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

Known gene sequences

Glass slide (chip)

Isolation RNA

Cy3 dye

Cy5 dye

Normal cell

Cancer cell

Matrix of expression

Gene 1

Gene 2

Gene N

Exp 1

Exp 2

Exp 3

E 1

E 2

E 3
Microarray gene expression data

Experiments (samples) $i$

Genes

Induced

Down-regulated

Repressed

Up-regulated

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

Analyzing microarray data

Supervised learning problems

Un-supervised learning

Genes

clinical trait (cancer/normal)

samples

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

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

Learn the mapping!

Learn the underlying model!

Why care about clustering?

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

Hierarchical clustering

- Compute all pairwise distances
- Merge closest pair
  - Easy
  - Depends on where to start the grouping
  - Trouble to interpret the “tree” structure

1. Euclidean distance
2. (Pearson’s) correlation coefficient
3. Etc etc...
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.
Clustering expression profiles

Co-regulated genes cluster together

Infer gene function

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

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 $e_1$-$Q$
    - $A$ and $B$ regulate the expression of $C$
      - ($A$ and $B$ are regulators of $C$)
- Bayesian networks

Regulatory network

- Bayesian network representation
  - $X_i$: expression level of gene $i$
  - $Val(X_i)$: continuous

- Interpretation
  - Conditional independence
  - Causal relationships

- Joint distribution
  - $P(X) = \prod P(X_i | Pa(X_i))$
CPD for Discrete Expression Level

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

<table>
<thead>
<tr>
<th></th>
<th>X5=high</th>
<th>X5=low</th>
</tr>
</thead>
<tbody>
<tr>
<td>X3=high, X4=high</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>X3=high, X4=low</td>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>X3=low, X4=high</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>X3=low, X4=low</td>
<td>0.2</td>
<td>0.8</td>
</tr>
</tbody>
</table>

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

Context specificity of gene expression

Context A
Basal expression level
Upstream region of target gene (X5)

Context B
Activator induces expression
Activator (X3) binding site

Context C
Activator + repressor decrease expression
Repressor (X4) binding site
Activator (X3) binding site
Context specificity of gene expression

Context A
Basal expression level

Context B
Activator induces expression

Context C
Activator + repressor decrease expression

Continuous-valued expression I

- Tree conditional probability distributions (CPD)
  - Parameters – mean (μ) & variance (σ²) of the normal distribution in each context
  - Represents combinatorial and context-specific regulation
Continuous-valued expression II

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

  $P(X_5|\text{Par}X_5:w) = N(\Sigma w_i x_i, \epsilon^2)$

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, $\Theta_S$: parameters for $S$, $D$: data
- Likelihood score
  \[
  p(D|S, \Theta_S) \leftarrow \hat{\Theta}_S = \arg\max_{\Theta_S} \ p(D|S, \Theta_S)
  \]
- How to overcome overfitting?
  - Reduce the complexity of the model
    - Bayesian score: $p(S|D) \propto p(D|S)p(S)$
  - Regularization
  - Simplify the structure
    - Module networks

Feature Selection Via Regularization

- Assume linear Gaussian CPD
- MLE: solve $\text{maximize}_w - (\sum w_i x_i - E_{\text{Targets}})^2$

Problem: This objective learns too many regulators
L₁ Regularization

“Select” a subset of regulators
- Combinatorial search?
- minimizeₜₜ \( w \left( \sum w_i x_i - E_{\text{Targets}} \right)^2 + \lambda \sum |w_i| \): convex optimization!
  \( \Rightarrow \) Induces sparsity in the solution \( w \) (Many \( w_i \)'s set to zero)

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

\[ P(E_{\text{Targets}} | x: w) = N(\sum w_i x_i, \sigma^2) \]

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.
The Module Networks Concept

Module 1

\[ X_1 \]

Module 2

\[ X_2 \rightarrow X_3 \rightarrow X_4 \]

Module 3

\[ X_5 \rightarrow X_6 \]

Linear CPDs

\[-3 \times \text{Repressor } X_4\]

\[0.5 \times \text{Activator } X_3\]

Tree CPDs

Activator expression

Repressor expression

Target gene expression

Regulation program

\[ \text{Activator } X_3 \rightarrow \text{true}\]

\[ \text{Repressor } X_4 \rightarrow \text{false}\]

\[ \text{Context A } \rightarrow \text{induced}\]

\[ \text{Context B } \rightarrow \text{repressed}\]

\[ \text{Context C } \]

Genes

Expression of candidate regulators

HAP4

CMK1

Heat Shock?

Context A

Context B

Context C

(\( \mu_A, \sigma_A \))

Experiments

Target gene

Regulation program

\( X_3 \rightarrow X_4 \rightarrow X_5 \rightarrow X_6 \)
Structure Learning – Bayesian Score & Tree CPD

- Find the structure $S$ that maximizes $P(S|D)$
  - $P(\text{Structure}|\text{Data}) \propto P(D|S) P(S)$
  - maximize $\log P(D|S) + \log P(S)$
  - $P(D|S) = \int P(D|S, \Theta) P(\Theta|S) d\Theta$

$P(S)$: prior distribution on the structure

Maximize $\log \int P(D|S, \Theta) P(\Theta|S) d\Theta + \log P(S)$

ML score: $\max_\Theta \log P(D|S, \Theta)$ → More prone to overfitting

Decomposability
- For a certain structure $S$, log $P(D|S)$
  - $\log \int P(D|S, \Theta) P(\Theta|S) d\Theta = \int \log P(X_1|\Theta_{m1}) P(X_2, X_3, X_4|X_1, \Theta_{m2}) P(X_5, X_6|X_3, X_4, \Theta_{m3}) \frac{d\Theta_{m1}}{d\Theta_{m2}} d\Theta_{m3}$
  - $\log \int P(X_2|\Theta_{m2}) P(X_3, X_4|X_1, \Theta_{m2}) P(\Theta_{m3}) d\Theta_{m2}$
  - $\log \int P(X_5, X_6|X_3, X_4, \Theta_{m3}) P(\Theta_{m3}) d\Theta_{m3}$

Module 1 score
Module 2 score
Module 3 score
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)

[Diagram showing gene interactions and regulatory networks]