Inferring Transcriptional Regulatory Networks from High-throughput Data

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

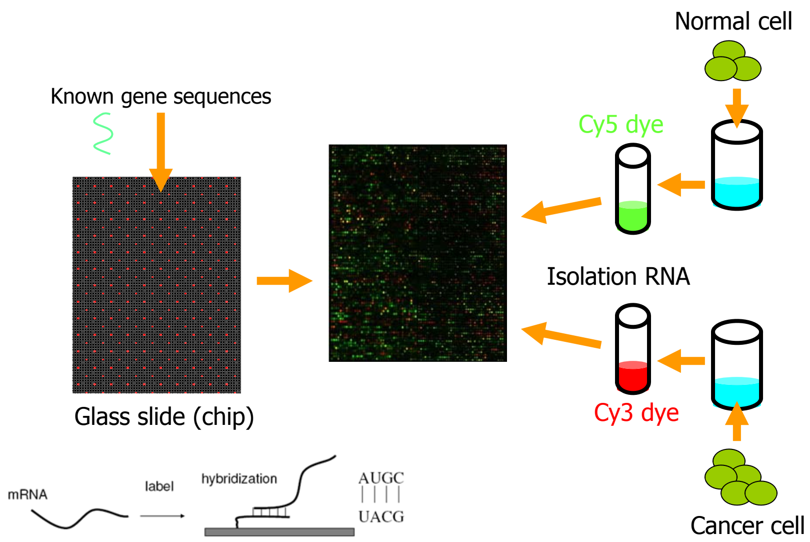
Instructor: Su-In Lee
TA: Christopher Miles

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

Outline

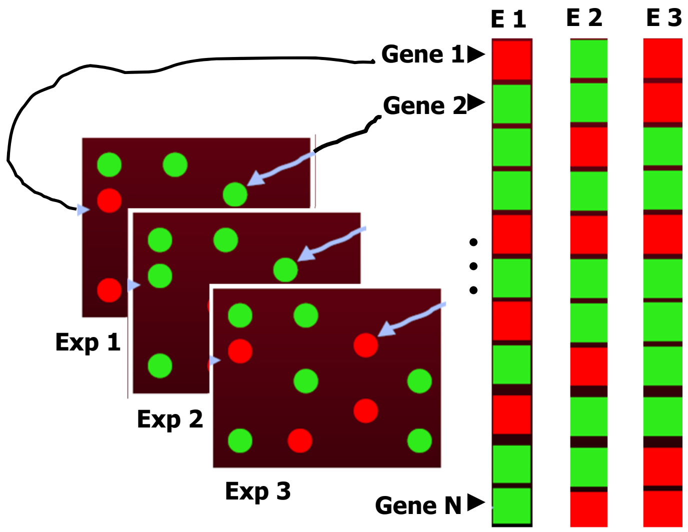
- Microarray gene expression data
  - Measuring the RNA level of genes
- Clustering approaches
- Beyond clustering
- Algorithms for learning regulatory networks
  - Application of probabilistic models
  - Structure learning of Bayesian networks
  - Module networks
- Evaluation of the method

Spot Your Genes



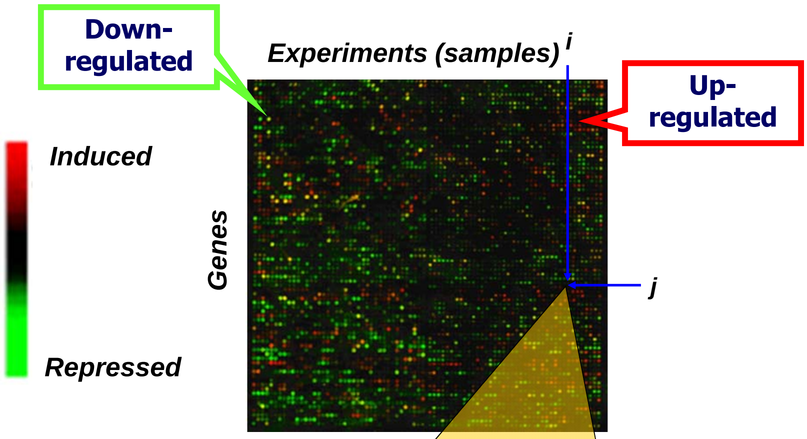
3

Matrix of expression



4

Microarray gene expression data

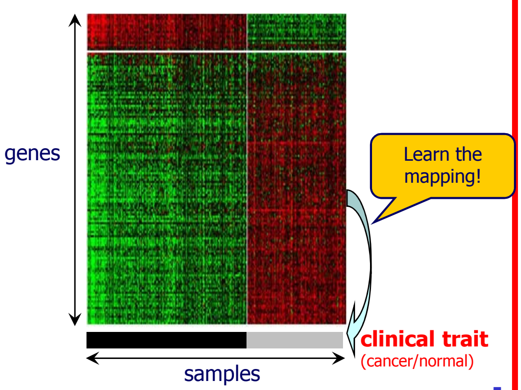


E_{ij} - RNA level of gene j in experiment i

5

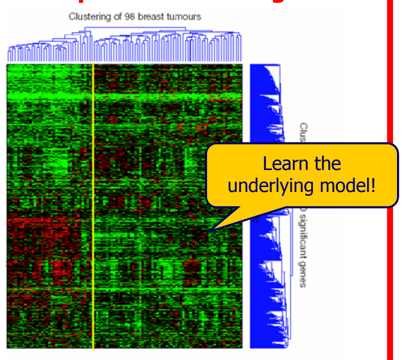
Analyzing micrarray data

Supervised learning problems



- Gene signatures can provide valuable diagnostic tool for clinical purposes
- Can also help the molecular characterization of cellular processes underlying disease states

Un-supervised learning

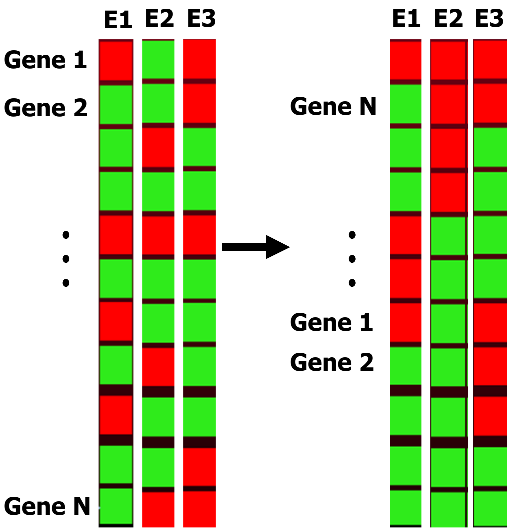


- Gene clustering can reveal cellular processes and their response to different conditions
- Sample clustering can reveal phenotypically distinct populations, with clinical implications

* Bhattacharjee et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS* (2001).

* van't Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* (2002).

Why care about clustering?



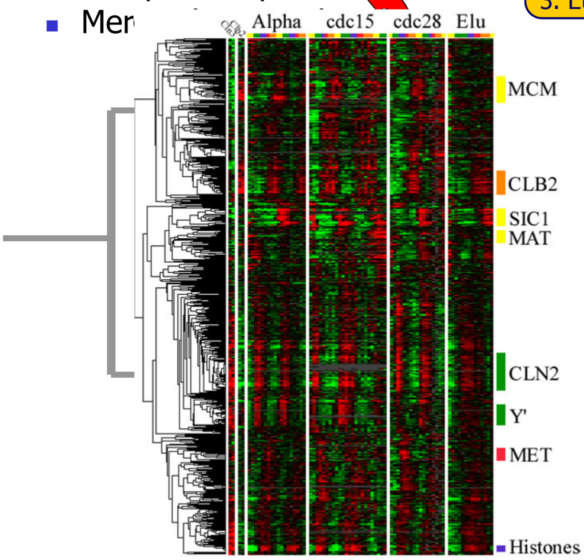
- Discover functional relation
  - Similar expression ⇒ functionally related
- Assign function to unknown genes
- Find which gene controls which other genes

7

Hierarchical clustering

- Compute all pairwise distances
- Mer

1. Euclidean distance
2. (Pearson's) correlation coefficient
3. Etc etc…

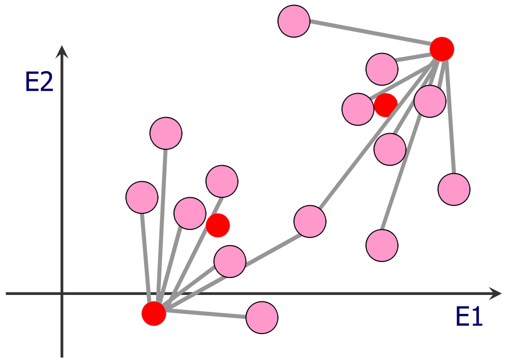


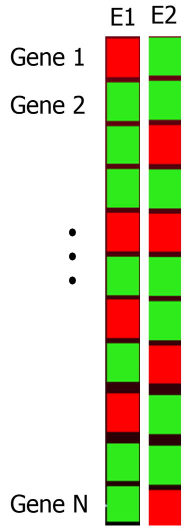
...

- Easy
- Depends on where to start the grouping
- Trouble to interpret the "tree" structure

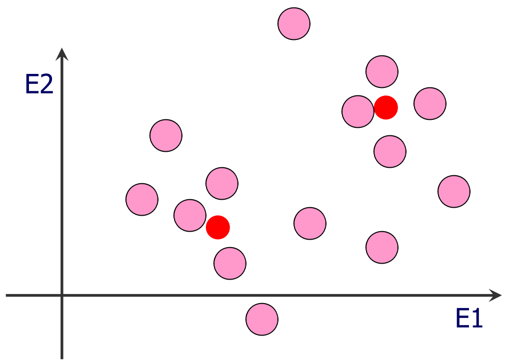
8

K-means clustering





K-means clustering



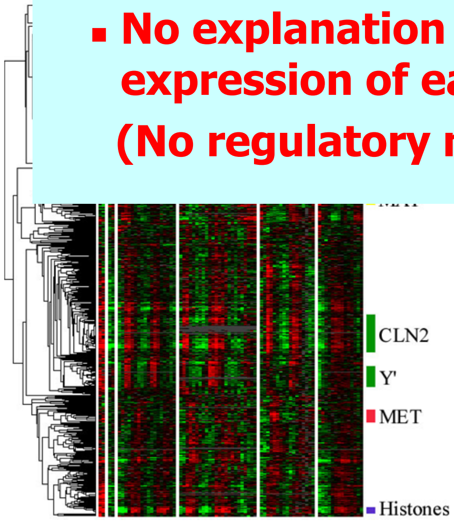
- Overall optimization
- How many (k) ?
- How to initiate ?
- Local minima

- Generally, heuristic methods have no established means to determine the "correct" number of clusters and to choose the "best" algorithm.

Cl

Limitations:

- **No explanation on what caused expression of each gene**
  (No regulatory mechanism)



Co-regulated genes cluster together

⇩

Infer gene function

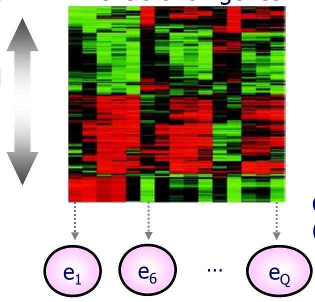
Beyond Clustering

- **Cluster**: set of genes with similar expression profiles
- **Regulatory module**: set of genes with shared regulatory mechanism

- **Goal:**
  - Automatic method for identifying candidate modules and their regulatory mechanism

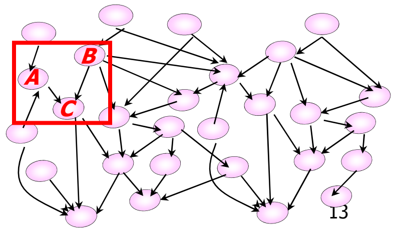
Inferring regulatory networks

"Expression data" – measurement of mRNA levels of all genes



- Infer the regulatory network that controls gene expression
  - Causal relationships among $e_{1\text{-}Q}$

    A and ***B*** regulate the expression of ***C***
    (A and B are regulators of C)



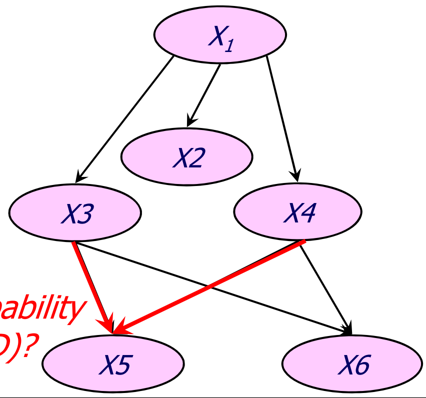
- Bayesian networks

13

Regulatory network

- Bayesian network representation
  - Xi: expression level of gene i
  - Val(Xi): continuous
- Interpretation
  - Conditional independence
  - Causal relationships
- Joint distribution
  - P(**X**) = $\prod_i P(X_i \mid Pa_{X_i})$

Conditional probability distribution (CPD)?



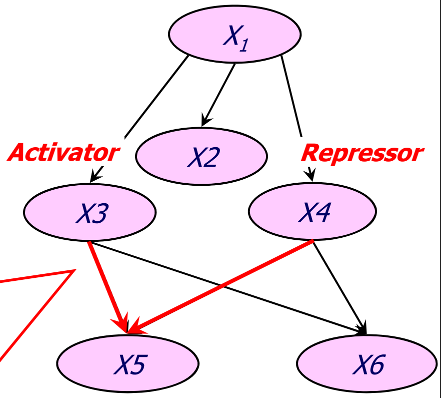
CPD for Discrete Expression Level

- After discretizing the expression levels to high/low,
  - Parameters – probability values in every entry

Table CPD

	X5=high	X5=low
X3=high, X4=high	0.3	0.7
X3=high, X4=low	0.95	0.05
X3=low, X4=high	0.1	0.9
X3=low, X4=low	0.2	0.8

parameters



- How about continuous-valued expression level?
  - Tree CPD; Linear Gaussian CPD

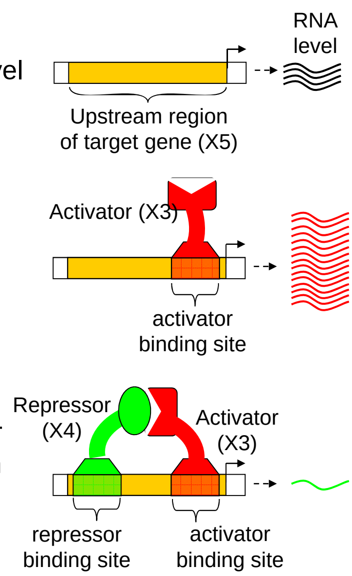
15

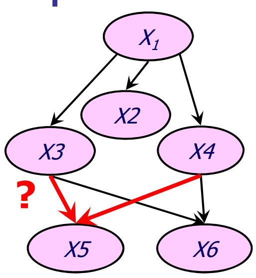
Context specificity of gene expression

Context A
Basal expression level

Context B
Activator induces expression

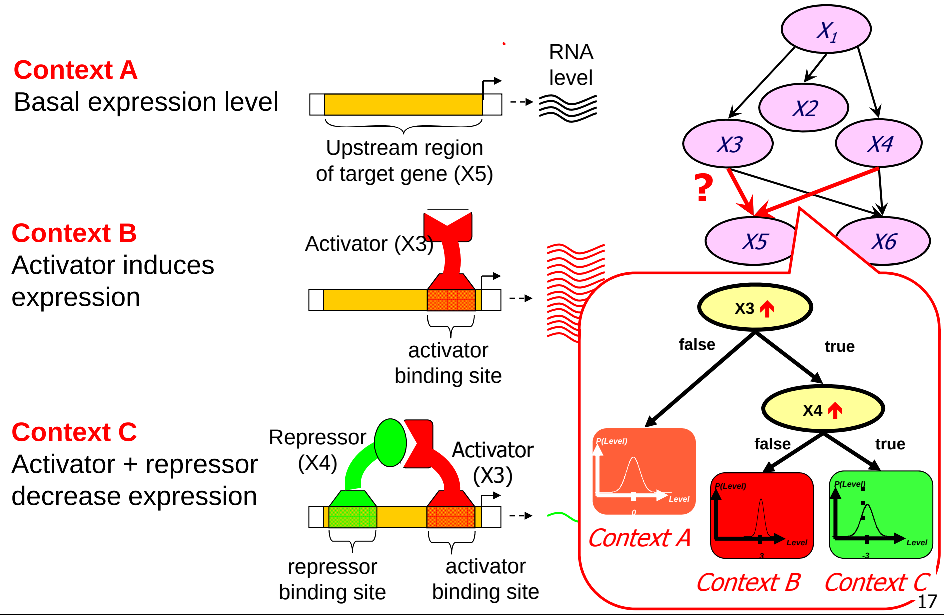
Context C
Activator + repressor decrease expression





16

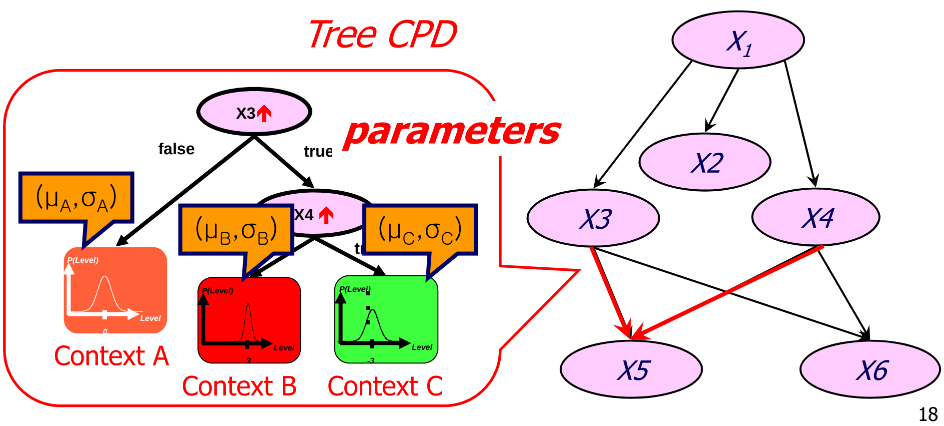
Context specificity of gene expression



17

Continuous-valued expression I

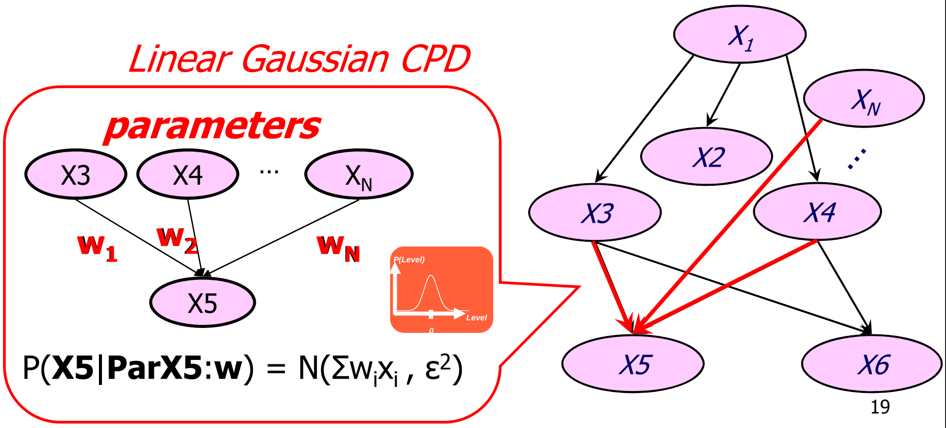
- Tree conditional probability distributions (CPD)
  - Parameters – mean (μ) & variance (σ^2) of the normal distribution in each context
  - Represents combinatorial and context-specific regulation



18

Continuous-valued expression II

- Linear Gaussian CPD
  - Parameters – weights $w_1,\ldots,w_N$ associated with the parents (regulators)

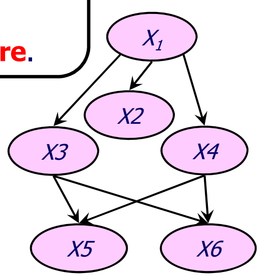


19

Learning

- Structure learning [Koller & Friedman]
  - Constraint based approaches
  - Score based approaches
  - Bayesian model averaging

Given **a set of all possible network structures** and the **scoring function** that measures how well the model fits the observed data, we try to **select the highest scoring network structure**.



- Scoring function
  - Likelihood score
  - Bayesian score

Scoring Functions

- Let S: structure, Θ_S: parameters for S, D: data
- Likelihood score

$p(D|S,\hat{\theta}_s) \leftarrow \quad \hat{\theta}_s = \mathrm{argmax}_{\theta_s}\, p(D|S,\theta_s)$

- How to overcome overfitting?
  - Reduce the complexity of the model
    - Bayesian score: P(**S**tructure|**D**ata)

$p(S|D) \propto p(D|S)\cdot p(S)$

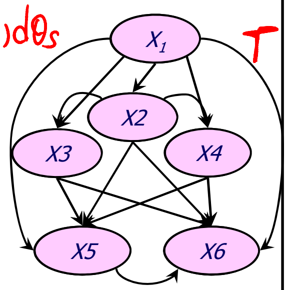
$= \int p(D|S,\theta_s)\, p(\theta_s|S)\, d\theta_s$

    - Regularization

$p(D|S,\theta_s)\; p(S,\theta_s)$

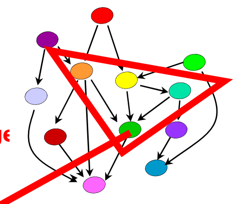
  - Simplify the structure
    - Module networks

Structure S

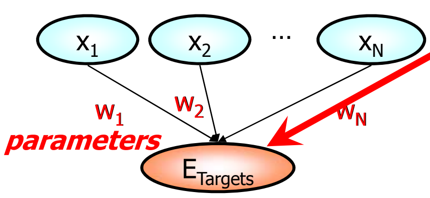


Feature Selection Via Regularization

- Assume linear Gaussian CPD
- MLE: solve **maximize$_w$ - ($\Sigma w_i x_i$ - E$_{Targe}$**



Regulatory network



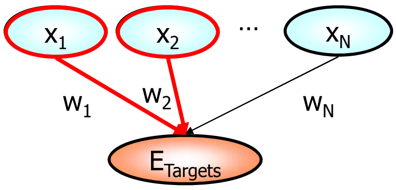
Candidate regulators (features)
Yeast: 350 genes
Mouse: 700 genes

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

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$_{\mathbf{w}}$ $(\Sigma w_i x_i - E_{Targets})^2 + \Sigma\ \mathbf{C}\ |w_i|$: convex optimization!
    $\Rightarrow$ Induces sparsity in the solution $\mathbf{w}$ (Many w_i's set to zero)

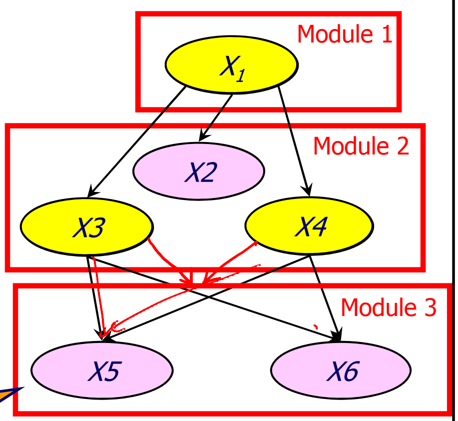


Candidate regulators (features)
Yeast: 350 genes
Mouse: 700 genes

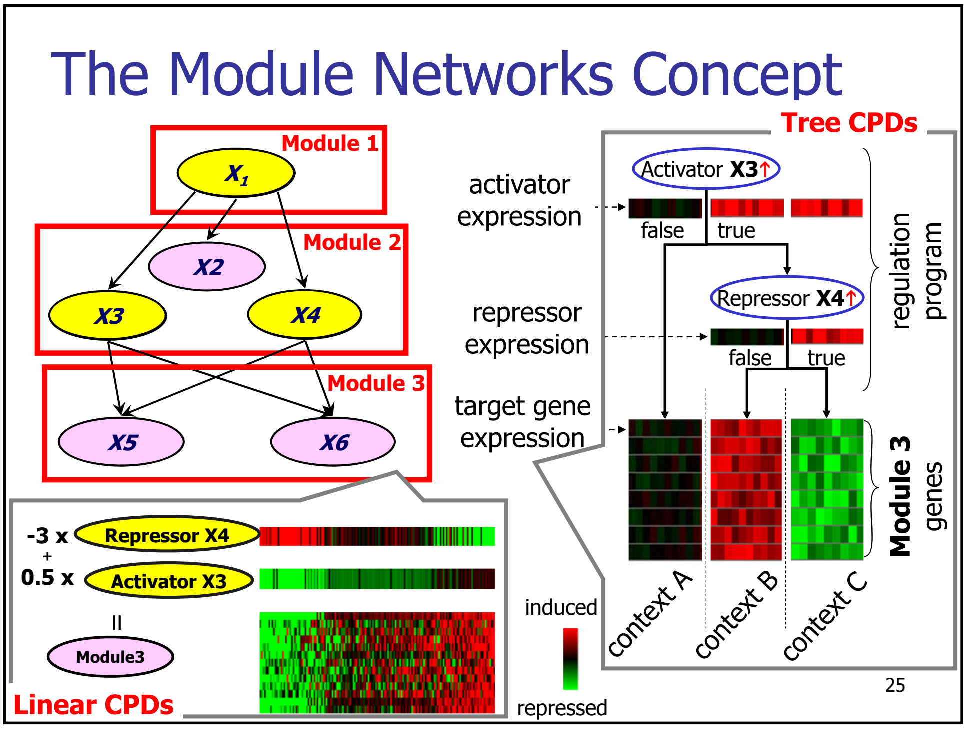
$$P(\mathbf{E}_{Targets} | \mathbf{x}:\mathbf{w}) = N(\Sigma w_i x_i, \varepsilon^2)$$

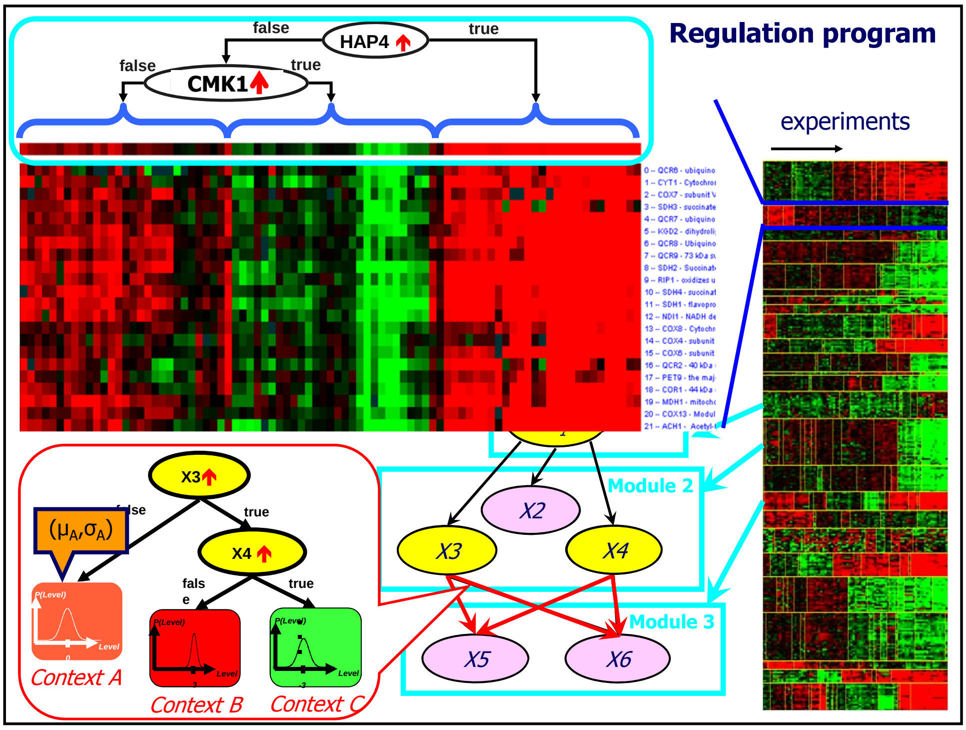
Modularity of Regulatory Networks

- Genes tend to be co-regulated with others by the same factors.
- Biologically more relevant
- More compact representation
  - Smaller number of parameters
  - Reduced search space for structure learning
- Candidate regulators
  - A fixed set of genes that can be parents of other modules.



Same module $\Rightarrow$ Share the CPD


The Module Networks Concept
Tree CPDs
Module 1
X1
Module 2
X2
X3
X4
Module 3
X5
X6
activator expression
repressor expression
target gene expression
Activator X3
false
true
Repressor X4
false
true
regulation program
Module 3 genes
context A
context B
context C
-3 x
Repressor X4
+
0.5 x
Activator X3
=
Module3
Linear CPDs
induced
repressed
25


Regulation program
false
HAP4
true
false
CMK1
true
experiments
X3
true
X4
true
(μA,σA)
Context A
Context B
Context C
Module 2
X2
X3
X4
Module 3
X5
X6

Structure Learning – Bayesian Score & Tree CPD

- Find the structure S that maximizes P(S|D)
  - P(**S**tructure|**D**ata) ∝ P(D|S) P(S)
  - maximize$_S$ log P(D|S) + log P(S)

  $\Rightarrow$ P(D|S) = $\int$ P(D|S,Θ_S) P(Θ_S|S) dΘ_S

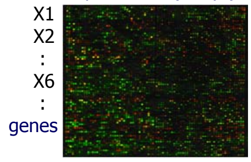
  P(S): prior distribution on the structure

  maximize$_S$ log $\int$ P(D|S,Θ_S) P(Θ_S|S) dΘ_S + log P(S)

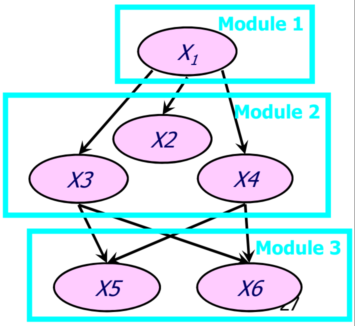
ML score: max$_\Theta$ log P(D|S,Θ)
$\Rightarrow$ More prone to overfitting

Data **D**

experiments (arrays)



Structure **S** ?



Structure Learning – Bayesian Score & Tree CPD

- Find the structure S that maximizes P(S|D)
  - P(**S**tructure|**D**ata) ∝ P(D|S) P(S)
  - maximize$_S$ log P(D|S) + log P(S)

  $\Rightarrow$ P(D|S) = $\int$ P(D|S,Θ_S) P(Θ_S|S) dΘ_S

  P(S): prior distribution on the structure

  maximize$_S$ log $\int$ P(D|S,Θ_S) P(Θ_S|S) dΘ_S + log P(S)
- Decomposability
  - For a certain structure S, log P(D|S)

  = log $\int$ P(D|S,Θ_S) P(Θ_S|S) dΘ_S

dΘ_{m1} dΘ_{m2} dΘ_{m3}

P(X1|Θ_{m1})P(X2,X3,X4|X1,Θ_{m2}) P(X5,X6|X3,X4,Θ_{m3})

P(Θ_{m1})P(Θ_{m2}) P(Θ_{m3})

module 1 score

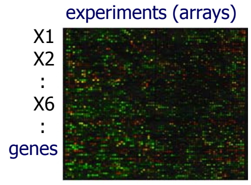
= log $\int$ P(X1|Θ_{m1}) P(Θ_{m1}) dΘ_{m1}

module 2 score + log $\int$ P(X2,X3,X4|X1,Θ_{m2}) P(Θ_{m2}) dΘ_{m2}

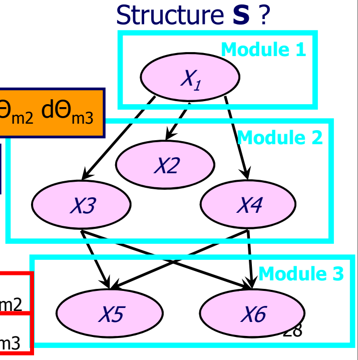
module 3 score + log $\int$ P(X5,X6|X3,X4,Θ_{m3}) P(Θ_{m3}) dΘ_{m3}

Data **D**

experiments (arrays)



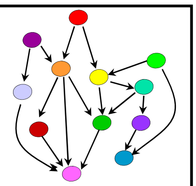
Structure **S** ?



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