Protein-Protein Interactions in Health and Disease

Franca Fraternali
Randall Division of Cellular and Molecular Biophysics
Interactions inside the cell...

PROTEOME
protein-protein interactions

METABOLISM
bio-chemical reactions
A Multiscale view of Protein-Protein Interactions


First Focus: **topological level**

Protein-Protein Interaction (PPI) Networks

**all interactions** → **interaction matrix**

<table>
<thead>
<tr>
<th></th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>a2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>a3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Useful for:**

- Studying inter-relationships of proteins (essential, crucial)
- The study of protein functions
- Detecting disease-related interactions (sub-networks)
- Mapping protein sub-networks to pathways
- Extracting phenotype-related sub-networks
- Complementing large scale siRNA screenings
Network Definition

The degree $k$ of protein $i$ in interaction network $c$ is defined as:

$$k_i(c) = \sum_j c_{ij}.$$  

Degree Distribution

i.e. probability of a protein to have degree $k$.

$$p(k) = \frac{1}{N} \sum_i \delta_{k, k_i(c)}$$
Relative degree-degree correlation DDC, i.e. the probability of two proteins with degree $k$ and $k'$ to be found linked in the graph.

The normalised DDC function $(k, k')$ of the network is defined as the ratio between the probability that two randomly picked nodes in $c$ with degrees $(k, k')$ are found to be connected, divided by what this probability would have been in large random networks with the same degree distribution as $c$.

The probabilities for large random networks can be calculated analytically. This results in the following definition for the normalised DDC function $\Pi(k, k')$:

$$\Pi(k, k') = \frac{\sum_{ij} c_{ij} \delta_{k_i(c), k'_i(c)} \delta_{k_j(c), k'_j(c)} \overline{k}}{[p(k) - N^{-1} \delta_{kk'}] p(k') \overline{Nkk'} \left(1 - \frac{1}{N}\right)}$$
Recent analyses of PPINs sparked a debate about the influence of the experimental method on the quality and biological relevance of the interaction data. Current experimental techniques, such as yeast two-hybrid (Y2H) and co–affinity purification combined with mass spectrometry (AP–MS), sample **subsets** of the interaction data space. These subsets show **very limited overlap**.
Degree-Degree Correlation $\triangleleft$ Yeast

Ito et al. 2001

Yu et al. 2008

Gavin et al. 2006

Krogan et al. 2006

Relative degree-degree correlation DDC
i.e. the probability of two proteins with degree $k$ and $k'$ to be found linked in the graph
Networks measured by the same experimental method cluster together, revealing a bias introduced by the methods that is seen to overrule species information. Although methodological biases have been acknowledged in the literature, we are now in a position to quantify their impact carefully by using an objective distance measure.
Analysis of PPINs using loop network motifs

Protein-protein interaction Network

Analysis by short loops of length 3, 4, 5, 6

Randomised Null Model
Markov Chain Graph Dynamics

Constraints:
i) degree distribution
ii) degree-degree correlation

1. Topological analysis
2. Functional analysis

Species | Method
---|---
C.elegans I | Y2H
C.elegans II | Y2H
C.jejuni | Y2H
H.pylori | Y2H
H.sapiens I | Y2H
H.sapiens II | Y2H
H.sapiens VIII | Y2H
M.loti | Y2H
P.falciparum | Y2H
S.cerevisiae I | Y2H
S.cerevisiae II | Y2H
S.cerevisiae III | Y2H
S.cerevisiae XII | Y2H
Synechocystis | Y2H
T.pallidum | Y2H
E.coli | AP-MS
H.sapiens III | AP-MS
S.cerevisiae IV | AP-MS
S.cerevisiae VI | AP-MS
S.cerevisiae VIII | AP-MS
S.cerevisiae IX | AP-MS
S.cerevisiae X | AP-MS
H.sapiens V | BP-MS
S.cerevisiae XI | PCA
D.melanogaster | Database
H.sapiens IV | Database
S.cerevisiae V | Data Integration
S.cerevisiae VII | Data Integration
H.sapiens VI | Database
H.sapiens VII | Database

H. sapiens V (BP-MS)

Reproduced from (Havugimana et al., 2012)
### Classification of PPINs by Topological Properties

<table>
<thead>
<tr>
<th>No.</th>
<th>Species</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C.elegans I</td>
<td>Y2H</td>
</tr>
<tr>
<td>2</td>
<td>C.elegans II</td>
<td>Y2H</td>
</tr>
<tr>
<td>3</td>
<td>C.jejuni</td>
<td>Y2H</td>
</tr>
<tr>
<td>4</td>
<td>H.pylori</td>
<td>Y2H</td>
</tr>
<tr>
<td>5</td>
<td>H.sapiens I</td>
<td>Y2H</td>
</tr>
<tr>
<td>6</td>
<td>H.sapiens II</td>
<td>Y2H</td>
</tr>
<tr>
<td>7</td>
<td>H.sapiens VIII</td>
<td>Y2H</td>
</tr>
<tr>
<td>8</td>
<td>M.lothi</td>
<td>Y2H</td>
</tr>
<tr>
<td>9</td>
<td>P.falciparum</td>
<td>Y2H</td>
</tr>
<tr>
<td>10</td>
<td>S.cerevisiae I</td>
<td>Y2H</td>
</tr>
<tr>
<td>11</td>
<td>S.cerevisiae II</td>
<td>Y2H</td>
</tr>
<tr>
<td>12</td>
<td>S.cerevisiae III</td>
<td>Y2H</td>
</tr>
<tr>
<td>13</td>
<td>S.cerevisiae XII</td>
<td>Y2H</td>
</tr>
<tr>
<td>14</td>
<td>Synechocystis</td>
<td>Y2H</td>
</tr>
<tr>
<td>15</td>
<td>T.pallidum</td>
<td>Y2H</td>
</tr>
<tr>
<td>16</td>
<td>E.coli</td>
<td>AP-MS</td>
</tr>
<tr>
<td>17</td>
<td>H.sapiens III</td>
<td>AP-MS</td>
</tr>
<tr>
<td>18</td>
<td>S.cerevisiae IV</td>
<td>AP-MS</td>
</tr>
<tr>
<td>19</td>
<td>S.cerevisiae VI</td>
<td>AP-MS</td>
</tr>
<tr>
<td>20</td>
<td>S.cerevisiae X</td>
<td>AP-MS</td>
</tr>
<tr>
<td>21</td>
<td>S.cerevisiae IX</td>
<td>AP-MS</td>
</tr>
<tr>
<td>22</td>
<td>H.sapiens V</td>
<td>BP-MS</td>
</tr>
<tr>
<td>23</td>
<td>S.cerevisiae XI</td>
<td>PCA</td>
</tr>
<tr>
<td>24</td>
<td>D.melanogaster</td>
<td>Database</td>
</tr>
<tr>
<td>25</td>
<td>H.sapiens IV</td>
<td>Database</td>
</tr>
<tr>
<td>26</td>
<td>S.cerevisiae V</td>
<td>Data Integration</td>
</tr>
<tr>
<td>27</td>
<td>S.cerevisiae VII</td>
<td>Data Integration</td>
</tr>
<tr>
<td>28</td>
<td>H.sapiens VI</td>
<td>Database</td>
</tr>
<tr>
<td>29</td>
<td>H.sapiens VII</td>
<td>Database</td>
</tr>
</tbody>
</table>

#### Loop Resilience

![Diagram showing loop resilience](image)

- Red line: 
  - `{k}`
  - Number of loops per edge decreases sharply.

- Blue line:
  - `{k, k'}`
  - The number of loops per edge decreases gradually.

Core Cluster of PPIN from Resilient Loops

U2 snRNP splicing

ATP-dependent helicase

Interleukin enhancer binding factors
Functional Consensus in *H. Sapiens* V (BP-MS) using Gene Ontology (GO)
Functional Enrichment in loop motifs

x-axis: Network, Loop3, Loop4, Loop5
y-axis: Frequency of GO terms

H. Sapiens V (BP-MS)

- mRNA metabolism
- gene expression
- viral process

Trend1
90

- organismal and
developmental processes
- DNA-templated transcription

Trend2
28

Trend3
144

- biosynthesis
- protein complex subunit
- localization (transport)

Trend4
171

- cell cycle / cell death
- regulation processes
- antigen processing

Trend5
56

H. Sapiens V (BP-MS)
- Ribosomal proteins

Network
Loop3
Loop4
Loop5
Network ~ Short Loops

Frequency of GO terms (%)
• cell cycle / cell death
• regulation processes
• antigen processing

• biosynthesis
• protein complex subunit
• localization (transport)
Second Focus: **Molecular Complexes**

Structural Coverage of Human PPINs

http://interactome3d.irbbarcelona.org/

---

**Single protein structures**

**Interaction protein pair structures**

**Chart 1**

![Chart 1](image1)

**Chart 2**

![Chart 2](image2)

---


Protein Binding Mechanism

Integrating 3D Structure Information
Identify Partner Proteins with Structures

Homolog structures of protein A and B are in the same PDB (biological unit)?

Yes

Identify surface and interface region of protein A and B by calculating solvent-accessible surface area (SASA).

(Kleinjung, J. and Fraternali, F., 2005 Nucleic Acids Research)

No

Discard partner protein B
From BIG to SMALLER data with atomic resolution

BIG data

• Large-scale genome sequencing projects as:
  ▪ the 1000 Genomes Project
  ▪ International HapMap Project
  ▪ the NHLBI Exome Sequencing Project
  ▪ CHARGE Consortium
  ▪ Cancer Genome Project
Implications at the molecular level
• **General nsSNPs** (*nsSNPs*\(^C\))
• **Germ-line disease nsSNPs** (*nsSNPs*\(^{GD}\))
• **Somatic cancer nsSNVs** (*nsSNVs*\(^{SC}\))
Three-dimensional reconstruction of protein networks provides insight into human genetic disease.
Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H.
Germ-line disease variants on structured regions are predicted damaging
enrichment close to PTM sites
enrichment - secondary structure

‘flexible’ regions

cancer-interface

GD-cancer-core
SCOP-Domains Functional annotation: GD

**General:** general and multiple functions; interactions with proteins/ions/lipids/small molecules

**Metabolism:** anabolic and catabolic processes; cell maintenance/homeostasis; secondary metabolism

**Information:** storage, maintenance of the genetic code; DNA replication/repair; general transcription/translation

**Regulation:** regulation of gene expression and protein activity; information processing in response to environmental input; signal transduction; general regulatory or receptor activity

**Extra-cellular processes:** inter-, extra-cellular processes, e.g. cell adhesion; organismal processes, e.g. blood clotting, immune system

**Intra-cellular processes:** cell motility/division; cell death; intra-cellular transport; secretion

SCOP-Domains Functional annotation SC

- RNA binding
- DNA-binding
- Kinases/phosphatases
- Signal transduction
- Other regulatory function
- Receptor activity
The Human Disease Network

Summary

• Both germ-line disease and cancer variants have tendency to occur at the interface region of proteins.

• Germ-line disease and cancer variants of the interface tend to be mostly located at flexible structural regions.

• A high number of disordered region variants locate close to PTM sites; for germ-line diseases and cancer the variants are close to ordered region PTM sites.

• The type of disease that a gene variant underpins may be intimately related to the function of the protein/domain implicated.
Conclusions

We have developed a new generation of precise and user-friendly computational tools to quantify PPIN topologies particularly useful for testing new protocols for the removal of experimental biases from PPIN datasets.

We have shown by a large-scale analysis of publicly available datasets that the present protein network data are strongly biased by their experimental methods, while still exhibiting species-specific similarity and reproducibility.

Introduced a new strategy to identify regulators of a signalling pathway in immune cells. Devised a targeted screen using bioinformatics databases identified PI3K-p85α and AKT1 as novel regulators of the oscillatory increase in CDC42 activity and prevent polarisation at the immunological synapse.

Characterised dynamical properties of Hub proteins and observed higher conformational flexibility for multi-partner residues in unbound proteins. Stronger communication exists between multi-partner residues.

Analysed the propensity for the localization of human missense variants on protein structures; highlighted different trends in germ-line diseases vs somatic cancer mutations, hypotheses on disease-protein-function relationships.
Thanks to.....

Luis Fernandez PhD
Ton Coolen, Alessia
Annibale IMMB KCL
Tony Ng, Leo Carlin
Cancer Studies, KCL
Jens Kleinjung NIMR
London