Protein Complex Inference enhanced by Text Mining

Lee Yu Ling Joanne

Overview

Hypothesis: Missing information might be found in Medline abstracts
Goal: Improve the prediction of protein complexes through text mining
Outline

- Background Information
- What was done
- Future Work
- Conclusion
- Questions

Background Information

- Real PPI
- Medline abstracts
- Text mining
- Abstract PPI
- PPI network
- Protein complex prediction model
- Evaluation
**PPI network**

Summarizes PPI data into graph
- Vertices represent proteins
- Edges represent protein interactions

**PPI data**

- **Experimental methods**
  - Yeast-two hybrid (Y2H)
  - Affinity Purification-Mass Spectrometry (AP-MS)
  - Protein Microarray

- **Non-experimental methods**
  - PPI database
  - Natural Language Processing (NLP)
    - Text mining
Background Information

Text mining

- Co-occurrences of two proteins in the same sentence (Co).
- Co and Dictionary of 4 verbs (Dict)
  - Interact, bind, complex, associate
  - Ono et al, 2001
- Bayesian Network (BN)
  - Chowdhary et al, 2009
Bayesian Network (BN)

- PPI triplet
  - 2 proteins + interacting word in same sentence
- Evaluated using trained BN and Bayes’ theorem.

\[
P(C|E) = \frac{P(E|C)P(C)}{P(E)}
\]

Bayes’ theorem

BN structure (Chowdhary et al, 2009)
Background Information

Dataset

Real PPIs
- Liu et al, 2009
  - 3295 proteins, 15900 interactions, 10458 interactions have common neighbours

Initial set of MEDLINE abstracts
- Li, 2008
  - 186798 non-empty abstracts

Augmenting set of MEDLINE abstracts
- 43516 non-empty abstracts
  - Mutually excludes the initial set of abstracts
Dataset

Reference complexes
- Liu et al, 2009
- Aloy (62 complexes), MIPS(164 complexes)
- AloyMIPS (213 complexes)
- Only complexes of size 4 and above
Protein Complex Prediction

- Markov Clustering (MCL)
  - van Dogen, 2000

- Molecular Complex Detection (MCODE)
  - Bader and Hogue, 2003

- Clustering based on Maximal Cliques (CMC)
  - Liu et al, 2009
  - Higher recall and precision

CMC

Step 1: Clique generation step

- PPIs → Iteratively weighting PPI network → Maximal cliques of min size = 4 → Rank cliques

Step 2: Clique merging step

- Inter-score < merge threshold → Yes: Merge cliques → No: Discard 1 clique
- Overlap-score ≥ overlap threshold → Yes: Rank cliques → No: Result

Result
Iterative weighting

Dataset

- Initial set
  - Augmenting set

- Medline abstracts

- Text mining

- Abstract PPI

- PPI network

- CMC

- Evaluation

Recall + Precision

Iteration = n
Merge thres = 0.25
Overlap thres = 0.5

<table>
<thead>
<tr>
<th>Number of iteration</th>
<th>Recall</th>
<th>Precision</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.403</td>
<td>0.015</td>
<td>0.294</td>
<td>0.026</td>
</tr>
<tr>
<td>10</td>
<td>0.403</td>
<td>0.016</td>
<td>0.281</td>
<td>0.025</td>
</tr>
<tr>
<td>20</td>
<td>0.403</td>
<td>0.016</td>
<td>0.281</td>
<td>0.025</td>
</tr>
<tr>
<td>30</td>
<td>0.403</td>
<td>0.016</td>
<td>0.281</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Recall and Precision for Co with different number of iteration
 Merge and overlap threshold

Dataset

- Initial set
- Augmenting set
- Medline abstracts
- Text mining
- Abstract PPI
- PPI network
- CMC

Evaluation

- AloyMIPS
- iteration = 20
- Merge thres = x
- Overlap thres = y

Recall + Precision

Precision-Recall of PPI network using Dict under different threshold values
Background Information

Dataset

- Initial set
- Augmenting set

Real PPI

AloyMIPS

Medline abstracts → Text mining → PPI network

CMC

Evaluation

Recall and Precision

- Recall: ratio of predicted clusters that match reference complexes
- Precision: ratio of reference complexes that match predicted clusters
Evaluation methods

Subset score
- Measures if 1 complex is a subset of another complex

Terminology
- High subset_score(Si,C) means large part of predicted cluster is a subset of reference complexes
- High subset_score(Ci,S) means large of reference complex is a subset of predicted complexes

Evaluation

Localization coherence (Lc)
- Cellular component of Gene Ontology (GO)
- Proteins that form complexes will seldom be in different cellular component
- Measures % of predicted clusters which have some % of proteins that occur together in the same cellular component
Background Information

Dataset

- Initial set
- Augmented set
- Medline abstracts
- Text mining
- PPI network
- CMC

Evaluation

- Recall + Precision
- Subset score
- Lc

Iteration = 20
Merge theses = 0.5
Overlap theses = 0.25

Outline

- Background Information
  - What was done
    - Which text mining method is best?
    - How to improve CMC?
    - How to deal with incomplete PPI data?
  
- Future Work
- Conclusion
- Questions
Which text mining is best?

![Diagram showing the process of text mining from Medline abstracts through to PPI network evaluation.]

### Table: Recall, precision and lc from 4 different PPI networks

<table>
<thead>
<tr>
<th>Method</th>
<th>Network size</th>
<th>Avg node degree</th>
<th>Number of clusters</th>
<th>Recall</th>
<th>Precision</th>
<th>Localization coherence (lc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI network of real PPIs</td>
<td>1836</td>
<td>3.86</td>
<td>186</td>
<td>0.474</td>
<td>0.333</td>
<td>At least 69% of clusters show 86% lc</td>
</tr>
<tr>
<td>PPI network of Dict</td>
<td>2594</td>
<td>3.02</td>
<td>482</td>
<td>0.249</td>
<td>0.095</td>
<td>At least 66% of clusters show 78% lc</td>
</tr>
<tr>
<td>Combined network of real PPIs and Dict</td>
<td>3225</td>
<td>4.02</td>
<td>617</td>
<td>0.549</td>
<td>0.154</td>
<td>At least 66% of clusters show 84% lc</td>
</tr>
<tr>
<td>PPI network of BN</td>
<td>1283</td>
<td>1.53</td>
<td>138</td>
<td>0.061</td>
<td>0.065</td>
<td>At least 60% of clusters show 80% lc</td>
</tr>
</tbody>
</table>

Recall, precision and lc from 4 different PPI networks
**Which text mining is best?**

- Largest increase in recall from Dict to real+Dict
  - Recall is likely to be limited by number of abstracts

- Highest recall in real+Dict
  - PPI abstracts may fill in missing edges of PPI network
  - Helps to predict more protein clusters that match the AloyMIPS

Graph of real complex 420
### Which text mining is best?

<table>
<thead>
<tr>
<th>Method</th>
<th>Network size</th>
<th>Avg node degree</th>
<th>Number of clusters</th>
<th>Recall</th>
<th>Precision</th>
<th>Localization coherence (lc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI network of real PPIs</td>
<td>1836</td>
<td>3.86</td>
<td>186</td>
<td>0.474</td>
<td>0.333</td>
<td>At least 69% of clusters show 86% lc</td>
</tr>
<tr>
<td>PPI network of Dict</td>
<td>2594</td>
<td>3.02</td>
<td>482</td>
<td>0.249</td>
<td>0.095</td>
<td>At least 66% of clusters show 78% lc</td>
</tr>
<tr>
<td>Combined network of real PPIs and Dict</td>
<td>3225</td>
<td>4.02</td>
<td>617</td>
<td>0.549</td>
<td>0.154</td>
<td>At least 66% of clusters show 84% lc</td>
</tr>
<tr>
<td>PPI network of BN</td>
<td>1283</td>
<td>1.53</td>
<td>138</td>
<td>0.061</td>
<td>0.065</td>
<td>At least 60% of clusters show 80% lc</td>
</tr>
</tbody>
</table>

Recall, precision and lc from 4 different PPI networks

### Which text mining is best?
- **Highest average node degree of real+Dict**
  - Combined network is better than individual
  - CMC uses clique finding strategy
Which text mining is best?

<table>
<thead>
<tr>
<th>Method</th>
<th>Network size</th>
<th>Avg node degree</th>
<th>Number of clusters</th>
<th>Recall</th>
<th>Precision</th>
<th>Localization coherence (lc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI network of real PPis</td>
<td>1836</td>
<td>3.86</td>
<td>186</td>
<td>0.474</td>
<td>0.333</td>
<td>At least 69% of clusters show 86% lc</td>
</tr>
<tr>
<td>PPI network of Dict</td>
<td>2594</td>
<td>3.02</td>
<td>482</td>
<td>0.249</td>
<td>0.095</td>
<td>At least 66% of clusters show 78% lc</td>
</tr>
<tr>
<td>Combined network of real PPis and Dict</td>
<td>3225</td>
<td>4.02</td>
<td>617</td>
<td>0.549</td>
<td>0.154</td>
<td>At least 66% of clusters show 84% lc</td>
</tr>
<tr>
<td>PPI network of BN</td>
<td>1283</td>
<td>1.53</td>
<td>138</td>
<td>0.061</td>
<td>0.065</td>
<td>At least 60% of clusters show 80% lc</td>
</tr>
</tbody>
</table>

Recall, precision and lc from 4 different PPI networks

Which text mining is best?

Analyzing predicted clusters

Venn diagram of correctly predicted clusters from 2 different networks

- Combined network is performing reasonably well
- 20% more predicted clusters in real+Dict network
Which text mining is best?
Analyzing PPI

- 32497 Dict abstract PPIs
- 15900 Real PPIs

Comparison result
- 32493 abstract PPIs not in real PPIs
- 15896 real PPIs not in abstract PPIs

The two set have little overlap
- Abstracts can fill missing PPI
- Considered too few abstracts

Which text mining is best?
Analyzing PPI

- Manual verification
  - Randomly choosing PPIs from abstracts

<table>
<thead>
<tr>
<th>Number of PPIs</th>
<th>Definitely not interact</th>
<th>Definitely interact</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>161</td>
<td>21</td>
<td>95</td>
<td>45</td>
</tr>
</tbody>
</table>

- Odds that an edge derived from abstract is real
  - 4:1
How to improve CMC?

Dataset

- Initial set
- Augmented set

Real PPI

- Abstract PPI network
- CMC

Evaluation

Recall + Precision
Subset score
Lc

Text mining

Medline abstracts

Dict

Iteratively weighting PPI network

Maximal cliques of min size = 4

Rank cliques

Step 1: Clique generation step

PPIs

Step 2: Clique merging step

Inter-score < merge threshold

Overlap-score ≥ overlap threshold

Discard 1 clique

Yes

No

Merge cliques

Result

Iteration = 20
Merge thres = 0.5
Overlap thres = 0.25

Dict

Iteration = 20
Merge thres = 0.5
Overlap thres = 0.25

Recall + Precision
Subset score
Lc
How to improve CMC?

Iterative removal of non-hub proteins

**Iteration n = 0 to 5**

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Network size</th>
<th>Avg node degree</th>
<th>Number of clusters</th>
<th>Recall</th>
<th>Precision</th>
<th>Localization coherence (lc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3225</td>
<td>4.02</td>
<td>617</td>
<td>0.549</td>
<td>0.154</td>
<td>At least 66% show 84% lc</td>
</tr>
<tr>
<td>1</td>
<td>1514</td>
<td>3.34</td>
<td>617+163=780</td>
<td>0.559</td>
<td>0.145</td>
<td>At least 69% show 84% lc</td>
</tr>
<tr>
<td>2</td>
<td>1339</td>
<td>3.42</td>
<td>780+29=809</td>
<td>0.559</td>
<td>0.142</td>
<td>At least 69% show 84% lc</td>
</tr>
<tr>
<td>3</td>
<td>999</td>
<td>2.89</td>
<td>809+77=886</td>
<td>0.563</td>
<td>0.132</td>
<td>At least 70% show 83% lc</td>
</tr>
<tr>
<td>4</td>
<td>901</td>
<td>2.88</td>
<td>886+30=916</td>
<td>0.563</td>
<td>0.13</td>
<td>At least 71% show 84% lc</td>
</tr>
<tr>
<td>5</td>
<td>783</td>
<td>2.65</td>
<td>916+41=957</td>
<td>0.563</td>
<td>0.126</td>
<td>At least 71% show 84% lc</td>
</tr>
</tbody>
</table>

Recall, precision and lc after different iteration of non-hub removal
subset score of AloyMIPS before and after iterated removal of non-hub proteins

- 4 complexes improved their score while none decreased

3D graph of subset evaluation after iterated removal of non-hub proteins
How to improve CMC?

Graph of real complex 520.20

How to improve CMC?

Graph of real complex 500 and predicted cluster C21
How to improve CMC?
- Cliques as a basis is stringent
- PPI data is incomplete

How to deal with incomplete PPI?

![Diagram explaining the process of dealing with incomplete PPI data.](image)
How to deal with incomplete PPI?

Future Work

- Evaluation by pathway coherence
- Predicting protein complexes based on largest k-connected sub-graphs
  - Connected sub-graph with size greater than k and will remain connected after deleting k nodes
- Improving the selection of abstracts for augmentation
  - Bayesian Inference
Conclusion

- 3 rule-based methods of PPI extraction
  - Co, Dict, BN
  - Real PPIs + Dict network fared better
- Noisy edges are pruned away by removing non-hub proteins
  - Prediction of greater number of complexes that were likely to be real
- Augmentation improved the prediction of some complexes

Questions