Inferring Protein-Signaling Networks

Lectures 14 – Nov 14, 2011
CSE 527 Computational Biology, Fall 2011
Instructor: Su-In Lee
TA: Christopher Miles
Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

Course Announcement

Outline

- Regulatory motif finding
  - More computational methods
    - Greedy search method (CONSENSUS)
    - Phylogenetic foot-printing method
    - Graph-based methods (MotifCut)
  - Before/after motif finding

- Inferring signaling networks

Drawbacks of Existing Methods

Independence assumption: biologically unrealistic
Perfectly conserved nucleotide dependency — ATG and CAT

Resulting PSSM

WRONG!
Overview: Graph-Based Representation

- Nodes: k-mers of input sequence
- Edges: pairwise k-mer similarity
- Motif search $\rightarrow$ maximum density subgraph

MotifCut Algorithm

- Convert sequence into a collection of k-mers
  - Each overlap/duplicate considered distinct

$k=3$
**Motif Graph Representation**

- Nodes are k-mers
- Edge weights are distances between k-mers
  - How the edge weights are determined? (later)

- Same k-mer node can appear multiple times.
  - If a certain k-mer appears frequently in the input sequences, there are many nodes for that k-mer.

- Finding over-represented similar k-mers → Finding maximum density subgraph (MDS)

**Motif Finding**

- Find highest density subgraph
  - Density is defined as sum of edge weights per node: graph density $\lambda = |E|/|V|$.
  - Find the maximum density subgraph (MDS)
Motif Dependency in MDS

MotifCut Algorithm

- Read input sequences
- Generate graph as previously described
  - K-mers are generated by shifting one base pair
  - Each k-mer in the sequence gets a node, including identical k-mers
  - Graph contains as many nodes as there are base pairs
  - Connect edges with weights based on distances between nodes
- Find maximum density subgraphs (MDSs)
Edge Weights

- **Semantics:** Edge weight is the likelihood of two k-mers to be in the same motif.

Use Hamming distance as a way to quantify distance between k-mers.

- Let’s make this a bit more precise:
  - For every pair of vertices \((v_i, v_j)\) create an edge with weight \(w_{ij}\)
  - \(w_{ij} = f(\text{Hamming distance between k-mers in } v_i, v_j)\)

\[
W_{ij} = \frac{\Pr(v_i \in M | v_j \in M) + \Pr(v_j \in M | v_i \in M)}{\theta(\Pr(v_i \in B)) + \theta(\Pr(v_j \in B))}
\]

- But how to compute \(\Pr(v_i \in M | v_j \in M)\)?

- Simulate it!
  - Way too many variables to account for analytically:
    - Background model, kmer length, hamming distance, etc...
Maximum Density Subgraph

- Standard graph theory method
  - Max-flow / min-cut: simple and easy to implement
  - However, its running time is $O(nm \log(n^2m))$, where $n$ is the number of vertices and $m$ is the number of edges

- Need faster method

- Developed heuristic approach that utilizes max-flow / min-cut method with modifications

MotifCut Algorithm

- Find the maximum density subgraph (MDS)
- MDS optimization

  - Remove all edges below a certain threshold
  - Pick one vertex (do this for every vertex)
  - Put back all neighboring edges for that vertex
  - Use standard algorithm to calculate densest subgraph
  - Repeat for every vertex
Synthetic Experiment Results

Yeast Test Results

- Gold standard data (Harbinson et al., 2004)
Outline

- Regulatory motif finding
  - More computational methods
    - Greedy search method (CONSENSUS)
    - Phylogenetic foot-printing method
    - Graph-based methods (MotifCut)
  - Before/after motif finding

- Inferring signaling networks

What After Motif Finding?

- Experiments to confirm results
- DNaseI footprinting & gel-shift assays
- Tells us which subsequences are the binding sites
Before Motif Finding

- How do we obtain a set of sequences on which to run motif finding?
- In other words, how do we get genes that we believe are regulated by the same transcription factor?
- Two high-throughput experimental methods: ChIP-chip and microarray.

19

- ChIP-chip
  - Take a particular transcription factor TF
  - Take hundreds or thousands of promoter sequences
  - Measure how strongly TF binds to each of the promoter sequences
  - Collect the set to which TF binds strongly, do motif finding on these

- Gene expression data
  - Collect set of genes with similar expression (activity) profiles and do motif finding on these.
Outline

- Regulatory motif finding
  - More computational methods
  - Before/after motif finding

- Inferring signaling networks
  - Signaling network
  - Flow cytometry
  - Bayesian networks

Gene Regulation

- Transcriptional regulation is one of many regulatory mechanisms in the cell

Focus of today’s lecture
Post-translational Modification

- Most proteins undergo some form of modification following translation.
- **Phosphorylation** is the most studied and best understood post-translation modification.
  - Addition of a phosphate (PO$_4^-$) group to a protein
  - It activates or deactivates many protein enzymes

  ![Phosphorylation Diagram]

- **Interventions** – artificially introducing chemicals which activate/repress the phosphorylation of a protein.

Cellular Signaling Networks

- **Cellular signaling**
  - *Part of a complex system of communication* that governs basic cellular activities and coordinates cell actions.
  - The ability of cells to perceive and correctly respond to their microenvironment is the basis of development, tissue repair, and immunity as well as normal tissue homeostasis.

![Cellular Signaling Network Diagram]

*Overview of signal transduction pathways*

Cellular Signaling Networks

- Reversible phosphorylation is a major regulatory mechanism controlling the signaling pathway.
  - Many signaling pathways, including the insulin/IGF-1 signaling pathway, transduce signals from the cell surface to downstream targets via tyrosine kinases and phosphatases.
- Elucidating complex signaling pathway phosphorylation events can be difficult.

Signaling Networks – Example

- Classic signaling network and points of intervention
- Human T cell (white blood cell)

Flow Cytometry

- Quantitatively measure as given proteins’ expression levels and their phosphorylation states.

Flow Laser

Cell suspension

Flow

Detector

Laser

Because each cell is treated as an independent observation, flow cytometric data provide a statistically large sample.

Relative protein levels per cell

Protein A Protein B Protein C

Protein A Protein B Protein C

Flow Cytometry Data

Signal Networks:
Flow Cytometry Data

Cells

Proteins

Intervention conditions

Phosphorylated protein levels

Gene expression levels

Repressor binding site

Activator binding site

Activator or Inhibitor

Gene expression levels

Regulatory Networks:
Gene Expression Data

Experiments

Genes

Module

RNA
Bayesian Networks

- Directionality via intervention

- Structure preservation

Bayesian Networks

- Directed Acyclic Graphs (DAGs)

Conditional independence

\[ P(\text{B} | \text{D}, \text{A}, \text{E}) = P(\text{B} | \text{A}, \text{E}) \]

Parents of B

\( (\text{B} \perp \text{D} | \text{A}, \text{E}) \)

Independent
Bayesian Networks

- Signal network (protein regulation)
  - Discrete phosphorylated protein levels

- Regulatory networks (gene regulation)
  - Continuous gene expression levels

Structure Learning

Signal Networks: Flow Cytometry Data

Learn DAG structure
**Overview**

Conditions (multi well format)  
Multiparameter Flow Cytometry

Datasets of cells  
- condition 'a'
- condition 'b'
- condition...‘n’

Bayesian Network Analysis

Influence diagram of measured variables


---

**Local Probability Model**

- **Conditional Probability Tables**

\[
D = \text{Data} \quad G = \text{Graph} \\
\theta = \text{CPT values for each node } X \\
\theta_{jk} = P( X_i = k | \text{Parents}(X_i) = j ) \\
N_{jk} = \# \text{ times } X_i = k \text{ and } \text{Parents}(X_i) = j \text{ in the Data}
\]

Conditional Probability Table (CPT)

| A  | B  | P(C=0|Pa) | P(C=1|Pa) |
|----|----|---------|---------|
| 0  | 0  | \theta_{00} | \theta_{01} |
| 0  | 1  | \theta_{00} | \theta_{01} |
| 1  | 0  | \theta_{10} | \theta_{11} |
| 1  | 1  | \theta_{10} | \theta_{11} |

\[
\theta_{jk} = \sum_{i} N_{ijk} \theta_{ijk}
\]
Maximum Likelihood Score

- Find \( G \) that maximizes:
  \[ P( \text{Data} = D \mid \text{Graph} = G, \Theta_{\text{MLE}}) \]

  \( K = \# \text{discrete levels of } X \)
  \( N_{ijk} = \# \text{ times } X_i = k \text{ and Parents}(X_i) = j \) in the Data

  \( \Theta_{ijk} = P( X_i = k \mid \text{Parents}(X_i) = j ) \)
  \( \Theta_{ijk}^{\text{ML}} = \frac{N_{ijk}}{\sum_k N_{ijk}} \)

Structure Score

- Bayesian score (Structure \mid Data)
  \[ = \log P(\text{Data} \mid \text{Structure}) + \log P(\text{Structure}) \]

- Decomposability
  \[ \log P(\text{Data} \mid \text{Structure}) = \sum_X \text{FamScore}(X, \text{Parents}(X) \mid \text{Data}) \]
Structure Score

\[ P(\text{Data}=D \mid \text{Graph}=G) = \int P(D \mid G, \theta) P(\theta \mid G) \, d\theta \]

Dirichlet prior \( \sim \text{Dir}(\alpha) \)

\( \theta_{ijk} = P(\text{Xi}=k \mid \text{Parents(Xi)}=j) \)

\( N_{ijk} = \# \text{times Xi}=k \text{ and Parents(Xi)}=j \text{ in the Data} \)


Structure Score

\[ P(D|G) = \prod_{i=1}^{\text{#proteins}} \prod_{j=1}^{\text{#parent states}} \frac{\Gamma(\Sigma_{k=1}^{K} \alpha_{ijk})}{\Gamma(\Sigma_{k=1}^{K} \alpha_{ijk} + N_{ijk})} \frac{\prod_{k=1}^{K} \Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \]

FamScore(\(X_i, Pa_j|D\)) = \log \prod_{j=1}^{\text{#parent states}} \frac{\Gamma(\Sigma_{k=1}^{K} \alpha_{ijk})}{\Gamma(\Sigma_{k=1}^{K} \alpha_{ijk} + N_{ijk})} \frac{\prod_{k=1}^{K} \Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}

Score(\(G|D\)) = \sum_{i=1}^{\text{#proteins}} \text{FamScore}(X_i, Pa_j|D)

Score(\(G|D\)) = \log P(D|G) + \log P(G)

Bayesian Score

- \(P(\text{Data}=D | \text{Graph}=G)\)
  - \(= \int P(D|G, \theta) P(\theta|G) \ d\theta\)
  - Multinomial
  - Dirichlet prior \(\sim \text{Dir}(\alpha)\)

- \(\Theta_{ijk}^{BS} = \frac{(N_{ijk} + \alpha_{ijk})}{\sum_k (N_{ijk} + \alpha_{ijk})}\)

"Imaginary" counts