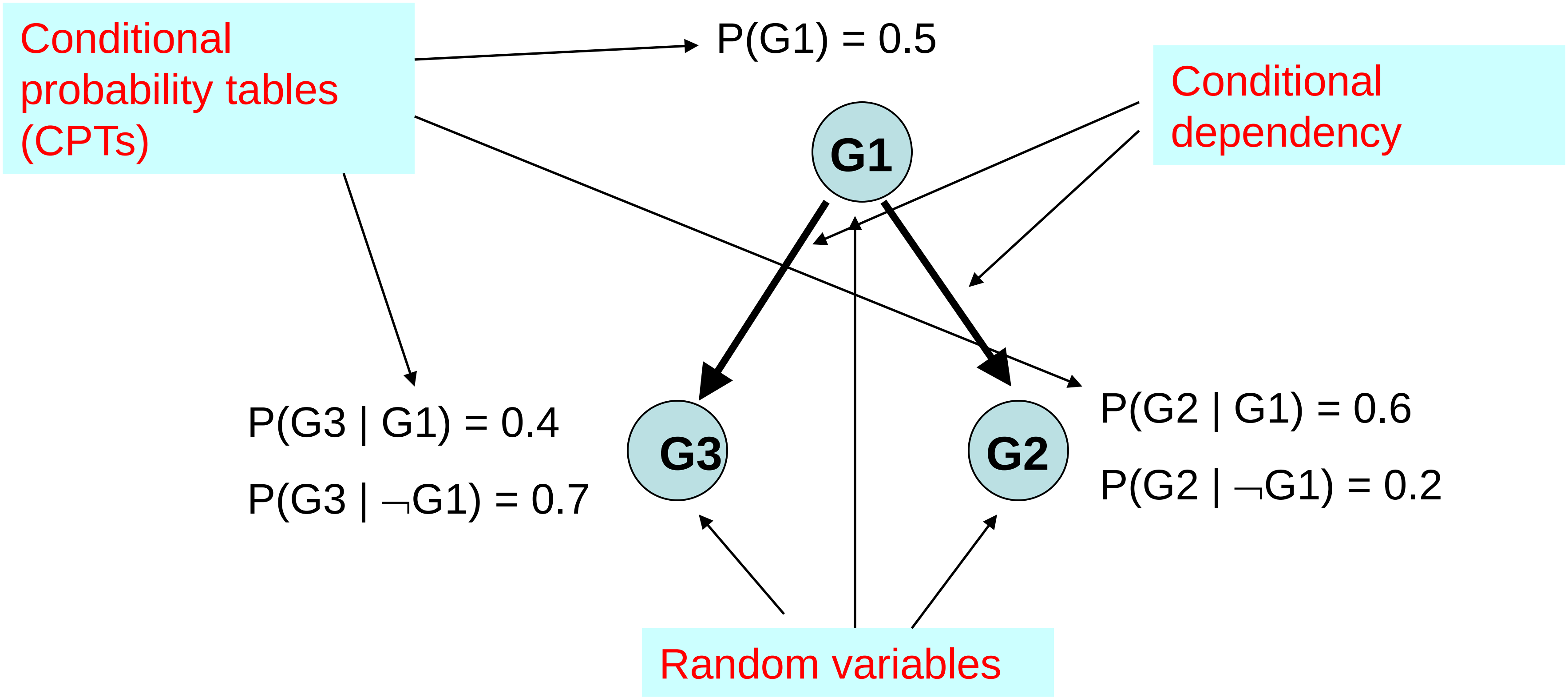
Bayesian networks

- Bayesian networks are *directed graphs* with nodes representing *random variables* and edges representing *dependency assumptions*
- Lets use our movie example: We would like to determine the joint probability for length, liked and slept in a movie



Bayesian networks: Notations

Bayesian networks are directed acyclic graphs.



Conditional probability tables (CPTs)

$P(G1) = 0.5$

Conditional dependency

$P(G3 \mid G1) = 0.4$

$P(G3 \mid \neg G1) = 0.7$

$P(G2 \mid G1) = 0.6$

$P(G2 \mid \neg G1) = 0.2$

Random variables

Bayesian networks: Notations

The Bayesian network below represents the following joint probability distribution:

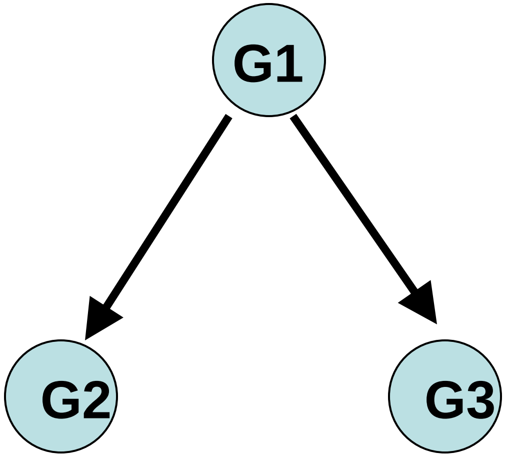
$$p(G1, G2, G3) = P(G1)P(G2 \mid G1)P(G3 \mid G1)$$

More generally Bayesian network represent the following joint probability distribution:

$$p(x_1 \cdots x_n) = \prod_i p(x_i \mid Pa(x_i))$$

The set of parents of x_i in the graph

P(G1) = 0.5



P(G2 | G1) = 0.4

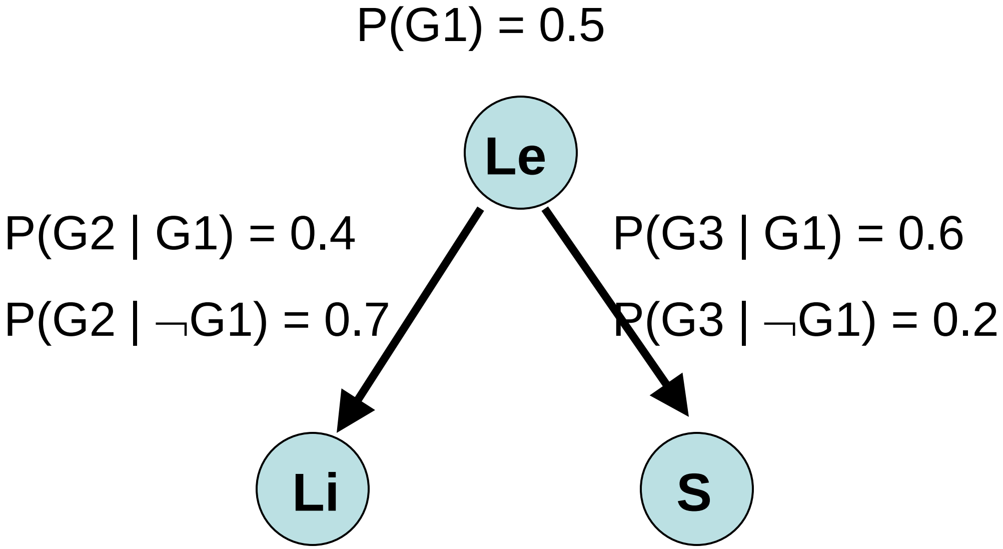
P(G2 | ¬G1) = 0.7

P(G3 | G1) = 0.6

P(G3 | ¬G1) = 0.2

Bayesian network: Inference

- Once the network is constructed, we can use algorithms for inferring the values of unobserved variables.
- For example, assume we only observed G2 and G3.
- Can we determine the value of G1?



P(G1) = 0.5

P(G2 | G1) = 0.4

P(G2 | ¬G1) = 0.7

P(G3 | G1) = 0.6

P(G3 | ¬G1) = 0.2

Methods for grouping genes in clusters and networks

- Clustering of expression data
  - Groups together genes with similar expression patterns
  - Does not reveal structural relations between genes
- Boolean networks
  - Deterministic models of the logical interactions between genes
  - Deterministic, static
- Linear models
  - Deterministic fully-connected linear model
  - Under-constrained, assumes linearity of interactions

So, Why Bayesian Networks?

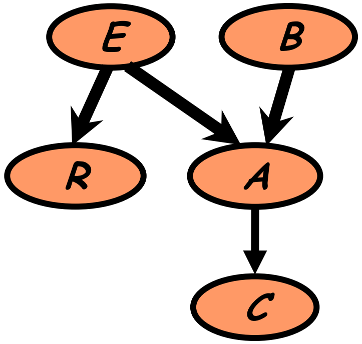
- Flexible representation of (**in**)**dependency structure** of multivariate distributions and interactions
- Natural for modeling **global processes** with **local interactions** => good for biology
- Clear probabilistic semantics
- Natural for statistical **confidence analysis** of results and answering of queries
- **Stochastic** in nature: models stochastic processes & deals (“sums out”) noise in measurements

Learning Bayesian Network

The goal:

- Given set of independent samples (***assignments*** to random variables), find the ***best*** (most likely) Bayesian Network
(both DAG and CPDs)

{ (B,E,A,C,R)=(T,F,F,T,F)
(B,E,A,C,R)=(T,F,T,T,F)
……..
(B,E,A,C,R)=(F,T,T,T,F) }



E	B	P(A | E,B)	
e	b	0.9	0.1
e	$\bar{b}$	0.2	0.8
$\bar{e}$	b	0.9	0.1
$\bar{e}$	$\bar{b}$	0.01	0.99

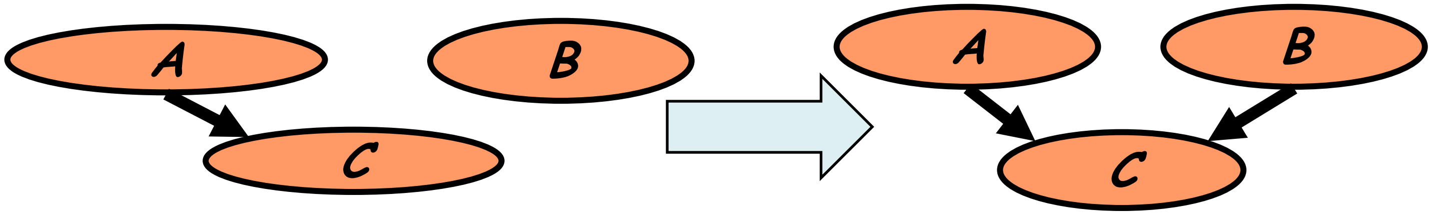
Learning Bayesian Network

• Learning of best CPTs *given DAG* is easy (collect statistics of values of each node given specific assignment to its parents). But…

•The structure (G) learning problem is NP-hard => heuristic search for best model must be applied, generally bring out a **locally** optimal network.

•It turns out, that the richer structures give higher likelihood P(D|G) to the data (adding an edge is always preferable), because…

Learning Bayesian Network



- If we add B to Pa(C) , we have more parametes to fit => more freedom => can always optimize SPD(C) , such that:

$$P(C \mid A) \leq P(C \mid A, B)$$

- But we prefer *simpler* (more explanatory) networks (Occam's razor!)

•Therefore, **practical** scores of Bayesian Networks compensate the likelihood improvement by a "penalty" on complex networks.

Modeling Biological Regulation

Variables of interest:

- Expression levels of genes
- Concentration levels of proteins
- Exogenous variables: Nutrient levels, Metabolite Levels, Temperature
- Phenotype information
- …

Bayesian Network Structure:

- Capture dependencies among these variables

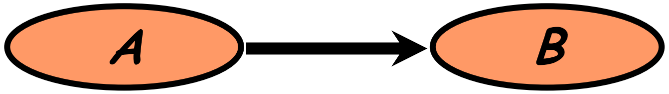
Possible Biological Interpretation

Measured expression level of each gene → **Random variables**

Gene interaction → **Probabilistic dependencies**

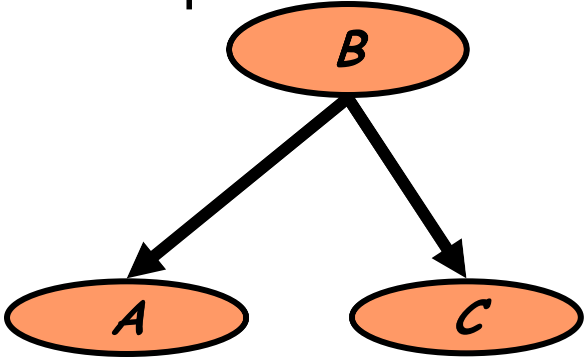
Interactions are represented by a graph:

- Each gene is represented by a node in the graph
- Edges between the nodes represent direct dependency

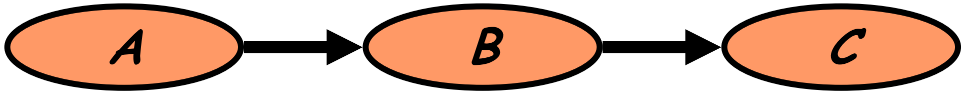


More Local Structures

- Dependencies can be mediated through other nodes

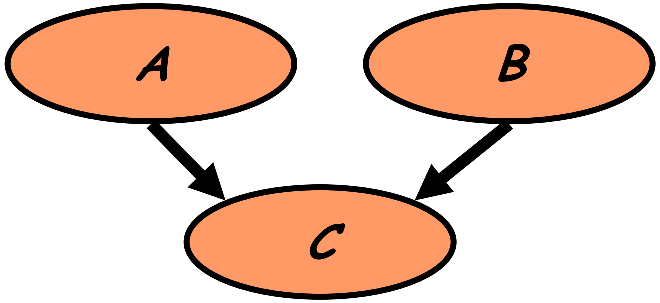


Common cause

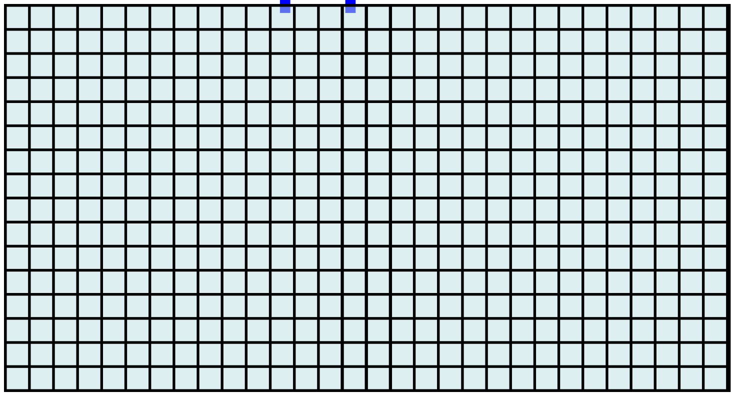


Intermediate gene

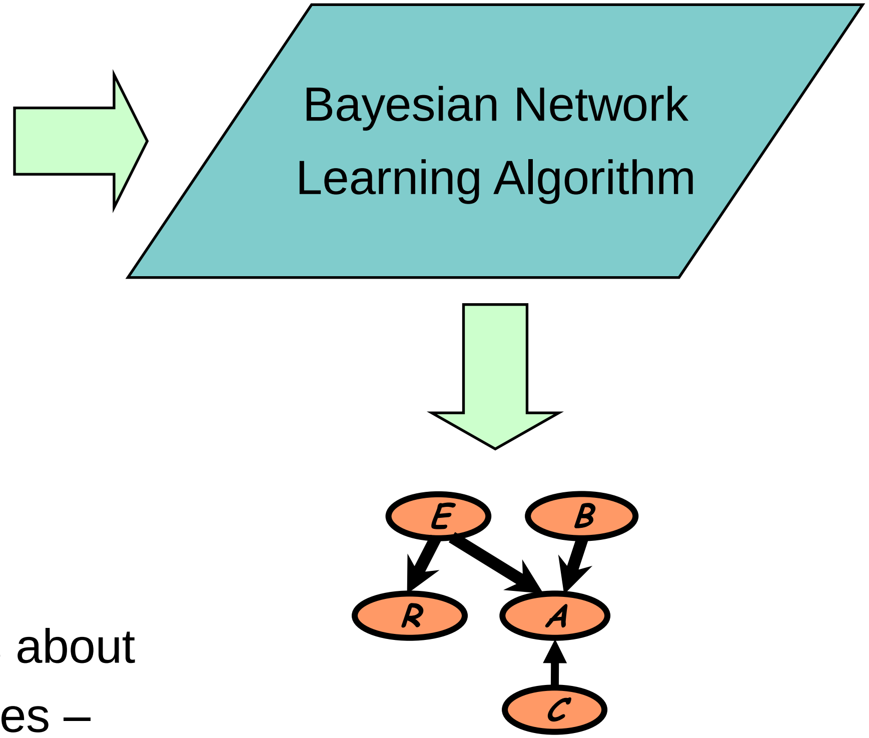
- Common/combinatorial effects:



The Approach of Friedman et al.



Expression data



Bayesian Network Learning Algorithm

Use learned network to make predictions about structure of the interactions between genes –

No prior biological knowledge is used!

Friedman et al, Using Bayesian networks to analyze expression data, *RECOMB* 2000

The Discretization Problem

◆The expression measurements are **real numbers**.

=> We can either discretize the values in order to learn general CPTs => lose information

=> If we don’t, we must assume some specific type of CPT (like regression based linear Gaussian models) => lose generality

Problem of Sparse Data

◆**There are much more genes than experiments** => many different networks suit the data well

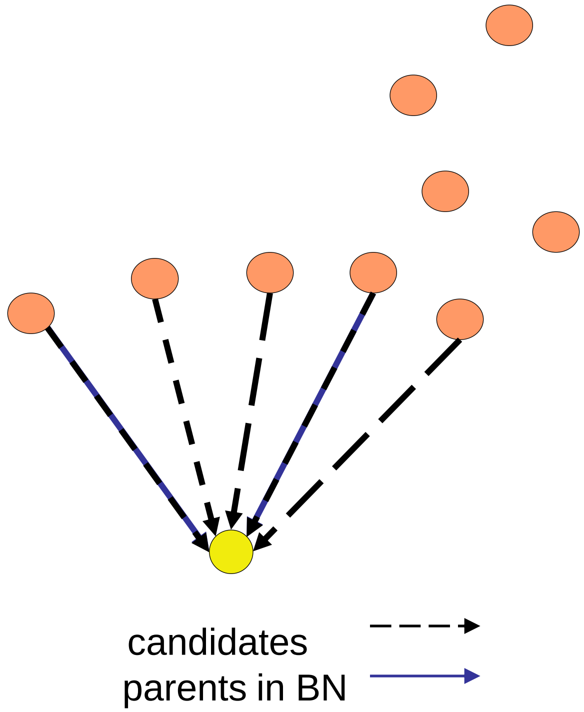
●**Shrink the network search space.** E.g., we can use the notion, that in biological systems each gene is regulated directly by only a few regulators.

●Don’t take for granted the resulting network, but instead **fetch from it pieces of reliable information**.

Learning With Many Variables

Sparse Candidate algorithm - efficient heuristic search, relies on sparseness of regulation nets.

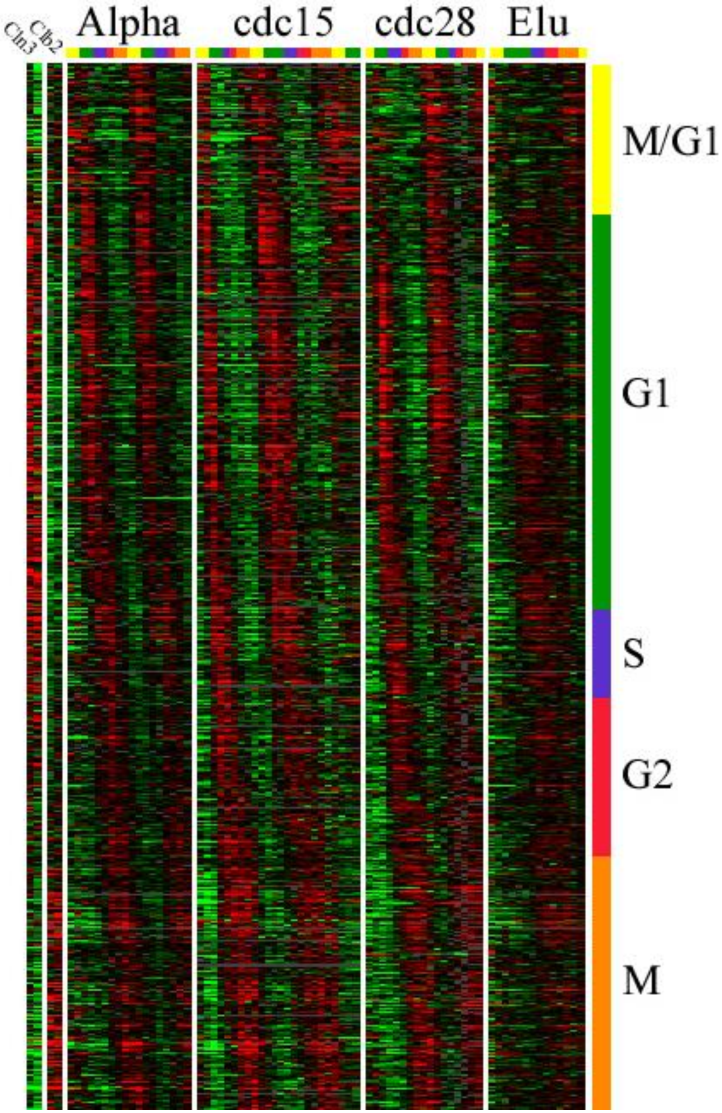
- For each gene, choose promising “candidate parents set” for direct influence for each gene
- Find (locally) optimal BN constrained on those parent candidates for each gene
- Iteratively improve candidate set



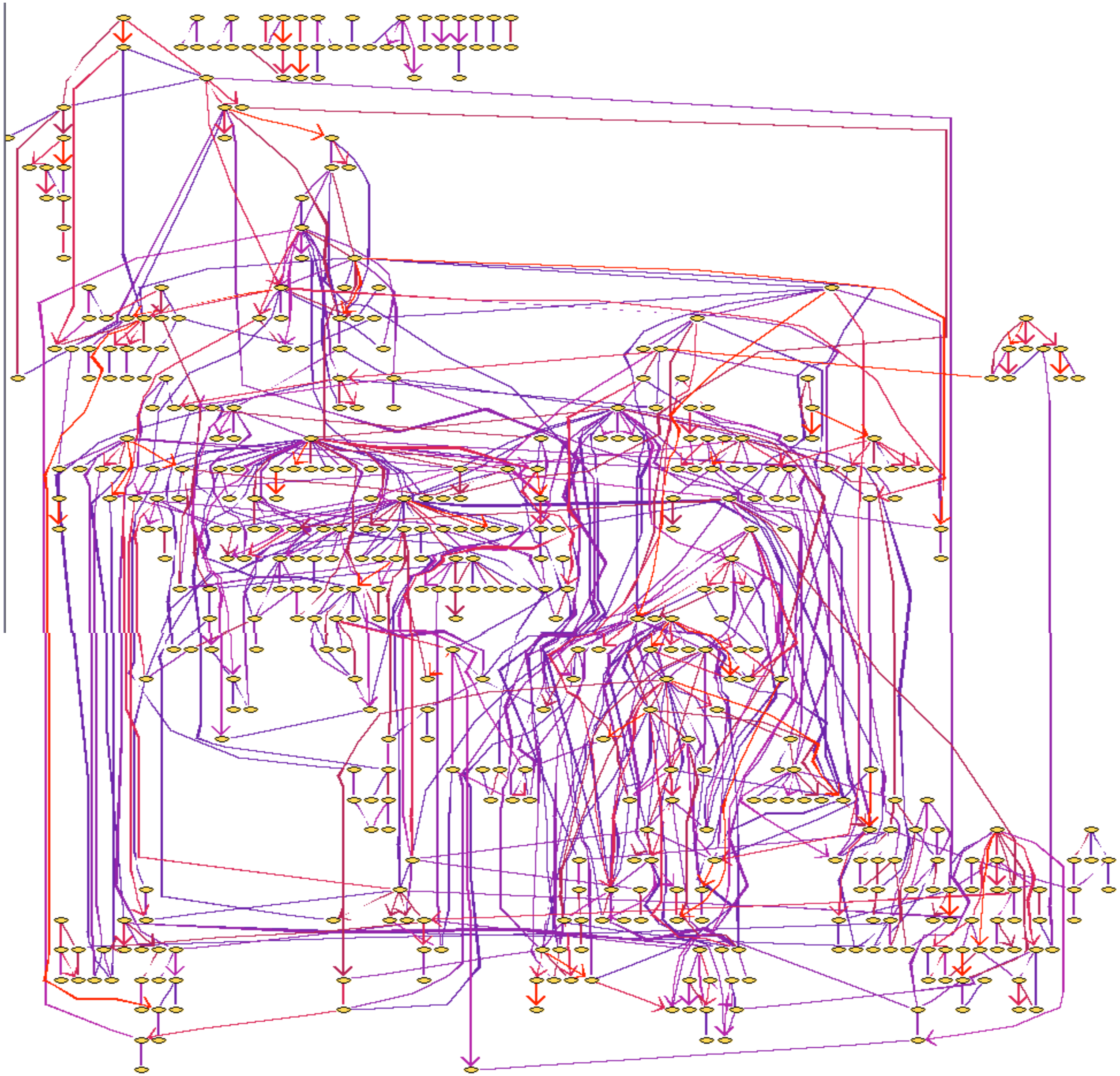
Experiment

Data from *Spellman et al.* (Mol.Bio. of the Cell 1998).

- Contains 76 samples of all the yeast genome:
  - Different methods for synchronizing cell-cycle in yeast.
  - Time series at few minutes (5-20min) intervals.
- *Spellman et al.* identified 800 cell-cycle regulated genes.



Network Learned



Challenge: Statistical Significance

Sparse Data

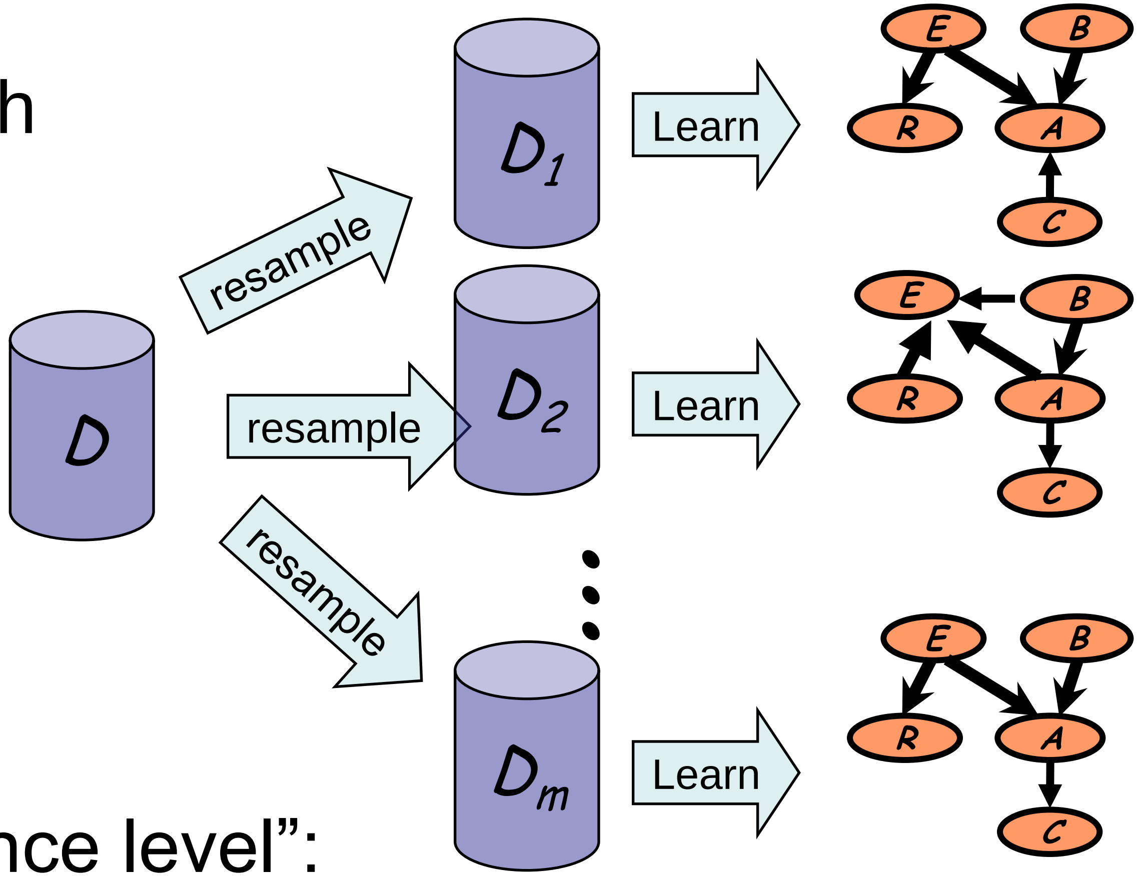
- Small number of samples
- “Flat posterior” -- many networks fit the data

Solution

- estimate confidence in network **features**
- E.g., two types of features
  - **Markov** neighbors**:** *X* **directly** interacts with *Y* *(through mutual edge or a mutual child)*
  - **Order** relations: *X* is a parent of *Y*

Confidence Estimates

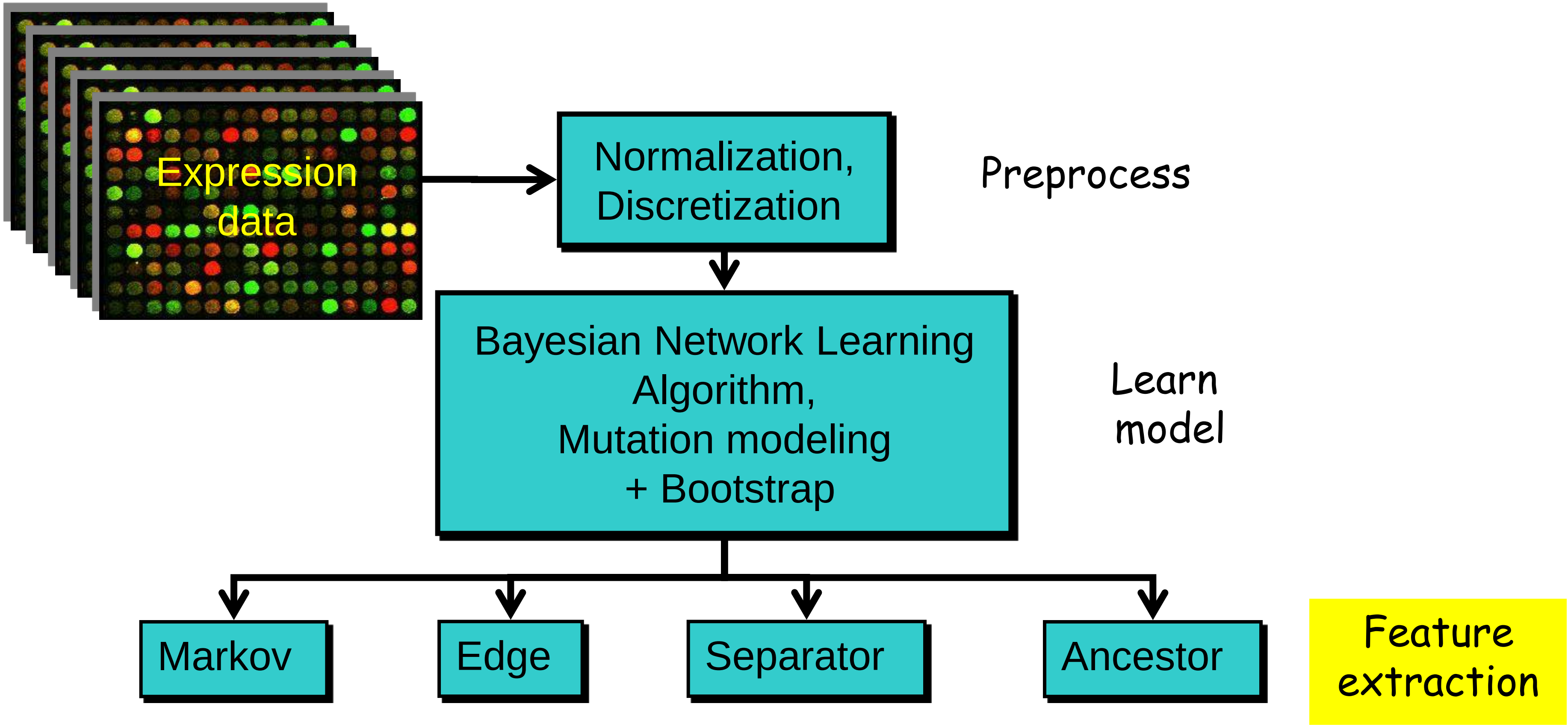
Bootstrap approach
[FGW, UAI99]



Estimate “Confidence level”:

$$C(f) = \frac{1}{m}\sum_{i=1}^{m} 1\{f \in G_i\}$$

In summary…



Expression data

Normalization, Discretization

Preprocess

Bayesian Network Learning Algorithm, Mutation modeling + Bootstrap

Learn model

Markov

Edge

Separator

Ancestor

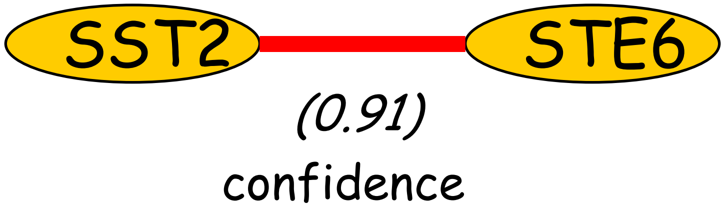
Feature extraction

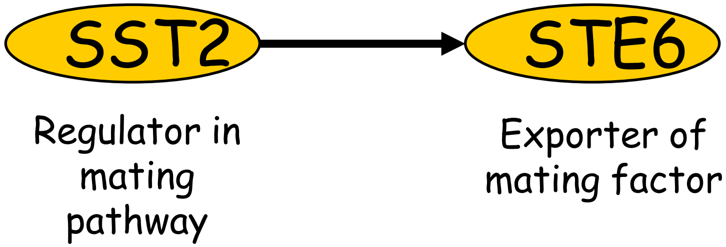
Result: a list of features with high confidence.
They can be biologically interpreted.

Resulting Features: Markov Relations

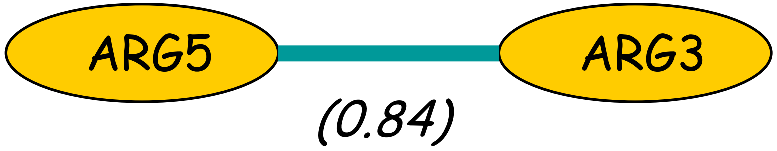
Question: Do *X* and *Y* directly interact?

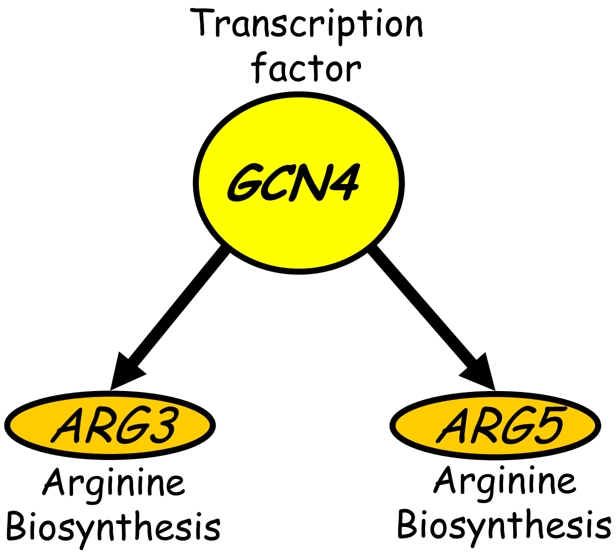
Parent-Child (one gene regulating the other)


SST2
STE6
(0.91)
confidence


SST2
STE6
Regulator in mating pathway
Exporter of mating factor

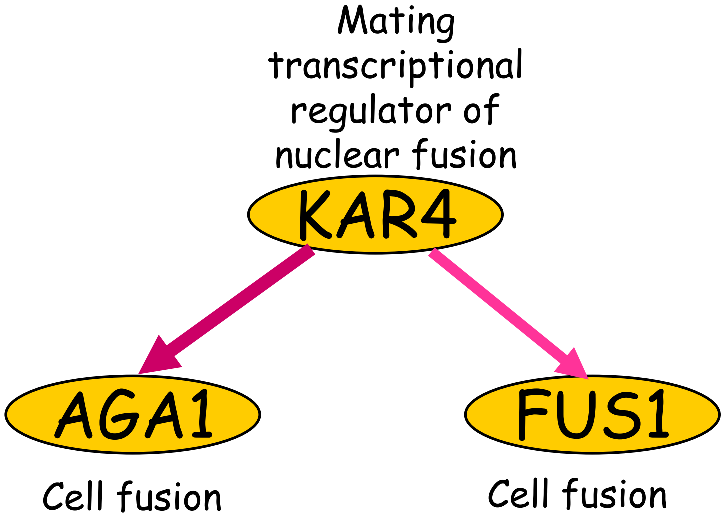
Hidden Parent (two genes co-regulated by a hidden factor)


ARG5
ARG3
(0.84)

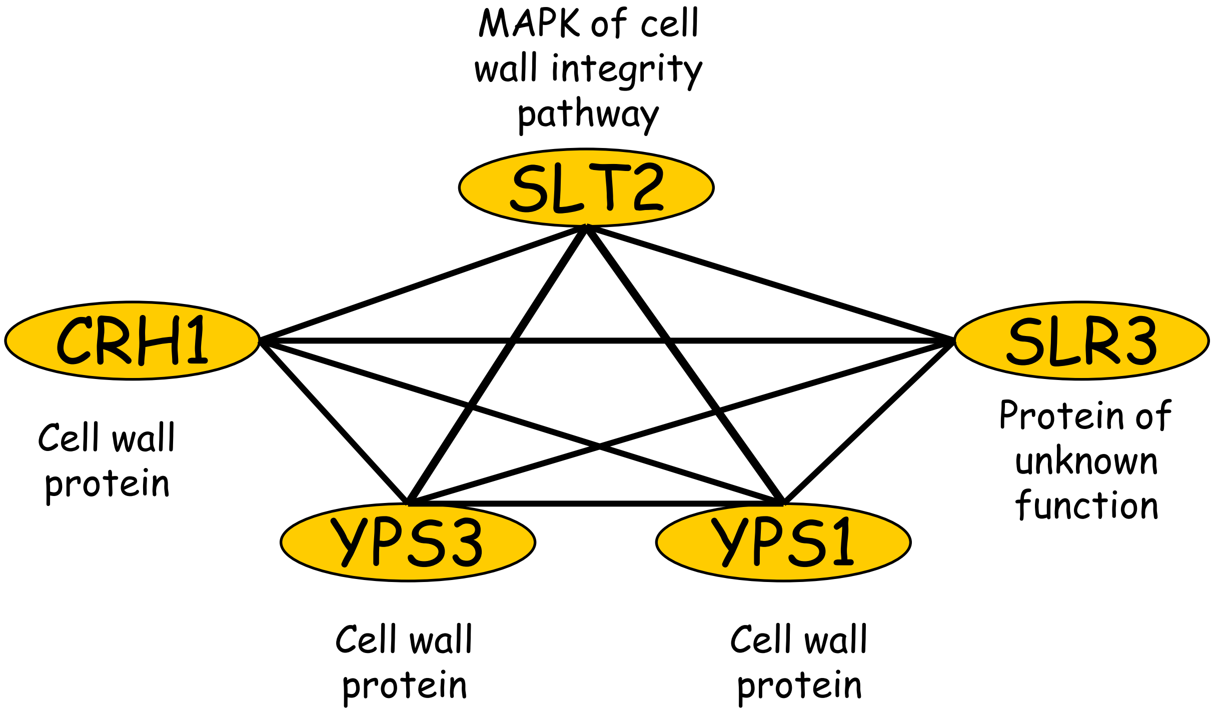

Transcription factor
GCN4
ARG3
Arginine Biosynthesis
ARG5
Arginine Biosynthesis

Resulting Features: Separators

- **Question:** Given that *X* and *Y* are indirectly dependent, who **mediates** this dependence?
- **Separator** relation:
  - *X* affects *Z* who in turn affects *Z*
  - *Z* regulates both *X* and *Y*

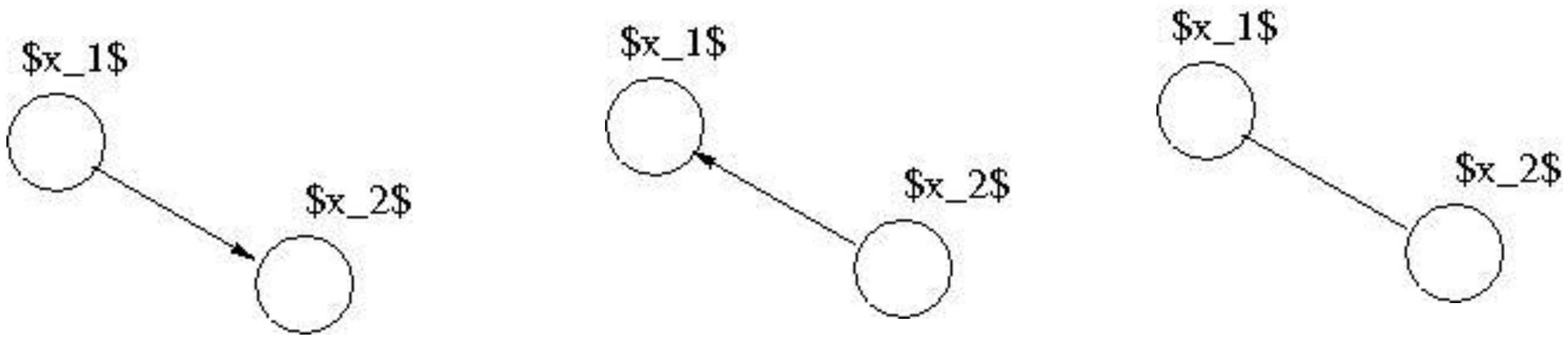


Separators: Intra-cluster Context



- All pairs have high correlation
- Clustered together

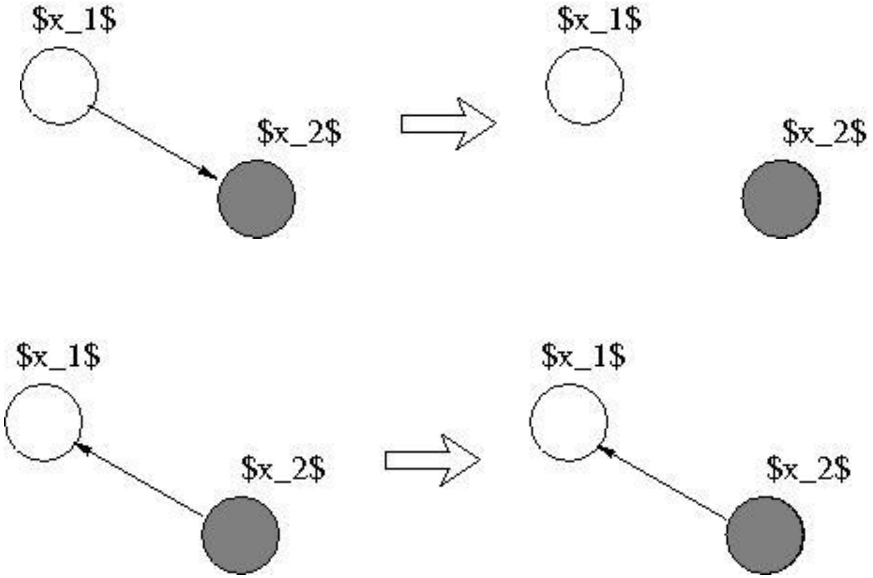
Dependencies and causality



- Since $P(x_1)P(x_2|x_1) = P(x_2)P(x_1|x_1) = P(x_1, x_2)$, we cannot immediately attach any causal interpretation to the probabilistic dependencies (e.g., if factor x_1 regulates x_2)

Causality

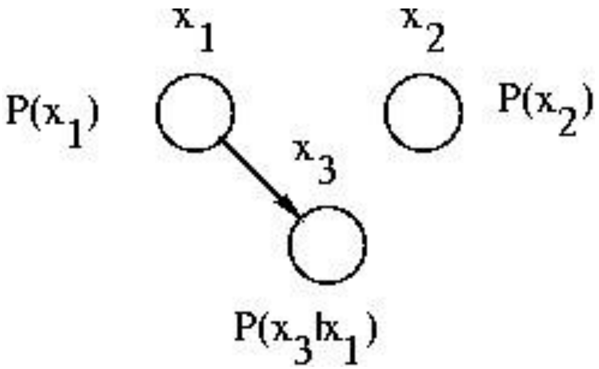
- We can use interventions (external manipulations) to disambiguate between possible causal interpretations
- For example: if we intervene to set the value of x_2 to a specific value (e.g., knock-out) then:



Extensions: Bayesian networks and regression

- Another way to deal with the continuous data is to use a different probabilistic model.

- For example, Gaussian linear regression:



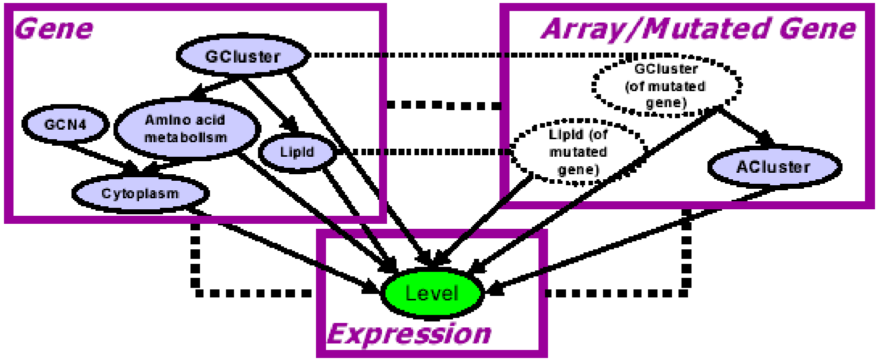
$$p(x_3 \mid x_1) : x_3 \sim N(\mu_3 + \alpha x_1, \sigma_3^2)$$
$$x_2 \sim N(\mu_2, \sigma_2^2)$$
$$x_1 \sim N(\mu_1, \sigma_1^2)$$

Rich probabilistic gene expression networks

- In many cases we have additional types of information that can be used for the network learning.
- In addition, the expression levels of multiple genes is commonly affected by their regulators and function
- Graphical models based on the idea of related ‘modules’ can be used to capture these notions.
- Specific model is termed Probabilistic Rational Model (PRM)
- Data sources includes:

  - Functional assignment for gene (from MIPS)

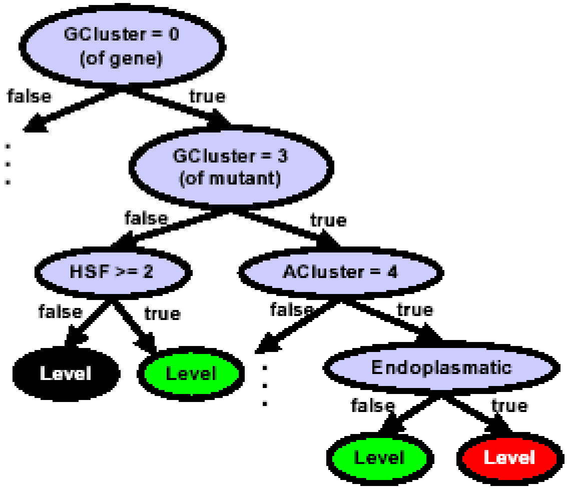
  - Binding site information for known TFs
- Gene classes are latent variables.
- Array classes are known (different class to each array).



Segal et al Nature Genetics 2003

Probability model

- Decision tree for each of the expression levels.
- Decision can be based on expression levels of other gene or on discrete values from the other data sources.
- Can use the node in the tree to determine parents for a given node.



Issues:

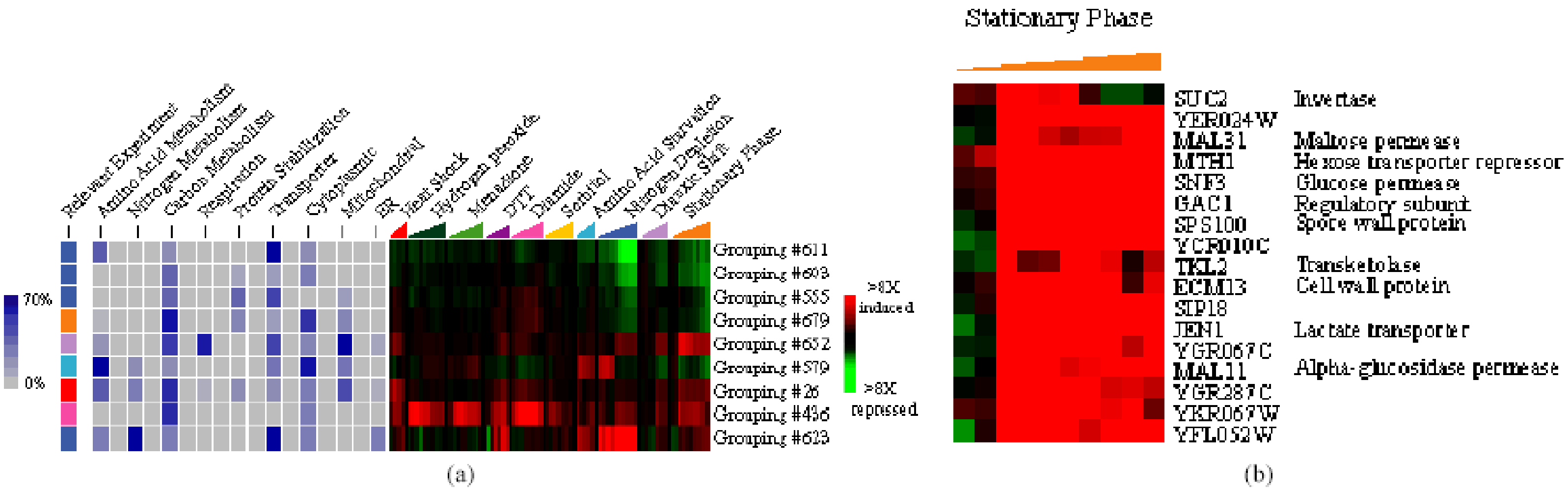
- Acyclic graph
- Learning the tree for each gene

Determining significance of results

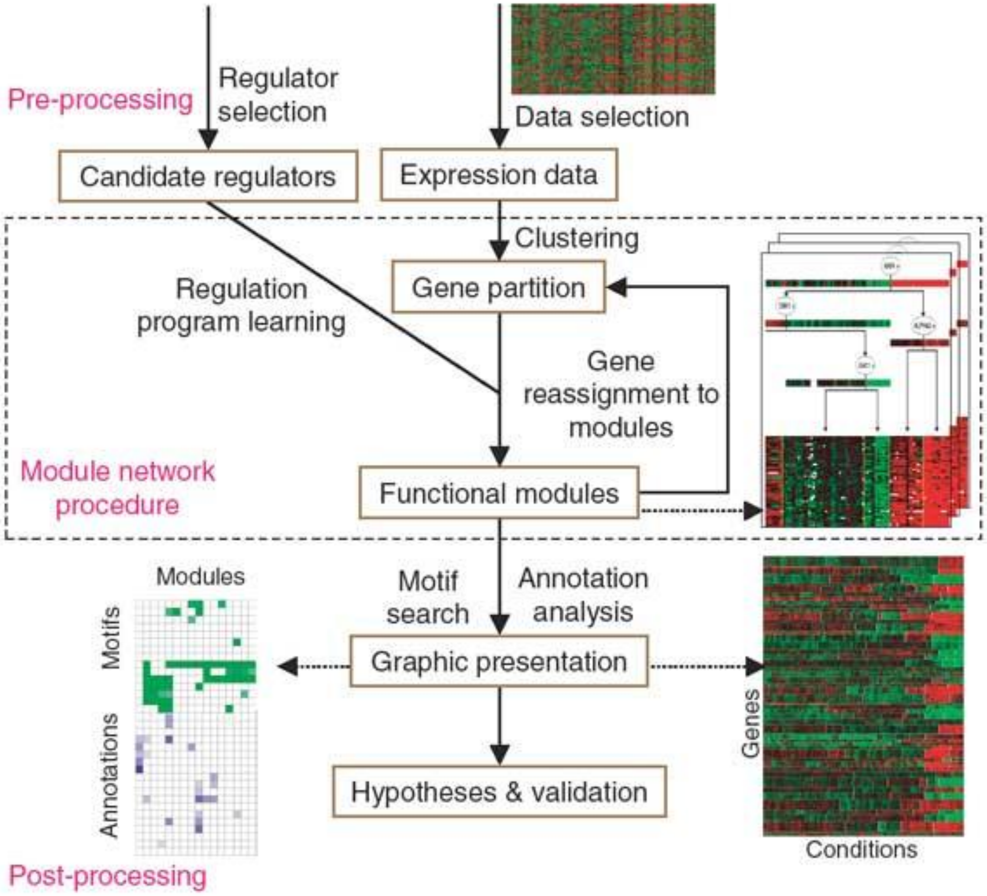
- Use permutation data to determine if the structure observed was present in the data.
- Apply the same algorithm to a randomized version of the data.
- Use likelihood of generated model to test the relevance of the learned structure.

Testing the clusters

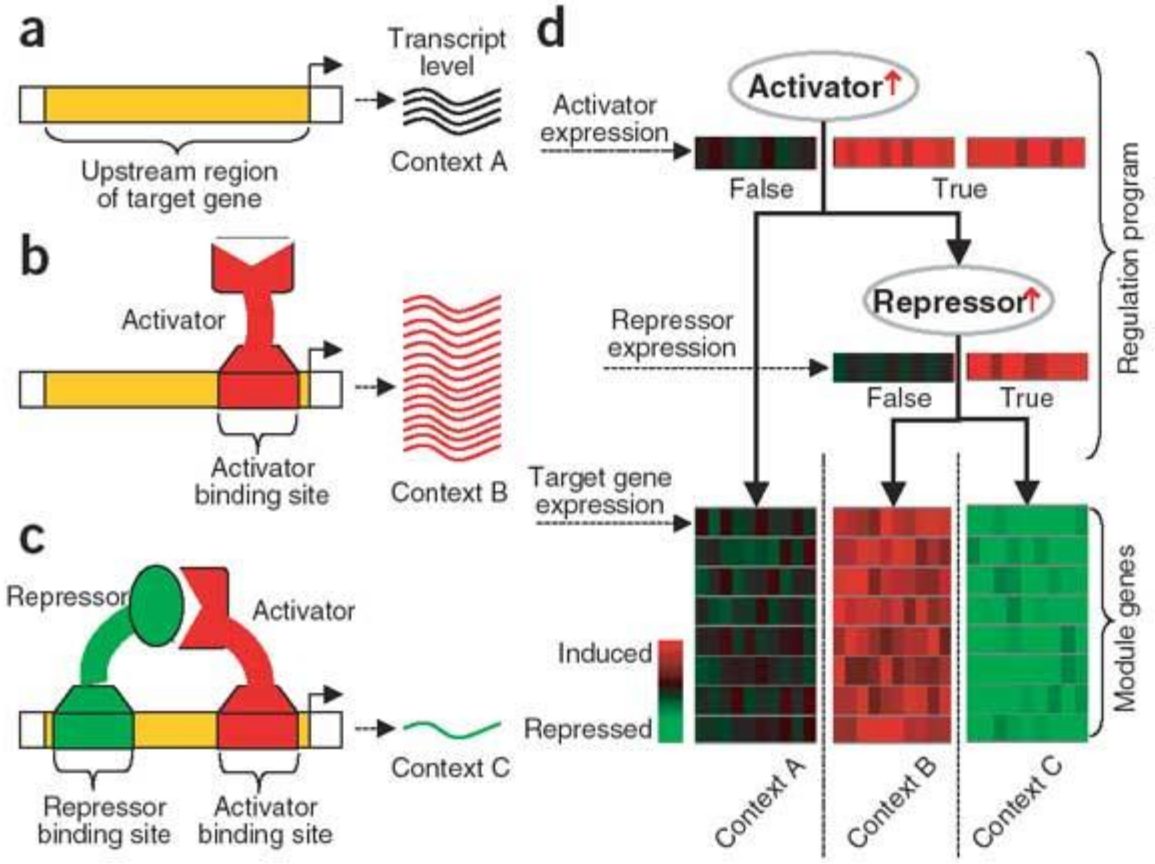
- Test the variance of the expression in each cluster.
- Remove functional annotation after initial step to allow for new annotations for unknown genes.



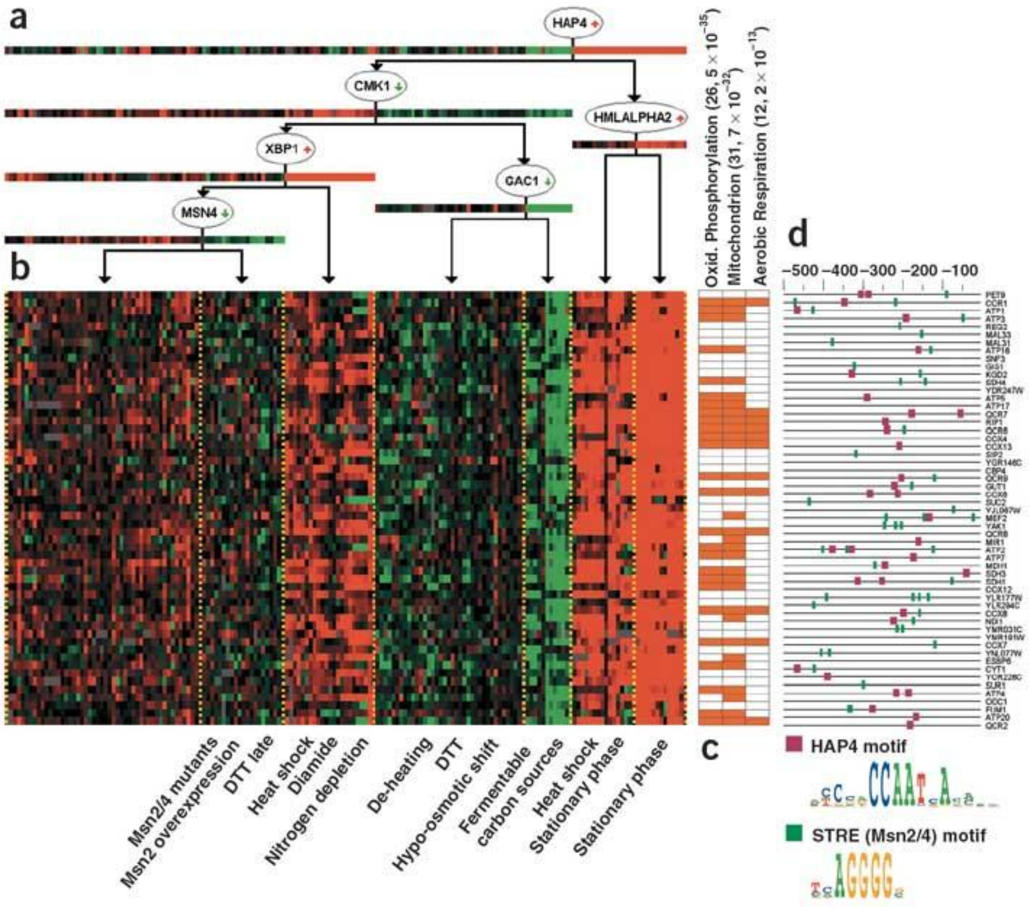
From modules to networks



Determine combinatorial control

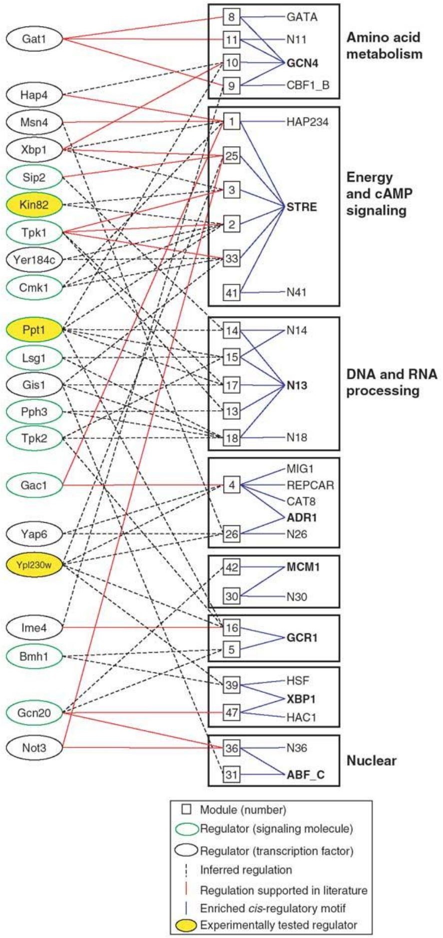


Resulting module



Segal et al Nature Genetics 2003

More combinatorial regulation



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)
Regulator (transcription factor)
Inferred regulation
Regulation supported in literature
Enriched cis-regulatory motif
Experimentally tested regulator