Identification of Signaling Pathways

Advanced Bioinformatics (BMI/CS 838)
March 17, 2015

Professor Tony Gitter
Goals for lecture

• Challenges of integrating high-throughput assays
• Connecting relevant genes/proteins with interaction networks
• ResponseNet algorithm
• Related signaling pathway prediction methods
High-throughput screening

• Which genes are involved in which cellular processes?
• Hit: gene that affects the phenotype
• Phenotypes include:
  • Growth rate
  • Cell death
  • Cell size
  • Intensity of some reporter
  • Many others
Types of screens

• Genetic screening
  • Test genes individually or in parallel
  • Knockout, knockdown (RNA interference), overexpression, CRISPR/Cas genome editing

• Chemical screening
  • Which genes are affected by a stimulus?
Differentially expressed genes

- Compare mRNA levels between control and treatment conditions
  - Previously microarrays, now RNA-seq
- Genes whose expression changes significantly are also involved in the cellular process
Interpreting screens

Screen hits

Differentially expressed genes

Very few genes detected in both
Assays reveal different parts of a cellular process
Assays reveal different parts of a cellular process

Differentially expressed genes

Genetic screen hits
Pathways connect the disjoint gene lists

- Can’t rely on pathway databases
- High-quality, low coverage

- Instead learn condition-specific pathways
- Combine data with generic physical interaction networks
Physical interactions

- Protein-protein

- Metabolic

- Protein-DNA (transcription factor-gene)

- Genes and proteins are different node types
Weighting interactions

• Probability-like confidence of the interaction

<table>
<thead>
<tr>
<th>Proteins</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MP2K1_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td>Temporarily not available for viewing in Netlity.</td>
</tr>
<tr>
<td>MK01_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td>Temporarily not available for viewing in Netlity.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidence</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source DB</td>
<td>Source ID</td>
<td>Interaction Type</td>
<td>PSI MI Code</td>
<td>PubMed ID</td>
<td>Detection Type</td>
<td>PSI MI Code</td>
</tr>
<tr>
<td>biogrid</td>
<td>857930</td>
<td>direct interaction</td>
<td>MI:0407</td>
<td>12788955</td>
<td>enzymatic study</td>
<td>MI:0415</td>
</tr>
<tr>
<td>ophid</td>
<td>17231</td>
<td>aggregation</td>
<td>MI:0191</td>
<td>11352917</td>
<td>confirmational text mining</td>
<td>MI:0024</td>
</tr>
<tr>
<td>ophid</td>
<td>17231</td>
<td>aggregation</td>
<td>MI:0191</td>
<td>15657099</td>
<td>deglycosylase assay</td>
<td>MI:1006</td>
</tr>
<tr>
<td>ophid</td>
<td>17234</td>
<td>aggregation</td>
<td>MI:0191</td>
<td>11352917</td>
<td>confirmational text mining</td>
<td>MI:0024</td>
</tr>
<tr>
<td>ophid</td>
<td>17234</td>
<td>aggregation</td>
<td>MI:0191</td>
<td>15657099</td>
<td>deglycosylase assay</td>
<td>MI:1006</td>
</tr>
<tr>
<td>biogrid</td>
<td>259225</td>
<td>direct interaction</td>
<td>MI:0407</td>
<td>12697810</td>
<td>t7 phage display</td>
<td>MI:0108</td>
</tr>
<tr>
<td>intact</td>
<td>EBI-8279991</td>
<td>phosphorylation reaction</td>
<td>MI:0217</td>
<td>23241949</td>
<td>biosensor</td>
<td>MI:0968</td>
</tr>
</tbody>
</table>

• Example evidence: edge score of 1.0
• 16 distinct publications supporting the edge
Identify connections within an interaction network

Yeger-Lotem2009
Hairball networks

- Networks are highly connected
- Can’t use naïve strategy to connect screen hits and differentially expressed genes

Yeger-Lotem 2009
Framing an optimization problem

• ResponseNet optimization goals
  • Connect screen hits and differentially expressed genes
  • Recover sparse connections
  • Prefer high-confidence interactions
Construct the interaction network
Transform to a flow problem
Weights and capacities on edges

\[ c_{Si} = \frac{|strength_i|}{\sum_{j \in Gen} |strength_j|} \]

\[ w_{ij}, c_{ij} \]

\[ c_{iT} = \frac{|\log_2(strength_i)|}{\sum_{j \in Tra} |\log_2(strength_j)|} \]

\( w_{ij} \) from interaction network confidence
Find the minimum cost flow

Prefer no flow on the high cost edges if alternative paths exist
Formal minimum cost flow

\[
\min_{f} \left( \sum_{i \in V', j \in V'} - \log(w_{ij}) \cdot f_{ij} \right) - (\gamma \cdot \sum_{i \in \text{Gen}} f_{Si})
\]

Subject to:

\[
\sum_{j \in V'} f_{ij} - \sum_{j \in V'} f_{ji} = 0 \quad \forall i \in V' - \{S, T\}
\]

\[
\sum_{i \in \text{Gen}} f_{Si} - \sum_{i \in \text{Tra}} f_{iT} = 0
\]

\[
0 \leq f_{ij} \leq c_{ij} \quad \forall (i, j) \in E'
\]
Linear programming

• Optimization problem is a linear program
• Canonical form

\[
\begin{align*}
\text{maximize} & \quad c^T x \\
\text{subject to} & \quad Ax \leq b \\
\text{and} & \quad x \geq 0
\end{align*}
\]

• Polynomial time complexity
• Many off-the-shelf solvers

Wikipedia
ResponseNet pathways

- Identifies pathway members that are neither hits nor differentially expressed
- Ste5 recovered when \textit{STE5} deletion is the perturbation
ResponseNet summary

• Advantages
  • Computationally efficient
  • Incorporates interaction confidence
  • Identifies plausible networks

• Disadvantages
  • Direction of flow is not biologically meaningful
  • Path length not considered
  • Requires sources and targets
Alternative pathway identification algorithms

• k-shortest paths
  • Ruths2007
  • Shih2012

• Random walks / network diffusion / circuits
  • Tu2006
  • eQTL electrical diagrams (eQED)
  • HotNet

• Integer programs
  • Signaling-regulatory Pathway INference (SPINE)
  • Chasman2014
Alternative pathway identification algorithms continued

• Path-based objectives
  • Physical Network Models (PNM)
  • Maximum Edge Orientation (MEO)
  • Signaling and Dynamic Regulatory Events Miner (SDREM)

• Steiner tree
  • Prize-collecting Steiner forest (PCSF)
  • Belief propagation approximation (msgsteiner)

• Hybrid approaches
  • PathLinker: random walk + shortest paths
  • ANAT: shortest path + Steiner tree
Recent developments in pathway discovery

- Multi-task learning: jointly model several related biological conditions
  - ResponseNet extension: **SAMNet**
  - Steiner forest extension: **Multi-PCSF**
  - SDREM extension: **MT-SDREM**

- Temporal data
  - ResponseNet extension: **TimeXNet**
  - Pathway synthesis
Condition-specific genes/proteins used as input

- Genetic hits (as causes or effects)
- Differentially expressed genes
- Transcription factors inferred from gene expression
- Proteomic changes (protein abundance or phosphorylation)
- Genetic variants or DNA mutations
- Receptors or sensory proteins
- Protein interaction partners
- Pathway databases or other prior knowledge
If you’re still interested

• Computational Network Biology
  • Fall 2015 special topics course
  • BMI 826/CS 838
  • Professor Sushmita Roy

• Talk to BMI faculty working on these problems
  • Professors Craven, Gitter, Roy, etc.