Protein Binding Networks: from Topology to Kinetics

Sergei Maslov
Brookhaven National Laboratory
Genome-wide protein binding networks

- Nodes - proteins
- Edges - protein-protein binding interactions
- Functions
  - structural
  - complexes/dimers
  - regulation/signaling
  - unknown?
  - etc

*C. elegans* PPI from
Li et al. (Vidal’s lab), *Science* (2004)
How much data is out there?

<table>
<thead>
<tr>
<th>Species</th>
<th>Set</th>
<th>nodes</th>
<th>edges</th>
<th># of sources</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td>HTP-PI</td>
<td>4,500</td>
<td>13,000</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>LC-PI</td>
<td>3,100</td>
<td>20,000</td>
<td>3,100</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>HTP-PI</td>
<td>6,800</td>
<td>22,000</td>
<td>2</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>HTP-PI</td>
<td>2,800</td>
<td>4,500</td>
<td>1</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>LC-PI</td>
<td>6,400</td>
<td>31,000</td>
<td>12,000</td>
</tr>
<tr>
<td></td>
<td>HTP-PI</td>
<td>1,800</td>
<td>3,500</td>
<td>2</td>
</tr>
<tr>
<td><em>H. pylori</em></td>
<td>HTP-PI</td>
<td>700</td>
<td>1,500</td>
<td>1</td>
</tr>
<tr>
<td><em>P. falciparum</em></td>
<td>HTP-PI</td>
<td>1,300</td>
<td>2,800</td>
<td>1</td>
</tr>
</tbody>
</table>
Yeast two-hybrid technique

uses two “hybrid proteins”: bait A* (A fused with Gal4p DNA-binding domain) and prey B* (B fused with Gal4p activation domain)

- **Cons:** wrong (very high) concentrations, localization (unless both proteins are nuclear), and even host organism (unless done in yeast)
- **Pros:** direct binding events
- **Main source of noise:** self-activating baits
Affinity capture + Mass Spectrometry

*(multi-)protein complex pulled out by affinity-tagged protein (bait)*

- Pros: in vivo concentrations and localizations
- Cons: binding interactions are often indirect
- Main source of noise: highly abundant and sticky proteins
# Breakup by experimental technique in yeast

<table>
<thead>
<tr>
<th>Technique</th>
<th>S. cerevisiae</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOGRID database</td>
<td>46063</td>
</tr>
<tr>
<td>Affinity Capture-Mass Spec</td>
<td>28172</td>
</tr>
<tr>
<td>Affinity Capture-RNA</td>
<td>55</td>
</tr>
<tr>
<td>Affinity Capture-Western</td>
<td>5710</td>
</tr>
<tr>
<td>Co-crystal Structure</td>
<td>107</td>
</tr>
<tr>
<td>FRET</td>
<td>43</td>
</tr>
<tr>
<td>Far Western</td>
<td>41</td>
</tr>
<tr>
<td>Two-hybrid</td>
<td>11935</td>
</tr>
<tr>
<td>Total</td>
<td>46063</td>
</tr>
</tbody>
</table>
What are the common topological features?

1. Broad distribution of the number of interaction partners (degree K) of individual proteins
2. Anti-correlation of degrees of interacting proteins
3. **Small-world-property**
   (follows from 1. for $<K^2>/<K>^2 > 2$)
Protein binding networks have small-world property

86% of proteins could be connected

83% in this plot

Large-scale Y2H experiment

Curated dataset from our study
Why small-world matters?

- Claims of “robustness” of this network architecture come from studies of the Internet where breaking up the network is a disaster.
- For PPI networks it is the OPPOSITE: interconnected networks present a problem.
- In a small-world network equilibrium concentrations of all proteins are coupled to each other.
- Danger of undesirable cross-talk.
Going beyond topology and modeling the equilibrium and kinetics

What is needed to model?

- A reliable network of reversible (non-catalytic) protein-protein binding interactions
  - √ CHECK! e.g. physical interactions between yeast proteins in the BIOGRID database with 2 or more citations

- Total concentrations and sub-cellular localizations of all proteins
  - √ CHECK! genome-wide data for yeast in 3 Nature papers (2003, 2003, 2006) by the group of J. Weissman @ UCSF
  - Left us with 1700 yeast proteins and ~5000 interactions

- *in vivo* dissociation constants $K_{ij}$
  - OOPS! 😞. High throughput experimental techniques are not there yet
Let’s hope it doesn’t matter

- The overall binding strength from the PINT database: 
  \( <1/K_{ij}> = 1/(5\text{nM}) \). In yeast: 1nM \( \sim \) 34 molecules/cell

- Simple-minded assignment: \( K_{ij} = \text{const} = 10\text{nM} \) 
  (also tried 1nM, 100nM and 1000nM)

- Evolutionary-motivated assignment: 
  \( K_{ij} = \max(C_i, C_j)/20 \): \( K_{ij} \) is only as small as needed to ensure binding

- All assignments of a given average strength give 
  ROUGHLY THE SAME RESULTS
Law of Mass Action

- \( \frac{dD_{AB}}{dt} = r_{on} F_A F_B - r_{off} D_{AB} \)
- In equilibrium \( D_{AB} = \frac{F_A F_B}{K_{AB}} \) where the dissociation constant \( K_{AB} = \frac{r_{off}}{r_{on}} \) has units of concentration
- Total concentration = free concentration + bound concentration
  \( C_A = F_A + \frac{F_A F_B}{K_{AB}} \); \( C_B = F_B + \frac{F_A F_B}{K_{AB}} \)
- \( F_A = \frac{C_A}{1 + F_B / K_{AB}} \); \( F_B = \frac{C_B}{1 + F_B / K_{AB}} \)
Law of Mass Action equilibrium of a PPI network

- In a network \( F_i = C_i / (1 + \sum_{\text{neighbors } j} F_j / K_{ij}) \)
- Even though it cannot be solved analytically it is easily solved numerically e.g. by iterations
- We use experimentally measured total concentrations \( C_i \) to calculate all unbound (free) \( F_i \) and all bound \( D_{ij} = F_i F_j / K_{ij} \) concentrations
Robustness with respect to assignment of $K_{ij}$

Bound concentrations: $D_{ij}$

Free concentrations: $F_i$

Spearman rank correlation: 0.89
Pearson linear correlation: 0.98

Spearman rank correlation: 0.89
Pearson linear correlation: 0.997
We simulate a twofold increase of the abundance $C_0$ of just one protein.

Proteins whose free concentration $F_i$ changes by $>20\%$ are considered to be significantly perturbed.

We refer to such proteins $i$ as concentration-coupled to the protein 0.

Look for cascading perturbations: changes in the total concentration $C_0$ of $P_0$ affects $F_1$ of its binding partner $P_1$, which in turn affects $F_2$ of its partner $P_2$, etc.
Indiscriminate cross-talk is suppressed

<table>
<thead>
<tr>
<th>L</th>
<th>variable $K_{ij}$, mean = 5nM</th>
<th>constant $K_{ij} = 1nM$</th>
<th>constant $K_{ij} = 10nM$</th>
<th>constant $K_{ij} = 0.1\mu M$</th>
<th>constant $K_{ij} = 1\mu M$</th>
<th>all pairs at distance L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2003</td>
<td>2469</td>
<td>1915</td>
<td>1184</td>
<td>387</td>
<td>8168</td>
</tr>
<tr>
<td>2</td>
<td>415</td>
<td>1195</td>
<td>653</td>
<td>206</td>
<td>71</td>
<td>29880</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>159</td>
<td>49</td>
<td>8</td>
<td>0</td>
<td>87772</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>60</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>228026</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>396608</td>
</tr>
</tbody>
</table>
What conditions make some long chains good conduits for propagation of concentration perturbations while suppressing it along the rest?
Resistor network analogy

- Conductivities $\sigma_{ij}$ – dimer (bound) concentrations $D_{ij}$
- Losses to the ground $\sigma_{iG}$ – free (unbound) concentrations $F_i$
- Electric potentials – relative changes in free concentrations $(-1)^L \delta F_i/F_i$
- Injected current – initial perturbation $\delta C_0$

SM, K. Sneppen, I. Ispolatov, q-bio/0611026
Perturbations propagate along dimers with large concentrations. They cascade down the concentration gradient and thus directional. Free concentrations of intermediate proteins are low.
Implications of our results
Cross-talk via small-world topology is suppressed, but...

- **Good news:** on average perturbations via reversible binding *rapidly decay*
- Still, the **absolute number** of concentration-coupled proteins is **large**
- In response to external stimuli **levels of several proteins** could be shifted. Cascading changes from these perturbations could either **cancel** or **magnify** each other.
- Our results could be used to extend the list of perturbed proteins measured e.g. in microarray experiments
Genetic interactions

- Propagation of concentration perturbations is behind many genetic interactions e.g. of the “dosage rescue” type.
- We found putative “rescued” proteins for 136 out of 772 such pairs (18% of the total, P-value $10^{-216}$).
Intra-cellular noise

- Noise is measured for total concentrations $C_i$ (Newman et al. Nature (2006))
- Needs to be converted in biologically relevant bound ($D_{ij}$) or free ($F_i$) concentrations
- Different results for intrinsic and extrinsic noise
- Intrinsic noise could be amplified (sometimes as much as 30 times!)
Could it be used for regulation and signaling?

- 3-step chains exist in bacteria: anti-anti-sigma-factors $\rightarrow$ anti-sigma-factors $\rightarrow$ sigma-factors $\rightarrow$ RNA polymerase

- Many proteins we find at the receiving end of our long chains are **global regulators** (protein degradation by ubiquitination, global transcriptional control, RNA degradation, etc.)
  - Other (catalytic) mechanisms spread perturbations even further
  - Feedback control of global protein abundance?
NOW BACK TO TOPOLOGY
What are the common topological features?

1. Broad distribution of the number of interaction partners of individual proteins
• What’s behind this broad distribution?

• Three explanations were proposed:
  • EVOLUTIONARY (duplication-divergence models)
  • BIOPHYSICAL (stickiness due to surface hydrophobicity)
  • FUNCTIONAL (tasks of vastly different complexity)

Evolutionary explanation: duplication-divergence models


- Network has to grow
- Divergence has to be asymmetric
  (K Evlampiev, H Isambert, q-bio.MN/0611070)
Gene duplication

Right after duplication

Pair of duplicated proteins

Shared interactions

After some time

Pair of duplicated proteins

Shared interactions
Traces of duplication in PPI networks

(a similar but smaller scale-plot vs $K_s$ in A. Wagner MBE 18, 1283 (2001)}
But: how important are duplications for shaping hubs?

Duplication-divergence models could still be OK if sequences diverge relatively fast.

Biophysical explanation: “stickiness” models


- Nodes have intrinsic “stickiness” $S_i$.
- Stickiness could have exponential or Gaussian PDF.
- Binding edge $i - j$ is drawn with probability $p_{ij} = F(S_i + S_j)$
- $F$ is some (soft) threshold function, e.g. $\exp(S_i + S_j - \mu)/(1 + \exp(S_i + S_j - \mu))$
- Network does not have to grow
There are just TOO MANY homodimers

- Null-model: $P_{self} \sim \langle k \rangle / N$

$$N_{\text{dimer}} = N \cdot P_{self} = \langle k \rangle$$

- Not surprising as homodimers have many functional roles

<table>
<thead>
<tr>
<th></th>
<th>$N_{\text{dimer}}$</th>
<th>$N^{(r)}_{\text{dimer}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>yeast</td>
<td>179</td>
<td>6.6 ± 0.2</td>
</tr>
<tr>
<td>worm</td>
<td>89</td>
<td>3.3 ± 0.1</td>
</tr>
<tr>
<td>fly</td>
<td>160</td>
<td>5.9 ± 0.1</td>
</tr>
<tr>
<td>human</td>
<td>1045</td>
<td>5.7 ± 0.1</td>
</tr>
</tbody>
</table>

\[ P_{dimer}(k) = 1 - (1 - p_{self})^k \]

**Fly:** two-hybrid data

\[ P_{self} \sim 0.003, \quad P_{\text{others}} \sim 0.0002 \]

**Human:** literature data

\[ P_{self} \sim 0.05, \quad P_{\text{others}} \sim 0.0002 \]

I. Ispolatov, A. Yuryev, I. Mazo, and SM, **33**, 3629 NAR (2005)
Our interpretation

- Both the number of interaction partners $K_i$ and the likelihood to self-interact are proportional to the same “stickiness” of the protein $S_i$ which could depend on
  - the number of hydrophobic residues on the surface
  - protein abundance
  - its’ popularity (in networks taken from many small-scale experiments)
  - etc.

- In random networks $p_{dimer}(K) \sim K^2$ not $\sim K$ like we observe empirically

Functional explanation: there are as many binding partners as needed for function

- Not an explanation: why difficulty of functions is so heterogeneous?
- Difficult to check: the function of many binding interactions is poorly understood (quite clear in transcriptional regulatory networks e.g. in *E. coli*).
- The 3rd explanation does not exclude the previous two: Evolution by duplications combined with pure Biophysics (stickiness) provide raw materials from which functional interactions are selected.
What are the common topological features?

- Broad distribution of the number of interaction partners (degree) of individual proteins
- Anti-correlation of degrees of interacting proteins
Central vs peripheral network architecture

central (hierarchical)

random

peripheral (anti-hierarchical)

What is the case for protein interaction network

Randomization

given complex network

random
Edge swapping (rewiring) algorithm

- Randomly select and rewire two edges
- Repeat many times

Metropolis rewiring algorithm

“energy” $E$ → “energy” $E + \Delta E$

- Randomly select two edges
- Calculate change $\Delta E$ in “energy function” $E = (N_{\text{actual}} - N_{\text{desired}})^2 / N_{\text{desired}}$
- Rewire with probability $p = \exp(-\Delta E / T)$
