Functional Pattern Mining from Genome-Scale Protein-Protein Interaction Networks

Young-Rae Cho, Ph.D.
Assistant Professor
Department of Computer Science
Baylor University
My Definition of Bioinformatics

- Hidden Knowledge (Pattern) Discovery

Computational Techniques
- Machine Learning Techniques
- Data Mining Algorithms
- Mathematical Models

Biological Data
- Genome
- Proteome
- Networks

Knowledge

Biomedical Applications
- Functional Characterization
- Disease Diagnosis
- Drug Development
Bioinformatics Milestone

Computational Biology → Functional Genomics → Systems Biology

Stage 1. Sequence Analysis
- DNA sequencing
- Homolog search
- Motif finding

Stage 2. Structure Analysis
- Protein folding
- Homolog search
- Binding site prediction

Stage 3. Genome Analysis
- Function prediction
- Gene clustering

Stage 4. System Analysis
- Network modeling
- Module finding
- Pathway prediction
Protein-Protein Interactions (PPIs)

- Permanent interactions vs. Transient interactions
- Determined by high-throughput experimental methods: mass spectrometry and yeast two-hybrid system

Interactome

- The entire set of PPIs in a species
- Problem?
  - Large scale & Unreliability
- Demand?
  - Computational, integrative approaches
Protein-Protein Interaction Networks

- A set of nodes $V$ represents proteins
- A set of edges $E$ represents interactions
- Undirected unweighed graph, $G(V,E)$
- Problem?
  - Complex connectivity
- Demand?
  - Computational, systematic approaches
Overview

- Protein Function Prediction
- Protein Complex Prediction
- Functional Module Detection
- Functional Hub Identification
- Functional Pathway Identification
Protein Function Prediction from PPI Networks

- **Previous Intuitions**
  - Utilize connectivity between proteins
    - Local neighborhood-based methods
  - Limited accuracy
    - Some interactions irrelevant to functional linkage

- **My Direction**
  - Utilize network motifs
    - Interconnection patterns occurring in PPI networks more frequently than in randomized networks
    - Core components for functional activities
    - Evolutionary conserved

<table>
<thead>
<tr>
<th># Motifs</th>
<th>Number of yeast motifs</th>
<th>Natural conservation rate</th>
<th>Random conservation rate</th>
<th>Conservation ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9,266</td>
<td>13.67%</td>
<td>4.63%</td>
<td>2.94</td>
</tr>
<tr>
<td>2</td>
<td>167,304</td>
<td>4.99%</td>
<td>0.81%</td>
<td>6.15</td>
</tr>
<tr>
<td>3</td>
<td>3,846</td>
<td>20.51%</td>
<td>1.01%</td>
<td>20.28</td>
</tr>
<tr>
<td>4</td>
<td>3,649,591</td>
<td>0.73%</td>
<td>0.12%</td>
<td>5.87</td>
</tr>
<tr>
<td>5</td>
<td>1,762,891</td>
<td>2.64%</td>
<td>0.18%</td>
<td>14.67</td>
</tr>
<tr>
<td>6</td>
<td>9,646</td>
<td>6.71%</td>
<td>0.17%</td>
<td>40.44</td>
</tr>
<tr>
<td>7</td>
<td>164,075</td>
<td>7.57%</td>
<td>0.17%</td>
<td>45.56</td>
</tr>
<tr>
<td>8</td>
<td>12,423</td>
<td>18.68%</td>
<td>0.12%</td>
<td>157.89</td>
</tr>
<tr>
<td>9</td>
<td>2,339</td>
<td>32.53%</td>
<td>0.08%</td>
<td>422.78</td>
</tr>
<tr>
<td>10</td>
<td>25,749</td>
<td>14.77%</td>
<td>0.05%</td>
<td>279.71</td>
</tr>
<tr>
<td>11</td>
<td>1,433</td>
<td>47.24%</td>
<td>0.02%</td>
<td>2,256.67</td>
</tr>
</tbody>
</table>

Wuchty, et al., 2003
Function Association Patterns

- **Subgraph Patterns**
  - 2-node subgraph
  - 3-node subgraph
  - 4-node subgraph

- **Functional Association Patterns**
  - Labeled subgraphs
  - Assigned a set of functions of a protein into the corresponding node as a label

- **Functional Association Pattern Mining**
  - Identification of functional association patterns occurring frequently

- **Function Prediction**
  - Predicting protein function by functional association pattern mining
Pattern Mining Process

- Assumption in Frequency
  - If a sub-pattern of a pattern \( p \) is not frequent, then \( p \) is not frequent
    → Downward closure
    → Applied the frequent item-set mining algorithm in the market basket problem

- Selective Joining
  - Merged two frequent \((k-1)\)-node patterns to generate a candidate \( k \)-node pattern
  - Considered the \((k-1)\)-node patterns which share a frequent \((k-2)\)-node sub-pattern

- Apriori Pruning
  - Filtered out infrequent patterns using a threshold of minimum frequency
  - Counted all isomorphic patterns using canonical forms
Pattern Mining Example

<table>
<thead>
<tr>
<th>candidate 2-node patterns</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>{f_1} - {f_1}</td>
<td>1</td>
</tr>
<tr>
<td>{f_1, f_2} - {f_1, f_2}</td>
<td>1</td>
</tr>
<tr>
<td>{f_1, f_3} - {f_1, f_3}</td>
<td>0</td>
</tr>
<tr>
<td>{f_1} - {f_1, f_2}</td>
<td>3</td>
</tr>
<tr>
<td>{f_1} - {f_1, f_3}</td>
<td>2</td>
</tr>
<tr>
<td>{f_1, f_2} - {f_1, f_3}</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>frequent 2-node patterns</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>{f_1} - {f_1, f_2}</td>
<td>3</td>
</tr>
<tr>
<td>{f_1} - {f_1, f_3}</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>frequent 3-node patterns</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>{f_1, f_2} - {f_1} - {f_1, f_3}</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>candidate 3-node patterns</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>{f_1} - {f_1, f_2} - {f_1}</td>
<td>1</td>
</tr>
<tr>
<td>{f_1, f_2} - {f_1} - {f_1, f_3}</td>
<td>1</td>
</tr>
<tr>
<td>{f_1} - {f_1, f_3} - {f_1}</td>
<td>1</td>
</tr>
<tr>
<td>{f_1, f_3} - {f_1} - {f_1, f_3}</td>
<td>0</td>
</tr>
<tr>
<td>{f_1, f_2} - {f_1} - {f_1, f_3}</td>
<td>3</td>
</tr>
</tbody>
</table>
Function Prediction Algorithm

- Pattern Analogue
  - Replaced only one node label in a pattern p with a different label

- Function Prediction
  - Given k, predicting function of an unknown protein by projecting each k-node frequent pattern

- Algorithm

  Prediction (G(V,E), F^{k-1}, F^k, v_u \in V)

  \( F_n^{k-1} \leftarrow \) selecting frequent patterns \( p_n^{k-1} \) including \( v_n \in N(v_u) \)

  \( P_u^k \leftarrow \) generating patterns \( p_u^k \) by extending \( p_n^{k-1} \in F_n^{k-1} \) to \( v_u \)

  if \( p_l^k \in P_u^k \) is a pattern analogue of \( p_l^k \in F_k \)

    \( f \leftarrow \) predicting function of \( v_u \) by the most frequent \( p_l^k \)

  end if

  return \( f \)
Function Prediction Accuracy

- **Evaluation Design**
  - Leave-one-out cross-validation
  - Exact match vs. Inclusive match
    - Exact match: \{ predicted functions \} \equiv \{ real functions \}
    - Inclusive match: \{ predicted functions \} \leq \{ real functions \}

- **Results**

- Cho and Zhang, *IEEE Transactions on Information Technology in Biomedicine (TITB)*, 2010
Overview

- Protein Function Prediction
- Protein Complex Prediction
- Functional Module Detection
- Functional Hub Identification
- Functional Pathway Identification
Protein Complex Prediction from PPI Networks

- **Previous Intuitions**
  - Search densely connected subgraphs
    - Graph clustering algorithms
  - Limited accuracy
    - Exclusion of proteins with low connectivity

- **My Direction**
  - Apply a seed-growth style approach
    - An initial seed cluster grows based a connectivity function
    - Searches core (dense region) and periphery (sparse region)
Graph Entropy

- Graph Entropy
  - An information-theoretic function of connectivity
  - General Notations
    - $p_i(v)$: probability of $v$ having inner links (edges from $v$ to the vertices in $V'$ of $G'(V', E')$)
    - $p_o(v)$: probability of $v$ having outer links (edges from $v$ to the vertices not in $V'$)
  - Definition
    - Vertex Entropy: $e(v) = -p_i(v) \log_2 p_i(v) - p_o(v) \log_2 p_o(v)$
    - Graph Entropy: $e(G(V, E)) = \sum_{v \in V} e(v)$

- Graph Entropy Example
Protein Complex Prediction Algorithm

- **Algorithm**
  - Greedy algorithm $\rightarrow$ Randomness, No parameters

- Creates an initial cluster including a vertex (seed) and its neighbors

- Removes vertices on cluster inner boundary iteratively to decrease graph entropy until it is minimal

- Adds vertices on cluster outer boundary iteratively to decrease graph entropy until it is minimal

- Outputs the cluster, and repeats all until there is no more vertex left
Protein Complex Prediction Accuracy

- Evaluation Design
  - F-measure
    - Mean of recall and precision
  - P-value
    - Hyper-geometric distribution

- Results

<table>
<thead>
<tr>
<th>seed selection method</th>
<th>average f-score</th>
<th>average p-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>random selection</td>
<td>0.452</td>
<td>4.63</td>
</tr>
<tr>
<td>highest degree</td>
<td>0.457</td>
<td>4.94</td>
</tr>
<tr>
<td>highest clustering coefficient</td>
<td>0.462</td>
<td>4.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>seed selection</th>
<th>average f-score</th>
<th>average p-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>random vertex</td>
<td>0.251</td>
<td>8.27</td>
</tr>
<tr>
<td>highest degree</td>
<td>0.311</td>
<td>8.56</td>
</tr>
<tr>
<td>highest clustering coefficient</td>
<td>0.320</td>
<td>8.41</td>
</tr>
</tbody>
</table>

- Kenley and Cho, ICDM 2012
Overview

- Protein Function Prediction
- Protein Complex Prediction
- Functional Module Detection
- Functional Hub Identification
- Functional Pathway Identification
Functional Module Detection from PPI Networks

- Previous Intuitions
  - Partition PPI networks
    - Graph clustering algorithms
  - Limited accuracy
    - Unreliable, complex connections
    - Non-overlapping clusters

- My Direction
  - Utilize an integrative approach
    - Semantic analytics using the genome-wide Gene Ontology (GO) data
  - Simulate information propagation
    - Information propagation from a selected source node through the network
Information Propagation Model

- **Path Strength Measurement**
  - Formula: \[ S(p) = \lambda \cdot w(v_0, v_1) \prod_{i=1}^{n-1} \frac{w(v_i, v_{i+1})}{d(v_i)} \]
  - Factors
    - Normalized edge weights
    - Inverse of path length
    - Inverse of node degree

- **Functional Impact Scoring**
  - Functional impact \( F_s(v_i) \) of a source \( v_s \) on \( v_i \) : Sum of strength of all possible paths from \( v_s \) to \( v_i \)

\[
S((v_s, v_1, \ldots, v_n)) = S((v_s, v_1, \ldots, v_{n-1})) \cdot \frac{w(v_{n-1}, v_n)}{d(v_{n-1})}
\]

\[
F_s(v_i) = \sum_{v_j \in N(v_i)} F_s(v_j) \cdot \frac{w(v_j, v_i)}{d(v_i)}
\]
Dynamic Propagation Algorithm

- Algorithm
  - Computation of cumulative functional impact of a source $v_s$ on all the other nodes
  - Repeated random walk simulation starting from a specific node $v_s$
  - Uses a minimum functional impact threshold
  - Stops when there are no more updates on cumulative functional impact on any nodes

- Example
Propagation Pattern Mining

- Process
Functional Module Detection Efficiency

- Evaluation in Synthetic Networks
  - Potential Factors
    - Number of nodes
    - Network density
    - Average node degree

- Results

![Graph 1: Run time vs. number of nodes]
- Constant density
- Constant average degree

![Graph 2: Run time vs. density]
- Constant number of nodes
- Constant average degree
Functional Module Detection Accuracy

- **Hypothesis**
  - If two proteins have coherent propagation patterns, then they are likely to perform the same functions.

- **Results**

<table>
<thead>
<tr>
<th>method</th>
<th>category</th>
<th>weighting</th>
<th># of clusters</th>
<th>average cluster size</th>
<th>accuracy ($-\log P$)</th>
<th>parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>flow pattern mining</td>
<td>flow-based</td>
<td>yes</td>
<td>14</td>
<td>11.20</td>
<td>5.47</td>
<td>$\theta_{10}r = 0.01$</td>
</tr>
<tr>
<td>edge-betweenness</td>
<td>hierarchical</td>
<td>no</td>
<td>43</td>
<td>9.67</td>
<td>4.62</td>
<td>min density = 0.2</td>
</tr>
<tr>
<td>clique percolation</td>
<td>density-based</td>
<td>no</td>
<td>52</td>
<td>5.50</td>
<td>3.72</td>
<td>$k = 3$</td>
</tr>
<tr>
<td>Markov Clustering</td>
<td>partition-based</td>
<td>yes</td>
<td>61</td>
<td>5.90</td>
<td>3.09</td>
<td>min size = 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Cho, Shi and Zhang, ICDM 2009
Overview

- Protein Function Prediction
- Protein Complex Prediction
- Functional Module Detection
- Functional Hub Identification
- Functional Pathway Identification
Functional Hub Identification from PPI Networks

- Previous Intuitions
  - Search high-degree nodes
  - Search centrally located nodes
    - Closeness / Betweenness
  - Limited accuracy
    - Unreliable, complex connections
    - Unable to find a hierarchy

- My Direction
  - Utilize an integrative approach
    - Semantic analytics using the genome-wide Gene Ontology (GO) data
  - Reconstruct the hierarchical structure of proteins
Functional Similarity Model

- **Path Strength**
  - Formula:
    \[
    S(p) = \lambda \cdot w(v_0, v_1) \prod_{i=1}^{n-1} \frac{w(v_i, v_{i+1})}{d(v_i)}
    \]

- **Functional Similarity**
  - \(k\)-length path strength:
    \[
    S_k(a, b) = \max_{v_0=a, v_k=b} S((v_0, v_1, \cdots, v_k))
    \]
  - Functional similarity:
    \[
    F(a, b) = \max_k S_k(a, b) \quad \text{where} \quad l \leq k \leq l+\theta
    \]

threshold
Network Conversion Algorithm

- **Algorithm**
  - Conversion of a complex PPI network into a hierarchical tree structure of proteins

- **Process**
  1. **Computes centrality for each node** \( a \)

\[
C(a) = \sum_{b \in V} F(a, b)
\]

  2. **Obtains a set of ancestor nodes** \( T(a) \) of \( a \)

\[
T(a) = \{ b \mid C(b) > C(a) \}
\]

  3. **Selects a parent node** \( p(a) \) of \( a \)

\[
p(a) = \begin{cases} 
  \text{null} & \text{if } T(a) = \emptyset \\
  \arg \max_{b \in T(a)} F(a, b) & \text{otherwise.}
\end{cases}
\]
Hub Confidence Measurement

- Algorithm
  - Quantification of functional “hubness” of proteins

- Process
  1. Obtains a set of child nodes $D(a)$ of $a$
     \[ D(a) = \{ b \mid p(b) = a \} \]
  2. Obtains a set of descendant nodes $L_a$ of $a$
     \[ L_a = \left( \bigcup_{b \in D(a)} L_b \right) \cup \{ a \} \]
  3. Computes the hub confidence $H(a)$ of $a$
     \[
     H(a) = \begin{cases} 
     \sum_{b \in L_a} f(a,b) & \text{if } p(a) = \text{null} \\
     \sum_{b \in L_a} f(a,b)/f(a,p(a)) & \text{otherwise.}
     \end{cases}
     \]

\[
\begin{array}{|c|}
\hline
\text{A} & 11.56 \\
\text{L} & 4.77 \\
\text{O} & 3.92 \\
\text{I} & 3.33 \\
\text{E} & 3.19 \\
\text{S} & 1.89 \\
\text{B} & 1.27 \\
\text{G} & 1.14 \\
\text{D} & 1.06 \\
\hline
\end{array}
\]
Topological Assessment of Functional Hubs

- Network Vulnerability Test
  - Random attack: Repeatedly disrupt a randomly selected node
  - Degree-based hub attack: Repeatedly disrupt the highest degree node
  - Functional hub attack: Repeatedly disrupt the node with the highest hub confidence
  - For each iteration, observe the largest component

- Results

![Graph showing network vulnerability test results]

- Fraction of largest component against number of nodes
- Lines represent:
  - Green: Random attack
  - Red: Degree-based hub attack
  - Blue: Structural hub attack

<table>
<thead>
<tr>
<th>Fraction of Largest Component</th>
<th>Number of Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0</td>
</tr>
<tr>
<td>0.95</td>
<td>20</td>
</tr>
<tr>
<td>0.90</td>
<td>40</td>
</tr>
<tr>
<td>0.85</td>
<td>60</td>
</tr>
<tr>
<td>0.80</td>
<td>80</td>
</tr>
<tr>
<td>0.75</td>
<td>100</td>
</tr>
<tr>
<td>0.70</td>
<td>120</td>
</tr>
<tr>
<td>0.65</td>
<td>140</td>
</tr>
<tr>
<td>0.60</td>
<td>160</td>
</tr>
</tbody>
</table>
Biological Assessment of Functional Hubs

- **Protein Lethality Test**
  - To determine lethal proteins by knock-out experiment
    (Lethality represents functional essentiality.)
  - Order proteins by degree and hub confidence
  - For every 10 proteins, observe the cumulative proportion of lethal proteins

- **Results**

![Graph showing the average lethality of structural and degree-based hubs against the number of hubs.]

- **Cho and Zhang, BMC Bioinformatics, 2010**
Overview

- Protein Function Prediction
- Protein Complex Prediction
- Functional Module Detection
- Functional Hub Identification
- Functional Pathway Identification
Functional Pathway Identification from PPI Networks

- Previous Intuitions
  - Search the strongest path between a source and a target
    - Computational inefficiency
  - Limited accuracy
    - Difference between signal transduction and interaction

- My Direction
  - Utilize an integrative approach
    - Semantic analytics using the genome-wide Gene Ontology (GO) data
  - Measurement of path frequency towards the target node
Frequent Path Mining

- **Notations**
  - \( l^x \): a list of length \( x \), \( l^x = (v_i, v_{i+1}, v_{i+2} \cdots v_{i+x}) \)
  - \( P^k \): a set of all paths of length \( k \) between a source and a target
  - \( L^x \): a set of path prefixes of \( P^k \) of length \( x \)
    \[
    L^x = \{ l^x \mid \exists p_i \in P^k, \text{s.t. } l^x \text{ is a path prefix of } p_i \text{ and } 0 \leq x \leq k \}.
    \]
  - A selection function \( f \): returning a subset of \( P^k \), having a specific prefix \( l^x \)
    \[
    f(P^k, l^x) = \{ p_i \mid p_i \in P^k \text{ and } l^x \text{ is a path prefix of } p_i \}.
    \]

- **Frequent Path Mining**
  - Given a path prefix \( l^x \), computes the support of the association rule, \( l^x \rightarrow v_i \)

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

\[
\text{Sup}(l^x \rightarrow v_i) = \sum_{v_j \in N(v_i) - l^x} \text{Sup}(l^x \circ v_i \rightarrow v_j)
\]
Functional Pathway Prediction Algorithm

- Algorithm
  - Mining frequent paths
  - Selecting multiple successors by an *expansion parameter*
  - Was the target selected?
    - no
    - yes
  - Merging the pathways
Efficiency Improvement of Path Mining

- Preprocessing
  - Quasi-clustering
    - Information propagation from the source node
    - Frequent path mining in a subgraph of the PPI network

- Approximation
  - Approximate support pre-computation
    - Estimation of cycles
    - Use of a pruning parameter
Functional Pathway Identification Performance

- **Accuracy**
  - Test for MAP Kinase signaling pathways

<table>
<thead>
<tr>
<th>Expansion Parameter $\gamma$</th>
<th>Average Recall</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.733</td>
<td>0.733</td>
</tr>
<tr>
<td>0.93</td>
<td>0.933</td>
<td>0.628</td>
</tr>
<tr>
<td>0.90</td>
<td>0.933</td>
<td>0.371</td>
</tr>
<tr>
<td>0.85</td>
<td>1.000</td>
<td>0.218</td>
</tr>
<tr>
<td>0.80</td>
<td>1.000</td>
<td>0.167</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Average Recall</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>$iSup$ (approximate support) ‡</td>
<td>0.881</td>
<td>0.327</td>
</tr>
<tr>
<td>color-coding</td>
<td>0.939</td>
<td>0.111</td>
</tr>
<tr>
<td>$iSup$ (quasi-clustering) * †</td>
<td>1.000</td>
<td>0.218</td>
</tr>
<tr>
<td>edge orientation *</td>
<td>1.000</td>
<td>0.142</td>
</tr>
</tbody>
</table>

- **Efficiency**

[Bar chart showing runtime (msec) for different methods across signaling network IDs.]
Overview

- Protein Function Prediction
- Protein Complex Prediction
- Functional Module Detection
- Functional Hub Identification
- Functional Pathway Identification
Conclusion

- Bioinformatics Research Trend
  - Gene-level → Genome-level
  - The local → The global
  - The particular → The universal

- Functional Characterization Process
  - Pre-genomic era: sequence → structure → function
  - Post-genomic era: interaction → network → function
Bioinformatics Program at Baylor

- Bioinformatics in Computer Science
References

- My personal webpage:  http://web.ecs.baylor.edu/faculty/cho/
- My lab webpage:  http://bionet.ecs.baylor.edu/
- Baylor, Bioinformatics program webpage:  http://www.ecs.baylor.edu/bioinformatics/
- Baylor, Institute of Biomedical Studies webpage:  http://www.baylor.edu/biomedical_studies/