Introduction
- Previous lectures assume networks to be analyzed are understood well enough to develop a detailed model.
- This lecture assumes that we have experimental data, and we wish to develop a model for this network.
- DNA microarrays can measure expression levels of thousands of mRNA targets simultaneously.
- Today 100s of samples may be run in a microarray experiment.
- Given this data and an abstract class of potential models for various network configurations, how can we decide the most likely network configuration that generated this data?
- Even largest experiments do not provide enough samples for high statistical significance.
- Current technology has high noise to signal ratio.

Methods
- Many analysis tools use clustering algorithms that group genes together that are expressed at the same time.
- Does not indicate which genes activate or repress other genes.
- New approaches needed to reveal structure of the transcriptional regulation system.
- Bayesian networks promising for analyzing gene expression.

Clustering
- Genes that are active at the same time are likely to be involved in the same regulatory process.
- Clustering assumes genes are grouped and within a group the genes produce the same expression profile.
- Due to noise and other uncertainties, groupings no longer clear.
- Goal: determine the original groupings of the genes.
- Assumes there exists a method to determine the pairwise distance between the expression profiles of any two genes.
- Many methods have been proposed for clustering.
**K-Means**
- Partitions $N$ genes into $K$ clusters.
- Begins with $K$ initial clusters of the genes either determined by the user or by random.
- For each cluster, computes its centroid (i.e., the average expression profile of the genes in a cluster).
- Reassigns each gene to cluster with centroid that is closest to the expression pattern of the gene.
- Centroids recalculated and process repeats until no change.

**Agglomerative Hierarchical Clustering**
- Begins with $N$ clusters each containing a single gene.
- Combines two clusters with smallest distance apart where distance is between their average expression profiles.
- Continues for $N - 1$ steps at which point all the genes are merged into a hierarchical tree.

**Algorithm Comparison**

**Bayesian Networks**
- Given expression data, $D$, learning techniques allow one to induce the network connectivity that best matches $D$.
- *Bayesian networks* are promising tool to learn connectivity.
- A *Bayesian network* represents a joint probability distribution.
- Represented with directed acyclic graph $G$ whose vertices correspond to random variables $X_1, \ldots, X_n$.
- Connections represent dependencies between random variables.
Dependence

- $P(X,Y)$ is joint distribution over two variables $X$ and $Y$.
- $X$ and $Y$ are independent if $P(X,Y) = P(X)P(Y)$ for all values of $X$ and $Y$ (equivalently, $P(X|Y) = P(X)$).
- When $X$ and $Y$ are dependent learning value of $Y$ gives us information about $X$.
- Correlation is sufficient but not necessary condition for dependence.
- When $X$ and $Y$ are dependent, this is represented in the Bayesian network by an arc between them.
- If arc directed from $X$ to $Y$, $X$ is a parent of $Y$.

Markov Assumption

- Associated with each variable $X_i$ is a conditional distribution given its parents in $G$.
- Graph $G$ encodes Markov Assumption, each variable $X_i$ is independent of its non-descendents given its parents in $G$.
- This is known as Conditional independence.

\[ \forall i, I(X_i; \text{NonDescendents}(X_i) \mid \text{Pa}(X_i)) \]

- Using Markov assumption, joint pdf can be decomposed:

\[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Pa}(X_i)) \]

Example Bayesian Networks

![Example Bayesian Network Diagram]

Simple Bayesian Network

- $P(A|B,C) = P(A|B)$
- We say $A$ and $C$ are conditionally independent
- $I(A;C|B)$
- $P(A, B, C) = P(A)P(B|A)P(C|B)$.
Another Bayesian Network

\[ I(A; E), I(B; D | A, E), I(C; A, D, E | B), \]
\[ I(D; B, C, E | A), I(E; A, D) \]
\[ P(A, B, C, D, E) = \]
\[ P(A)P(B | A, E)P(C | B)P(D | A)P(E) \]

A is common cause of B and D
If A not measured, hidden common cause

Learning Bayesian Networks
- Given training set \( D = \{x^1, \ldots, x^N\} \) of independent instances of \( X \), find a network \( B = (G, \Theta) \) that best matches \( D \).
- Evaluate using Bayesian scoring metric:
  \[ \text{Score}(G : D) = \log P(G | D) = \log P(D | G) + \log P(G) + C \]
where \( C \) is constant and \( P(D | G) = \int P(D | G, \Theta)P(\Theta | G)d\Theta \) is the marginal likelihood.
- Choice of priors \( P(G) \) and \( P(\Theta | G) \) determines score.
- Given priors and data, learning amounts to finding structure \( G \) that maximizes the score.

Learning Bayesian Networks (cont)
- NP-hard so use heuristics like greedy random search.
- For example, beginning with some initial network, a greedy random search would select an edge to add, delete, or reverse.
- It would then compute this networks score, and if it is better than the previous network, then it would keep the change.
- This process is repeated until no improvement is found for some number of steps.
Efficient Learning Algorithms
- Number of graphs is super-exponential in number of variables.
- Sparse candidate algorithm identifies small number of candidate parents for each gene based on local statistics.
- Pitfall is early choices can overly restrict the search space.
- Adapting the candidate sets during the search can help.

Applying Bayesian Networks to Expression Data
- By learning Bayesian network, can answer questions like which genes depend on which other genes.
- Difficulty is data is for thousands of genes but often only a few dozen samples, but on positive side, networks typically sparse.
- A set of plausible networks needs to be considered.
- May attempt to characterize features common in the set.
- Markov relations: Is $Y$ in the Markov blanket of $X$?
- Order relations: Is $X$ an ancestor of $Y$? (or cause?)

Discretization
- Expression level of each gene modeled as a random variable.
- Need to define local probability model for each variable.
- Discretize gene expression into 3 categories: significantly lower (-1), similar to (0), or significantly greater (+1) than control.
- Discretizing can lose information, but more levels can be used if more resolution in experimental data.
- Control expression level either determined experimentally or the average expression level can be used.
- Meaning of “significantly” defined by setting threshold to ratio between measured expression and control.

Estimating Statistical Confidence in Features
- Confidence is likelihood that a feature is actually true.
- Want to compute $P(G_i|D)$ but too many possible networks.
- Instead can use bootstrap method:
  - For $i = 1 \ldots m$
    - Re-sample with replacement, $N$ instances from $D$, denote $D_i$.
    - Apply learning procedure on $D_i$ to find network $G_i$.
  - For each feature $f$ calculate $\text{confidence}(f) = \frac{1}{m} \sum_{i=1}^{m} f(G_i)$,
    where $f(G)$ is 1 if $f$ is a feature of $G$, 0 otherwise.
Discussion

- Clustering approaches can only find correlations.
- Bayesian analysis can potentially discover causal relationships and interactions between genes.
- Probabilistic semantics good for noisy biological systems.
- Focus on extracting features rather than find single model.
- Theory can assist with experimental design.
- Use Dynamic Bayesian Networks for temporal expression data.

Time Series Data

- Series of measurements of gene activity over a time period.
- Can be performed on both wild-type and mutated cells.
- Time series and mutational data essential to determining which genes activate or repress other genes.

Learning Causal Patterns

- A Bayesian network represents correlative relationships, but ultimately interested in knowing causal relationships.
- In a causal network, parents of a variable are interpreted as its immediate causes.
- Causal Markov assumption: given values of variable’s immediate causes, it is independent of earlier causes.
- Causal networks model not only distribution of observations but also effects of interventions.
- In causal networks, $X \rightarrow Y$ and $Y \rightarrow X$ are not equivalent.
- The true causal network can be any in the equivalence class of Bayesian networks.

Learning Causal Networks

- Bayesian analysis of time series data can determine causal networks.
- Method presented here uses Bayesian scoring metric a gene’s parents, but does not search for best fit network.
- Not only the context of when a gene is high or low gives information, but also the context in which it rises.
- Looks for percentage of times a child gene rises in a given configuration to determine activation and repression behavior.
- Also uses child’s own level in the calculations.
- Result is directed graph representation of the genetic circuit.
Causal Network Inference Method

- Time Series Data $E$
- Initial Network $N_1$
- Activation and Repression Thresholds $T$
- Encodings $L$
- Optional

Genetic Network Graph Model

- $N = (S,C)$ where $S$ is species in the experimental data.
- $C \subseteq S \times I \times S$ denote edges or connections between species.
- $I \subseteq \{a,r\}$ where $a$ represents activation and $r$ repression.
- Activation edges shown with $\rightarrow$ and repression edges with $\leftarrow$.

Genetic Network Graph Model Example

- Possible Four-Gene Networks
- Possible Ten-Gene Networks

Time Series Data Model

- $D = (S,E)$ where $S$ is the set of species being monitored, and $E$ is the set of experiments.
- Each $\Delta \in E$ is of form $(W,M,\delta)$, where $W \in \mathbb{R}$ is weight of experiment, $M$ are mutated species, and $\delta$ are the data points.
- Should include both wild-type and mutational experiments.
- Weights, $W$, used to determine the relative importance of each experiment (sum should be 1).
- A data point $\alpha \in \delta$ is of form $(\tau,\nu)$, $\tau \in \mathbb{R}$ is a time point and $\nu \in \mathbb{R}^{|S|}$ is a vector over $S$, which is a value for each species.
Synthetic Time Series Data Example

Experiment 1

<table>
<thead>
<tr>
<th>Time</th>
<th>Lac</th>
<th>CI</th>
<th>Tet</th>
<th>GFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>8</td>
<td>15</td>
<td>17</td>
<td>11</td>
</tr>
</tbody>
</table>

The GenNet Algorithm

1: void GenNet(Species S, Expt E, Net C, Thresh T, Enc L)
2: \((E, L) := \text{DiscretizeSpace}(S, E, L)\)
3: for all \(s \in S\) do
4: \(P := \text{GetInitialParents}(s, S, E, C, T, L)\)
5: \(Parents_s := \text{CheckMultipleParents}(s, P, S, E, T, L)\)
6: end for

Selecting Initial Parents

1: void GetInitialParents(s, S, E, C, T, L)
2: \(E := E - \{\text{experiments in which } s \text{ is mutated}\}\)
3: Parents := 0
4: for all \(p \in S\) do
5: if \((p, a, s) \in C \text{ and } (p, r, s) \in C\) then
6: \(votes_a, votes_r, votes_u := 0\)
7: for all level \(l \in L\) where \(l_p \in L\) do
8: \(prob_0 = P(s \mid x_p = 0, v_p = \text{level})\)
9: for all \(l \in (l_p - \{0\}) \text{ where } l_p \in L\) do
10: \(prob_0 = P(s \mid x_p = l, v_p = \text{level}), \text{diff} = \frac{prob_1}{prob_0}\)
11: if \(\text{diff} > T_s\) then \(votes_a + +\)
12: else if \(\text{diff} < T_r\) then \(votes_r + +\)
13: else \(votes_a + +\)
14: end if
15: if \(votes_a > votes_r + votes_u\) then Parents := Parents \cup \{p\}\)
16: end for
17: end if
18: end for
19: return Parents
### Slide 33

**Example: Selecting Initial Parents**
- The following tables show the probability of GFP rising given potential parent levels.
- Repression and activation bounds are set at 0.9 and 1.05.

<table>
<thead>
<tr>
<th>Lac</th>
<th>Differences</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>L0</td>
<td>1 / 0</td>
<td>Activation: 3</td>
</tr>
<tr>
<td>L1</td>
<td>2 / 0</td>
<td>Repression: 0</td>
</tr>
<tr>
<td>L2</td>
<td></td>
<td>Not Related: 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CI</th>
<th>Differences</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>L0</td>
<td>1 / 0</td>
<td>Activation: 3</td>
</tr>
<tr>
<td>L1</td>
<td>2 / 0</td>
<td>Repression: 6</td>
</tr>
<tr>
<td>L2</td>
<td></td>
<td>Not Related: 3</td>
</tr>
</tbody>
</table>

### Slide 34

**Removing Unrelated Parents**

1. CheckMultipleParents(x, P, E, F, L)
2. E := E − {experiments in which x is mutated}
3. Q := BreakIntoSmallerSubsets(P)
4. Parents := Ø
5. for all y ∈ Q do
6. for all p ∈ y do
7. votez, votes := 0
8. for all pc where pc assigns a level to each species in P − (p) do
9. for all level ∈ l, where l ∈ P do
10. prob := P(x | v_p = 0, v_z = level, v_p−y = pc)
11. for all pl ∈ (1 − (0)) where l_p ∈ L do
12. prob := P(x | v_p = pl, v_z = level, v_p−y = pc)
13. diff := prob / prob
14. if diff > T then votez = ++
15. else if diff < T then votes = ++
16. else votez = ++
17. end for
18. end for
19. end for
20. if votez > votes, then Parents := Parents ∪ {p}
21. else if votes > votes + votez, then Parents := Parents ∪ {p}
22. end for
23. end for
24. return Parents

### Slide 35

**Examples: Removing Unrelated Parents**
- Repression and activation bounds are set at 0.75 and 1.15.

| GFP Level 0 | 0.8 0.4 0.1 0.8 0.5 0.8 1.2 2.3 2.3 2.2 |
| GPP Level 1 | 2 3 3 4 0 3 1 2 3 3 3 6 1 1 5 |
| GPP Level 2 | 25 26 18 17 22 22 12 23 23 44 |

### Slide 36

**Evaluation Procedure**
Results: 4 Gene Networks

- Inspired by synthetic networks created by Guet et al.
- Genes activate and repress each other in very specific ways.
- Encoding levels for each species broken into 3 distinct states based on time spent in each level during the experiments.
- For example, the species Tet is typically below 20 molecules, so 3 levels chosen are 0 to 8, 8 to 12, and above 12.

Results: 4 Gene Networks (cont)

- There are 18 four-gene networks under Guet et al.’s framework.
- With 20 experiments per network, all 72 connections are correctly identified.
- With 10 experiments per network, 67 or 93 percent are correctly identified.
- Several possible sources of network errors:
  - Not enough data to fully discover all connections.
  - Stochasticity of data may introduce too much noise.
  - Transitive edges are formed when multiple genes are connected by a path.

Results: 10 Gene Networks

- Also applied to 10 high level 10 gene randomized networks that follow Guet et al. framework.
- Each gene has one promoter from the set of genes.

<table>
<thead>
<tr>
<th>Number of Experiments</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23/90</td>
<td>23/100</td>
</tr>
<tr>
<td>5</td>
<td>52/102</td>
<td>52/100</td>
</tr>
<tr>
<td>10</td>
<td>71/102</td>
<td>71/100</td>
</tr>
<tr>
<td>25</td>
<td>89/103</td>
<td>89/100</td>
</tr>
<tr>
<td>50</td>
<td>98/101</td>
<td>98/100</td>
</tr>
</tbody>
</table>
Sources

- Friedman
- Sachs, Science 2005
- Barker and Myers