Increasing Confidence of Protein-Protein Interactomes

Limsoon Wong
(Based on work of/with Jin Chen, Kenny Chua, Wynne Hsu, Mong Li Lee, See-Kiong Ng, Rintaro Saito, Wing-Kin Sung)

Outline

• Reliability of experimental protein-protein interaction data

• Identification of false positives
  – Interaction generality
  – Interaction generality 2
  – Interaction pathway reliability
  – FS Weight
  – Meso-scale network motifs
How Reliable are Experimental Protein-Protein Interaction Data?

Figure credit: Jeong et al. 2001

Why Protein Interactions?

• Complete genomes are now available
• Knowing the genes is not enough to understand how biology functions

• Proteins, not genes, are responsible for many cellular activities
• Proteins function by interacting w/ other proteins and biomolecules

GENOME

PROTEOME

“INTERACTOME”

Slide credit: See-Kiong Ng
High-Tech Expt PPI Detection Methods

- Yeast two-hybrid assays
- Mass spec of purified complexes (e.g., TAP)
- Correlated mRNA expression
- Genetic interactions (e.g., synthetic lethality)
- ...

**FACT:** Generating *large amounts* of experimental data about protein-protein interactions can be done with ease.

Key Bottleneck

- Many high-throughput expt detection methods for protein-protein interactions have been devised
- But ...

High-throughput approach sacrifice quality for quantity: *false negatives*, & *false positives*
Some Protein Interaction Data Sets
Sprinzak et al., JMB, 327:919-923, 2003

Large disagreement betw methods

- GY2H: genome-scale Y2H
- 2M, 3M, 4M: intersection of 2, 3, 4 methods

<table>
<thead>
<tr>
<th>Experimental method category</th>
<th>Number of interacting pairs</th>
<th>Co-localization (%)</th>
<th>Co-cellular role (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: All methods</td>
<td>9347</td>
<td>64</td>
<td>49</td>
</tr>
<tr>
<td>A0: Small-scale Y2H</td>
<td>1861</td>
<td>73</td>
<td>62</td>
</tr>
<tr>
<td>A1: GY2H Uetz et al., (published results)</td>
<td>956</td>
<td>66</td>
<td>45</td>
</tr>
<tr>
<td>A2: GY2H Ito et al., (core)</td>
<td>798</td>
<td>64</td>
<td>40</td>
</tr>
<tr>
<td>A3: GY2H Ito et al., (all)</td>
<td>8695</td>
<td>41</td>
<td>15</td>
</tr>
<tr>
<td>B: Physical methods</td>
<td>71</td>
<td>98</td>
<td>95</td>
</tr>
<tr>
<td>C: Genetic methods</td>
<td>1052</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>D1: Biochemical, for vitro</td>
<td>614</td>
<td>87</td>
<td>79</td>
</tr>
<tr>
<td>D2: Biochemical, chromatography</td>
<td>648</td>
<td>93</td>
<td>88</td>
</tr>
<tr>
<td>E1: Immunological, direct</td>
<td>1025</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>E2: Immunological, indirect</td>
<td>34</td>
<td>100</td>
<td>93</td>
</tr>
<tr>
<td>2M: Two different methods</td>
<td>2340</td>
<td>87</td>
<td>85</td>
</tr>
<tr>
<td>3M: Three different methods</td>
<td>1212</td>
<td>92</td>
<td>94</td>
</tr>
<tr>
<td>4M: Four different methods</td>
<td>570</td>
<td>95</td>
<td>95</td>
</tr>
</tbody>
</table>

% of TP based on shared cellular role (I = .95)
% of TP based on shared cellular role (I = 1)
% of TP based on co-localization

TP = ~50%
Objective

• Some high-throughput protein interaction expts have as much as 50% false positives

• Can we find a way to rank candidate interaction pairs according to their reliability?

• How do we do this?
  – Would knowing their neighbours help?
  – Would knowing their local topology help?
  – Would knowing their global topology help?

Would knowing their neighbours help?
The story of interaction generality
An Observation

- It seems that configuration a is less likely than b in protein interaction networks
- Can we exploit this?

Interaction Generality
Saito et al., NAR, 30:1163-1168, 2002

Given an edge $X \leftrightarrow Y$ connecting two proteins, $X$ and $Y$, the “interaction generality” measure $\text{ig}^0(X \leftrightarrow Y)$ of this edge as defined as:

$$\text{ig}^0(X \leftrightarrow Y) = 1 + \left| \{X' \leftrightarrow Y' \in G \mid X' \in \{X,Y\}, \ deg^0(Y') = 1 \} \right|$$

where $\text{deg}^0(U) = |\{V : U \leftrightarrow V \in G\}|$ is the degree of the node $U$ in the undirected graph $G$.

$\text{ig}(YDR412W \leftrightarrow \text{GLC7}) = 1 + \# \text{ of yellow nodes}$
Assessing Reliability Using Interaction Generality

• Recall configuration a is less likely than b in protein interaction networks
• The smaller the “ig” value of a candidate interaction pair is, the more likely that interaction is

There are 229 pairs in Ito having ig = 1. Of these, 66 (or 34%) are also reported by Uetz

Interacting pairs c’mon to Ito et al. & Uetz et al. are more reliable
• Also have smaller “ig”
⇒ “ig” seems to work
Evaluation wrt Co-localization

~60% of pairs in Ito having ig=1 are known to have common localization

- Interaction pairs having common cellular localization are more likely
- Also have lower “ig”
⇒ “ig” seems to work

Evaluation wrt Co-cellular Role

- Interaction pairs having common cellular role are more likely
- Also have lower “ig”
⇒ “ig” seems to work

A: before restrict to pairs with “ig = 1”
B: after restrict to pairs with “ig = 1”
Would knowing their local topology help?

The story of interaction generality 2

Existence of Network Motifs
Milo et al., Science, 298:824-827, 2002

- A network motif is just a local topological configuration of the network
- “Detected” in gene regulation networks, WWW links, etc.

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>$N_{\text{real}}$</th>
<th>$N_{\text{rand}} \pm SD$</th>
<th>$Z$ score</th>
<th>$N_{\text{real}}$</th>
<th>$N_{\text{rand}} \pm SD$</th>
<th>$Z$ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene regulation (transcription)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. coli</td>
<td>424</td>
<td>519</td>
<td>40</td>
<td>7 ± 3</td>
<td>10</td>
<td>203</td>
<td>47 ± 12</td>
<td>13</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>685</td>
<td>1,652</td>
<td>70</td>
<td>13 ± 4</td>
<td>14</td>
<td>1812</td>
<td>300 ± 46</td>
<td>41</td>
</tr>
</tbody>
</table>

Observed 70 times in S. cerevisiae
Observed ~11 times in random data
5 Possible Network Motifs

- Classify a protein $C$ that directly interacts with the pair $A \leftrightarrow B$ according to these 5 topological configurations.

A New Interaction Generality


The improved interaction generality measure $\ell_g^\beta(X \leftrightarrow Y)$ is defined as a weighted sum of the 5 local topological configurations $\tau_1, \ldots, \tau_5$ as

$$\ell_g^\beta(X \leftrightarrow Y) = \sum_{i=1}^{5} \lambda_i \cdot \left| \{X' | X' \leftrightarrow Y' \in \emptyset, Y' \in \{X, Y\}, \tau_i^g(X', X \leftrightarrow Y)\} \right|$$

where $\lambda_i$ is the weight for configuration $\tau_i$, and $\tau_i^g(X', X \leftrightarrow Y)$ means $X'$ is in configuration $\tau_i$ in graph $g$ wrt $X \leftrightarrow Y$. 
~95% of pairs having ig₂ = –6 have common cellular roles

“ig₂” correlates well to common cellular roles, localization, & expression

“ig₂” seems to work better than “ig”

Would knowing their global topology help?
The story of interaction pathway reliability
Some “Reasonable” Speculations

- A true interacting pair is often connected by at least one alternative path (reason: a biological function is performed by a highly interconnected network of interactions)

- The shorter the alternative path, the more likely the interaction (reason: evolution of life is through “add-on” interactions of other or newer folds onto existing ones)

Therefore...

Conjecture:
“An interaction that is associated with an alternate path of reliable interactions is likely to be reliable.”

Idea:
Use alternative interaction paths as a measure to indicate functional linkage between the two proteins

Slide credit: See-Kiong Ng
Interaction Pathway Reliability

The "interaction pathway reliability" measure $ipr^G(X \leftrightarrow Y)$ is defined as

$$ipr^G(X \leftrightarrow Y) = \max_{\phi \in \mathcal{G}} \prod_{(U \leftrightarrow V) \in \phi} \left( 1 - \frac{ig^G(U \leftrightarrow V)}{ig_{\text{max}}^G} \right)$$

where $ig_{\text{max}}^G = \max \{ig^G(X \leftrightarrow Y) \mid (X \leftrightarrow Y) \in \mathcal{G} \}$ is the maximum interaction generality value in $G$, and $\Phi^G(X,Y)$ is the set of all possible non-reducible paths between $X$ and $Y$, but excluding the direct path $X \leftrightarrow Y$. Here, a path $\phi$ connecting $X$ and $Y$ is non-reducible if there is no shorter path $\psi'$ connecting $X$ and $Y$ that shares some common intermediate nodes with the path $\phi$.

IPM is also called IRAP, "Interaction Reliability by Alternate Pathways"

Non-reducible Paths

- Non-reducible paths are
  - $A \leftrightarrow F \leftrightarrow E$
  - $A \leftrightarrow B \leftrightarrow E$
- Reducible paths are
  - $A \leftrightarrow B \leftrightarrow C \leftrightarrow D \leftrightarrow E$
  - $A \leftrightarrow B \leftrightarrow C \leftrightarrow E$
Evaluation wrt Common Cellular Role, etc

- "ipr" correlates well to common cellular roles, localization, & expression
- "ipr" seems to work better than "ig2"

At the ipr threshold that eliminated 80% of pairs, ~85% of the remaining pairs have common cellular roles.

Evaluation wrt
- Common Cellular Role, etc
- "ipr" correlates well to common cellular roles, localization, & expression
- "ipr" seems to work better than "ig2"

Stability in Protein Networks
Maslov & Sneppen, Science, 296:910-913, 2002

Part of the network of physical interactions reported by Ito et al., PNAS, 2001

- According to Maslov & Sneppen
  - Links betw high-connected proteins are suppressed
  - Links betw high- & low-connected proteins are favoured
- This decreases cross talks & increases robustness
Evaluation wrt “Many-few” Interactions

- Number of “Many-few” interactions increases when more “reliable” IPR threshold is used to filter interactions
- Consistent with the Maslov-Sneppen prediction

Evaluation wrt “Cross-Talkers”

- **A MIPS functional cat:**
  - | 02 | ENERGY
  - | 02.01 | glycolysis and gluconeogenesis
  - | 02.01.01 | glycolysis methylglyoxal bypass
  - | 02.01.03 | regulation of glycolysis & gluconeogenesis

- **First 2 digits is top cat**
- **Other digits add more granularity to the cat**
  ⇒ Compare high- & low- IPR pairs that are not co-localised to determine number of pairs that fall into same cat. If more high-IPR pairs are in same cat, then IPR works
Evaluation wrt “Cross-Talkers”

- **For top cat**
  - 148/257 high-IPR pairs are in same cat
  - 65/260 low-IPR pairs are in same cat
- **For fine-granularity cat**
  - 135/257 high-IPR pairs are in same cat.
  - 37/260 low-IPR pairs are in same cat

⇒ IPR works
⇒ IPR pairs that are not co-localized are real cross-talkers!

---

**Example Cross Talkers**

<table>
<thead>
<tr>
<th>ProteinA</th>
<th>Cellular Localization</th>
<th>ProteinB</th>
<th>Cellular Localization</th>
<th>Functional Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDR299w</td>
<td>nucleolar-protein</td>
<td>YLR096w</td>
<td>cytoplasm-centriole</td>
<td>Nuclear transport</td>
</tr>
<tr>
<td>VOL04c</td>
<td>nucleolus-recombination</td>
<td>YMR174c</td>
<td>spindle body-spindle pole component</td>
<td>Centrosome</td>
</tr>
<tr>
<td>YDL154c</td>
<td>nucleolus-recombination</td>
<td>YBR183c</td>
<td>cytoplasm-endo.</td>
<td>Meiosis and budding</td>
</tr>
<tr>
<td>VGL192w</td>
<td>nucleolus-protein</td>
<td>YHR057c</td>
<td>cytoplasm-mitotic</td>
<td>Development of nuclear-mitotic spore</td>
</tr>
<tr>
<td>YDR299w</td>
<td>nucleolar-protein</td>
<td>YPR065w</td>
<td>cytoplasm-ER-cytoplasm</td>
<td>Both in nuclear transport</td>
</tr>
<tr>
<td>YEL029c</td>
<td>vacuole-phosphorylated</td>
<td>YPL029c</td>
<td>cytoskeleton-muscin</td>
<td>Protein targeting and budding</td>
</tr>
</tbody>
</table>

Examples of interactions with high IRAP values (≥ 0.95) between non-co-localized proteins (“cross-talkers”) involved in the same cellular pathway.
Can local topology do better?  
The story of FS Weight

Guilt by Association of Common Interaction Partners

- Two proteins that have a large proportion of their interaction partners in common are likely to directly interact also

- In fact, this is a special case of the “alternative paths” used in the IPR index, because length-1 alternative paths = shared interaction partners
Czekanowski-Dice Distance

- **Functional distance between two proteins** \cite{Brun et al, 2003}

\[
D(u, v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v| + |N_u \cap N_v|}
\]

- $N_k$ is the set of interacting partners of $k$
- $X \Delta Y$ is symmetric diff betw two sets $X$ and $Y$
- Greater weight given to similarity

⇒ Similarity can be defined as

\[
S(u, v) = 1 - D(u, v) = \frac{2X}{2X + (Y + Z)}
\]

Is this a good measure if $u$ and $v$ have very diff number of neighbours?

Functional Similarity Estimate: FS-Weighted Measure

- **FS-weighted measure**

\[
S(u, v) = \frac{2|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v|} \times \frac{2|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v|}
\]

- $N_k$ is the set of interacting partners of $k$
- Greater weight given to similarity

⇒ Rewriting this as

\[
S(u, v) = \frac{2X}{2X + Y} \times \frac{2X}{2X + Z}
\]
Another way to improve using local topology information

The story of meso-scale network motifs
Motivation for “Meso Scale”

- These motifs are very local and very small
- Many processes in biological network are “meso-scale” (5-25 proteins)

⇒ Maybe we should also use meso-scale motifs?

What is a network motif?

- A network motif \( g \) in a PPI network \( G \) is a connected unlabelled undirected topological pattern of inter-connections that is repeated and unique in \( G \)

- Repeated: \( f_g \), the number of occurrences of \( g \) in \( G \), is more than threshold \( F \)

- Unique: \( s_g \), the number of times \( f_g \) exceeds \( f_{g,\text{rand},i} \) over total number of randomized networks considered, is more than threshold \( S \)
Example

Figure 1: Example graph $G$.

$\mathcal{T}_1$ $\mathcal{T}_2$ $\mathcal{T}_3$ $\mathcal{T}_4$

Figure 2: Size $2$ to size $5$ trees.

$\mathcal{T}_5$ $\mathcal{T}_6$ $\mathcal{T}_7$ $\mathcal{T}_8$

Figure 3: 3 trees $\mathcal{T}_5$ $\mathcal{T}_6$ $\mathcal{T}_7$.

Figure 4: Occurrences of $\mathcal{T}_4$ in $G$.

NeMoFinder: Discovery of Meso-Scale Motifs

![Graphical representation of NeMoFinder](image)
Motif Strength and PPI Reliability

- Strength of a size k motif $g$ is
  
  $$MS^k(g) = \frac{s_g \times f_g}{\max_k}$$

- Motif-strength PPI reliability index is a pair of possibly interacting protein $X \leftrightarrow Y$ is
  
  $$I(X \leftrightarrow Y) = \sum_{i=0}^{K} \sum_{i=0}^{n} MS^k(g_i) \times k$$

where $\max_k$ is max value of $s_g \times f_g$ over all size-k motifs

where $g_i$ are motifs involving the edge $X \leftrightarrow Y$, and $k$ is size of $g_i$

Evaluation wrt Common Cellular Role, etc

- Motif-strength PPI reliability index correlates well to common cellular roles, localization, & expression

  $$\Rightarrow$$ works as well as “ipr”
Some Observations

• Meso-scale motifs are more reliable than small local motifs (c.f. “ig_2”)
• Similar performance to “ipr”, but may have advantages if network is sparse (i.e., where few alternate paths are present)

• Btw, this is the first time size-12 network motifs are known to be extracted from yeast PPI network

Conclusions and Suggestions
Conclusions

• There are latent local & global network “motifs” that indicate likelihood of PPIs

• These network “motifs” can be exploited in computational elimination of false positives from high-throughput PPI data

• FS-Weight, CD-Distance & meso-scale motifs are effective topologically-based computational measure for assessing the reliability (false positives) of PPIs

Follow-Up Works

• Expectation maximization can be applied on FS-Weight, CD-Distance, etc to further increase their power for detecting false positives

• FS-Weight, CD-Distance, etc can be adapted to detect false negatives
Readings

- J. Chen et al, “NeMoFinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs”, *Proc. KDD 2006*
- G. Liu, J. Li, L. Wong. “Assessing and predicting protein interactions using both local and global network topological metrics”, *Proc GIW2008*

Any Questions?