Network Biology

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2021
Daifeng Wang
daifeng.wang@wisc.edu
Goals for lecture

• Biological networks
• Challenges of integrating high-throughput assays
• Connecting relevant genes/proteins with interaction networks
• ResponseNet algorithm
• Evaluating pathway predictions
• 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-Lotem2009
Identify connections within an interaction network

Yeger-Lotem2009
Biological Network Properties

• **Degree**: number of neighbors of a node

• **Power law degree distribution**
  – Most nodes have low degrees
  – Few highly connected nodes (hubs)

• **Robust to random attacks**
  – e.g., structure resilient to mutations
  – Mutations in hubs can damage the network

• **Modular organization**
  – High clustering coefficient (short paths)
  – Efficient signal propagation
Power law degree distribution

- Probability of finding a highly connected node decreases exponentially with $K$

$$P(K) \sim K^{-\gamma}$$


- a) A. fulgidus (Archae)
- b) Bacterium
- c) C. elegans (Eukaryote),
- d) averaged over 43 organisms
Modularity

- Small highly connected cohesive clusters that combine to form larger units
- Communication between clusters through hubs
- Hierarchical modularity overlaps with known metabolic functions

Modularity $Q$: measurement on strength of network division

$$Q = \frac{1}{2m} \sum_{i,j} \left( W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

- **Normalization** $m$: total number of edges
- **Edge weight** between nodes $i$ and $j$
- **Expected edge weight** that would go between $i$ and $j$

**Clustering goal:** assign each node a module to maximize “modularity” as an objective function (module is a group of highly connected nodes)

Clustering coefficient

Measures the average probability that two neighbors of a node are connected

\[ C_I = \frac{n_I}{\binom{k}{2}} = \frac{2n_I}{k \cdot (k - 1)} \]

- \( n_I \): # edges between node \( I \)'s neighbors
- \( k \): # of neighbors of \( I \)
High degree nodes -> low clustering coefficient CC
Network’s modularity -> CC averaged over all nodes
Metabolic networks have high intrinsic modularity

Network centralities

Topological importance of a node

G. Iacono et al., Genome Biology 20 (2019)
Network problems

• Network inference
  – Infer network structure

• Motif finding
  – Identify common subgraph topologies

• Pathway or module detection
  – Identify subgraphs of genes that perform the same function or active in same condition

• Network comparison, alignment, querying

• Conserved modules
  – Identify modules that are shared in networks of multiple species/conditions
Network motifs

• Problem: Find subgraph topologies that are statistically more frequent than expected
• Brute force approach
  – Count all topologies of subgraphs of size m
  – Randomize graph (retain degree distribution) and count again
  – Output topologies that are over/under represented

*Feed-forward loop*: over-represented in regulatory networks

*not very common*
Gene regulatory network motifs

Network modules

- Modules: dense (highly-connected) subgraphs (e.g., large cliques or partially incomplete cliques)
- Problem: Identify the component modules of a network
- Difficulty: definition of module is not precise
  - Hierarchical networks have modules at multiple scales
  - At what scale to define modules?
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
Network flow problem

• Finding an optimal route by minimizing transportation costs from LA to NYC
  – $c_{i,j}$, the cost between City $i$ and City $j$
  – $f_{i,j} = 1$ if in route, $= 0$ if not
  – $\text{argmin}_f \sum c_{i,j} * f_{i,j}$ s.t. constraints

https://www.visualcapitalist.com/u-s-interstate-highways-transit-map/
ResponseNet optimization goals

- Connect screen hits and differentially expressed genes
- Recover sparse connections
- Identify intermediate proteins missed by the screens
- Prefer high-confidence interactions

- Minimum cost flow formulation can meet these objectives
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

<table>
<thead>
<tr>
<th>Protein</th>
<th>Species</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP2K1_HUMAN</td>
<td>Homo sapiens</td>
<td>Temporarily not available for viewing in Netility.</td>
</tr>
<tr>
<td>MK01_HUMAN</td>
<td>Homo sapiens</td>
<td>Temporarily not available for viewing in Netility.</td>
</tr>
</tbody>
</table>

### Evidence

<table>
<thead>
<tr>
<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>
</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
Weights and capacities on edges

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

\[ c_{ij} = 1 \]

Flow capacity

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

\( w_{ij} \) from interaction network confidence

\((w_{ij}, c_{ij})\)
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

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

Positive flow on an edge incurs a cost.

Cost is greater for low-weight edges.

Flow on an edge.

Parameter controlling the amount of flow from the source.
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:

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

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

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

Subject to:

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

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

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

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

Subject to:

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

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

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

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

Subject to:

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

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

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

- Optimization problem is a linear program
- Canonical form

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

- Polynomial time complexity
- Many off-the-shelf solvers
- **Practical Optimization: A Gentle Introduction**
  - Introduction to linear programming
  - Simplex method
  - Network flow
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
Evaluating pathway predictions

Ritz2016  https://www.nature.com/articles/npisba20162.pdf
Evaluating pathway predictions

Ritz2016
Evaluating pathway predictions

- PR curves can evaluate node or edge recovery but not the global pathway structure
Evaluation beyond pathway databases

• 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.

  • Chilibot

  Our studies reveal a novel mechanism in which phosphorylation of STAT3 is mediated by a constitutively active JNK2 [MAPK9] isoform, JNK2 [MAPK9] ¹⁻. Ref: Oncogene, 2011, PMID: 20871632

  • 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
  - Yes
    - Sources and targets?
      - No
        - Spanning tree
      - Yes
        - Steiner tree

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 and ST-Steiner
  – Temporal Pathway Synthesizer
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