Inferring Cellular Networks
Using Probabilistic Graphical Models

Jianlin Cheng, PhD
University of Missouri
2010
Bayesian Network Software

Demo
References


Research in molecular biology is undergoing a revolution

- mRNA transcript quantities
- protein-protein
- protein-DNA interactions
- chromatin structure
- Protein quantities
- Protein localization
- Protein modification
Challenge

• Provide methodologies for transforming high-throughput heterogeneous data sets into biological insights about the underlying mechanisms
• Data is noisy
• Data integration
• Generate Hypothesis
Biological Networks – Gene Regulatory Networks

Legend: A transcription factor molecule binds to the DNA at its binding site, and thereby regulates the production of a protein from a gene.
Figure 1: Rich media gene modules network
Signal Transduction Network

- Ras
- Raf-1
- Sos
- Grb2
- MEK
- MAPK
- MKP
- MAPK mRNA
- NUCLEUS
- Gene Expression
- TNFα
- Insulin Receptor
- IRS-1
- IRS-2
- PI3K
- p85
- p110
- P70S6K
- PHAS-1
- Glut4
- Glycogen Synthesis
- Protein Synthesis
- Mitogenesis

PTPases
Translocation
Glucose
Protein Interaction Network
Model-Based Approaches VS Procedure Approaches

• **Procedure**: Binding sites – Gene expression. (a) cluster co-expressed genes to find common sites (b) group genes with similar binding sites and test if they are coexpressed

• **Declarative**: design a model that describes the relations between the two types of data. Learn parameter from data and make predictions
Probabilistic Models

- Stochasticity for measurement noise
- Learning Algorithms
- Select model that fits the actual observations
- Inference
- Make predictions
- Generate insights and hypothesis
Modeling Examples

• Hidden Markov Model for sequence analysis
• Probabilistic Graphical Model for cellular networks
Advantages

• Concise language for describing probability distributions over the observations
• Approaches to learning from data that are derived from basic well-understood principles
• Use of observations to fill in model details
• Provide principles for combining multiple local models into a joint global model
• Declarative nature provides an advantage to extend model to account for additional aspects of the system
Infer Gene Regulatory Network from Gene Expression Data
Model for gene expression and cis-regulatory elements

• **Assumptions 1**: genes can be partitioned into clusters of coexpressed genes, and the genes in each cluster have a typical expression level in each array.

• **Assumption 2**: arrays are partitioned into array clusters, which capture relevant biological context, and that the expression of a gene is roughly the same in the arrays that belong to the same array cluster
Random Variables

• $X_{g,a}$, where $g$ is an index over gene and $a$ is an index over arrays

• $\text{GeneCluster}_g$: denotes the cluster assignment of gene $g$

• $\text{ArrayCluster}_a$ denotes the cluster assignment of array $a$.

• Assumption: the expression of gene $g$ in array $a$ depends on the value of $\text{GeneCluster}_g$ and $\text{ArrayCluster}_a$
Regular Bayesian Networks
Conditional Distribution
Learning Models from Data

• Parameter estimation – maximum likelihood problem (P(data | model))

• Model selection: select among different model structures to find one that best reflects the dependencies in the domain. P(model | data)
• The model just described can achieve high likelihood if the cluster and gene assignment partitions the original measurements into blocks with approximately uniform expression within each block.
• Expectation Maximization procedure that iterates between an **E-step**, which uses current parameters to find the probabilistic cluster assignment of genes and arrays, and an **M-step**, which re-estimates the distribution within each gene/array cluster combination on the basis of this assignment.
Reconstruction of Regulatory Networks

• A key challenge in gene expression analysis is the reconstruction of regulatory networks.
• Distinguish correlation and regulation
• Direct and in-direct regulation
Challenges of Gene Bayesian Network

• Massive number of variables (genes)
• Small number of samples (dozens)
• Sparse networks (only a small number of genes directly affect one another)
• Two crucial aspects: computational complexity and statistical significance of relations in learned models

N. Friedman, 2005
Approach 1: Learning BN from Gene Expression Data

Measured expression level of each gene (discretized) → Random variables Affecting on another

Learn parameters (conditional probabilities) from data
Learn structure (casual relation) from data
Make inference given a learned BN model

N. Friedman, 2005
Gene Bayesian Network

**Qualitative Part:**
Directed acyclic Graph (DAG)
- Nodes – random variables
- Edges – direct (causal) influence

**Quantitative Part**
- Local conditional probability

| E  | B  | P(A|E,B) |
|----|----|----------|
| 0  | 1  | 0.9      |
| 0  | 0  | 0.2      |
| 1  | 0  | 0.9      |
| 0  | 0  | 0.01     |
| 1  | 1  | 0.9      |
| 0  | 0  | 0.01     |
Solutions

• **Sparse candidate algorithm** (by Nir Friedman): Choose a small candidate set for direct influence for each gene. Find optimal BN constrained on candidates. Iteratively improve candidate set.

• **Bootstrap confidence estimate**: use re-sampling to generate perturbations of training data. Use the number of times a relation (or feature) is repeated among networks learned from these datasets to estimate confidence of Bayesian network features.
Data: 76 samples of 250 cell-cycle related genes in yeast genome
Discretized into 3 expression levels. Run 100 bootstrap using sparse learning algorithm.
Compute the confidence of features (relations). Most high confident relations make bio-sense.
Co-Regulation

• A key regulation mechanism involves binding of transcription factors to promoter regions of genes.
• Identify the transcription factor binding sites in the promoter region of genes that can explain observed co-expression.
Module Network Approach

A regulatory module is a set of genes that are regulated in concert by a shared regulation program.

A regulation program specifies the behavior of the genes in the module as a function of the expression level of a small set of regulators.
Regulatory Model

$R_{g,j}$ as depending on the promoter sequence $Seq_g$
Integration of Sequence and Expression Data

• The parameters of this conditional probability characterize the specific motif recognized by the transcription factor. This extension allows us to learn the characterization of the binding site while learning how its presence influences gene expression.
Procedure

• **Inputs**: a gene expression data set and a large precompiled set of candidate regulatory genes for the corresponding organism (independent of data set) containing both known and putative transcription factors and signal transduction molecules

• **Goal**: search for a partition of genes into modules and for a regulation program for each module

• **Output**: a list of modules and associated regulation programs
• **Results:** apply the method to Yeast gene expression data set consisting of 2355 genes and 173 arrays.

• Each inferred modules contained a functionally coherent set of genes (metabolic pathways, oxidative stress, cell cycle-related processes, etc)

• Many module has a match between predicted regulator and its known cis-regulatory binding motif.
One Example
Evaluation of Module Content and Regulation Program

• We evaluate all 50 modules to test whether the proteins encoded by genes in the same module had related functions. We scored the functional/biological coherence of each module according to percentage of its genes covered by annotations. Most of modules had a coherence level above 50%.
<table>
<thead>
<tr>
<th>Module</th>
<th>G C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Respiration and carbon regulation</td>
<td>55</td>
</tr>
<tr>
<td>2. Energy and osmotic stress I</td>
<td>31</td>
</tr>
<tr>
<td>3. Glycolysis and TCA cycle</td>
<td>37</td>
</tr>
<tr>
<td>4. Amino acid metabolism I</td>
<td>39</td>
</tr>
<tr>
<td>5. Amino acid metabolism II</td>
<td>37</td>
</tr>
<tr>
<td>6. Nuclear</td>
<td></td>
</tr>
<tr>
<td>7. Nuclear</td>
<td></td>
</tr>
<tr>
<td>8. Nuclear</td>
<td></td>
</tr>
<tr>
<td>9. Nuclear</td>
<td></td>
</tr>
<tr>
<td>10. Nuclear</td>
<td></td>
</tr>
<tr>
<td>11. Nuclear</td>
<td></td>
</tr>
<tr>
<td>12. Nuclear</td>
<td></td>
</tr>
<tr>
<td>13. Nuclear</td>
<td></td>
</tr>
<tr>
<td>14. Nuclear</td>
<td></td>
</tr>
<tr>
<td>15. Nuclear</td>
<td></td>
</tr>
<tr>
<td>16. Nuclear</td>
<td></td>
</tr>
<tr>
<td>17. Nuclear</td>
<td></td>
</tr>
<tr>
<td>18. Nuclear</td>
<td></td>
</tr>
<tr>
<td>19. Nuclear</td>
<td></td>
</tr>
<tr>
<td>20. Nuclear</td>
<td></td>
</tr>
<tr>
<td>21. Nuclear</td>
<td></td>
</tr>
<tr>
<td>22. Nuclear</td>
<td></td>
</tr>
<tr>
<td>23. Nuclear</td>
<td></td>
</tr>
<tr>
<td>24. Nuclear</td>
<td></td>
</tr>
<tr>
<td>25. Nuclear</td>
<td></td>
</tr>
<tr>
<td>26. Nuclear</td>
<td></td>
</tr>
<tr>
<td>27. Nuclear</td>
<td></td>
</tr>
<tr>
<td>28. Nuclear</td>
<td></td>
</tr>
<tr>
<td>29. Nuclear</td>
<td></td>
</tr>
<tr>
<td>30. Nuclear</td>
<td></td>
</tr>
<tr>
<td>31. Nuclear</td>
<td></td>
</tr>
<tr>
<td>32. Nuclear</td>
<td></td>
</tr>
<tr>
<td>33. Nuclear</td>
<td></td>
</tr>
<tr>
<td>34. Nuclear</td>
<td></td>
</tr>
<tr>
<td>35. Nuclear</td>
<td></td>
</tr>
<tr>
<td>36. Nuclear</td>
<td></td>
</tr>
<tr>
<td>37. Nuclear</td>
<td></td>
</tr>
<tr>
<td>38. Nuclear</td>
<td></td>
</tr>
<tr>
<td>39. Nuclear</td>
<td></td>
</tr>
<tr>
<td>40. Nuclear</td>
<td></td>
</tr>
<tr>
<td>41. Nuclear</td>
<td></td>
</tr>
<tr>
<td>42. Nuclear</td>
<td></td>
</tr>
<tr>
<td>43. Nuclear</td>
<td></td>
</tr>
<tr>
<td>44. Nuclear</td>
<td></td>
</tr>
<tr>
<td>45. Nuclear</td>
<td></td>
</tr>
<tr>
<td>46. Nuclear</td>
<td></td>
</tr>
<tr>
<td>47. Nuclear</td>
<td></td>
</tr>
<tr>
<td>48. Nuclear</td>
<td></td>
</tr>
<tr>
<td>49. Nuclear</td>
<td></td>
</tr>
<tr>
<td>50. Nuclear</td>
<td></td>
</tr>
</tbody>
</table>

**Legend:**
- Enrichment for motif known to participate in regulation by respective regulator
- Partial evidence
- Regulator known to have a role under the predicted condition
- Partial evidence
- Regulator known to regulate module genes or their implied process
- Partial evidence
Candidate regulators

• Compiled a set of 466 candidate regulators annotated in Yeast Genome and Proteome databases

• Use Yeast gene expression data set consisting of 173 microarrays that measure responses to various stress conditions.

• We downloaded these data in log (base 2) ratio to control format from Stanford Microarray Database. Chose a subset of 2355 genes that have a significant change in gene expression under the measured stress conditions
• **Protein annotations:** downloaded Gene Ontology and Munich Information center for Protein Sequence (MIPS) function and KEGG.

• **Regulation program:** Regression tree (decision nodes and leaf nodes); the model semantics is that given a gene \( g \) in the module and an array \( a \) in a context, the probability of observing some expression value for a gene in array is governed by the normal distribution specified for the context.
Learning Module Networks

• In each iteration, the procedure searches for a regulation program for each module and then reassign each gene to the module whose program best predicts its behavior. Repeated until it converges.

• Search for the model with the highest score by using the EM algorithm.
EM Algorithm

- **M-Step**: given a partition of genes into modules and learns the best regulation program (regression tree) for each module. The regulation program is learned through a combinatorial search over the space of trees. The tree is grown from the root to its leaves. At any given node, the query that best partitions the gene expression into two distinct distribution is chosen.
• **E-step**: given the inferred regulation programs, we determine the module whose associated regulation program best predicts each gene’s behavior. Select the module whose program gives the gene’s expression profile the highest probability and re-assign the gene to this module.

• We initialize our modules to 50 clusters using Pcluster, a hierarchical agglomerative clustering. We then applied the EM algorithm to this starting point, refining both the gene partition and the regulatory program.
Evaluating statistical significance of modules

- All of the statistical evaluations were done and visualized in GeneXPress. The tool can evaluate the output of any clustering program for enrichment of gene annotations and motifs
Annotation enrichment

• We associated each gene with the processes in which it participates. Resulted in 923 GO categories, 208 MIPS categories, and 87 KEGG pathways. For each module and for each annotation, we calculated the fraction of genes in the module associated with that annotation and used the hypergeometric distribution to calculate a P-value for this fraction.
Promoter Analysis

• We search for motifs (represented as Position-Specific Scoring Matrices) within 500 bp upstream of each gene. We downloaded TRANSFAC, containing 34 known function cis-regulatory motifs. We also use a motif finder to find 50 potentially novel motifs.
Motif Combination

• We searched for statistically significant occurrences of motif pairs. We constructed a motif pair attribute, which assigns a “true” value for each gene if and only if both motifs of the pair are found in the upstream region of that gene. For each module and for each motif pair attribute, we calculated the fraction of genes in the module associated with that attribute and used the hypergeometric distribution to calculate a P value for this fraction.
Regulator Annotations

• We associate regulators with annotations and binding sites in the same way we associate with these attributes to the modules. Because a regulator may regulate more than one module, its targets consist of the union of the genes in all modules predicted to be regulated by that regulator. We tested the targets of each regulator for enrichment of the same motifs and gene annotations as above using the hypergeometric P value.