Protein-Protein-Interaction Networks

Ulf Leser, Samira Jaeger
This Lecture

• **Protein-protein interactions**
  - Characteristics
  - Experimental detection methods
  - Databases
• **Biological networks**
Motivation

• Many cellular mechanisms rely on the physical binding of two or more proteins
  – E.g. signal transduction, gene regulation, metabolism, ...
  – May be transient or permanent
  – May have a directed effect (regulates) or undirected (bind)

• Changes in protein structure may hinder binding and thus perturb natural cellular processes
  – Influence on all “downstream” proteins, i.e., proteins reachable through a path of interactions

• Interactome – set of all PPIs in a cell (type, species, ...)

Ulf Leser: Bioinformatics, Summer Semester 2013
Context-dependency

- Protein-protein interaction: *Physical contact* with molecular docking
- Many proteins can interact – but under which *conditions*?
  - Cell type, cell cycle phase and state
  - Environmental conditions
  - Developmental stage
  - Protein modification
  - Presence of cofactors and other binding partners
  - ...
Experimental detection methods

- PPIs have been studied extensively using different experimental methods
- Many are small-scale: Two given proteins in a given condition
- **High-throughput methods**
  - Yeast two-hybrid assays (Y2H)
  - Tandem affinity purification and mass spectrometry (TAP-MS)
Yeast two-hybrid screens

• **Readout:** Certain gene is expressed iff two proteins bind

• **Method**
  – Test if protein A (bait) is interacting with protein B (prey)
  – Choose a transcription factor T whose expression requires two bindings: DNA binding and activation
  – Bait is fused to DNA binding domain of T
  – Prey is fused to activating domain of T
  – Both are expressed in genetically engineered yeast cells
  – If A binds to B, T is expressed and can be detected

![Diagram of yeast two-hybrid system]

Bait protein  Prey protein

BD  AD

Promoter

RNA Polymerase

Transcription of reporter gene
Properties

• Advantages
  – Throughput: Many preys can be tested with same bait (and vice versa)
  – Can be automated – high coverage of interactome
  – Very sensitive

• Problems
  – High rate of false positives (up to 50%)
    • Artificial environment: Yeast cells
    • No post-translational modifications
    • No protein transport
    • Unclear if proteins in vivo are ever expressed at the same time
    • ...
  – Fusion influences binding behavior – false negatives
Tandem affinity purification and mass spectrometry

1. Tag the protein of interest

2. Protein binds in its natural environment

3. Complexes are fished by affinity chromatography

4. Purification

5. Purified protein complexes

6. Identification of associated proteins by mass spectrometry
Properties

• Advantages
  – Can capture PPI in natural conditions
  – Single bait can detect many interactions in one experiment
  – Few false positives

• Disadvantages
  – Tag may hinder PPI – false negatives
  – Purification and MS are delicate processes (see next lecture)
  – Complexes are not identified
  – Internal structure of complex is not resolved
Matrix / Spokes Model

- Direct interactions can not be distinguished from interactions mediated by other proteins in a complex
- **Matrix model**: infers interactions between all proteins of a purified complex $\rightarrow \frac{N\times(N-1)}{2}$
- **Spokes model**: infers only interactions between the bait and the co-purified proteins $\rightarrow N-1$

<table>
<thead>
<tr>
<th># Proteins</th>
<th>Matrix</th>
<th>Spokes</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>45</td>
<td>9</td>
</tr>
<tr>
<td>80</td>
<td>3540</td>
<td>79</td>
</tr>
</tbody>
</table>
PPI Databases [KP10]

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually curated “source” DBs
  - Gather data from low-throughput methods
PPI Databases

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually curated "source" DBs
- DBs integrating other DBs and HT data sets
PPI Databases

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually curated “source” DBs
- DBs integrating others and HT data sets
- Predicted interactions
PPI Databases

- There are >300 BDBs related to PPI and pathways
  - See http://www.pathguide.org

- Manually curated “source” DBs
- DBs integrating other and HT data sets
- Predicted interactions
- Pathway DBs (consisting of PPI)
A Mess [KP10]

- Different definitions of a PPI
  - Binary, physical interaction
  - Complexes
  - Transient, functional association
- Some integrated DBs have imported more data than there is in the sources
- Databases overlap to varying degrees
- Different reliability of content
- Literature-curated DBs do not exhibit higher quality than HT [CYS08]
  - Re-annotation reveals inconsistencies, subjective judgments, errors in gene name assignment, ...
## Concrete Examples

<table>
<thead>
<tr>
<th>Database</th>
<th>Species</th>
<th>Proteins</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>IntAct</td>
<td>No restriction</td>
<td>53.276</td>
<td>271.764</td>
</tr>
<tr>
<td>BioGrid</td>
<td>No restriction</td>
<td>30.712</td>
<td>131.638</td>
</tr>
<tr>
<td>DIP</td>
<td>No restriction</td>
<td>23.201</td>
<td>71.276</td>
</tr>
<tr>
<td>MINT</td>
<td>No restriction</td>
<td>31.797</td>
<td>90.505</td>
</tr>
<tr>
<td>HPRD</td>
<td>Human only</td>
<td>30.047</td>
<td>39.194</td>
</tr>
<tr>
<td>MMPPI</td>
<td>Mammals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STRING</td>
<td>No restriction</td>
<td>2.590.259</td>
<td></td>
</tr>
<tr>
<td>UniHI</td>
<td>Human only</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPID</td>
<td>Human only</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Experimentally verified
- Experimentally verified and / or predicted
This Lecture

- Protein-protein interactions
- **Biological networks**
  - Scale-free graphs
  - Cliques and dense subgraphs
  - Centrality and diseases
Some Fundamental Observations

- Proteins that are **close in the network** have higher chance to share function
- **Central proteins** are vital
- Complexes form **dense subgraphs**
- **Functional modules** are subgraphs
- **Certain subgraphs** can be found significantly more often than expected by chance (why?)
Protein-protein interaction networks

- Networks are represented as directed graphs
- Definition of a graph: $G = (V,E)$
  - $V$ is the set of nodes (proteins)
  - $E$ is the set of edges (interactions)
- Computational representation

Adjacency lists
{ (A,C), (A,D), (B,D), (C,A), (C,D) (D,B) , (D,C), (D,A) }

Adjacency matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Degree distribution

- **Degree distribution** $P(k)$: relative frequency of nodes with exactly $k$ links (in or out)
- Used to define different classes of networks
- Common distributions
  - Poisson
    - Random networks
    
    
    $$P(k) \sim \frac{\lambda^k}{k!} e^{-\lambda}$$
  - Power-law
    - Scale-free networks
    
    $$P(k) \sim k^{-\gamma}$$

Barabasi et al., 2004
Scale-free Networks

- Biological networks are (presumably) scale-free
  - Few nodes are highly connected (hubs)
  - Most nodes have very few connections
- Also true for many other graphs: electricity networks, public transport, social networks, ...
- Evolutionary explanation
  - Growth: Networks grow by addition of new nodes
  - Preferential attachment: new nodes prefer linking to highly connected nodes
    • Possible explanation: Gene duplication – interaction with same targets
  - Older nodes have more changes to connect to nodes
  - Hub-structure emerges naturally
Other Biological Networks

- **Regulatory networks**: How genes and transcription factors influence expression of each other
  - TF influence genes and other TFs
  - Edges with semantic: activate / inhibit
  - Important, for instance, in cell differentiation

- **Signal networks**: Molecular reaction to external stimulus
  - Transient interactions including small molecules
  - Temporal dimension important
  - Important, for instance, in oncology

- Metabolic networks
- Ecological networks
- ...
Modular network organization

- Cellular function is carried out in a **modular manner**
- Function is reflected in a modular network structure

Don‘t be fooled by layout

**Modules must be dense, not close**

Costanzo et al., Nature, 2010
Clustering Coefficient

- Modules (cluster) are densely connected groups of nodes
- **Cluster coefficient** $C$ reflects network modularity by measuring tendency of nodes to cluster (‘triangle density’)

$$C_v = \frac{2 E_v}{d_v (d_v - 1)} \quad \rightarrow \quad C = \frac{1}{|V|} \sum_{v \in V} C_v$$

- $E_v = \text{number of edges between neighbors of } v$
- $d_v = \text{number of neighbors of } v$

- $\frac{d_v (d_v - 1)}{2} = \text{maximum number of edges between neighbors } d_v$
Example

- Cluster coefficient is a measure for the entire graph
- We also want to find modules, i.e., regions in the graph with very high cluster coefficient
- A **clique** is a maximal complete subgraph, i.e., a set of nodes where every pair is connected by an edge

\[
C_v = 10/10 = 1 \quad \quad C_v = 3/10 = 0.3 \quad \quad C_v = 0/10 = 0
\]
Functional Modules

Ribosome subunits – Translation

Pathways in cancer

MAPK/VEGF/Erb B signaling pathway

Proteasome subunits – Protein degradation

Protein transport

Gorasp2: Gorasp1
Rabia: Rab1b
Usp1

Hspa1, Hspa5, Hspa8, Hspa10, Hspa2

Rsp5, Rsp9

Protein degradation

Protein transport

Ribosome subunits

Translation
Finding Modules / Cliques

- Finding all **maximal cliques** in a graph is a highly complex problem
  - NP-complete
- Finding "**quasi-cliques**" is equally complex
  - Cliques with some missing edges
  - Same as subgraphs with high cluster coefficient
- Various heuristics
  - E.g. a good quasi-clique probably contains a (smaller) clique

```plaintext
build set $S_2$ of all cliques of size 2
i:= 2;
repeat
  i := i+1;
  $S_i$ := Ø;
  for j := 1 to $|S_{i-1}|$
    for k := i+1 to $|S_{i-1}|$
      $T$ := $S_{i-1}[j] \cap S_{i-1}[k]$;
      if $|T|$=$i-1$ then
        $N$ := $S_{i-1}[j] \cup S_{i-1}[k]$;
        if $N$ is a clique then
          $S_i$ := $S_i \cup N$;
        end if;
      end if;
    end for;
  end for;
until $|S_i|$ = 0:
```
Example

- 4-cliques: (1,3,4,5) – (1,3,4,6) – (1,3,4,7) - ...

- Merge-Phase

  \[ |(1,3,4,6) \cap (1,3,4,7)| = 3 \]
  \[ (1,3,4,6) \cup (1,3,4,7) = (1,3,4,6,7) \]
  Edge (6,7) exists

  5-clique

  \[ |(1,3,4,5) \cap (1,3,4,6)| = 3 \]
  \[ (1,3,4,5) \cup (1,3,4,6) = (1,3,4,5,6) \]
  Edge (5,6) does not exist

  No clique
This Lecture

- Protein-protein interactions
- Biological networks
  - Scale-free graphs
  - Cliques and dense subgraphs
  - Centrality and diseases
Network centrality

- **Central proteins** exhibit interesting properties
  - Essentiality – knock-out is lethal
  - Much higher evolutionary conservation
  - Often associated to (certain types of) human diseases

- Various measures exist
  - Degree centrality: Rank nodes by degree
  - Betweenness-centrality: Rank nodes by number of shortest paths between any pair of nodes on which it lies
  - Closeness-centrality: Rank nodes by their average distance to all other nodes
  - PageRank
  - ...
Network-based Disease Gene Ranking

Known Disease Genes (Seeds)

HRAS, RB1, KRAS2, FGFR3

Human Protein Interactions

Indirect Interactions

Functional Associations

Predicted Function

Network Centrality Analysis

Ranking

Disease candidates
Centrality of Seeds in (OMIM) Disease Networks

Fraction of seeds among top k% proteins; ~600 diseases from OMIM

\[ d_1 = \text{direct interactions} \]
\[ d_2 = \text{direct and indirect interactions} \]
Cross-Validation

- If a disease gene is not yet known – can we find it?
Further Reading

Example