Goals for lecture

• Challenges of integrating high-throughput assays
• Connecting relevant genes/proteins with interaction networks
• ResponseNet algorithm
• Classes of 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 transcript levels between control and treatment conditions

• Genes whose expression changes significantly are also involved in the cellular process

• Alternatively, differential protein abundance or phosphorylation
Interpreting screens

Screen hits

Differentially expressed genes

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

Database representation of a "pathway"
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 computationally
- Combine data with generic physical interaction networks
Physical interactions

• Protein-protein interactions (PPI)

• Metabolic

• Protein-DNA (transcription factor-gene)

• Genes and proteins are different node types
Hairball networks

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

Yeger-Lotem 2009
Identify connections within an interaction network

Yeger-Lotem 2009
How to define a computational “pathway”

• **Given:**
  • Partially directed network of known physical interactions (e.g. PPI, kinase-substrate, TF-gene)
  • Scores on source nodes
  • Scores on target nodes

• **Do:**
  • Return directed paths in the network connecting sources to targets
ResponseNet optimization goals

• Connect screen hits and differentially expressed genes
• Recover sparse connections
• Identify intermediate proteins missed by the screens
• Prefer high-confidence interactions
Construct the interaction network
Transform to a flow problem
Max flow on graphs

Each edge can tolerate different level of flow or have different preference of sending flow along that edge.

Pump flow from source

Incoming and outgoing flow conserved at each node

Flow conserved to target
Weighting interactions

• Probability-like confidence of the interaction

| Proteins | | | |
|----------|--------------|-------------------------------|
| MP2K1_HUMAN | Homo sapiens | Temporarily not available for viewing in Netlity. |
| MK01_HUMAN | Homo sapiens | Temporarily not available for viewing in Netlity. |

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Source DB</th>
<th>Source ID</th>
<th>Interaction Type</th>
<th>PSI MI Code</th>
<th>PubMed ID</th>
<th>Detection Type</th>
<th>PSI MI Code</th>
</tr>
</thead>
<tbody>
<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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></td>
</tr>
</tbody>
</table>

• Example evidence: edge score of 1.0
• 16 distinct publications supporting the edge
Weights and capacities on edges

\[ c_{Si} = \frac{|\text{strength}_i|}{\sum_{j \in \text{Gen}} |\text{strength}_j|} \]

\[ c_{ij} = 1 \]

Flow capacity

\[ c_{iT} = \frac{|\log_2(\text{strength}_i)|}{\sum_{j \in \text{Tra}} |\log_2(\text{strength}_j)|} \]

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

Return the edges with non-zero flow

Prefer no flow on the low-weight edges if alternative paths exist
Formal minimum cost flow

\[
\min\left(\sum_{f} - \log(w_{ij}) \cdot f_{ij} - (\gamma \cdot \sum_{i \in \text{Gen}} f_{Si})\right)
\]

Positive flow on an edge incurs a cost
Flow on an edge
Cost is greater for low-weight edges
Parameter controlling the amount of flow from the source
Formal minimum cost flow

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

Subject to:

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

Flow coming in to a node equals flow leaving the node
Formal minimum cost flow

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

Subject to:

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

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

Flow leaving the source equals flow entering the target
Formal minimum cost flow

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

Subject to:

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

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

Flow is non-negative and does not exceed edge capacity

\[
0 \leq f_{ij} \leq c_{ij} \quad \forall (i, j) \in E'
\]
Formal minimum cost flow

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

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

• Practical Optimization: A Gentle Introduction
  • Introduction to linear programming
  • Simplex method
  • Network flow

Wikipedia
ResponseNet pathways

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

• Advantages
  • Computationally efficient
  • Integrates multiple types of data
  • Incorporates interaction confidence
  • Identifies biologically plausible networks

• Disadvantages
  • Direction of flow is not biologically meaningful
  • Path length not considered
  • Requires sources and targets
  • Dependent on completeness and quality of input network
Evaluating pathway predictions

• Unlike PIQ, we don’t have a complete gold standard available for evaluation

• Can simulate “gold standard” pathways from a network

• Compare relative performance of multiple methods on independent data
  • Top secret example
Evaluating pathway predictions

Ritz2016
Evaluating pathway predictions

Ritz2016
Evaluating pathway predictions

• Natural language processing can also help semi-automated evaluation

  • **Literome**

  PMID: 14611643

  WNK1, the kinase mutated in an inherited high-blood-pressure syndrome, is a novel PKB (protein kinase B)/Akt substrate.

  ... that PKB mediates the ... of WNK1 at ... (details)

  • **Chilibot**


  • **iHOP**

  Akt1 ✴, but not Akt2, phosphorylates palladin ✴ at Ser507 in a domain that is critical for F-actin bundling. [2010]
Classes of pathway prediction algorithms

Are edges important?

- No
  - Network diffusion
  - No
    - Spanning tree
  - Yes
    - Sources and targets?
      - No
        - Steiner tree
      - Yes
        - Next slide...
Classes of pathway prediction algorithms

- Have sources and targets
  - What path properties are important?
    - Total path length or score
    - Total source-target connectivity
    - Connectivity in minimum cost network
    - Complex properties
      - Integer program
      - Symbolic solver
      - Graphical model
    - Shortest paths
    - Network flow
    - Steiner tree
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

• 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)
  • Omics Integrator implementation
• Hybrid approaches
  • PathLinker: random walk + shortest paths
  • ANAT: shortest paths + 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**
  • Steiner forest extension
  • **Temporal Pathway Synthesizer** (unpublished)
Condition-specific genes/proteins used as input

• Genetic screen hits (as causes or effects)
• Differentially expressed genes
• Transcription factors inferred from gene expression
• Proteomic changes (protein abundance or post-translational modifications)
• Kinases inferred from phosphorylation
• Genetic variants or DNA mutations
• Enzymes regulating metabolites
• Receptors or sensory proteins
• Protein interaction partners
• Pathway databases or other prior knowledge