Protein interaction networks

Ulf Leser and Samira Jaeger
Molecular interactions – Motivation

- Proteins mediate their function in complex interplay with other molecules through molecular interactions.

Enzymes bind substrates to catalyze biochemical reactions.

α and β-hemoglobin chains assemble into hetero-tetramers for transporting oxygen from lungs to tissues.

Transcription factors bind the DNA to induce transcription.
Protein-protein interactions – Motivation

• Important class of biomolecular interactions are **protein-protein interactions**

• Virtually **all cellular mechanisms rely on the physical binding of two or more proteins** to accomplish a particular task:
  • Critical role in cellular processes, e.g. signal transduction, gene regulation, cell cycle control and metabolism
  • Alterations in protein interactions perturb natural cellular processes and contribute to many diseases, such as cancer and AIDS

• Identifying all physical interactions within an organisms – the **interactome** – essential towards understanding the complex molecular relationships in living systems
This Lecture

- Protein-protein interactions
  - Characteristics
  - Experimental detection methods
  - Databases

- Protein-protein interaction networks
  - Characteristics
  - Applications
    - Protein function prediction
Protein-protein interactions – Characteristics

- Protein interaction defined as physical contact with molecular docking
  - Non-covalent contacts between side chains driven by hydrophobic effects, hydrogen bonds and electrostatic interactions

- Any two proteins can interact – but on what conditions?

- Important aspect is the biological context:
  - Cell type, cell cycle phase and state
  - Environmental conditions
  - Developmental stage
  - Protein modification
  - Presence of cofactors and other binding partners
Protein-protein interactions – Characteristics

- Protein interactions differ in diverse structural and functional characteristics, e.g. composition, affinity and life time of the association

- **Strength** depicts whether an interaction is permanent or transient

- **Specificity** refers to the selective binding of interaction partners

- **Type of interacting subunits** specifies whether an interaction forms hetero-oligomer with several different subunits or homo-oligomer with only one type of protein subunit
Experimental detection methods

- Protein-protein interactions have been studied extensively by different experimental methods

- Small-scale techniques
  - Yeast two-hybrid assays (Y2H)
  - Tandem affinity purification and mass spectrometry (TAP-MS)

- Large-scale techniques

- Cell assay
  - *in vivo* vs. *in vitro*

- Type of interaction
  - binary vs. complex

- Type of characterization

Piehler, 2005
Yeast two-hybrid screens

- Y2H is a molecular genetic tool, in which an interaction reconstitutes a transcription factor that activates expression of reporter genes.
- Transcription factors require two domains: DNA binding domain (BD) and an activation domain (AD).

Expression of fusion proteins in yeast cell

Transcription of reporter gene
Yeast two-hybrid screens

Benefits

• Large-scale analysis
• Sensitive \textit{in vivo} technique
• Identification of direct, transient and unstable interactions
• Genetic code of any fusion protein may be introduced into yeast cells

Drawbacks

• Poor reliability
  – High false positive rate up to 50\% (!)
  – High false negative rate
• Analysis of proteins in nucleus rather than in their native compartment
• Stable expression of fusion protein might be a problem
• Essential post-translational modification of non-yeast proteins may not be carried out
Tandem affinity purification and mass spectrometry

- TAP-MS involves biochemical isolation of protein complexes and subsequent identification of their constituting proteins using mass spectrometry.

**Diagram:**
- **Bait**
- Expression of fusion proteins
- IgG column
- Purification of protein complexes by affinity chromotography
- Calmodulin column
- Purified protein complexes
- Identification of associated proteins by mass spectrometry
Matrix and spoke model

- Direct interactions can not be distinguished from interactions mediated by other proteins in a complex
- How many interactions are detected from a TAP-MS purification?

  - **Matrix model**: infers interactions between all proteins of a purified complex → \((N^2 - 1)/2\)
  - **Spokes model**: infers only interactions between the bait and the co-purified proteins → \(N - 1\)

<table>
<thead>
<tr>
<th># Proteins</th>
<th>Matrix</th>
<th>Spoke</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>
Tandem affinity purification and mass spectrometry

Benefits

• Large-scale analysis
• Detection of protein complexes/interactions in correct cellular environment and detect several members of a complex

Drawbacks

• Data can not be directly translated into binary interactions
• Protein complexes not present under given conditions are missed
• Loosely associated proteins of a complex might be washed of during purification
• Protein targeting might interfere with complex formation
## Protein-protein interaction databases

<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></td>
<td>(372)</td>
<td></td>
<td></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></td>
<td>(630)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniHI</td>
<td>Human only</td>
<td>22.307</td>
<td>200.473</td>
</tr>
<tr>
<td>OPID</td>
<td>Human only</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Experimentally verified protein interactions
- Experimentally verified and predicted protein interactions
This Lecture

- Protein-protein interactions
  - Characteristics
  - Experimental detection methods
  - Databases

- Protein-protein interaction networks
  - Characteristics
  - Applications
    - Protein function prediction
Protein-protein interaction networks

- Binary interaction data can be assembled to protein-protein interaction networks
- Networks are represented as graphs
- Definition of a graph: $G = (V, E)$
  - $V$ is the set of nodes (proteins)
  - $E$ is the set of edges (interactions)
- Computational representation of graphs:

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>

Adjacency lists: (ordered) pairs of nodes

{ (A,C), (A,D), (B,D), (C,A), (C,D), (D,B), (D,C), (D,A) }
Graph-theoretic concepts

- A graph is defined by $G = (V,E)$, where $V$ is the set of nodes and $E$ is the set of edges connecting pairs of nodes.

- The distance between two nodes is the number of edges on the shortest path between them.
  - Diameter is the maximum distance between any two nodes.

- The neighborhood of a node is the set of nodes connected to it.
  - $n$-neighborhood of a node is the set of nodes with distance $n$.

- A clique is a fully connected subgraph, a subgraph in which every two nodes are connected by an edge.
  - $k$-core is a subgraph where each node has at least $k$ interactions.

- The density is the fraction of edges a graph has given all possible pairs of nodes.

$$D_G = \frac{2|E|}{|V|(|V|-1)}$$
Protein-protein interaction networks

• Why study protein interaction networks?
  • Elucidate the relationship between network structure and biological function
  • Discover novel protein function
  • Identify functional modules and conserved interaction patterns
  • Associate proteins with phenotypes or disease
  • Study pharmacological drug-target relationships
Topological network properties

- Topology of a network reveals its organization on different levels
  - Local and global characteristics provide insights in network evolution, stability and dynamics

- Common properties of biological networks
  - Small world property
  - Clustering coefficient
  - Degree distribution
  - Network centrality

- Modular network organization
Network centrality

- Network centrality analysis identifies interesting elements/proteins within a network.
- Quantitative measure to determine a protein's relative position in a network.

Example – Degree centrality:
- Degree of a node = number of edges to other nodes.

\[ C_D(v) = \frac{\text{deg}(v)}{|V| - 1} \]

- High centrality in interaction networks correlates with:
  - Gene essentiality
  - Evolutionary importance
  - Conservation rate
  - Likelihood to cause human disease
Degree distribution

- Degree distribution $P(k)$: probability that a node has exactly $k$ links
  - Count the number of nodes $N(k)$ with $k = 1, 2, \ldots$ links and divide by $N$
  - Allows to distinguish between different network classes

- Common network distributions
  - Poisson: $P(k) = \frac{e^{-d} d^k}{k!}$
  - Exponential: $P(k) \sim e^{-k/d}$
  - Power-law: $P(k) \sim k^{-\gamma}$

Barabasi et al., 2004
Scale-free networks

• **Scale-free networks**, e.g. protein interaction networks, follow a power law distribution: \( P(k) \sim k^{-\gamma} \), with degree exponents \( 2<\gamma<3 \)
  • Characterized by a small number of highly connected nodes known as hubs
• Scale-free topology typical feature of interaction networks
  • Most proteins participate in few interactions and few proteins in dozens
• Resistent to random failure, but prone to vulnerable attacks especially against hubs
Scale-free networks

- **Evolutionary origin** of scale-free networks
  - Growth: networks emerge through addition of new nodes
  - Preferential attachment: new nodes prefer to link to more connected nodes

- In interaction networks: scale-free property is thought to originate from gene duplications

![Network before duplication](image1)

- ![Gene duplication](image2)

![Network after duplication](image3)
Modular network organization

- Complex cellular function is carried out in a highly modular manner
- Modular organization is reflected in a modular network structure

Costanzo et al., Nature, 2010
Clustering coefficient

- Modules (or cluster) are densely connected groups of nodes
- Cluster coefficient $C$ reflects a network’s potential modularity and characterizes the tendency of nodes to cluster (‘triangle density’)
  \[
  C_v = \frac{2E_v}{d_v(d_v-1)} \quad \rightarrow \quad <C> = \frac{1}{|V|} \sum_{v \in V} C_v
  \]
  - $E_v$ = number of edges between neighbors of $v$
  - $d_v$ = number of neighbors of $v$
    \[
    \frac{d_v(d_v-1)}{2} = \text{maximum number of edges between neighbors } d_v
    \]
- Example:

  - $C_v = 10/10 = 1$
  - $C_v = 3/10 = 0.3$
  - $C_v = 0/10 = 0$
Functional modules

Modules comprise proteins that work together to achieve a specific function.

Ribosome subunits – Translation

Pathways in cancer

MAPK/VEGF/Erb B signaling pathway

Proteasome subunits – Protein degradation
Functional modules

- Two types of modules are distinguished in interaction networks
  - Protein complexes: proteins that interact with each other at the same time and place
  - Functional modules: proteins that participate in the same process but interacting at different times and places

- Finding functional modules → find densely connected subgraphs:
  - Network clustering
  - Network alignment
  - Decompose networks into subnetworks according to particular topological properties
Network-based protein function prediction

- Knowing a protein's function is essential for understanding biological processes, cellular mechanisms, evolutionary changes and the onset of diseases
- Protein interactions reflect the biological role of proteins within the cells
- Neighboring proteins in a network are likely to share function (guilt-by-association)
- Function might be inferred:
  1. By transferring known functions from directly or indirectly interacting proteins.
  2. Based on the protein complexes a protein belongs to.
Network-based protein function prediction

- Study the set of neighbors
- Consider position of the protein within its neighborhood,
- Consider position of the protein in the entire interaction network

Module-assisted annotation scheme

Direct annotation scheme
Direct annotation scheme

- **Correlation between network and functional distance**: the closer two proteins are in the network, the more similar are their functional annotation

- **Majority-rule** based on most common function(s) annotated to the direct interaction partners of a protein – proteins are associated with the most frequent functions of its direct neighbors
Module-assisted annotation scheme

• Based on hypothesis: cellular function is organized in a highly modular manner

• Module-based function assignment:
  1. First compute clusters (or modules) within the protein network.
  2. Proteins in a cluster are associated with annotations that are enriched within the module
     • Common functional annotations shared by the majority of the module’s proteins
     • Over-represented function that are enriched in a cluster according to the hypergeometric distribution
Direct vs. Module-based methods

- **Direct methods**
  + Accurate predictions
  + Provide high coverage
    - More sensitive to high level of false positives

- **Module-based methods**
  + More robust to missing or false interactions
  + Performance improves in networks with less functional coverage
    - Predict function only in dense network regions → reduced coverage
    - Less accurate than simple direct methods
  
- Both methods work within a species, which disregards functional information available in evolutionary related other species

Combining modularity, conservation, and direct interactions in one method
Combine modularity, conservation, and direct interactions

- Assumption: structural conservation in networks correlates with functional conservation in networks which can be exploited for predicting protein functions

![Diagram showing data integration, network comparison, and protein function prediction.](image)
Network Comparison

Species 1

Species 2

Species 3

Identification of orthologous groups

Identification of interologs

Assembly to CCS
CCS-based function prediction

- Analyze proteins within CCS that are defined by evolutionary conserved processes

- Combine comparative cross-species genomics and functional linkage within species-specific networks, predict function from
  - Orthology relationships
  - Direct interaction partners

![Diagram showing evolutionary conserved CCS]

Evolutionary conserved CCS = Functionally coherent?
1) **Exploiting Orthology Relationships across multiple species**

Species 1

Species 2

Identify non-characterized proteins

Infer function

**CCS-based function prediction**
CCS-based function prediction

2) Exploiting Protein Neighbours – Based on functional similarity between proteins

\[
\begin{align*}
\text{GO}_\text{Sim}(x,y) &> t \\
\text{GO}_\text{Sim}(x,z) &< t \\
\text{GO}_\text{Sim}(y,z) &> t
\end{align*}
\]

Candidate GO terms = \{GO_1, GO_2, ..., GO_n\}

<table>
<thead>
<tr>
<th>protein</th>
<th>GO Sim</th>
<th>GO_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>candidate</td>
<td>a</td>
<td>-</td>
</tr>
<tr>
<td>training</td>
<td>x</td>
<td>0.7</td>
</tr>
<tr>
<td>y</td>
<td>0.75</td>
<td>-</td>
</tr>
<tr>
<td>z</td>
<td>0.71</td>
<td>✓</td>
</tr>
</tbody>
</table>
**CCS-based function prediction**

**Combining modularity, conservation and interaction**

- **Species 1**
  - Predict along interactions
  - + Increased coverage
  - - Disregarding power of comparative genomics
  - - False positive PPIs

- **Species 2**
  - Predict by orthology relationships
  - + More robust to missing of false interactions
  - + Good performance in networks with less functional coverage
  - + Exploit knowledge of well-studied species
  - + High precision
  - - Limited coverage
  - - Restricted to proteins with characterized orthologs
Summary

• Analysis of protein interaction data and protein interaction network facilitates the understanding of cellular organization, function and processes.

• Public databases provide large repositories of interaction data of varying quality and quantity.

• Successfully used to infer novel function and disease associations from interaction partners.