CENG 734
Advanced Topics in Bioinformatics

Week 11
Reconstruction of Signaling Pathways

Fall 2010-2011
1. Describe how Team 42 converted article level protein-protein interaction information in the training data to sentence level information.
Pathway Reconstruction

• The problem:

<table>
<thead>
<tr>
<th>Knockdown</th>
<th>Effect on Gene 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 2</td>
<td>activated</td>
</tr>
<tr>
<td>Gene 3</td>
<td>activated</td>
</tr>
<tr>
<td>Gene 4</td>
<td>not activated</td>
</tr>
<tr>
<td>Genes 2 &amp; 3</td>
<td>not activated</td>
</tr>
</tbody>
</table>
Fig. 1. Average number of minimal acyclic graphs which generate the same pattern of observations at a single downstream node after knockdowns, plotted over network size.
Assumptions for Identifiability

- Feedback loops are not allowed, i.e., only acyclic directed graphs are considered.
- The network has to contain a connection between the stimulated node and the readout node.
- There is no direct connection from the stimulated node to the readout node.
- Every node in the network except the stimulated node and the readout node has to have indegree and outdegree at least 1.
Overview of the Methods

• A Bayesian network with Boolean variables is used for pathway inference.
• The probability of a node being active is described by a sigmoid function of the weighted sum of the incoming regulations.
• Mode hopping Markov chain Monte Carlo is used to evaluate the posterior distribution over model parameters given the knockdown data.
• The likelihood is approximated using stochastic simulation.
Statistical modeling

- Each protein is described by a Boolean variable, $x_i \in \{0, 1\}$

\[
p\{x_i(t) = 1|x(t-1)\} = \frac{1}{1 + \exp\left(-\gamma \left(w_i^0 + \sum_{j=1}^{n} w_{j,i}x_j\right)\right)}
\]
The Current State of the Network

• For a network of $n$ nodes, the current state can be described by a binary number with $n$ digits.
• There are $2^n$ possible different states.
• If the weights on the edges are known, we can compute transition probabilities between states.
  – $M$: a matrix of size $2^n \times 2^n$. 
Transition Probabilities

\[ M_{i,j} = p \left\{ x(t) = \eta^{(j)} | x(t-1) = \eta^{(i)} \right\} \]

\[ = \prod_{k=1}^{n} p \left\{ x_k(t) = \text{active}(k, \eta^{(j)}) | x(t-1) = \eta^{(i)} \right\} \]
Probability of being in a state at time $T$

$$p(\eta(T)|M,p_0) = p'_0M^T \delta(\eta(T))$$
Effect of knock-downs

- A single knock-down of a node $k$ is modeled by removing the states from the M matrix in which $k$ is active.
Network Construction

• The likelihood of observed data $D$:

$$p\left\{ D|w, w^0, T \right\} = \prod_{k=1}^{K} p\left( \eta^{(k)}(T)|M^{-k}, p_0 \right)$$
Bayesian Estimation

\[ p\left\{ w, w^0, T | \mathcal{D} \right\} = \frac{p\left\{ \mathcal{D} | w, w^0, T \right\} \pi(w, w^0, T)}{p(\mathcal{D})} \]
Prior probabilities

• Are assumed to be independent
• And drawn from a distribution which simulates sparsity of the networks.
Incomplete Observations

\[ \mathcal{D}' = \left\{ x^{(k)}_{\Psi}(T) \right\}_{k=1}^K \]

\[ p\left( \mathcal{D}' | w, w^0, T \right) = \prod_{k=1}^{K} \sum_{\eta \in \mathcal{B}_k} p\left( \eta(T) | M^{-k}, p_0 \right) \]
Estimating the posterior

- By sampling from the posterior using Metropolis-Hastings

A1 Initialize $\theta(0) = (w(0), w^0(0), T(0))$, $t = 0$.
A2 Sample $\tilde{\theta}$ from a proposal distribution $q(\cdot | \theta(t))$.
A3 Compute $\alpha = \min \left(1, \frac{p\{D' | \tilde{\theta}\} \pi(\tilde{\theta})q(\theta(t)|\tilde{\theta})}{p\{D' | \theta(t)\} \pi(\theta(t))q(\tilde{\theta}|\theta(t))} \right)$.
A4 Accept $\theta(t + 1) = \tilde{\theta}$ with probability $\alpha$, otherwise stay at the old point $\theta(t + 1) = \theta(t)$.
A5 Increase $t$ and go to A2, until enough points sampled
Approximating the Likelihood

\[ p\{\mathcal{D}|\theta\} \approx \frac{1}{N} \sum_{m=1}^{N} I(S_m = \mathcal{D}) \]
Approximating the Likelihood

B1  Initialize $\theta(0) = (w(0), w^0(0), T(0)), \ t = 0$.

B2  Sample $\tilde{\theta}$ from a proposal distribution $q(\cdot | \theta(t))$.

B3  Simulate a dataset $S$ using the stochastic model described by Equations (1) and (2), with parameters $\tilde{\theta}$.

B4  If $S \neq S'$ (based only on the observed nodes in $S'$), let $\theta(t+1) = \theta(t)$, increase $t$ and go to B2.

B5  Compute $\gamma = \min\left(1, \frac{\pi(\tilde{\theta})q(\theta(t)|\tilde{\theta})}{\pi(\theta(t))q(\tilde{\theta} | \theta(t))}\right)$.

B6  Accept $\theta(t+1) = \tilde{\theta}$ with probability $\gamma$, otherwise stay at the old point $\theta(t+1) = \theta(t)$.

B7  Increase $t$ and go to B2, until enough points are sampled.
Mode Hopping Monte Carlo

- They estimate the modes of the distribution by grid search and the distributions are sampled taking the modes into account.
Results

• Simulated results on a 5 node network
  – 126 minutes for the exact computation of the likelihood
  – 38 minutes for approximation
  – Identified two distinct topologies

• Real results on the JAK/STAT pathway
  – 18 hours
  – Identified the real network and also provided some hypotheses
Discussion

• Alternative modeling ideas?
• Speeding up?
• Anything else?
Next Week

• 2 papers to read about PPI network alignment: