Inferring Transcriptional Regulatory Networks from Gene Expression Data II

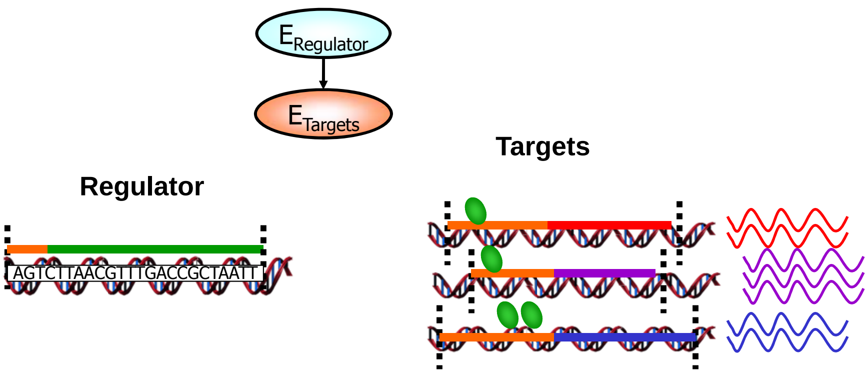
Lectures 9 – Oct 26, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

Review: Gene Regulation

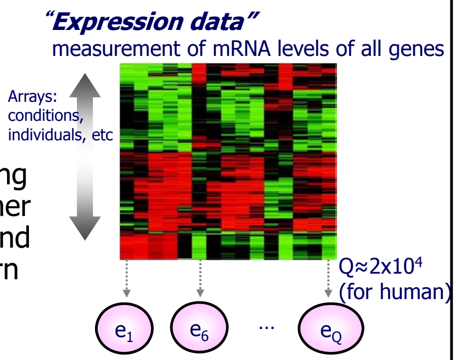
- Expression level of Regulator controls the expression levels of Targets it binds to.
- Regulator's expression is predictive of Targets' expression

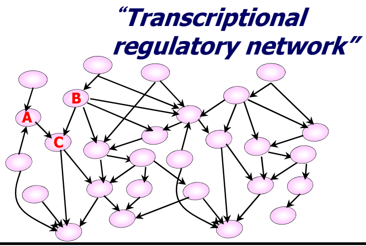


Segal et al., Nature Genetics 2003; Lee et al., PNAS 2006

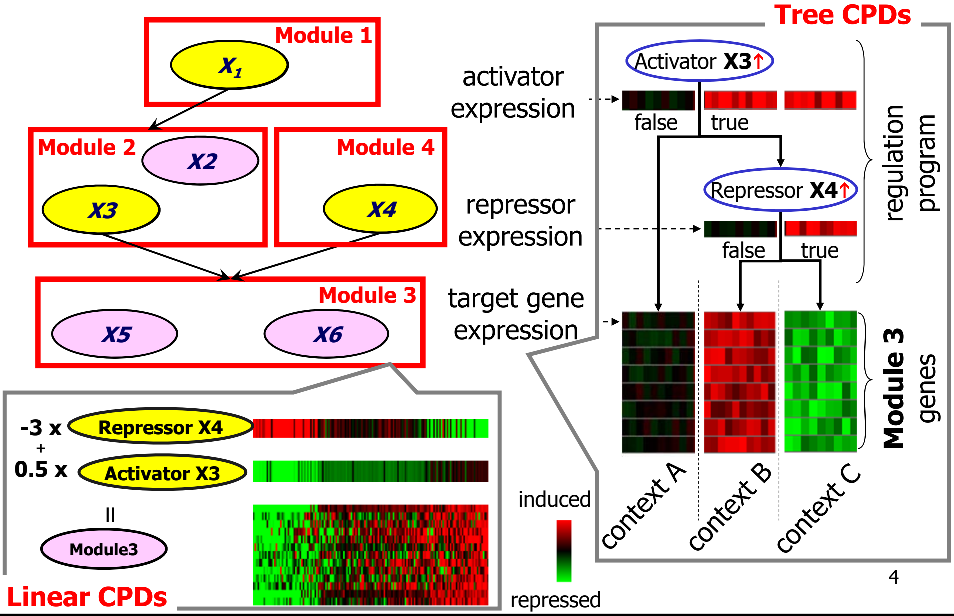
Review: Inferring the regulatory networks

- Input:
  - Gene expression data
- Output:
  - Bayesian network representing how genes regulate each other at the transcriptional level, and how these interactions govern gene expression levels.
- Algorithm:
  - Score-based structure learning of Bayesian networks
  - Challenge: Too many possible structures
  - Solutions: Control the complexity of the structure, Module networks





Review: The Module Networks Concept



4

Outline

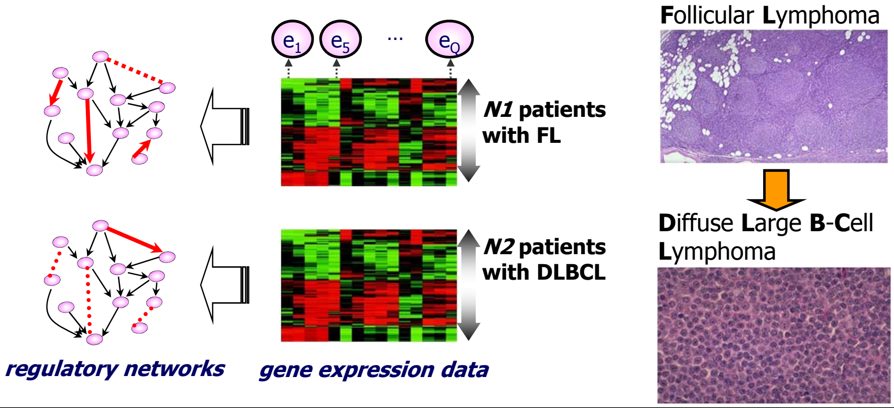
- Motivation
  - Why are we interested in inferring the regulatory network?
- Algorithms for learning regulatory networks
  - Tree-CPDs with Bayesian score
  - Linear Gaussian CPDs with regularization
- Evaluation of the method
  - Statistical evaluation
  - Biological interpretation
- Advanced topics
  - Many models can get similar scores. Which one would you choose?
  - A gene can be involved in multiple modules.
  - Possible to incorporate prior knowledge?

Motivations

- "Gene A regulates gene B's expression"
  - Important basic biology questions
- "Gene A regulates gene B in condition C"
  - Condition C: disease states (cancer/normal, subtypes), phenotypes, species (evolutionary processes), developmental stages, cell types, experimental conditions
- Example application (C: disease states)
  - Understanding histologic transformation in lymphoma
    - Lymphoma: the most common type of blood cancer in the US.
  - Transformation of **F**ollicular **L**ymphoma (FL) to **D**iffuse **L**arge **B**-**C**ell **L**ymphoma (DLBCL)
    - Occurs in 40-60% of patients; Dramatically worse prognosis
  - Goal: Infer the mechanisms that drive transformation.

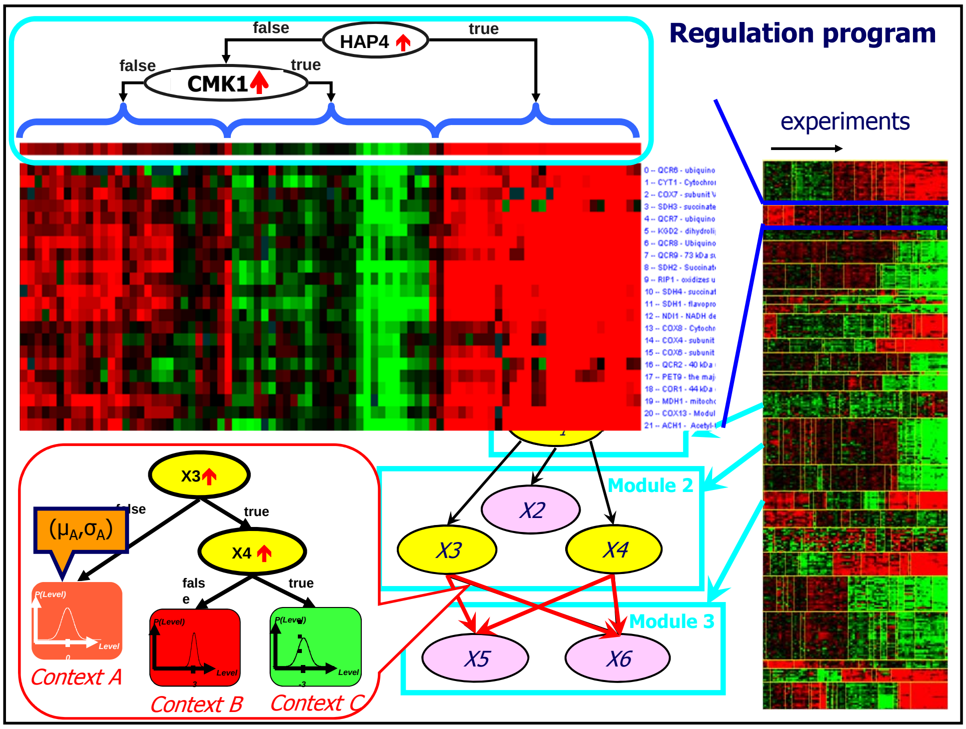
Predicting Cancer Transformation

- Network-based approach
  - Different network features ⇒ transformation mechanism?
    - Early diagnosis of transformation; therapeutic implications?
  - Share as many networks features as possible ⇒ more robust than inferring two networks separately.



Outline

- Motivation
- Algorithms for learning regulatory networks
  - Tree-CPDs with Bayesian score
  - Linear Gaussian CPDs with regularization
- Evaluation of the method
  - Statistical evaluation
  - Biological interpretation
- Advanced topics
  - Many models can get similar scores. Which one would you choose?
  - A gene can be involved in multiple modules.
  - Possible to incorporate prior knowledge?

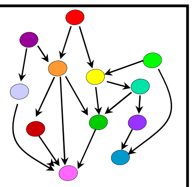


Learning

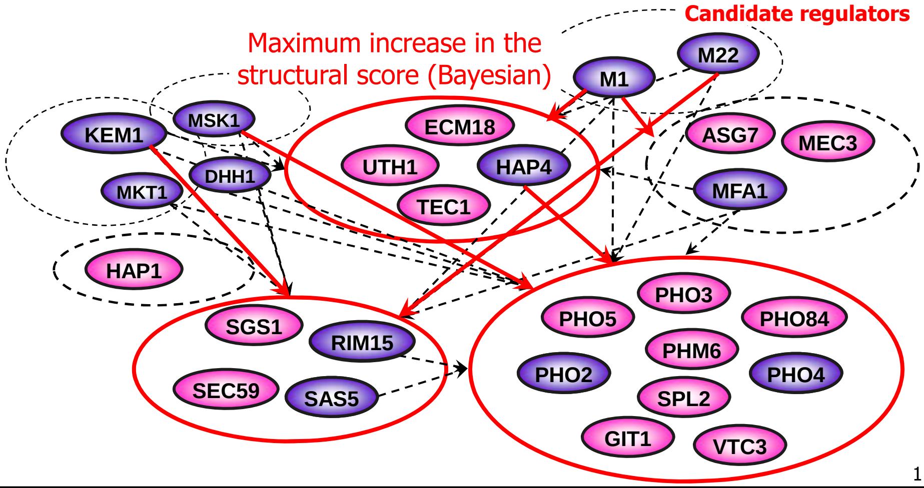
- Structure learning
  - Find the structure that maximizes **Bayesian score** log P(S|D) (or via regularization)

- Expectation Maximization (EM) algorithm
  - M-step: Given a partition of the genes into modules, **learn the best regulation program (tree CPD)** for each module.
  - E-step: Given the inferred regulatory programs, we **reassign genes into modules** such that the associated regulation program best predicts each gene's behavior.

Learning Regulatory Network



- Iterative procedure (EM)
  - Cluster genes into modules (E-step)
  - Learn a regulatory program for each module (tree model) (M-step)

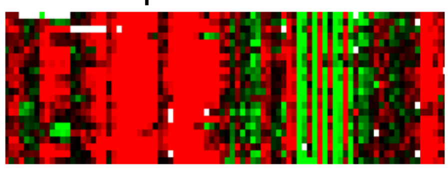


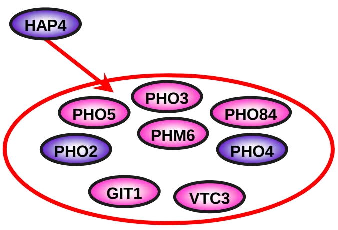
11

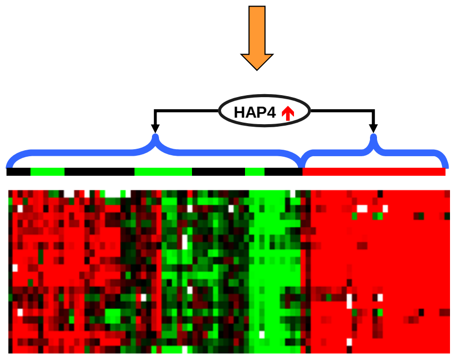
From Expression to Regulation (M-step)

- Combinatorial search over the space of trees

Arrays sorted in original order →





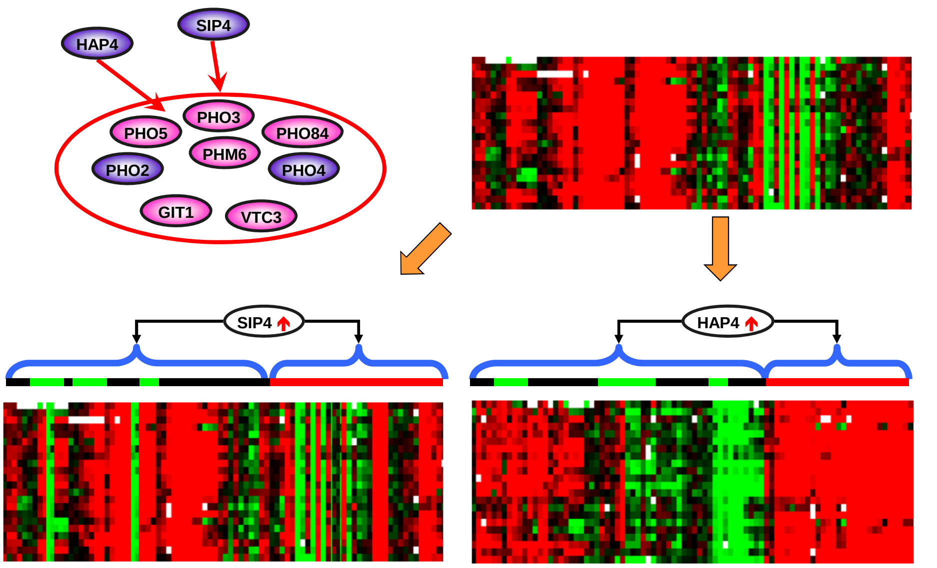


Arrays sorted according to expression of HAP4 →

Segal et al. Nat Genet 2003

12

From Expression to Regulation (M-step)



Segal et al. Nat Genet 2003

Learning Control Programs

- **Score:**

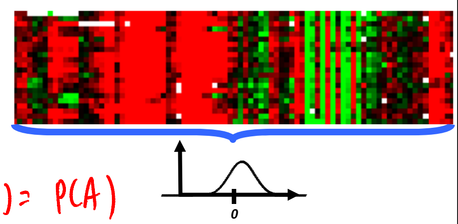
log $P(M \mid D) \propto$

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

$= \frac{P(D|M)P(M)}{P(D)}$

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

$\sum P(A,B) = P(A)$

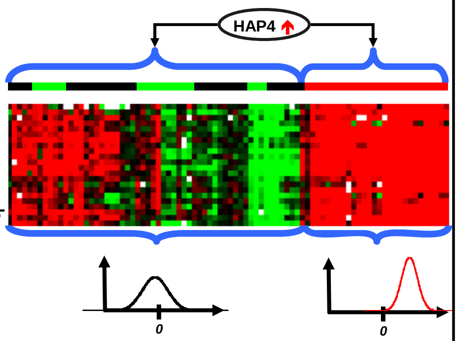


- **Score of HAP4 split:**

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

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

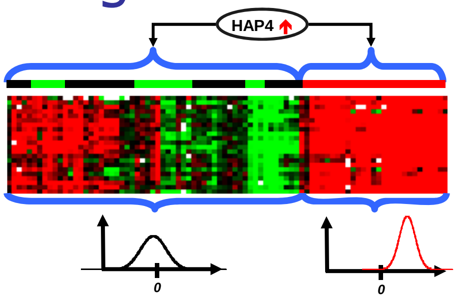
+ log $\int P(D_{HAP4\downarrow} \mid M,\mu,\sigma)P(\mu,\sigma)\, d\mu d\sigma$

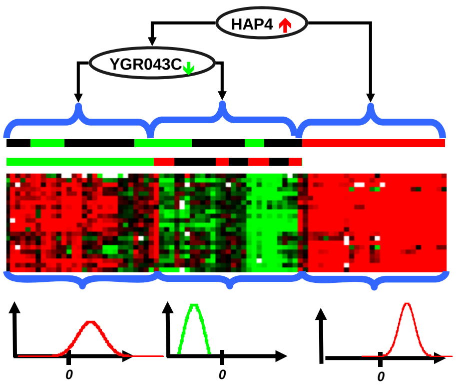


Segal et al. Nat Genet 2003

Learning Control Programs

- Split as long as the score improves





- **Score of HAP4 split:**

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

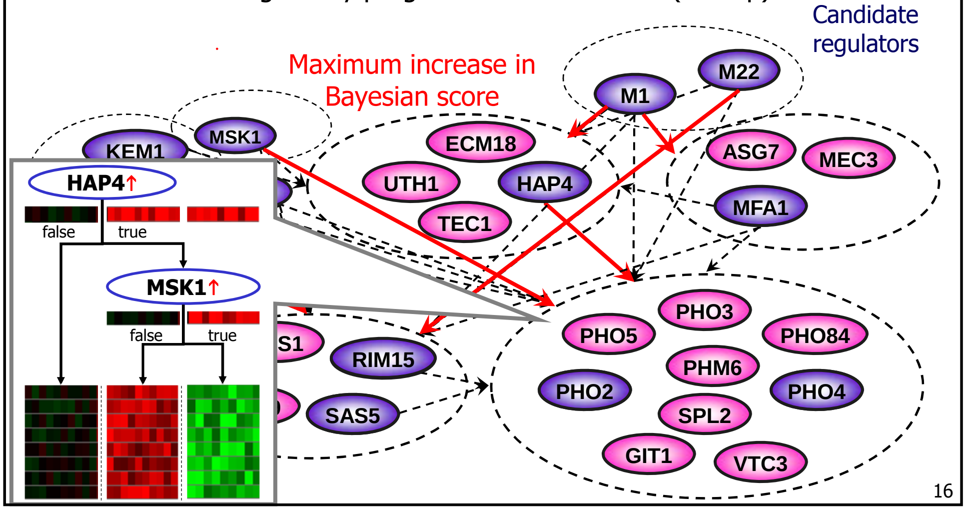
- **Score of HAP4/YGR043C split:**

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

15

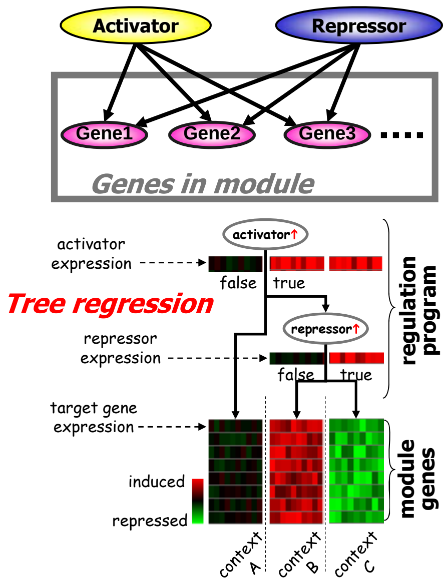
Review – Learning Regulatory Network

- Iterative procedure
  - Cluster genes into modules (E-step)
  - Learn a regulatory program for each module (M-step)



16

Module Networks*

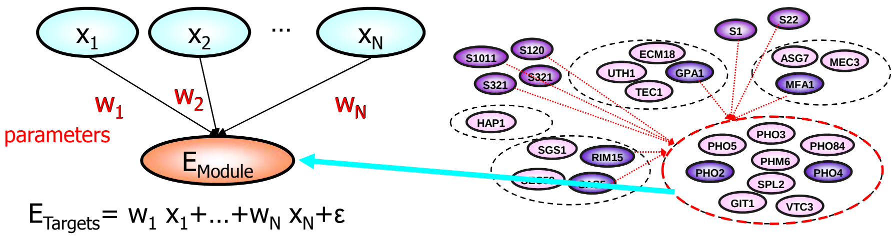


- Learning quickly runs out of statistical power
  - Poor regulator selection lower in the tree
  - Many correct regulators not selected
- Arbitrary choice among correlated regulators
- Combinatorial search
  - Multiple local optima

* Segal et al., Nature Genetics 2003

Regulation as Linear Regression

minimize$_w$ $(w_1x_1 + \dots w_Nx_N - E_{Module})^2$



$E_{Targets} = w_1\, x_1 + \dots + w_N\, x_N + \varepsilon$

- But we often have very large *N*
- … and linear regression gives them all nonzero weight!

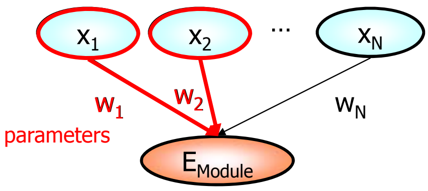
Problem: This objective learns too many regulators

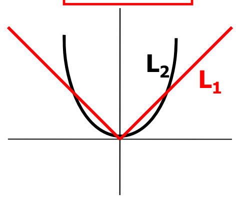
18

Lasso* (L_1) Regression

L_1 term

minimize$_{\mathbf{w}}$ $(w_1x_1 + \ldots w_Nx_N - E_{Module})^2 + \Sigma\, \mathbf{C}\, |w_i|$



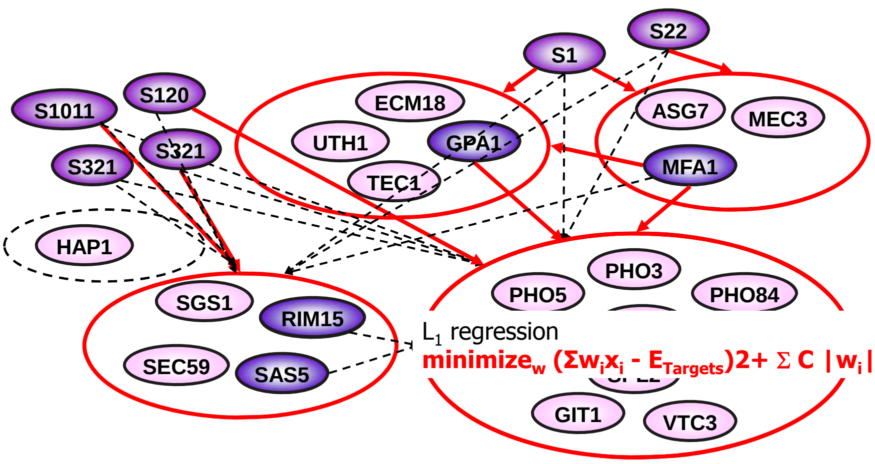


- Induces sparsity in the solution **w** (many w_i's set to zero)
  - Provably selects "right" features when many features are irrelevant
- Convex optimization problem
  - No combinatorial search
  - Unique global optimum
  - Efficient optimization

* Tibshirani, 1996

Learning Regulatory Network

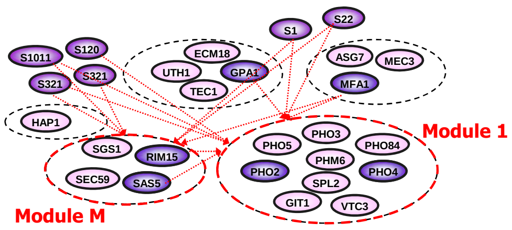
- Cluster genes into modules
- Learn a regulatory program for each module

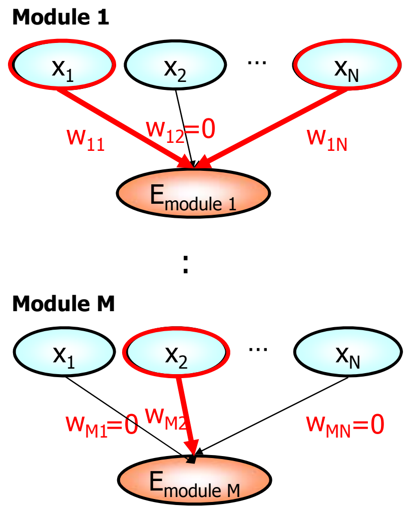


Lee et al., PLoS Genet 2009

Learning the regulatory network

- Multiple regression tasks





$\text{minimize}_{\mathbf{w1}}\ (\Sigma\ w_{1n}x_n - E_{module1})^2 + \Sigma\ \mathbf{C}|\mathbf{w_{1n}}|$

:

$\text{minimize}_{\mathbf{wn}}\ (\Sigma\ w_{Mn}x_n - E_{moduleM})^2 + \Sigma\ \mathbf{C}|\mathbf{w_{Mn}}|$

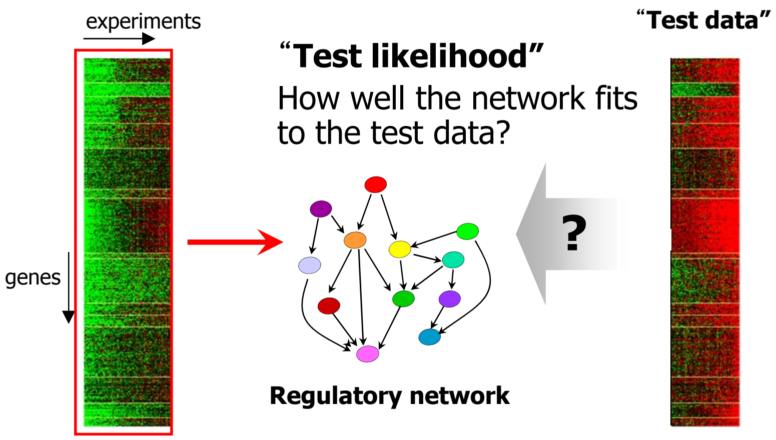
Outline

- Motivation
- Algorithms for learning regulatory networks
- Evaluation of the method
  - Statistical evaluation
  - Biological interpretation
- Advanced topics
  - Many models can get similar scores. Which one would you choose?
  - A gene can be involved in multiple modules.
  - Possible to incorporate prior knowledge?

22

Statistical Evaluation

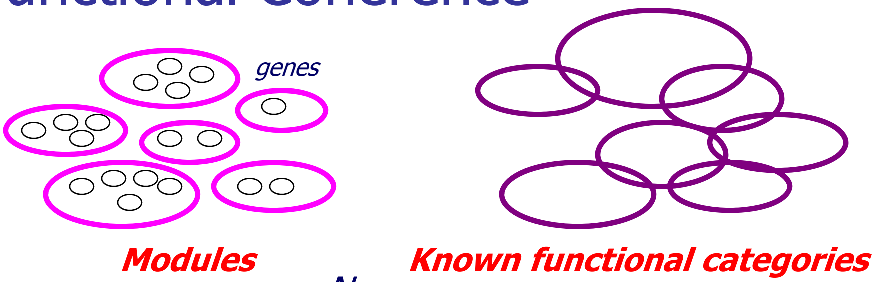
- Cross-validation test
  - Divide the data (experiments) into training and test data
  - Compute the likelihood function for the **Test data**

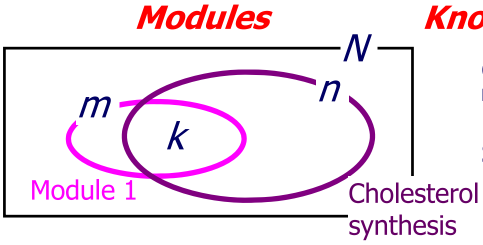


Module Evaluation Criteria

- Are the module genes functionally coherent?
- Do the regulators have regulatory roles in the predicted conditions C (see slide 6) ?
- Are the genes in the module known targets of the predicted regulators?
- Are the regulators consistent with the *cis*-regulatory motifs (TF binding sites) found in promoters of the module genes?

Functional Coherence

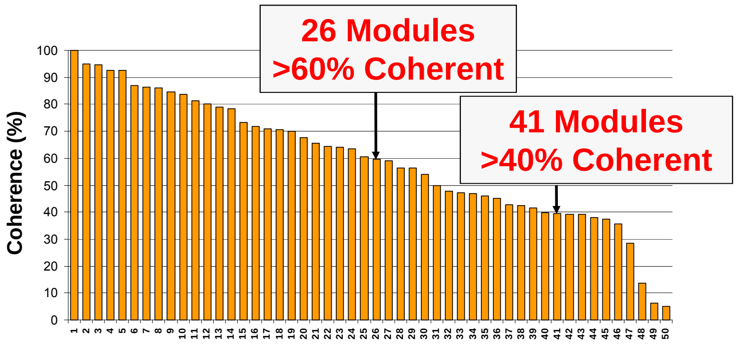


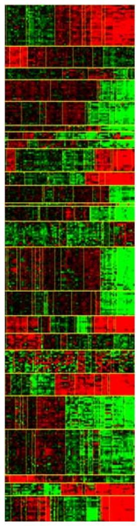


Gene ontology (GO)
http://www.geneontology.org/
Predicted targets of regulators
Sharing TF binding sites
:

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

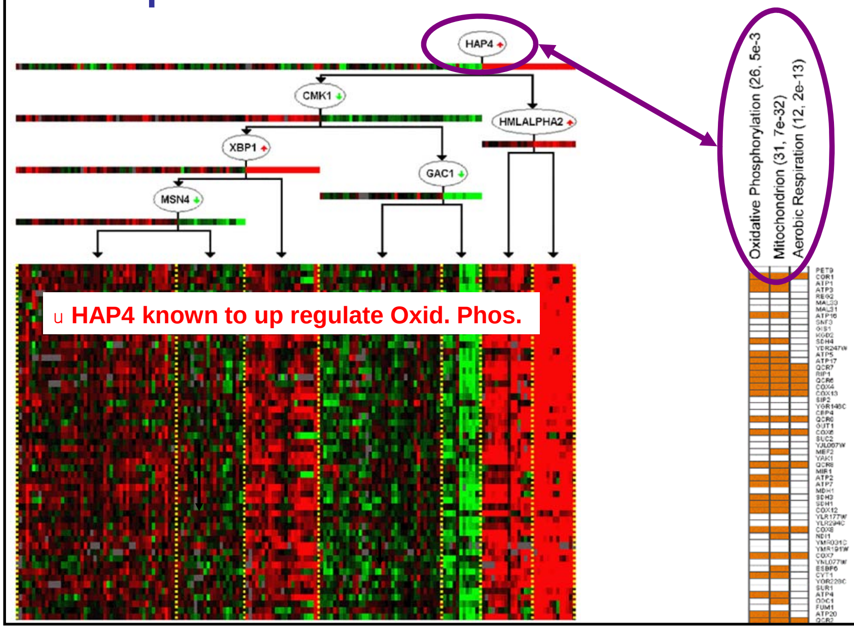
Module Functional Coherence





- **Metabolic:** AA, respiration, glycolysis, galactose
- **Stress:** Oxidative stress, osmotic stress
- **Cellular localization:** Nucleas, ER
- **Cellular processes:** Cell cycle, sporulation, mating
- **Molecular functions:** Protein folding, RNA & DNA processing, trafficking

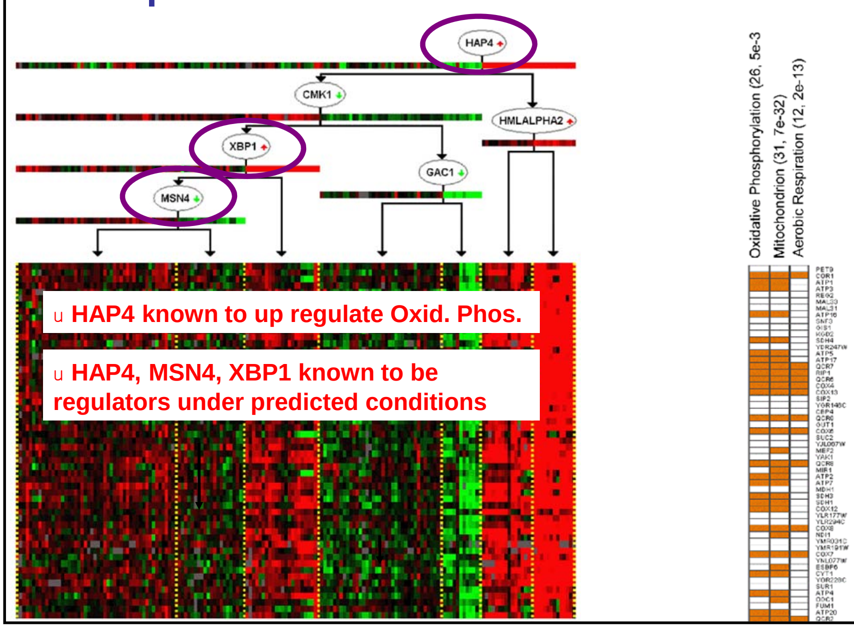
Respiration Module



- HAP4 known to up regulate Oxid. Phos.

27

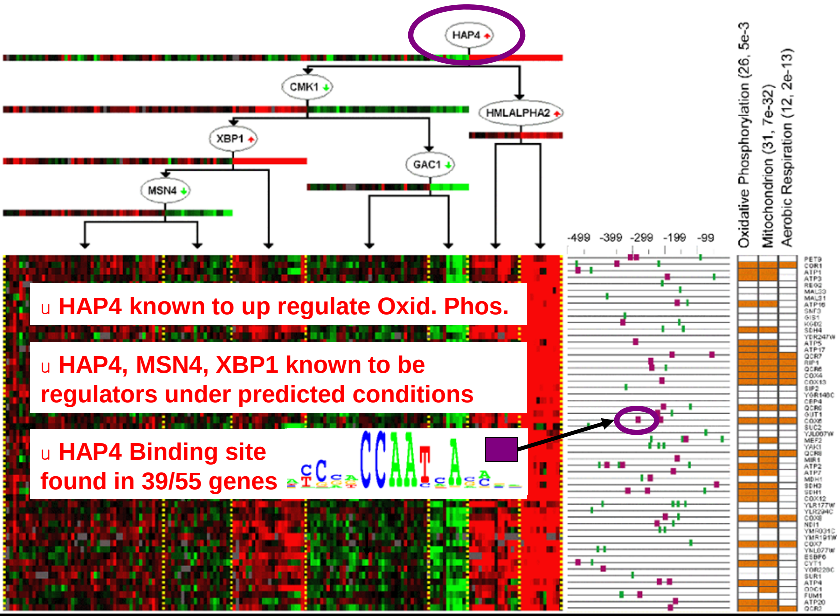
Respiration Module



- HAP4 known to up regulate Oxid. Phos.
- HAP4, MSN4, XBP1 known to be regulators under predicted conditions

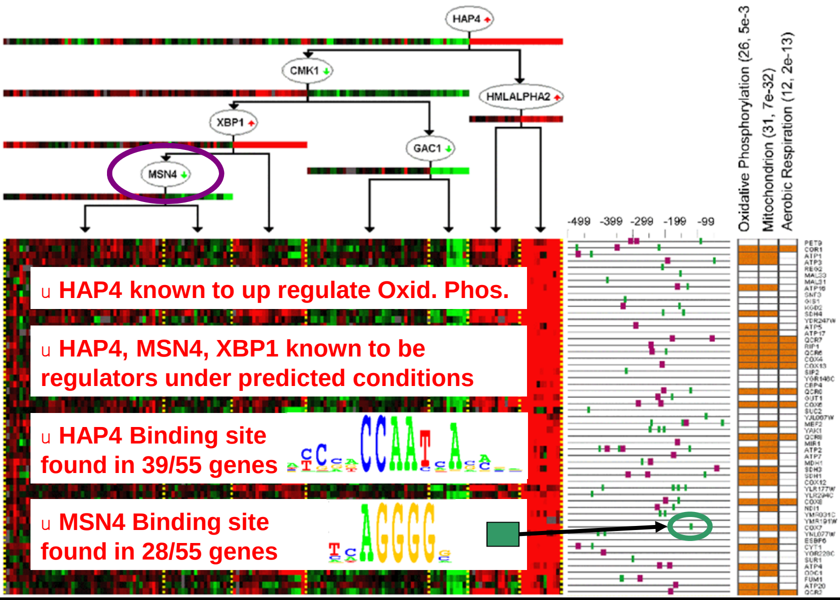
28

Respiration Module



29

Respiration Module



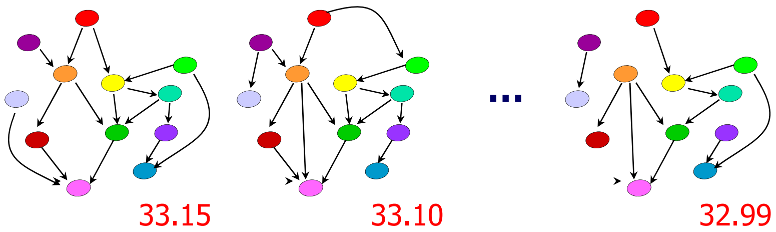
30

Outline

- Motivation
- Algorithms for learning regulatory networks
- Evaluation of the method
- Advanced topics
  - Many models can get similar scores. Which one would you choose?
  - A gene can be involved in multiple modules.
  - Possible to incorporate prior knowledge?

Structural learning via boostrapping

- Many networks that achieve similar scores

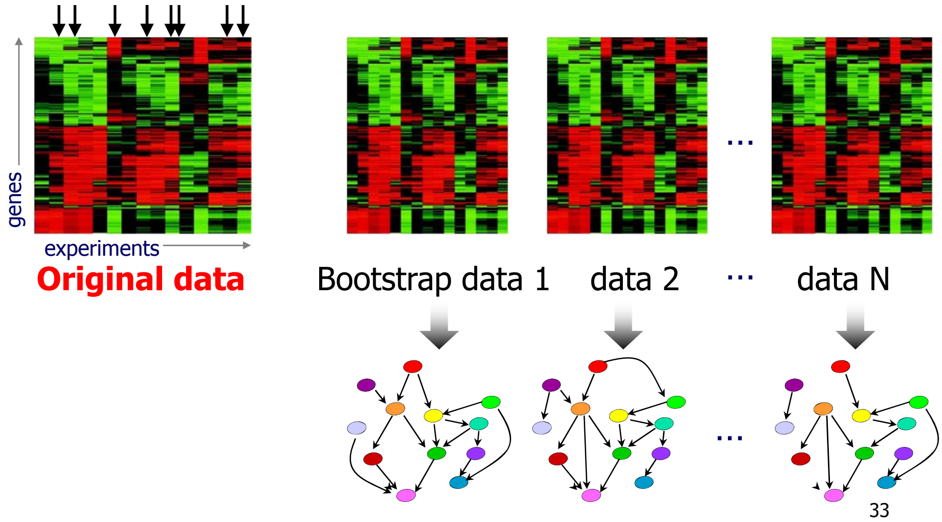


- Which one would you choose?
  - Estimate the robustness of each network or each edge.
  - How?? Learn the networks from **multiple datasets**.

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

Bootstrapping

- Sampling with replacement



33

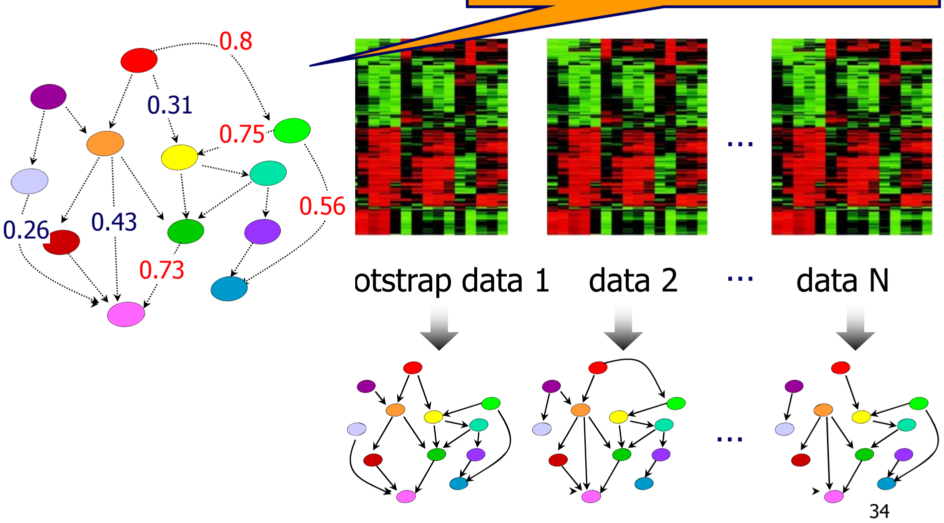
Inferring sub-networks from perturbed expression profiles, Pe'er et al. Bioinformatics 2001

Bootstrapping

- Sampling with replacement

- Estimated confidence of each edge *i*

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

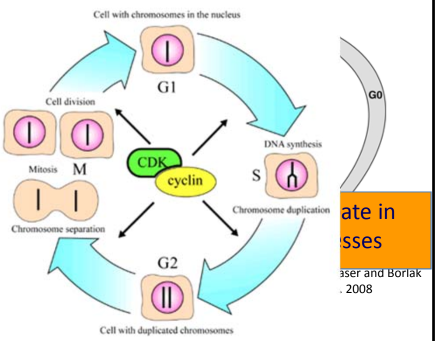


34

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

Overlapping Processes

- The living cell is a complex system
  - Example, the cell cycle
    - Cell cycle: the series of events that take place in a cell leading to its division and duplication.
  - Genes functionally relevant to cell cycle regulation in the specific cell cycle phase

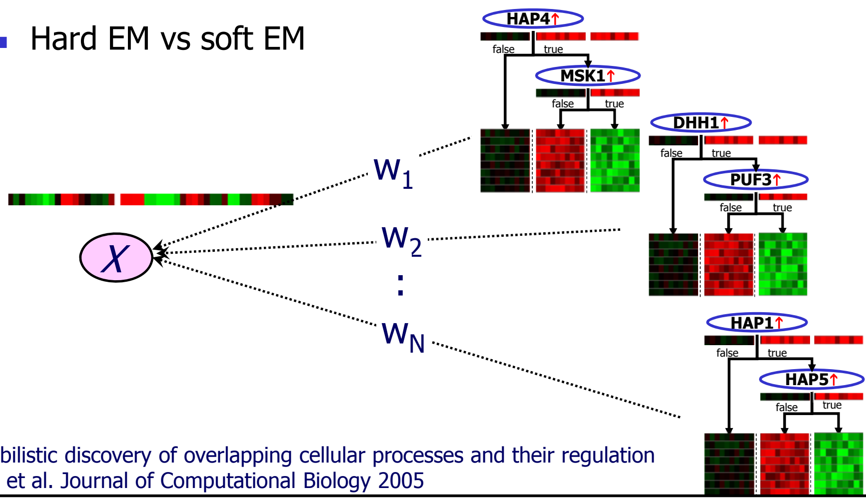


- Mutually exclusive clustering as a common approach to analyzing gene expression
  - (+) genes likely to share a common function
  - (-) group genes into mutually exclusive clusters
  - (-) no info about genes relation to one another

35

Decomposition of Processes…

- Model an expression level of a gene as **a mixture of regulatory modules**.
- Hard EM vs soft EM



Probabilistic discovery of overlapping cellular processes and their regulation
Battle et al. Journal of Computational Biology 2005