Analysis and visualization of protein–protein interactions

Olga Vitek
Assistant Professor
Statistics and Computer Science
Outline

1. Protein–protein interactions
2. Using graph structures to study protein–protein interactions
3. Clustering of graphs
4. Evaluation of clusters
Life begins with Cell

- A cell is a smallest structural unit of an organism that is capable of independent functioning
- All cells have some common features
DNA, RNA, and the Flow of Information

This model is known as the “central dogma”
Why should we study proteins?

- **Proteins**: large molecules made up of amino acids
  - accomplish most of the function of the living cells
    - by interacting (i.e. entering in physical contact) with other molecules
  - linear structures fold into 3–dimensional shapes
    - the structure is used to accomplish the function
Proteins accomplish function by forming complexes

- A protein complex is a group of tightly interacting proteins
  - also called functional module
  - protein interactions within the complex help accomplish its function

- Example: exosome
  - a complex of 11 proteins
  - degrades RNA molecules
  - ring structure ensures the function

Discovery of the complex helps understanding its function

- Example: exosome
  - first discovered in yeast
  - helped discovering an equivalent complex in humans
  - has clinical implications
    - target of autoimmune disease
    - chemotherapies for cancer block its activity

- Knowledge of protein complexes speeds up biological and clinical research
Complexity of a bacterial cell

Often study simpler “model” organisms to gain insight into the function of the cell
Modern technologies determine protein interactions on large scale

- New terms
  - **Proteome**: all proteins that exist in an organism
  - **Interactome**: all protein–protein interactions

- New questions
  - Interactions of individual proteins
  - Network-wide patterns of interactions

- New challenges
  - Large, complex, noisy datasets
  - Computational approaches are key
New technologies determine protein–protein interactions on a large scale

Ho et al., Nature, 2002

Such datasets are being increasingly produced, and are publicly available
Outline

1. Protein–protein interactions
2. Using graph structures to study protein–protein interactions
3. Clustering of graphs
4. Evaluation of clusters
Representing protein–protein interactions using graphs

Interaction attributes:
- type
- confidence
- direction

experimentally determined interaction

protein A

protein B

Protein attributes:
- name
- function
- quantitative data

Protein attributes:
- name
- function
- quantitative data
Representing protein–protein interactions using graphs

Interaction attributes:
• type
• confidence
• direction

experimentally determined interaction

protein A

protein B

Protein attributes:
• name
• function
• quantitative data

Protein attributes:
• name
• function
• quantitative data

Experimental artifact
Use graphs to represent the large-scale information on proteins, interactions and their attributes.
Graph-based representation of protein–protein interactions

- View data as a graph
  - Proteins are nodes and interactions are edges
  - Nodes have attributes
    - e.g. known function
  - Directed edges
    - experimental artifact
The interactions are determined by tag–affinity purification (TAP)

- A protein (“bait”) is labeled by a chemical
- The bait forms its interactions (collects “prey”)
- The bait, and all other proteins in the complex are isolated
- All components of the complex are identified by mass spectrometry

Kumar & Snyder, Nature, 2002
The interactions are determined by tag–affinity purification (TAP)

1. Tagging
2. Purification
3. Identification
The technology yields false positive and false negative interactions

- Can not distinguish between various types of complexes
  - chain
  - star
  - complete graph

- Use the “spoke” model to represent results of experiments
  - directed edges from “bait” to “prey”
  - multiple proteins in a complex can be used as a bait
    - direction of edges reflects experimental design, but not the underlying biology
Global graph–based summaries: degree of a node

Degree of a node: the number of edges that the node has to other nodes

- **degree distribution**: fraction of nodes in the network with a different degree
- **mean degree**: average degree over all nodes

Each node is labeled with its degree

http://en.wikipedia.org/wiki/Degree_(graph_theory)
Degree distribution: Gavin et al., 2002

- Only a few nodes have a large number of edges
Global graph-based summaries: clustering coefficient

- **Clustering coefficient of a node:** the fraction of the neighbors of a node that are also neighbors
- **Clustering coefficient of a network:** average clustering coefficient over all neighbors

http://en.wikipedia.org/wiki/Clustering_coefficient
Mean degree vs clustering coefficient of experimental networks

Two technologies:
- tag affinity purification
- yeast-two-hybrid
Conclusion from these summaries for protein interaction networks:

- Most nodes have a low degree (i.e. few neighbors)
- Some nodes have a high clustering coefficient (i.e. their neighbors are also neighbors)
- Of interest are protein clusters (i.e. groups of proteins that interact with each other more closely than outside the group)
  - close interactions can help infer biological function
  - challenge: large and noisy datasets
Outline

1. Protein–protein interactions

2. Using graph structures to study protein–protein interactions

3. Clustering of graphs

4. Evaluation of clusters
Our goal: find protein clusters in the large and noisy interaction graph

Gavin et al., Nature, 2002
Step 1: “de-noise” the interaction graph

- We are more confident in protein interactions if they are determined using multiple baits
  - remove isolated subgraphs
  - determine connected components
    - subgraphs where there is a directed path from each protein to every other protein

Gavin et al., Nature, 2002
Step 1: “de-noise” the interaction graph

- We are more confident in protein interactions if they are determined using multiple baits
  - remove isolated subgraphs
  - determine connected components
    - subgraphs where there is a directed path from each protein to every other protein

Gavin et al., Nature, 2002
Step 2: based on the graph topology, find protein clusters in the connected components

- Finding clusters
  - ignore directions of edges
  - use Markov Cluster (MCL) algorithm for clustering
- The output are sets of closely interacting proteins
- Not every protein is expected to cluster

Gavin et al., Nature, 2002
Step 2: based on the graph topology, find protein clusters in the connected components.

Output of a clustering procedure

Exosome example: additional proteins were found by clustering the network.

Gavin et al., Nature, 2002

Gavin et al., Nature, 2006