Protein Complexes Prediction by sampling

Osamu Maruyama

Kyushu University

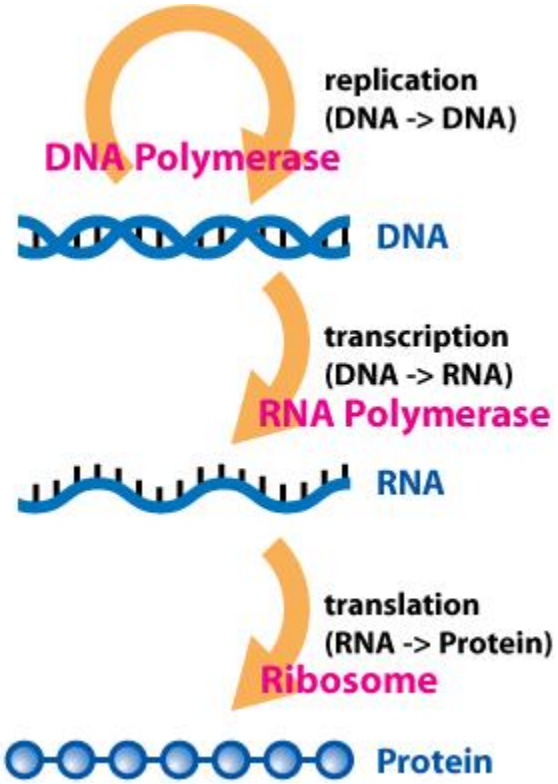
Contents

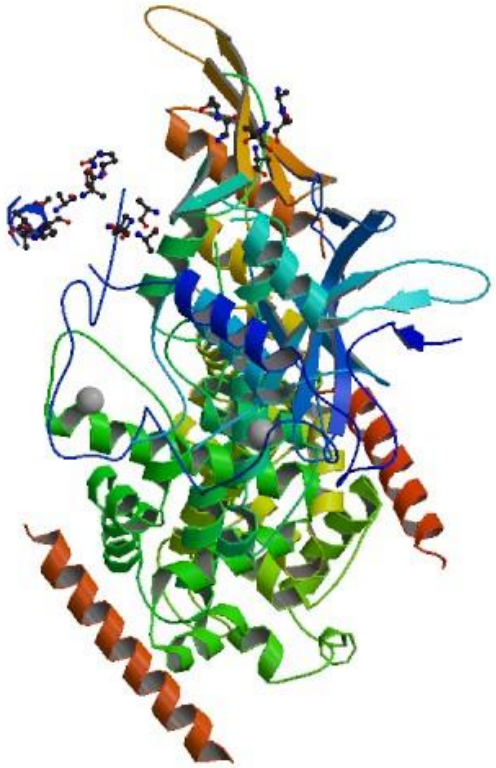
1. What's the protein complex prediction problem?
2. First challenging difficulty – Small protein complexes are the majority
3. PPSampler2 (Proteins' Partition)
4. Second challenging difficulty – Known complexes are overlapped with each other.
5. RocSampler (Regularizing overlaps of complexes)
6. Concluding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

Contents

DNA, protein, and protein complex

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

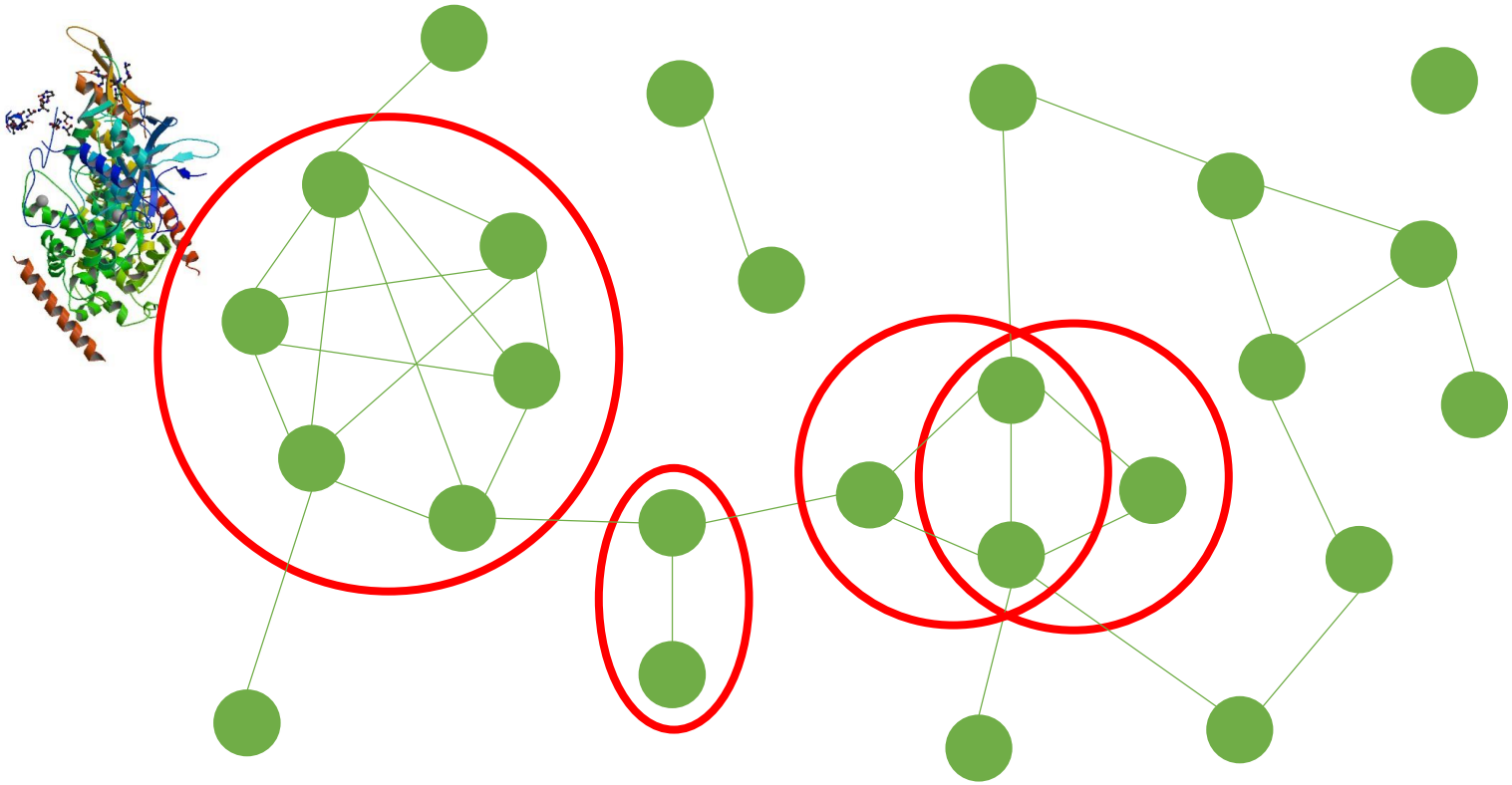




A protein complex is a
set of proteins connected
by protein-protein interactions

Problem: Protein complex prediction

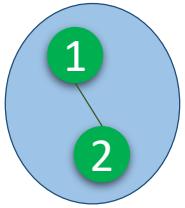
Input: a protein-protein interaction (PPI) network (edge-weighted undirected graph).



Contents

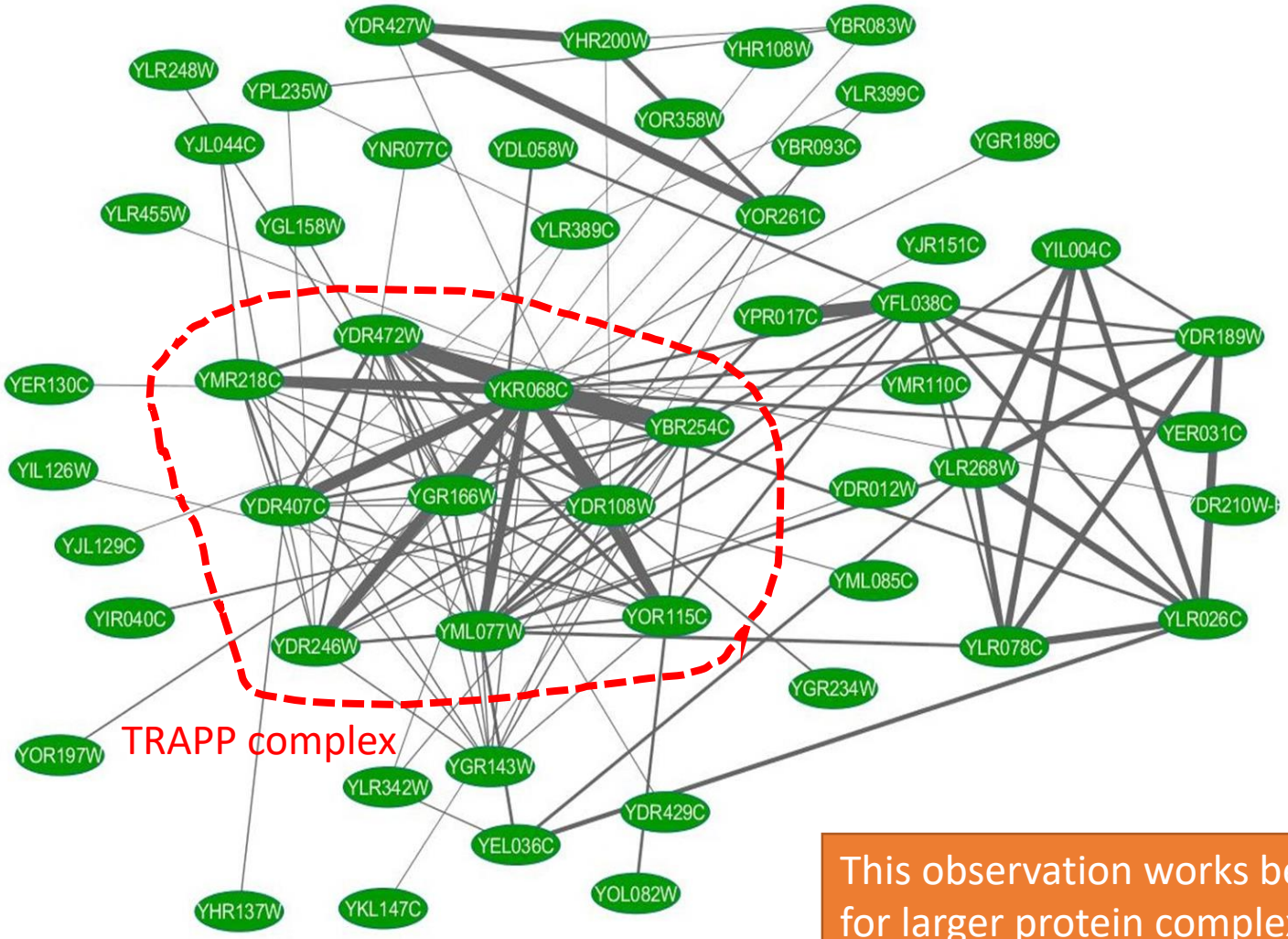
First challenging difficulty

- 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

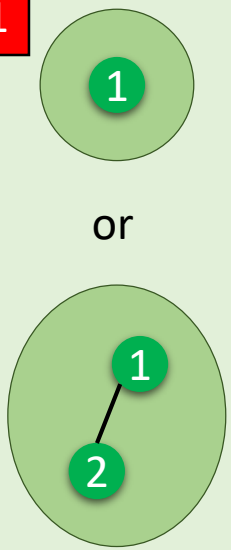


This observation works better for larger protein complexes.

Typical approach: cluster-expansion method

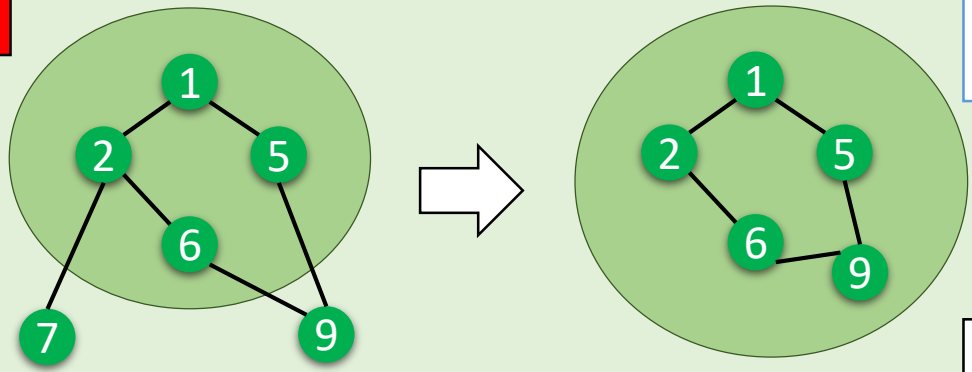
affinity:
- cohesiveness
- random walk
- etc

1



Every protein or PPI forms an initial cluster

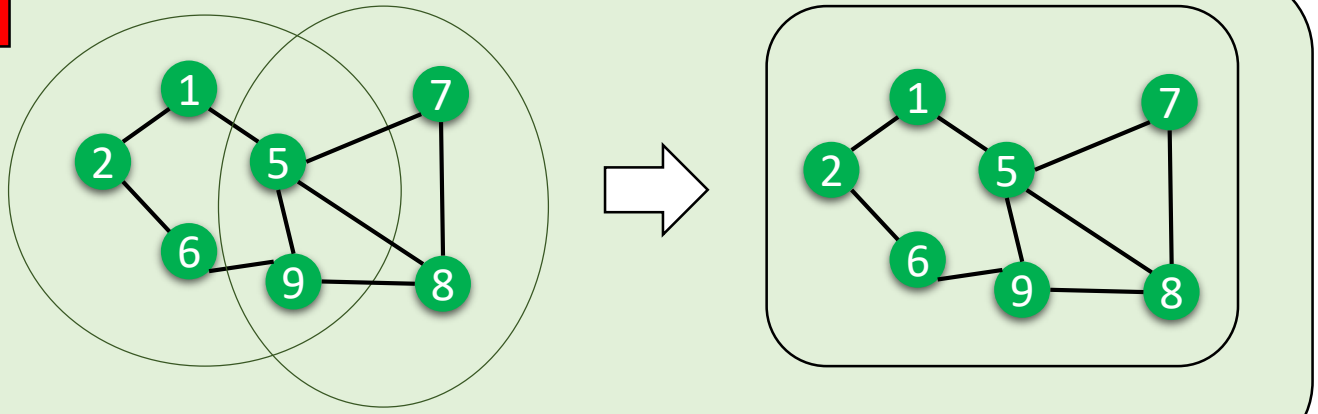
2



cluster-expansion process

stopped by a simple criterion

3



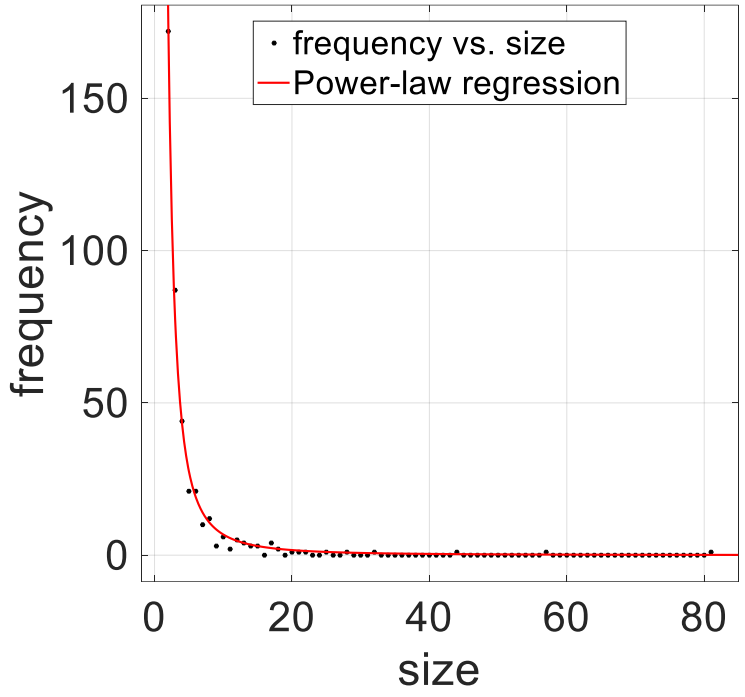
post-processing for overlapping clusters

This has no mechanism to control the sizes of predicted clusters

Contents

How to predict small complexes

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

$$\psi_\gamma(s) = \frac{s^{-\gamma}}{\sum_{t=2}^{S_{\max}} t^{-\gamma}}$$

: two-side truncated power-law distribution

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

: fraction of predicted clusters of size s

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

where

$$h_{clu-size,s}(X, \gamma) = \left(\psi_\gamma(s) - \psi_X(s)\right)^2$$

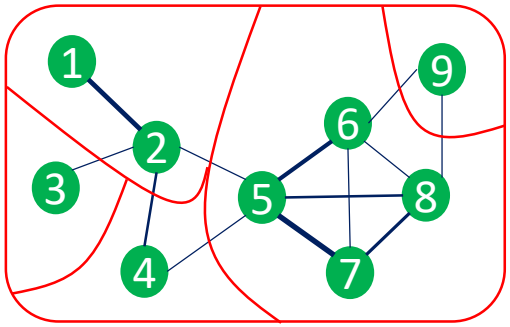
Widita and Maruyama *BMC Systems Biology* 2013, **7**(Suppl 6):S14
http://www.biomedcentral.com/1752-0509/7/S6/S14

BMC Systems Biology

RESEARCH Open Access

PPSampler2: Predicting protein complexes more accurately and efficiently by sampling

Chasanah Kusumastuti Widita[1], Osamu Maruyama[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 $f(X) =$

Optimization term

$h_{clu-den}(X)$

Sum of the densities of predicted clusters

Regularization terms

$+b(X)$

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.

$+c_{pro-num} \cdot h_{pro-num}(X)$

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 λ is a given target.

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

force of fitting the distribution of the sizes of predicted clusters to the power-law distribution of the scaling-exponent γ

$h_{clu-den}(X)$

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

where

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

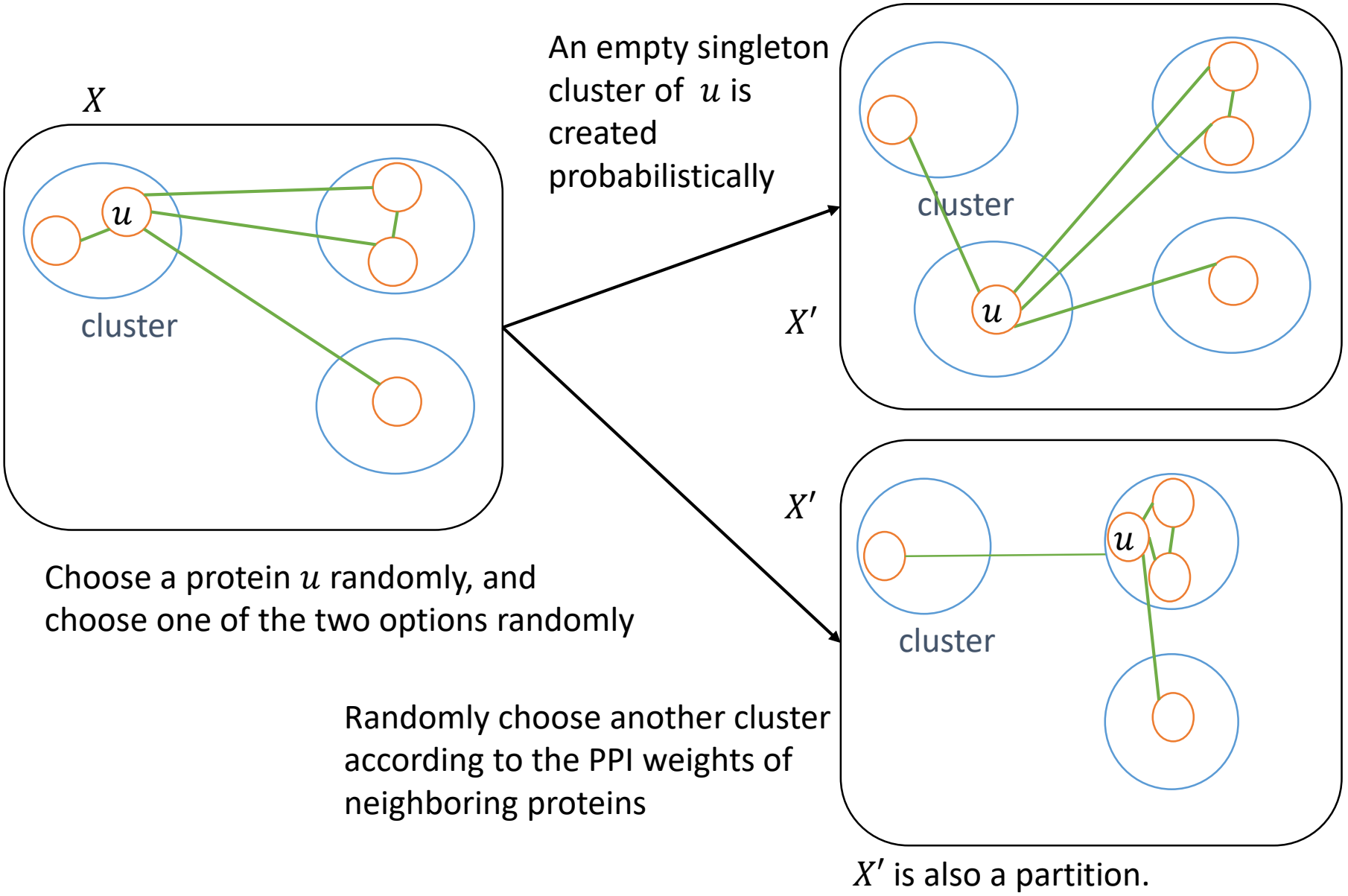
$$h_{clu-den}(X) = -\sum_{x \in X} density(x)$$

- The standard density is divided by $|x| \cdot (|x| - 1)/2.$
- $\sqrt{|x|}$ is used to alleviate excessively severer evaluation of a larger cluster.

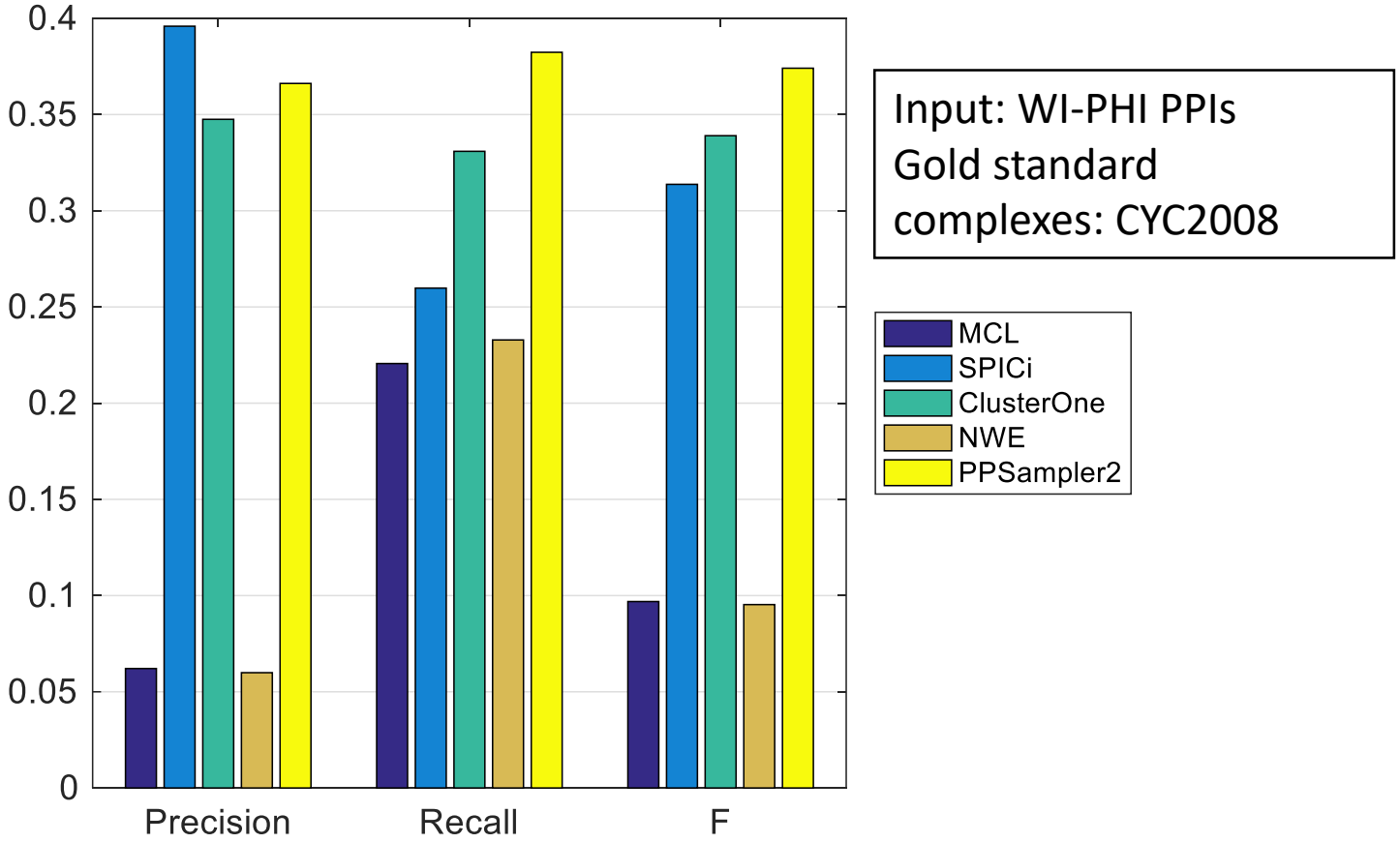
$b(X)$

- Let x be a subset of proteins, called a predicted cluster.
- Constraint 1: $|x| \leq S_{\max}$ (Max size of predicted clusters)
- Constraint 2: the vertex-induced subgraph of G by x is connected
- $b(x) = \begin{cases} 0 & \text{if both constraints are satisfied} \\ \infty & \text{otherwise} \end{cases}$
- $b(X) = \sum_{x \in X} b(x)$

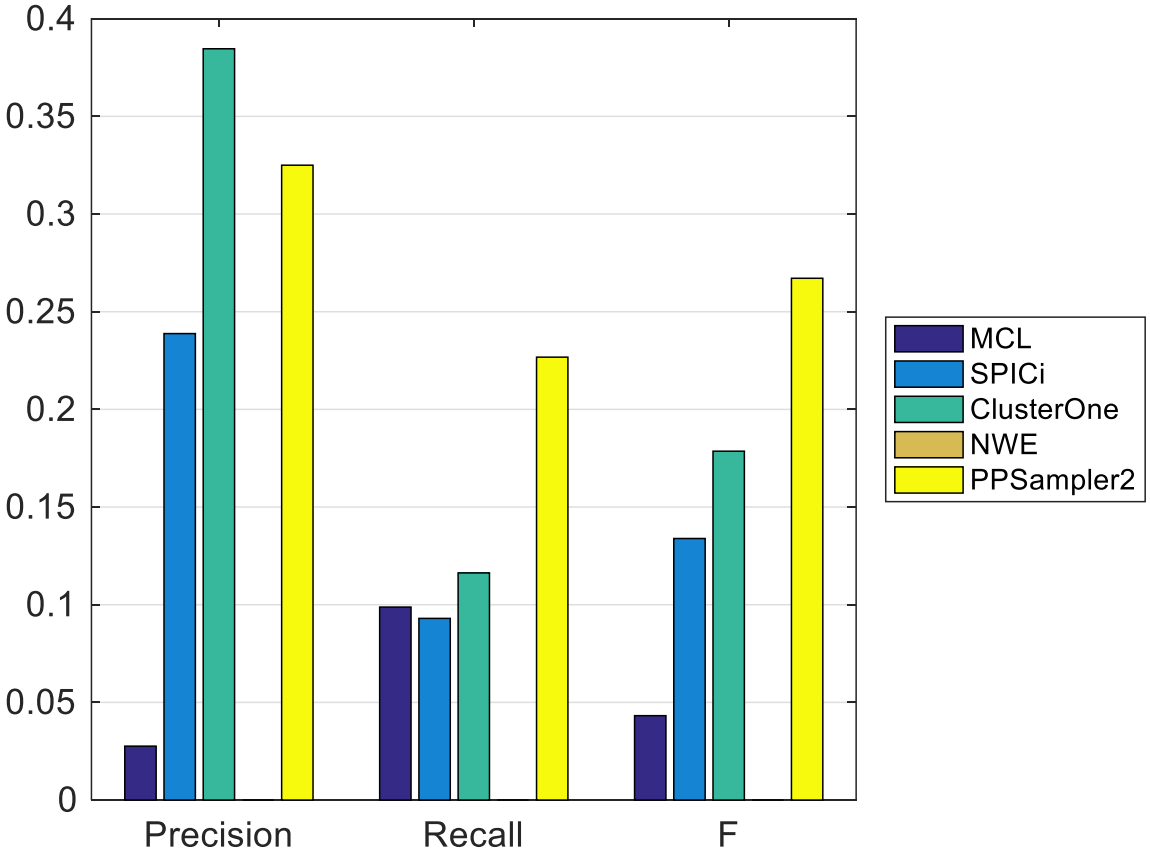
Proposal distribution $Q(X'|X)$



Performance comparison



Performance comparison on size-2 complexes by exact matching



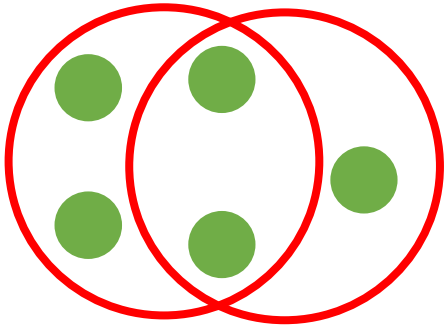
Contents

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.

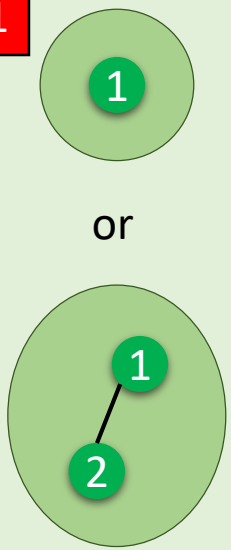


Overlap size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Frequency	151	22	9	13	4	1	10	1		1	1	1				1	1

Typical approach: cluster-expansion method

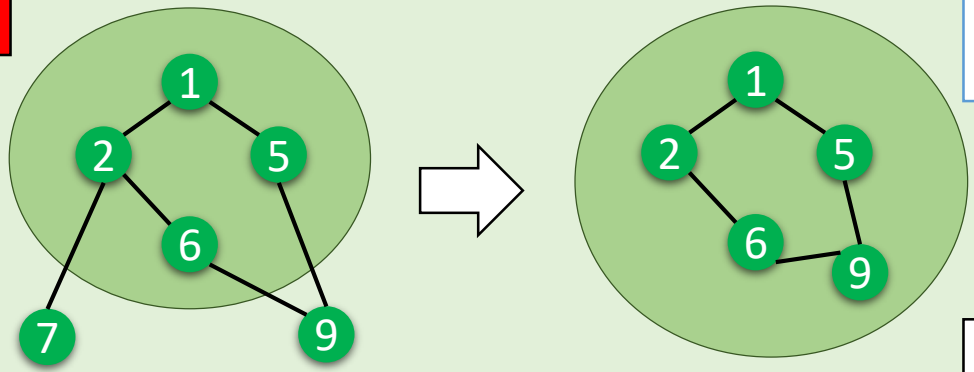
affinity:
- cohesiveness
- random walk
- etc

1



Every protein or PPI forms an initial cluster

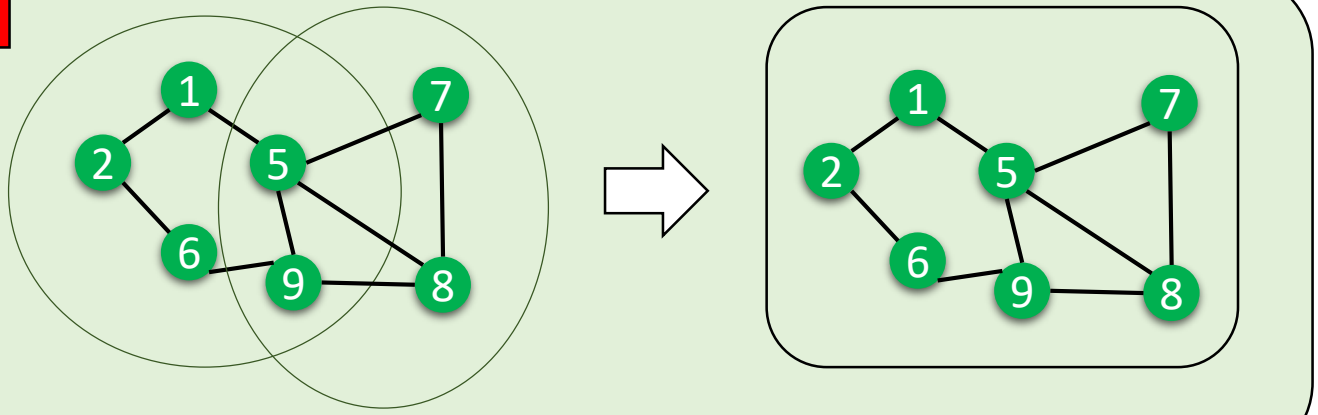
2



cluster-expansion process

stopped by a simple criterion

3

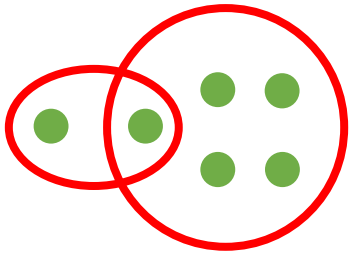


post-processing for overlapping clusters

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

Contents

$h_{clu-dis}(X)$: Term regularizing overlaps between predicted clusters



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

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

where

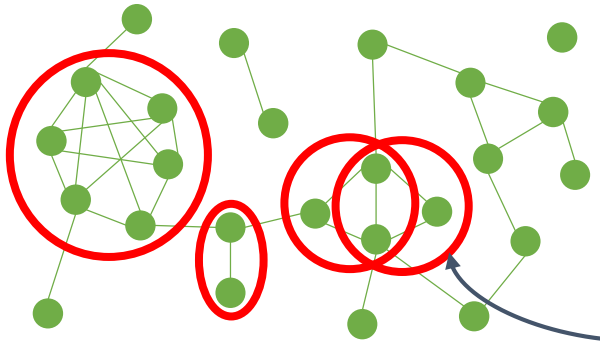
$J(x, x') = \frac{|x \cap x'|}{|x \cup x'|}$ (Jaccard index)

Parameter $\beta \in \{0.2, 0.3, 0.4\}$ is optimized for a PPI network

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

Our new sampler: RocSampler (Regularizing Overlaps of Clusters) [2016]

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



+ γ

A new term regularizing overlaps of predicted clusters is designed

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

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

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

Optimization term

$h_{clu-den}(X)$	Sum of the densities of predicted clusters

Regularization terms

$+b(X)$	Two Boolean basic constraints on predicted clusters
$+c_{pro-num} \cdot h_{pro-num}(X)$	Minimize the number of proteins of the predicted clusters; $h_{pro-num}(X) = \lvert \cup_{x \in X} x \rvert^2$
$+c_{clu-size} \cdot \sum_{s=2}^{S_{max}} h_{clu-size,s}(X, \gamma)$	force of fitting the distribution of the sizes of predicted clusters to the power-law distribution of the scaling-exponent γ
$+c_{hy} \cdot h_{hy}(\gamma)$	Prior of γ; $h_{hy}(\gamma) = (\gamma - \gamma_0)^2$ with $\gamma_0 = 2.5$
$+c_{clu-dis} \cdot h_{clu-dis}(X)$	Penalize overlaps between predicted clusters

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

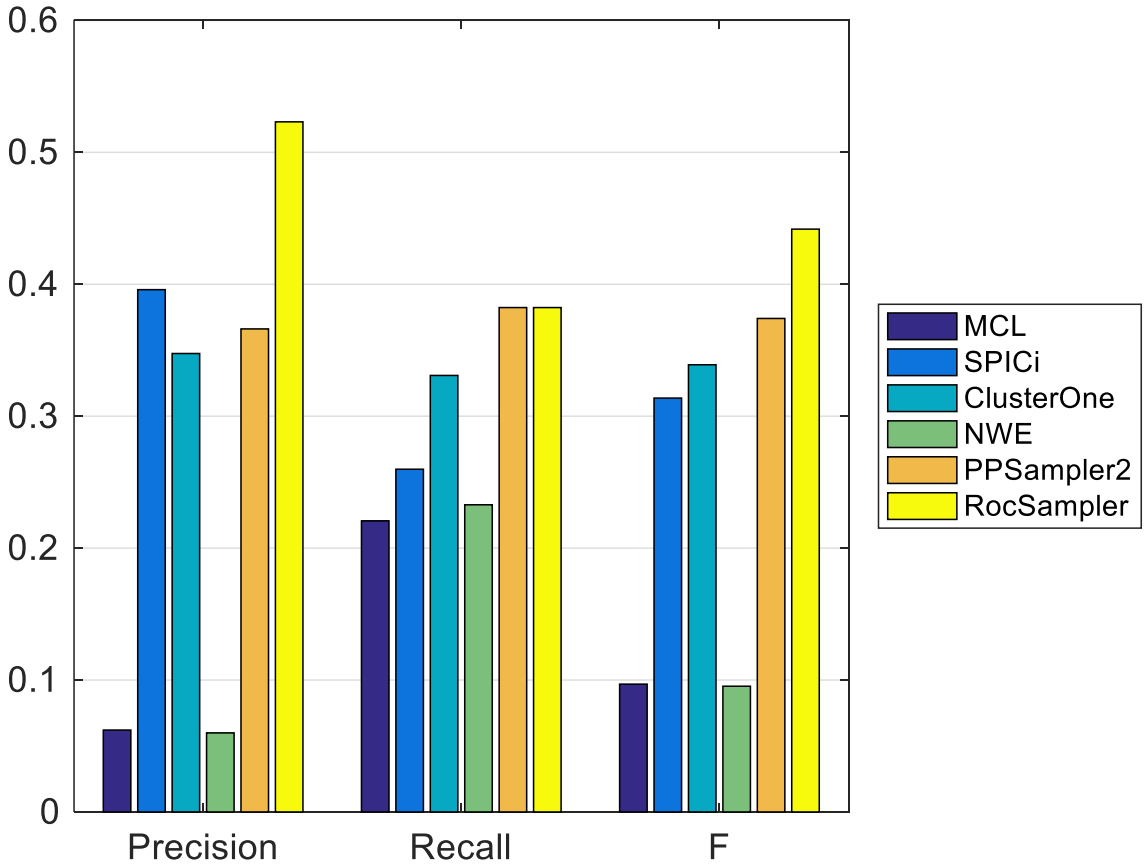
1. Randomly choose one of the 4 options
    1. Randomly add a new clusters of size 2 to X
    2. Randomly remove a cluster of size 2 in X
    3. Randomly add a new protein to a cluster in X
    4. Randomly remove a protein of a cluster in X
2. $\gamma = \min\{10^{-10}, \gamma + \varepsilon\}$ where $\varepsilon \sim N(0,0.001)$

Computational experiment

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

Parameter values of protein complex prediction methods are optimized.

Performance comparison on WI-PHI



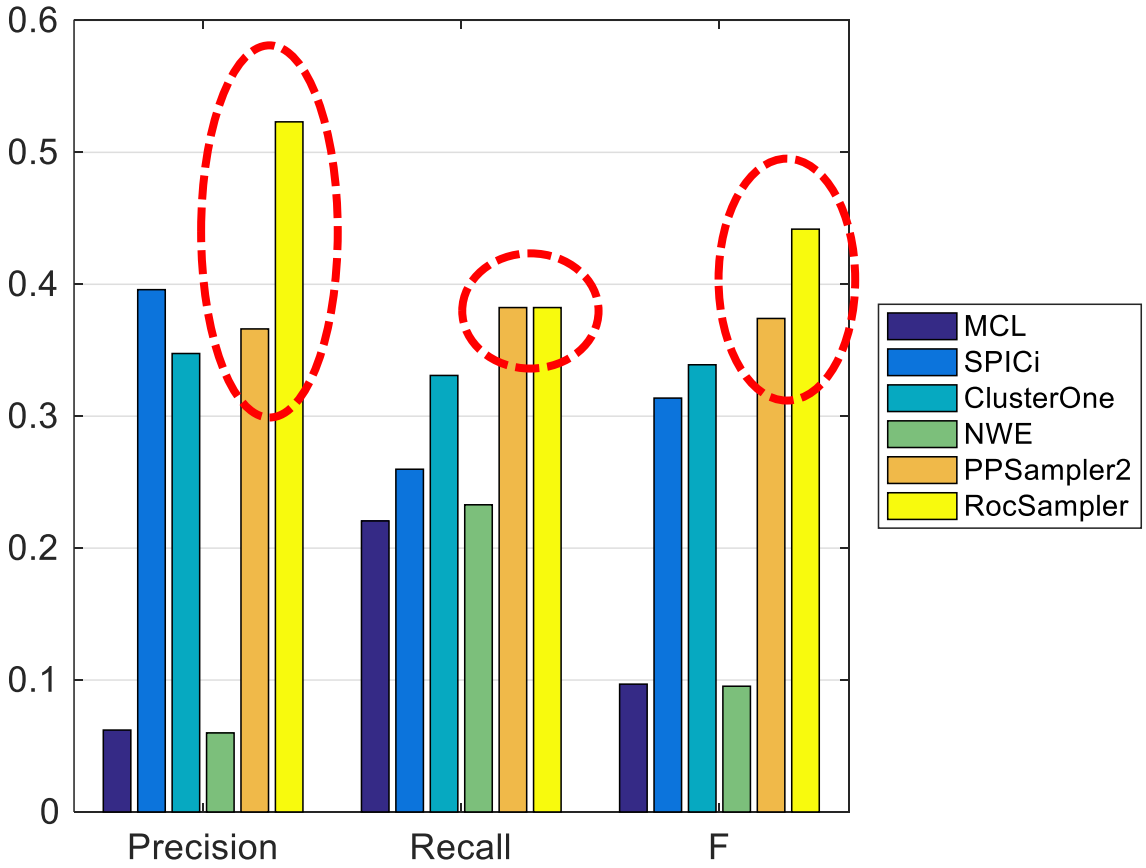
Estimated value of γ

- The scaling exponent of the power-law regression curve of CYC2008 is 2.02.
- The estimated value of γ is 1.91.
- Note that PPSampler2 uses γ to be 2.

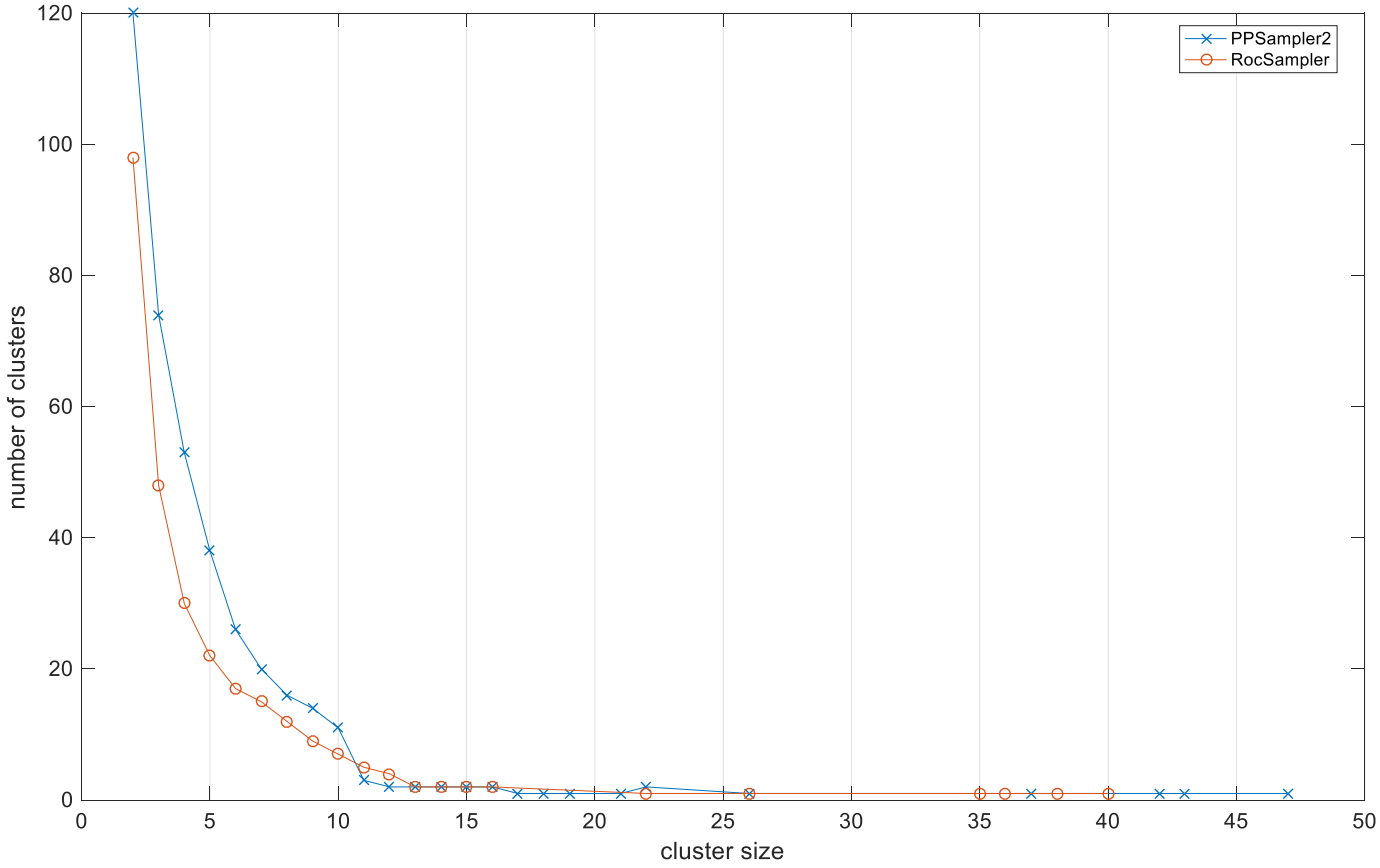
Difference in performance between PPSampler2 and RocSampler
Precision:

PPSampler2: 145/396 = 0.37

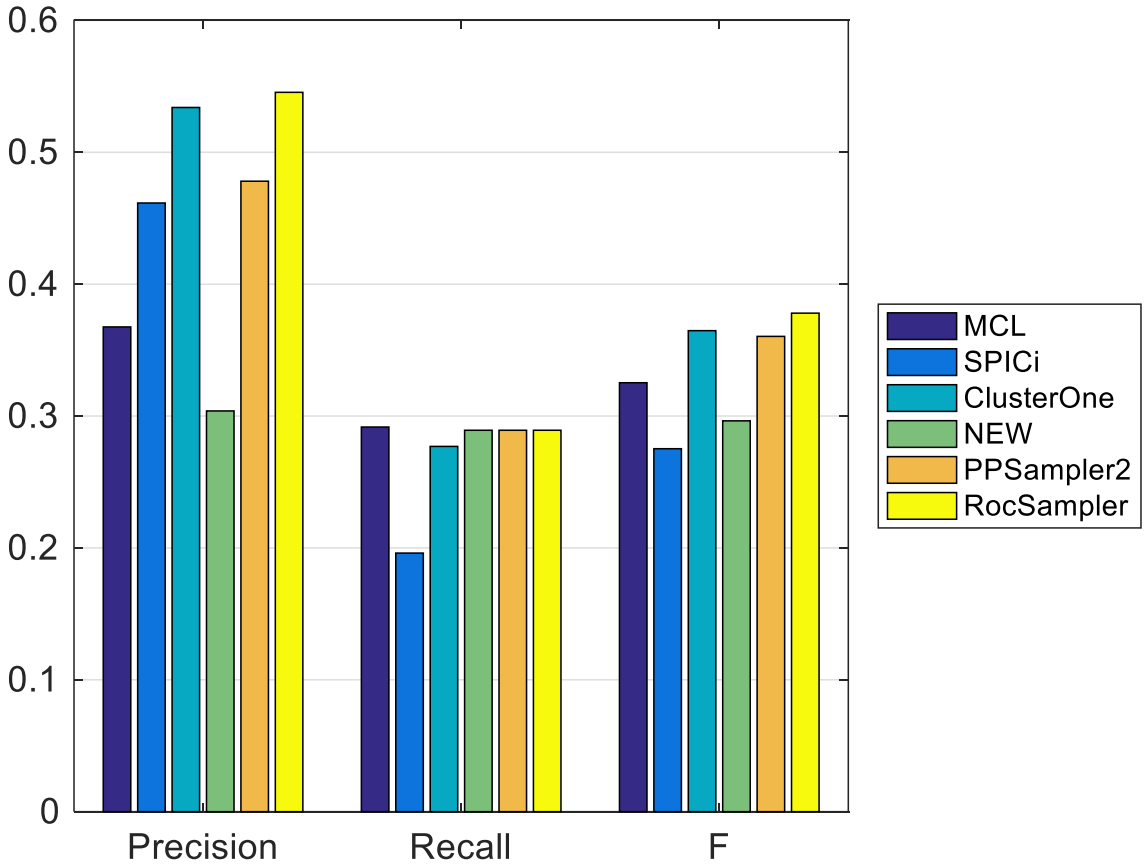
RocSampler: 147/281 = 0.52



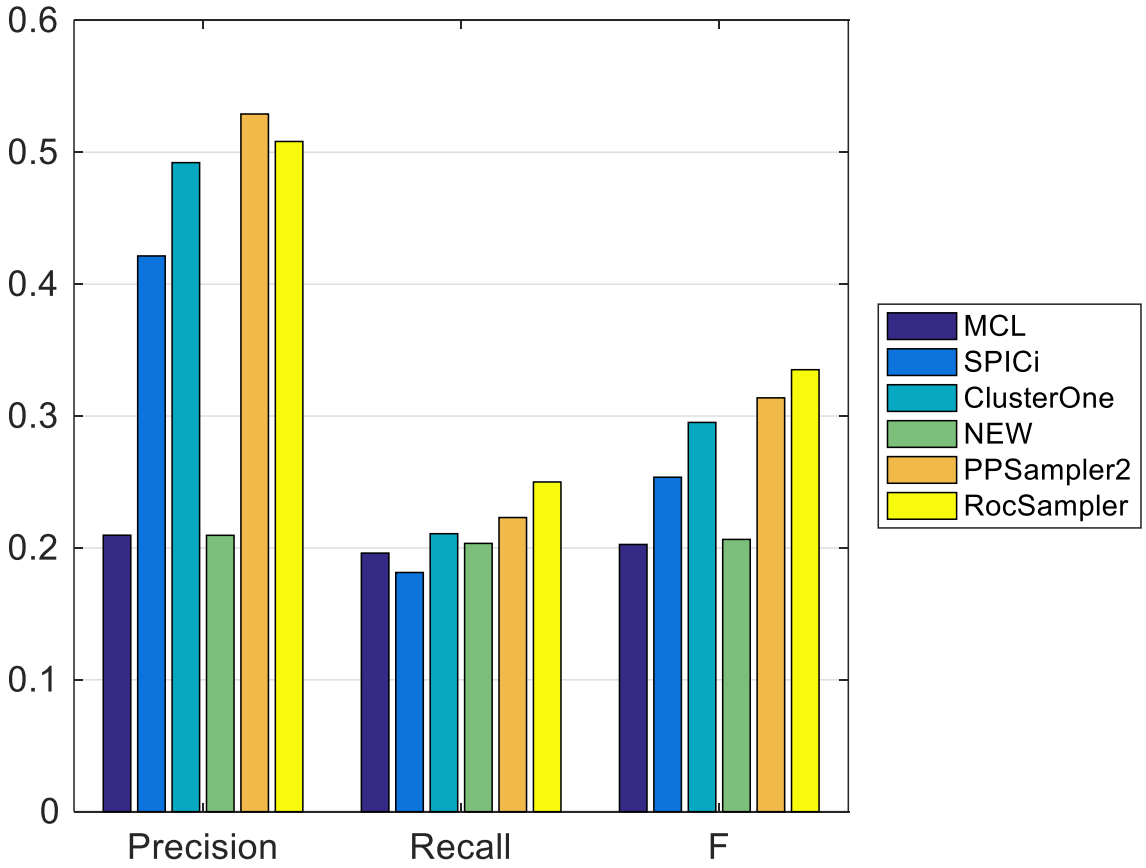
PPSampler2 predicted more (insignificant) small clusters



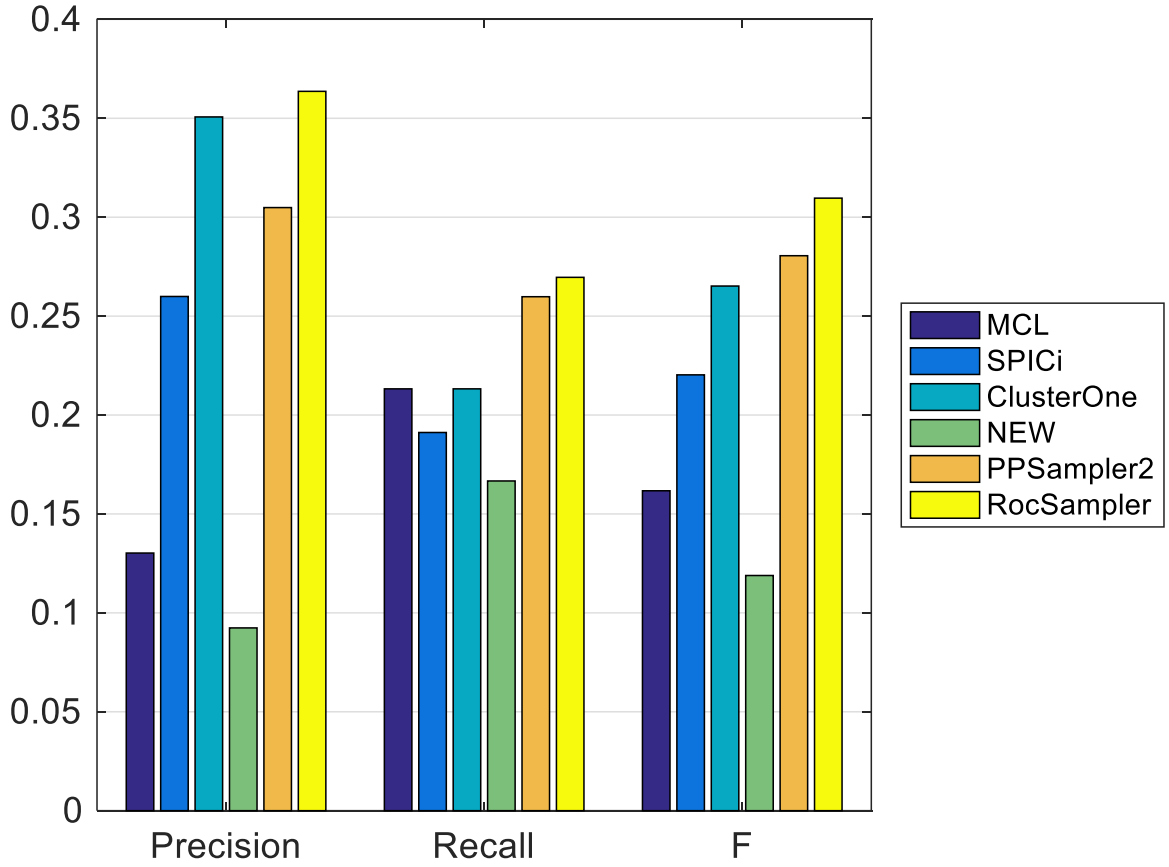
On Collins2007



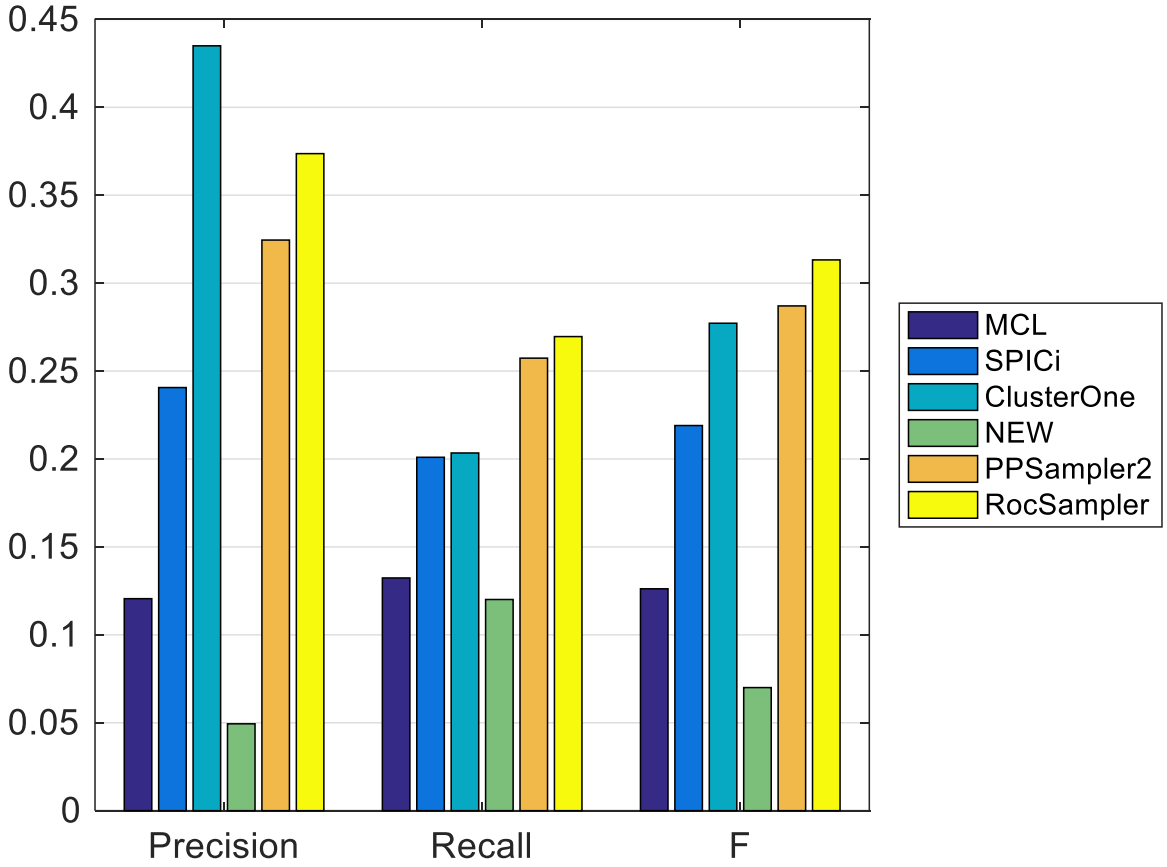
On Gavin2006



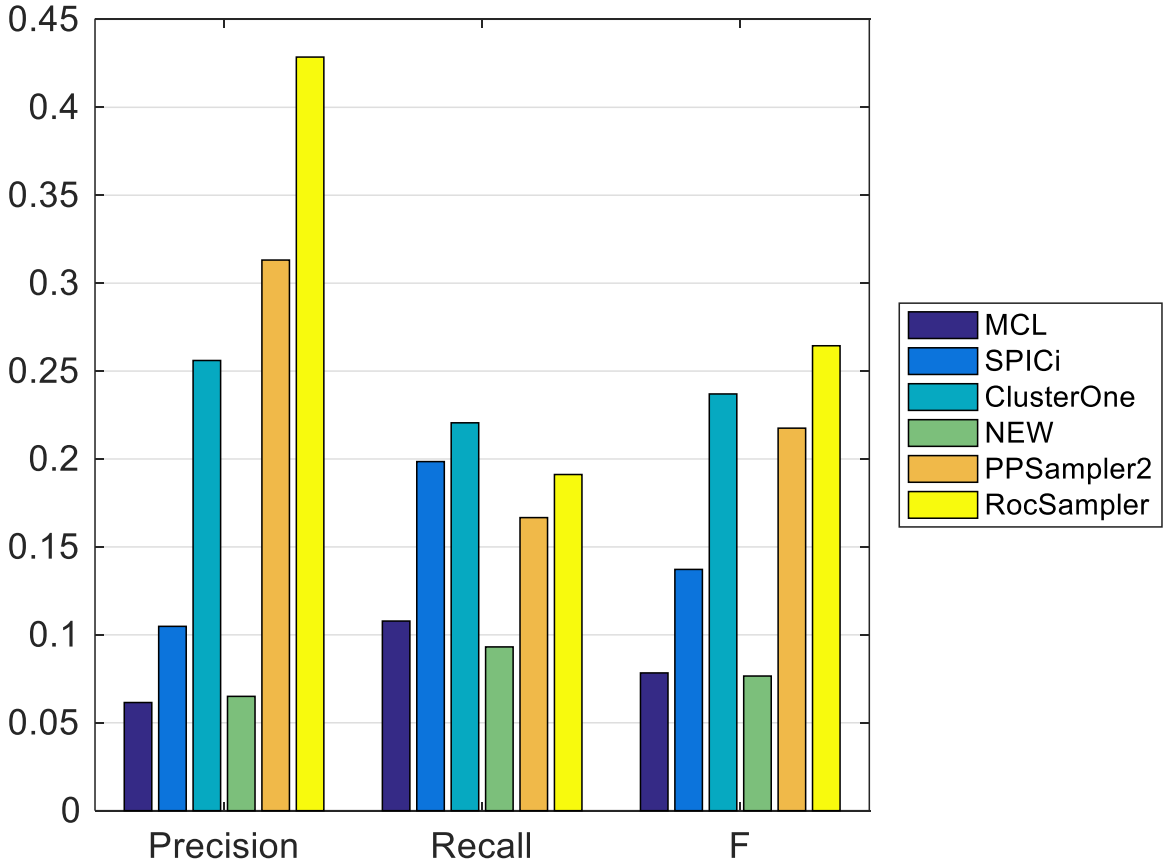
On Krogan2006Core



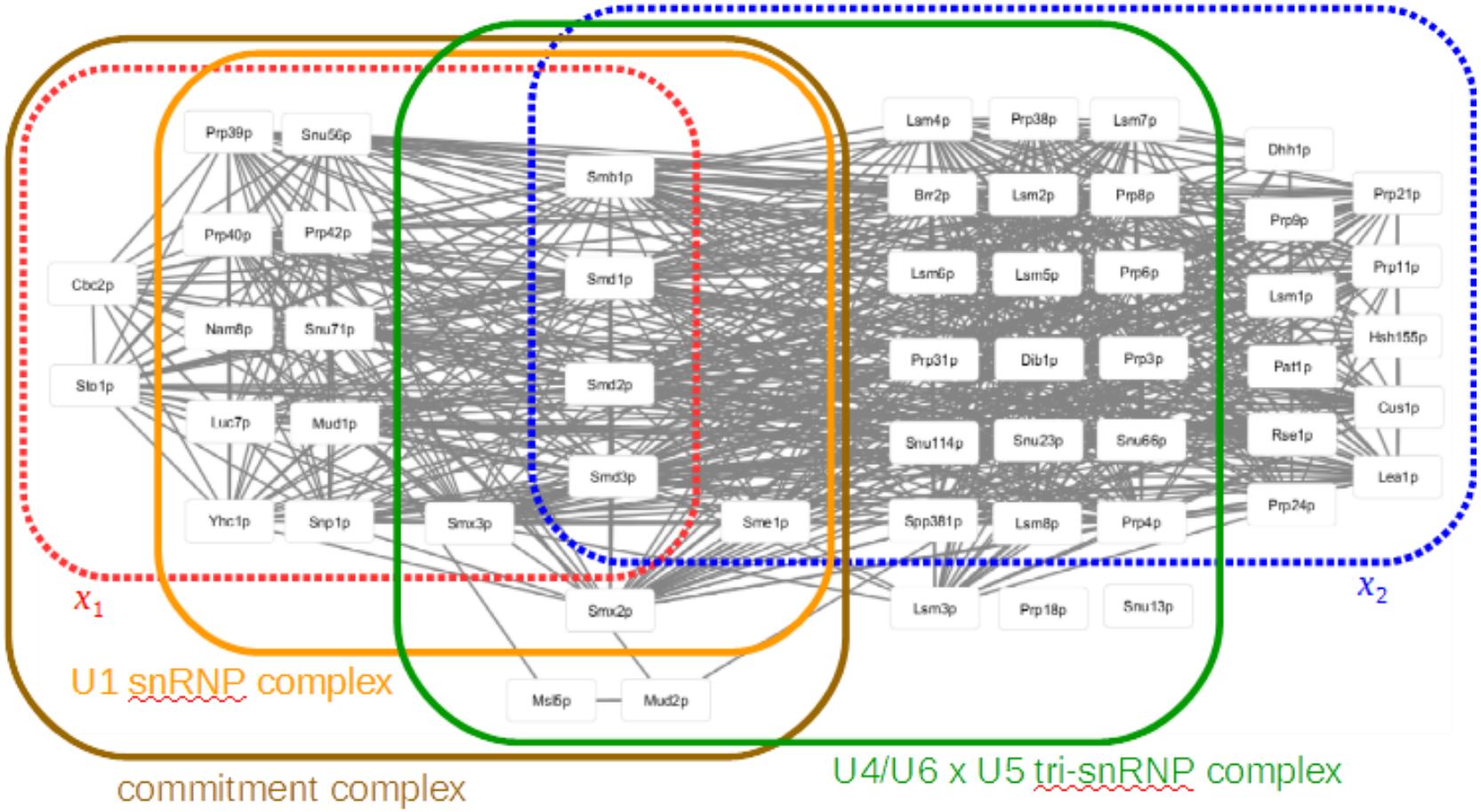
On Krogan2006Extended



On BioGRID (PPIs are unweighted)



On Collins PPIs



Contents

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)