Interaction Network Analysis

Young-Rae Cho
Associate Professor
Department of Computer Science
Baylor University

Biological Networks

➢ Definition
  • Maps of biochemical reactions, interactions, regulations between genes or proteins

➢ Importance
  • Show the mechanisms of molecular function in a cell
  • Key resource for functional characterization of molecules in a systematic view

➢ Examples
  • Metabolic networks
  • Protein-protein (or genetic) interaction networks
  • Gene regulatory networks
  • Signal transduction networks
Overview

- **Backgrounds**
- Network Modeling
- Interaction / Function Prediction
- Protein Complex / Functional Module Detection
- Signaling Pathway / Functional Pathway Prediction
- Essential Protein Selection
- Conserved System Identification

Protein-Protein Interactions (1)

- **Protein Structures**
  - Primary
  - Secondary
    - β-strand
    - α-helix
  - Tertiary
  - Quaternary

protein-protein interactions

monomer, dimer, trimer, ..., l-mer
Protein-Protein Interactions (2)

- **Protein-Protein Interactions**
  - Physical interactions to form protein complexes
  - Functional relationships to perform the same molecular functions

- **Interactome**
  - The entire set of protein-protein interactions in the genome scale
  - Problem: large scale
  - Requires systematic, computational analysis

Protein Interaction Networks

- **Interaction Networks**
  - Undirected, unweighted graph representation with a set of nodes $V$ as proteins and a set of edges $E$ as interactions between them
  - Problem: complex connectivity

- **Example**
Interaction Determination (1)

- **Types of Interactions**
  - Permanent (stable) interactions
  - Transient interactions

- **Experimental Methods**
  - Mass spectrometry
    - Identification of components in protein complexes
  - Two-hybrid system
    - Determination of binary protein-protein interactions
    - Problem: a large number of false positives
  - Protein microarray

Interaction Determination (2)

- **Computational Methods**
  - Sequence-based approaches
    - Phylogenetic profile analysis
    - Gene fusion analysis
    - Gene neighborhood analysis
  - Structure-based approaches
    - Homolog search
    - Interface similarity search
  - Expression-based approaches
    - Expression profile correlation
  - Interaction-based approaches
    - Interaction prediction from known interactions
Overview

- Backgrounds
- Network Modeling
- Interaction / Function Prediction
- Protein Complex / Functional Module Detection
- Signaling Pathway / Functional Pathway Prediction
- Essential Protein Selection
- Conserved System Identification

Random Graph Model (1)

- Erdős and Rényi, 1960 (E-R Model)
  - Random graph as N nodes connected by m edges that are randomly chosen from N(N-1)/2 possible edges
  - m = p(N(N-1)/2) where p is the probability of each pair of nodes being connected
  - Degree distribution P(k) = \( \binom{N-1}{k} p^k (1-p)^{N-1-k} \)
    - Degree of v: the number of links from v to other nodes
    - Degree distribution P(k): probability that a node has k links
  - Expected number of nodes with degree k
    \[
    E(X_k) = N P(k) = N \binom{N-1}{k} p^k (1-p)^{N-1-k} = \lambda_k \rightarrow P(X_k = r) = e^{-\lambda_k} \frac{\lambda_k^r}{r!}
    \]
    (Poisson distribution)
Example of E-R Model
- Poisson distribution with \( N = 1000 \) and \( p = 0.0015 \)

Random Graph Model (2)

Barabasi and Albert, 1999 (B-A Model)
- Focused on network dynamics based on these two steps:
  - Growth: networks are continuously expanded by the addition of new nodes with a link to the nodes already present
  - Preferential attachment: the nodes are likely to be linked to high-degree nodes
- Power law degree distribution: \( P(k) \sim k^{-\gamma} \) where \( 2 < \gamma < 3 \)
- Features
  - A very few high-degree nodes and many low-degree nodes
    \( \rightarrow \) hub-oriented network structure
  - Very low average shortest path length \( \rightarrow \) small-world network

Scale-Free Network Model (1)
Scale-Free Network Model (2)

- Example of B-A Model
  - Power-law degree distribution with the best fit of $\gamma = 2.9$ on the dashed line

![Graph showing power-law degree distribution](image)

Modular Network Model

- Modular Networks
  - Verified by high average clustering coefficient
    - Density of a graph $G(V,E)$: the number of actual edges over the number of all possible edges, $D(G) = 2|E| / |V||V|-1$
    - Clustering coefficient of a node $v$: the density of a subgraph with neighbors of $v$
  - Dense intra-connections among the nodes in the same modules
  - Sparse inter-connections between two nodes in different modules

![Diagram of modular network](image)
Hierarchical Networks

- Integrated of scale-free topology with modular structure
- Hierarchy of modules is controlled by hubs
- Clustering coefficient distribution $C$
  - Scale-free network & Modular network: $C$ is independent of degree $k$
  - Hierarchical network: $C \sim k^{-1}$

Schematic View

- Scale-free Networks
- Modular Networks
- Hierarchical Networks
Overview

- Backgrounds
- Network Modeling
  - Interaction / Function Prediction
- Protein Complex / Functional Module Detection
- Signaling Pathway / Functional Pathway Prediction
- Essential Protein Selection
- Conserved System Identification

Research Topics

- Function Prediction
- Protein Complex Detection
- Conserved Pattern Identification
- Essential Protein Selection
- Signaling Pathway Prediction
Phylogenetic Profile Analysis (1)

Main idea

- If two proteins in one organism have orthologs in another organism, they are likely to interact with each other and be functionally linked.

Genome

<table>
<thead>
<tr>
<th>Protein</th>
<th>E. Coli (EC)</th>
<th>S. Cerevisiae (SC)</th>
<th>B. Subtilis (BS)</th>
<th>H. Influenzae (HI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>P3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>P5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P6</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Profile Clusters

- P4 1 0 0
- P1 1 0 1
- P5 1 1 1
- P3 0 1 1
- P6 0 1 1

Conclusion

- P2 and P7 are functionally linked.
- P3 and P6 are functionally linked.

Phylogenetic Profile Analysis (2)

Process

- Input: a pair of proteins
- Find orthologs across species
- Build scoring matrices
- Calculate correlation between two proteins
Interaction-based Local Analysis

- **Majority-Based Method**
  - Assigning majority of functions of interacting partners to the unknown gene
  - Called "guilt-by-association"

Interaction-based Global Analysis

- **Extension of Majority-Based Method**
  - Assigning functions of all unknown genes in an interaction network
  - Requires optimization
Common Neighbor Analysis

- **Ratio of Common Interacting Partners**
  - If two proteins have many common interacting partners, they are likely to interact with each other
  - Jaccard coefficient: \( S(x, y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|} \)
  - Geometric coefficient: \( S(x, y) = \frac{|N(x) \cap N(y)|^2}{|N(x)| \cdot |N(y)|} \)
  - Dice coefficient: \( S(x, y) = \frac{2|N(x) \cap N(y)|}{|N(x)| + |N(y)|} \)
  - Simpson coefficient: \( S(x, y) = \frac{|N(x) \cap N(y)|}{\min(|N(x)|, |N(y)|)} \)
  - Hyper-geometric coefficient (P-value): \( P = \frac{\binom{T}{Z} \cdot \frac{T-Z}{T} \cdot \frac{T-|N(x)|}{T-|N(x)|-Z} \cdot \frac{T-|N(y)|}{T-|N(y)|-Z}}{\binom{T}{|N(x)|} \cdot \frac{T}{T-|N(x)|} \cdot \frac{T}{T-|N(y)|}} \)

Overview

- **Backgrounds**
- **Network Modeling**
- **Interaction / Function Prediction**
  - *Protein Complex / Functional Module Detection*
- **Signaling Pathway / Functional Pathway Prediction**
- **Essential Protein Selection**
- **Conserved System Identification**
Clustering Interaction Networks

- **Protein Complex**
  - A group of proteins having physical interactions at the same place, same time

- **Functional Module**
  - A group of proteins having the same function
  - A group of proteins having interactions even at different place, different time

- **Protein Complex / Functional Module Detection**
  - Clustering interaction networks by graph clustering algorithms
  - Detecting densely connected subgraphs
  - Grouping densely connected proteins and adding peripheral proteins

### Maximum Clique

- **Algorithm**
  - Find all maximum sized cliques
  - Use antimonotonic property
    - If a subset of set A is not a clique, then the set A is not a clique

```
size 2 cliques: (AB), (AC), (AE), ..... 
size 3 cliques: (ABC), (ACE), ..... 
size 4 cliques: (JKLM) 
```
Clique Percolation

- **Definitions**
  - Two $k$-cliques are adjacent if they share $(k-1)$ vertices where $k$ is the number of vertices in each clique.
  - A $k$-clique chain is a sub-graph comprising the union of a sequence of adjacent $k$ cliques.

- **Algorithm**
  1. Find all $k$-cliques
  2. Find all maximal $k$-clique chains by iterative merging adjacent $k$-cliques

- **Reference**

Hierarchical Approaches

- **Bottom-Up (Agglomerative) Approaches**
  - Start with each vertex as a cluster
  - Iteratively merge the closest clusters
  - Require a distance function between two clusters

- **Top-Down (Divisive) Approaches**
  - Start with the whole graph as a cluster
  - Recursively divide up the clusters
  - Require a cutting algorithm
Merging by Shortest Path Length

- **Main Idea**
  - Agglomerative approach using single-link distance

- **Algorithm**
  1. Select two closest vertices from different clusters based on the shortest path length between them
  2. Merge two clusters that include the selected vertices
  3. Repeat (1) and (2) until the shortest path length reaches a threshold

Merging by Common Neighbors (1)

- **Main Idea**
  - Agglomerative approach using the similarity based on common neighbors
  - More common neighbors two vertices share, more similar they are

- **Algorithm**
  1. Find the most similar vertices from different clusters based on a similarity function
  2. Merge the two clusters if the merged cluster reaches a density threshold
  3. Repeat (1) and (2) until no more clusters can be merged

- **Similarity Functions**
  - Jaccard coefficient: \( S(x,y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|} \)
  - Geometric coefficient: \( S(x,y) = \frac{|N(x) \cap N(y)|^2}{|N(x)||N(y)|} \)
Merging by Common Neighbors (2)

- More Similarity Functions
  - Dice coefficient: \( S(x, y) = \frac{2|N(x) \cap N(y)|}{|N(x)| + |N(y)|} \)
  - Simpson coefficient: \( S(x, y) = \frac{|N(x) \cap N(y)|}{\min(|N(x)|, |N(y)|)} \)
  - Maryland bridge coefficient: \( S(x, y) = \frac{1}{2} \left( \frac{|N(x) \cap N(y)|}{|N(x)|} + \frac{|N(y) \cap N(x)|}{|N(y)|} \right) \)

- Reference

Minimum Cut

- Definitions
  - Cut: a set of edges whose removal disconnects the graph
  - Minimum cut: a cut with minimum number of edges

- Algorithm
  - Recursively find the minimum cut

- Parameter
  - Minimum density threshold
  - Minimum size threshold
Betweenness Cut

- Betweenness
  - Measurement of vertices or edges located between clusters

- Algorithm
  1. Iteratively eliminate a vertex or an edge with the highest Betweenness value until the graph is separated
  2. Recursively apply (1) into each subgraph
  3. Repeat (1) and (2) until all subgraphs reach a density threshold

- Reference

Seed Growth

- Main Idea
  - Search for local optimization
  - Use a modularity (density) function
  - Types of seeds
    - Random seeds: selected randomly
    - Core seeds: selected by degree or clustering coefficient

- Algorithm
  1. Select a seed as an initial cluster S
  2. Add a neighbor to S repeatedly to find the max modularity (or density)
  3. Return S if the modularity of S > threshold
  4. Repeat (1), (2) and (3) to find a set of clusters
Graph Entropy (1)

- Definitions
  - An example of seed-growth algorithms
  - General notation
    - Inner links of $v$ in $G'(V',E')$: edges from $v$ to the vertices in $V'$
      $\rightarrow p_i(v)$: probability of $v$ having inner links
    - Outer links of $v$ in $G'(V',E')$: edges from $v$ to the vertices not in $V'$
      $\rightarrow p_o(v)$: probability of $v$ having outer links
  - Definitions
    - Vertex entropy: $e(v) = -p_i(v) \log_2 p_i(v) - p_o(v) \log_2 p_o(v)$
    - Graph entropy: $e(G(V,E)) = \sum_{v \in V} e(v)$
    - Find the minimum graph entropy during seed growth

Graph Entropy (2)

- Example

(a) $e(a) = e(b) = e(c) = 0$
$e(d) = 0.81$
$e(f) = 0.00$
$e(g) = e(h) = e(k) = 0$

(b) $e(a) = e(b) = e(c) = e(d) = 0$
$e(f) = 1.00$
$e(g) = 0.92$
$e(h) = e(k) = 0$
Graph Entropy (3)

- Algorithm
  1. Select a seed node, and include all neighbors of the seed node into a seed cluster.
  2. Iteratively remove a neighbor if removal decreases graph entropy.
  3. Iteratively add a node on the outer boundary of a current cluster if addition decreases graph entropy.
  4. Output the cluster with the minimal graph entropy.
  5. Repeat (1), (2), (3), and (4) until no seed node remains.

- Reference

Overview

- Backgrounds
- Network Modeling
- Interaction / Function Prediction
- Protein Complex / Functional Module Detection
  - Signaling Pathway / Functional Pathway Prediction
- Essential Protein Selection
- Conserved System Identification
Predicting Signaling Pathways

- Signaling Pathway
  - A series of proteins having signaling and response relationship

- Signaling Network
  - A combined form of linear signaling pathways
  - A directed acyclic graph

- Signaling Pathway / Signaling Network Prediction
  - Given starting and ending nodes, searching the strongest paths
  - Combining the strongest paths to form a signaling network

Path Frequency-Based Approach

- Main Idea
  - Measuring path frequency towards the target on a PPI network
  - Applying a greedy algorithm to repeatedly select a link making the most paths to the target

- Frequent Path Mining
  - Given a path prefix $l_x$, computes the support of the association rule, $l_x \rightarrow v_i$

  \[
  \text{Sup}(l^* \rightarrow v_i) = |f(P^*, l^* \circ v_i)|
  \]

  \[
  \text{Sup}(l^* \rightarrow v_i) = \sum_{v_j \in N(v_i) \setminus l^*} \text{Sup}(l^* \circ v_i \rightarrow v_j)
  \]
Network Motif-Based Approach

- **Main Idea**
  - Mapping network motifs into a PPI network
  - Applying a greedy algorithm to repeatedly select a link which fits the most network motifs

- **Network Motif Mapping**
  - Calculates motif occurrence scores by counting the matches
  - Computes motif strength by summing the weights of the matches

Overview

- **Backgrounds**
- **Network Modeling**
- **Interaction / Function Prediction**
- **Protein Complex / Functional Module Detection**
- **Signaling Pathway / Functional Pathway Prediction**
  - *Essential Protein Selection*
- **Conserved System Identification**
Measuring Essentiality of Proteins

- Essential Genes / Proteins
  - Functional core genes or proteins in a functional module
  - Hubs in an interaction network
  - Significance in biomedical applications: drug target detection

- Measurement Essentiality of Genes / Proteins
  - Local centrality measures
    - Degree
    - Clustering coefficient
  - Global centrality measures
    - Closeness
    - Betweenness

Local Centrality

- Clustering Coefficient of $v_i$
  - The density of a sub-graph $G'(V',E')$ where $V'$ is the set of neighbors of $v_i$
  - $C(v_i) = \frac{\sum_{j \in N(v_i)} \{v_j, v_i\}}{|N(v_i)|(|N(v_i)| - 1)}$
  - Measuring the effectiveness of $v_i$ on denseness

- Average Clustering Coefficient of $G(V,E)$
  - Average of the clustering coefficients of all vertices in $V$
  - Maximum is 1
  - Measuring the modularity of $G$

Global Centrality

- Closeness, $C_c(v_i)$
  - Detects the vertices located in the center of a graph
  - $C_c(v_i) = \frac{1}{\sum_{j \neq i} |p_i(v_i, v_j)|}$
    where $|p_i(v_i, v_j)|$ is the shortest path length between $v_i$ and $v_j$

- Betweenness, $C_B(v_i)$
  - Detects the vertices located between two clusters
  - $C_B(v_i) = \sum_{s \in V, t \in V, s \neq t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$
    where $\sigma_{st}$ is the number of shortest paths between $s$ and $t$, and
    $\sigma_{st}(v_i)$ is the number of shortest paths between $s$ and $t$, which pass
    through the vertex $v_i$

Overview

- Backgrounds
- Network Modeling
- Interaction / Function Prediction
- Protein Complex / Functional Module Detection
- Signaling Pathway / Functional Pathway Prediction
- Essential Protein Selection
- Conserved System Identification