**Roadmap**

- Cellular networks
  - Co-Expression networks
  - Gene Regulatory networks
  - Protein-protein interaction networks
  - Metabolic networks
- Analysis of networks
- Random models
- Some computational challenges
Relationship between networks

Projections over phenomenological gene network models of the whole biochemical system. [Brazhnik et al, 2002 TIBS]

Co-expression network (yeast)

1729 nodes, 4000 edges
[Magwene and Kim, Genome Biology 2004]
Co-expression networks

- Nodes of co-expression networks are genes
- Undirected edges
- Add edge (A,B) if gene A and B have similar mRNA expression profiles (according to some chosen distance)

\[ A \quad B \]

- Also called gene expression networks

Co-expression networks

- Collect mRNA abundance for a variety of conditions, tissues, etc.
- All pair of genes are considered, and a measure of association is computed on each pair
- Distance measures include Spearman/Pearson's correlation and mutual information
- Pairwise interactions higher than a threshold are considered significant and are represented by edges
Co-expression networks
Regulatory network (yeast)

Illustration of the interactions between transcription factors and promoters of other transcription factor genes as determined by ChIP analysis. Bold arrows indicate an interaction that has been verified by PCR assay; blue boxes correspond to genes that have a cell cycle periodicity of expression.

[Horak et al, Genes & Development, 16:3017-3033, 2002]

Regulatory networks

• Nodes of regulatory networks are genes
• Directed graph
• Edge (A,B) if gene A is known to regulate gene B

• Also called transcription factor network
Genetic “circuits”

Regulatory networks

- Gene regulatory networks can be constructed directly from ChIP-Seq experiments
- A chromatin immunoprecipitation (ChIP) assay determines whether proteins (including transcription factors) bind to a particular region on the chromatin
- Or indirectly from gene expression data
Relation between regulatory networks and expression networks

Protein-protein interaction network

Red = lethal
Green = non-lethal
Yellow = unknown
Protein-protein interaction network

- Nodes of PPI networks are proteins
- Undirected edges
- Add edge (A,B) if protein A and B are known to interact (bind/dock/etc.)

Protein-protein interaction network

- Protein interaction data can be obtained by
  - yeast two-hybrid
  - coimmunoprecipitation (ChIP)
  - in vitro binding assays
  - “mining” pub-med abstracts
  ... and more
Yeast two-hybrid assay

• Take a pair (X,Y) of proteins of interest
• Fuse protein X to a DNA-binding domain
• Fuse Y to a transcription activator domain
• If the pair interacts, then a functional transcription factor will form
• Measure interaction by checking for the presence of a reporter gene’s transcript

Yeast two-hybrid assay

• Benefits
  • *In vivo*
  • Is independent of endogenous protein levels
  • High-throughput methods available
• Drawbacks
  • Can only test two proteins at a time
  • Proteins are not in their native environment
Protein function

• PPIs can help us discover the functions of unknown proteins
• If the function of one protein is known, the function of its binding partners are likely to be related ("guilt by association")
• Thus, having a good method for detecting interactions can allow us to use a small number of proteins with known function to characterize the function of new proteins
• Hubs – essential?

Relation between regulatory networks, expression networks, and PPI networks
**Metabolic networks**

- A metabolic network represents the complete set of metabolic and physical processes that determine the physiological and biochemical properties of a cell.
- Nodes are metabolites (amino acids, sugars and lipids, and the energy required to synthesize them and to use them in creating proteins and cellular structures).
- Directed edge \((A,B)\) if metabolite \(A\) transforms in (or affects) metabolite \(B\) through a chemical reaction.

---

**Metabolic networks**

- Metabolism is a step-by-step modification of the initial molecule to shape it into another product. The result can be used in one of three ways:
  - Stored by the cell
  - Used immediately, as a metabolic product
  - Initiate another metabolic pathway, called a flux\(^*\) generating step

\(\ast\) rate of turnover of molecules through a metabolic pathway
**Metabolic networks**

A → B → C → E

- Active reaction
- Inactive reaction (very low or zero flux)

**Metabolic networks**

A → B → C → E

- Exchange flux
- Internal flux
- System Boundary

Flux – The production or consumption of mass per unit area per unit time
Metabolic network (yeast)

- Changes in metabolism as diauxic shift proceeds
  - red genes are the ones turned on
  - green ones are the ones turned off

[DeRisi, Iyer, and Brown, Science, 278:680-686]

KEGG (Metabolic network database)
Relation between regulatory networks, expression networks, PPI and metabolic networks

Summary

- **Co-Expression networks**
  Vertices: genes
  Edges: co-expression

- **Gene regulatory networks**
  Vertices: genes
  Edges: regulatory influences

- **Metabolic networks**
  Vertices: metabolites
  Edges: reactions, consumption, production
  Degenerate networks: enzymes, metabolites

- **Protein-protein interaction networks**
  Vertices: proteins
  Edges: physical interactions
Basic graph (network) concepts

• Directed/undirected
• Weighted/unweighted
• (Strongly) connected/disconnected
• We deal with networks that have at most one edge/link between two nodes
• $n =$ number of nodes
• $m =$ number of edges/links
Network metrics

- Degree distribution (scale-free exponent)
- Characteristic path length (small-world)
- Clustering coefficient
- (Betweenness) centrality

Degree

- Most elementary characteristic of a node
- In directed networks there is a in-degree and a out-degree
- An undirected network with $n$ nodes and $m$ edges (links, arcs) has average degree $\langle k \rangle = 2m/n$
- Nodes with high degree are called hubs
Degree distribution

• $P(k)$ = degree distribution
  [Frequency of nodes with degree $k$]

• Scale-free network when $P(k)$ follows a power law

Scale-free networks

• In scale-free networks most of the nodes have small degree, but a few nodes with very large degree (hubs) hold the network connected
• The smaller is $\gamma$, the more important the role of the hubs in the network
• For most biological networks, the value of $\gamma$ is in (2,3)
Degree Distribution of yeast PPI (2005)

CCDF of degree distribution for yeast PPI (2005)
**Definition of shortest path**

- Given two nodes \( u, v \) in the network, there could be multiple paths (or none) from \( u \) to \( v \)
- The *shortest path* on an unweighted network is the path that traverses the smallest number of edges
- The *shortest path* on a weighted network is the path that has the smallest total weight on the edges traversed
- In directed graph, the shortest path \( u, v \) is in general different from the shortest path \( v, u \)

**Characteristic path length**

- \( L_{(u,v)} \) is the length of the shortest path between vertices \( u \) and \( v \)
- The average characteristic path length \( <L> \) of a graph is the average of the \( L_{(u,v)} \) for all pairs \( u, v \)
- Networks with small values of \( <L> \) are said to have the *small world property* (navigability)
Clustering coefficient

- In many networks, if node $u$ is connected to node $v$, and node $v$ is connected to node $z$, then is likely that $u$ is connected to $z$. In other words, $u, v, z$ form a triangle.
- The extent of this phenomenon can be quantified using the clustering coefficient.
- The clustering coefficient of node $v$ is the ratio between the number of triangles that go through node $v$ and the total number of triangles that could pass through node $v$ (should all neighbors of $v$ be connected to each other).

\[
C_v = \frac{2n_v}{k_v(k_v-1)}
\]

where $k_v$ is the degree of node $v$, and $n_v$ is the number of neighbor pairs of $v$ that are connected.
Roadmap

- The average clustering coefficient $<C>$ characterizes the overall tendency of nodes to form clusters or groups
- $C(k)$ is defined to be the average clustering coefficient for nodes with degree $k$
- For many networks $C(k) \sim 1/k$ which indicates a network with hierarchical characteristics

Statistics on PPI networks (2005)

Table 1. Statistic summary: $n$ is the number of vertices, $m$ is the number of edges, $\bar{k}$ is the average degree, $L$ is the characteristic path length, and $C$ is the clustering coefficient.

<table>
<thead>
<tr>
<th></th>
<th>Yeast</th>
<th>Fly</th>
<th>AS990220</th>
<th>skitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>4687</td>
<td>6926</td>
<td>4686</td>
<td>9200</td>
</tr>
<tr>
<td>$m$</td>
<td>15138</td>
<td>20745</td>
<td>8772</td>
<td>28957</td>
</tr>
<tr>
<td>$\bar{k}$</td>
<td>6.4596</td>
<td>5.9905</td>
<td>3.7439</td>
<td>6.2927</td>
</tr>
<tr>
<td>$L$</td>
<td>4.18519</td>
<td>4.45931</td>
<td>3.72621</td>
<td>3.118</td>
</tr>
<tr>
<td>$C$</td>
<td>0.126</td>
<td>0.0154</td>
<td>0.3786</td>
<td>0.6212</td>
</tr>
</tbody>
</table>
Centrality

• There are various measures of the centrality of a node that determine the relative its importance within the graph
  • Degree centrality
  • Betweenness centrality
  • Closeness centrality
  • Eigenvector centrality

Betweenness centrality

• The betweenness of a node is defined as the number of shortest paths running through it among all shortest paths between all pairs of vertices in a graph

Hue (from red=0 to blue=max) shows the node betweenness
Random graphs models

- Erdös-Renyi (1960) ‘random-random’
- Barabasi-Albert (1999) ‘scale-free’
- Hierarchical (2002)

Random-random model properties

- For large $n$, there is a giant connected component if the average connectivity (number of links per node) is larger than one
- Degree distribution $P(k)$ is Poisson distributed (most nodes have approximately the same number of links $<K>$, and nodes that significantly deviate from the average are extremely rare)
- The clustering coefficient $C(k)$ is independent from the node degree
- The average path length $<L>$ in the giant component scales as $<L> \approx \log n$
Small-world [Watts-Strogatz 1998]

- Input: number of nodes $n$, mean degree $K$ (even), parameter $p$ in $[0,1]$
- The model constructs an undirected graph with $n$ nodes and $nK/2$ edges in the following way
  - Construct a regular ring lattice, with $n$ nodes each connected to $K$ neighbors, $K/2$ on each side
  - For every node $n_i$ consider every edge $(n_i, n_j)$ with $i < j$, and rewire it with probability $p$. Rewiring is done by replacing $(n_i, n_j)$ with $(n_i, n_k)$ where $k$ is chosen with uniform probability from all possible values that avoid loops and link duplication

Small-world [Watts-Strogatz 1998]

- Start with a regular network with $n$ vertices
- Rewire each edge with probability $p$

For $p=0$ (Regular Networks)
- high clustering coefficient
- high characteristic path length

For $p=1$ (Random Networks)
- low clustering coefficient
- low characteristic path length

What happens for intermediate values of $p$?
Watts-Strogatz model properties

- The degree distribution in the case of the ring lattice is just a Dirac delta function centered at K. It is the limiting case of a Poisson distribution.
- The average path length \( \langle L \rangle \) when \( p=0 \) is \( n/2K \)
- In the limiting case of \( p \to 1 \) the graph converges to a classical random graph with \( \langle L \rangle = \log n / \log K \)
**Scale-free** [Barabasi & Albert 1999]

- **Input:** $n$, $m_0$, $m$
- **Construction of a scale-free graph (preferential attachment)**
  - Construct a complete graph with $m_0 > 2$ nodes
  - Repeat the following $n-m_0$ times
    - Create a new node $v$
    - Select $m$ nodes from the current graph
    - Connect $v$ to these $m$ nodes, where each of $m$ nodes is selected randomly according to the probability proportional to its degree [i.e., $\text{deg } v_i / \sum_j \text{deg } v_j$]

**Random network vs. Scale-free network**

Random network

$$P(k) = e^{\lambda k} k!$$

Scale-free network

$$P(k) \propto k^{-\gamma}$$
Random-random vs. Scale-free

(a) Random-random graph with \( N = 10 \) and \( p = 0.2 \)
(b) A new node (red) connects to two existing nodes in the network (black) at time \( t+1 \). This new node is much more likely to connect to highly connected nodes, a phenomenon called preferential attachment.
(c) The network connectivity can be characterized by the probability \( P(k) \) that a node has \( k \) links. For random graphs, \( P(k) \) is strongly peaked at \( k = \langle k \rangle \) and decays exponentially for large \( k \).
(d) A scale-free network does not have a peak in \( P(k) \), and decays as a power law \( P(k) \sim k^{-\gamma} \) at large \( k \).
(e) A random network - most nodes have approximately the same number of links.
(f) Most nodes in a scale-free network have one or two links, but a few nodes have many links; this guarantees that the system is fully connected.

Scale-free [Barabasi & Albert 1999]

- Degree distribution \( P(k) \approx k^{-3} \) (the probability that a node is highly connected is higher than in a random graph – presence of hubs).
- The average path length scales as
  \[ L \approx \log \log n \]
- Clustering coefficient (experimentally)
  \[ C \approx n^{-0.75} \] [\( C(k) \) is independent from \( k \)]
**Hierarchical Model** (Barabasi et al. 2002)
- Deterministic construction of scale-free network
- Fractal (self-similarity)
- But, $\gamma < 2.58$

*Geometric models*

- Proposed as a random model for PPI by (Przulj et al.)
- Set up a $k$-dimensional hypercube $[0,1]^k$
- Generate $n$ nodes as points randomly distributed in the hypercube
- Connect two nodes $i,j$ if their Euclidean distance is below a threshold $T$
- Strong agreement with real PPI using graphlet counts
Data mining on biological networks

- Clustering graphs (finding “communities”)
- Discovering unusually frequent subgraphs (network motifs)
- Discovering (statistically significant) paths
- Discovering highly-connected components (quasi-cliques)
- Integrative (i.e., multi-network) approaches
- Comparative (i.e., multi-species) approaches
Graph Clustering

Highly-connected subgraphs

Spirin, Mirny, PNAS 100, 12123 (2003)
Network motifs

- Enumerate all small size motifs in a network
- Compare the motif frequencies with what is expected by chance (using a random model)
- Identify motifs which are under or over-represented
- What is the biological meaning of observed patterns?

Understanding the dynamic (System Biology)

- Formal analysis using System theory to characterize stability (or other properties) of the network
- Simulations
  - Network ↦ State transition diagram
  - Hypothesis: “A cell type is an attractor in gene expression state space”
  - Be careful not to write the desired behaviors into the simulation
In-silico Biology?