341 Introduction to Bioinformatics: Biological Networks

March 4, 2010

Interplay of network topology and biological function (cont.)

RECAP: topics covered in the previous lecture:

1 Lethality and centrality in PPI networks

2 Specificity and stability in the topology of PPI networks

3 Gene essentiality and the topology of PPI networks

4 Functional topology in PPI networks

(Continued from last class.)

- Distinction functional classes of proteins (e.g.: transcription, DNA-repair, metabolism, etc.) have different network properties, e.g. higher or lower degree in the PPI network
- Highly connected subgraphs (subgraphs which are dense in edges) tend to be protein complexes (i.e., groups of proteins which do a particular function together when they bind)
- In conclusion, there is a structure-function relationship in PPI networks.

5 Protein function prediction from PPI networks

- Proteins interact to perform a function

- Since PPI networks represent interactions, it is reasonable to use PPI network topology to predict protein function

Types of methods used to predict protein function:

5.1 Direct methods:

Proteins that lie closer to each other in the network are more likely to have similar function. Two examples are:

5.1.1 Majority-rule method: (Schwikowski and Fields, 2000, Nature Biotechnology)

- Annotate protein of an unknown function with the most common function(s) (such as transcription, DNA-repair, etc.) of its annotated neighbors.
- Advantages:
  - although it is a simple method, it works well.
- Disadvantages:
  - it doesn't assign significance values to predicted functions
  - only a limited network topology is considered
  - it fails to differentiate between proteins at different distances from the target protein

5.1.2 Network flow-based method: (Nabieva et. al., Bioinformatics, 2005)

- each functionally annotated protein is considered as a source of a functional flow
- the spread of the functional flow through the network is simulated over time
- each unannotated protein is assigned a score for having a given function based on the amount of flow it received during simulation

5.2 Cluster-based methods:

- Protein complex prediction: Dense network regions are a sign of a common involvement of proteins in certain biological processes and are candidates for protein complexes or functional modules.
- In these methods, we partition the network into clusters.

- Then we assign the entire cluster with a function based on the functions of its annotated members.

Examples of clustering methods (algorithms):

5.2.1 Highly connected subgraphs (HCS) method: (used in Przulj et. al., 2004)

It is based on the identification of highly connected subgraphs.

5.2.2 Restricted neighborhood search clustering (RNSC): (used in King et. al., 2004)

It is based on the following principle:

- we start with certain nodes (i.e. cluster)
- we keep swapping nodes in/out from this cluster
- when certain (previously specified) conditions are met (e.g. cluster density) then we keep these nodes
- otherwise we carry on searching

5.2.3 MCODE: (Bader and Hogue, 2003)

We won't discuss this method as it is intricate.

5.2.4 Hierarchical clustering:

We will discus it in detail later in the course. Here, it is applied to network data. The following distance metric was used:

- Pairwise distances between nodes along a shortest path:
  The assumption is that the smaller the distance between the proteins, the more "similar" they are.
- Czekanowski's dice distance: (Brun et. al., 2003)
  Assigns the maximum distance to two proteins with no common neighbors and distance 0 to those interacting with exactly the same neighbors.

5.2.5 Uncovering Biological Network Function via Graphlet Degree Signatures: (Przulj and Milenkovic, Cancer Informatics, 2008)

- Biological function of a protein and its local network structure (as described by graphlet degree vectors, a.k.a. "node signatures," covered in previous classes) are closely related.
- Proteins with topologically similar neighborhoods are clustered together and the resulting clusters are statistically significantly enriched in:

  - protein complexes
  - biological function
  - sub-cellular localization
  - tissue expression (in human)
  - involvement in (human) disease

- Used to predict function and new proteins involved in disease.

6 Biological networks in disease:

Essential genes have higher degrees. Now we can ask the question: Do disease/cancer-related genes also have high degrees? However, we need to be aware of the bias in the network data, since researchers are more interested in genes/proteins related to disease and hence these genes are more studied which results in them having higher degrees.

The following approaches were taken:

6.1 Jonsson and Bates, Bioinformatics, 2006

They demonstrated a greater connectivity and centrality of cancer genes compared to non-cancer ones. However, Goh et. al., PNAS (2007) pointed that the relationship between disease genes and their degrees needs more attention in the following sense. Initially they observed a correlation between the disease genes and their degrees, but later they found out that this was due to a large percent of essential genes in the disease gene class.

6.2 Milenkovic and Przulj, Cancer Informatics, 2008

They observed that disease/cancer genes tend to have very similar network neighborhoods in terms of their topology.

6.3 Disease network (Goh et. al., PNAS 2007) and drug-target network (Yildirim et al. (Marc Vidal's group), Nature Biotechnology, 2007)

Interesting readings of further interest, but no time to cover them in class.