Motivation

- Nature of high-throughput PPI expts
  - Proteins are taken out of their natural context!

- Can a protein interact with so many proteins simultaneously?

- A big “hub” and its “spokes” should probably be decomposed into subclusters
  - Each subcluster is a set of proteins that interact in the same space and time
  - Viz., a protein complex
Plan

• Motivation and Approaches

• PPI Network Cleansing based on PPI Topology

• Impact of Cleansing on PPI-based Protein Complex Prediction Methods

Approaches
Approaches to PPI-Based Protein Complex Prediction

<table>
<thead>
<tr>
<th>Type</th>
<th>RNSC</th>
<th>MCODE</th>
<th>MCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering, local search cost based</td>
<td>Local neighborhood density search</td>
<td>Flow simulation</td>
<td></td>
</tr>
<tr>
<td>Multiple assignment of protein</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Weighted edge</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- And several other methods....
- Recall vs precision is poor

Cause of Low Recall/Precision

<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>All: All methods</td>
<td>9347</td>
<td>64</td>
<td>49</td>
</tr>
<tr>
<td>A: Small scale Y2H</td>
<td>1801</td>
<td>73</td>
<td>62</td>
</tr>
<tr>
<td>A0: GY2H Uetz et al. (published results)</td>
<td>956</td>
<td>66</td>
<td>45</td>
</tr>
<tr>
<td>A1: GY2H Uetz et al. (unpublished results)</td>
<td>516</td>
<td>53</td>
<td>33</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>3695</td>
<td>41</td>
<td>13</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>1092</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>D1: Biochemical, in 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>D3: Immunological, direct</td>
<td>1025</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>D2: 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>93</td>
</tr>
</tbody>
</table>

Sprinzak et al., *JMB*, 327:919-923, 2003

- High level of noise
  ⇒ Need to clean up before protein complex prediction

Talk at OSCADD09, IMTECH, 22-26 March 2009. Copyright © 2009 by Limsoon Wong
Cause of Low Recall/Precision

- Trouble with “non-ball-like” complexes
  ⇒ Clique merging? Relative density? Core-n-attachment?

PPI Network Cleansing based on PPI Topology
Measures that correlate with function homogeneity and localization coherence

- Two proteins participating in same biological process are more likely to interact
- Two proteins in the same cellular compartments are more likely to interact

CD-distance & FS-Weight: Based on concept that two proteins with many interaction partners in common are likely to be in same biological process & localize to the same compartment

Czekanowski-Dice Distance (Brun et al, 2003)

- Given a pair of proteins \((u, v)\) in a PPI network
  - \(N_u\) = the set of neighbors of \(u\)
  - \(N_v\) = the set of neighbors of \(v\)
- \(CD(u,v) = \frac{2 | N_u \cap N_v |}{| N_u | + | N_v |}\)
- Consider relative intersection size of the two neighbor sets, not absolute intersection size
  - Case 1: \(|N_u|= 1\), \(|N_v|= 1\), \(|N_u \cap N_v| = 1\), CD\((u,v)\)=1
  - Case 2: \(|N_u|= 10\), \(|N_v|= 10\), \(|N_u \cap N_v| = 10\), CD\((u,v)\)=1
FSWeight (Chua et al, 2006)

- Try to overcome weakness of CD-distance

\[
FS(u,v) = \frac{2 | N_u \cap N_v |}{| N_u | + | N_u \cap N_v | + \lambda_u} \times \frac{2 | N_u \cap N_v |}{| N_v | + | N_u \cap N_v | + \lambda_v}
\]

- \( \lambda_u \) and \( \lambda_v \) penalize proteins with few neighbors
  - \( \lambda_u = \max \{0, \frac{\sum_{x \in G} | N_x |}{| V |} - | N_u |\} \), \( \lambda_v = \max \{0, \frac{\sum_{x \in G} | N_x |}{| V |} - | N_v |\} \)

- Suppose average degree is 4, then
  - Case 1: \( |N_u| = 1, |N_v| = 1, |N_u \cap N_v| = 1 \), \( FS(u,v) = \frac{4}{25} = 0.16 \)
  - Case 2: \( |N_u| = 10, |N_v| = 10, |N_u \cap N_v| = 10 \), \( FS(u,v) = 1 \)

---

Evaluation wrt Common Cellular Role, etc
A simpler formulation seems to work too…

Iterated CD-Distance (Liu et al, 2008)

- Variant of CD-distance that penalizes proteins with few neighbors

\[ wL(u,v) = \frac{2 |N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v} \]

\[ \lambda_u = \max\{0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_u|\} \]

\[ \lambda_v = \max\{0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_v|\} \]

- Suppose average degree is 4, then
  - Case 1: $|N_u| = 1$, $|N_v| = 1$, $|N_u \cap N_v| = 1$, $wL(u,v) = 0.25$
  - Case 2: $|N_u| = 10$, $|N_v| = 10$, $|N_u \cap N_v| = 10$, $wL(u,v) = 1$
A thought…

\[ wL(u,v) = \frac{2 |N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v} \]

• Weight of interaction reflects its reliability

⇒ Can we get better results if we use this weight to re-calculate the score of other interactions?

Iterated CD-Distance (Liu et al, 2006)

• \( wL^0(u,v) = 1 \) if \((u,v) \in G\), otherwise \( wL^0(u,v)=0 \)

• \( wL^1(u,v) = \frac{|N_u \cap N_v| + |N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v} \)

• \( wL^k(u,v) = \frac{\sum_{x \in N_u \cap N_v} wL^{k-1}(u,x) + \sum_{x \in N_u \cap N_v} wL^{k-1}(v,x)}{\sum_{x \in N_u} wL^{k-1}(u,x) + \lambda_u \sum_{x \in N_v} wL^{k-1}(v,x) + \lambda_v} \)

• \( \lambda^k_u = \max\{0, \sum_{x \in N_u} \sum_{y \in N_x} wL^{k-1}(x,y) - \sum_{x \in N_u} wL^{k-1}(u,x) \} \}

• \( \lambda^k_v = \max\{0, \sum_{x \in N_v} \sum_{y \in N_x} wL^{k-1}(x,y) - \sum_{x \in N_v} wL^{k-1}(v,x) \} \)
Validation

• DIP yeast dataset
  – Functional homogeneity is 32.6% for PPIs where both proteins have functional annotations and 3.4% over all possible PPIs
  – Localization coherence is 54.7% for PPIs where both proteins have localization annotations and 4.9% over all possible PPIs

• Let’s see how much better iterated CD-distance is over the baseline above, as well as over the original CD-distance/FS-weight

How many iteration is enough?

• Iterated CD-distance achieves best performance wrt functional homogeneity at k=2
• Ditto wrt localization coherence (not shown)
How many iteration is enough?

<table>
<thead>
<tr>
<th>noise level</th>
<th>$k$</th>
<th>$#$ common PPIs</th>
<th>$\text{avg rank diff}$</th>
<th>$\text{avg score diff}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>1</td>
<td>5669</td>
<td>540.21</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5870</td>
<td>144.86</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>5849</td>
<td>67.00</td>
<td>0.01</td>
</tr>
<tr>
<td>300%</td>
<td>1</td>
<td>5322</td>
<td>881.77</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5664</td>
<td>367.45</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>5007</td>
<td>249.85</td>
<td>0.02</td>
</tr>
<tr>
<td>500%</td>
<td>1</td>
<td>5081</td>
<td>1013.14</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5502</td>
<td>625.46</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>5008</td>
<td>317.33</td>
<td>0.05</td>
</tr>
<tr>
<td>1000%</td>
<td>$k=1$</td>
<td>4472</td>
<td>1187.10</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>$k=2$</td>
<td>5101</td>
<td>1021.69</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>$k=20$</td>
<td>5264</td>
<td>614.66</td>
<td>0.13</td>
</tr>
</tbody>
</table>

- Iterative CD-distance at diff $k$ values on noisy network
  $\Rightarrow$ # of iterations depends on amt of noise

Identifying False Positive PPIs

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%

- Iterated CD-distance is an improvement over previous measures for assessing PPI reliability
Identifying False Negative PPIs

- Iterated CD-distance is an improvement over previous measures for predicting new PPIs

5-Fold Cross-Validation

- **DIP core dataset**
  - Ave # of proteins in 5 groups: 986
  - Ave # of interactions in 5 training datasets: 16723
  - Ave # of interactions in 5 testing datasets: 486591
  - Ave # of correct answer interactions: 307

- **Measures:**
  - sensitivity = TP/(TP + FN)
  - specificity = TN/(TN + FP)
    - #negatives >> #positives, specificity is always high
    - >97.8% for all scoring methods
  - precision = TP/(TP + FP)
5-Fold X-Validation

- Iterated CD-distance is an improvement over previous measures for identifying false positive & false negative PPIs

Impact of Cleansing on PPI-Based Protein Complex Prediction Methods
### PPI-Based Complex Prediction Algorithms

<table>
<thead>
<tr>
<th>Type</th>
<th>RNSC</th>
<th>MCODE</th>
<th>MCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering, local search cost based</td>
<td>Local neighborhood density search</td>
<td>Flow simulation</td>
<td></td>
</tr>
<tr>
<td>Multiple assignment of protein</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Weighted edge</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- Issue: recall vs precision has to be improved
- Does a “cleaner” PPI network help?
- How to capture non-ball-like complexes?

### Cleaning PPI Network

- Modify existing PPI network as follow
  - Remove level-1 interactions with low weight
  - Add level-2 interactions with high weight
- Then run RNNS, MCODE, MCL, ..., as well as our own method CMC
Clustering based on Maximal Cliques

- Remove noise edges in input PPI network by discarding edges having low iterated CD-distance
- Augment input PPI network by addition of missing edges having high iterated CD-distance
- Predict protein complex by finding overlapping maximal cliques, and merging/removing them
- Score predicted complexes using cluster density weighted by iterated CD-distance

Validation Experiments

- Matching a predicted complex \( S \) with a true complex \( C \)
  - \( V_S \): set of proteins in \( S \)
  - \( V_C \): set of proteins in \( C \)
  - \( \text{Overlap}(S, C) = \frac{|V_S \cap V_C|}{|V_S \cup V_C|} \)
  - \( \text{Overlap}(S, C) \geq 0.5 \)
- Evaluation
  - Precision = matched predictions / total predictions
  - Recall = matched complexes / total complexes
- Datasets: combined info from 6 yeast PPI expts
  - #interactions: 20461 PPI from 4671 proteins
  - #interactions with >0 common neighbor: 11487
**Effecting of Cleaning on CMC**

- Cleaning by Iterated CD-distance improves recall & precision of CMC

**Noise Tolerance of CMC**

- If cleaning is done by iterating CD-distance 20 times, CMC can tolerate up to 500% noise in the PPI network!
Effect of Cleansing on MCL

- MCL benefits significantly from cleaning too
- Ditto for other protein complex prediction methods

### CMC vs Others

<table>
<thead>
<tr>
<th>Clustering Method</th>
<th>Adjust/CD</th>
<th>Aloy (#complexes: 63)</th>
<th>MIPS (#complexes: 162)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>avg size</td>
<td>#clusters</td>
<td>precision</td>
</tr>
<tr>
<td>CMC</td>
<td></td>
<td>0.83</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>121</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>124</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>146</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>64</td>
<td>56</td>
</tr>
<tr>
<td>CFluxor</td>
<td></td>
<td>13.84</td>
<td>0.228</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12.84</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12.66</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11.84</td>
<td>44</td>
</tr>
<tr>
<td>MCL</td>
<td></td>
<td>9.40</td>
<td>0.638</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12.84</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.31</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.75</td>
<td>52</td>
</tr>
<tr>
<td>MCode</td>
<td></td>
<td>7.31</td>
<td>0.840</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.42</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.67</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.28</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 3. The impact of the iterative scoring method on the performance of four clustering methods. For CMC, MCL, and CFluxor, we return only the top 4000 interactions, and no new interactions are added. For MCode, we retain all the interactions with non-zero score and add top 3000 new interactions with the highest score. The 3rd column is the number of iterations k of the iterative scoring method, and k=0 means the PPI network is unweighted. The 3rd column is the number of clusters generated, the 4th and 5th columns in the average size and co-localization score of generated clusters.

Characteristics of Unmatched Clusters

• At k = 2 …
• 85 clusters predicted by CMC do not match complexes in Aloy and MIPS

• Localization coherence score ~90%
• 65/85 have the same informative GO term annotated to > 50% of proteins in the cluster

⇒ Likely to be real complexes

What have we learned?

• Guilt by association of common interaction partners is useful for predicting
  – PPI cellular localization
  – Missing PPIs
  – Protein complexes

• Acknowledgement
  – Kenny Chua, Guimei Liu
Readings


- G. Liu, J. Li, L. Wong. “Assessing and predicting protein interactions using both local and global network topological metrics”, *Proc GIW2008*


Any Question?