Protein Complexes Prediction by sampling

Osamu Maruyama
Kyushu University
Contents

1. What’s the protein complex prediction problem?
2. First challenging difficulty – Small protein complexes are the majority
3. PPSampler2 (Proteins’ Partition)
4. Second challenging difficulty – Known complexes are overlapped with each other.
5. RocSampler (Regularizing overlaps of complexes)
6. Concluding remarks
1. *What’s the protein complex prediction problem?*
2. First challenging difficulty – Small protein complexes are the majority
3. PPSampler2 (Proteins’ Partition)
4. Second challenging difficulty – Known complexes are overlapped with each other.
5. RocSampler (Regularizing overlaps of complexes)
6. Concluding remarks
DNA, protein, and protein complex

Our body has a $10^{11}$ (hundred billion) cells.

A protein complex is a set of proteins connected by protein-protein interactions.
Problem: Protein complex prediction

Input: a protein-protein interaction (PPI) network (edge-weighted undirected graph).
Contents

1. What’s the protein complex prediction problem?
2. First challenging difficulty – Small protein complexes are the majority
3. PPSampler2 (Proteins’ Partition)
4. Second challenging difficulty – Known complexes are overlapped with each other.
5. RocSampler (Regularizing overlaps of complexes)
6. Concluding remarks
First challenging difficulty

• Small complexes (of size 2 and 3) are the majority of known complexes
  • CYC2008 (yeast protein complex database) has 408 complexes.
    • 172 (42%) are of size 2.
    • 87 (21%) are of size 3.
    • A human protein complex database has similar ratios.
• It is relatively difficult to identify small complexes.

The internal structure is poor!
Observation: Dense subgraphs are often overlapped with known protein complexes. This observation works better for larger protein complexes.
Typical approach: cluster-expansion method

1. Every protein or PPI forms an initial cluster

2. Cluster-expansion process:
   - Initial clusters
   - Expansion process
   - Stopped by a simple criterion

3. Post-processing for overlapping clusters

Affinity:
- Cohesiveness
- Random walk
- Etc.

This has no mechanism to control the sizes of predicted clusters
Contents

1. What’s the protein complex prediction problem?
2. First challenging difficulty – Small protein complexes are the majority
3. PPSampler2 (Proteins’ Partition)
4. Second challenging difficulty – Known complexes are overlapped with each other.
5. RocSampler (Regularizing overlaps of complexes)
6. Concluding remarks
How to predict small complexes

The distribution of sizes of known complexes is approximated by a power-law distribution.

• A human protein complex database has the same property.

• This property can be used as a prior knowledge.
Design of a term for regularizing the distribution of sizes of predicted clusters

A regularization term gives a force to fit the distribution of sizes of predicted clusters to a target power-law distribution.

\[ \psi_Y(s) = \frac{s^{-\gamma}}{\sum_{t=2}^{S_{\text{max}}} t^{-\gamma}} \]

: two-side truncated power-law distribution

\[ \psi_X(s) = \frac{|\{x \in X||x| = s\}|}{|X|} \]

: fraction of predicted clusters of size \( s \)

\[ \sum_{s=2}^{S_{\text{max}}} h_{\text{clu-size},s}(X, \gamma) \]

where

\[ h_{\text{clu-size},s}(X, \gamma) = \left( \psi_Y(s) - \psi_X(s) \right)^2 \]
A sample, $X$, is a partition of all proteins in an input PPI network. Any overlaps are not allowed.

Samples are generated from $P(X)$ by a Metropolis-Hastings algorithm.

$$P(X) \propto \exp \left( - \frac{f(X)}{T} \right)$$
### Scoring function $f(X) =$

<table>
<thead>
<tr>
<th>Optimization term</th>
<th>Regularization terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_{clu-den}(X)$</td>
<td>Sum of the densities of predicted clusters</td>
</tr>
</tbody>
</table>

**Optimization term**

**Regularization terms**

- $+ b(X)$
- $+ c_{pro-num} \cdot h_{pro-num}(X)$
- $+ c_{clu-size} \cdot \sum_{s=2}^{S_{max}} h_{clu-size,s}(X, \gamma)$

**Two Boolean basic constraints on predicted clusters:**
1. $\forall x \in X, |x| \leq S_{max}$
2. All $x \in X$ should be connected via PPIs.

Minimize the number of proteins of the predicted clusters:

$$h_{pro-num}(X) = \left( \lambda - \left| \bigcup_{x \in X} x \right| \right)^2$$

where $\lambda$ is a given target.

force of fitting the distribution of the sizes of predicted clusters to the power-law distribution of the scaling-exponent $\gamma$
\[ h_{\text{clu-den}}(X) \]

\[
density(x) = \frac{w(x)}{\sqrt{|x|}}
\]

where

\[
w(x) = \sum_{u,v \in x} w(u, v)
\]

\[
h_{\text{clu-den}}(X) = - \sum_{x \in X} \density(x)
\]

- The standard density is divided by \( |x| \cdot (|x| - 1)/2 \).

- \( \sqrt{|x|} \) is used to alleviate excessively severer evaluation of a larger cluster.
Let $x$ be a subset of proteins, called a predicted cluster.

- **Constraint 1:** $|x| \leq S_{\text{max}}$ (Max size of predicted clusters)
- **Constraint 2:** the vertex-induced subgraph of $G$ by $x$ is connected

- $b(x) = \begin{cases} 0 & \text{if both constraints are satisfied} \\ \infty & \text{otherwise} \end{cases}$
- $b(X) = \sum_{x \in X} b(x)$
Proposal distribution $Q(X'|X)\$.

Choose a protein $u$ randomly, and choose one of the two options randomly.

An empty singleton cluster of $u$ is created probabilistically.

Randomly choose another cluster according to the PPI weights of neighboring proteins.

$X'$ is also a partition.
Performance comparison

Input: WI-PHI PPIs
Gold standard complexes: CYC2008
Performance comparison on size-2 complexes by exact matching
Contents

1. What’s the protein complex prediction problem?
2. First challenging difficulty – Small protein complexes are the majority
3. PPSampler2 (Proteins’ Partition)
4. Second challenging difficulty – Known complexes are overlapped with each other
5. RocSampler (Regularizing overlaps of complexes)
6. Concluding remarks
Second challenging difficulty

Some known complexes are overlapped with each other.

CYC2008 has 216 overlaps between two complexes with 112 complexes.

<table>
<thead>
<tr>
<th>Overlap size</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>151</td>
<td>22</td>
<td>9</td>
<td>13</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Typical approach: cluster-expansion method

1. Every protein or PPI forms an initial cluster.

2. Cluster-expansion process:
   - In the first iteration, two clusters are formed, where 1 and 2 are in cluster 1, and 3, 4, and 5 are in cluster 2.
   - In the second iteration, cluster 1 expands to include 6, 7, and 8, while cluster 2 remains unchanged.
   - In the third iteration, cluster 2 expands to include 9, while cluster 1 remains unchanged.

3. Post-processing for overlapping clusters:
   - The final clusters are formed by merging the overlapping regions of clusters 1 and 2.

This approach, only by chance, finds good overlapping clusters.

Affinity:
- Cohesiveness
- Random walk
- Etc.

Stopped by a simple criterion.
Contents

1. What’s the protein complex prediction problem?
2. First challenging difficulty – Small protein complexes are the majority
3. PPSampler2 (Proteins’ Partition)
4. Second challenging difficulty – Known complexes are overlapped with each other
5. RocSampler (Regularizing overlaps of complexes)
6. Concluding remarks
$h_{clu-dis}(X)$: Term regularizing overlaps between predicted clusters

$$m_{x,x'} = \min\{|x|, |x'|\}$$

$$h_{clu-dis}(x,x') = \begin{cases} J(x,x') & m_{x,x'} \leq 3 \text{ and } |x \cap x'| \leq 1, \\ \infty & \text{or } m_{x,x'} \geq 4 \text{ and } \frac{|x \cap x'|}{m_{x,x'}} \leq \beta, \\ \infty & \text{otherwise} \end{cases}$$

where

$$J(x,x') = \frac{|x \cap x'|}{|x \cup x'|} \text{ (Jaccard index)}$$

$$h_{clu-dis}(X) = \sum_{x,x' \in X} h_{clu-dis}(x,x')$$

Parameter $\beta \in \{0.2, 0.3, 0.4\}$ is optimized for a PPI network.
Our new sampler: RocSampler (Regularizing Overlaps of Clusters) [2016]

Samples are generated from $P(X, \gamma)$ by Metropolis-Hastings algorithm-based simulated annealing algorithm.

$$P(X, \gamma) \propto \exp \left( - \frac{f(X, \gamma)}{T} \right)$$

A new term regularizing overlaps of predicted clusters is designed

The scaling exponent of the power-law is also optimized by sampling

$\gamma$
**Scoring function** $f(X, \gamma) =$

<table>
<thead>
<tr>
<th>Optimization term</th>
<th>Sum of the densities of predicted clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_{clu\text{-}den}(X)$</td>
<td></td>
</tr>
<tr>
<td>$+ b(X)$</td>
<td>Two Boolean basic constraints on predicted clusters</td>
</tr>
<tr>
<td>$+ c_{pro\text{-}num} \cdot h_{pro\text{-}num}(X)$</td>
<td>Minimize the number of proteins of the predicted clusters; $h_{pro\text{-}num}(X) =</td>
</tr>
<tr>
<td>$+ c_{clu\text{-}size} \cdot \sum_{s=2}^{s_{\text{max}}} h_{clu\text{-}size,s}(X, \gamma)$</td>
<td>force of fitting the distribution of the sizes of predicted clusters to the power-law distribution of the scaling-exponent $\gamma$</td>
</tr>
<tr>
<td>$+ c_{hy} \cdot h_{hy}(\gamma)$</td>
<td>Prior of $\gamma$; $h_{hy}(\gamma) = (\gamma - \gamma_0)^2$ with $\gamma_0 = 2.5$</td>
</tr>
<tr>
<td>$+ c_{clu\text{-}dis} \cdot h_{clu\text{-}dis}(X)$</td>
<td>Penalize overlaps between predicted clusters</td>
</tr>
</tbody>
</table>

**Regularization terms**
Simulated annealing

\[ T_0 = 1 \]
\[ T_l = T_{l-1} \times 0.999999 \]

If \( T \) is replaced with \( T_l \), the Metropolis-Hastings algorithm turns to be a simulated annealing algorithm.
Proposal function of Metropolis-Hastings algorithm

1. Randomly choose one of the 4 options
   1. Randomly add a new clusters of size 2 to $X$
   2. Randomly remove a cluster of size 2 in $X$
   3. Randomly add a new protein to a cluster in $X$
   4. Randomly remove a protein of a cluster in $X$

2. $\gamma = \min\{10^{-10}, \gamma + \varepsilon\}$ where $\varepsilon \sim N(0,0.001)$
# Computational experiment

<table>
<thead>
<tr>
<th></th>
<th>#Protein</th>
<th>#PPI</th>
<th>Degree</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>WI-PHI</td>
<td>5,953</td>
<td>49,607</td>
<td>16.7</td>
<td>N/A</td>
</tr>
<tr>
<td>Collins</td>
<td>1,622</td>
<td>9,074</td>
<td>11.2</td>
<td>Top 9,074</td>
</tr>
<tr>
<td>Krogan core</td>
<td>2,708</td>
<td>7,123</td>
<td>5.3</td>
<td>0.273</td>
</tr>
<tr>
<td>Krogan extended</td>
<td>3,672</td>
<td>14,317</td>
<td>7.8</td>
<td>0.101</td>
</tr>
<tr>
<td>Gavin</td>
<td>1,855</td>
<td>7,669</td>
<td>8.3</td>
<td>5</td>
</tr>
</tbody>
</table>

Parameter values of protein complex prediction methods are optimized.
Performance comparison on WI-PHI
Estimated value of $\gamma$

- The scaling exponent of the power-law regression curve of CYC2008 is 2.02.
- The estimated value of $\gamma$ is 1.91.
- Note that PPSampler2 uses $\gamma$ to be 2.
Difference in performance between PPSampler2 and RocSampler

Precision:
- PPSampler2: $\frac{145}{396} = 0.37$
- RocSampler: $\frac{147}{281} = 0.52$
PPSampler2 predicted more (insignificant) small clusters
On Collins2007
On Gavin2006
On Krogan2006Core
On Krogan2006Extended
On BioGRID (PPIs are unweighted)
On Collins PPIs
Contents

1. What’s the protein complex prediction problem?
2. First challenging difficulty – Small protein complexes are the majority
3. PPSampler2 (Proteins’ Partition)
4. Second challenging difficulty – Known complexes are overlapped with each other
5. RocSampler (Regularizing overlaps of complexes)
6. Concluding remarks
Concluding remarks

- First challenging difficulty
  - Small protein complexes are the majority
  - PPSampler2 (Proteins’ Partition)
    - We designed a term for fitting the distribution of sizes of predicted clusters to a power-law distribution.

- Second challenging difficulty
  - Known complexes are overlapped with each other
  - RocSampler (Regularizing overlaps of complexes)
    - We designed a term for regularizing overlaps between predicted clusters.

- But overlapping clusters are found only on the Collins PPI network among 5 networks.
  - Collins PPI network might has a special feature.
  - The current regularizer for overlaps is on the way?
Collaborators

Chasanah Kusumastuti Widita (PhD student)
Graduate School of Mathematics, Kyushu University
PPSampler2 (2013)

Yuki Kuwahara (Master course student)
Graduate School of Mathematics, Kyushu University
RocSampler (2016)