Network Analysis and System Biology with Omics Data

Zhenqiu Liu
Samuel Oschin Comprehensive Cancer Institute
Cedars Sinai Medical Center
Outline

• System biology: core concepts and basic ideas
• Networks: concepts and statistics
• Correlation and causal networks in biology
• Gene set enrichment analysis and network visualization with Cytoscape
• Other software
Why Systems Biology?

The omics revolution:
• Traditional computational analysis methods and tools used in (molecular) biology were not adequate anymore.

• Omics also made clear that biology is extremely complex!
  
  We still know so very little about biology

• We do know:
  • Biological functions cannot be understood by looking at a single gene or protein.
  • Biological functions are a property of the biological system and must be studied at the systems level.
System biology and omics

Biology  |  Biotechnology  |  Bioinformatics  |  Biologist
---|---|---|---
DNA  |  Genomics  |  Data storage  |  Experiments
RNA  |  Transcriptomics  |  Data handling  |  Results
protein  |  Proteomics  |  Data preprocessing  |  Knowledge
metabolite  |  Metabolomics  |  Data analysis  |

Integrative biology or Systems biology  |  Data integration  |  Data interpretation  |

Informatics  |  ICT infrastructure  |
Systems Biology. What Is It?

- A branch of science that seeks to integrate different levels of information to understand how biological systems function.

- L. Hood: “Systems biology defines and analyses the interrelationships of all of the elements in a functioning system in order to understand how the system works.”

- It is not (only) the number and properties of system elements but their relations!!
More on Systems Biology

Essence of living systems is flow of mass, energy, and information in space and time.

The flow occurs along specific networks

- Flow of mass and energy (*metabolic networks*)
- Flow of information involving DNA (*transcriptional regulation networks*)
- Flow of information not involving DNA (*signaling networks*)

The Goal of Systems Biology:
To understand the flow of mass, energy, and information in living systems.
More on Systems Biology

What is a System:
  dynamics of its components,
  interaction of components,
  we need modeling to understand the mechanism

*human brain* can be thought by the interaction of brain cells,
*a single brain cell* is incapable of the property of thought.
Networks and the Core Concepts of Systems Biology

(i) Complexity emerges at all levels of the hierarchy of life

(ii) System properties emerge from interactions of components

(iii) The whole is more than the sum of the parts.
How to Describe a System As a Whole?

Networks - The Language of Complex Systems
Air Transportation Network
The World Wide Web
Large graphs = Networks
Networks

• Represent relationships
  – Various interactions
• Useful for discovering relationships in large data sets
  – Better than tables in Excel
• Visualize multiple data types together
  – See interesting patterns
• Network analysis
What is a Network?

Network is a mathematical structure composed of points connected by lines.

Network Theory <-> Graph Theory

- Network <-> Graph
- Nodes <-> Vertices (points)
- Links <-> Edges (Lines)

A network can be build for any functional system

System vs. Parts = Networks vs. Nodes
Network Concepts

- A network can be **connected** (presented by a single component) or **disconnected** (presented by several disjoint components).

- Networks having no cycles are termed **trees**. The more cycles thenetwork has, the more **complex** it is.
Undirected and Directed Networks

Undirected:

Directed:
Network Concepts

- Module: Modules are subsets of nodes that are tightly connected to each other.
- Clusters on the graph
Module discovery problem

- Divide a network into relatively densely connected sub-networks (Modules)
Statistical features of networks

- **Vertex degree distribution** (the degree of a vertex is the number of nodes connected with it via an edge)
Scale-free networks

most of the nodes have only a few links. A few nodes with a very large number of links, which are often called hubs, hold these nodes together. Networks with a power degree distribution are called scale-free.

It is the same distribution of wealth following Pareto’s 20-80 law: Few people (20%) possess most of the wealth (80%), most of the people (80%) possess the rest (20%)
Comparing Random and Scale-free distribution

- In the random network, the five nodes with the most links (in red) are connected to only 27% of all nodes (green). In the scale-free network, the five most connected nodes (red) are connected to 60% of all nodes (green) (source: Nature)
Attack Tolerance

- Complex systems maintain their basic functions even under errors and failures (cell → mutations; Internet → router breakdowns)
Hubs

Attacks to hubs can rapidly destroy the network
Statistical features of networks

Given a pair of nodes, compute the shortest path between them

- **Average shortest distance between two vertices**
- **Diameter**: maximal shortest distance
Shortest-Path between nodes
Maximal Shortest-Path
Six Degrees of Separation

• How many degrees of separation are they between two random people in the world, when friendship networks are considered?
• Everyone in the world is connected by at most six links
• Which path should we take?
• Shortest path by breadth first search
  – If two nodes are connected, will find the shortest path between them
• Are two proteins connected? If so, how?
• Biologically relevant?

http://www.time.com/time/techtime/200406/community.html
What we have learned?

• System: Interactions of components
• Network Concepts:
  • Directed and undirected networks
• Modules
• Hubs
• Six degrees of separation
Biological Networks

A. Intra-Cellular Networks
   Protein interaction networks
   Metabolic Networks
   Signaling Networks
   Gene Regulatory Networks
   Composite networks
   Networks of Modules, Functional Networks
   Disease networks

B. Inter-Cellular Networks
   Neural Networks

C. Organ and Tissue Networks

D. Ecological Networks

E. Evolution Network
Biological Networks (A)

- **Gene regulatory network**: two genes are connected if the expression of one gene modulates expression of another one by either activation or inhibition.
- **Protein interaction network**: proteins that are connected in physical interactions or metabolic and signaling pathways of the cell;
- **Metabolic network**: metabolic products and substrates that participate in one reaction;
Cell reproduction, metabolism, and responses to the environment are all controlled by proteins;
Each gene is responsible for constructing a single protein;
Some genes manufacture proteins which control the rate at which other genes manufacture proteins (either promoting or suppressing);
Hence some genes regulate other genes (via the proteins they create);
Protein-protein interaction networks

- Yeast PPI network
- Nodes – proteins
- Edges – interactions

The color of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown).
Protein Interaction Network

The yeast protein interaction network seems to reveal some basic graph theoretic properties:

- The frequency of proteins having interactions with exactly $k$ other proteins follows a \textit{power law}.
- The network exhibits the \textit{small world phenomena}: can reach any node within small number of hops, usually 4 or 5 hops
- \textbf{Robustness}: Resilient and have strong resistance to failure on random attacks and vulnerable to targeted attacks.
Protein-protein interaction (PPI) networks

• A *protein-protein interaction (PPI)* usually refers to a physical interaction, i.e., binding between proteins
• Can be other associations of proteins such as functional interactions – e.g., synthetic lethality
**Protein-protein interaction (PPI) networks**

- PPIs are very important for structure and function of a cell:
  - Participate in signal transduction
    - Play a role in many diseases (e.g., cancer)
  - Can be *stable interactions* forming a *protein complex*
    (a form of a quaternary protein structure, set of proteins which bind to do a particular function, e.g., ribosome, hemoglobin – illus)
Metabolic Networks

Two types of nodes: enzymes and substrates
- Reactions can be directional or bidirectional
- Bipartite graph, reactions are not connected and substrates are not connected

Berg et al. Biochemistry
New York: W. H. Freeman and Co.; c2002
Example of Gene Regulation Networks

Stem cell differentiation regulation

• Nodes are genes and transcription factors
  Interactions can be directional or bidirectional
• Interactions can be activation or inhibition

Cell Signaling Pathways

- Nodes are proteins, metabolites, lipids, second messengers, or peptides
- Interactions designate information flow, can be activation or inhibition, and are direct and physical

Network Representations

<table>
<thead>
<tr>
<th>Relationships</th>
<th>Optional weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 ↔ A2</td>
<td>1</td>
</tr>
<tr>
<td>A1 ↔ A3</td>
<td>3</td>
</tr>
<tr>
<td>A2 ↔ A3</td>
<td>1</td>
</tr>
<tr>
<td>A2 ↔ A4</td>
<td>2</td>
</tr>
<tr>
<td>A3 ↔ A4</td>
<td>1</td>
</tr>
<tr>
<td>A3 ↔ A5</td>
<td>1</td>
</tr>
<tr>
<td>A3 ↔ A7</td>
<td>1</td>
</tr>
<tr>
<td>A5 ↔ A4</td>
<td>1</td>
</tr>
<tr>
<td>A5 ↔ A6</td>
<td>1</td>
</tr>
<tr>
<td>A6 ↔ A8</td>
<td>1</td>
</tr>
<tr>
<td>A6 ↔ A9</td>
<td>2</td>
</tr>
<tr>
<td>A8 ↔ A9</td>
<td>3</td>
</tr>
</tbody>
</table>

Network:
- Node A1 connected to A2
- Node A2 connected to A3 and A4
- Node A3 connected to A4, A5, and A7
- Node A4 connected to A5
- Node A5 connected to A6
- Node A6 connected to A8
- Node A7 connected to A8, A9

Heatmap:
- Colors represent the weight of connections.
- Node A3 connected to A2, A1, A7, A4, A5, A6, A8, A9.
- Edge weights and node degrees are shown in the network.
Graphs: Adjacency Matrix

• Example:

```
A | 1  2  3  4  
---|---|---|---|---|
1 | 0  5  6  0  
2 | 5  0  9  0  
3 | 6  9  0  4  
4 | 0  0  4  0  
```

Weighted graph
Obtaining biological networks

• Direct experimental methods
  – Protein-protein interaction networks
    • Yeast-2-hybrid
    • Tandem affinity purification
    • Co-immunoprecipitation
  – Protein-DNA interaction
    • Chromatin Immunoprecipitation (followed by microarray or sequencing, ChIP-chip, ChIP-seq)
  – High level of noises (false-positive and false-negative)

• Computational prediction methods
  – Often cannot differentiate direct and indirect interactions
Why networks?

- Studying genes/proteins on the network level allows us to:
  - Assess the role of individual genes/proteins in the overall pathway
  - Evaluate redundancy of network components
  - Identify candidate genes involved in genetic diseases
  - Sets up the framework for mathematical models

For complex systems, the actual output may not be predictable by looking at only individual components:

The whole is greater than the sum of its parts
What can we do with these molecular networks?
Using the position in networks to describe function

Guilt by association

Finding the causal regulator (the "Blame Game")

Courtesy of Mark Gerstein
Networks and Disease

- **Identification of disease Associatee subnetworks** – identification of disease subnetworks that are transcriptionally active in disease.

- **Subnetwork-based diagnosis** – source of biomarkers for disease classification, identify interconnected genes whose aggregate expression levels are predictive of disease state.

- **Subnetwork-based gene association** – map common pathway mechanisms affected by collection of genotypes.
Examples: Predict Metastasis

• If metastasis is likely => aggressive adjuvant therapy
  – How to decide the likelihood?

• Traditional predictive factors are not good
Recently: Gene Marker Sets

• Examine genome-wide expression profiles
  – Score individual genes for how well they discriminate between different classes of disease
    • Establish gene expression signature
  – Problem: # genes >> # patients
Pathway Expression vs. PPI Subnetwork as Marker

• Score known pathways for coherence of gene expression changes?
  – Majority of human genes not yet assigned to a definitive pathway

• Large Protein-Protein Interaction networks recently became available
  – Extract subnetworks from PPI networks as markers
Subnetwork Marker Identification: Data Used

- 2 separate cohorts of breast cancer patients
  - van 't Veer et. al, and Wang et. al.
  - Roughly half had developed metastasis
- Used Protein-Protein Interaction network obtained by assembling a pooled dataset
  - 57,235 interactions among 11,203 proteins
Goal: Find Significantly Discriminative Subnetworks

- Use a scoring system to search for subnetworks highly discriminative of metastasis
Results: Correspondence to hallmarks of cancer

- For two datasets of 295 and 286 patients, 149 and 243 (resp.) discriminative subnets found
- 47% and 65% of subnets enriched for common biological process
- 66 and 153 subnets were enriched for processes involved in major events of cancer progression
Network Construction from microarray data

Construct Co-expression network

vs

Clustering
Biological Causal Networks

- Gene regulation:
  - Activators increase gene production
    \[ X \rightarrow Y \]
  - Repressors decrease gene production
    \[ X \rightarrow^\uparrow Y \]
Biological Networks

• Gene regulation:
  – Negative feedback loop:
  – Positive feedback loop:
Biological Networks

- Nodes are proteins (or genes)
A gene regulatory network can be represented by a directed graph;

- **Node** represents a gene;
- **Directed edge** stands for the modulation (regulation) of one node by another:
  - e.g. arrow from gene X to gene Y means gene X affects expression of gene Y
Learning Causal Relationships

• High-throughput genetic technologies empowers to study **how genes interact with each other**;

• Learning gene causal relationship is important:
  – Turning on a gene can be achieved directly or through other genes, which have causal relationship with it.
Causality vs. Correlation

Example: *rain* and *falling_barometer*

- Observed that they are either *both* true or *both* false, so they are related. Then write

\[
    \text{rain} = \text{falling\_barometer}
\]

- Neither *rain* causes *falling\_barometer* nor vice-versa.
- Thus if one wanted *rain* to be true, one could not achieve it by somehow forcing *falling\_barometer* to be true. This would have been possible if *falling\_barometer* caused *rain*.

- We say that the relationship between *rain* and *falling\_barometer* is *correlation*, but *not cause*.
Learning Causal Relationship with Steady State Data

• How to infer causal relationship?
  – In wet-labs, knocking down the possible subsets of a gene;
  – Use time series gene expression data;

• Problem?
  – Human tissues gene expression data is only available in the steady state observation;

• (IC) algorithm by Pearl et al to infer causal information but not in biological domain;
One example of Network Constructed

Pirin causatively influences WNT5A – “In order to maintain the level of WNT5A we need to directly control WNT5A or through pirin”.

Causal connection between WNT5A and MART-1 “WNT5A directly causes MART-1”
Example of Network Construction

• A list of biological important genes
• 143 genes corresponding to IBD associated SNPs
• Genes are put into STRING and DAVID
• Combining PPI and molecular function together
Validation with knockout signatures

• In each experiment, one gene was knocked-out, and the expression levels of the remainder genes in control and knocked-out strains were interrogated for differential expression.

• The set of differentially expressed genes form the knock-out signature (ko-signature) of the knocked-out gene (ko-gene).

• The ko-signature represents a validated set of causal relations.
Pathways

• Pathways are subsets of networks,
• Pathways are networks of interactions,
• Pathways are related to a known physiological process or complete function.
• Example:
  KEGG pathways
One Example
Pathway Databases

- KEGG (Kyoto Encyclopedia of Genes and Genomes)
  - Institute for Chemical Research, Kyoto University
- PathDB
  - National Center for Genomic Resources
- SPAD: Signaling PAtthway Database
  - Graduate School of Genetic Resources Technology. Kyushu University.
- Cytokine Signaling Pathway DB.
  - Dept. of Biochemistry. Kumamoto Univ.
- EcoCyc and MetaCyc
  - Stanford Research Institute
- BIND (Biomolecular Interaction Network Database)
  - UBC, Univ. of Toronto
What we have learned?

• Biological networks and subnetworks
• Subnetworks and phenotype associations
• Causal networks in biology.
• Pathways
Network Analysis using Cytoscape

Find biological processes underlying a phenotype

Network Analysis

Network Information

Databases

Literature

Expert knowledge

Experimental Data
Cytoscape

http://www.cytoscape.org

• Help
  – Tutorials, case studies
  – Mailing lists for discussion
  – Documentation, data sets

• 10,000s users, 5000 downloads/month

• >160 Plugins/Apps Extend Functionality
  – The app store: http://apps.cytoscape.org/
  – Build your own, requires programming
Network Software Tools

- Pajek
- Cytoscape
  - General-purpose, open-source software environment for the large scale integration of molecular interaction network data
  - Network visualization (a variety of automated network layout algorithms)
  - Links a network to molecular interaction and functional databases (transfer of annotations)
  - Data integration
  - ...

- GraphCrunch
- tYNA
- FANMOD
- mFinder
Gene Set Enrichment Analysis (GSEA)

Ranked Gene List

UP (A > B)

DOWN (B > A)

Gene-sets

Enrichment in Condition A vs. B

<table>
<thead>
<tr>
<th>Gene-set</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Cycle</td>
<td>0.0001</td>
</tr>
<tr>
<td>EGF Pathway</td>
<td>0.003</td>
</tr>
<tr>
<td>Spindle</td>
<td>0.007</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Enrichment in Condition B vs. A

<table>
<thead>
<tr>
<th>Gene-set</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteasome</td>
<td>0.0002</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>0.005</td>
</tr>
<tr>
<td>Caspase</td>
<td>0.009</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Enrichment Map

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA1</td>
<td>+7.5</td>
</tr>
<tr>
<td>ABA2</td>
<td>+5.3</td>
</tr>
<tr>
<td>BCD3</td>
<td>+4.1</td>
</tr>
<tr>
<td>FDE4</td>
<td>-3.7</td>
</tr>
<tr>
<td>ADA1</td>
<td>-3.0</td>
</tr>
<tr>
<td>CGE6</td>
<td>-5.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene-set</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Cycle</td>
<td>0.0001</td>
</tr>
<tr>
<td>EGF Pathway</td>
<td>0.003</td>
</tr>
<tr>
<td>Spindle</td>
<td>0.007</td>
</tr>
<tr>
<td>Proteasome</td>
<td>0.0001</td>
</tr>
<tr>
<td>Apoptosis</td>
<td>0.003</td>
</tr>
<tr>
<td>Caspase</td>
<td>0.007</td>
</tr>
</tbody>
</table>

Overlap

$|A \cap B| / \min(|A|,|B|)$
Software for gene enrichment analyses

There are many programs available for annotating modules (or set of genes):

• EASE: http://david.abcc.ncifcrf.gov/ease/ease1.htm
• ToppFun: http://toppgene.cchmc.org/
• ChiliBot: http://www.chilibot.net/
• WebGestalt: http://bioinfo.vanderbilt.edu/webgestalt/
• Ingenuity: http://www.ingenuity.com/
• GSEA: http://www.broadinstitute.org/gsea/index.jsp
• UGET: http://genome.ucla.edu/projects/UGET
• STRING: http://string.embl.de/
• Galaxy: https://main.g2.bx.psu.edu/
Protein Databases

- **Swiss-Prot** (non-redundant database):
  - **TrEMBL** (translations of EMBL nucleotide sequences
    - not yet integrated into Swiss-Prot):

- The number of entries keeps rapidly growing mainly due to large scale sequencing projects.
Protein Interaction Databases

- **Species-specific**
  - FlyNets - Gene networks in the fruit fly
  - MIPS - Yeast Genome Database
  - RegulonDB - A Database On Transcriptional Regulation in E. Coli
  - SoyBase
  - PIMdb - Drosophila Protein Interaction Map database

- **Function-specific**
  - Biocatalysis/Biodegradation Database
  - BRITE - Biomolecular Relations in Information Transmission and Expression
  - COPE - Cytokines Online Pathfinder Encyclopaedia
  - Dynamic Signaling Maps
  - EMP - The Enzymology Database
  - FIMM - A Database of Functional Molecular Immunology
  - CSNDB - Cell Signaling Networks Database
Protein Interaction Databases

- **Interaction type-specific**
  - DIP - Database of Interacting Proteins
  - DPIInteract - DNA-protein interactions
  - Inter-Chain Beta-Sheets (ICBS) - A database of protein-protein interactions mediated by interchain beta-sheet formation
  - Interact - A Protein-Protein Interaction database
  - GeneNet (Gene networks)

- **General**
  - BIND - Biomolecular Interaction Network Database
  - BindingDB - The Binding Database
  - MINT - a database of Molecular INTeractions
  - PATIKA - Pathway Analysis Tool for Integration and Knowledge Acquisition
  - PFBP - Protein Function and Biochemical Pathways Project
  - PIM (Protein Interaction Map)
Conclusions and Acknowledgement

• System Biology and network analysis
• Relatively new and very exciting area
• Need to understand the main concepts
• Software, software, software
• Need help of a bioinformatician

Some slides were adapted from Dr. Bader, UT