Statistical methods for inferring the gene regulatory networks – Part II

Lecture 2 – June 5th, 2014
GENOME 541, Spring 2014

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

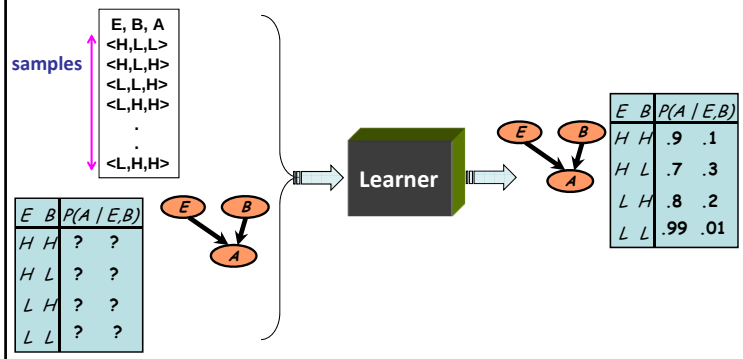
1

Outline

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

2

Known structure, complete data

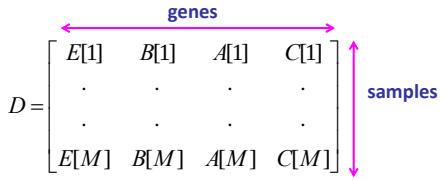


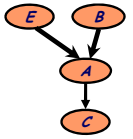
- Network structure is specified
  - Learner needs to estimate parameters

3

Learning parameters

- Training data has the form:



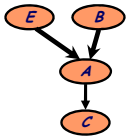


4

Likelihood function

- Assume i.i.d. samples
- Likelihood function is defined as:

$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta)$$



genes

samples

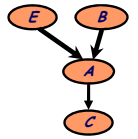
$$\begin{bmatrix} E[1] & B[1] & A[1] & C[1] \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ E[M] & B[M] & A[M] & C[M] \end{bmatrix}$$

Likelihood function

- Joint distribution can be decomposed as:

$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta)$$

$$= \prod_m \begin{pmatrix} P(E[m] : \Theta) \times \\ P(B[m] : \Theta) \times \\ P(A[m] \mid B[m], E[m] : \Theta) \times \\ P(C[m] \mid A[m] : \Theta) \end{pmatrix}$$



genes

samples

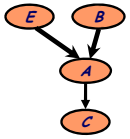
$$\begin{bmatrix} E[1] & B[1] & A[1] & C[1] \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ E[M] & B[M] & A[M] & C[M] \end{bmatrix}$$

Likelihood function

- Reordering terms, we got

$$L(\Theta : D) = \prod_m P(E[m], B[m], A[m], C[m] : \Theta)$$

$$= \begin{pmatrix} \prod_m P(E[m] : \Theta_E) \times \\ \prod_m P(B[m] : \Theta_B) \times \\ \prod_m P(A[m] \mid B[m], E[m] : \Theta_{A|B,E}) \times \\ \prod_m P(C[m] \mid A[m] : \Theta_{C|A}) \end{pmatrix}$$



- Parameters can be estimated for each variable independently!

$$\begin{bmatrix} E[1] & B[1] & A[1] & C[1] \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ E[M] & B[M] & A[M] & C[M] \end{bmatrix}$$

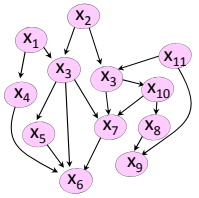
General Bayesian networks

- Generalization for any Bayesian network:

$$L(\Theta : D) = \prod_m P(x_1[m], \ldots, x_n[m] : \Theta)$$

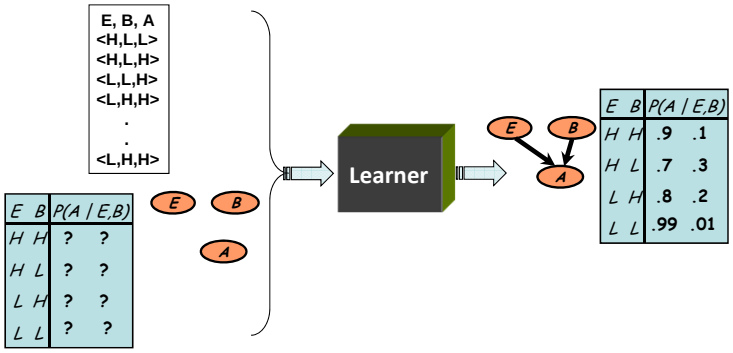
$$= \prod_m \prod_i P(x_i[m] \mid Pa_i[m] : \Theta_i)$$

$$= \prod_i L_i(\Theta_i : D)$$



- Parameters can be estimated for each variable independently!

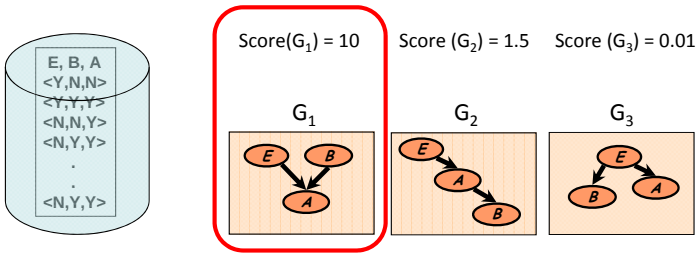
Unknown structure, complete data



- Network structure is not specified
  - Learner needs to estimate both structure and parameters

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.

10

Structure score

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

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

Marginal likelihood | Likelihood | Prior distribution over parameters

- Penalized likelihood score

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

11

Decomposability of scores

- Likelihood score

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

- Bayesian score

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

12

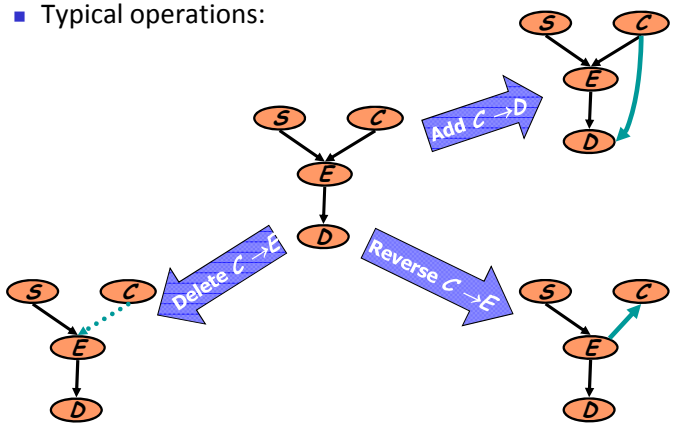
Search for optimal network structure

- 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.

13

Search for optimal network structure

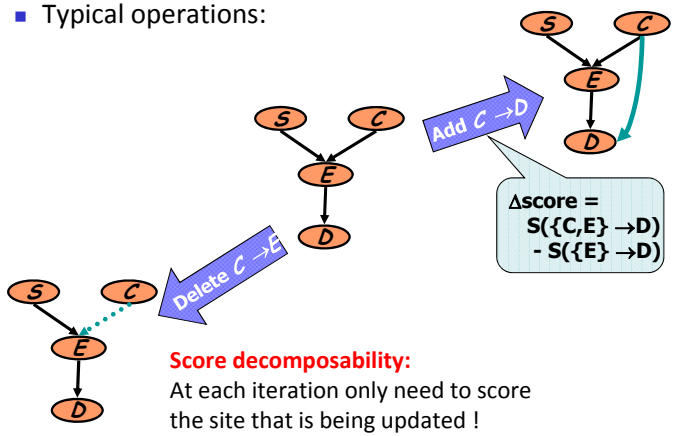
- Typical operations:



14

Search for optimal network structure

- Typical operations:



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

15

Outline

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

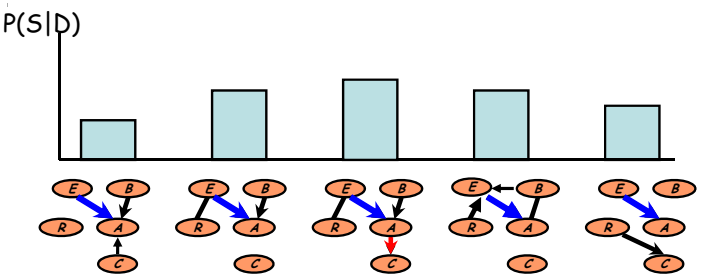
16

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

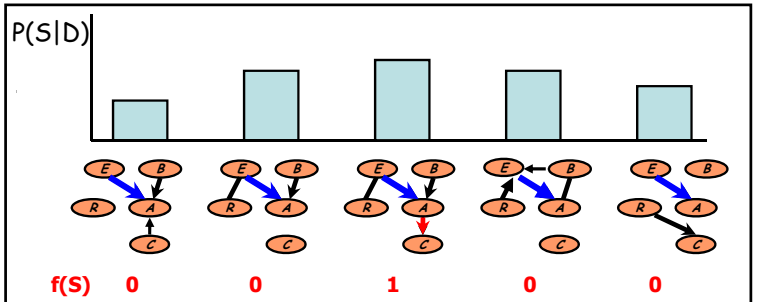
17

Model averaging



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

18



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

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

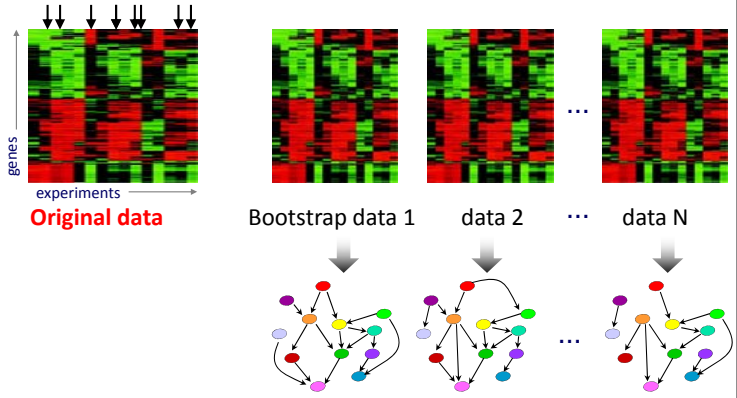
- We are interested in computing

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

19

Bootstrapping

- Sampling with replacement



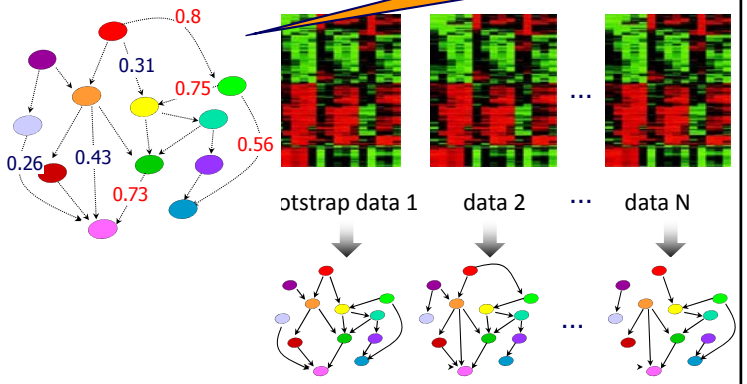
Inferring sub-networks from perturbed expression profiles, Pe'er et al. Bioinformatics 2001

20

Bootstrapping

- Sampling with replacement

Estimated confidence of each edge *i* = # networks that contain the edge / total # networks (N)



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

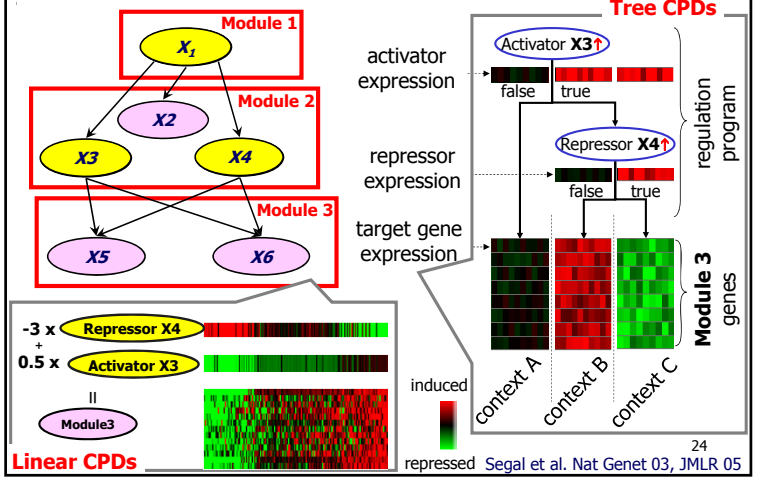
Outline

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

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

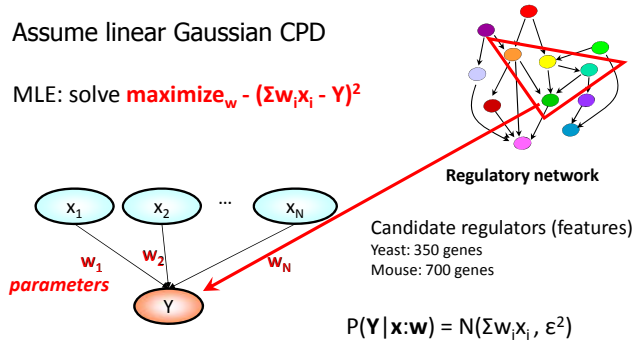
The Module networks concept



Segal et al. Nat Genet 03, JMLR 05

Feature selection via regularization

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



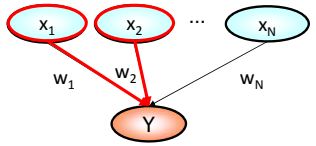
Candidate regulators (features)
Yeast: 350 genes
Mouse: 700 genes

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

Problem: This objective learns too many regulators

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$_w$ $(\Sigma w_i x_i - Y)^2 + \Sigma C |w_i|$: convex optimization!
    $\Rightarrow$ Induces sparsity in the solution **w** (Many w_i's set to zero)

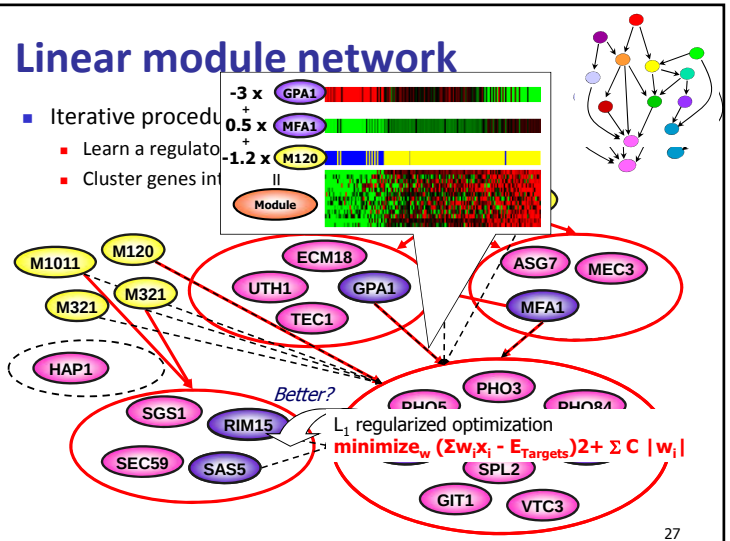


Candidate regulators (features)
Yeast: 350 genes
Mouse: 700 genes

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

Linear module network

- Iterative procedu
  - Learn a regulato
  - Cluster genes int



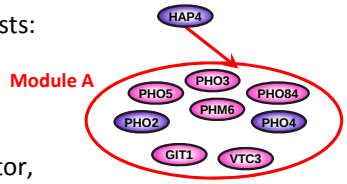
Lee et al., PLoS Genet 2009

Outline

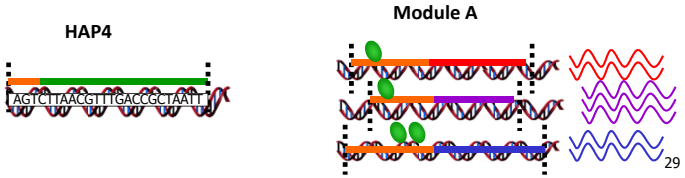
- Basic concepts on Bayesian networks
- Probabilistic models of gene regulatory networks
- Learning algorithms
- Recent probabilistic approaches to reconstructing the regulatory networks
- Evaluation
  - Predicted co-regulated groups of genes
  - Putative regulator-regulatees

Predicted regulatory interaction I

- Say that your network suggests:



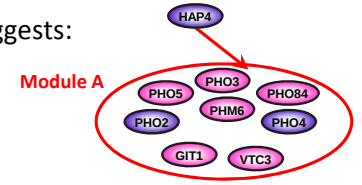
- If HAP4 is a transcription factor,
  - Targets should have a binding site for HAP4.
  - Or there should be different kind of evidence that HAP4 binds to genes in Module A (chip-chip or chip-seq data).



29

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.

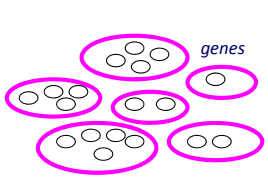
30

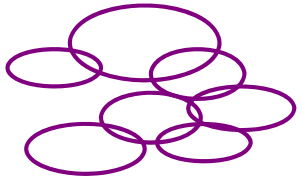
Create functional categories

- For each Gene Ontology (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]

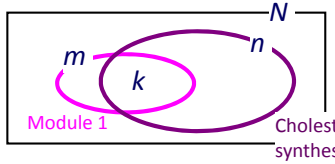
Functional categories

Functional coherence





Modules ***Known functional categories***



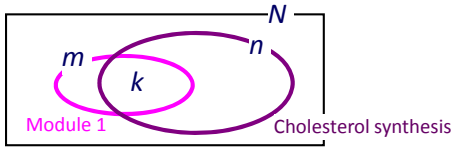
Gene ontology (GO)
http://www.geneontology.org/
Predicted targets of regulators
Sharing TF binding sites
:

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

32

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.

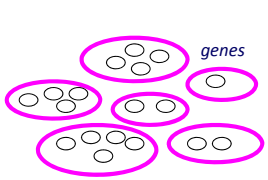


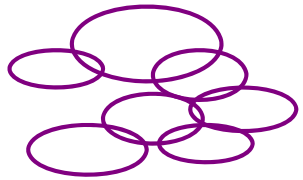
- How significant is the overlap?
  - Calculate p-value = P(# overlap ≥ k | m, n, N; two groups are independent), based on the hypergeometric distribution
  - What p-values are considered to be significant?

33

Multiple hypothesis testing

- Say that there are 200 modules and 3000 functional categories





Modules ***Known 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

34

Summary

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

35