Analyzing and visualizing 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. Function–based 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
    - accomplish their function 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

- **Proteome**: an ensemble of proteins that exist in an organism

- **Interactome**: an ensemble of all protein–protein interactions

Ubiquitin–conjugating enzyme E2 G1 (PDB entry 2AWF)
Complexity of a bacterial cell

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

Such datasets are being increasingly produced, and are publicly available
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
- A large data set
  - hard to visualize
  - hard to analyze by hand
- Computational methods are key
  - cytoscape
  - graph clustering
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

- Finding “cohesive components” (i.e. clusters) in the protein interaction graph helps understand protein function

Discovery of the complex helps biological and clinical research

- 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 this research
  - different complexes are formed under different biological conditions
Outline

1. Protein–protein interactions
2. Using graph structures to study protein–protein interactions
3. Function–based evaluation of clusters
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 technology yields false positive and false negative interactions

- Can not distinguish between various types of complexes

- 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
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

- Proteins within a complex interact with each other more closely than between complexes
  - a number of algorithms searches for “cohesive components” of the graph
  - Markov Cluster (MCL) is one such algorithm
  - directions of edges are ignored at this stage
- The output are sets of closely interacting proteins

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

- Proteins within a complex interact with each other more closely than between complexes
  - a number of algorithms searches for “cohesive components” of the graph
  - Markov Cluster (MCL) is one such algorithm
  - directions of edges are ignored at this stage
- The output are sets of closely interacting proteins

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

- Proteins within a complex interact with each other more closely than between complexes
  - a number of algorithms searches for “cohesive components” of the graph
  - Markov Cluster (MCL) is one such algorithm
  - directions of edges are ignored at this stage
- The output are sets of closely interacting proteins

Protein cluster for the exosome complex

Gavin et al., Nature, 2006
Outline

1. Protein–protein interactions
2. Using graph structures to study protein–protein interactions
3. Function–based evaluation of clusters
Quality of clusters can be evaluated based on the available biological knowledge

- Proteins form complexes to perform a biological function
- A good clustering algorithm clusters together proteins with a similar biological function
  - if we can uncover known protein clusters, we can better trust new clusters that we discover

Do these proteins have a similar biological function?
Functional similarity between proteins can be quantified using specifically designed metrics

- Databases and literature contain description on protein function
  - usually in form of text or ontologies

- Can translate functional description in a number quantifying functional similarity between two proteins

- Lin introduced one such metric
  - $1 = \text{similar, } 0 = \text{different}$
    - pre-calculated for the project in form of matrices

- Can compare values of the metric within and between clusters
  - a good clustering procedure will have high values within a cluster, and low between clusters

D. Lin, 15th Int. Conf. on Machine Learning, 1998
Distribution of values of a similarity metric can be visualized using a boxplot.

Boxplot of similarity metrics

Clustering result

Gavin et al., Nature, 2002
Distribution of values of a similarity metric can be visualized using a boxplot.

Boxplot of similarity metrics

Within-cluster  Between-clusters

Least similar

Most similar

Each edge contributes one similarity value

Clustering result

Gavin et al., Nature, 2002
Distribution of values of a similarity metric can be visualized using a boxplot.

Each edge contributes one similarity value.

Boxplot of similarity metrics

Clustering result

Gavin et al., Nature, 2002
Distribution of values of a similarity metric can be visualized using boxplots

![Boxplot Diagram]

**Conclusion:**
- In this dataset, proteins within clusters have a higher pairwise similarity than between clusters

Gavin et al., Nature, 2002
Can detail the distribution of within-cluster similarity for each cluster.

Gavin et al., Nature, 2002
Bonus

- How is the functional similarity metric by Lin calculated?
  - similarity values are pre-calculated for the project
    - this material is not required to carry out the project
The Lin similarity metric is calculated using publicly available database Gene Ontology (GO)

- GO is a set of structural vocabularies
  - describe various aspects of what is known about a molecule in a cell
  - has three vocabularies:
    - (i) molecular function (MF) of a molecule
    - (ii) the broader biological process (BP) that the molecule is involved in;
    - (iii) the cellular compartment (CC) that the molecule acts in

- Structured as a directed acyclic graph
  - children terms have more specific information regarding a molecule than the parent
  - species-independent

Yet another use of graphs!
Graphical representation of GO

- Example: exosome
  - Create the Cellular compartment (CC) vocabulary induced by proteins in the exosome complex

  more specific terms are on top of the graph

- The semantic similarity metric between two proteins is a measure of the specificity of GO terms shared between the two proteins

  Protein IDs in the exosome complex → Gene IDs → All GO CC ids mapped to these genes in the literature
Since exosome exists in both nucleus and cytoplasm, its functional description can be more detailed.

The semantic similarity metric between two proteins is a measure of specificity of GO terms shared between the two proteins.
Definition of semantic similarity between two proteins

- For a protein $c$, define its probability $p(c) = \frac{freq(c)}{N}$
  i.e. # proteins mapped to its the most specific GO term / total number of proteins in the dataset

- For two proteins $c_1$ and $c_2$, define the probability of minimal subsumer $p_{ms}(c_1, c_2) = \min_{c \in S(c_1, c_2)} \{p(c)\}$
  i.e. # proteins mapped to the most specific GO term shared between the proteins / total number of proteins in the experiment

- Then the Lin similarity metric is defined as
  $$\text{sim}(c_1, c_2) = \frac{2 \times [\ln p_{ms}(c_1, c_2)]}{\ln p(c_1) + \ln p(c_2)}$$
  - it takes into account both the GO information, and the proteins typically observed in the dataset