Statistical methods for inferring the gene regulatory networks – Part II

Lecture 2 – May 16th, 2013
GENOME 541, Spring 2013

Su-In Lee
GS & CSE, UW
suinlee@uw.edu

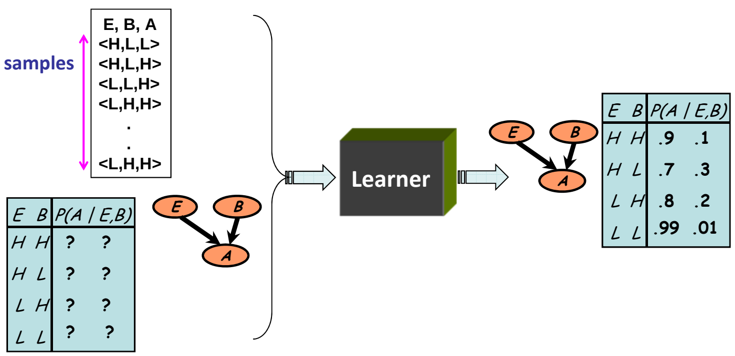
1

Outline (5/14, 5/16)

- Basic concepts on Bayesian networks
- Probabilistic models of gene regulatory networks
- Learning algorithms ← *Today*
- Evaluation
- Recent probabilistic approaches to reconstructing the regulatory networks

2

Known structure, complete data

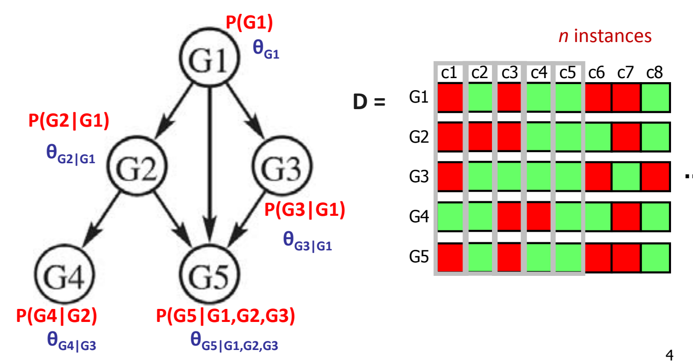


- Network structure is specified
  - Learner needs to estimate parameters
- Data does not contain missing values

3

Training data

- Learn the parameters based on **D**



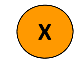
4

LET'S CONSIDER THE SIMPLEST EXAMPLE.

5

The *Thumbtack* example

- Parameter estimation for a single variable



- Variable
  - X - an outcome of a thumbtack toss
  - Val(X) = {head, tail}
- Data
  - A set of thumbtack tosses: x[1] … x[M]

6

Maximum likelihood estimation

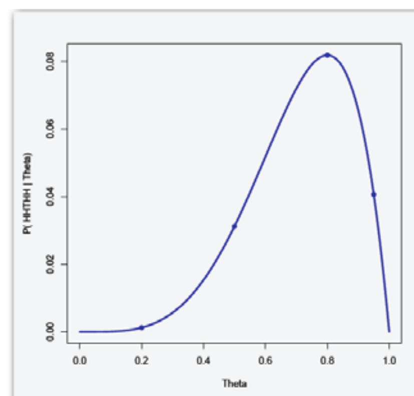
- Say that P(x=head) = Θ, P(x=tail) = 1-Θ
  - P(HHTTHHH…<M_h heads, M_t tails>; Θ) =
- **Definition:** The likelihood function
  - L(Θ : D) = P(D; Θ)
- Maximum likelihood estimation (MLE)
  - Given data D=HHTTHHH…<M_h heads, M_t tails>, find Θ that maximizes the likelihood function L(Θ : D).

7

Likelihood function

Probability of HHTHH, given P(H) = θ:

θ	$θ^4(1-θ)$
0.2	0.0013
0.5	0.0313
0.8	0.0819
0.95	0.0407

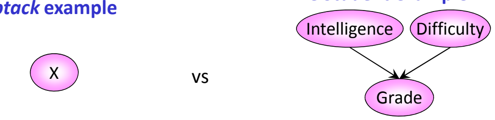


8

MLE for the *Thumbtack* problem

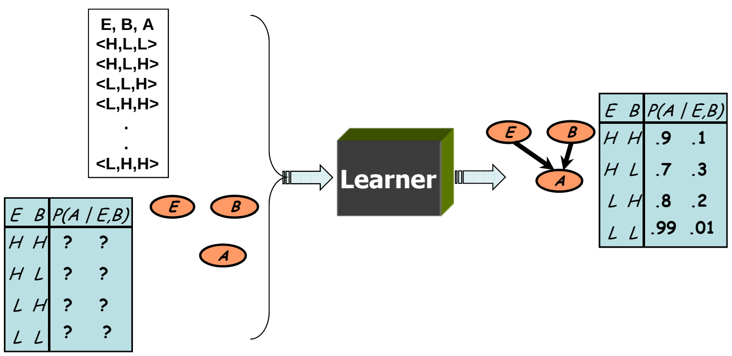
- Given data D=HHTTHHH…<M_h heads, M_t tails>
  - MLE solution $\theta^* = M_h / (M_h+M_t)$.
- Proof:

Bayesian Network with table CPDs



	The *Thumbtack* example	The *Student* example
Joint distribution	P(X)	P(I,D,G) =
Parameters	θ	$\theta_I, \theta_D, \theta_{G\mid I,D}$
Data	D: {H…x[m]…T}	D: {(i^1,d^0,g^1)…(i[m],d[m],g[m])…}
Likelihood function **L(θ:D) = P(D;θ)**	$\theta^{Mh}(1-\theta)^{Mt}$	$\theta_{I=i^1}^{M_{I=i^1}} \theta_{I=i^0}^{M_{I=i^0}} \theta_{D=d^1}^{M_{D=d^1}} \theta_{D=d^0}^{M_{D=d^0}} \theta_{G=g^1\mid I=i^1,D=d^1}^{M_{G=g^1\mid I=i^1,D=d^1}} \cdots$
MLE solution	$\hat{\theta} = \frac{Mh}{Mh+Mt}$	$\theta_{G=g^1\mid I=i^1,D=d^0} = \frac{M_{G=g^1,I=i^1,D=d^0}}{M_{I=i^1,D=d^0}}$

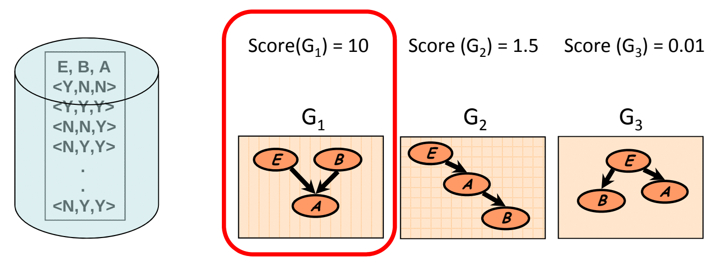
Unknown structure, complete data



- Network structure is not specified
  - Learner needs to estimate both structure and parameters
- Data does not contain missing values

Score-based learning

- Define scoring function that measures how well a certain structure fits the observed data.



- Search for a structure that maximizes the score.

Structure score

- Likelihood score: $P(D|S, \hat{\theta}_S)$ Maximum likelihood parameters
- Bayesian score
  - Average over all possible parameter values

$$P(D\,|\,S) = \int P(D\,|\,S,\theta)P(\theta\,|\,S)\,d\theta$$

Marginal likelihood | Likelihood | Prior distribution over parameters

- Penalized likelihood score

$$\log P(D|S,\theta_S) - C \cdot \text{model complexity}(S, \theta_S, D)$$

Decomposability of scores

- Likelihood score

$L(\Theta : D) = \prod_i L_i(\Theta_i : D)$ (see slide 11)

- Bayesian score

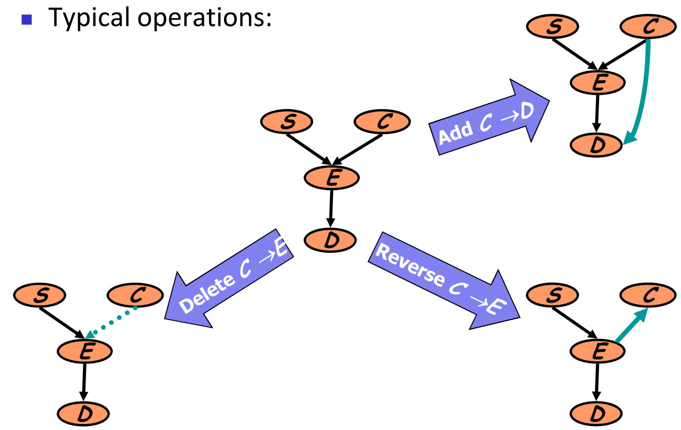
$$\begin{aligned} P(D\,|\,S) &= \int P(D\,|\,S,\theta)P(\theta\,|\,S)\,d\theta \\ &= \int_{\Theta_1 \ldots \Theta_k} \prod_i \left(\prod_m P(x_i[m]\,|\,Pa_i[m] : \Theta_i) \right) P(\Theta_i : S)\; d\Theta \\ &= \prod_i \int_{\Theta_i} \left(\prod_m P(x_i[m]\,|\,Pa_i[m] : \Theta_i) \right) P(\Theta_i : S) d\Theta_i \\ &= \prod_i \text{BayesianScore}(\Theta_i : D) \end{aligned}$$

Search for optimal network structure

- Start with a given network structure.
  - Empty network
  - Best simple structure (e.g. tree)
  - A random network
- At each iteration
  - Evaluate all possible changes
  - Apply change based on score
- Stop when no modification improves the score.

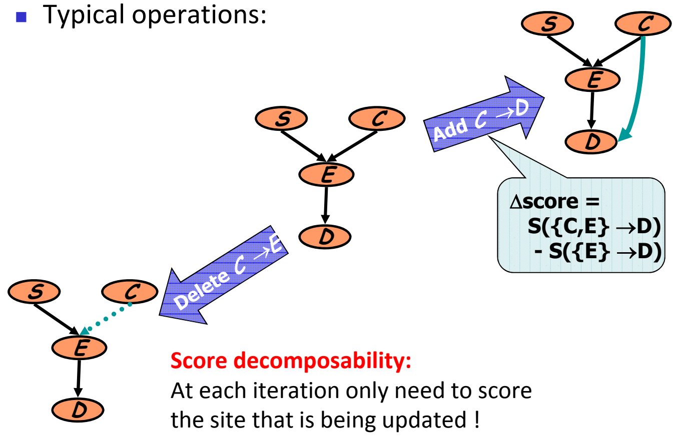
Search for optimal network structure

- Typical operations:



Search for optimal network structure

- Typical operations:



Score decomposability:
At each iteration only need to score the site that is being updated !

17

Outline

- Basic concepts on Bayesian networks
- Probabilistic models of gene regulatory networks
- Learning algorithms
  - Parameter learning
  - Structure learning
  - Structure discovery
- Evaluation
- Recent probabilistic approaches to reconstructing the regulatory networks

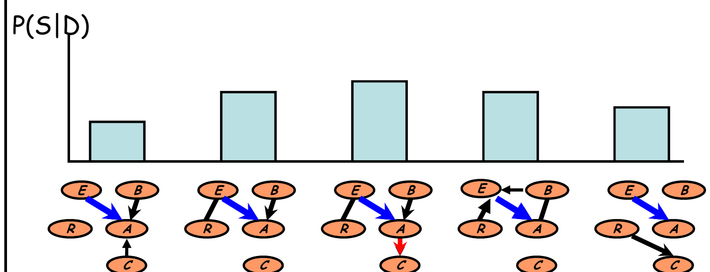
18

Structure discovery

- **Task:** Discover structural properties
  - Is there a direction connection between X and Y?
  - Does X separate between two "subsystems"?
  - Does X causally affect Y?
- Example: scientific data mining
  - Disease properties and symptoms
  - Interactions between the expression of genes

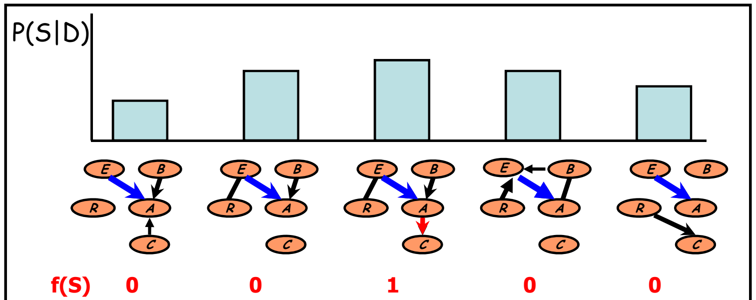
19

Model averaging



- There may be many high-scoring models
- Answer should not be based on any single model
- Want to average over many models

20



- Define a structural feature f(S) of a model S.
  - For example:

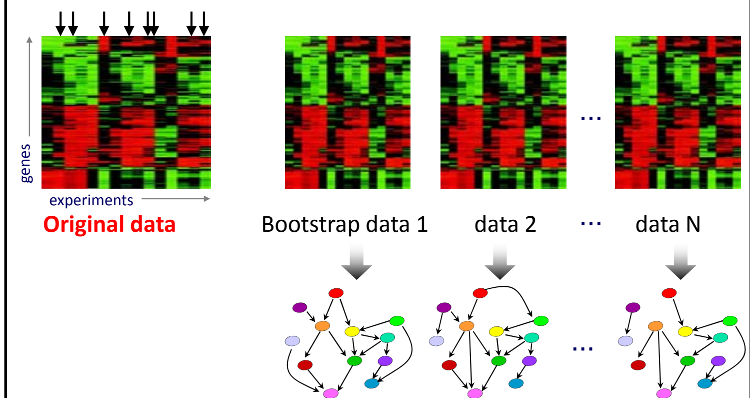
$$f(S) = \begin{cases} 1 & \text{if a graph S has } A \rightarrow C \\ 0 & \text{otherwise} \end{cases}$$

- We are interested in computing

$$E_{P(S|D)}[f(S)] = \sum_S f(S)P(S \mid D)$$

Bootstrapping

- Sampling with replacement



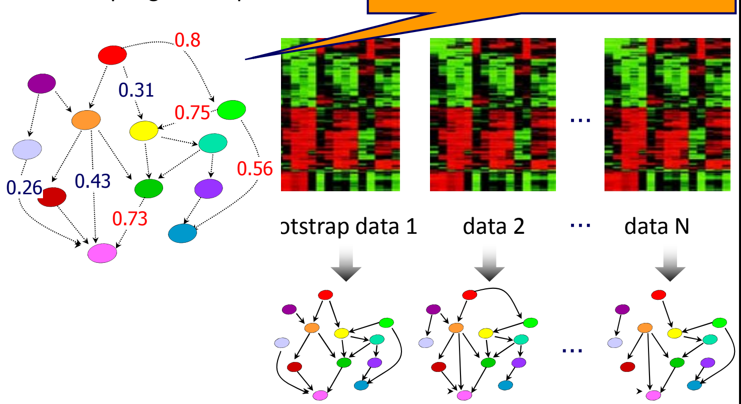
Inferring sub-networks from perturbed expression profiles, Pe'er et al. Bioinformatics 2001

Bootstrapping

- Sampling with replacement

- Estimated confidence of each edge *i*

$$= \frac{\text{\# networks that contain the edge}}{\text{total \# networks (N)}}$$



Inferring sub-networks from perturbed expression profiles, Pe'er et al. Bioinformatics 2001

Outline

- Basic concepts on Bayesian networks
- Probabilistic models of gene regulatory networks
- Learning algorithms
- Evaluation
  - Predicted co-regulated groups of genes
  - Putative regulator-regulatees
- Recent probabilistic approaches to reconstructing the regulatory networks

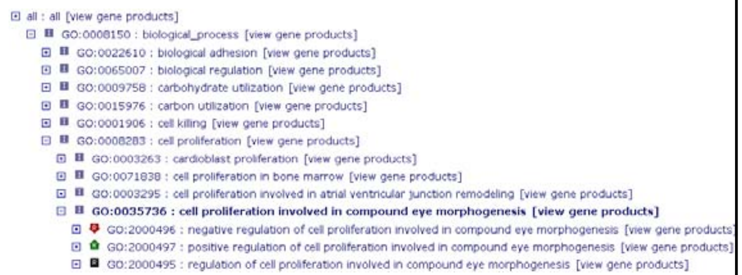
Functional coherence of gene clusters

- Gene Ontology (GO) [http://www.geneontology.org/]
  - The GO database provides a controlled vocabulary to describe gene and gene product attribute in any organism.
  - Set of biological phrases (GO terms) which are applied to genes
  - Organized as three separate ontologies
    - Molecular functions
    - Biological processes
    - Cellular components
  - Each gene may
    - Have more than one in molecular function.
    - Take part in more than one biological process.
    - Act in more than one cellular component.

25

Structure of ontologies

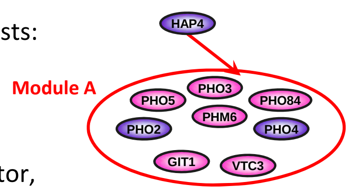
- Shows the relationship between different terms
  - One term may be a more specified description of another more general term.
  - Shows hierarchies of the terms (directed acyclic graph).
  - Each child-term is a member of its parent-term



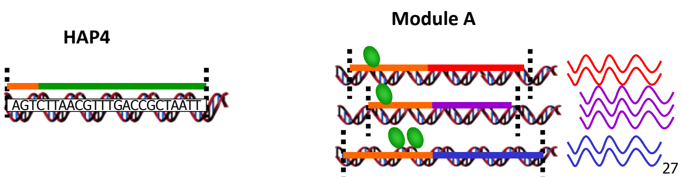
26

Predicted regulatory interaction I

- Say that your network suggests:



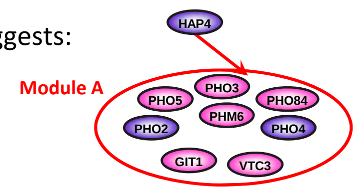
- If HAP4 is a transcription factor,
  - Targets should have a binding site for HAP4.
  - Or there should be different kind of evidence that HAP4 binds to genes in Module A (chip-chip or chip-seq data).



27

Predicted regulatory interaction II

- Say that your network suggests:

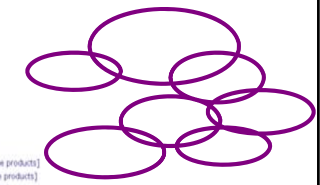


- If HAP4 really regulates module A, deletion (or overexpression) of HAP4 should lead to significant up/down- regulation of genes in module A.
  - There are many publicly available gene expression data that measure expression of genes after deleting/over-expressing a certain gene.

28

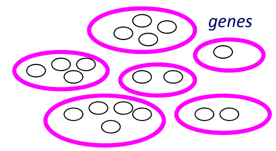
Create functional categories

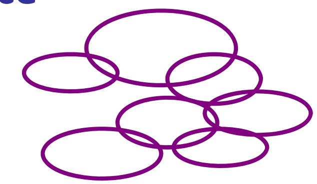
- For each GO term,
  - Genes that have the same GO term form a functional category
- Other gene annotation systems
  - KEGG: Kyoto Encyclopedia of Genes and Genomes [http://www.genome.jp/kegg/]
  - Molecular Signature Database [http://www.broadinstitute.org/gsea/msigdb/index.jsp]

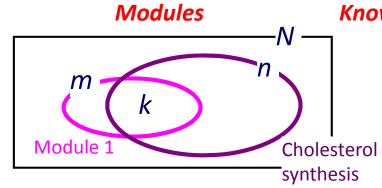


Functional categories

Functional coherence





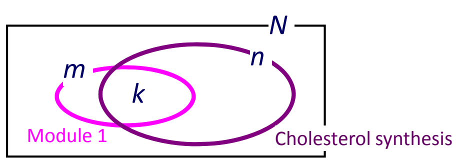


Gene ontology (GO)
http://www.geneontology.org/
Predicted targets of regulators
Sharing TF binding sites
:

- How significant is the overlap?
  - Calculate P(# overlap ≥ k | m, n, N; two groups are independent) based on the hypergeometric distribution

Examples

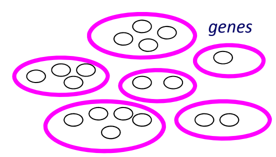
- Say N=1000, m=100, n=200 genes
  - If k = 40 genes in the intersection, p-value = 2.7410e-07.
  - If k = 30, p-value = 0.0039
  - If k = 20, p-value = 0.4394.

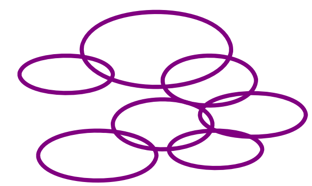


- How significant is the overlap?
  - Calculate p-value = P(# overlap ≥ k | m, n, N; two groups are independent), based on the hypergeometric distribution
  - What p-values are considered to be significant?

Multiple hypothesis testing

- Say that there are 200 modules and 3000 functional categories





- How many hypotheses are we testing?
  - 200 x 3000 = 600,000
  - Is p-value of 0.001 significant? (p-value=0.001: frequency of observing the # genes in intersection by random.)
- P-values should be "corrected"
  - Bonferroni correction: min(1, p-value x # hypotheses)
  - FDR correction: control false discovery rate

Outline

- Basic concepts on Bayesian networks
- Probabilistic models of gene regulatory networks
- Learning algorithms
- Evaluation
  - Predicted co-regulated groups of genes
  - Putative regulator-regulatees
- Recent probabilistic approaches to reconstructing the regulatory networks

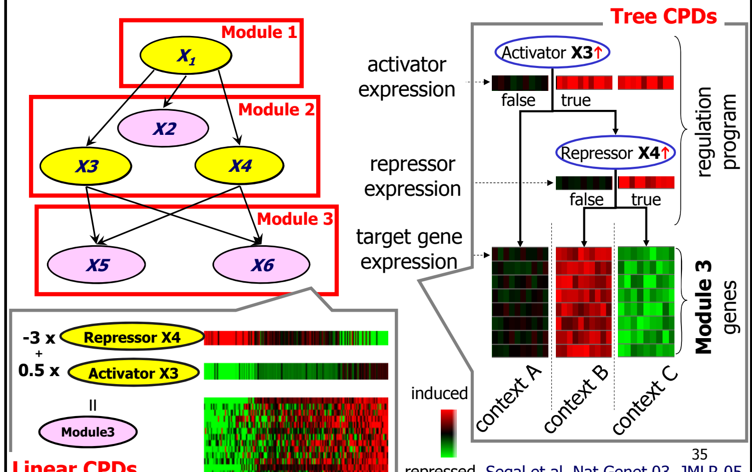
33

Challenges

- Too large search space
  - For a network with *n* genes, what is the number of possible structures? $\sim 3^{n^2/2}$
- Computationally costly
- Heuristic approaches may be trapped to local maxima.
- Biologically motivated constraints can alleviate the problems
  - Module-based approach
  - Only the genes in the candidate regulators list can be parents of other variables

34

The Module networks concept

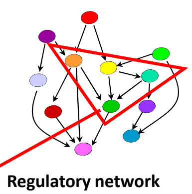


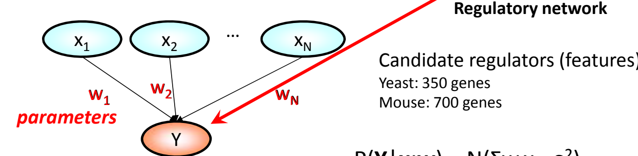
35

Segal et al. Nat Genet 03, JMLR 05

Feature selection via regularization

- Assume linear Gaussian CPD
- MLE: solve **maximize$_w$ - $(\Sigma w_i x_i - Y)^2$**





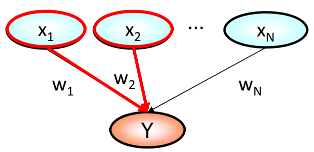
$P(\mathbf{Y}|\mathbf{x}:\mathbf{w}) = N(\Sigma w_i x_i, \varepsilon^2)$

Problem: This objective learns too many regulators

36

L_1 regularization

- "Select" a subset of regulators
  - Combinatorial search?
  - Effective feature selection algorithm: L_1 regularization (LASSO)
    [Tibshirani, J. Royal. Statist. Soc B. 1996]
  - minimize$_{\mathbf{w}}$ $(\Sigma w_i x_i - Y)^2 + \Sigma\, C\, |w_i|$: convex optimization!
    $\Rightarrow$ Induces sparsity in the solution $\mathbf{w}$ (Many w_i's set to zero)

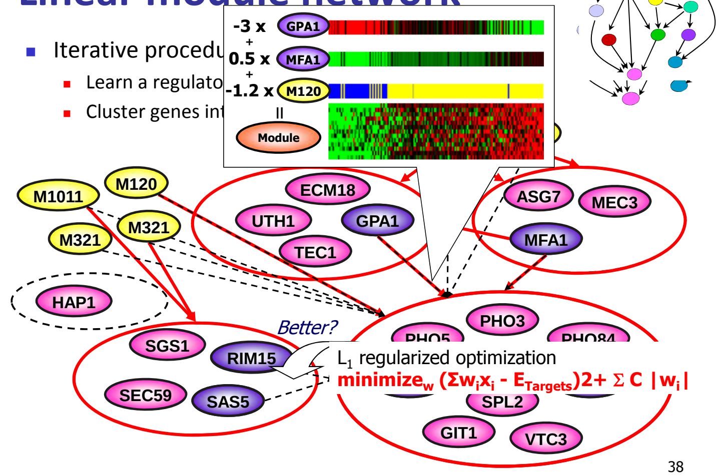


Candidate regulators (features)
Yeast: 350 genes
Mouse: 700 genes

$P(Y|\mathbf{x}:\mathbf{w}) = N(\Sigma w_i x_i, \varepsilon^2)$

Linear module network

- Iterative procedu
  - Learn a regulato
  - Cluster genes int



Lee et al., PLoS Genet 2009

Let's consider the module network with tree CPDs…

LEARNING

Learning module networks

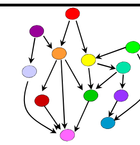
- Score-based learning – Find the structure that maximizes **Bayesian score** log P(S|D) (or via regularization)
- "Hidden" variables
  - How genes are organized into modules is not known.
  - Expectation Maximization (EM) algorithm: Repeat
    - E-step: filling in hidden variables
    - M-step: parameter estimation

Learning module networks

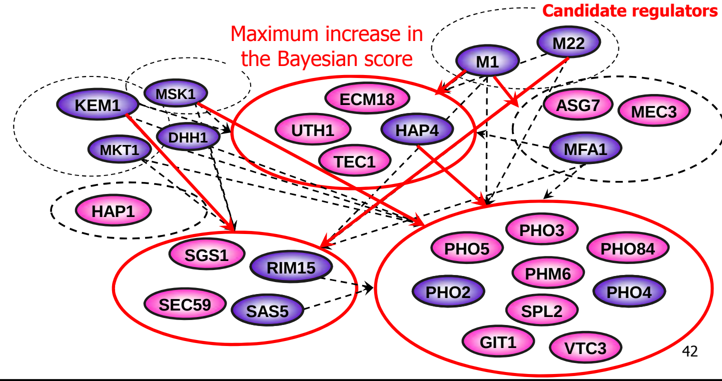
- Learning algorithm
  - Initialization: Group genes by (k-means) clustering into modules
  - M-step: Given a partition of the genes into modules, **learn the best regulation programs (tree CPD)** for modules.
  - E-step: Given the inferred regulatory programs, we **reassign genes into modules** such that the associated regulation program best predicts each gene's behavior.
  - Repeat until convergence.

41

Learning module networks



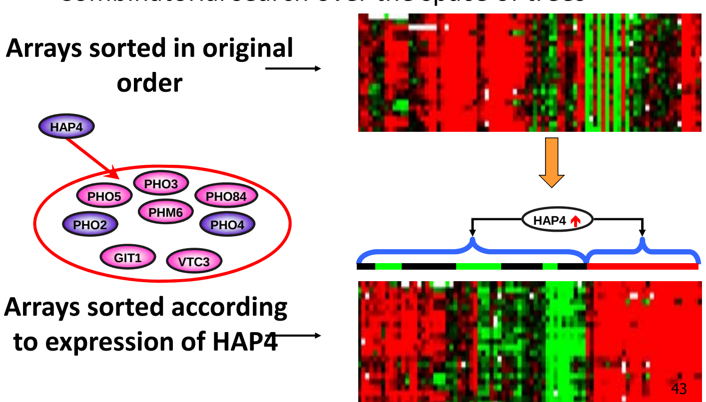
- Iterative procedure (EM-steps)
  - Cluster genes into modules (E-step)
  - Learn regulatory programs for modules (tree CPD) (M-step)



42

M-step: Learning regulatory programs

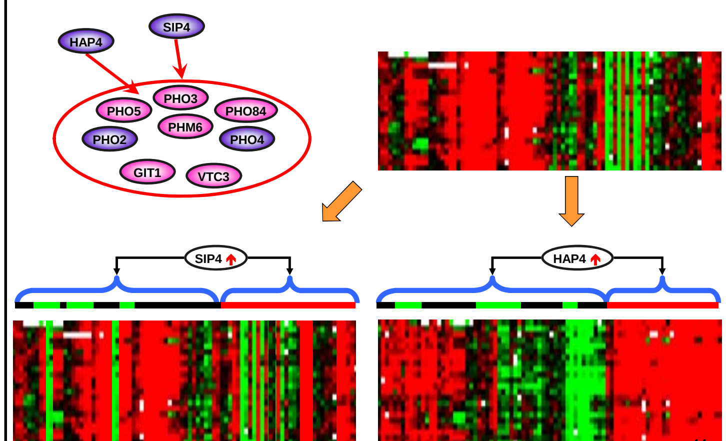
- Combinatorial search over the space of trees



43

Segal et al. Nat Genet 2003

M-step: Learning regulatory programs



44

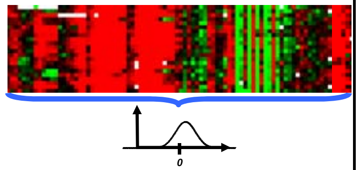
Segal et al. Nat Genet 2003

M-step: Learning regulatory programs

- **Score:**

log $P(M \mid D) \propto$

$\log \int P(D \mid M,\mu,\sigma)P(\mu,\sigma)\, d\mu d\sigma$

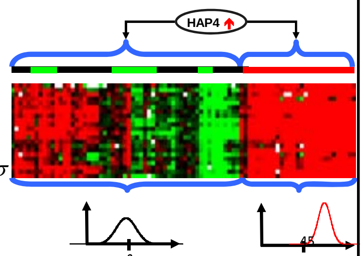


- **Score of HAP4 split:**

log $P(M \mid D) \propto$

$\log \int P(D_{HAP4\uparrow} \mid M,\mu,\sigma)P(\mu,\sigma)\, d\mu d\sigma$

$+ \log \int P(D_{HAP4\downarrow} \mid M,\mu,\sigma)P(\mu,\sigma)\, d\mu d\sigma$

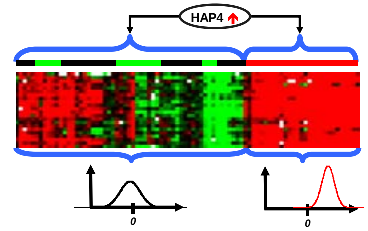


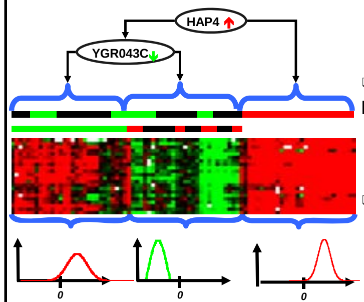
Segal et al. Nat Genet 2003

45

M-step: Learning regulatory programs

- Split as long as the score improves





- **Score of HAP4 split:**

log $P(M \mid D) \propto$

$\log \int P(D_{HAP4\uparrow} \mid M,\mu,\sigma)P(\mu,\sigma)\, d\mu d\sigma$

$+ \log \int P(D_{HAP4\downarrow} \mid M,\mu,\sigma)P(\mu,\sigma)\, d\mu d\sigma$

- **Score of HAP4/YGR043C split:**

log $P(M / D) \propto$

$\log \int P(D_{HAP4\uparrow} / M,\mu,\sigma)P(\mu,\sigma)\, d\mu d\sigma$

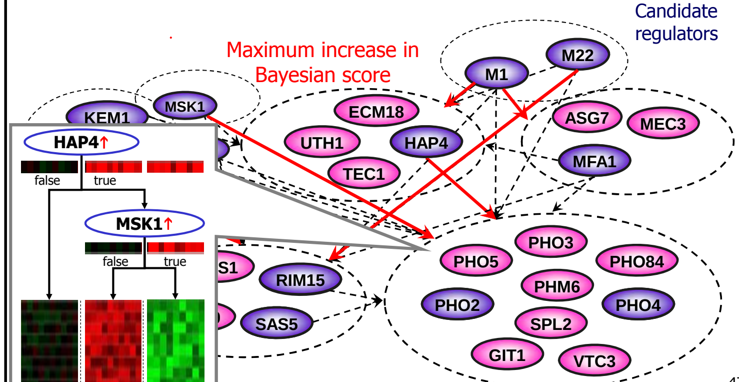
$+ \log \int P(D_{HAP4\downarrow}\, D_{YGR043C\uparrow} / M,\mu,\sigma)P(\mu,\sigma)\, d\mu d\sigma$

$+ \log \int P(D_{HAP4\downarrow}\, D_{YGR043C\downarrow} / M,\mu,\sigma)P(\mu,\sigma)\, d\mu d\sigma$

46

Learning module networks

- Iterative procedure
  - Cluster genes into modules (E-step)
  - Learn regulatory programs for modules (M-step)



47

Summary

- Basic concepts on Bayesian networks
- Probabilistic models of gene regulatory networks
- Learning algorithms
- Evaluation
- Recent probabilistic approaches to reconstructing the regulatory networks

48