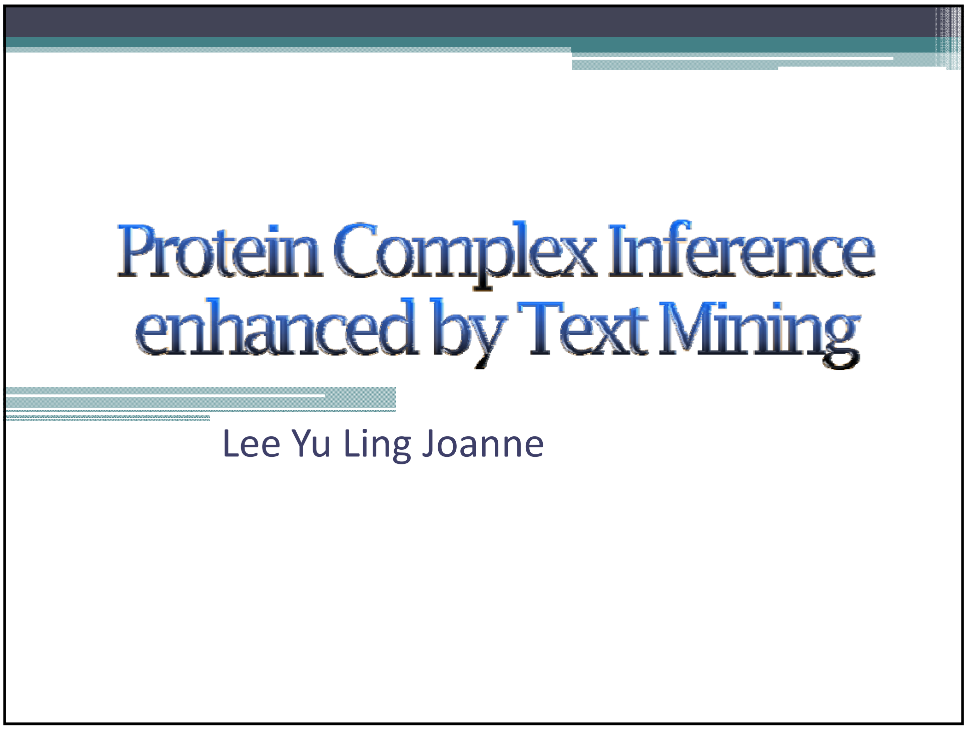
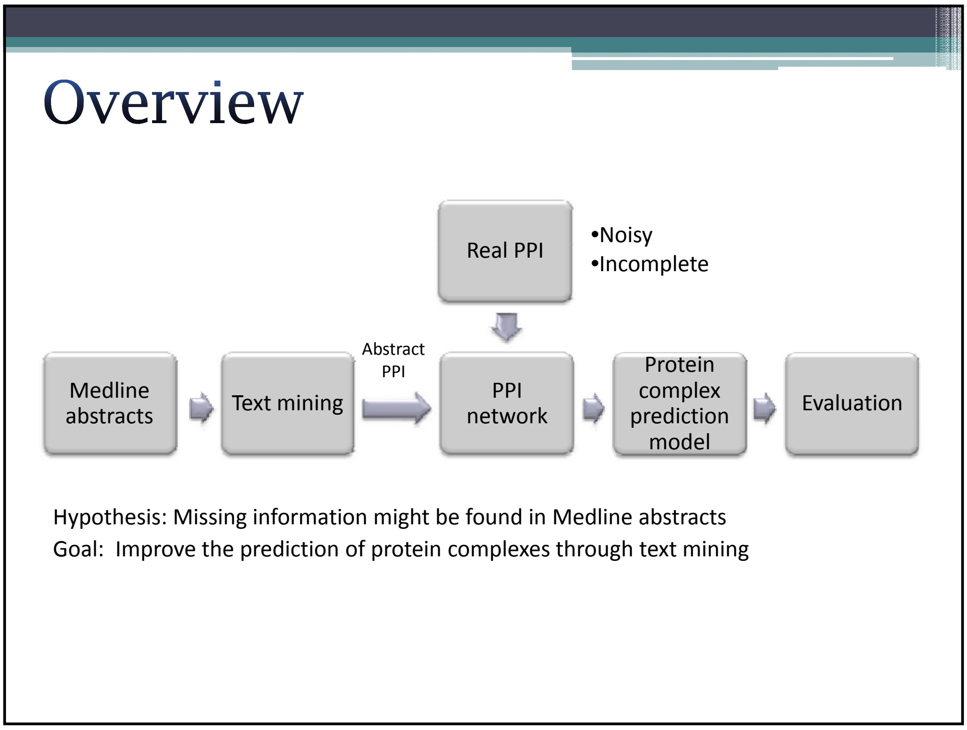


Protein Complex Inference enhanced by Text Mining

Lee Yu Ling Joanne



Overview

Real PPI

•Noisy
•Incomplete

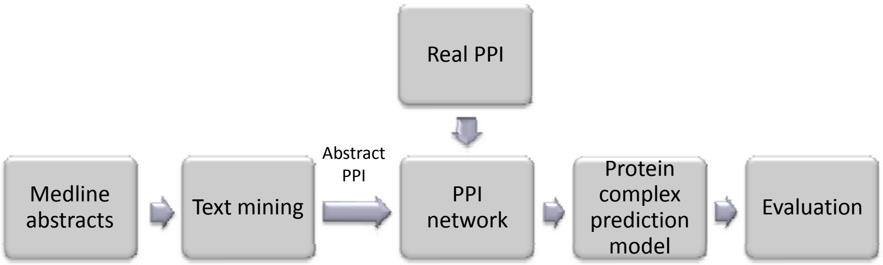
Medline abstracts → Text mining → Abstract PPI → PPI network → Protein complex prediction model → Evaluation

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

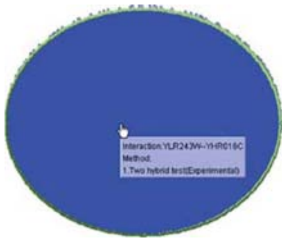
Outline

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

Background Information



PPI network

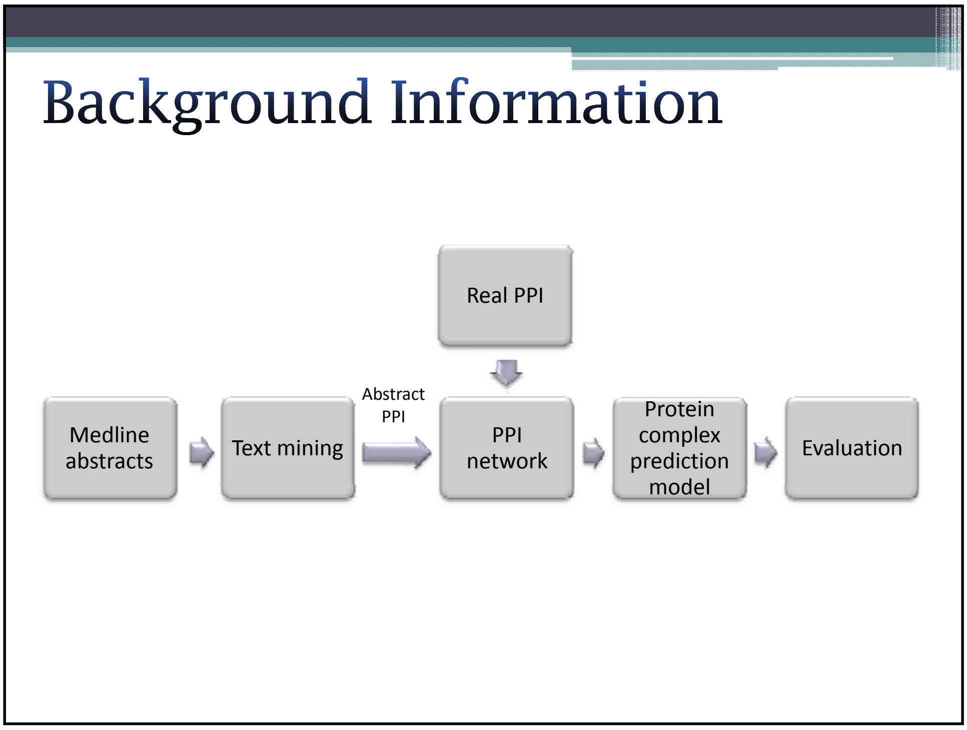


PPI network (Yeast-two hybrid) of yeast (Hu et al, 2004)

- 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

Real PPI

Medline abstracts

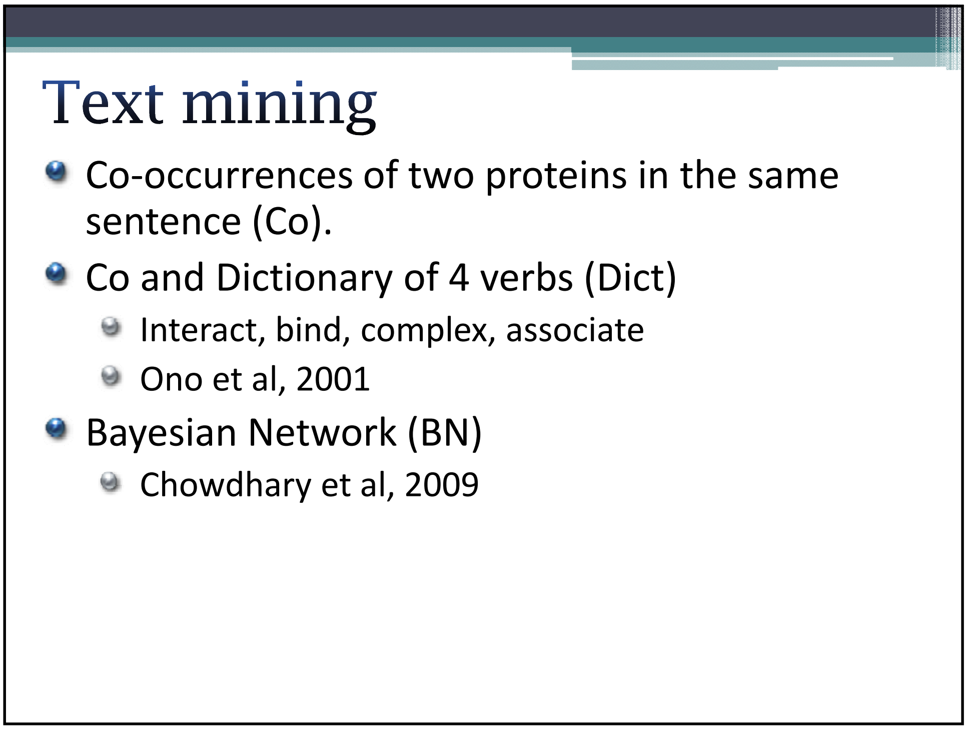
Text mining

Abstract PPI

PPI network

Protein complex prediction model

Evaluation



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

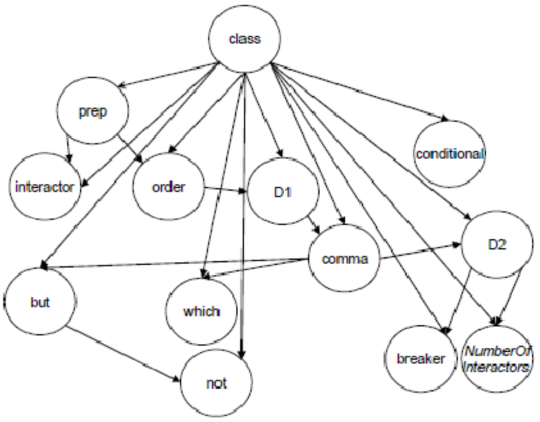
Bayesian Network (BN)

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

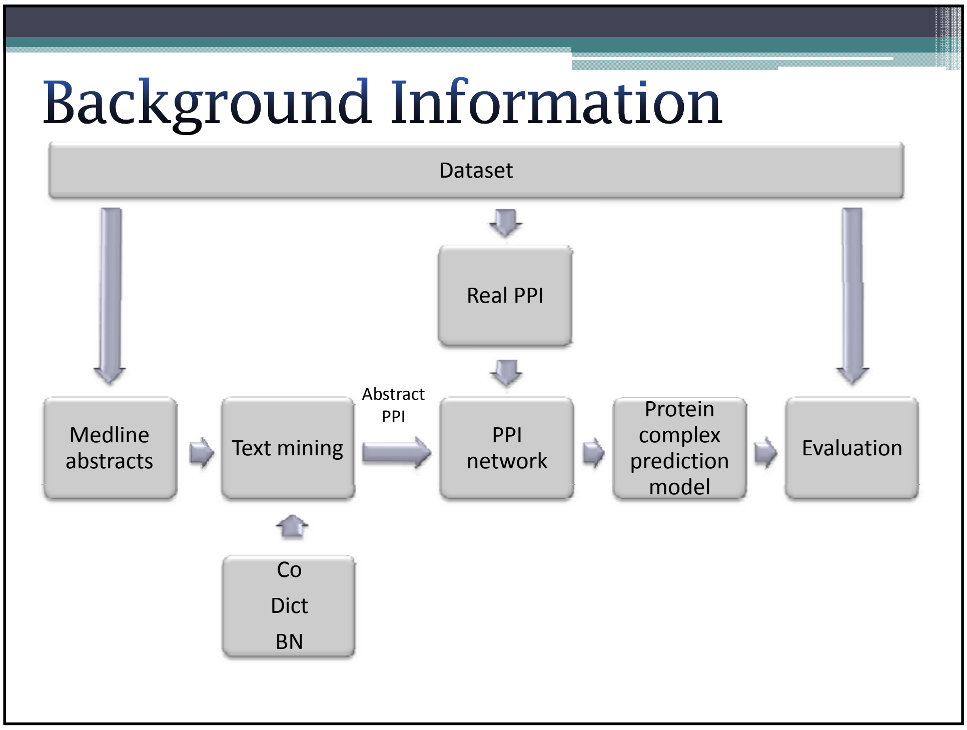
$$P(C|E) = \frac{P(E|C)P(C)}{P(E)}$$

Bayes' theorem

Bayesian Network (BN)



BN structure (Chowdhary et al, 2009)



Background Information

Dataset

Real PPI

Medline abstracts

Text mining

Abstract PPI

PPI network

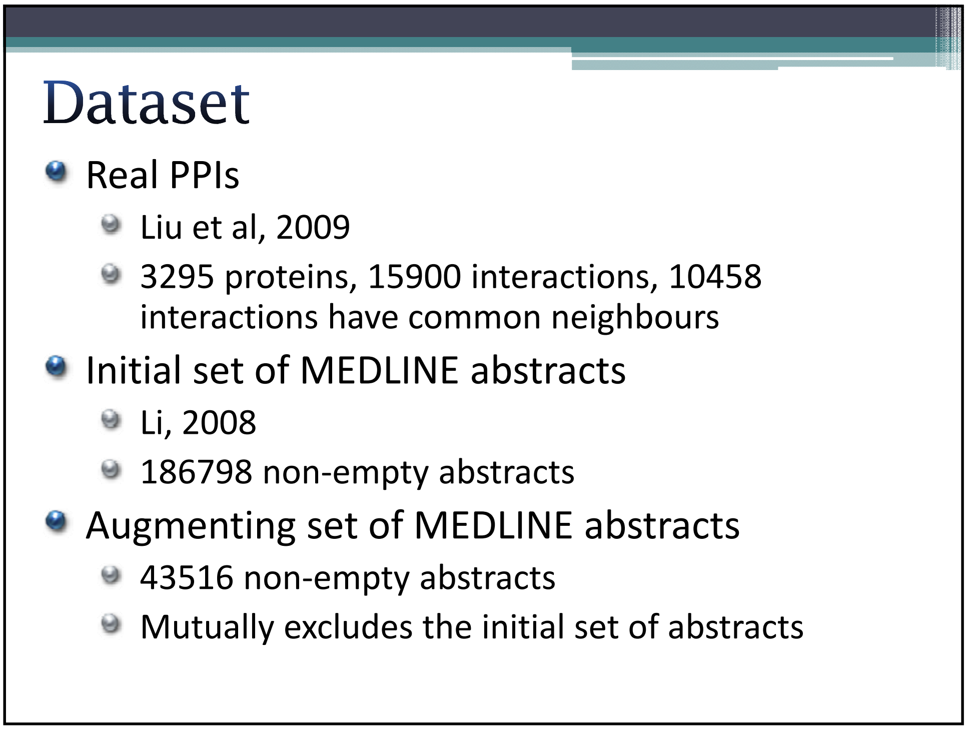
Protein complex prediction model

Evaluation

Co

Dict

BN



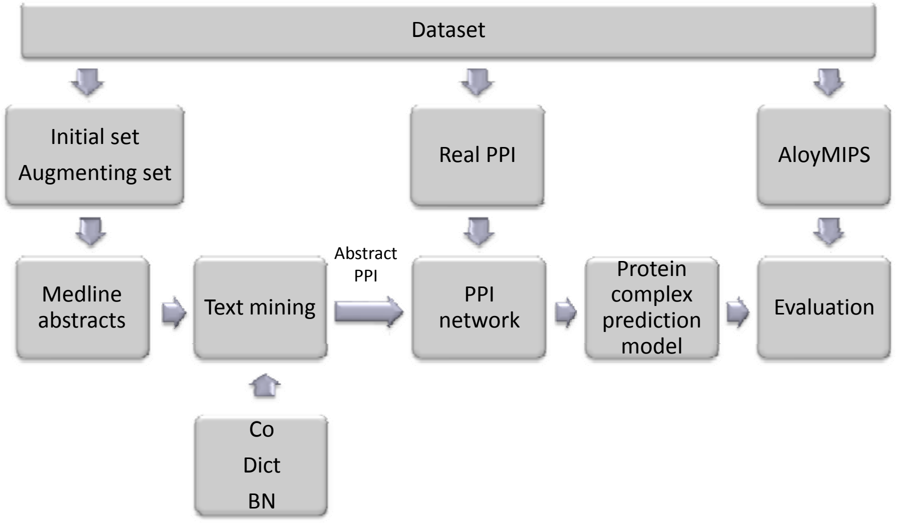
Dataset

- 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

Background Information



Dataset

Initial set
Augmenting set

Real PPI

AloyMIPS

Medline abstracts

Text mining

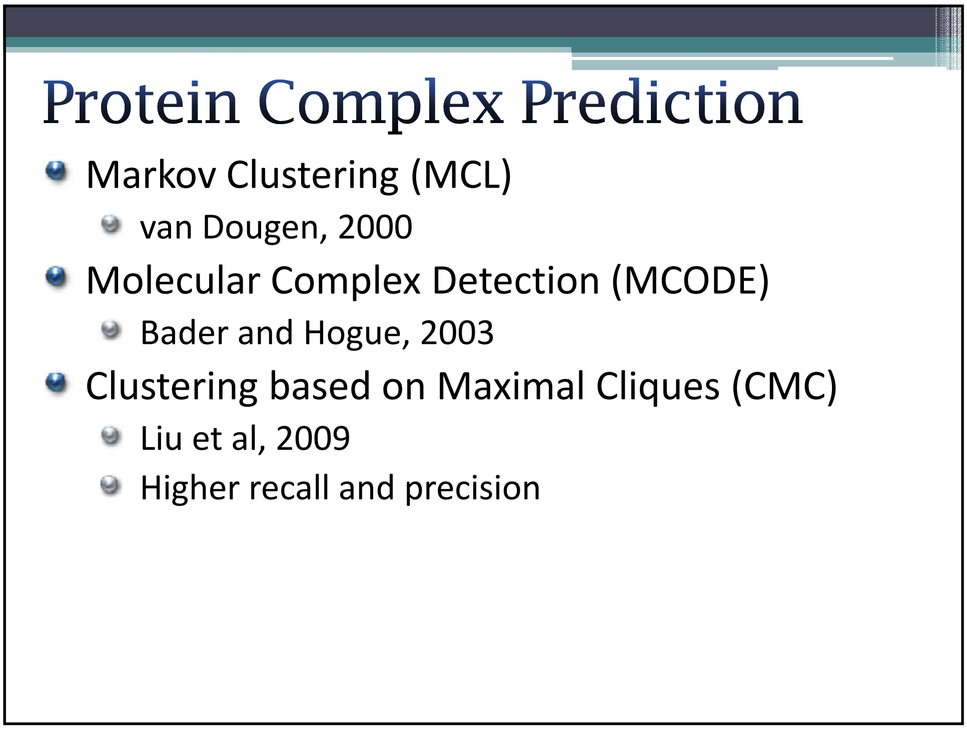
Abstract PPI

PPI network

Protein complex prediction model

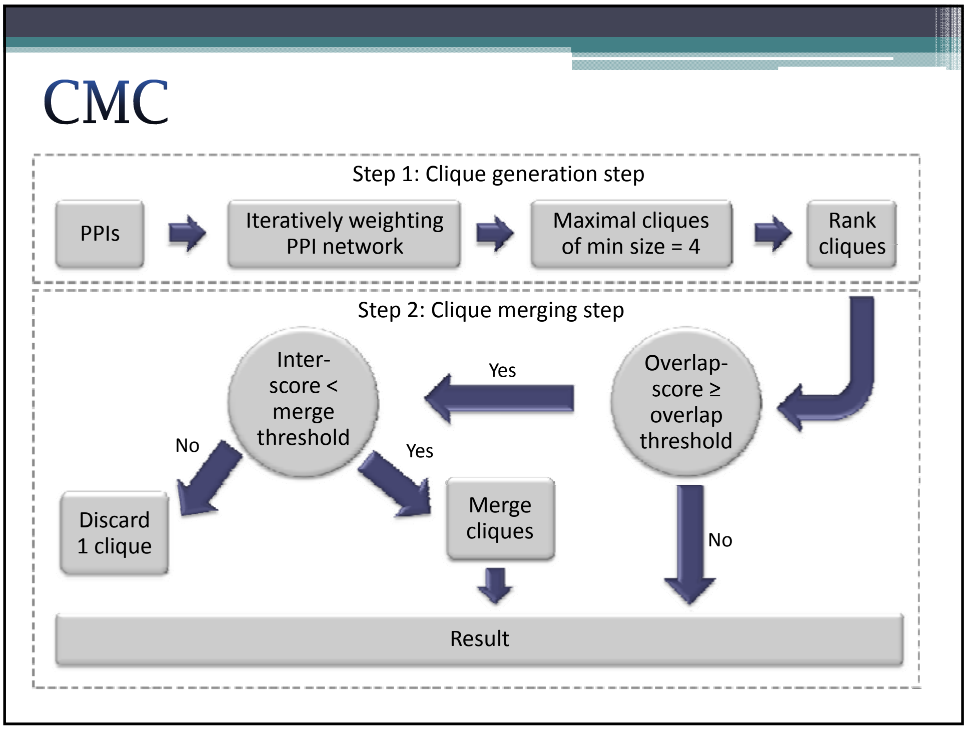
Evaluation

Co
Dict
BN



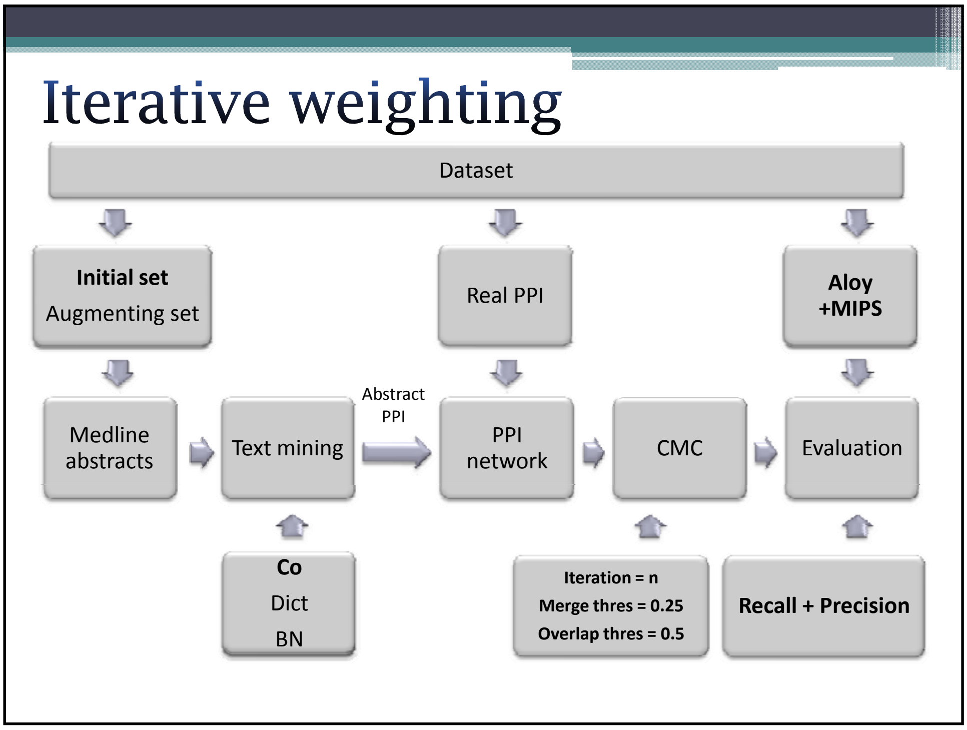
Protein Complex Prediction

- Markov Clustering (MCL)
  - van Dougen, 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
Overlap-score ≥ overlap threshold
Yes
Inter-score < merge threshold
No
Discard 1 clique
Yes
Merge cliques
No
Result



Iterative weighting

Dataset

Initial set
Augmenting set

Real PPI

Aloy
+MIPS

Medline abstracts

Text mining

Abstract PPI

PPI network

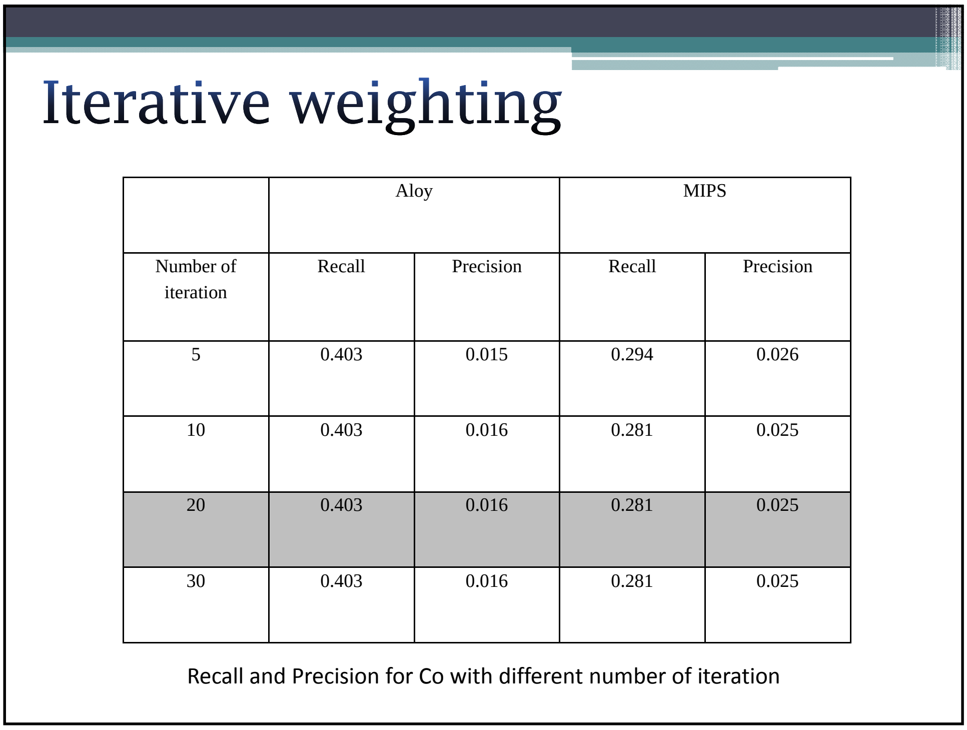
CMC

Evaluation

Co
Dict
BN

Iteration = n
Merge thres = 0.25
Overlap thres = 0.5

Recall + Precision

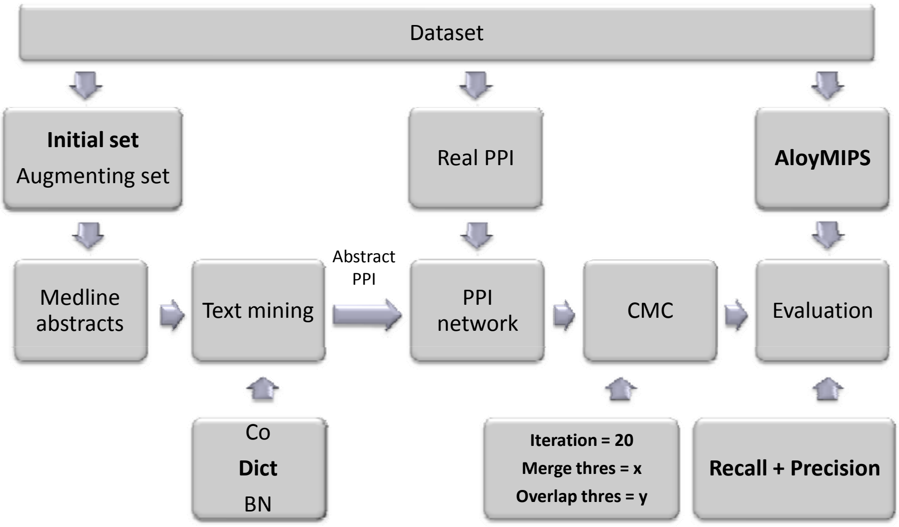


Iterative weighting

	Aloy		MIPS	
Number of iteration	Recall	Precision	Recall	Precision
5	0.403	0.015	0.294	0.026
10	0.403	0.016	0.281	0.025
20	0.403	0.016	0.281	0.025
30	0.403	0.016	0.281	0.025

Recall and Precision for Co with different number of iteration

Merge and overlap threshold



Dataset

Initial set
Augmenting set

Real PPI

AloyMIPS

Medline abstracts

Text mining

Abstract PPI

PPI network

CMC

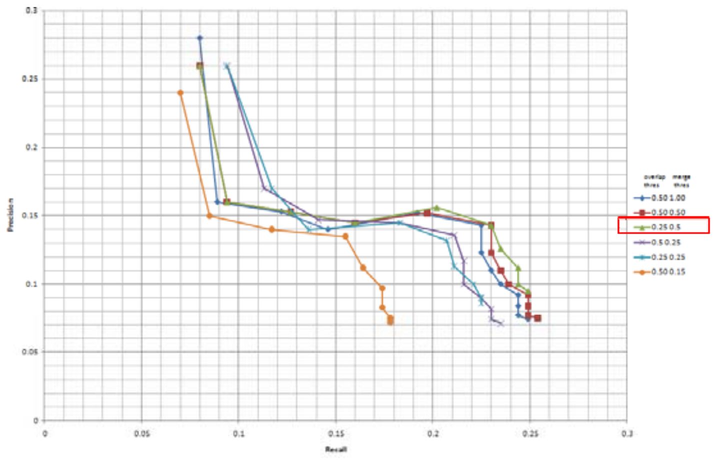
Evaluation

Co
Dict
BN

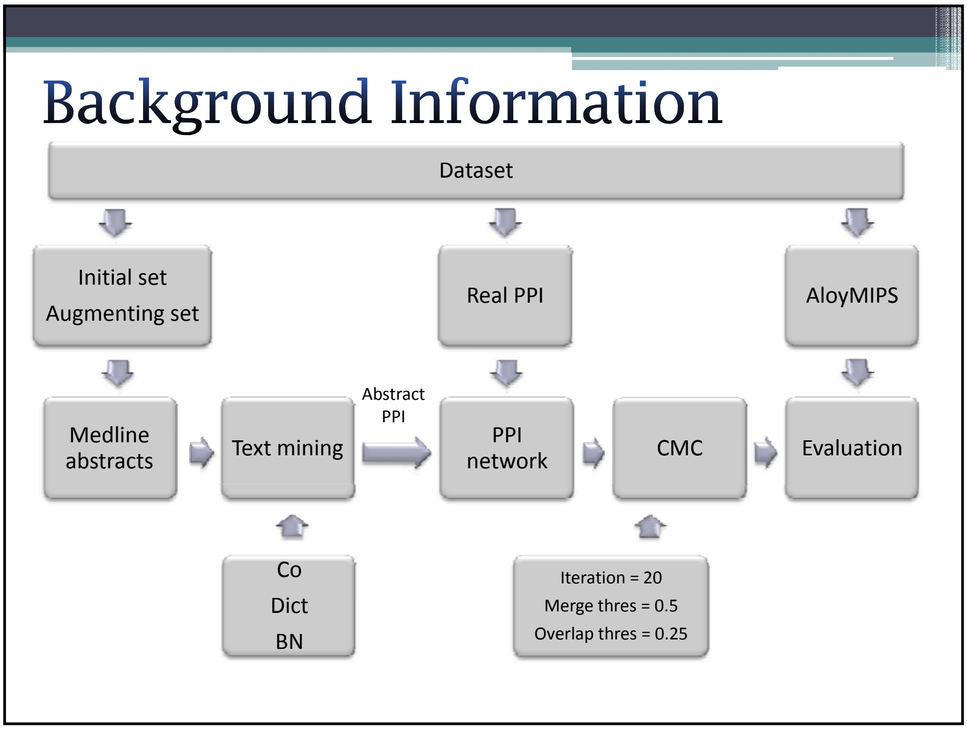
Iteration = 20
Merge thres = x
Overlap thres = y

Recall + Precision

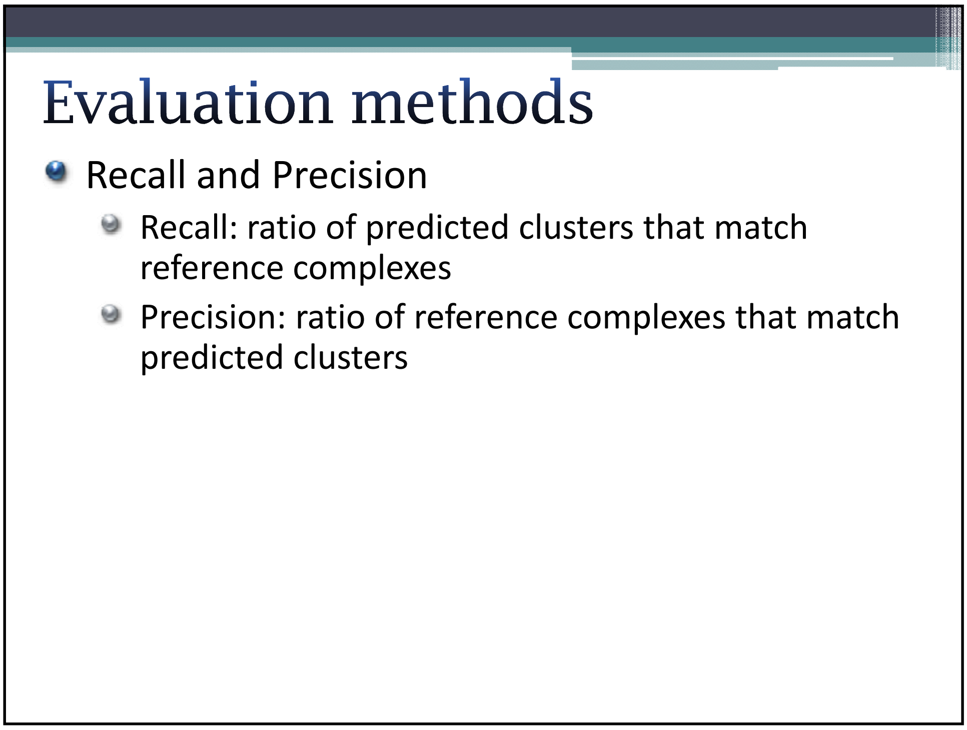
Merge and overlap threshold



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

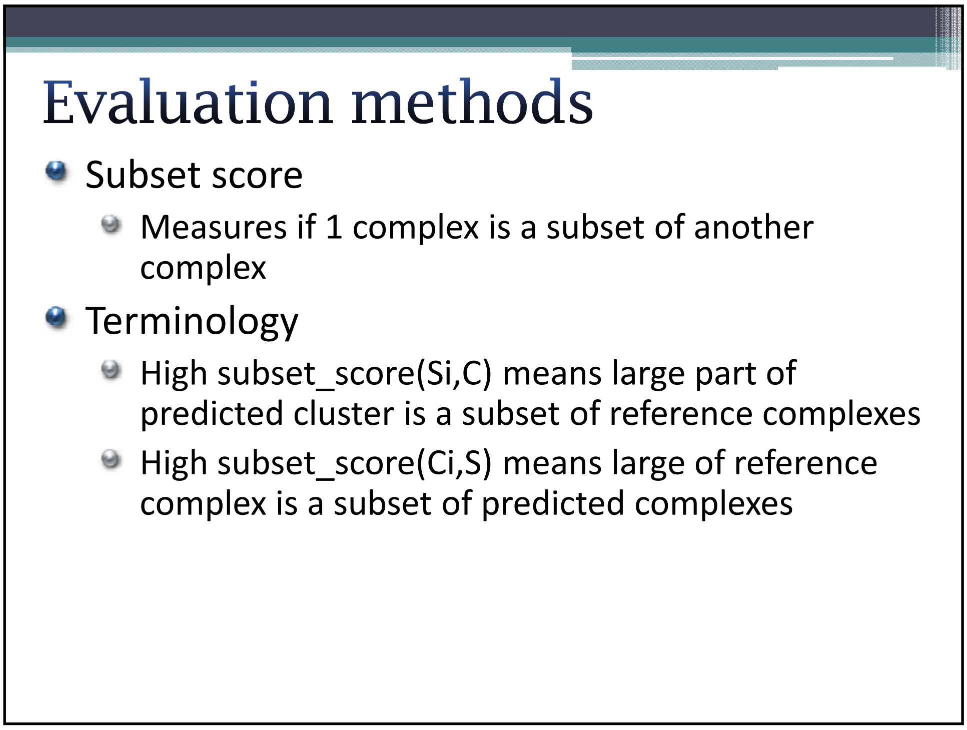


Background Information



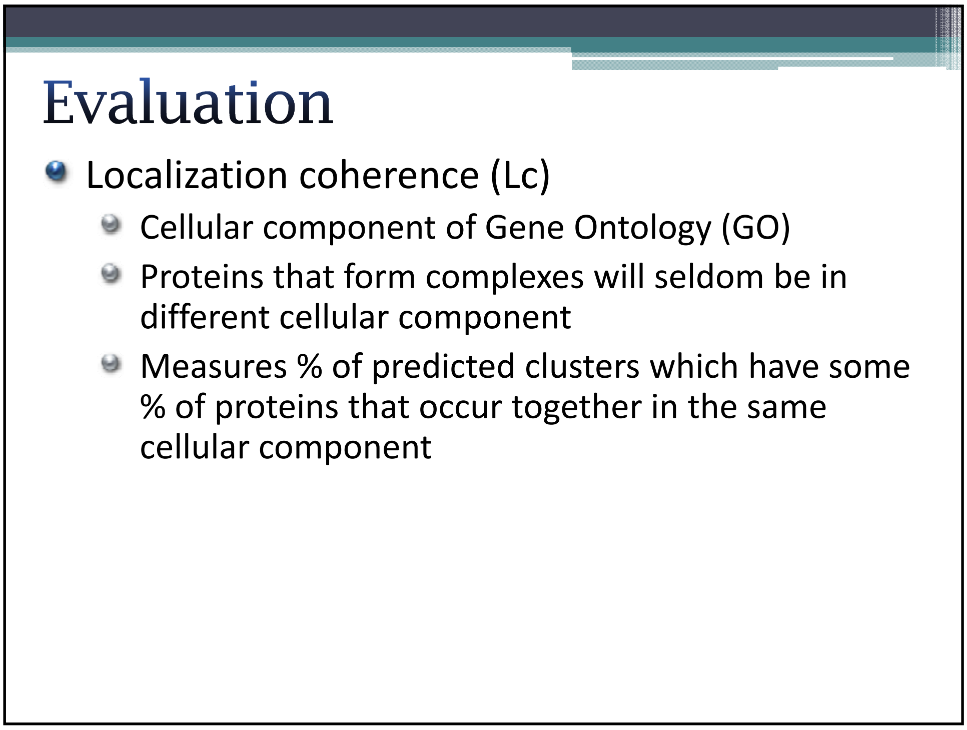
Evaluation methods

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



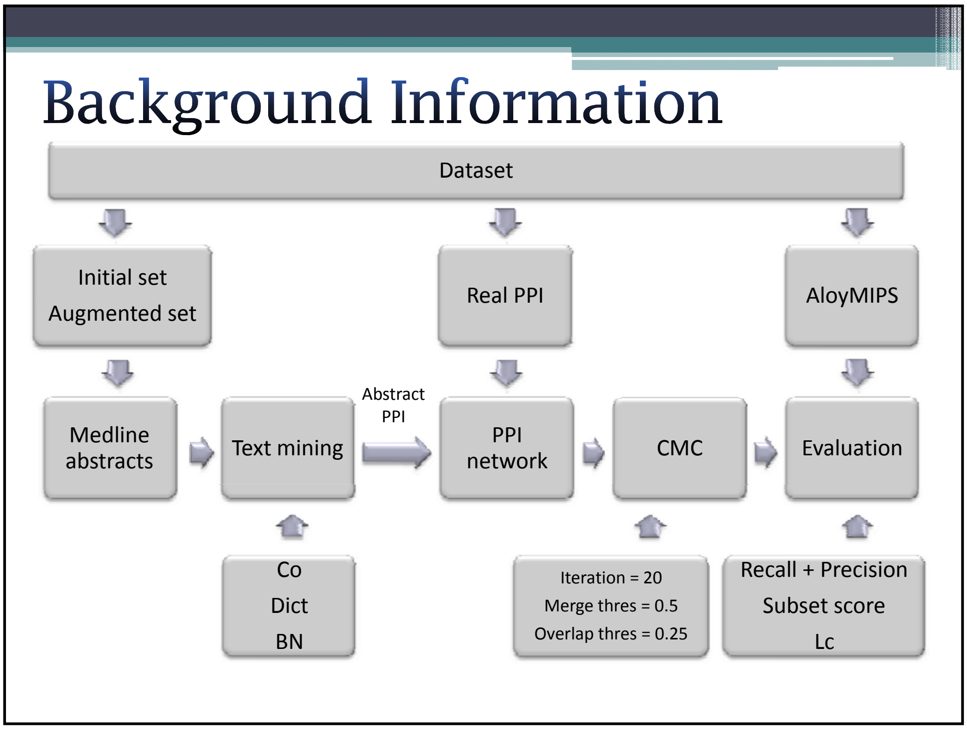
Evaluation methods

- 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

Real PPI

AloyMIPS

Medline abstracts

Text mining

Abstract PPI

PPI network

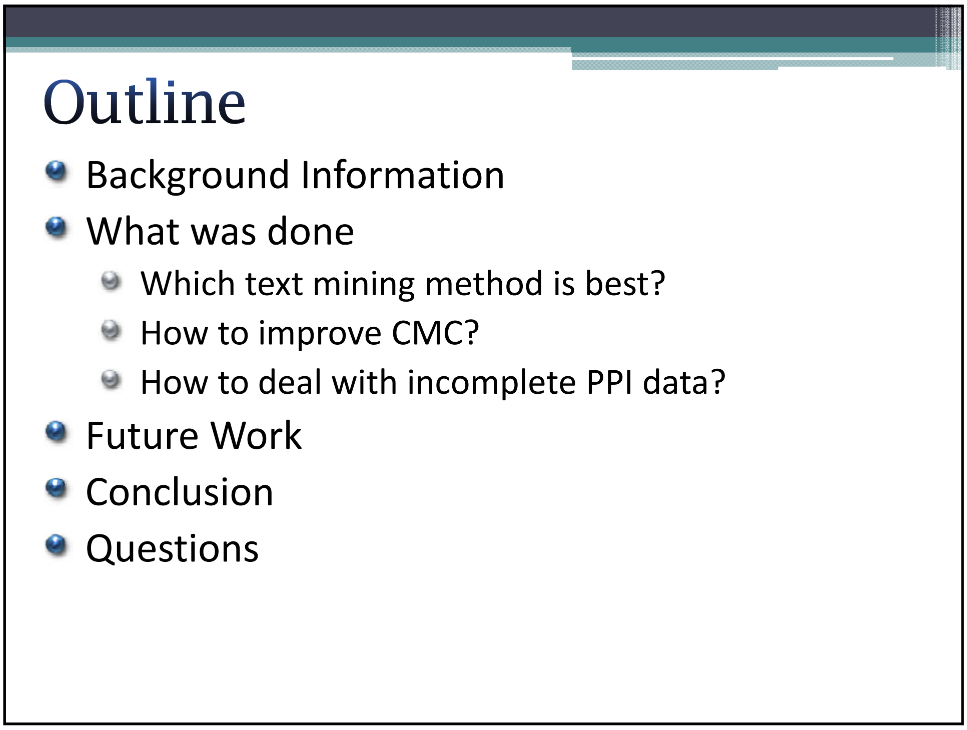
CMC

Evaluation

Co
Dict
BN

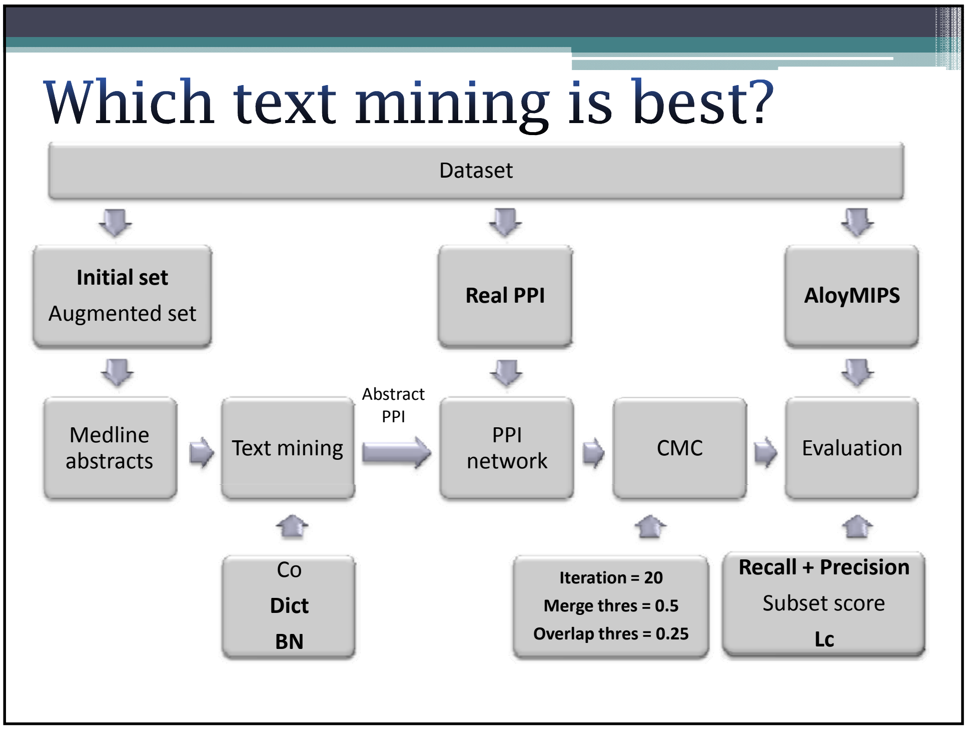
Iteration = 20
Merge thres = 0.5
Overlap thres = 0.25

Recall + Precision
Subset score
Lc

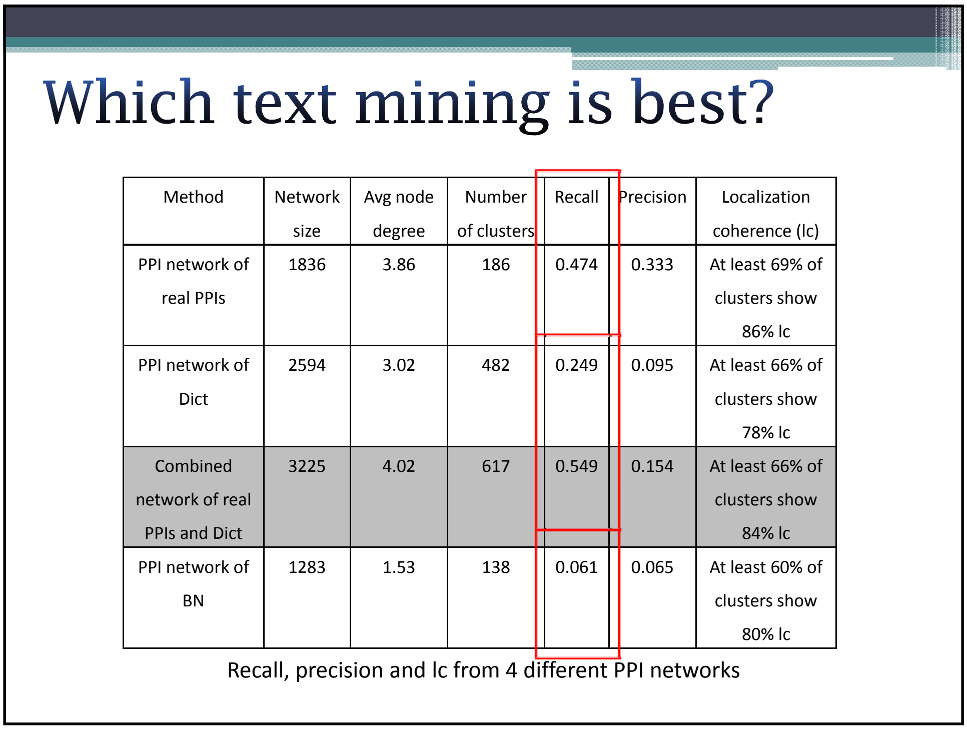


Outline

- 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?



Which text mining is best?

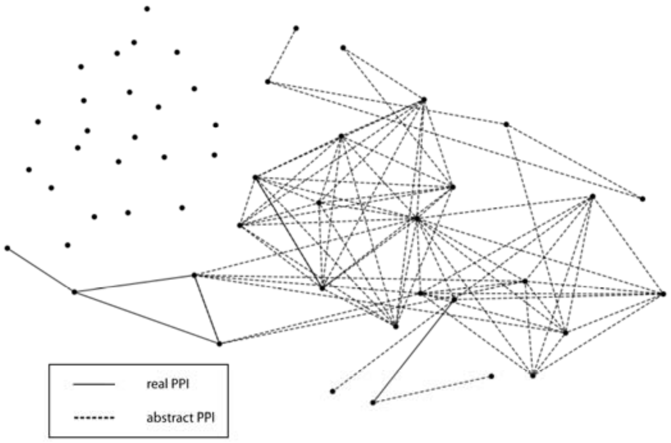
Method	Network size	Avg node degree	Number of clusters	Recall	Precision	Localization coherence (lc)
PPI network of real PPIs	1836	3.86	186	0.474	0.333	At least 69% of clusters show 86% lc
PPI network of Dict	2594	3.02	482	0.249	0.095	At least 66% of clusters show 78% lc
Combined network of real PPIs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

Recall, precision and lc from 4 different PPI networks

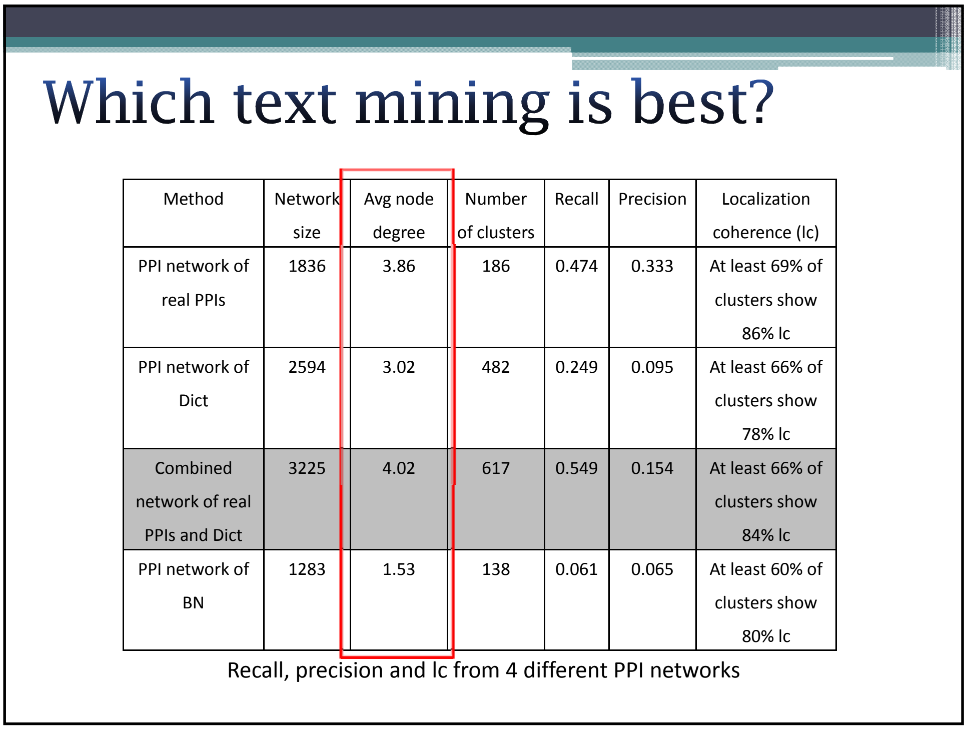
Which text mining is best?

- 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

Which text mining is best?



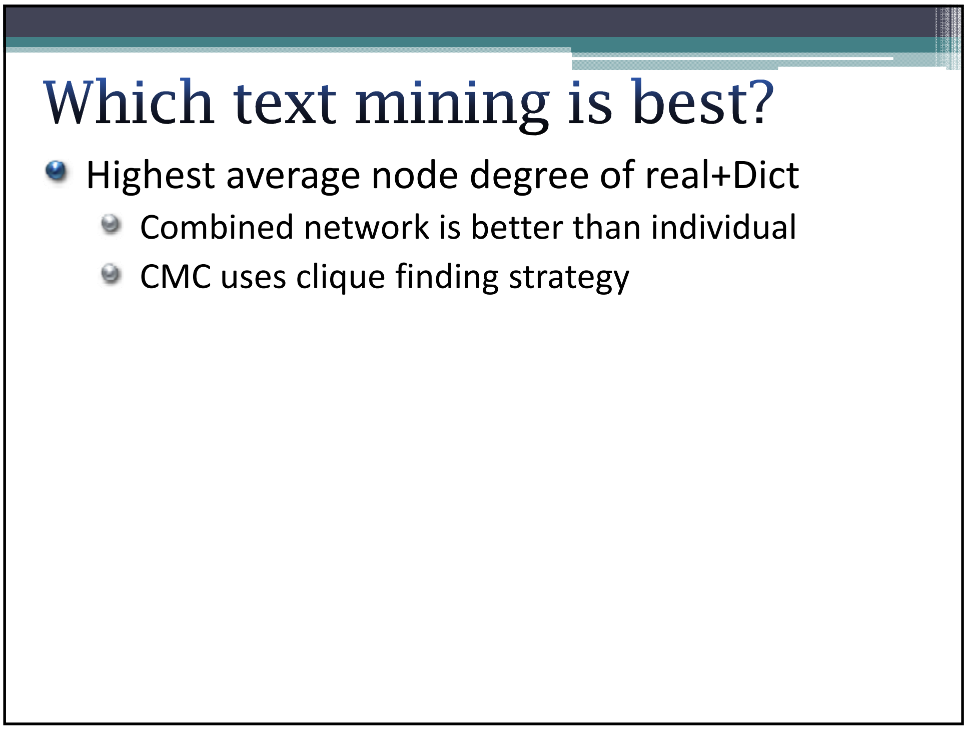
Graph of real complex 420



Which text mining is best?

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

Recall, precision and lc from 4 different PPI networks



Which text mining is best?

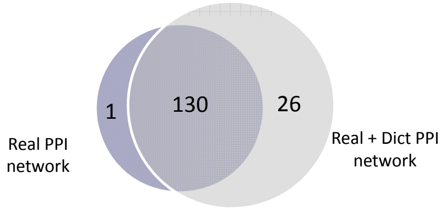
- Highest average node degree of real+Dict
  - Combined network is better than individual
  - CMC uses clique finding strategy

Which text mining is best?

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

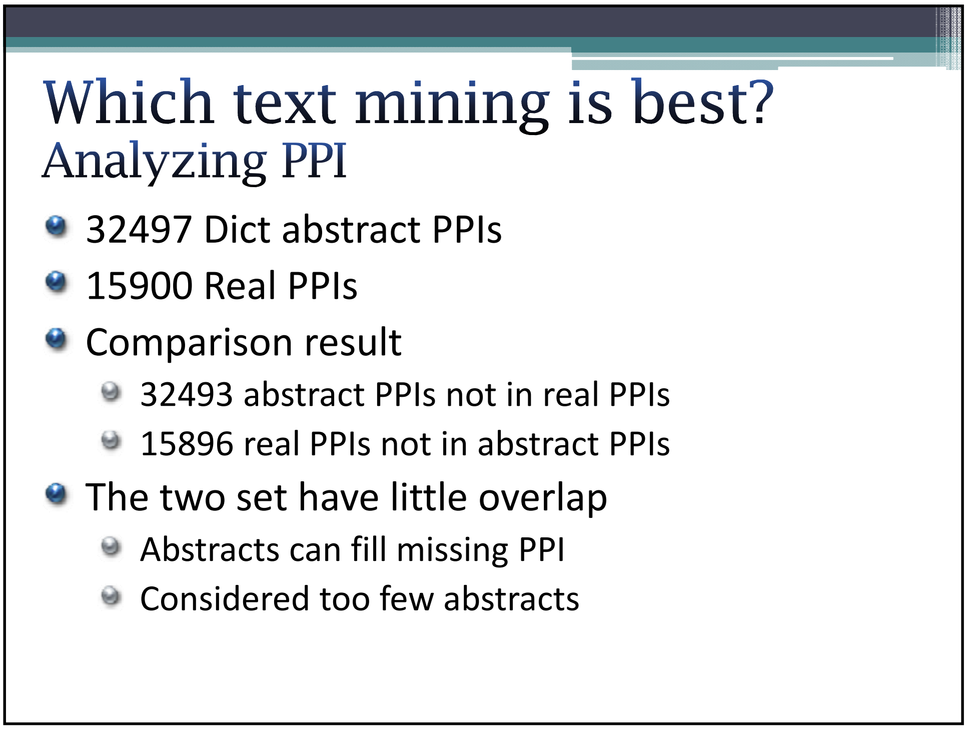
Recall, precision and lc from 4 different PPI networks

Which text mining is best?
Analyzing predicted clusters



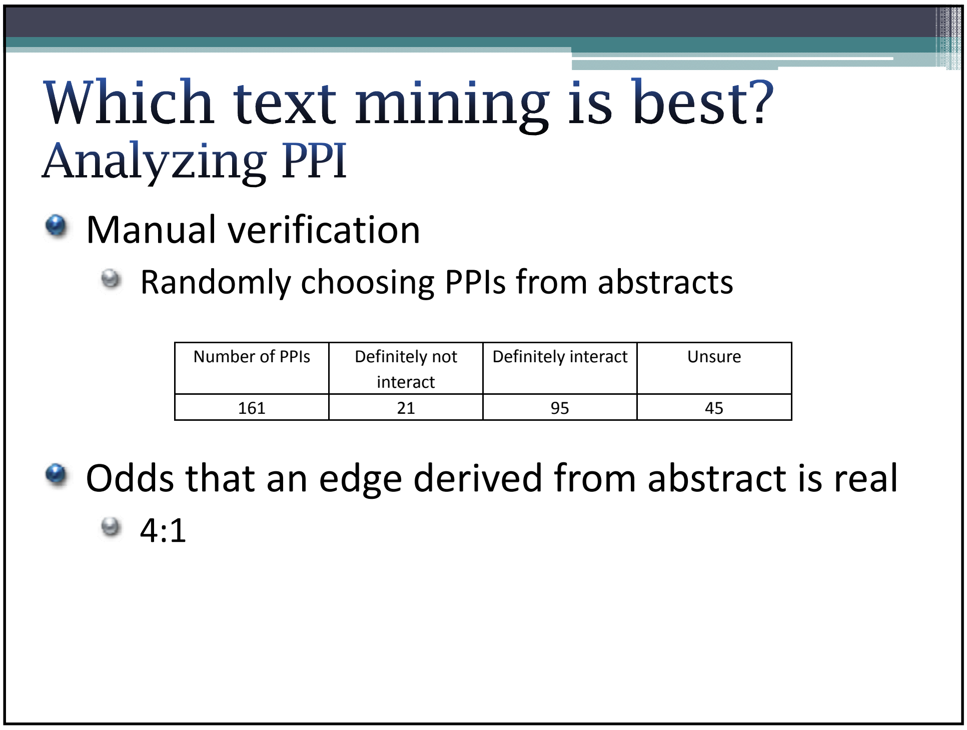
Venn diagram of correctly predicted clusters from 2 different networks

- 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



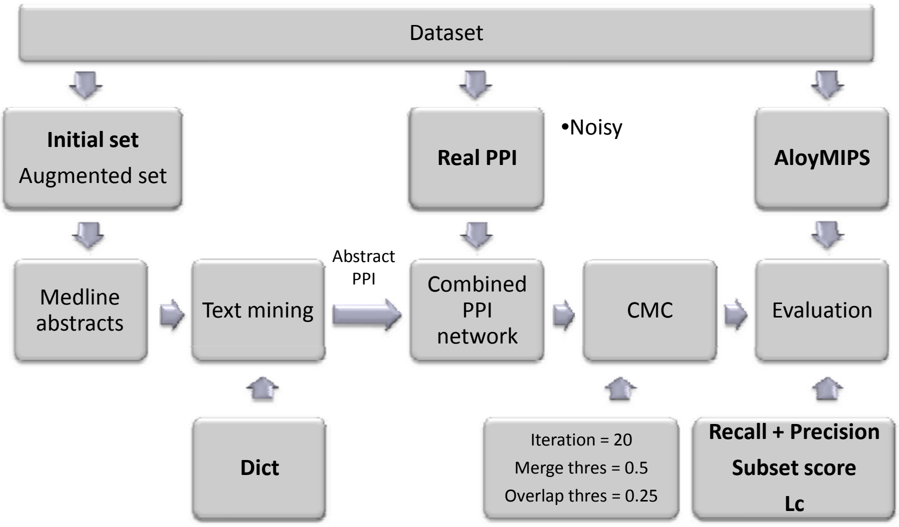
Which text mining is best?
Analyzing PPI

- Manual verification
  - Randomly choosing PPIs from abstracts

Number of PPIs	Definitely not interact	Definitely interact	Unsure
161	21	95	45

- Odds that an edge derived from abstract is real
  - 4:1

How to improve CMC?



Dataset

Initial set

Augmented set

Real PPI

•Noisy

AloyMIPS

Medline abstracts

Text mining

Abstract PPI

Combined PPI network

CMC

Evaluation

Dict

Iteration = 20

Merge thres = 0.5

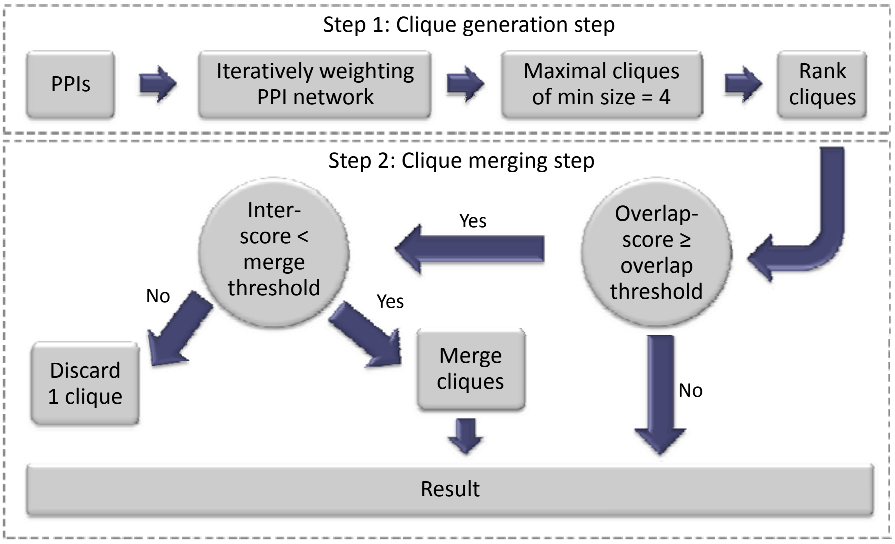
Overlap thres = 0.25

Recall + Precision

Subset score

Lc

CMC



Step 1: Clique generation step

PPIs

Iteratively weighting PPI network

Maximal cliques of min size = 4

Rank cliques

Step 2: Clique merging step

Overlap-score ≥ overlap threshold

Yes

No

Inter-score < merge threshold

No

Yes

Discard 1 clique

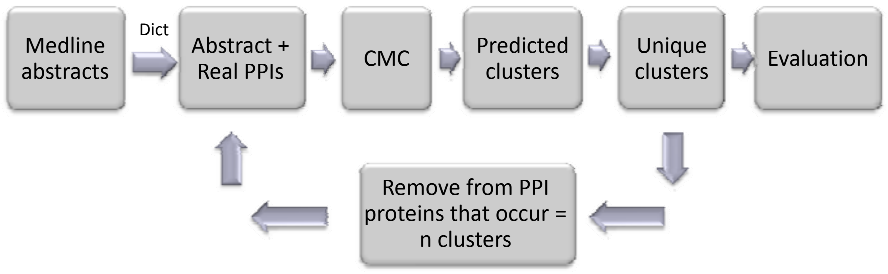
Merge cliques

Result

How to improve CMC?

- Iterative removal of non-hub proteins

Iteration n = 0 to 5

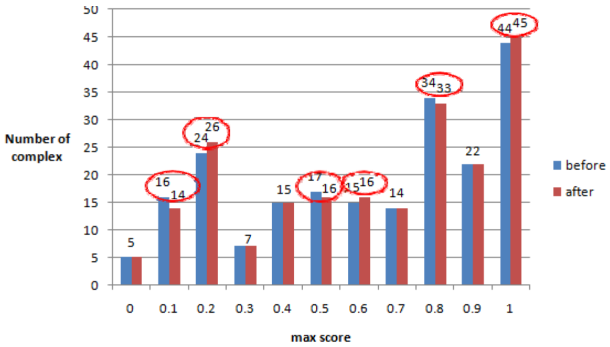


How to improve CMC?

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

Recall, precision and lc after different iteration of non-hub removal

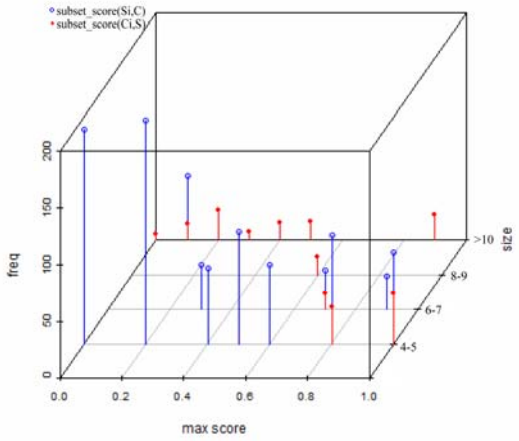
How to improve CMC?



subset score of AloyMIPS before and after iterated removal of non-hub proteins

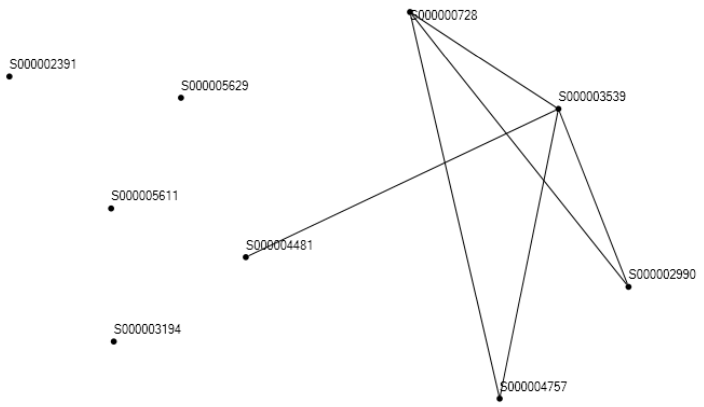
- 4 complexes improved their score while none decreased

How to improve CMC?



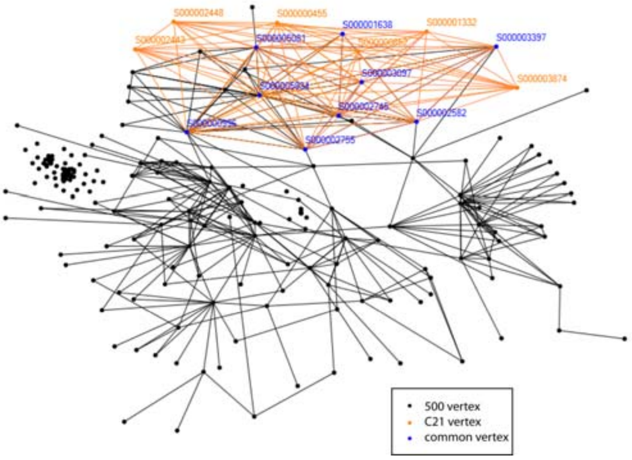
3D graph of subset evaluation after iterated removal of non-hub proteins

How to improve CMC?



Graph of real complex 520.20

How to improve CMC?

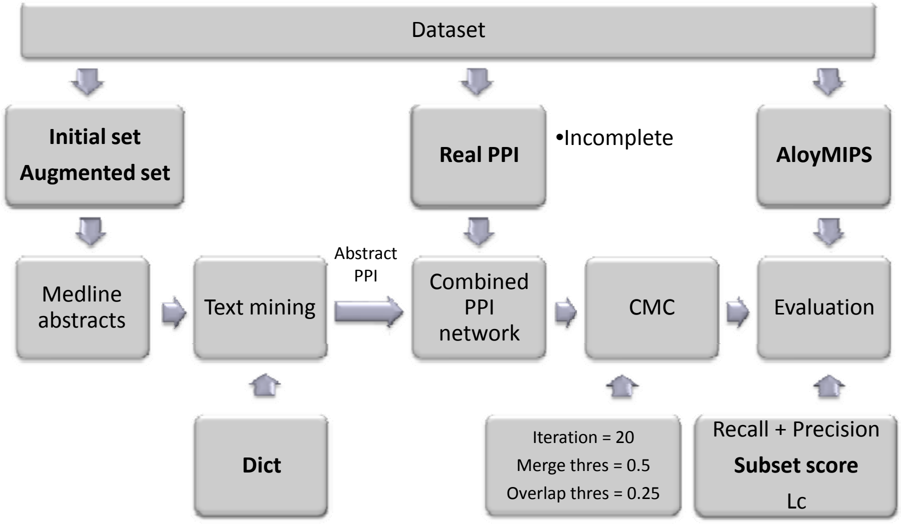


Graph of real complex 500 and predicted cluster C21

How to improve CMC?

- Cliques as a basis is stringent
- PPI data is incomplete

How to deal with incomplete PPI?



Dataset

Initial set

Augmented set

Real PPI

•Incomplete

AloyMIPS

Medline abstracts

Text mining

Abstract PPI

Combined PPI network

CMC

Evaluation

Dict

Iteration = 20

Merge thres = 0.5

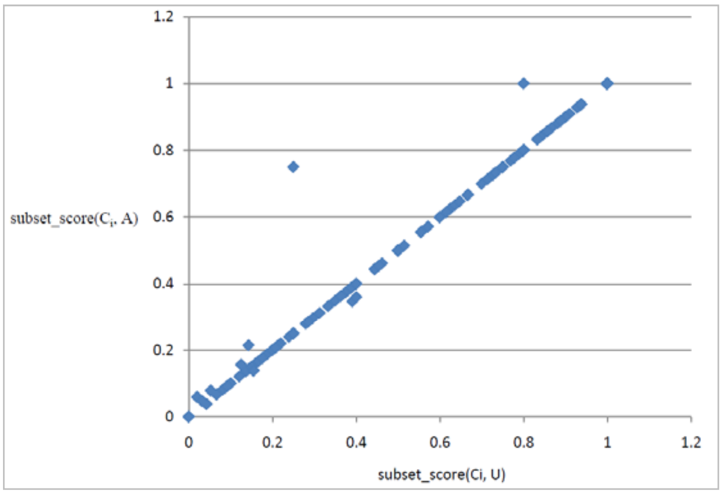
Overlap thres = 0.25

Recall + Precision

Subset score

Lc

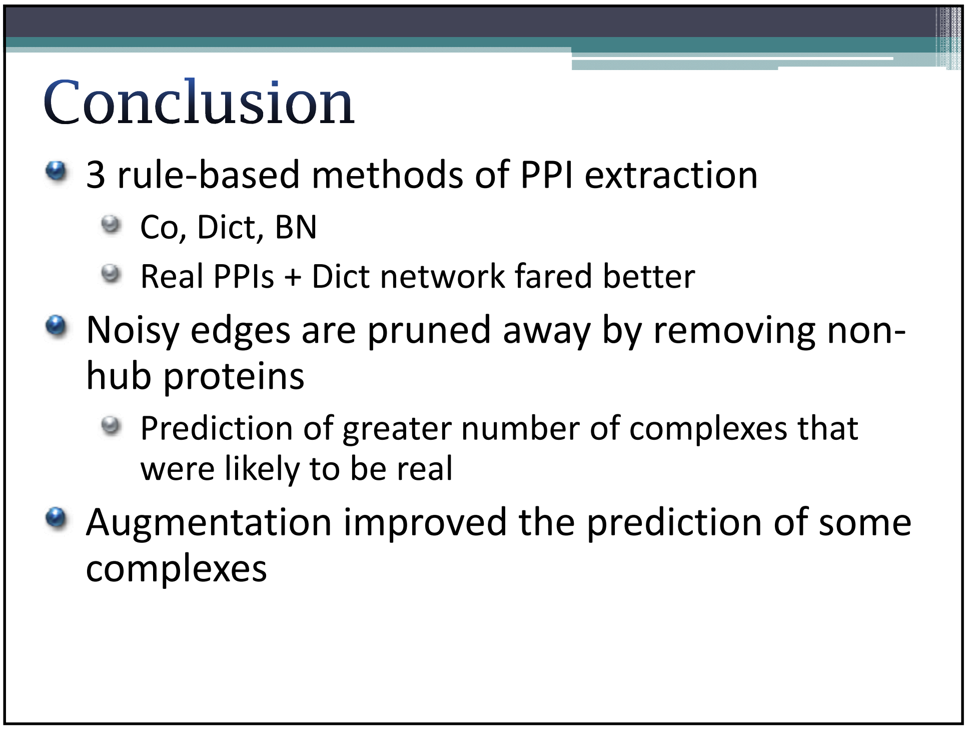
How to deal with incomplete PPI?



subset_score of AloyMIPS after augmentation vs before augmentation

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