CS4220: Knowledge Discovery Methods for Bioinformatics
Unit 6: Protein Complex Prediction

Limsoon Wong
Lecture Outline

• Overview of protein complex prediction
• A case study: MCL-CAw
• Impact of PPIN cleansing
• Detecting overlapping complexes
• Detecting sparse complexes
• Detecting small complexes
Overview of Protein Complex Detection from PPIN
“Assemblies” of Interacting Proteins

- Proteins interact to form “protein assemblies”
- These assemblies are like “protein machines”
  - Highly coordinated parts
  - Highly efficient

Protein assemblies
- Complexes
- Functional modules
- Intricate, ubiquitous, control many biological processes

Source: Sriganesh Srihari
Protein Interaction Networks

- Proteins come together & interact
- The collection of these interactions form a Protein Interaction Network or PPIN

Collection of such interactions in an organism

Valuable source of knowledge

Source: Sriganesh Srihari
Detection & Analysis of Protein Complexes in PPIN

- **PPIN derived from several high-throughput expt**
- **Space-time info is lost**
- **Space-time info is “recovered”**
- **Identifying embedded complexes**
- **Individual complexes (Some might share proteins)**
- **Embedded complexes identified from PPIN**
- **Entire module might be involved in the same function/process**

Source: Sriganesh Srihari
Identifying Complexes from PPIN: The Complete Picture

1. Affinity purification followed by MS for identifying “baits” and “preys” (in vitro)
2. Arriving at a close approximation to the in vivo network
3. Identifying complexes from the PPI network
Taxonomy of Protein Complex Prediction Methods

Complex detection methods

- Solely graph clustering
  - Merging and growing clusters
    - MCODE (Bader et al., 2003)
    - LCMA (Li et al., 2005)
    - CMC (Liu et al., 2009)
    - HACO (Wang et al., 2009)
  - Network partitioning
    - MCL (Dongen, 2000)
    - Pu et al. (2007)
    - Friedel et al. (2008)

- Graph clustering + Biological information
  - Core-attachment
  - Functional homogeneity
  - Evolutionary conservation
  - Co-operative and exclusive interactions

Kind of information used

Kind of algorithmic techniques used

Network alignment

- Merging and growing clusters
  - COACH (Wu et al., 2009)
  - CORE (Leung et al., 2009)
  - HUNTER (Chin et al., 2010)

- Merging and growing clusters
  - PCP (Chua et al., 2007)
  - DECAFF (Li et al., 2007)

- Network partitioning
  - RNSC (King et al., 2005)

- Network alignment
  - Ozawa et al. (2010)
  - Jung et al. (2010)

Source: Sriganesh Srihari
As researchers try to improve basic graph clustering techs, they also incorporate bio insights into the methods.
Graph Clustering: MCODE

- Good visualization
  - MCODE offered as a “plug-in” to Cytoscape
- Produces very few clusters
  - High accuracy, but low recall
- Performs well on highly filtered high-density PPIN
  - Low tolerance to noise

Graph Clustering: MCL

Repeated inflation and expansion separates the network into multiple dense regions

\[
(\Gamma M)_{pq} = (M_{pq})^{\gamma} / \sum_{i=1}^{k} (M_{iq})^{\gamma}
\]

• Popular software for general graph clustering
• Reasonably good for protein complex detection
• Highly scalable and fast; robust to noise


Markov Chains

To see how this works, an example:

In one time step, a random walker at node 1 has a 33% chance of going to node 2, 3, & 4, and 0% chance to nodes 5, 6, or 7.

From node 2, 25% chance for 1, 3, 4, 5 and 0% for 6 and 7.

Creating a transition matrix gives:

\[
\begin{pmatrix}
0 & .25 & .33 & .33 & 0 & 0 & 0 \\
.33 & 0 & .33 & .33 & .33 & 0 & 0 \\
.33 & .25 & 0 & .33 & 0 & 0 & 0 \\
.33 & .25 & .33 & 0 & 0 & 0 & 0 \\
0 & .25 & 0 & 0 & 0 & .5 & .5 \\
0 & 0 & 0 & .33 & 0 & .5 & 0 \\
0 & 0 & 0 & 0 & .33 & .5 & 0
\end{pmatrix}
\]

Also can be looked at as a probability matrix!
Markov Chains

- A simpler example: \[
\begin{pmatrix}
.6 & .2 \\
.4 & .8
\end{pmatrix}
\]

- Next time step: \( t_0 \rightarrow t_1 \rightarrow t_2 \)

\[
1 \rightarrow 1 \rightarrow 1 \quad + \quad 1 \rightarrow 2 \rightarrow 1
\]

\[
.6 \times .6 + .4 \times .2 = .44
\]

\[
\begin{pmatrix}
.6 & .2 \\
.4 & .8
\end{pmatrix} \quad \begin{pmatrix}
.6 & .2 \\
.4 & .8
\end{pmatrix} = \begin{pmatrix}
.44 & .28 \\
.56 & .72
\end{pmatrix} \quad \begin{pmatrix}
.35 & .32 \\
.65 & .68
\end{pmatrix} \rightarrow \begin{pmatrix}
.34 & .33 \\
.66 & .66
\end{pmatrix}
\]

\[
\text{eventually} \quad \begin{pmatrix}
.33 & .33 \\
.66 & .66
\end{pmatrix}
\]
MCL

- "Flow is easier within dense regions than across sparse boundaries, however, in the long run this effect disappears."
- During the earlier powers of the Markov Chain, the edge weights will be higher in links that are within clusters, and lower between the clusters.
- This means there is a correspondence between the distribution of weight over the columns and the clusterings.
MCL

- MCL deliberately boosts this effect by
  - Stopping partway in the Markov Chain
  - Then adjusting the transitions by columns. For each vertex, the transition values are changed so that
    - Strong neighbors are further strengthened
    - Less popular neighbors are demoted.

- This adjusting can be done by raising a single column to a non-negative power, and then re-normalizing.

- This operation is named “Inflation”

- (Taking the Markov Chain powers is named “Expansion”)
Evolutionary Insight: Conserved Subnests

- **Assumption**
  - Complexes are evolutionarily conserved

- **Form orthology network out of PPINs from multiple species**
- **Identify conserved subnetworks**
- **Verify if these are complexes**

Functional Info: RNSC & DECAFF

- **RNSC**
  - Iterative clustering based on optimizing a cost function
  - Post-process based on size, edge-density, & functional homogeneity

- **DECAFF**
  - Li et al. *CSB 2007*, pp. 157-168
  - Find dense local neighborhoods and identify local cliques
  - Merge cliques to produce candidate complexes
  - Post-process based on functional homogeneity
Core-Attachment Structure: COACH

- Perform well on high-density PPIN
  - Higher recall than MCODE & MCL
- List cores & attachments separately

Mutually Exclusive PPIs: SPIN

- +15% in precision & +10% in recall for MCL & MCODE using SPIN
- Limitation: Insufficient amt of domain-domain interaction data

Fig. 6. Comparisons among the known complexes and clusters predicted by LCMA's based on PPIN and SPIN. The gray ovals represent known complexes from MIPS, the quadrangle is a PPIN cluster, and the dotted quadrangles are SPIN clusters. A protein that appears in several complexes is underlined.

Jung et al., Bioinformatics, 26(3):385-391, 2010
Comparative Assessment

- Methods arranged in chronological order

⇒ Over the years, F1-measure have improved!

Li et al., *BMC Genomics*, 11(Suppl 1):S3, 2010

(Noisy, sparse, old)  (Good quality, dense, new)
“Plug” into Chronological Classification

- Adding biological info improves F1
Challenges

- Recall & precision of protein complex prediction algo’s have lots to be improved
- Does a “cleaner” PPI network help?
- How to capture “high edge density” complexes that overlap each other?
- How to capture “low edge density” complexes?
- How to capture small complexes?
A Case Study: MCL-CAw
Core-Attachment Modularity in Yeast Complexes

- **Cores**
  - High interactivity among each other
  - Highly co-expressed
  - Main functional units of complexes

- **Attachments**
  - Not co-expressed w/ cores all the time
  - Attach to cores & aid them in their functions
  - May be shared across complexes

Identify \textbf{dense} regions within PPI network

Identify \textbf{core-attachment} structures within these regions

Extract them out, discard the rest

---

Srihari et al. MCL-CAw: A refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure. \textit{BMC Bioinformatics}, 11:504, 2010
MCL-CAw: Main Steps

- Cluster PPI network using MCL hierarchically
- Identify core proteins within clusters
- Filter noisy clusters
- Recruit attachment proteins to cores
- Extract out complexes
- Rank the complexes

Step 1: Cluster by MCL Hierarchically

Why MCL?
• Simple, robust, scalable
• Find dense regions reasonably well
• Work on weighted networks

Why hierarchical?
• Some clusters are large, & amalgamate smaller ones
• Hierarchical clustering identifies these smaller clusters

Step 2: Identify Core Proteins in Clusters

- **Set of cores within a cluster:**
  - Essentially a $k$-core
  - But, with some additional restrictions

Protein $p \in \text{Core} \ (C_i)$ if:
1. $p$ has high degree w.r.t. $C_i$
2. $p$ has more neighbors within $C_i$ than outside

Protein $p \in \text{Core} \ (C_i)$ if:
1. In-degree of $p$ w.r.t. $C_i \geq \text{Avg in-degree of } C_i$
2. In-degree of $p$ w.r.t. $C_i > \text{Out-degree of } p \text{ w.r.t. } C_i$
   (Considering weighted degrees)
Step 3: Filter Noisy Clusters

- In accordance with our assumption that every complex we predict must have a core

- Discard noisy clusters (i.e., those w/o core)

Step 4: Identify Attachments to Cores

- Protein $p$ is an attachment to an acceptor cluster, if
  1. Non-core
  2. Has strong interactions with core proteins
  3. Stronger the interactions among cores, stronger have to be the interactions of $p$
  4. Large core sets, strong interactions to some, or weaker to many

$\text{Interactions}(p, \text{Core}(C_j)) \propto \text{Interactions}(\text{Core}(C_j))$
Step 4: Identify Attachments to Cores

\[ \text{Interactions}(p, \text{Core}(C_j)) \propto \text{Interactions}(\text{Core}(C_j)) \]

Protein \( p \in \text{Donor cluster } C_i \) is an attachment to Acceptor Core \( (C_j) \), if:

\[ l(p, \text{Core}(C_j)) \geq \alpha^* l(\text{Core} (C_j)) * [ |\text{Core}(C_j)| / 2 ]^{-\gamma} \]

Parameters \( \alpha \) and \( \gamma \) used to control effect of right-hand side
Step 5: Extract Complexes

Complex \( C = Core(C) \cup Attach(C) \)

- Attachment proteins may be shared between complexes
Step 6: Rank Predicted Complexes

- **Weighted density-based ranking of complexes**
  - Reliability of interactions within complex C
  - Size of complex C
  - Weighted density
    
    \[ \frac{\sum (\text{wt of interactions})}{|C| \times (|C| - 1)} \]

- **Unweighted density** → Blindly favors small complexes or complexes with large # of interactions
- **Weighted density** → More reliable complexes ranked higher

PPI Datasets for Evaluation of MCL-CAw

• **Unscored,**
  – G+K: Gavin and Krogan datasets combined

• **Scored**
  – G+K (ICD): Scoring G+K network by iterated CD distance
  – A few other edge weighting schemes are also used

If you don’t remember CD-distance, please refer to last lecture!
“Gold Standard” Benchmarks Complexes

- **CYC 08: 408 complexes**

- **MIPS: 313 complexes**,  

- **Aloy: 101 complexes**,  

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#cmplx</th>
<th>#proteins</th>
<th>size max</th>
<th>size avg</th>
<th>size median</th>
<th>density avg</th>
<th>density median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aloy</td>
<td>63</td>
<td>544</td>
<td>34</td>
<td>9.22</td>
<td>7</td>
<td>0.865</td>
<td>0.944</td>
</tr>
<tr>
<td>CYC08</td>
<td>148</td>
<td>1115</td>
<td>81</td>
<td>8.84</td>
<td>6</td>
<td>0.831</td>
<td>0.944</td>
</tr>
<tr>
<td>MIPS</td>
<td>156</td>
<td>1171</td>
<td>95</td>
<td>14.86</td>
<td>9</td>
<td>0.565</td>
<td>0.564</td>
</tr>
<tr>
<td>Combined</td>
<td>305</td>
<td>1543</td>
<td>95</td>
<td>11.85</td>
<td>7</td>
<td>0.692</td>
<td>0.800</td>
</tr>
</tbody>
</table>

Size > 3  
Measured based on BioGrid yeast physical PPIN
Evaluation of MCL-CAw

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. CMC</td>
<td>1.146</td>
<td>1.000</td>
</tr>
<tr>
<td>2. HACO</td>
<td>0.899</td>
<td>0.785</td>
</tr>
<tr>
<td>3. MCL-CAw</td>
<td>0.800</td>
<td>0.700</td>
</tr>
<tr>
<td>4. CORE</td>
<td>0.757</td>
<td>0.661</td>
</tr>
<tr>
<td>5. MCLO</td>
<td>0.734</td>
<td>0.641</td>
</tr>
<tr>
<td>6. MCL</td>
<td>0.717</td>
<td>0.626</td>
</tr>
<tr>
<td>7. COACH</td>
<td>0.515</td>
<td>0.450</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. MCL-CAw</td>
<td>1.595</td>
<td>1.000</td>
</tr>
<tr>
<td>2. HACO</td>
<td>1.536</td>
<td>0.962</td>
</tr>
<tr>
<td>3. CMC</td>
<td>1.516</td>
<td>0.950</td>
</tr>
<tr>
<td>4. MCLO</td>
<td>1.414</td>
<td>0.886</td>
</tr>
<tr>
<td>5. MCL</td>
<td>1.411</td>
<td>0.884</td>
</tr>
</tbody>
</table>

- CORE and COACH assume only unweighted networks

- **F1 values have increased for all methods upon scoring**
Strengths of MCL-CAw

• **Perform better than MCL**
  – Demonstrate effectiveness of adding biological insights (core-attachment structure)

• **Respond well to most affinity scoring schemes**
  – Always ranked among top 3 on all scored / weighted networks
  – Weighting of edges improves performance of MCL-Caw and other methods
    • Good to incorporate reliability info of the edges!
Limitations of MCL-CAw

• Amalgamation of closely-interacting complexes
  – “Inherited” from MCL
  – Lowers the recall

• Undetected sparse complexes
  – “Inherited” from MCL
  – Does not work when PPI is sparse
  – Less sensitive to very sparse complexes

• Undetected small complexes (size < 4)
  – Discards small predicted complexes as many are FP
Impact of PPIN Cleansing on Protein Complex Prediction
Cleaning PPI Network

• Modify existing PPI network as follow
  – Remove interactions with low weight
  – Add interactions with high weight

• Then run RNSC, MCODE, MCL, …, as well as our own method CMC
CMC: Clustering of 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


If you don’t remember CD-distance, please refer to last lecture!
Some Details of CMC

- Iterated CD-distance is used to weigh PPI’s

\[ w^k(u, v) = \frac{\sum_{x \in N_u \cap N_v} (w^{k-1}(x, u) + w^{k-1}(x, v))}{\sum_{x \in N_u} w^{k-1}(x, u) + \lambda^k_u + \sum_{x \in N_v} w^{k-1}(x, v) + \lambda^k_v} \]

- Clusters are ranked by weighted density

\[ \text{score}(C) = \frac{\sum_{u \in C, v \in C} w(u, v)}{|C| \cdot (|C| - 1)} \]

- Inter-cluster connectivity is used to decided whether highly overlapping clusters are merged or (the lower weighted density ones) removed

\[ \text{inter-score}(C_1, C_2) = \sqrt{\frac{\sum_{u \in (C_1 - C_2)} \sum_{v \in C_2} w(u, v)}{|C_1 - C_2| \cdot |C_2|} \cdot \frac{\sum_{u \in (C_2 - C_1)} \sum_{v \in C_1} w(u, v)}{|C_2 - C_1| \cdot |C_1|}} \]
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: 20,461 PPI from 4,671 proteins
  - #interactions with >0 common neighbor: 11,487
Effecting of Cleaning on CMC

- Cleaning by Iterated CD-distance improves recall & precision 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 methods...

Table 3. The impact of the iterative scoring method on the performance of four clustering methods. For CMC, MCL and CFinder, we retain only the top-6000 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 2nd 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 column is 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 $\sim 90\%$
- 65/85 have the same informative GO term annotated to $> 50\%$ of proteins in the cluster

$\Rightarrow$ Likely to be real complexes
Detecting Overlapping Protein Complexes from Dense Regions of PPIN
Overlapping Complexes in Dense Regions of PPIN

- Dense regions of PPIN often contain multiple overlapping protein complexes
- These complexes often got clustered together and cannot be corrected detected

- Two ideas to cleanse PPI network
  - Decompose PPI network by localisation GO terms
  - Remove big hubs

Idea I: Split by Localization GO Terms

• A protein complex can only be formed if its proteins are localized in same compartment of the cell

⇒ Use **general** cellular component (CC) GO terms to decompose a given PPI network into several smaller PPI networks

• Use “general” CC GO terms as it is easier to obtain rough localization annotation of proteins
  – How to choose threshold $N_{\text{GO}}$ to decide whether a CC GO term is “general”?
Effect of $N_{GO}$ on Precision

- Precision always improves under all $N_{GO}$ thresholds
Effect of $N_{GO}$ on Recall

- Recall drops when $N_{GO}$ is small due to excessive info loss
- Recall improves when $N_{GO} > 300$  
  $\Rightarrow$ Good to decompose by general CC GO terms

Idea II: Remove Big Hubs

- Hub proteins are those proteins that have many neighbors in the PPI network

- Large hubs are likely to be “date hubs”; i.e., proteins that participate in many complexes
  - Likely to confuse protein complex prediction algo

⇒ Remove large hubs before protein complex prediction
  - How to choose threshold $N_{hub}$ to decide whether a hub is “large”?
Effect of $N_{\text{hub}}$ on Recall

- Recall is affected when $N_{\text{hub}}$ is small, due to high info loss
- Not much effect on recall when $N_{\text{hub}}$ is large

Table 4. Number of hub proteins and PPIs removed under different $N_{\text{hub}}$.
Effect of $N_{hub}$ on Precision

- Precision of MCL & RNSC not much change
- Precision of IPCA & CMC improve greatly

Combining the Two Ideas

1. Let $\mathcal{C}$ be the set of clusters generated. Initially $\mathcal{C}$ is empty.

2. Remove hub proteins that have at least $N_{\text{hub}}$ neighbors from the given PPI network $G$. Let $G'$ be the resultant network.

3. Let $g_1, \ldots, g_m$ be the localization GO terms that are selected using threshold $N_{\text{GO}}$. For each $g_i$, do the following:
   
   • Remove proteins that are not annotated with $g_i$ from $G'$. Let $G''_i$ be the resultant network.
   • Apply a complex discovery algorithm on $G''_i$ to find clusters. Let $\mathcal{C}_i$ be the set of clusters generated.

   • $\mathcal{C} = \mathcal{C} \cup \mathcal{C}_i$

4. Remove duplicated clusters from $\mathcal{C}$. 

Effect of Combining $N_{GO}$ & $N_{hub}$


- RNSC doesn’t benefit further
- MCL, IPCA & CMC all gain further
Conclusions

- RNSC performs best (F1 = 0.353) on original PPI network; it also benefits much from CC GO term decomposition, but not from big-hub removal

- CMC performs best (F1 = 0.501) after PPI network preprocessing by CC GO term decomposition and big-hub removal

- But many complexes still cannot be detected…

Many complexes not detectable. Why?

- Among 305 complexes, 81 have density < 0.5, and 42 have density < 0.25
Many complexes not detectable. Why?

- **18 complexes w/ more than half of their proteins being isolated**
  - *Isolated vertex* connects to no other vertices in the complex

- **144 complexes w/ more than half of their proteins being loose**
  - *Loose vertex connects to < 50% of other vertices in the complex*

---

Many complexes not detectable. Why?

- For all four algo’s, 90% of detected complexes have a density > 0.5
- But many undetected complexes have a density < 0.5, and also have many loose vertices

Detecting Protein Complexes from Sparse Regions of PPIN
Sparse Complexes

~ 25% sparse complexes – “scattered” or low density

ANY algorithm based solely on topological will miss these sparse complexes!!

Source: Sriganesh Srihari
Noisy & Transient PPIs

• **Noise in PPI data**
  – Spuriously-detected interactions (false positives), and missing interactions (false negatives)

• **Transient interactions**
  – Many proteins that actually interact are not from the same complex, they bind temporarily to perform a function

• **Also, not all proteins in the same complex may actually interact with each other**
Cytochrome BC1 Complex

- Involved in electron-transport chain in mitochondrial inner membrane

- Discovery of this complex from PPI data is difficult
  - Sparseness of the complex’s PPI subnetwork
    - Only 19 out of 45 possible interactions were detected between the complex’s proteins
  - Many extraneous interactions detected with other proteins outside the complex
    - E.g., UBI4 is involved in protein ubiquitination, and binds to many proteins to perform its function.

![Complex Interaction Diagram]

Figure 1: PPI subgraph of the mitochondrial cytochrome bc1 complex. Nineteen interactions were detected between the ten proteins from the complex, while many extraneous interactions were detected. Five example proteins from transient interactions are shown: NAB2 and UBI4 are involved in mRNA polyadenylation and protein ubiquitination, while PET1, SHY1, and COX1 are mitochondrial membrane proteins that are also involved in the electron transport chain. The extraneous interactions around the complex make its discovery difficult. All such network figures were generated by Cytoscape [30].
Key idea to deal with sparseness

Augment physical PPI network with other forms of linkage that suggest two proteins are likely to integrate

Supervised Weighting of Composite Networks (SWC)

- Data integration
- Supervised edge weighting
- Clustering

Overview of SWC

1. Integrate diff data sources to form composite network

2. Weight each edge based on probability that its two proteins are co-complex, using a naïve Bayes model w/ supervised learning

3. Perform clustering on the weighted network

• Advantages
  – Data integration increases density of complexes
    • co-complex proteins are likely to be related in other ways even if they do not interact
  – Supervised learning
    • Allows discrimination betw co-complex and transient interactions
  – Naïve Bayes’ transparency
    • Model parameters can be analyzed, e.g., to visualize the contribution of diff evidences in a predicted complex

1. Integrate multiple data sources

- **Composite network**: Vertices represent proteins, edges represent relationships between proteins
- **There is an edge between proteins** $u$, $v$, if and only if $u$ and $v$ are related according to any of the data sources

<table>
<thead>
<tr>
<th>Data source</th>
<th>Database</th>
<th>Scoring method</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI</td>
<td>BioGRID, IntACT, MINT</td>
<td>Iterative AdjustCD.</td>
</tr>
<tr>
<td>L2-PPI (indirect PPI)</td>
<td>BioGRID, IntACT, MINT</td>
<td>Iterative AdjustCD</td>
</tr>
<tr>
<td>Functional association</td>
<td>STRING</td>
<td>STRING</td>
</tr>
<tr>
<td>Literature co-occurrence</td>
<td>PubMed</td>
<td>Jaccard coefficient</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Yeast</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Pairs</td>
<td>% co-complex</td>
</tr>
<tr>
<td>PPI</td>
<td>106328</td>
<td>5.8%</td>
</tr>
<tr>
<td>L2-PPI</td>
<td>181175</td>
<td>1.1%</td>
</tr>
<tr>
<td>STRING</td>
<td>175712</td>
<td>5.7%</td>
</tr>
<tr>
<td>PubMed</td>
<td>161213</td>
<td>4.9%</td>
</tr>
<tr>
<td>All</td>
<td>531800</td>
<td>2.1%</td>
</tr>
</tbody>
</table>
2. Supervised edge-weighting

- Treat each edge as an instance, where features are data sources and feature values are data source scores, and class label is “co-complex” or “non-co-complex”

<table>
<thead>
<tr>
<th>PPI</th>
<th>L2 PPI</th>
<th>STRING</th>
<th>Pubmed</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.56</td>
<td>451</td>
<td>0</td>
<td>“co-complex”</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>“non-co-complex”</td>
</tr>
</tbody>
</table>

- Supervised learning:
  1. Discretize each feature (Minimum Description Length discretization⁷)
  2. Learn maximum-likelihood parameters for the two classes:
    \[ P(F = f | \text{co-comp}) = \frac{n_{c,F=f}}{n_c} \]
    \[ P(F = f | \text{non-co-comp}) = \frac{n_{-c,F=f}}{n_{-c}} \]
    for each discretized feature value \( f \) of each feature \( F \)

- Weight each edge \( e \) with its posterior probability of being co-complex:
  \[
  \text{weight}(e) = P(\text{co-comp}|F_1 = f_1, F_2 = f_2, \ldots) \\
  = \frac{P(F_1 = f_1, F_2 = f_2, \ldots | \text{co-comp})P(\text{co-comp})}{Z} \\
  = \frac{\prod_i P(F_i = f_i | \text{co-comp})P(\text{co-comp})}{Z} \\
  = \frac{\prod_i P(F_i = f_i | \text{co-comp})P(\text{co-comp})}{\prod_i P(F_i = f_i | \text{co-comp})P(\text{co-comp}) + \prod_i P(F_i = f_i | \text{non-co-comp})P(\text{non-co-comp})}
  \]
3. Complex Discovery

- Weighted composite network used as input to clustering algorithms
  - CMC, ClusterONE, IPCA, MCL, RNSC, HACO
- Predicted complexes scored by weighted density

- The clustering algo’s generate clusters with low overlap
  - Only 15% of clusters are generated by two or more algo’s

⇒ Voting-based aggregative strategy, COMBINED:
  - Take union of clusters generated by the diff algo’s
  - Similar clusters from multiple algo’s are given higher scores
  - If two or more clusters are similar (Jaccard ≥ 0.75), then use the highest scoring one and multiply its score by the # of algo’s that generated it
Experiments

• **Weighting approaches:**
  – SWC vs BOOST, TOPO, STR, NOWEI

• **Evaluate performance on the 6 clustering algos and the COMBINED clustering strategy**

• **Real complexes for training and testing:** CYC200814 for yeast, CORUM15 for human

• **Evaluation**
  – How well co-complex edges are predicted
  – How well predicted complexes match real complexes

**Evaluation wrt Co-Complex Prediction**

**Figure 2** Precision-recall graph for classification of co-complex edges using the five weighting schemes. (a) Classification of yeast co-complex edges. SWC and BOOST achieve the highest recall through data integration. TOPO has high precision for its top-scoring edges, but these are clustered in a few complexes. SWC achieves higher precision than STR, except when too many edges are considered. BOOST classifies edges categorically, giving high scores to one set of edges with about 50% recall and 35% precision, and low scores to the remainder. (b) Classification of human co-complex edges. Recall and precision for human is much lower than for yeast. TOPO has higher precision than SWC, but its predicted edges are clustered in fewer complexes. BOOST classifies edges categorically, and its high-scoring edges achieve 7% recall, with comparable precision with SWC. NOWEI has slightly higher precision than STR, which has the lowest precision.
Evaluation wrt Yeast Complex Prediction
Evaluation wrt Human Complex Prediction
Example: Yeast BC1 Complex

PPI network

Composite network

- **protein from complex**
- **protein outside complex**
- **PPI**
- **STRING**
- **PUBMED**
- **SWC weighted edge**
- **predicted cluster**

SWC-weighted network

Likelihood network
Example: Human BRCA1-A complex

- SWC found a complex that included 5 extra proteins, of which 3 (BABAM1, BRE, BRCC3) have been included in the BRCA1-A complex.
High-confidence Predicted Complexes

Yeast

# of predictions

Biological process coherence

Human

# of predictions

Biological process coherence
Two Novel Predicted Complexes

(a) Yeast
- MMS1
- MMS22
- RTT101
- RTT107

(b) Human
- HCN4
- HCN2
- HCN3
- HCN1

- Novel yeast complex: Annotated with DNA metabolic process and response to stress, forms a complex called Cul8-RING which is absent in our ref set
- Novel human complex: Annotated with transport process, Uniprot suggests it may be a subunit of a potassium channel complex
## Novel complexes predicted

**Yeast**

<table>
<thead>
<tr>
<th>Biological process</th>
<th># complexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein metabolic process</td>
<td>49</td>
</tr>
<tr>
<td>RNA metabolic process</td>
<td>36</td>
</tr>
<tr>
<td>DNA metabolic process</td>
<td>15</td>
</tr>
<tr>
<td>Small molecule metabolic process</td>
<td>23</td>
</tr>
<tr>
<td>Regulation of metabolic process</td>
<td>11</td>
</tr>
<tr>
<td>Regulation of gene expression</td>
<td>8</td>
</tr>
<tr>
<td>Organelle organization</td>
<td>40</td>
</tr>
<tr>
<td>Transport</td>
<td>43</td>
</tr>
<tr>
<td>Response to stress</td>
<td>20</td>
</tr>
<tr>
<td>Response to chemical stimulus</td>
<td>7</td>
</tr>
<tr>
<td>Cell cycle process</td>
<td>11</td>
</tr>
</tbody>
</table>

**Human**

<table>
<thead>
<tr>
<th>Biological process</th>
<th># complexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein metabolic process</td>
<td>32</td>
</tr>
<tr>
<td>RNA metabolic process</td>
<td>29</td>
</tr>
<tr>
<td>DNA metabolic process</td>
<td>4</td>
</tr>
<tr>
<td>Small molecule metabolic process</td>
<td>19</td>
</tr>
<tr>
<td>Regulation of metabolic process</td>
<td>74</td>
</tr>
<tr>
<td>Regulation of gene expression</td>
<td>34</td>
</tr>
<tr>
<td>Organelle organization</td>
<td>19</td>
</tr>
<tr>
<td>Transport</td>
<td>38</td>
</tr>
<tr>
<td>Response to stress</td>
<td>28</td>
</tr>
<tr>
<td>Response to chemical stimulus</td>
<td>32</td>
</tr>
<tr>
<td>Cell cycle process</td>
<td>14</td>
</tr>
</tbody>
</table>
Conclusions

• Naïve-Bayes data-integration to predict co-complexed proteins
  – Use of multiple data sources increases density of complexes
  – Supervised learning allows discrimination betw co-complex and transient interactions

• Tested approach using 6 clustering algo’s
  – Clusters produced by diff algo’s have low overlap, combining them gives greater recall
  – Clusters produced by more algo’s are more reliable
Detecting Small Protein Complexes
There are many small complexes. Density-based methods cannot predict them from PPI networks.

Source: Osamu Maruyama

Strategy used by PPSampler to better identify small complexes

- Random sampling and constrain the distribution of predicted clusters wrt complex size!

Source: Osamu Maruyama
PPSampler (Proteins’ Partition Sampler)

- \( C \): partition of the set of all proteins
- \( f(C) \), scoring function of \( C \)
- \( P(C) \), probability of \( C \)

\[
P(C) \propto \exp\left(-\frac{f(C)}{T}\right)
\]

\( T \): temperature parameter

- Samples are generated from \( P(C) \) by Metropolis-Hastings algo

Source: Osamu Maruyama
Scoring function $f(C)$

$$f(C) \equiv -f_1(C) \cdot f_2(C) \cdot f_3(C)$$

- $f_1(C)$, weight of interactions in clusters of $C$
- $f_2(C)$, distribution of number of clusters of a size in $C$
- $f_3(C)$, number of proteins within clusters of size 2 or more in $C$

Source: Osamu Maruyama
Scoring subfunction $f_1(C)$

- $f_1(C) = \sum_{d \in C} f_1(d)$

$w(u, v)$: weight of the interaction between proteins $u$ and $v$.
$N$: upper bound on the size of a cluster

$$f_1(d) = \begin{cases} 
0 & \text{if } |d| = 1 \\
-\infty & \text{else if } |d| > N \text{ or } \exists u \in d, \forall v(\neq u) \in d, w(u, v) = 0 \\
\sum_{u, v(\neq u) \in d} w(u, v) & \text{otherwise}
\end{cases}$$

Source: Osamu Maruyama
Scoring subfunction $f_2(C)$

$$f_2(C) \equiv \prod_{i=2}^{N} \frac{1}{1 + i^2 \cdot (\psi(i) - \psi_C(i))^2}$$

$\psi_C(i)$: relative frequency of clusters of size $i$ in $C$

$\psi(i)$: predefined target number for relative frequency for size $i$

$\psi(i) \equiv \frac{i^{-\gamma}}{Z}$

$\gamma$: power-law param.

$Z$: norm. const.

Source: Osamu Maruyama
Scoring subfunction $f_3(C)$

$s(C)$: number of proteins within clusters of size 2 or more in $C$

$\lambda$: predefined target number for $s(C)$

$$f_3(C) \equiv \frac{1}{1 + \frac{(s(C) - \lambda)^2}{10^3}}$$

Source: Osamu Maruyama
Evaluation Metrics

\[ ov(s, u) = \frac{5}{\sqrt{7} \cdot 9} = 0.63 \]

\[
ov(s, u) = \begin{cases} 
\frac{|s \cap u|}{\sqrt{|s| \cdot |u|}} & \text{if } |s \cap u| \geq 2 \\
0 & \text{otherwise.}
\end{cases}
\]

\( s \) and \( u \) are matched if \( ov(s, u) \geq \eta \).

\[
\text{precision}(C, K, \eta) \equiv \frac{|\{d | d \in C, \exists k \in K, ov(d, k) \geq \eta\}|}{|C|}
\]

\[
\text{recall}(C, K, \eta) \equiv \frac{|\{k | k \in K, \exists d \in C, ov(k, d) \geq \eta\}|}{|K|}
\]

\[
\text{F–measure}(C, K, \eta) \equiv 2 \cdot \frac{\text{precision}(C, K, \eta) \cdot \text{recall}(C, K, \eta)}{\text{precision}(C, K, \eta) + \text{recall}(C, K, \eta)}
\]

Source: Osamu Maruyama
Experimental Configuration

The set of PPIs of WI-PHI (Kiener et al., 2007) is given as input to the above algorithms. It contains 49,607 non-self interactions with 5953 proteins. The mean of the degrees of the proteins is 16.7. Each interaction has a weight representing the reliability of the interaction, which is determined from various heterogeneous data sources, including results of tandem affinity purification coupled to mass spectrometry (TAP-MS), large-scale yeast two-hybrid studies, and small-scale experiments stored in dedicated databases. The weights of those interactions ranged from 6.624488528 to 146.5512397. Note that, if the weight of an interaction is higher, it is more reliable.

Source: Osamu Maruyama

• PPIs: WI-PHI
• Complexes: CYC2008
• Threshold: $\eta=0.45$
• Parameters for PPSampler: $\gamma=2$ in f2, $\lambda=2000$ in f3
Results for Small Complexes

Source: Osamu Maruyama

Result: Size 2

<table>
<thead>
<tr>
<th></th>
<th>MCL</th>
<th>MCODE</th>
<th>DPCLus</th>
<th>CMC</th>
<th>COACH</th>
<th>RRW</th>
<th>NWE</th>
<th>PPSampler</th>
</tr>
</thead>
<tbody>
<tr>
<td># protein</td>
<td>462</td>
<td>6</td>
<td>2</td>
<td>12</td>
<td>0</td>
<td>3648</td>
<td>1264</td>
<td>258</td>
</tr>
<tr>
<td># cluster</td>
<td>231</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>1824</td>
<td>632</td>
<td>129</td>
</tr>
<tr>
<td>precision</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
<td>0.20</td>
<td>0.30</td>
</tr>
<tr>
<td>recall</td>
<td>0.46</td>
<td>0.03</td>
<td>0.34</td>
<td>0.19</td>
<td>0.41</td>
<td>0.35</td>
<td>0.48</td>
<td>0.33</td>
</tr>
<tr>
<td>F1</td>
<td>0.06</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.11</td>
<td>0.29</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Result: Size 3

<table>
<thead>
<tr>
<th></th>
<th>MCL</th>
<th>MCODE</th>
<th>DPCLus</th>
<th>CMC</th>
<th>COACH</th>
<th>RRW</th>
<th>NWE</th>
<th>PPSampler</th>
</tr>
</thead>
<tbody>
<tr>
<td># protein</td>
<td>456</td>
<td>162</td>
<td>120</td>
<td>616</td>
<td>60</td>
<td>309</td>
<td>162</td>
<td>180</td>
</tr>
<tr>
<td># cluster</td>
<td>152</td>
<td>54</td>
<td>40</td>
<td>216</td>
<td>20</td>
<td>103</td>
<td>54</td>
<td>60</td>
</tr>
<tr>
<td>precision</td>
<td>0.09</td>
<td>0.11</td>
<td>0.08</td>
<td>0.02</td>
<td>0</td>
<td>0.44</td>
<td>0.80</td>
<td>0.53</td>
</tr>
<tr>
<td>recall</td>
<td>0.63</td>
<td>0.05</td>
<td>0.44</td>
<td>0.18</td>
<td>0.60</td>
<td>0.55</td>
<td>0.57</td>
<td>0.55</td>
</tr>
<tr>
<td>F1</td>
<td>0.15</td>
<td>0.07</td>
<td>0.13</td>
<td>0.04</td>
<td>0</td>
<td>0.49</td>
<td>0.67</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Source: Yong Chern Han
Performance Comparison: A Cautionary Tale

Prediction of all complexes

- Data: WI-PHI yeast PPI, weighted by reliability. ~50,000 PPIs, scores from 6.6 – 146.5
- Paper reported that PPSampler achieves higher precision and recall than CMC
- Found that this is true only if the data is not normalized to 0-1 range! When data is normalized, then CMC outperforms PPSampler!

Source: Osamu Maruyama

Source: Yong Chern Han
Must Read


• [cmc] Liu et al. **Complex Discovery from Weighted PPI Networks.** *Bioinformatics*, 25(15):1891--1897, 2009

• Liu et al. **Decomposing PPI Networks for Complex Discovery.** *Proteome Science*, 9(Suppl. 1):S15, 2011

• [MCL-CAw] Srihari et al. **MCL-CAw: A refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure.** *BMC Bioinformatics*, 11:504, 2010

Good to Read

Acknowledgements

• A lot of the slides for this lecture were adapted from ppt files given to me by Sriganesh Srihari and Osamu Maruyama

• A lot of the results presented here are from the work of Liu Guimei, Yong Chern Han, and Sriganesh Srihari

Lui Guimei  Yong Chern Han