Canadian Bioinformatics Workshops

www.bioinformatics.ca

In collaboration with
Cold Spring Harbor Laboratory
&
New York Genome Center
Module 4
Depth on networks and pathways

Robin Haw
High-throughput Biology: From Sequence to Networks
April 27-May 3, 2015
**Why Pathway/Network Analysis?**

- Dramatic data size reduction: 1000’s of genes => dozens of pathways.
- Increase statistical power by reducing multiple hypotheses.
- Find meaning in the “long tail” of rare cancer mutations.
- Tell biological stories:
  - Identifying hidden patterns in gene lists.
  - Creating mechanistic models to explain experimental observations.
  - Predicting the function of unannotated genes.
  - Establishing the framework for quantitative modeling.
  - Assisting in the development of molecular signatures.

---

**What is Pathway/Network Analysis?**

- Any analytic technique that makes use of biological pathway or molecular network information to gain insights into a biological system.
- A rapidly evolving field.
- Many approaches.
Pathways vs Networks

- Reactome & KEGG
  - explicitly describe biological processes as a series of biochemical reactions.
  - represents many events and states found in biology.

Reaction-Network Databases
KEGG

- KEGG is a collection of biological information compiled from published material → curated database.
- Includes information on genes, proteins, metabolic pathways, molecular interactions, and biochemical reactions associated with specific organisms.
- Provides a relationship (map) for how these components are organized in a cellular structure or reaction pathway.
Reactome

- Open source and open access pathway database
- Curated human pathways encompassing metabolism, signaling, and other biological processes.
- Every pathway is traceable to primary literature.
- Cross-reference to many other bioinformatics databases.
- Provides data analysis and visualization tools
What is an Interaction Network?

- An Interaction Network is a collection of:
  - Nodes (or vertices).
  - Edges connecting nodes (directed or undirected, weighted, multiple edges, self-edges).
- Nodes can represent proteins, genes, metabolites, or groups of these (e.g. complexes) - any sort of object.
- Edges can be either physical or functional interactions, activators, regulators, reactions - any sort of relations.
Types of Interactions Networks

Degree: the number of edges that a node has.

Closeness: measure how close a node to all other nodes in the network.

Betweenness: quantify the number of all shortest paths that pass thorough a node.

Weighting the Nodes

The node with the highest degree in the graph

The nodes with the highest closeness

The nodes with the highest betweenness

CentiScePe cytoscape plugin can calculate various graph measures, including closeness, betweenness, etc
Network Databases

- Can be built automatically or via curation.
- More extensive coverage of biological systems.
- Relationships and underlying evidence more tentative.
- Popular sources of curated networks:
  - BioGRID – Curated interactions from literature; 57,065 genes, 572,682 interactions.
  - IntAct – Curated interactions from literature; 86,812 genes, 349,690 interactions.
  - MINT – Curated interactions from literature; 53,553 genes, 241,458 interactions.

IntAct

7,708 binary interactions found for search term 'tp53'

<table>
<thead>
<tr>
<th>Interactions (7708)</th>
<th>Browser</th>
<th>Lists</th>
<th>Interaction Viewer</th>
<th>Molview</th>
<th>Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter out the node expanded co-complexes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7708 negative interactions matching the query (link now, set exported in MINT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Your query also matches 8423 interaction evidences from 7 other databases.

IntAct
Types of Pathway/Network Analysis

- **Goal**
  1. Enrichment of f and gene sets
  2. De novo subnetwork construction and clustering
  3. Pathway-based modeling

- **Tools**
  - What biological processes are altered in this cancer?
  - Are new pathways altered in this cancer? Are there clinically-relevant tumour subtypes?
  - How are pathway activities altered in a particular patient? Are there targetable pathways in this patient?

Visualization and Analysis Tools for Biological Networks

- **Tools**
  - Cytoscape
  - NAVIGaTOR
  - Osprey

Module 4 bioinformatics.ca
Open Data Exchange

**SBML**
- Level 2.3
- Standard graphical languages for representing biological processes and interactions
- Open access interchange format for computer models of biochemical pathways, reactions and networks.

**BioPAX**
- Level 2 & 3
- Standard language that aims to enable integration, exchange, visualization and analysis of biological pathway data.

**PSICQUIC**
- PSI-MITAB is the data exchange format.

Using Reactome with other Network Visualisation and Analysis Tools

- CellDesigner (SBML)
- Cytoscape (BioPAX)
- VANTED (SBGN)
2) De Novo Subnetwork Construction & Clustering

• Apply list of altered genes, proteins, RNAs to a biological network.
• Identify “topologically unlikely” configurations.
  - E.g. a subset of the altered genes are closer to each other on the network than you would expect by chance.
• Extract clusters of these unlikely configurations.
• Annotate the clusters.

Network clustering

• Clustering can be defined as the process of grouping objects into sets called clusters (communities or modules), so that each cluster consists of elements that are similar in some ways.

• Network clustering algorithm is looking for sets of nodes [proteins] that are joined together in tightly knit groups.

• Cluster detection in large networks is very useful as highly connected proteins are often sharing similar functionality.
Popular Network Clustering Algorithms

- Girvin-Newman
  - a hierarchical method used to detect communities by progressively removing edges from the original network
- Markov Cluster Algorithm
  - a fast and scalable unsupervised cluster algorithm for graphs based on simulation of (stochastic) flow in graphs
- HotNet
  - Finds “hot” clusters based on propagation of heat across metallic lattice.
  - Avoids ascertainment bias on unusually well-annotated genes.
- HyperModules Cytoscape App
  - Find network clusters that correlate with clinical characteristics.
- Reactome FI Network Cytoscape App
  - Offers multiple clustering and correlation algorithms (including HotNet, PARADIGM and survival correlation analysis)

Typical output of a network clustering algorithm

This hypothetical subnetwork was decomposed onto 6 clusters.

Different clusters are marked with different colors. Cluster 6 contains only 2 elements and could be ignored in the further investigations.

Clusters are mutually exclusive meaning that nodes are not shared between the clusters.
Reactome Functional Interaction (FI) Network and ReactomeFIViz app

- No single mutated gene is necessary and sufficient to cause cancer.
  - Typically one or two common mutations (e.g. TP53) plus rare mutations.
- Analyzing mutated genes in a network context:
  - reveals relationships among these genes.
  - can elucidate mechanism of action of drivers.
  - facilitates hypothesis generation on roles of these genes in disease phenotype.
- Network analysis reduces hundreds of mutated genes to < dozen mutated pathways.

What is a Functional Interaction?

- Convert reactions in pathways into pair-wise relationships
  - **Functional Interaction**: an interaction in which two proteins are involved in the same reaction as input, catalyst, activator and/or inhibitor, or as components in a complex

**FI network construction: part I**

Data sources for annotated FIs:
- Reactome [1]
- Panther [2]
- KEGG [3]
- NCI-Biocarta [4]
- TRED [5]

REFERENCES
4. NCI Pathway Interaction Database [http://pid.nci.nih.gov]

**FI network construction: part II**

Data sources for predicted FIs:
- Human PPI
- Pfam domain interaction
- GO BP sharing
- Worm PPI
- Fly PPI
- Yeast PPI
- Mouse PPI
- Lee’s gene expression
- Prieto’s gene expression

Naïve Bayes Classifier

Annotated FIs

Predicted FIs

FI Network
Composition of the FI network (2014)

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Total FIs</th>
<th>SwissProt Ids</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathway</td>
<td>205,158</td>
<td>8,804</td>
<td>44%</td>
</tr>
<tr>
<td>Predicted</td>
<td>131,061</td>
<td>9,081</td>
<td>45%</td>
</tr>
<tr>
<td>Total</td>
<td>336,219</td>
<td>11,780</td>
<td>58%</td>
</tr>
</tbody>
</table>

Projecting Experimental Data onto FI Network
>200 Recurrently Mutated Genes in 52 Pancreatic Cancers

Module 1: Hedgehog, TGFβ signaling
Module 2: p53 signaling
Module 3: Wnt & Cadherin signaling
Module 4: Translation
Module 5: Axon guidance
Module 6: Ca2+ signaling
Module 7: ECM, focal adhesion, integrin signaling
Module 8: MHC class II antigen presentation
Module 9: Rho GTPase signaling
Module 10: Spliceosome

Functional Interaction Clustering Reveals Modular Structure
Module-Based Prognostic Biomarker in ER+ Breast Cancer

Measure levels of expression of the genes in this network module

Guanming Wu

3) Pathway-Based Modeling

• Apply list of altered {genes, proteins, RNAs} to biological pathways.
• Preserve detailed biological relationships.
• Attempt to integrate multiple molecular alterations together to yield lists of altered pathway activities.
• Pathway modeling shades into Systems Biology
## Types of Pathway-Based Modeling

- Partial differential equations/boolean models, e.g. CellNetAnalyzer
  - Mostly suited for biochemical systems (metabolomics)
- Network flow models, e.g. NetPhorest, NetworKIN
  - Mostly suited for kinase cascades (phosphorylation info)
- Transcriptional regulatory network-based reconstruction methods, e.g. ARACNe (expression arrays)
- Probabilistic graph models (PGMs), e.g. PARADIGM
  - Most general form of pathway modeling for cancer analysis at this time.

## Probabilistic Graphical Model (PGM) based Pathway Analysis

- Integrate multiple ‘omics’ data types into pathway context using PGM models
  - CNV
  - mRNA
  - Mutation, Protein, etc
- Find significantly impacted pathways for diseases
- Link pathway activities to patient phenotypes
PARADIGM: Pathways as Factor Graphs

A. Data on a single patient is integrated for a single gene using a set of four different biological entities for the gene describing the DNA copies, mRNA and protein levels, and activity of the protein. B. PARADIGM models various types of interactions across genes including transcription factors to targets (upper-left), subunits aggregating in a complex (upper-right), post-translational modification (lower-left) and sets of genes in a family performing redundant functions (lower-right). C. Toy example of a small sub-pathway involving P53, an inhibitor MDM2, and the high level process, apoptosis as represented in the model.


PARADIGM Applied to GBM Data


Fig. 8. Clustering of IPAs for TOGA GBM. Each column corresponds to a single sample, and each row to a biologic entity. Color bars beneath the hierarchical clustering tree denote clusters used for Figure 9.
PARADIGM: Good & Bad News

• Bad News
  – Distributed in source code form & hard to compile.
  – No pre-formatted pathway models available.
  – Scant documentation.
  – Takes a long time to run.

• Good News
  – Reactome cytoscape app supports PARADIGM (alpha testing).
  – Includes Reactome-based pathway models.
  – We have improved performance; working on further improvements.

Pathway/Network Database URLs

• BioGRID
  – http://www.thebiogrid.org
• IntAct
  – http://www.ebi.ac.uk/intact/
• KEGG
  – http://www.genome.jp/kegg
• MINT
  – http://mint.bio.uniroma2.it
• Reactome
  – http://www.reactome.org
• Pathway Commons
  – http://www.pathwaycommons.org
• Wiki Pathways
  – http://wikipathways.org
De novo network construction & clustering

• GeneMANIA
  – http://www.genemania.org
• HotNet
  – http://compbio.cs.brown.edu/projects/hotnet/
• HyperModules
  – http://apps.cytoscape.org/apps/hypermodules
• Reactome Cytoscape FI App
  – http://apps.cytoscape.org/apps/reactomefis

Pathway Modeling

• CellNetAnalyzer
  – http://www.ebi.ac.uk/research/saez-rodriguez/software
• NetPhorest/NetworKIN
• ARACNe
• PARADIGM
  – http://paradigm.five3genomics.com/
We are on a Coffee Break & Networking Session