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
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
Details on the MCL algorithm

- Think of edges as water pipes
  - information circulates through edges in a flow
  - iteratively remove edges which contain little information flow
  - edge removal is controlled by a granularity parameter

- Consequences of the algorithms
  - non-overlapping clusters
  - deterministic cluster membership of nodes

http://www.micans.org/mcl/
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 used to evaluate the quality of clusters

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

- The specific tasks are:
  - link each protein to its functional annotation
  - check if the pair shares functional description
    - can conduct statistical tests whether the overlap is larger than what is expected at random
    - translate functional descriptions in a number that quantifies functional similarity
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

Protein IDs in the exosome complex ➔ Gene IDs ➔ All GO CC ids mapped to these genes in the literature

  ▪ 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

Gene IDs

Protein IDs in the exosome complex

exosome (RNase complex) ➔ intracellular part ➔ protein complex ➔ intracellular ➔ cell part ➔ cell ➔ cellular_component
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.
Lin metric 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 dataset

- 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

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.

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 a boxplot.

Boxplot of similarity metrics

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

In this dataset, proteins within clusters have a higher pairwise similarity than between clusters

Conclusion:

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

Gavin et al., Nature, 2002