CS4220: Knowledge Discovery Methods for Bioinformatics

Unit 6: Protein-Complex Prediction

Wong Limsoon

NUS National University of Singapore

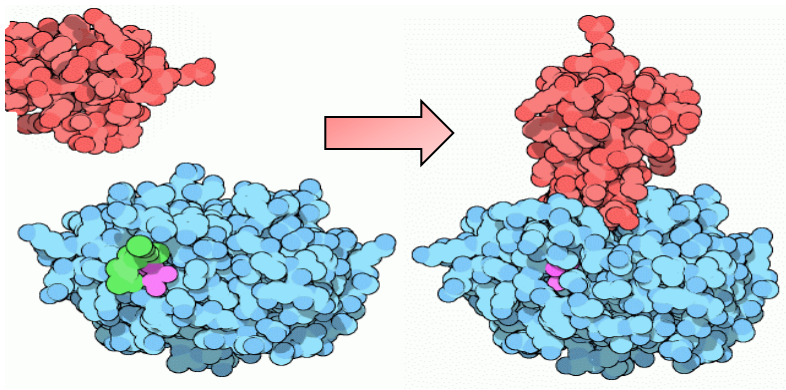
Lecture Outline

- **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

NUS
National University
of Singapore

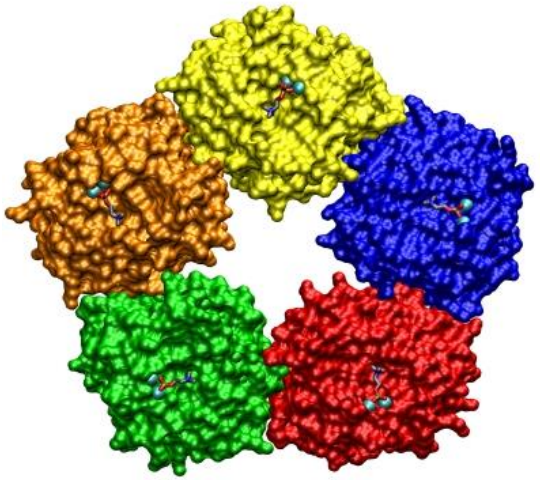
"Assemblies" of Interacting Proteins



Individual proteins come together and interact

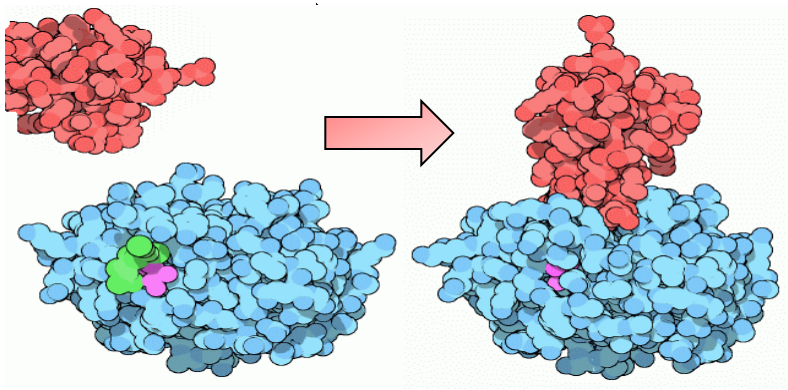
- **Proteins interact to form "protein assemblies"**
- **These assemblies are like "protein machines"**
  - Highly coordinated parts
  - Highly efficient

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



Protein assembly of multiple proteins

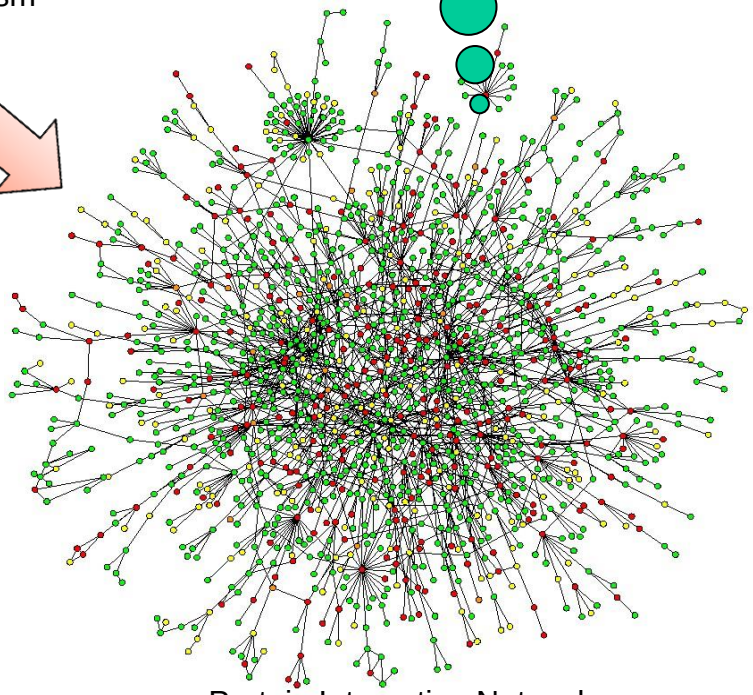
Protein Interaction Networks



Individual proteins come together and interact

Collection of such interactions in an organism

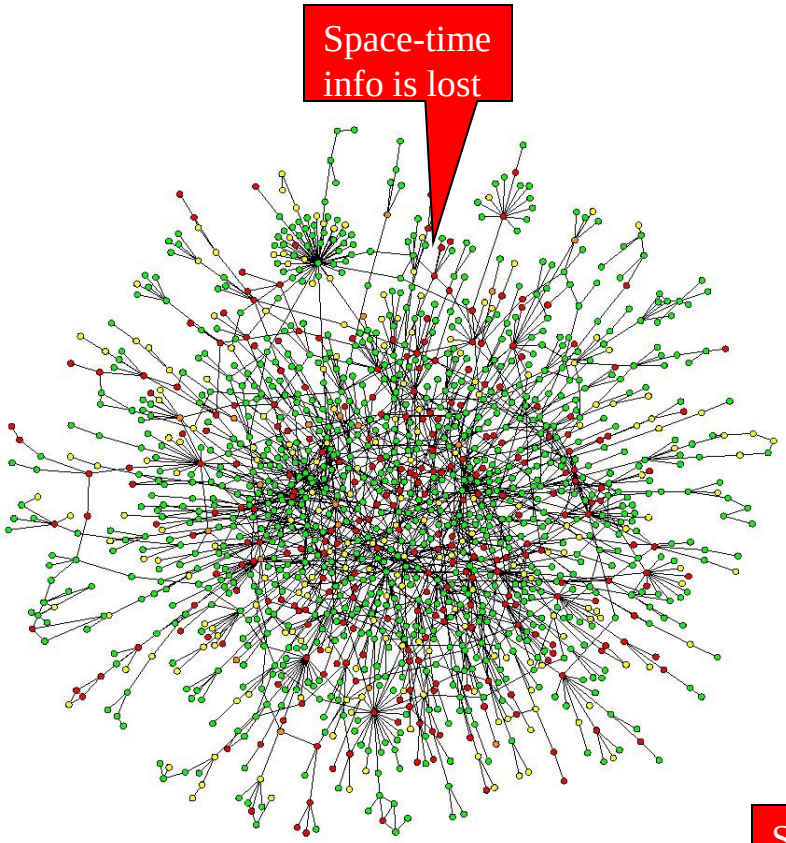
PPIN
Valuable source of knowledge



Protein Interaction Network

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

Detection & Analysis of Protein Complexes in PPIN



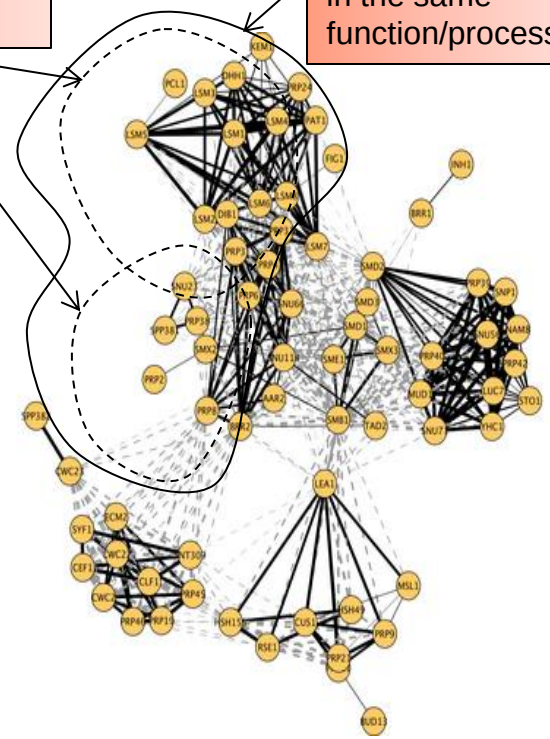
PPIN derived from several high-throughput expt

Identifying embedded complexes

Space-time info is "recovered"

Individual complexes (Some might share proteins)

Entire module might be involved in the same function/process



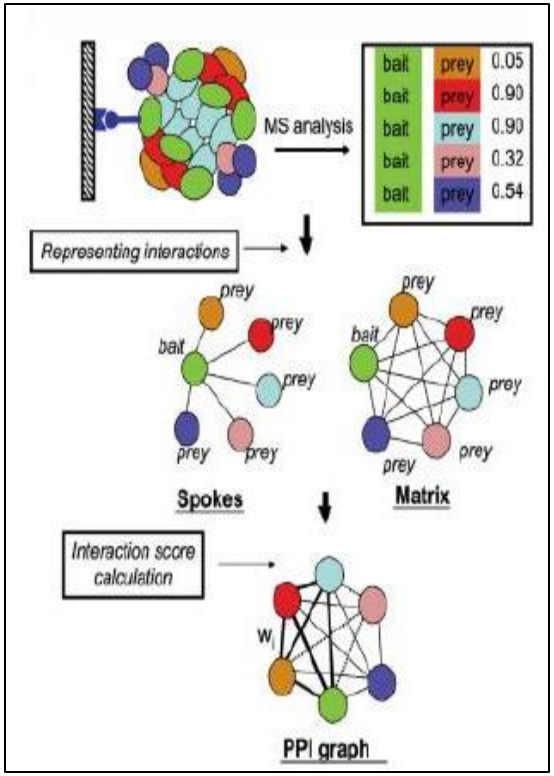
Embedded **complexes** identified from PPIN

Identifying Complexes from PPIN: The Complete Picture

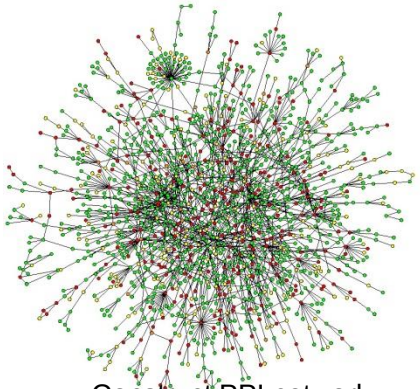
1. Affinity purification followed by MS for identifying “baits” and “preys” (*in vitro*)

2. Arriving at a close approximation to the *in vivo* network

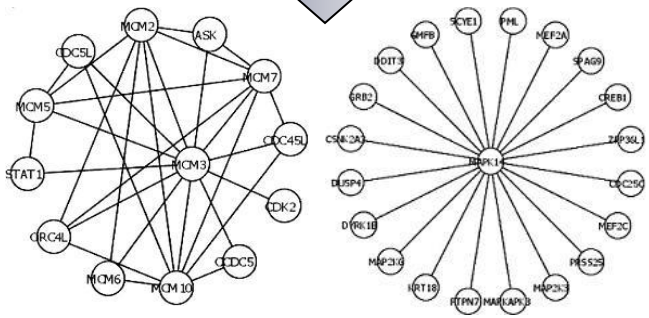
3. Identifying complexes from the PPI network



Computational techniques

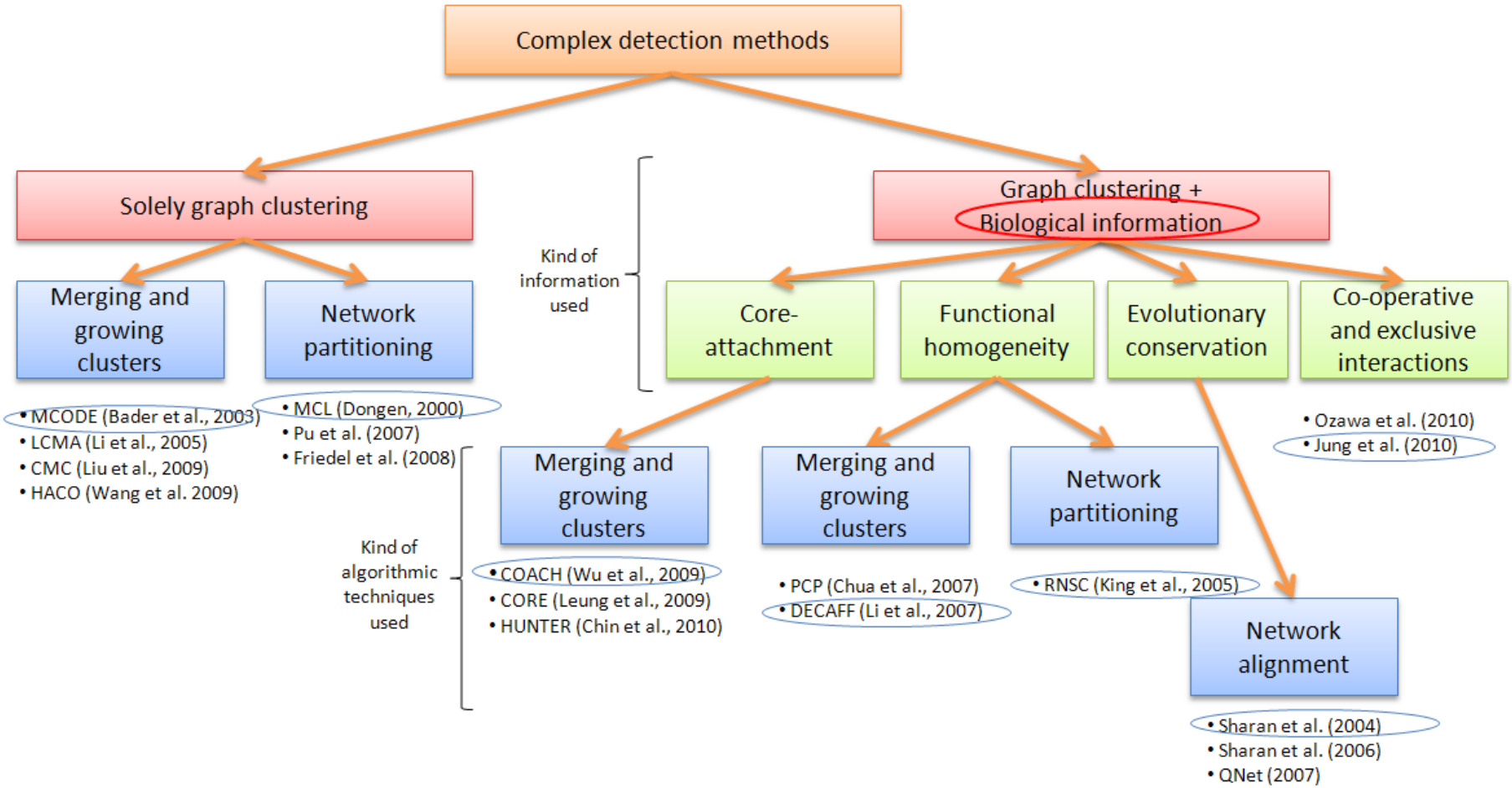


Construct PPI network

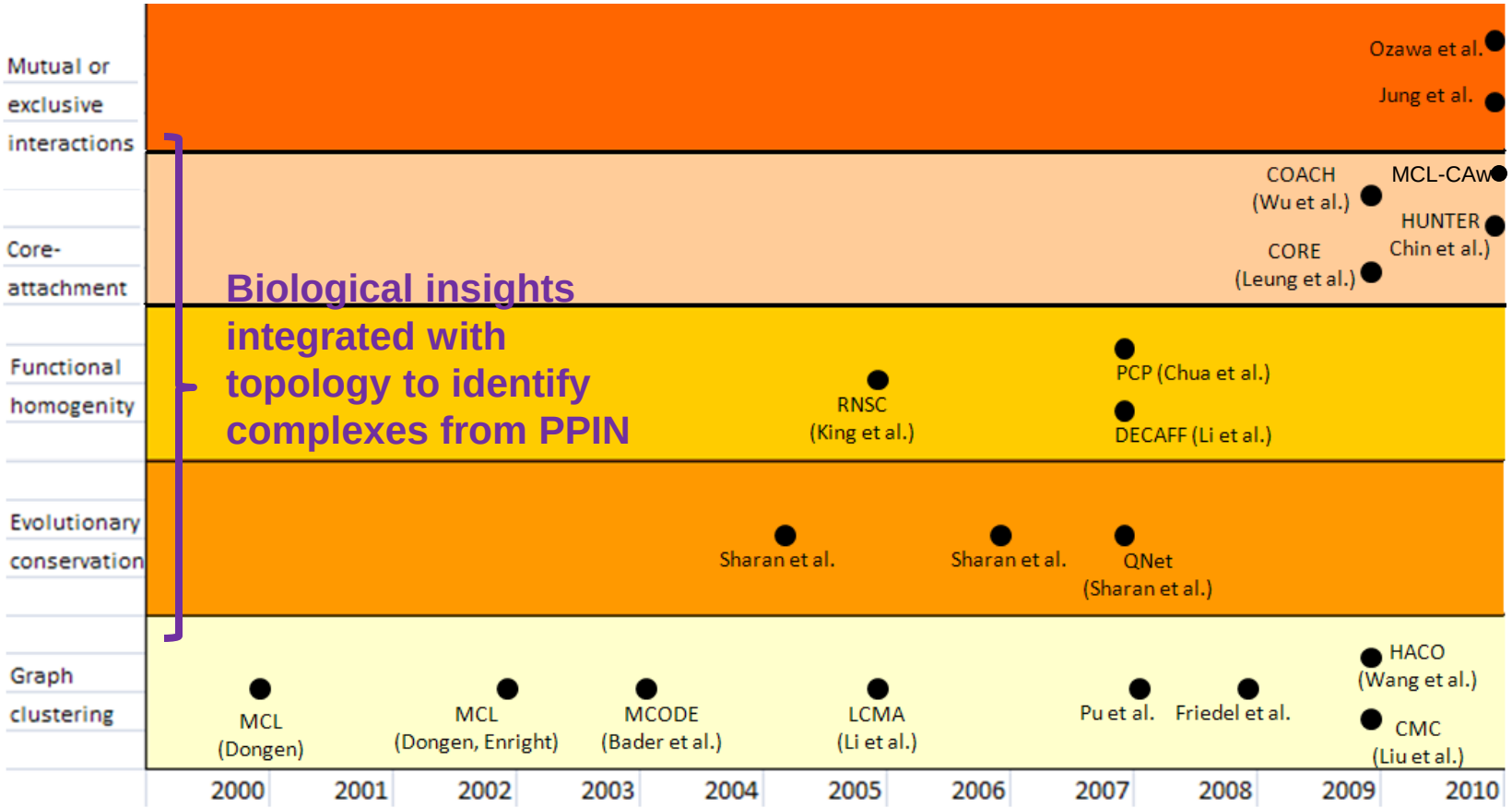


Detect & evaluate complexes

Taxonomy of Protein-Complex Prediction Methods



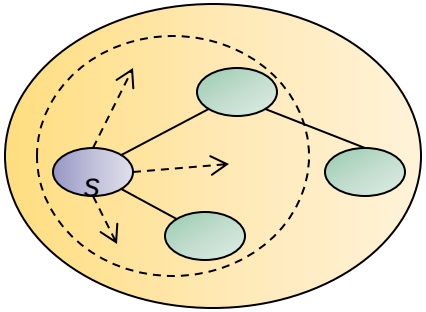
Chronology of Protein-Complex Prediction Methods



- **As researchers try to improve basic graph clustering techs, they also incorporate bio insights into the methods**

Bader & Hogue, "An automated method for finding molecular complexes in large protein interaction networks". *BMC Bioinformatics*, 4:2, 2003

Graph Clustering: MCODE



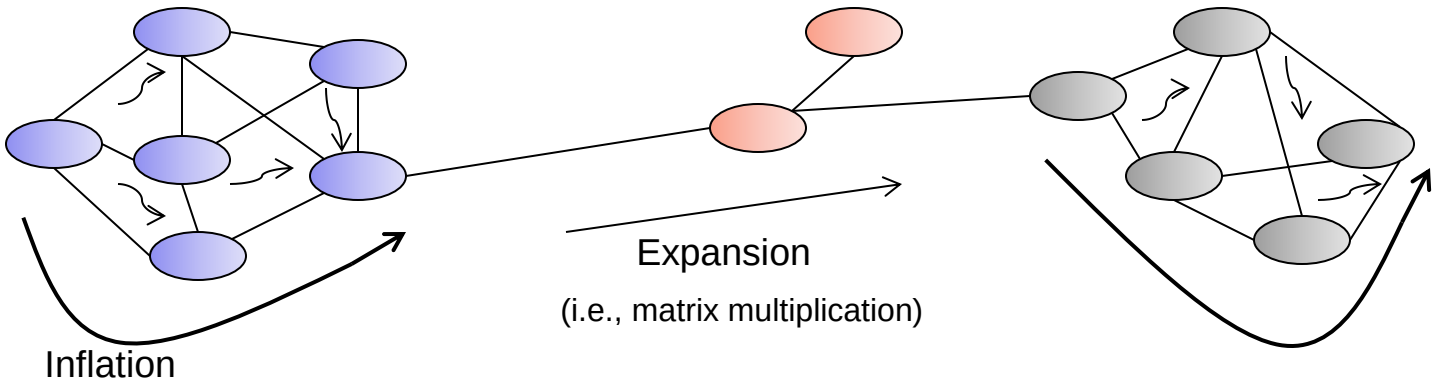
Seed a complex and look in neighborhood

- Weight vertices by density of their immediate neighbourhood
- Select vertices in decreasing order of weights
- 'Seed' a complex using vertex *s*
- Look in neighborhood of *s*
  - Vertex Weight Parameter
- Add vertices to "grow" the complex

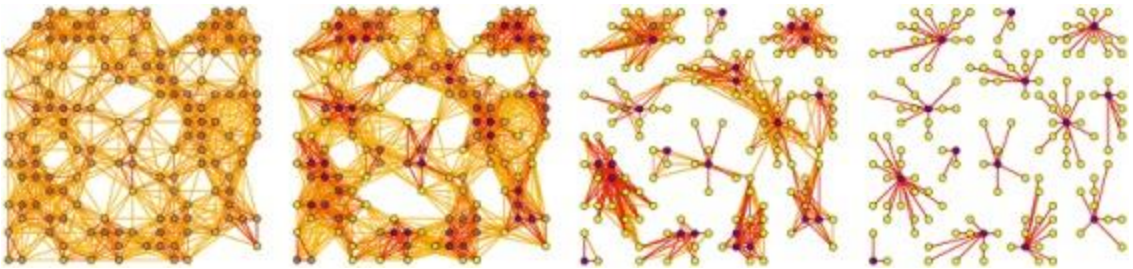
- **Good visualization**
  - MCODE offered as a "plug-in" to Cytoscape
- **Produces very few clusters**
  - High accuracy, but low recall
- **Performs well on highly filtered high-density PPIN**
  - Low tolerance to noise

Pereira-Leal et al. "Detection of functional modules from protein interaction networks", *Proteins: Structure, Function, and Bioinformatics*,54:49-57, 2004

Graph Clustering: MCL



$$(\Gamma_r M)_{pq} = (M_{pq})^r / \sum_{i=1}^{k} (M_{iq})^r$$



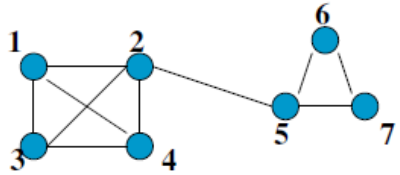
Repeated inflation and expansion separates the network into multiple dense regions

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

Nice slides on MCL, http://www.cs.ucsb.edu/~xyan/classes/CS595D-2009winter/MCL_Presentation2.pdf

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:

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
1 & 0 & .25 & .33 & .33 & 0 & 0 & 0 \\
2 & .33 & 0 & .33 & .33 & .33 & 0 & 0 \\
3 & .33 & .25 & 0 & .33 & 0 & 0 & 0 \\
4 & .33 & .25 & .33 & 0 & 0 & 0 & 0 \\
5 & 0 & .25 & 0 & 0 & 0 & .5 & .5 \\
6 & 0 & 0 & 0 & 0 & .33 & 0 & .5 \\
7 & 0 & 0 & 0 & 0 & .33 & .5 & 0
\end{array}
$$

(notice each column sums to one)

Also can be looked at as a probability matrix!

http://www.cs.ucsb.edu/~xyan/classes/CS595D-2009winter/MCL_Presentation2.pdf

Markov Chains

- A simpler example: $\begin{pmatrix} .6 & .2 \\ .4 & .8 \end{pmatrix}$
- Next time step: $t_0 \rightarrow t_1 \rightarrow t_2$

$1 \rightarrow 1 \rightarrow 1 \;+\; 1 \rightarrow 2 \rightarrow 1$

$.6 * .6 + .4 * .2 = .44$

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

$$\xrightarrow{\text{eventually}} \begin{pmatrix} .33 & .33 \\ .66 & .66 \end{pmatrix}$$

http://www.cs.ucsb.edu/~xyan/classes/CS595D-2009winter/MCL_Presentation2.pdf

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.

http://www.cs.ucsb.edu/~xyan/classes/CS595D-2009winter/MCL_Presentation2.pdf

MCL

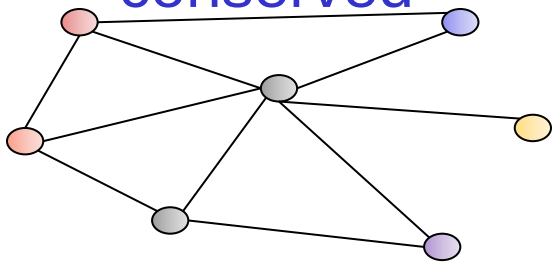
- MCL deliberately boosts this affect by
  - Stopping partway in the Markov Chain
  - Then adjusting the transitions by columns.
    For each vertex, the transition values are changed so that
    - Strong neighbors are further strengthened
    - Less popular neighbors are demoted.
- This adjusting can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

http://www.cs.ucsb.edu/~xyan/classes/CS595D-2009winter/MCL_Presentation2.pdf

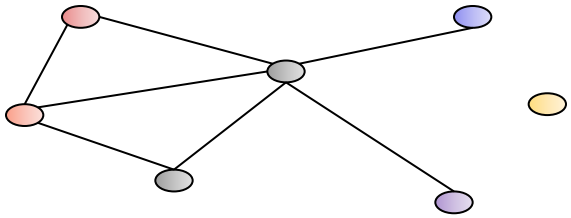
Hirsh & Sharan. "Identification of conserved protein complexes based on a model of protein network evolution". *Bioinformatics*, 23(2):e170-e176, 2007

Evolutionary Insight: Conserved Subnets

- **Assumption**
  - Complexes are evolutionarily conserved

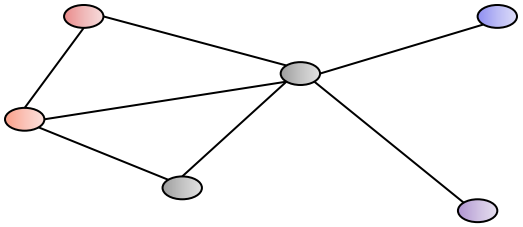


PPI from species 1



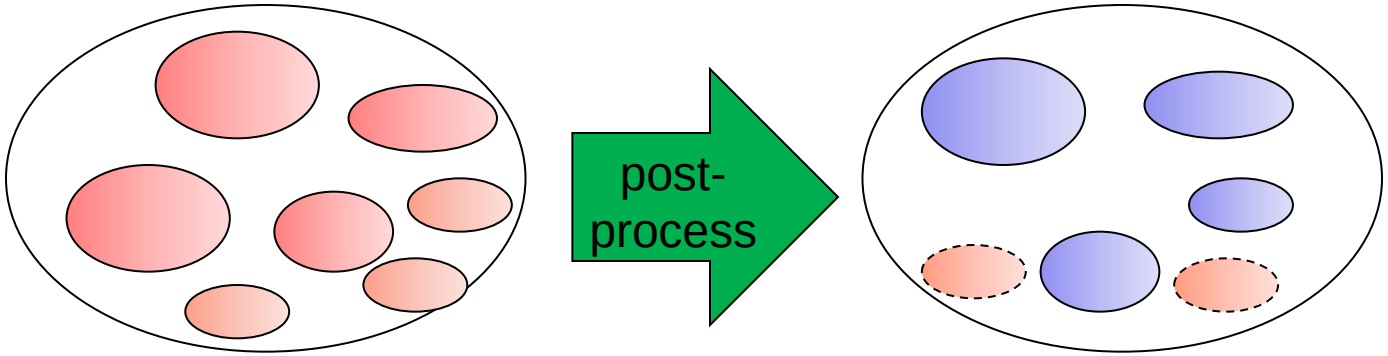
PPI from species 2

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



Orthology network

Functional Info: RNSC & DECAFF

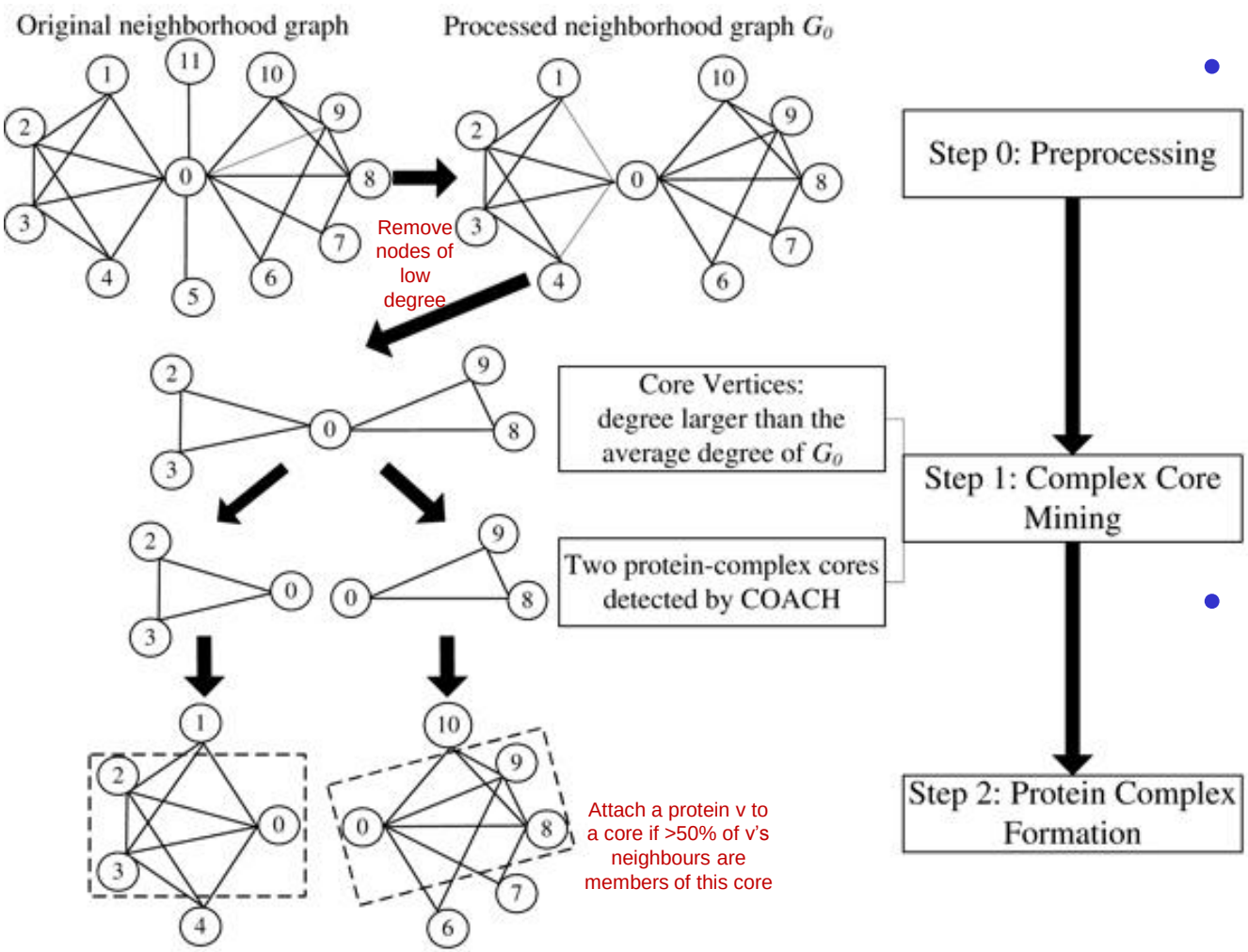


Identify dense candidate complexes

Functionally coherent complexes

- **RNSC**
  - King et al. *Bioinformatics*, 20(17):3013-3020, 2004
  - 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

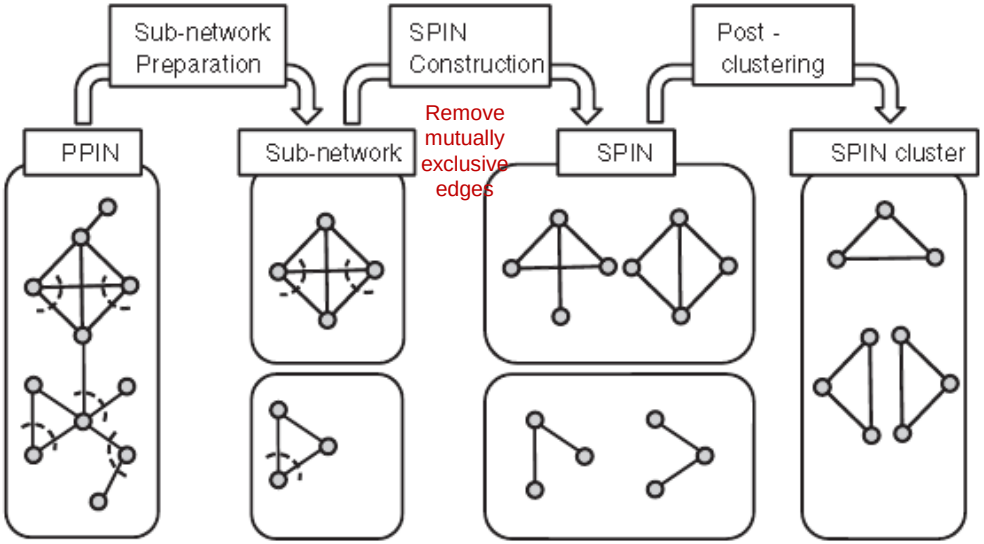
Core-Attachment Structure: COACH



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

Wu et al., "A core-attachment based method to detect protein complexes in PPI networks". *BMC Bioinformatics*, 10:169, 2009

Mutually Exclusive PPIs: SPIN



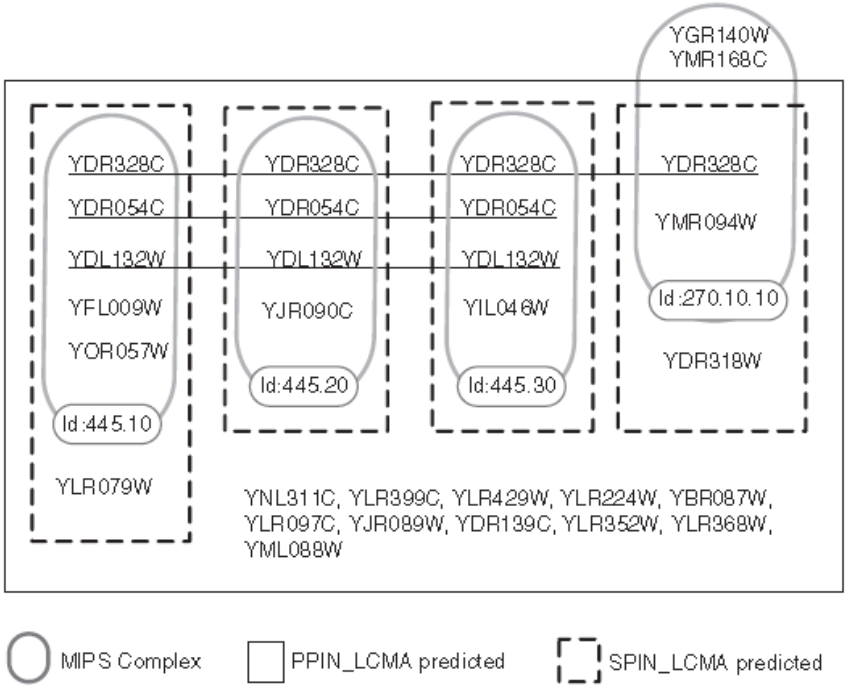


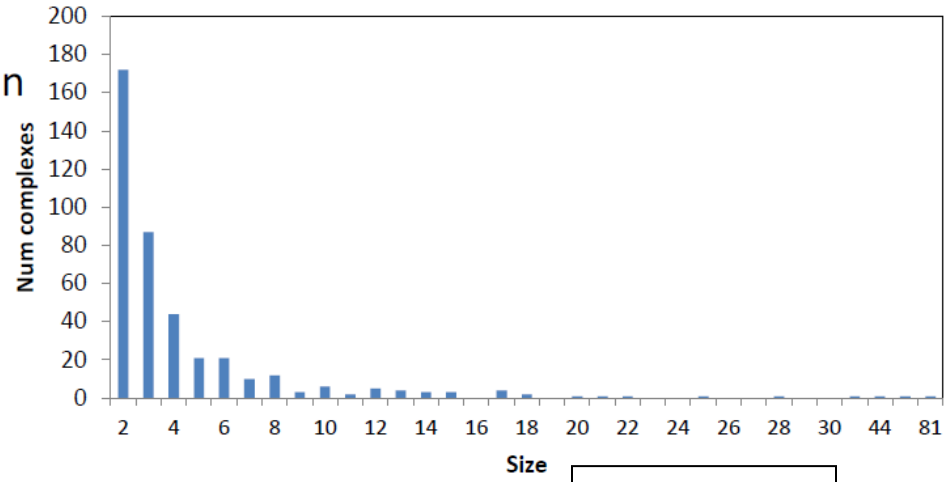
Fig. 6. Comparisons among the known complexes and clusters predicted by LCMAs based on PPIN and SPIN. The gray ovals represent known complexes from MIPS, the quadrangle is a PPIN cluster, and the dotted quadrangles are SPIN clusters. A protein that appears in several complexes is underlined.

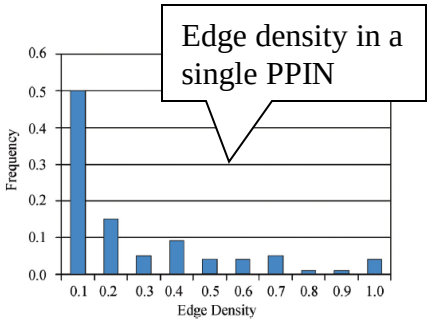
- **+15% in precision & +10% in recall for MCL & MCODE using SPIN**
- **Limitation: Insufficient amt of domain-domain interaction data**

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

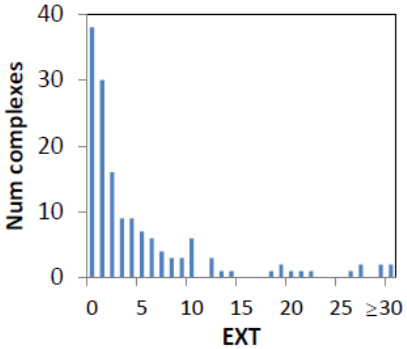
Statistics of Yeast Complexes

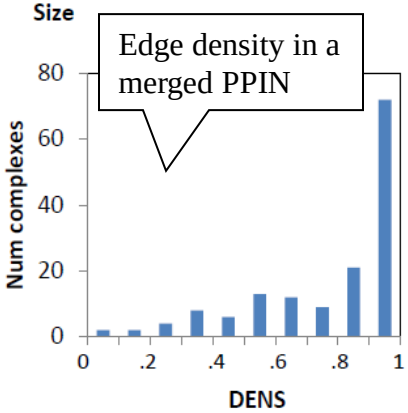
(a) Size distribution





(b) Large complexes





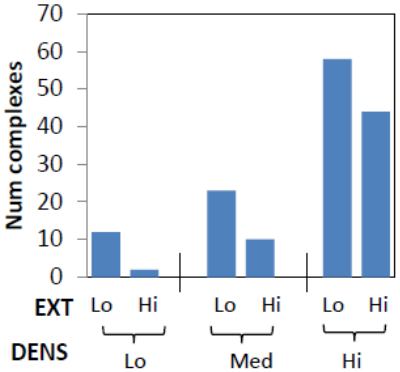


Figure 2.4: Statistics of the yeast reference complexes, from the CYC2008 database. (a) The size distribution of the complexes. (b)EXT (number of highly-connected external proteins) and DENS (density) distributions of large complexes.

Performance of Protein Complex Prediction Methods

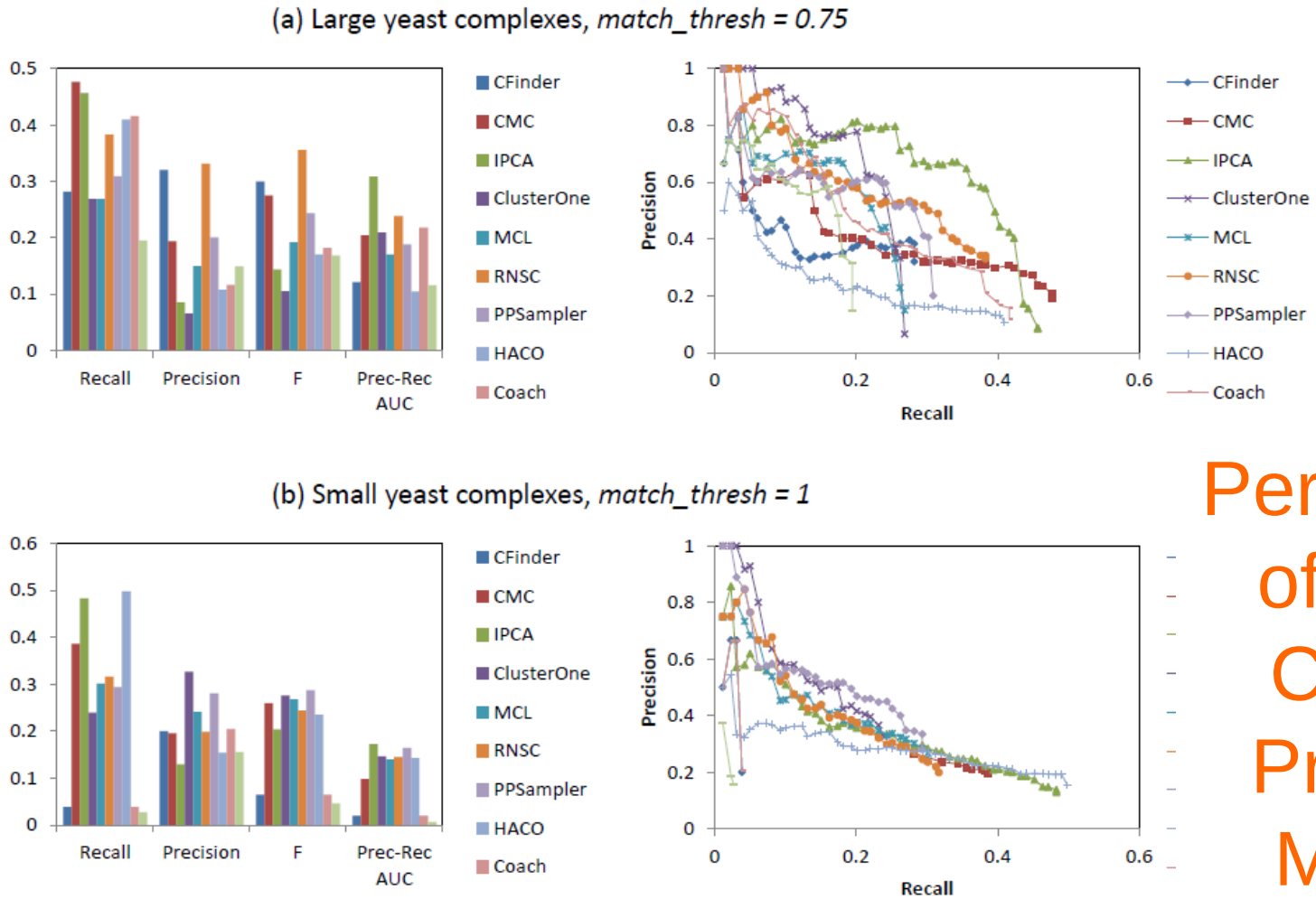


Figure 2.6: Performance of the ten clustering algorithms on prediction of yeast complexes, with (a) $match\_thresh = 0.75$ for large complexes, (b) $match\_thresh = 1$ for small complexes. The left chart shows the precision, recall, F score, and AUC of the precision-recall graph. The right chart shows the precision-recall graph.

W ey dly on

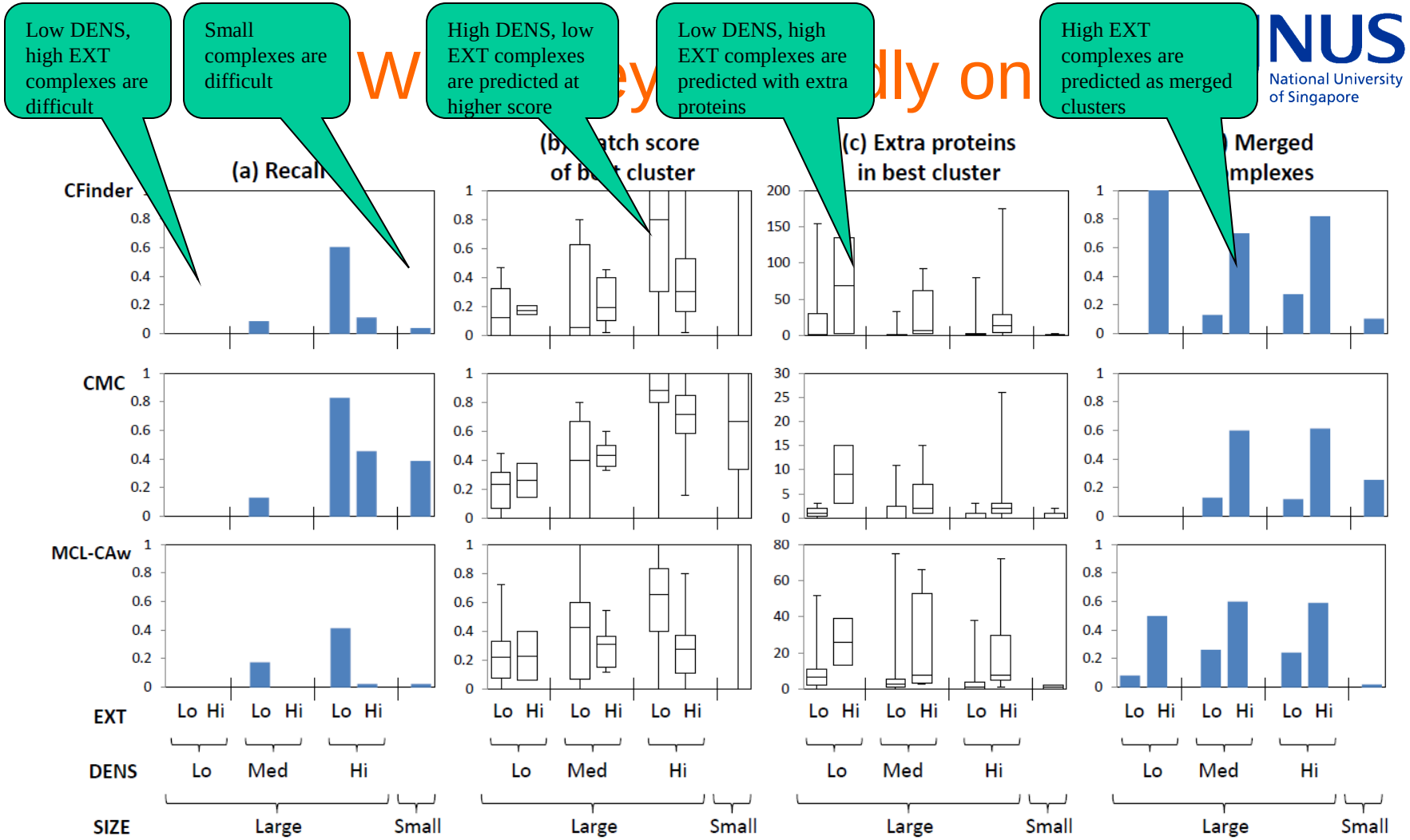


Figure 2.8: Performance of complex-discovery algorithms on yeast complexes, stratified by size, DENS, and EXT. The x-axis of each chart corresponds to the different stratified groups of complexes, given at the bottom of the figure.

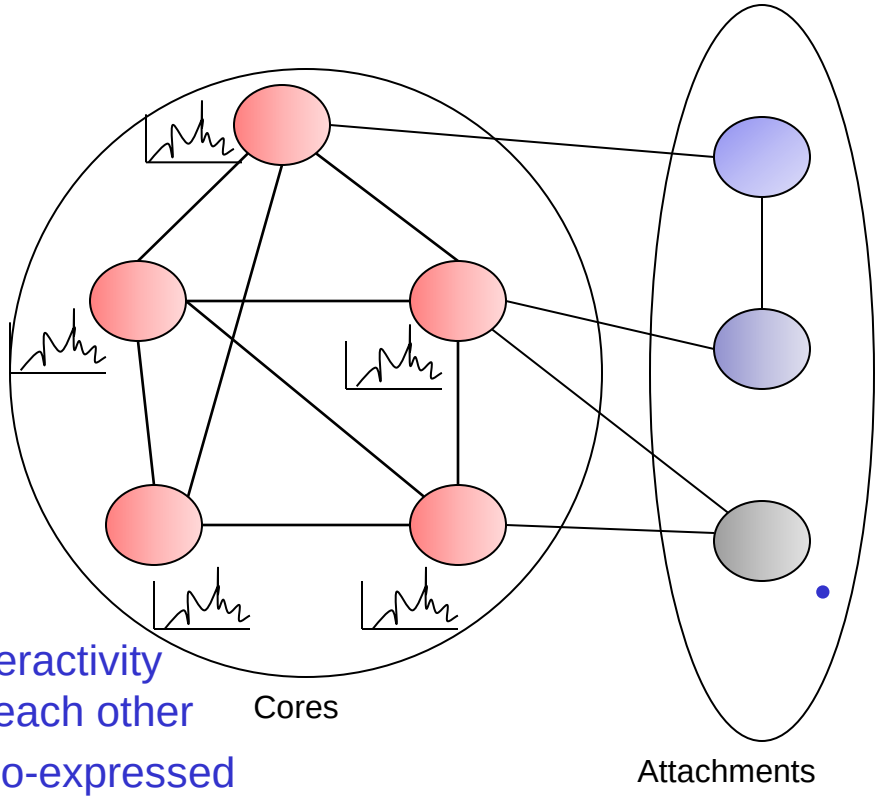
Challenges

- **Recall & precision of protein complex prediction algo's have lots to be improved**
  - Does a "cleaner" PPIN 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

NUS
National University
of Singapore

Core-Attachment Modularity in Yeast Complexes

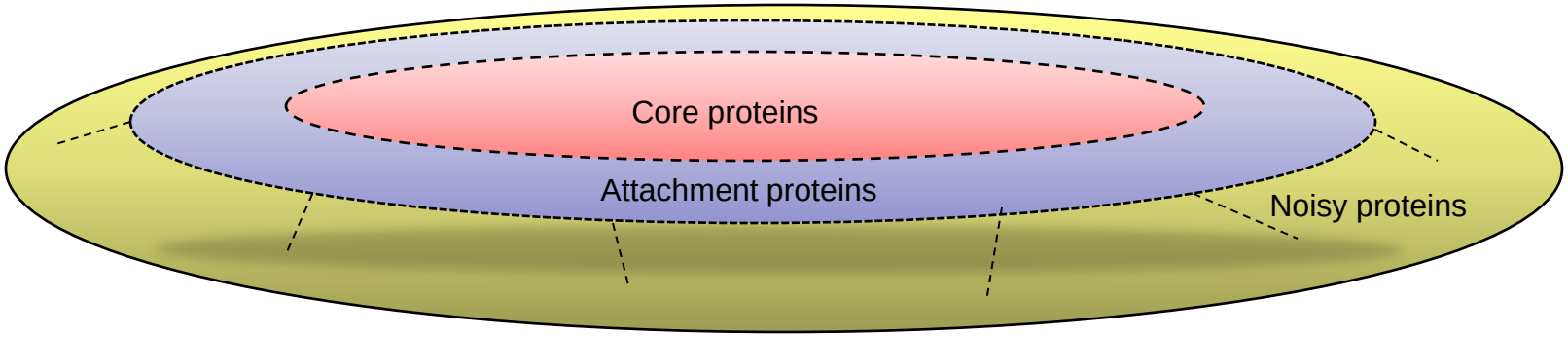


- **Cores**
  - High interactivity among each other
  - Highly co-expressed
  - Main functional units of complexes
- **Attachments**
  - Not co-expressed w/ cores all the time
  - Attach to cores & aid them in their functions
  - May be shared across complexes

Gavin *et al.*, "Proteome survey reveals modularity of the yeast cell machinery", *Nature*, 440:631-636, 2006

MCL-CAw: Key Idea

Identify **dense** regions within PPI network → Identify **core-attachment** structures within these regions → **Extract** them out, discard the rest

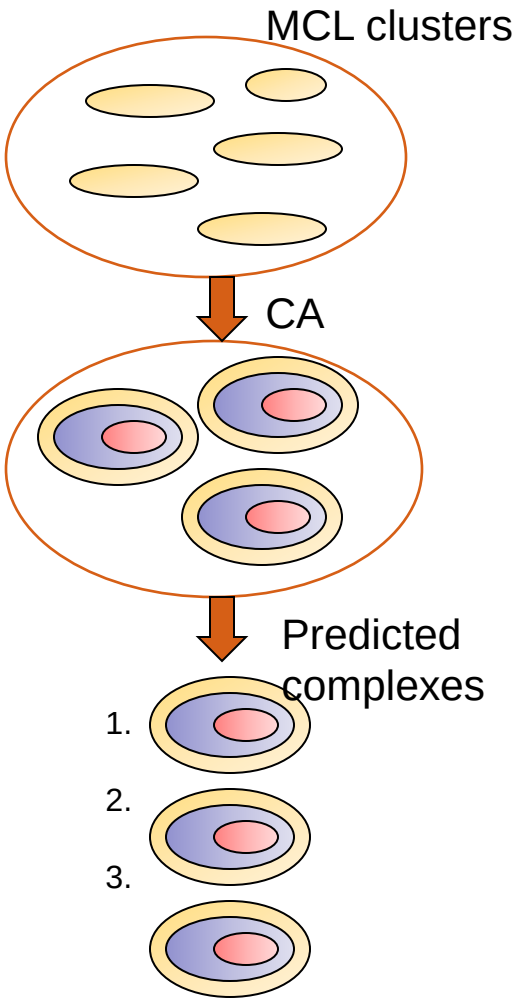


Dense region within a PPI network

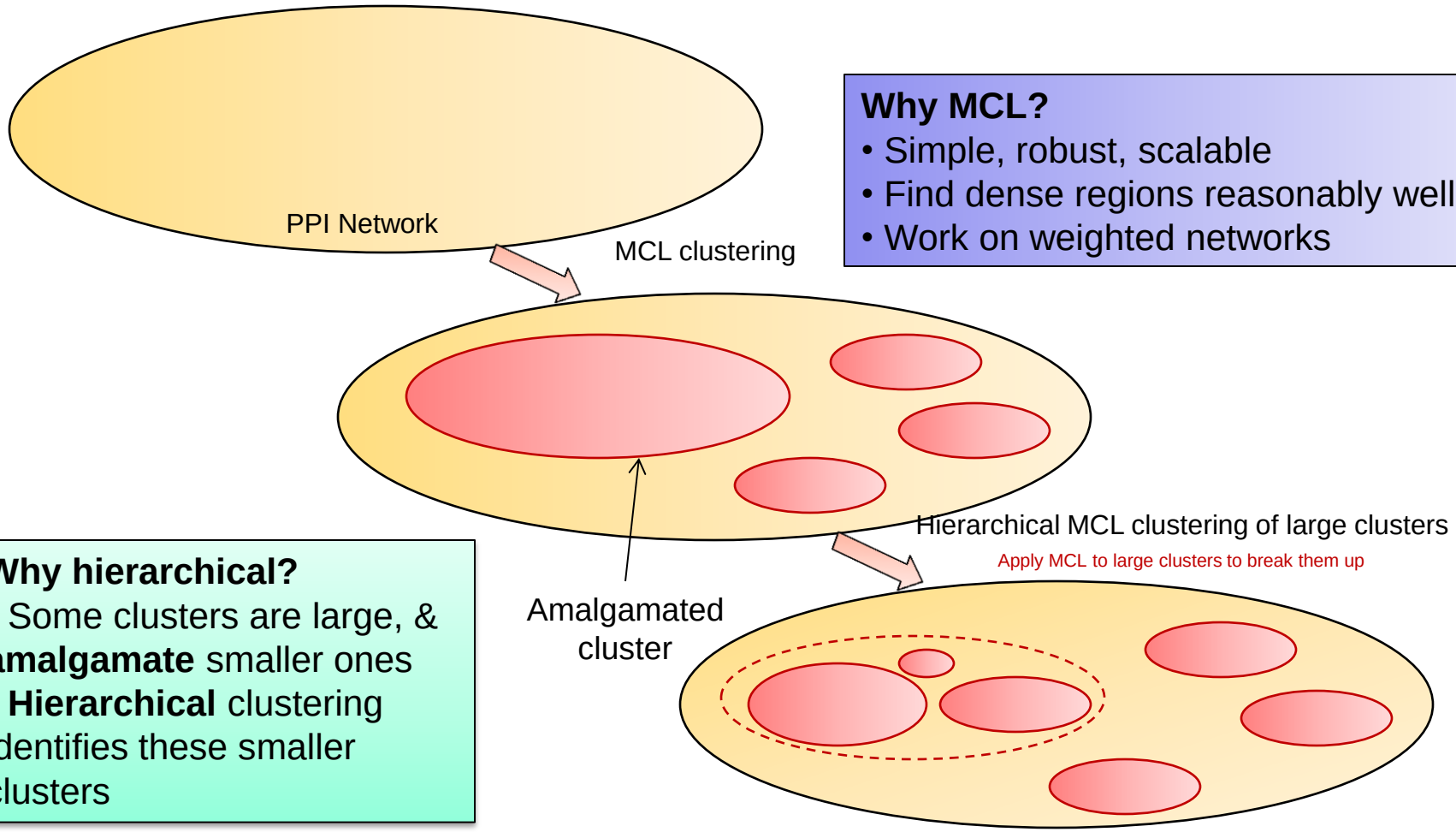
Srihari et al. MCL-CAw: A refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure. *BMC Bioinformatics*, 11:504, 2010

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

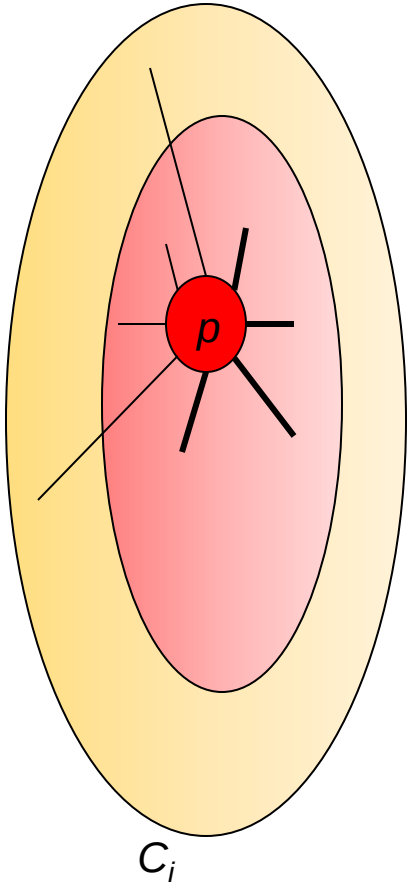
Expect every complex we predict to have a **core**

Protein $p \in$ Core (C_i) *if:*

1. p has high degree w.r.t. C_i
2. p has more neighbors within C_i than outside

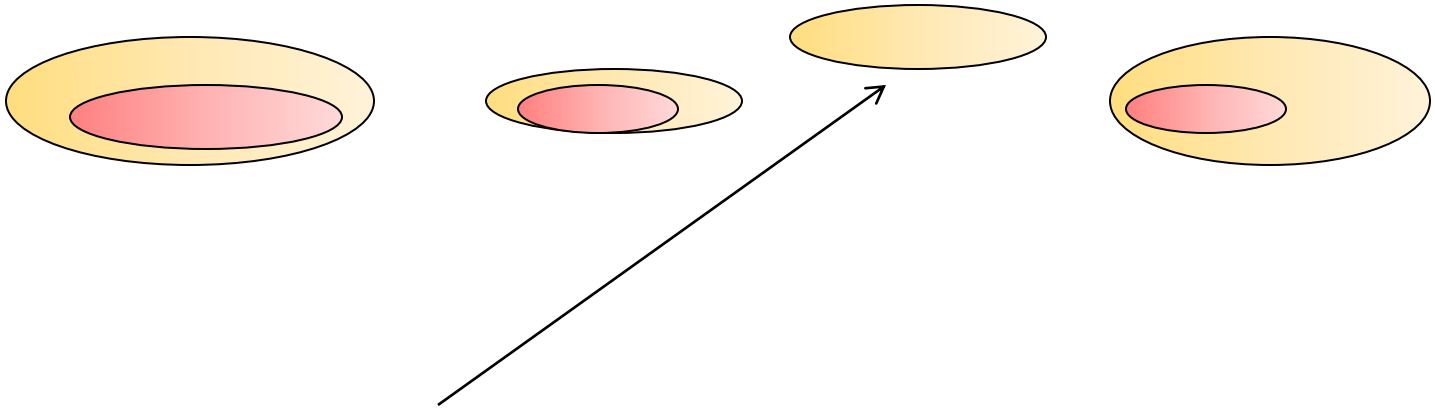
Protein $p \in$ Core (C_i) *if:*

1. *In-degree of p w.r.t.* $C_i \geq$ *Avg in-degree of* C_i
2. *In-degree of p w.r.t.* $C_i >$ *Out-degree of p w.r.t.* C_i
   (Considering weighted degrees)



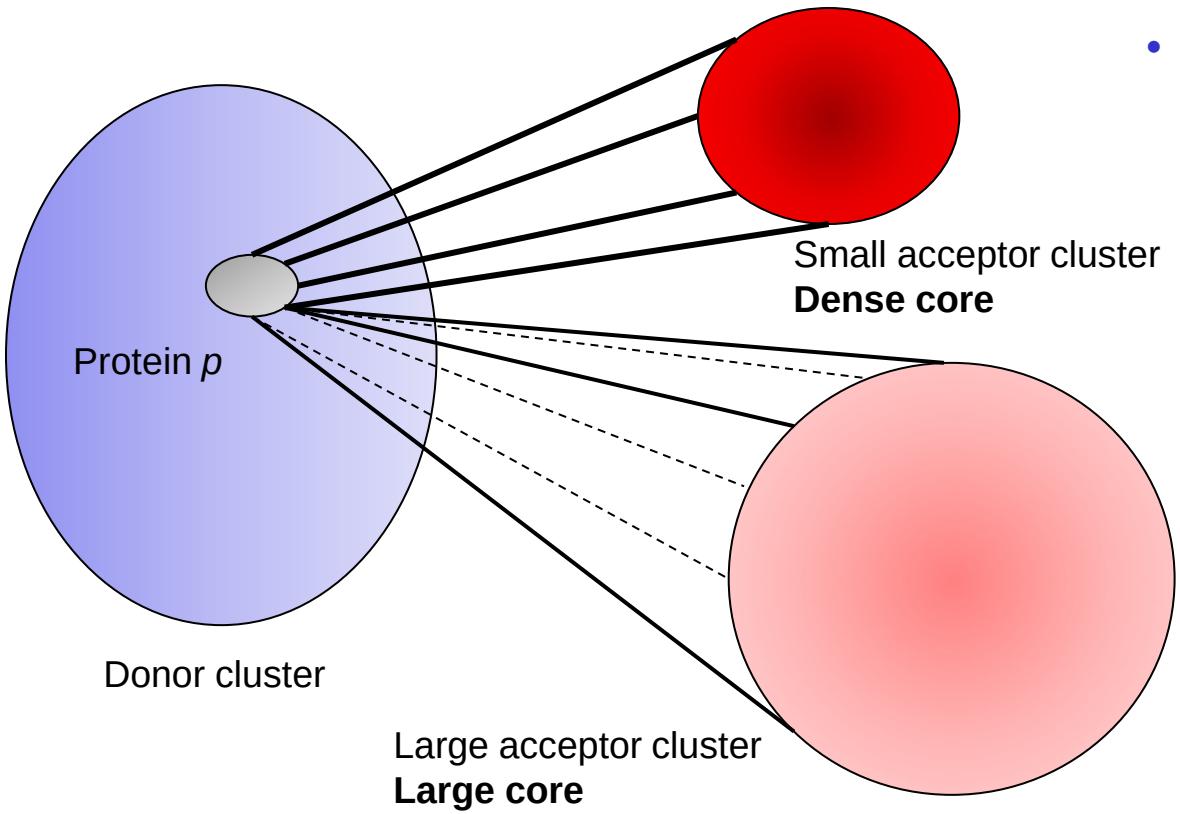
Step 3: Filter Noisy Clusters

- **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

$$Interactions\,(p,\ Core(C_j)) \propto Interactions(Core(C_j))$$

Step 4: Identify Attachments to Cores

$$Interactions(p, Core(C_j)) \propto Interactions(Core(C_j))$$

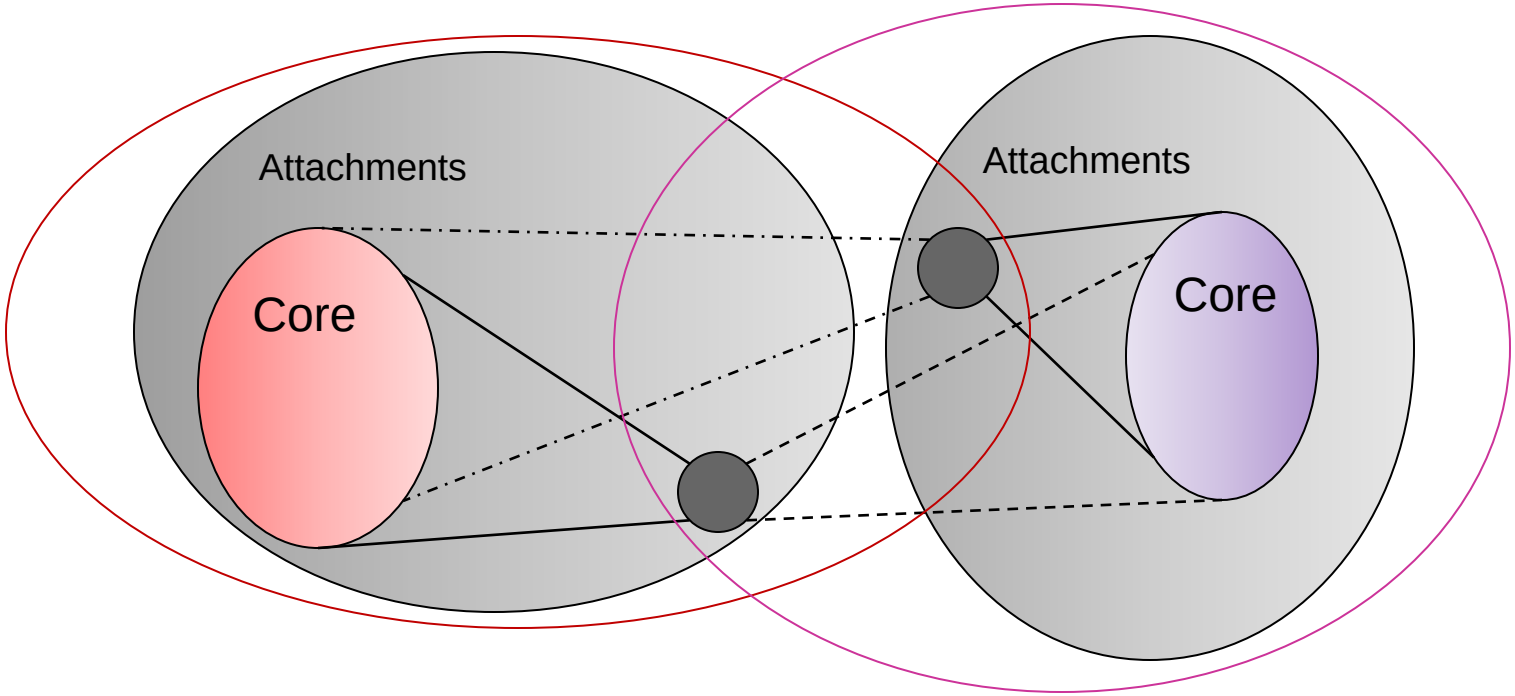
Protein p ∈ Donor cluster Ci is an attachment to Acceptor Core (Cj), if:

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

Parameters α and γ used to control effect of right-hand side

Step 5: Extract Complexes

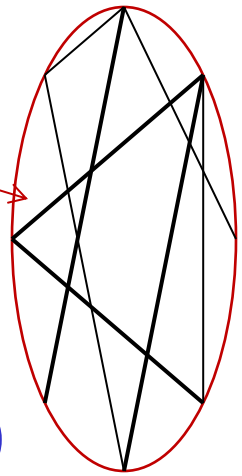
Complex $C = Core(C) \cup Attach(C)$



- **Attachment proteins may be shared betw complexes**

Step 6: Rank Predicted Complexes

- **Weighted density-based ranking of complexes**
  - Reliability of interactions within complex C
  - Size of complex C
  - Weighted density



$$= \Sigma(\text{wt of interactions}) / (|C| * (|C|-1))$$

- 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
  - Gavin et al., *Nature*, 440:631-636, 2006
  - Krogan et al., *Nature*, 440:637-643, 2006

If you don't remember CD-distance, please refer to last lecture!

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

"Gold Standard" Benchmarks Complexes

- **CYC 08: 408 complexes**
  - Pu et al., *Nucleic Acids Res.*, 37(3):825-831, 2009
- **MIPS: 313 complexes,**
  - Mewes et al., *Nucleic Acids Res.*, 32(Database issue):41–44, 2006
- **Aloy: 101 complexes,**
  - Aloy et al., *Science*, 303(5666):2026–2029 2004

			size			density	
Datasets	#cmplx	#proteins	max	avg	median	avg	median
Aloy	63	544	34	9.22	7	0.865	0.944
CYC08	148	1115	81	8.84	6	0.831	0.944
MIPS	156	1171	95	14.86	9	0.565	0.564
Combined	305	1543	95	11.85	7	0.697	0.800

Size > 3

Measured based on BioGrid yeast physical PPIN

Evaluation of MCL-CAw

G+K

	Method	F1	Norm
1.	CMC	1.146	1.000
2.	HACO	0.899	0.785
3.	MCL-CAw	0.800	0.700
4.	CORE	0.757	0.661
5.	MCLO	0.734	0.641
6.	MCL	0.717	0.626
7.	COACH	0.515	0.450

G+K (ICD)

	Method	F1	Norm
1.	MCL-CAw	1.595	1.000
2.	HACO	1.536	0.962
3.	CMC	1.516	0.950
4.	MCLO	1.414	0.886
5.	MCL	1.411	0.884

- CORE and COACH assume only unweighted networks

Adding the F1 scores across all three benchmarks and normalizing against the best

- **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!**

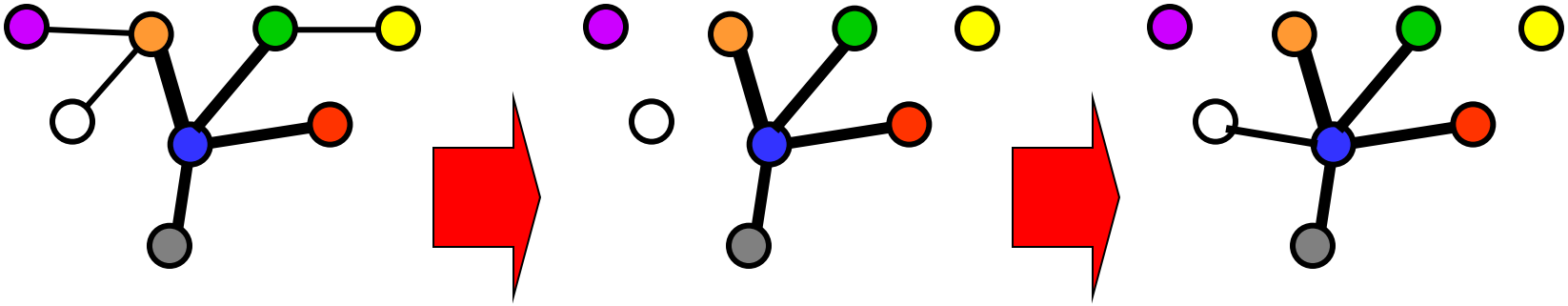
Limitations of MCL-CAw

- **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

NUS
National University
of Singapore

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**

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

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 the 1st lecture!

Some Details of CMC

- **Iterated CD-distance is used to weigh PPI's**

$$w^k(u,v) = \frac{\sum_{x \in N_u \cap N_v} (w^{k-1}(x,u) + w^{k-1}(x,v))}{\sum_{x \in N_u} w^{k-1}(x,u) + \lambda_u^k + \sum_{x \in N_v} w^{k-1}(x,v) + \lambda_v^k}$$

- **Clusters are ranked by weighted density**

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

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

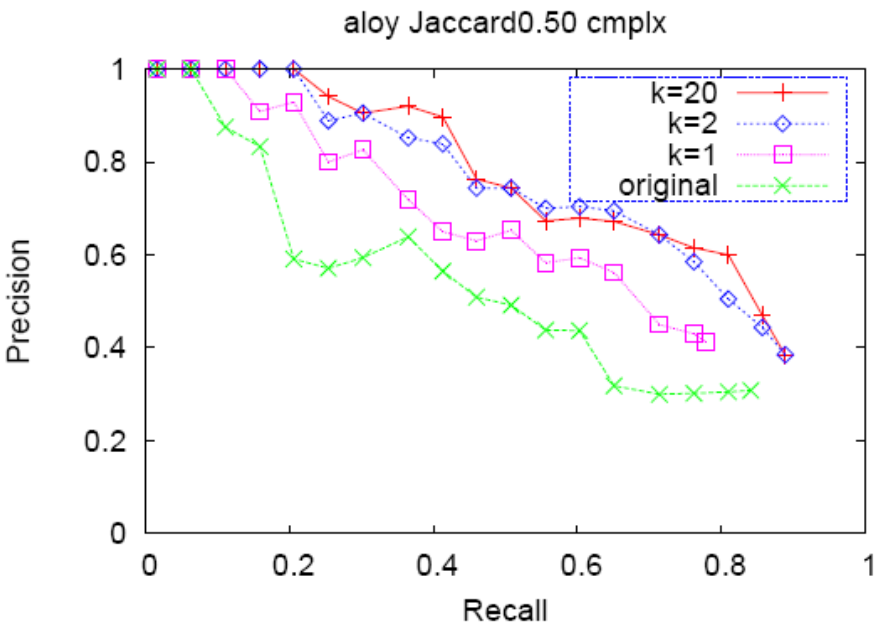
$$inter\text{-}score(C_1, C_2)$$

$$= \sqrt{\frac{\sum_{u \in (C_1 - C_2)} \sum_{v \in C_2} w(u,v)}{|C_1 - C_2| \cdot |C_2|} \cdot \frac{\sum_{u \in (C_2 - C_1)} \sum_{v \in C_1} w(u,v)}{|C_2 - C_1| \cdot |C_1|}}$$

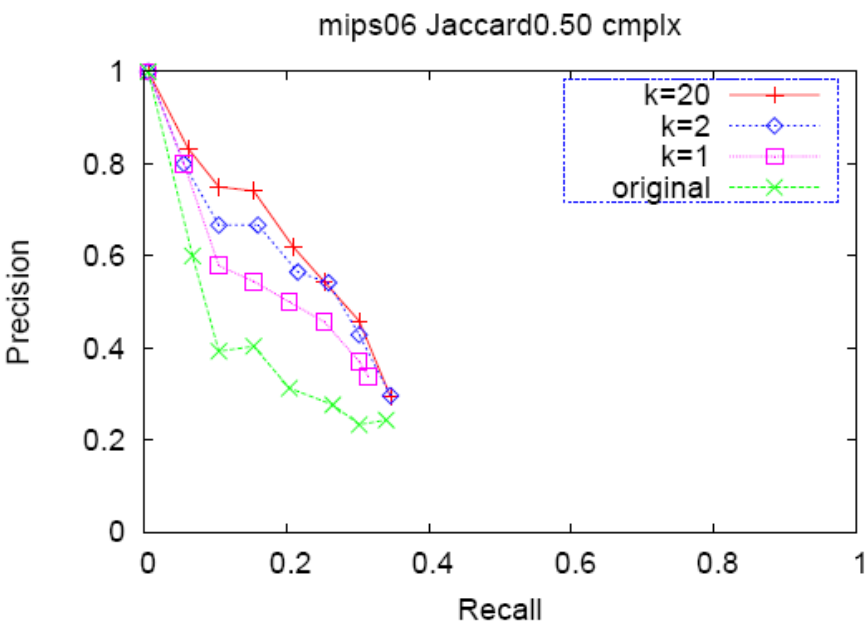
Validation Experiments

- **Matching a predicted complex S with a true complex C**
  - Vs: set of proteins in S
  - Vc: set of proteins in C
  - Overlap(S, C) = $|Vs \cap Vc| / |Vs \cup Vc|$, 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



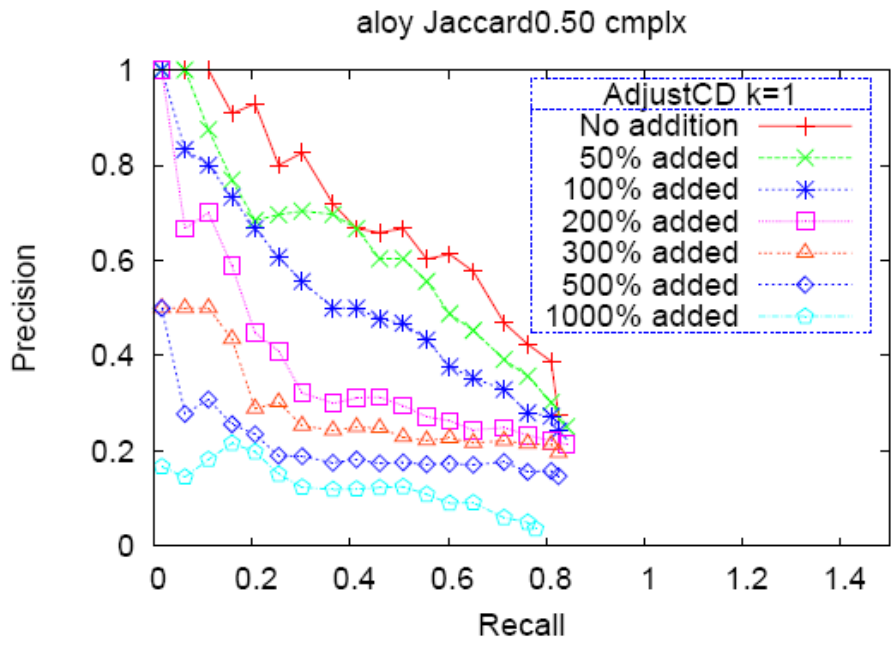
(a) Aloy, $match\_thres$=0.50

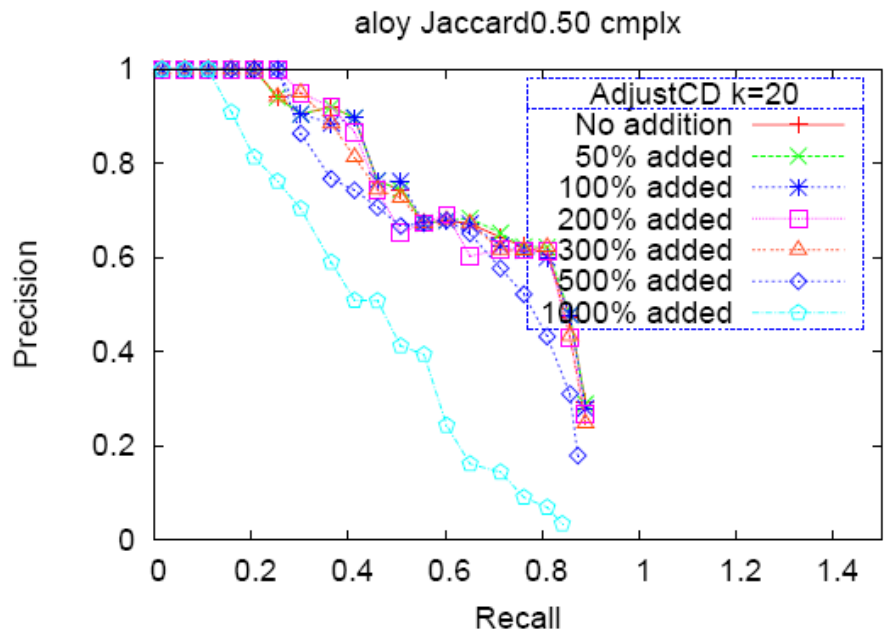


(b) MIPS, $match\_thres$=0.50

- **Cleaning by Iterated CD-distance improves recall & precision of CMC**

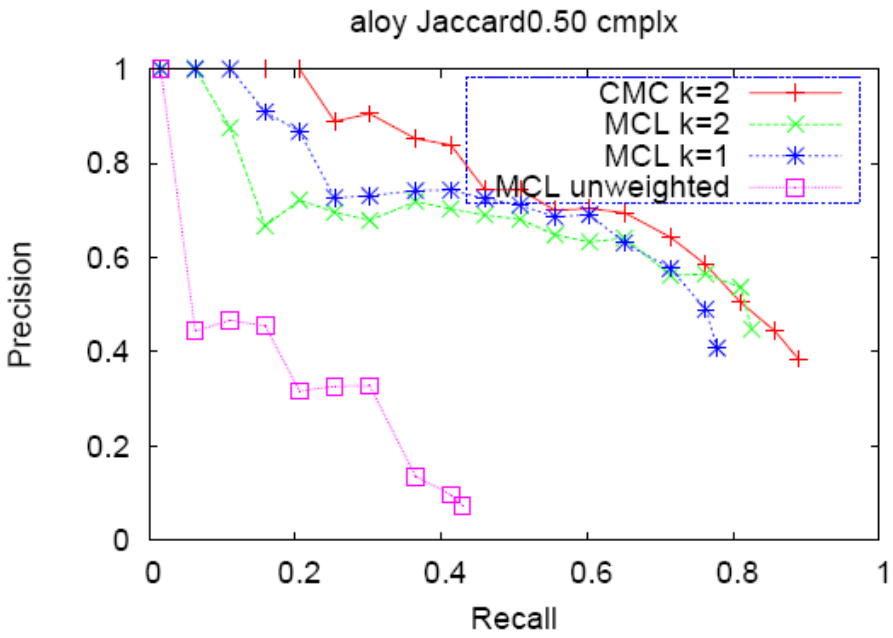
Noise Tolerance of CMC

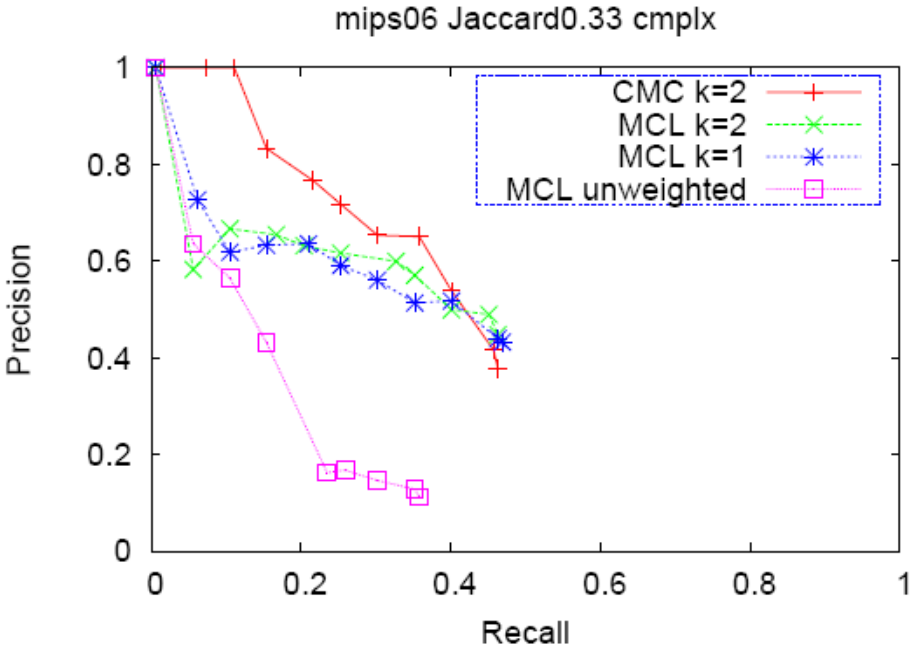




- **If cleaning is done by iterating CD-distance 20 times, CMC can tolerate up to 500% noise in the PPI network!**

Effect of Cleansing on MCL





- **MCL benefits significantly from cleaning too**

Ditto for other methods…

scoring method: AdjustCD					*match_thres*=0.50							
					Aloy (#complexes: 63)				**MIPS (#complexes: 162)**			
clustering methods	k	#clusters	avg size	loc_ score	#matched clusters	precision	#matched complxes	recall	#matched clusters	prec	#matched complxes	recall
CMC	0	172	9.83	0.823	53	0.308	53	0.841	42	0.244	55	0.340
	1	121	9.42	0.897	50	**0.413**	49	0.778	41	**0.339**	51	0.315
	2	148	8.50	0.899	57	0.385	**56***	**0.889**	44	0.297	**56***	**0.346**
	20	146	8.78	0.891	56	0.384	**56***	**0.889**	43	0.295	**56***	**0.346**
CFinder	0	103	13.84	0.528	39	0.379	38	0.603	34	0.330	40	0.247
	1	76	12.86	0.724	38	**0.500**	38	0.603	30	**0.395**	34	0.210
	2	95	11.66	0.713	44	0.463	**43**	**0.683**	36	0.379	46	0.284
	20	95	11.77	0.718	44	0.463	**43**	**0.683**	37	0.389	**49**	**0.302**
MCL	0	372	9.40	0.638	27	0.073	27	0.429	30	0.081	37	0.228
	1	120	10.18	0.848	49	0.408	49	0.778	40	0.333	**51**	**0.315**
	2	116	10.31	0.856	52	**0.448**	**52**	**0.825**	41	**0.353**	**51**	**0.315**
	20	110	10.75	0.849	49	0.445	49	0.778	37	0.336	47	0.290
MCode	0	61	7.31	0.849	20	0.328	20	0.317	18	0.295	22	0.136
	1	103	7.42	0.913	35	0.340	**35**	**0.556**	30	0.291	**39**	**0.241**
	2	88	8.67	0.897	34	**0.386**	34	0.540	29	**0.330**	**39**	**0.241**
	20	82	10.28	0.838	29	0.354	29	0.460	23	0.280	32	0.198

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 ~90%**
- **65/85 have the same informative GO term annotated to > 50% of proteins in the cluster**

⇒ **Likely to be real complexes**

Detecting Overlapping Protein Complexes from Dense Regions of PPIN

NUS
National University of Singapore

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

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

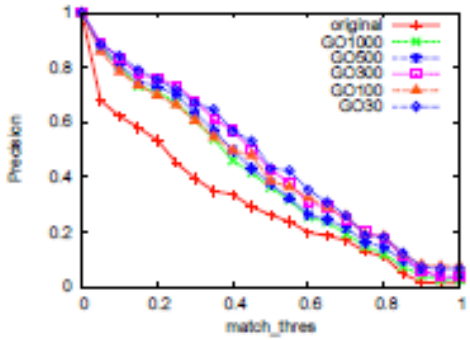
Idea I: Split by Localization GO Terms

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

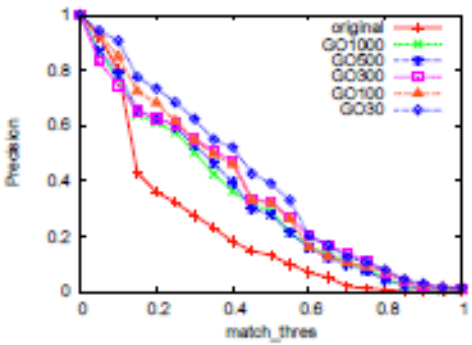
$\Rightarrow$ **Use general cellular component (CC) GO terms to decompose a given PPI network into several smaller PPI networks**

- **Use "general" CC GO terms as it is easier to obtain rough localization annotation of proteins**
  - How to choose threshold N_{GO} to decide whether a CC GO term is "general"?

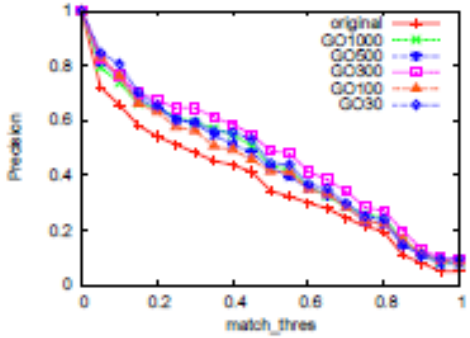
Effect of N_{GO} on Precision



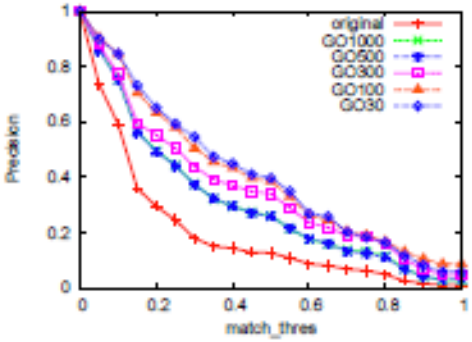
(b) MCL precision



(f) IPCA precision



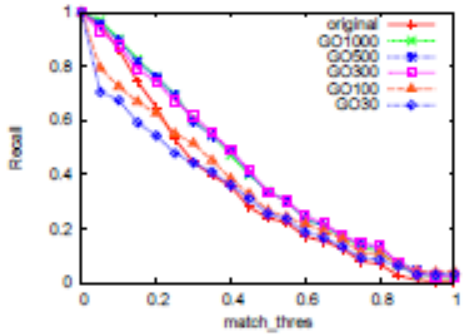
(d) RNSC precision



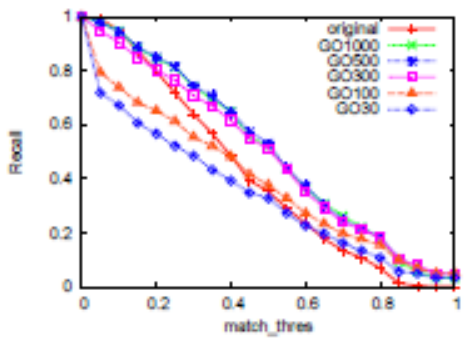
(h) CMC precision

- **Precision always improves under all N_{GO} thresholds**

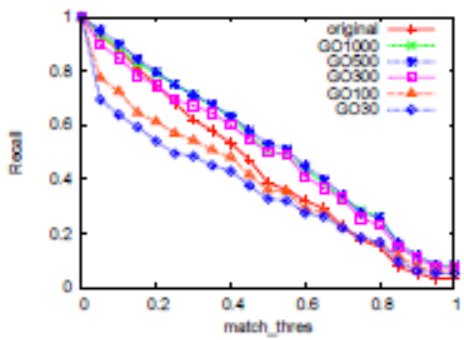
Effect of N_{GO} on Recall



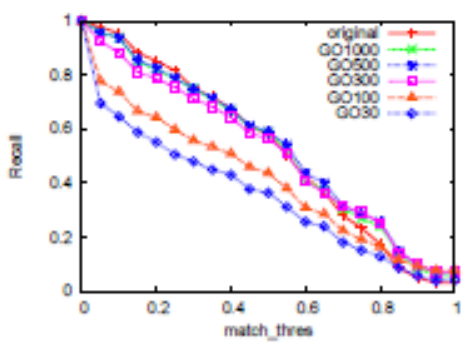
(a) MCL recall



(e) IPCA recall



(c) RNSC recall



(g) CMC recall

- **Recall drops when N_{GO} is small due to excessive info loss**

N_{GO}	#GO terms selected	#proteins discarded	#PPIs discarded
1000	6	2065	27145
500	10	2192	27474
300	10	2481	33425
100	28	3022	39989
30	57	3461	43638

Table 3. Number of GO terms selected under different N_{GO} values.

- **Recall improves when N_{GO} >300**

$\Rightarrow$ **Good to decompose by general CC GO terms**

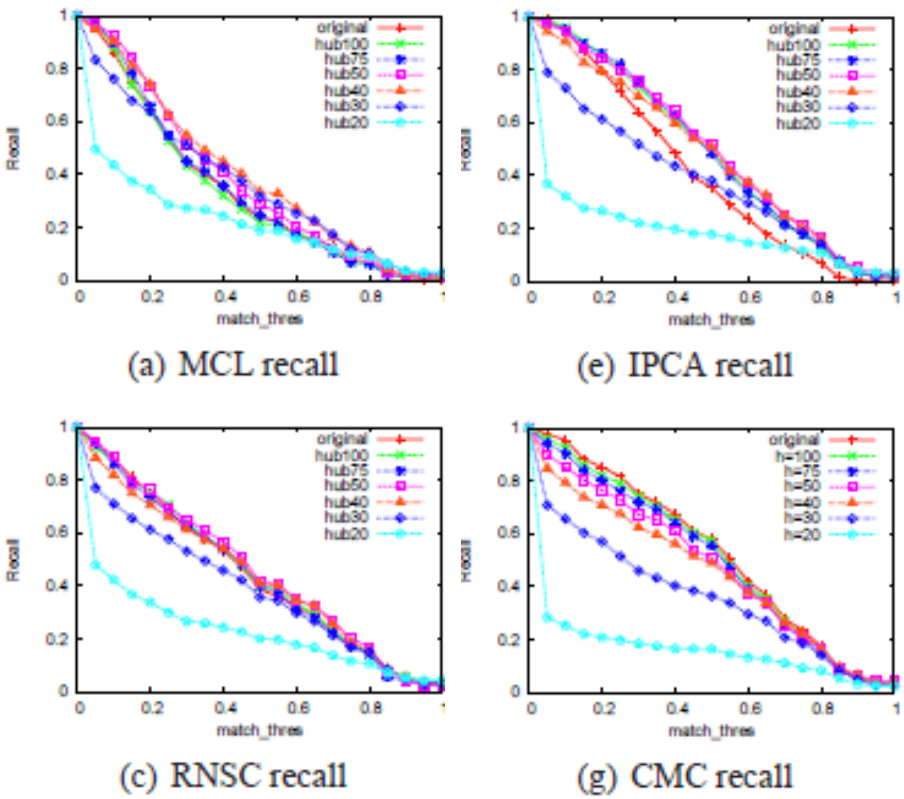
Idea II: Remove Big Hubs

- **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

$\Rightarrow$ **Remove large hubs before protein complex prediction**
  - How to choose threshold N_{hub} to decide whether a hub is "large"?

Effect of N_{hub} on Recall



(a) MCL recall (e) IPCA recall

(c) RNSC recall (g) CMC recall

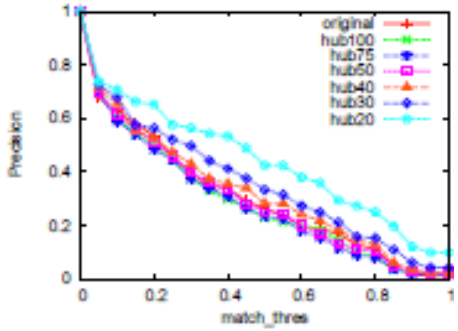
- **Recall is affected when N_{hub} is small, due to high info loss**

N_{hub}	#hub proteins removed	#PPIs removed
100	97	19292
75	207	26331
50	446	35632
40	651	40534
30	996	45568
20	1550	49775

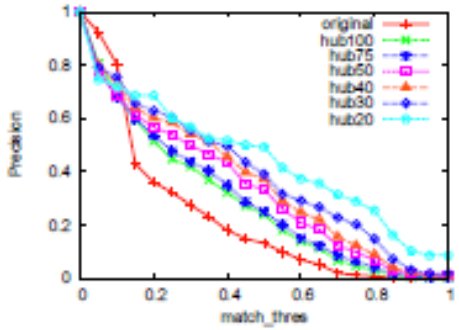
Table 4. Number of hub proteins and PPIs removed under different N_{hub}.

- **Not much effect on recall when N_{hub} is large**

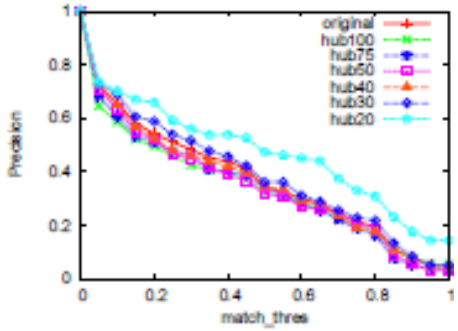
Effect of N_{hub} on Precision



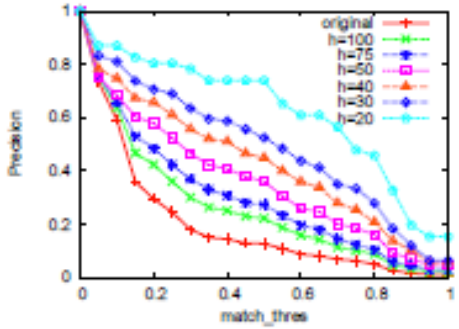
(b) MCL precision



(f) IPCA precision



(d) RNSC precision



(h) CMC precision

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

algorithm	original	hub100	hub75	hub50	hub40	hub30	hub20
MCL	0.623	0.720	0.754	0.796	0.831	0.851	0.919
RNSC	0.847	0.839	0.839	0.846	0.885	0.894	0.928
IPCA	0.640	0.758	0.776	0.853	0.892	0.897	0.906
CMC	0.771	0.835	0.845	0.875	0.898	0.922	0.905

Table 5. Localization coherence score of generated clusters when different N_{hub} values are used for removing hub proteins.

Combining the Two Ideas

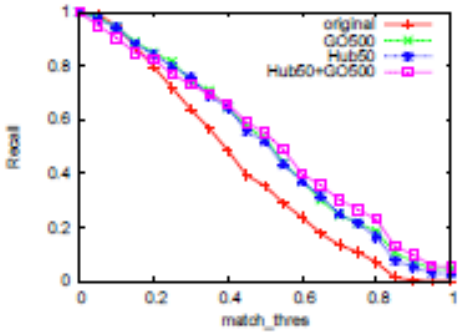
1. Let $\mathcal{C}$ be the set of clusters generated. Initially $\mathcal{C}$ is empty.
2. Remove hub proteins that have at least N_{hub} neighbors from the given PPI network G. Let G' be the resultant network.
3. Let $g_1, \cdots, g_m$ be the localization GO terms that are selected using threshold N_{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}

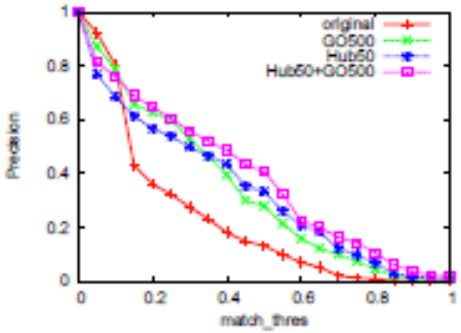
Table 5 - F1-measure of the four algorithms when *match_thres*=0.5

	original	Hub50	GO500	Hub50+GO500
MCL	0.250	0.272	0.354	0.406
RNSC	0.353	0.347	0.471	0.436
IPCA	0.191	0.405	0.368	0.469
CMC	0.207	0.421	0.359	0.501

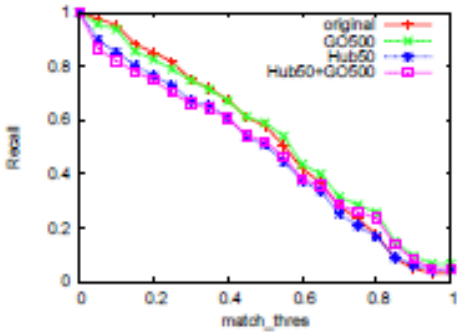
- **RNSC doesn't benefit further**
- **MCL, IPCA & CMC all gain further**



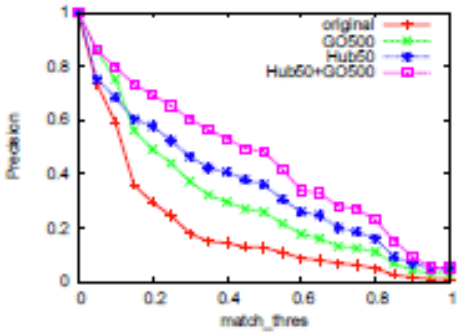
(e) IPCA recall



(f) IPCA precision



(g) CMC recall



(h) CMC precision

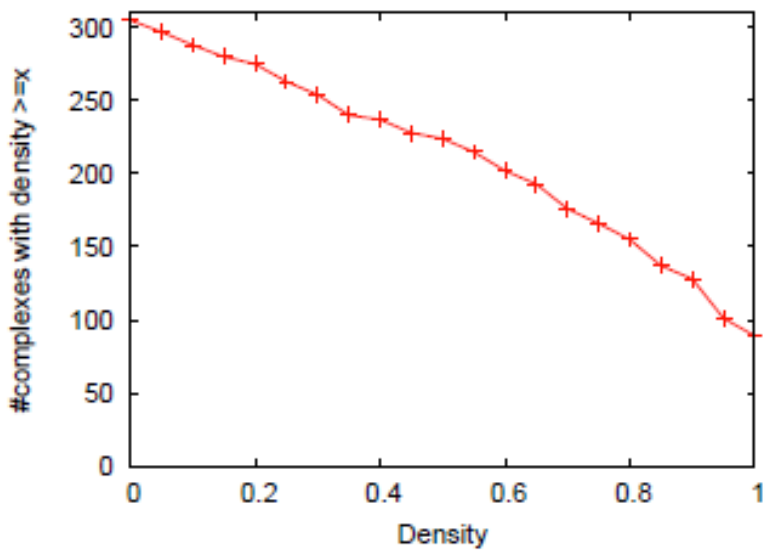
Conclusions

Table 5 - F1-measure of the four algorithms when $match\_thres$=0.5

	original	Hub50	GO500	Hub50+GO500
MCL	0.250	0.272	0.354	0.406
RNSC	0.353	0.347	0.471	0.436
IPCA	0.191	0.405	0.368	0.469
CMC	0.207	0.421	0.359	0.501

- **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…**

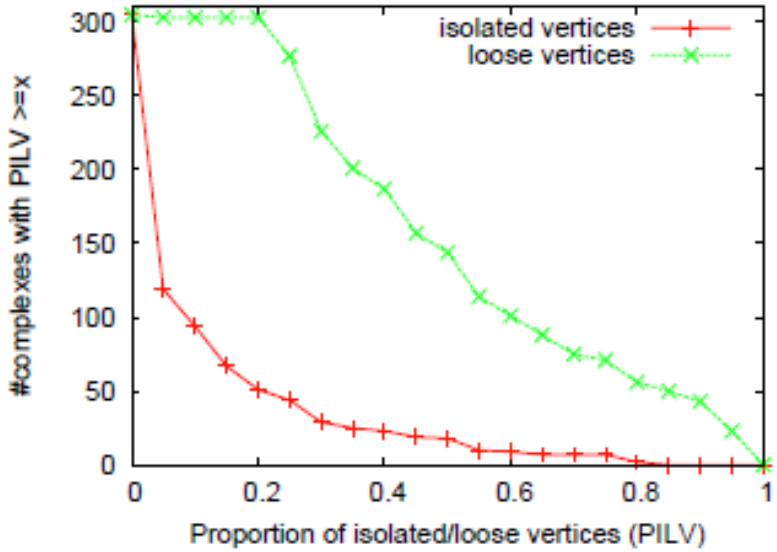
Why many complexes are not detectable



(a) density

- **Among 305 complexes, 81 have density < 0.5, and 42 have density < 0.25**

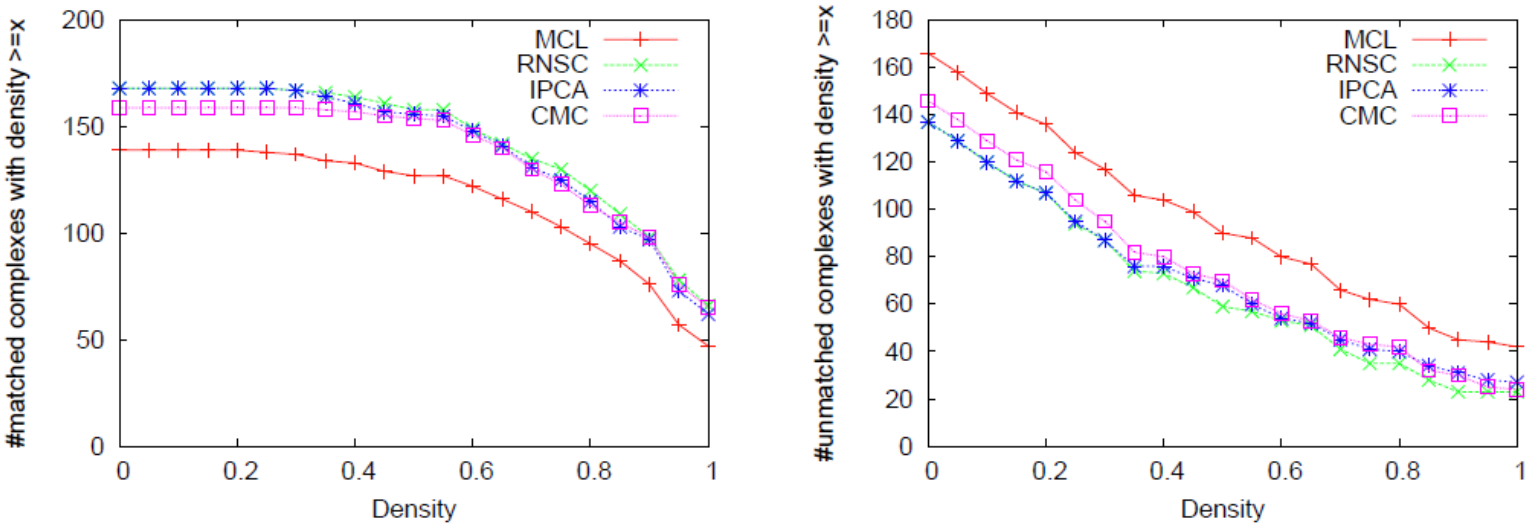
Why many complexes are not detectable



(b) connectivity

- **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

Why many complexes are not detectable



(a) detected complexes (b) undetected complexes

- **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

NUS
National University of Singapore

Sparse Complexes

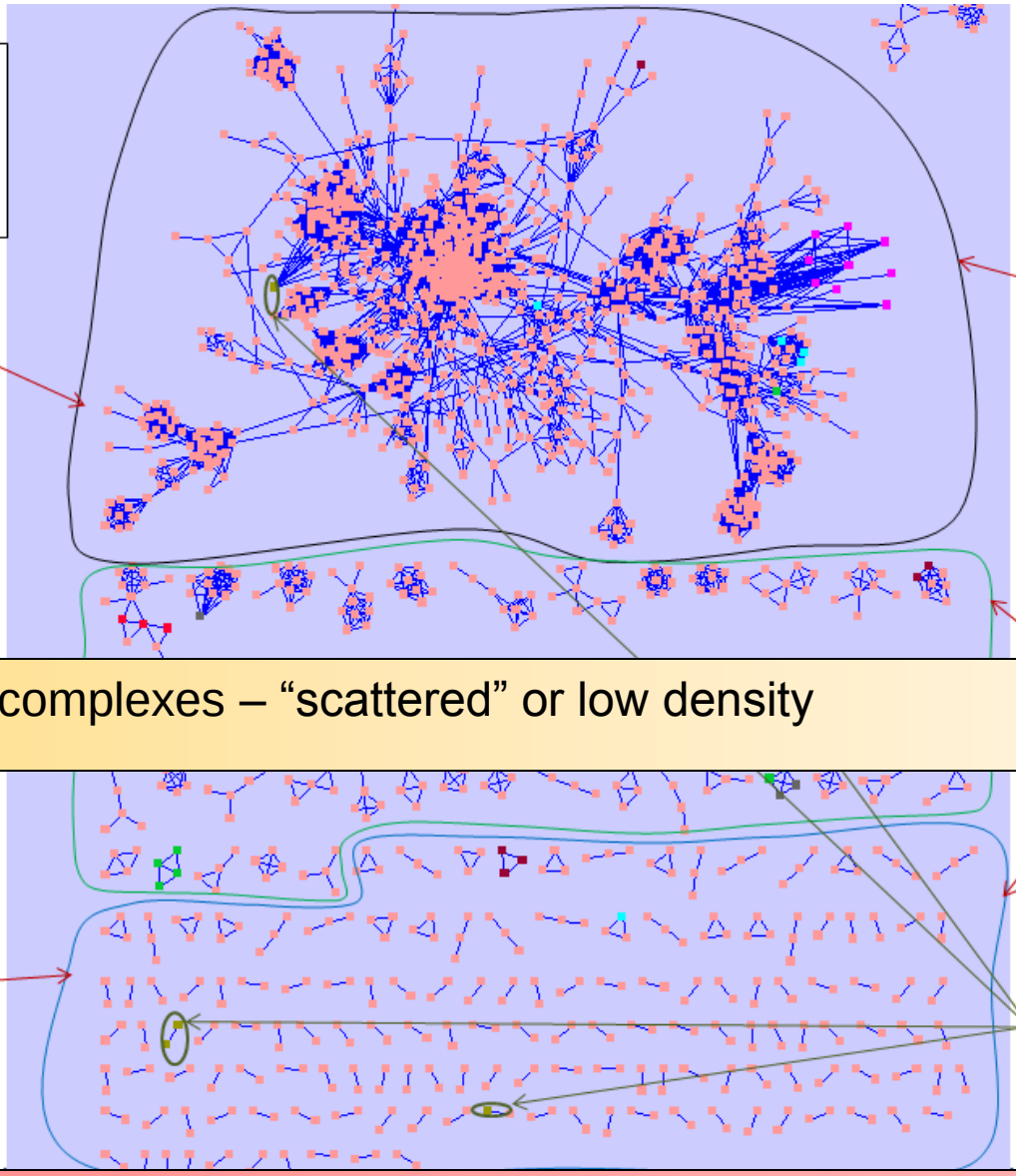
The Consolidated (3.19) network (Collins et al., 2007); #proteins: 1622; #interactions: 9704

Main large component

#prot: 1034; #int: 8377

Total MIPS complexes in the network: **123**

#MIPS complexes in main large component: **89**



~ 25% sparse complexes – "scattered" or low density

#MIPS complexes "scattered" in medium and small components : **34**

#prot: 588; #int: 696

Small components (size 2 to 3): **140**

"Scattered" MIPS complex 510.190.110 (CCR4 complex)

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

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.**

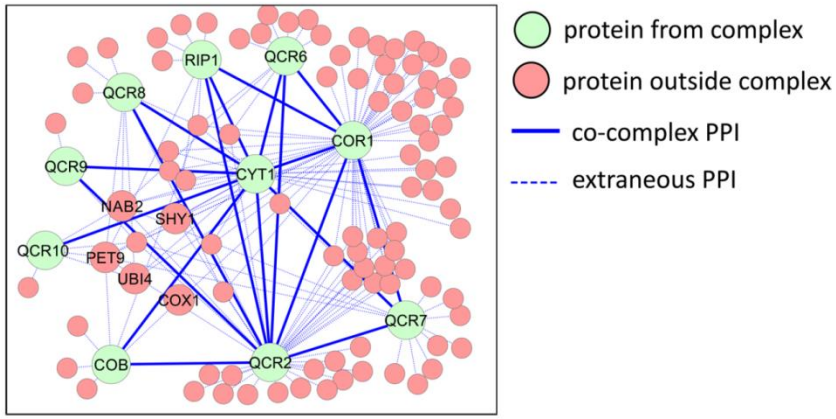


Figure 1 PPI subgraph of the mitochondrial cytochrome bc1 complex. Nineteen interactions were detected between the ten proteins from the complex, while many extraneous interactions were detected. Five example proteins from transient interactions are shown: NAB2 and UBI4 are involved in mRNA polyadenylation and protein ubiquitination, while PET9, SHY1, and COX1 are mitochondrial membrane proteins that are also involved in the electron-transport chain. The extraneous interactions around the complex makes its discovery difficult. All such network figures were generated by Cytoscape [30].

Yong et al. "Supervised maximum-likelihood weighting of composite protein networks for complex prediction". *BMC Systems Biology*, 6(Suppl 2):S13, 2012

- **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 Sources

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

Data source	Database	Scoring method
PPI	BioGRID, IntACT, MINT	Iterative AdjustCD.
L2-PPI (indirect PPI)	BioGRID, IntACT, MINT	Iterative AdjustCD
Functional association	STRING	STRING
Literature co-occurrence	PubMed	Jaccard coefficient

	Yeast			Human		
	# Pairs	% co-complex	coverage	# Pairs	% co-complex	coverage
PPI	106328	**5.8**%	**55**%	48098	10%	14%
L2-PPI	181175	1.1%	18%	131705	5.5%	20%
STRING	175712	5.7%	89%	311435	3.1%	27%
PubMed	161213	4.9%	70%	91751	4.3%	11%
All	531800	**2.1**%	**98**%	522668	3.4%	49%

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"**

PPI	L2 PPI	STRING	Pubmed	Class
0	0.56	451	0	"co-complex"
0.1	0	25	0	"non-co-complex"

- Supervised learning:
  1. Discretize each feature (Minimum Description Length discretization[7])
  2. Learn maximum-likelihood parameters for the two classes:

$$P(F = f|co-comp) = \frac{n_{c,F=f}}{n_c} \qquad P(F = f|non-co-comp) = \frac{n_{\neg c,F=f}}{n_{\neg c}}$$

  for each discretized feature value f of each feature F

- Weight each edge e with its posterior probability of being co-complex:

$$\begin{aligned} & weight(e) \\ &= P(co-comp|F_1 = f_1, F_2 = f_2, \ldots) \\ &= \frac{P(F_1 = f_1, F_2 = f_2, \ldots|co-comp)P(co-comp)}{Z} \\ &= \frac{\prod_i P(F_i = f_i|co-comp)P(co-comp)}{Z} \\ &= \frac{\prod_i P(F_i = f_i|co-comp)P(co-comp)}{\prod_i P(F_i = f_i|co-comp)P(co-comp) + \prod_i P(F_i = f_i|non-co-comp)P(non-co-comp)} \end{aligned}$$

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

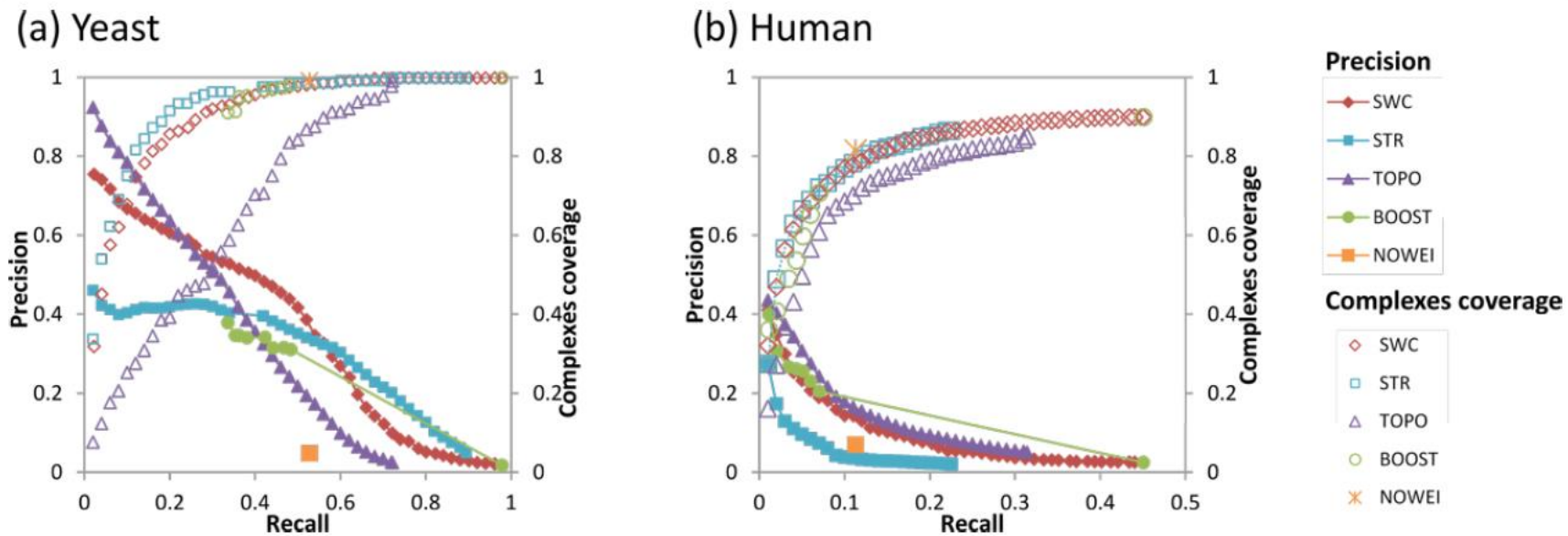
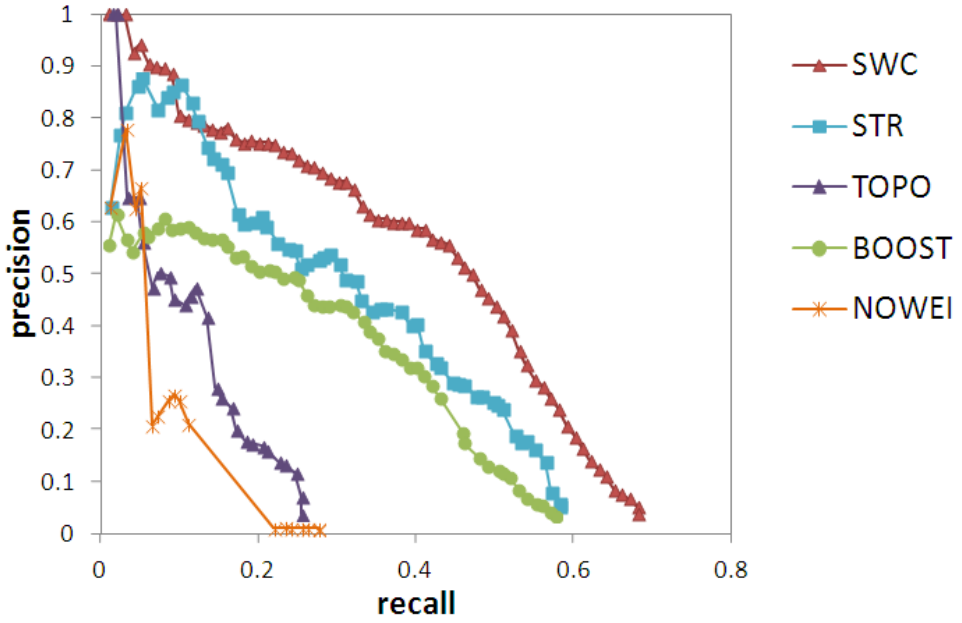
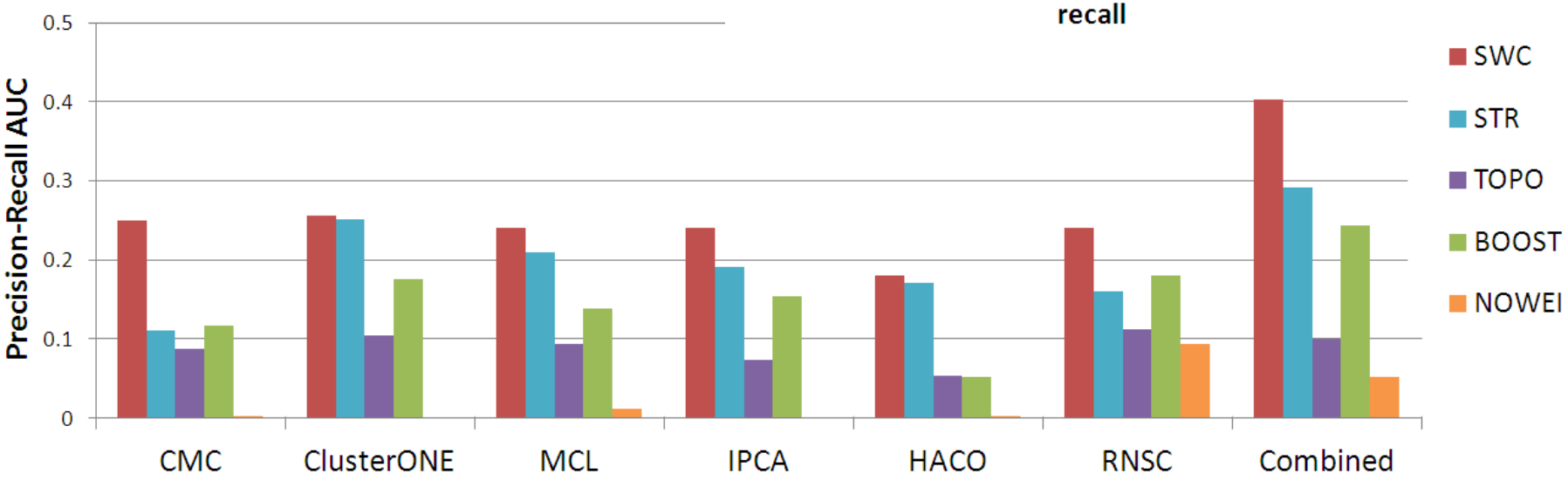


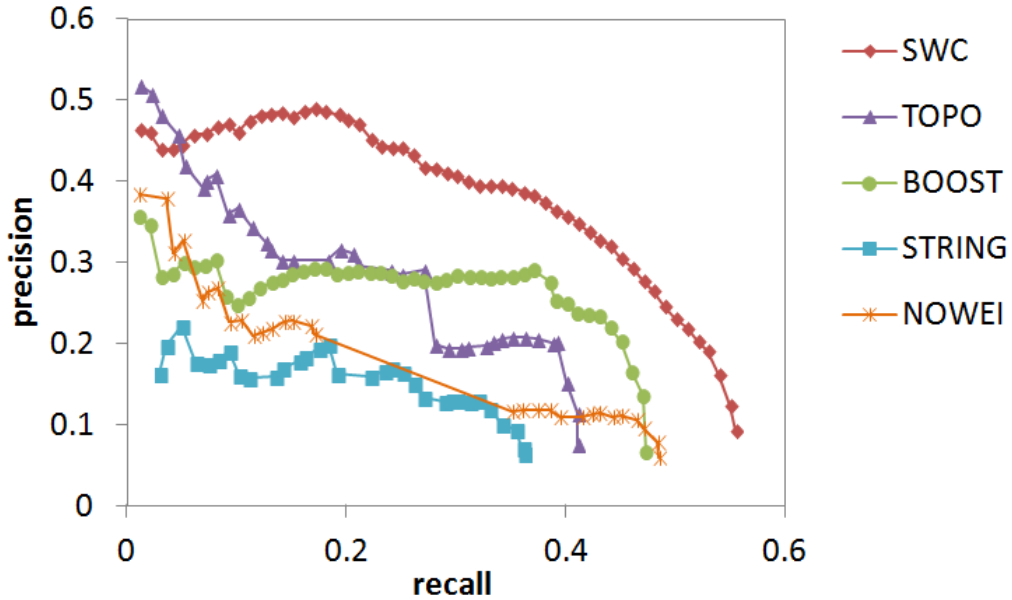
Figure 2 Precision-recall graph for classification of co-complex edges using the five weighting schemes. (a) Classification of yeast co-complex edges. SWC and BOOST achieve the highest recall through data integration. TOPO has high precision for its top-scoring edges, but these are clustered in a few 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.

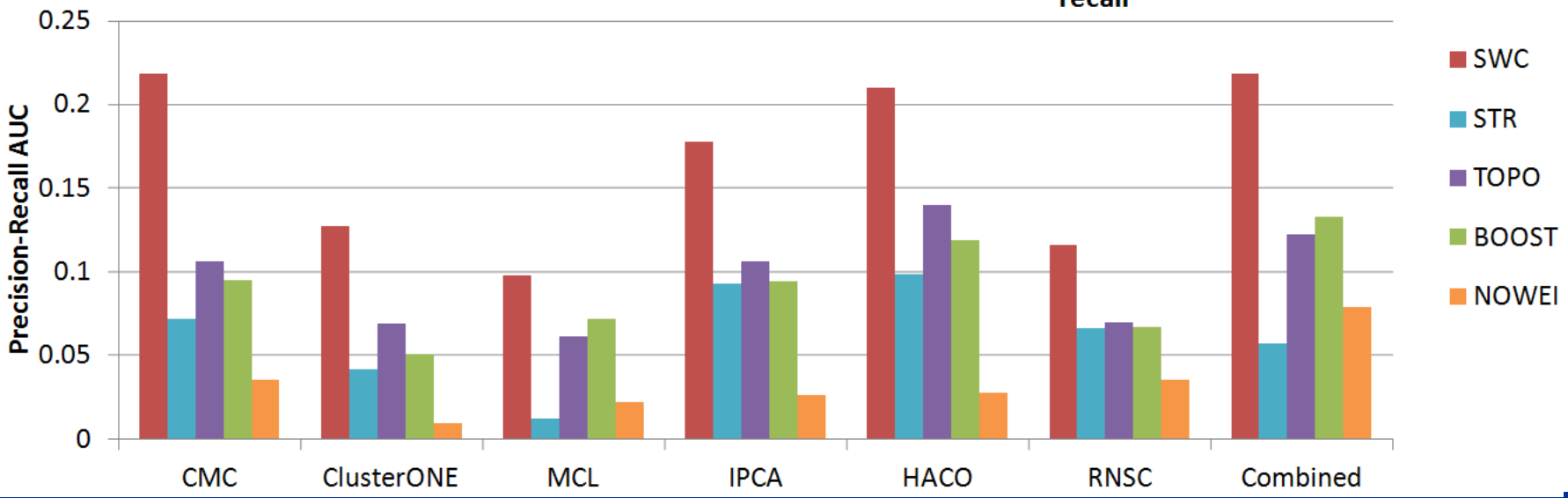
Evaluation wrt Yeast Complex Prediction



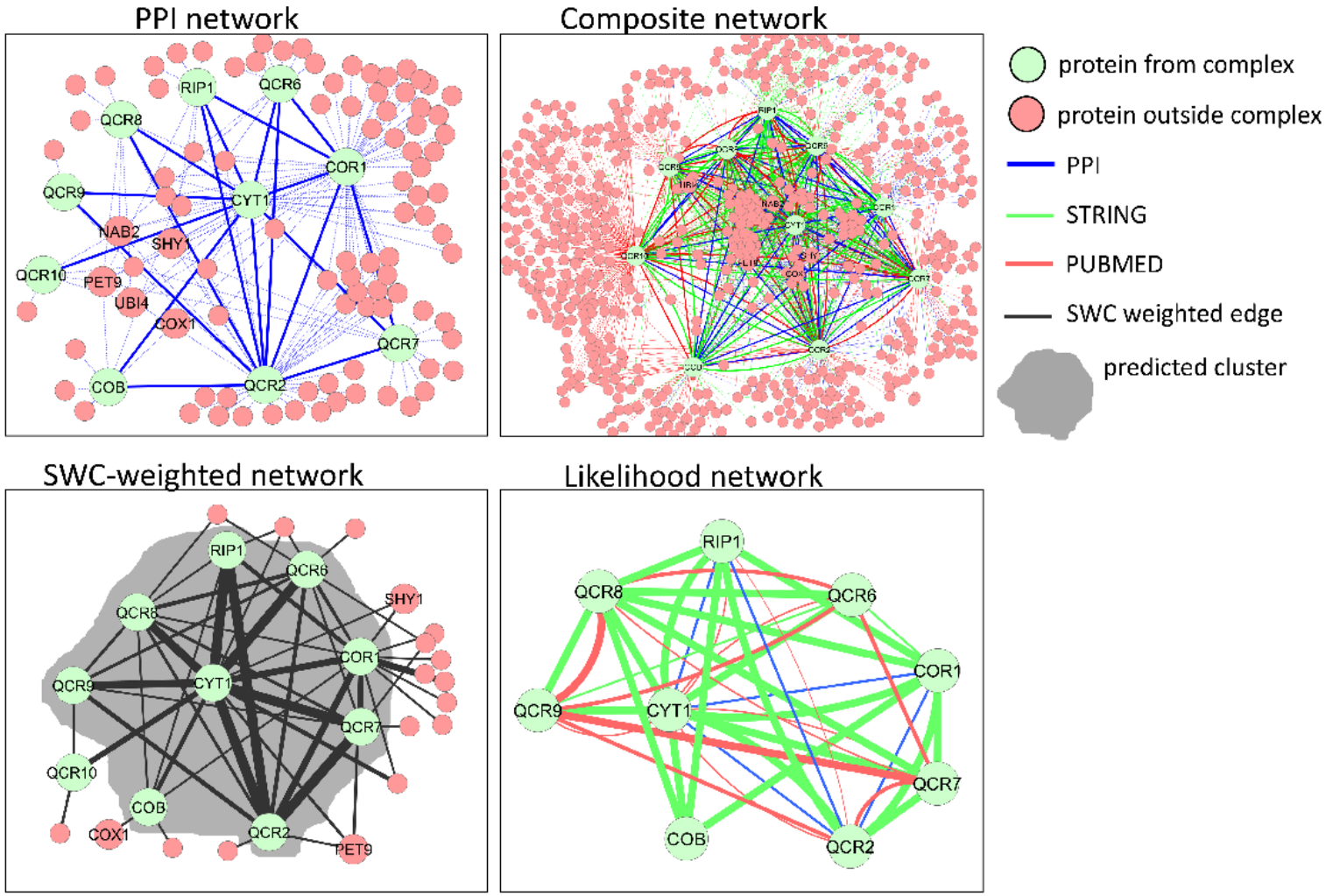


Evaluation wrt Human Complex Prediction

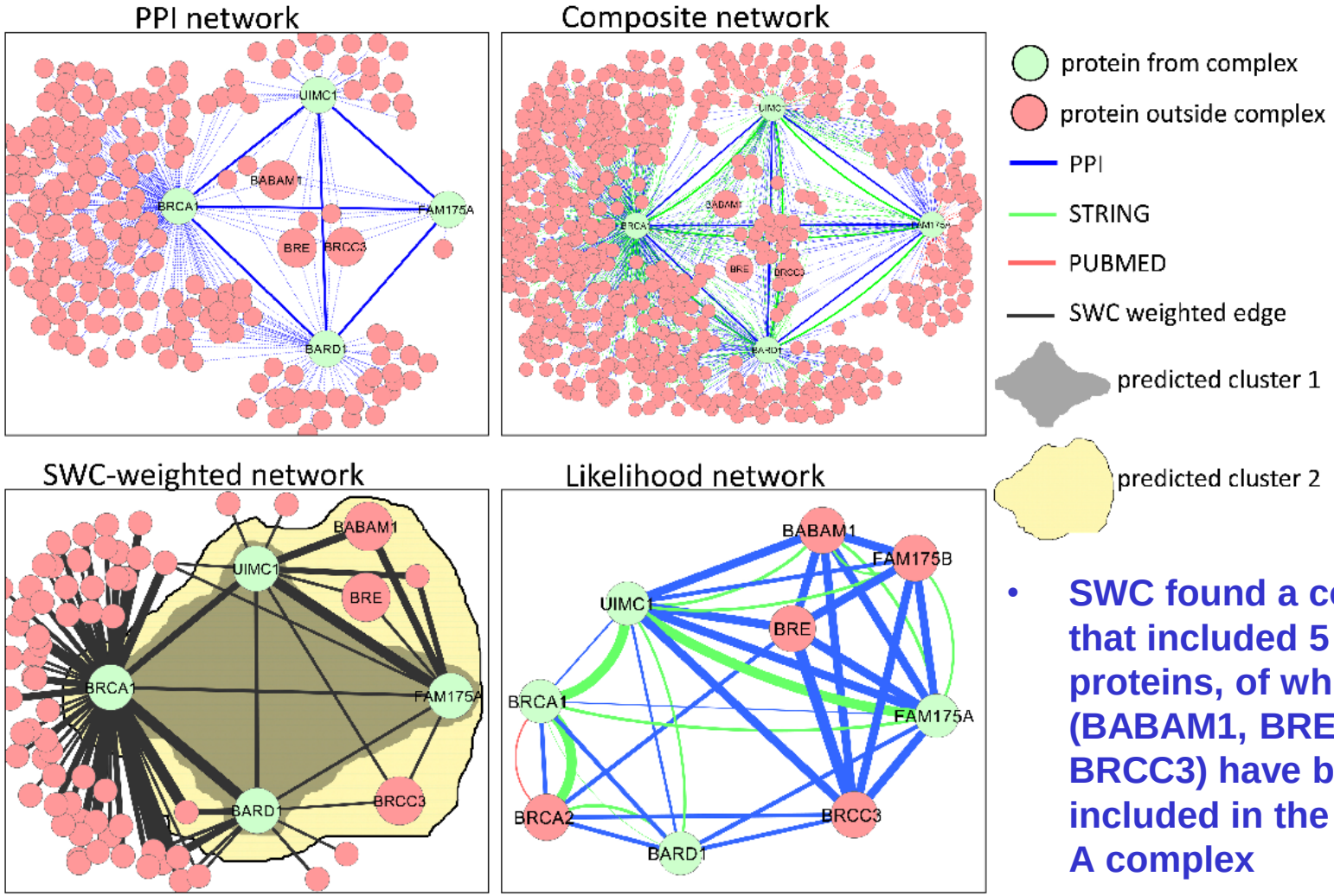




Yeast BC1 Complex



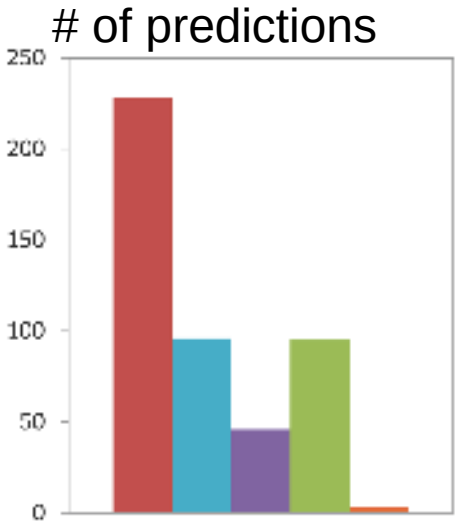
Human BRCA1-A complex

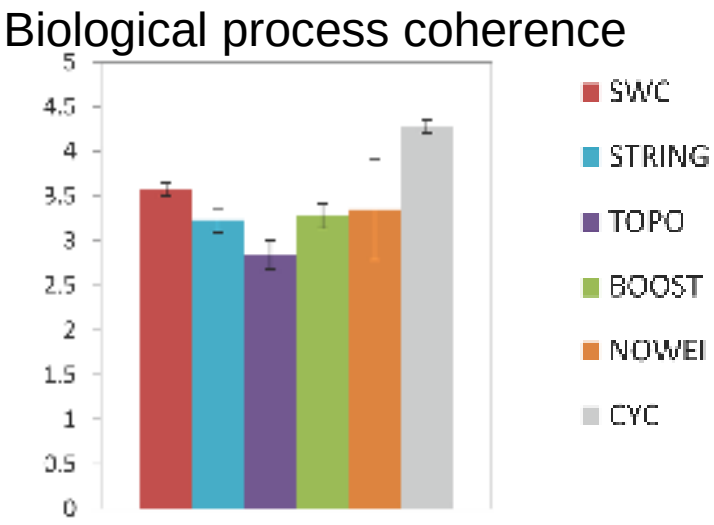


- **SWC found a complex that included 5 extra proteins, of which 3 (BABAM1, BRE, BRCC3) have been included in the BRCA1-A complex**

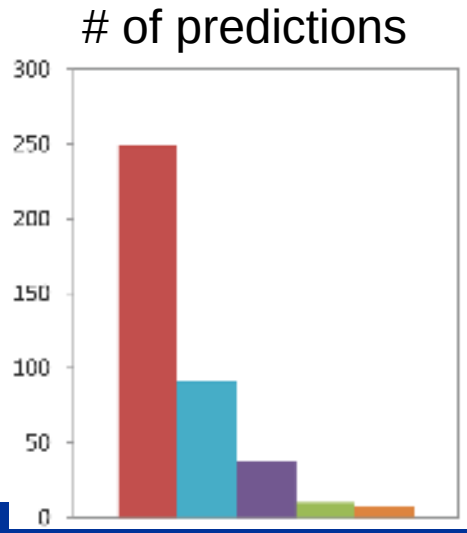
High-Confidence Predicted Complexes

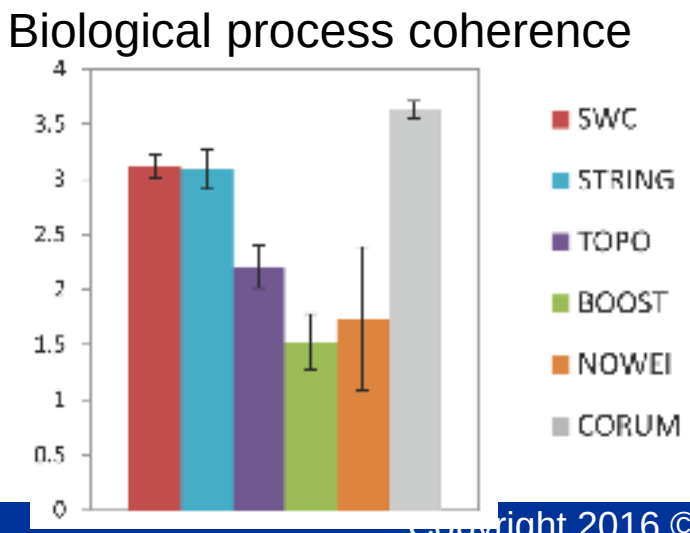
Yeast



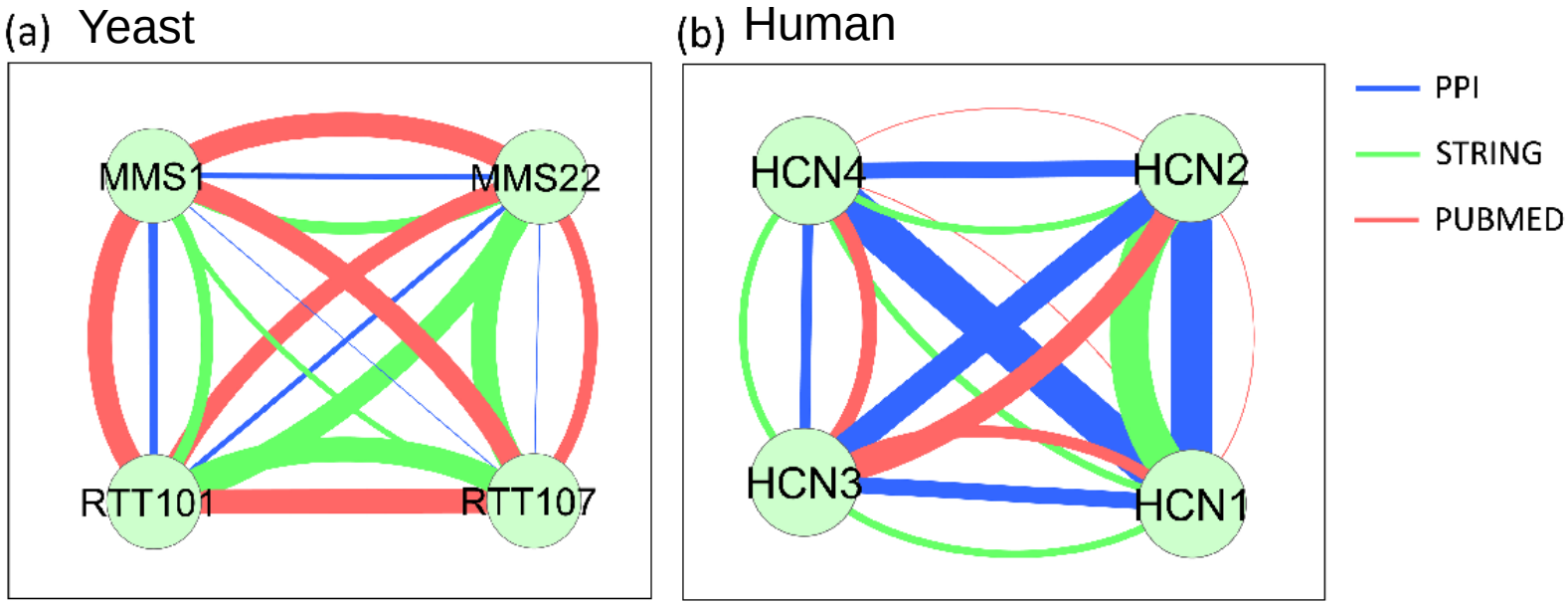


Human





Two Novel Predicted Complexes



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

Novel Complexes Predicted

Yeast

Biological process	# complexes
Protein metabolic process	49
RNA metabolic process	36
DNA metabolic process	15
Small molecule metabolic process	23
Regulation of metabolic process	11
Regulation of gene expression	8
Organelle organization	40
Transport	43
Response to stress	20
Response to chemical stimulus	7
Cell cycle process	11

Human

Biological process	# complexes
Protein metabolic process	32
RNA metabolic process	29
DNA metabolic process	4
Small molecule metabolic process	19
Regulation of metabolic process	74
Regulation of gene expression	34
Organelle organization	19
Transport	38
Response to stress	28
Response to chemical stimulus	32
Cell cycle process	14

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

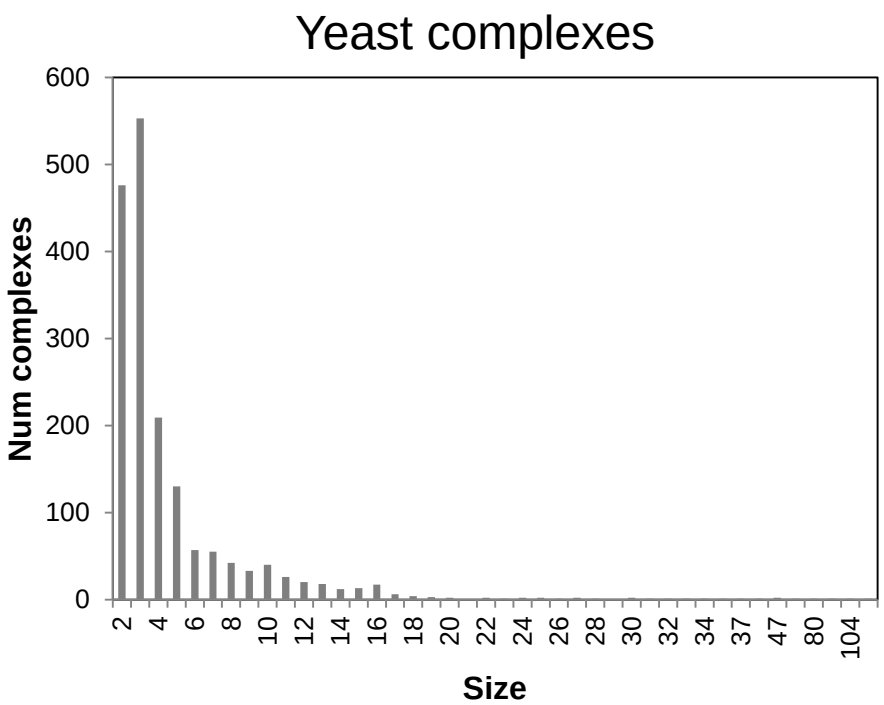
- **Tested approach using 6 clustering algo's**
  - Clusters produced by diff algo's have low overlap, combining them gives greater recall
  - Clusters produced by more algo's are more reliable

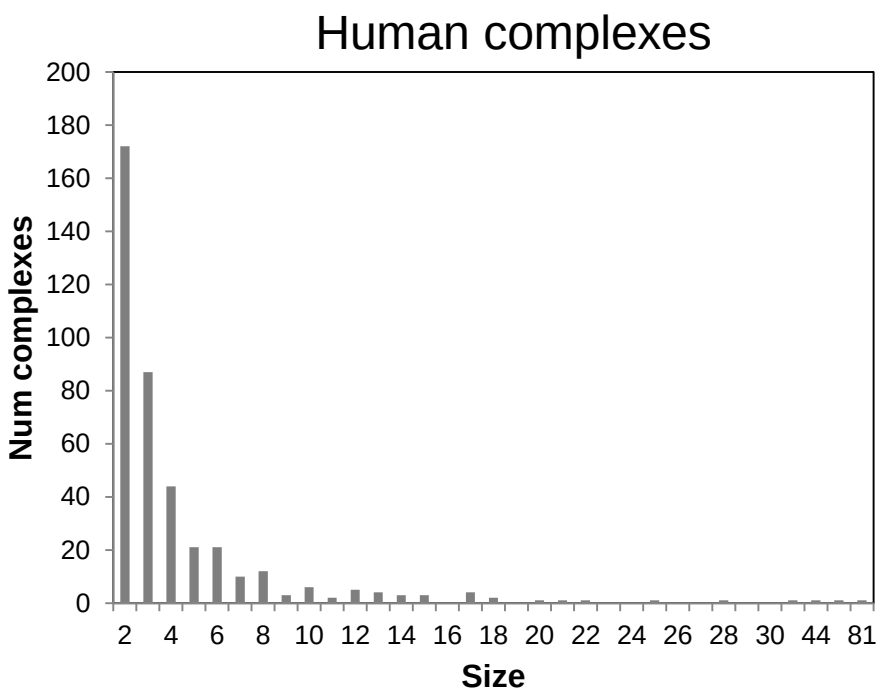
Detecting Small Protein Complexes

NUS
National University
of Singapore

Motivation

- **Size of protein complexes follows a power-law distribution, meaning that most complexes are small (ie. 2 or 3 distinct proteins)**





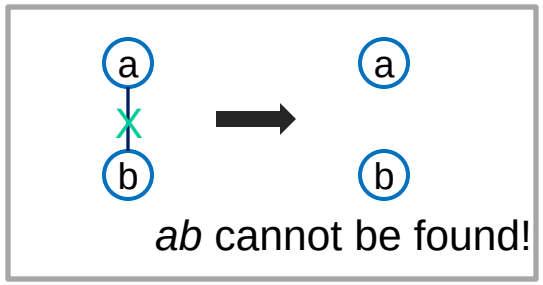
Small Complexes, Big Challenges

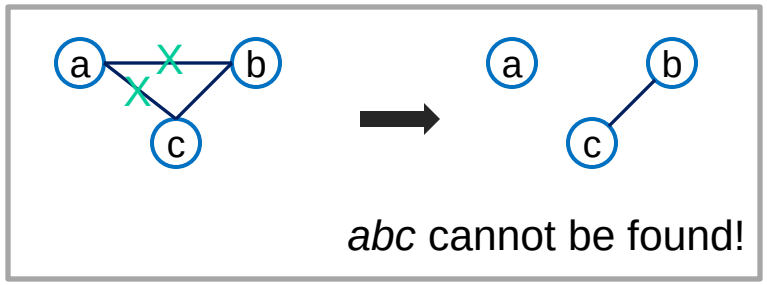
- **Traditionally, complexes are predicted by searching for dense clusters in a PPI network**

- **For small complexes, topological characteristics like density are problematic**
  - A fully-dense size-2 complex is an edge
  - A fully-dense size-3 complex is a triangle
  - But there are many edges and triangles in the PPI network that are not complexes

Small Complexes, Big Challenges

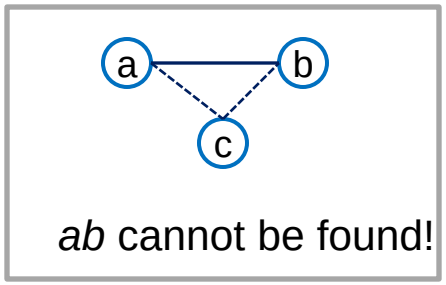
- **Sensitive to missing edges**
  - One missing edge disconnects a size-2 complex
  - Two missing edges disconnect a size-3 complex

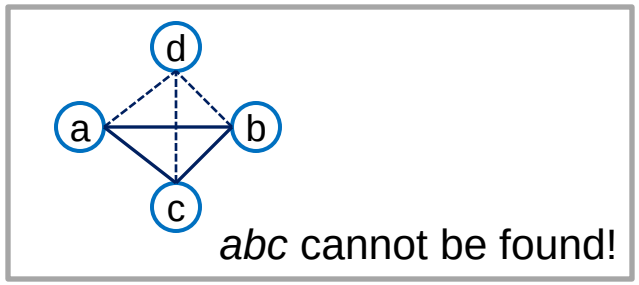




Small Complexes, Big Challenges

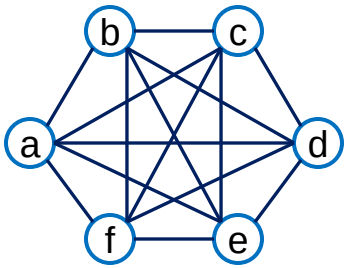
- **Sensitive to extraneous edges**
  - Two extraneous edges embed a size-2 complex in a size-3 clique
  - Three extraneous edges embed a size-3 complex in a size-4 clique





Small Complexes, Big Challenges

- **Predicted complexes are scored using their internal weights to give them some reliability measure, eg. using weighted density. This reliability is averaged out over the internal weights of the candidate complex**



Size-6 complex: Score is averaged over 15 edge weights

- **Scores of small complexes are sensitive to the correct edge weights, since only one or three edges weights are used**

Size-2 complex: Score depends on just 1 edge weight. It is very sensitive to its value

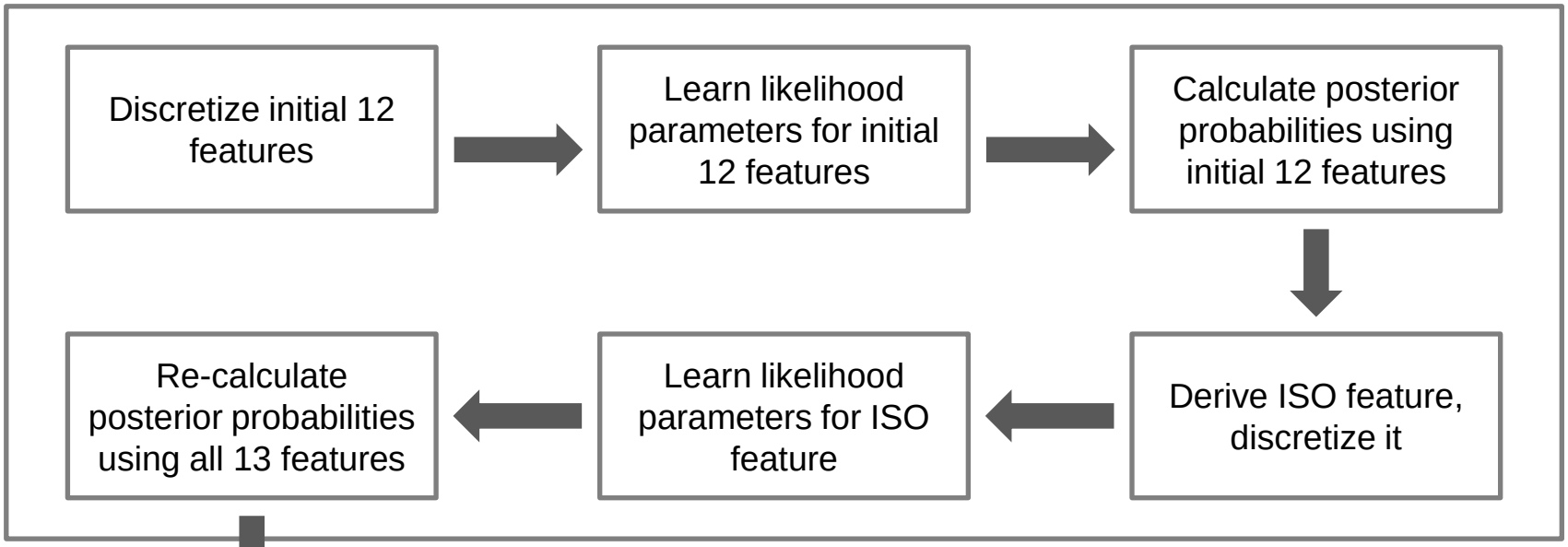
Small Complexes, Big Challenges

- **Previously used data integration and supervised learning successfully for predicting large complexes (SWC2)**

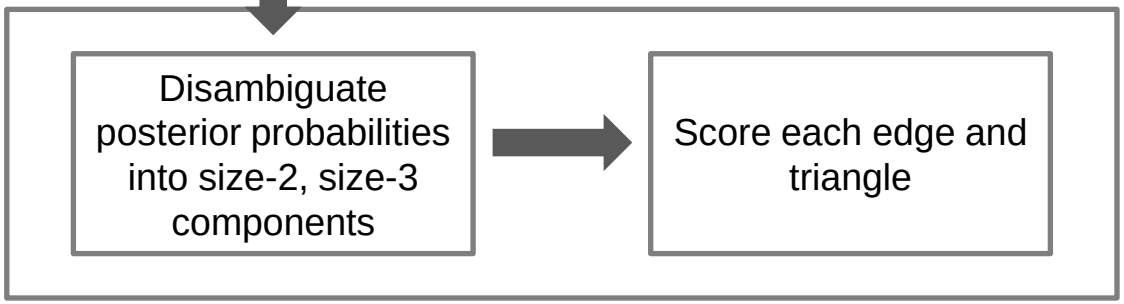
- **It does not work well for small complexes**
  - Small complexes have different topological features compared to large complexes
  - Learned model corresponds to large complexes, not small complexes, as large complexes have much more edges

Two-Stage Approach

1. Size-specific supervised weighting (SSS)



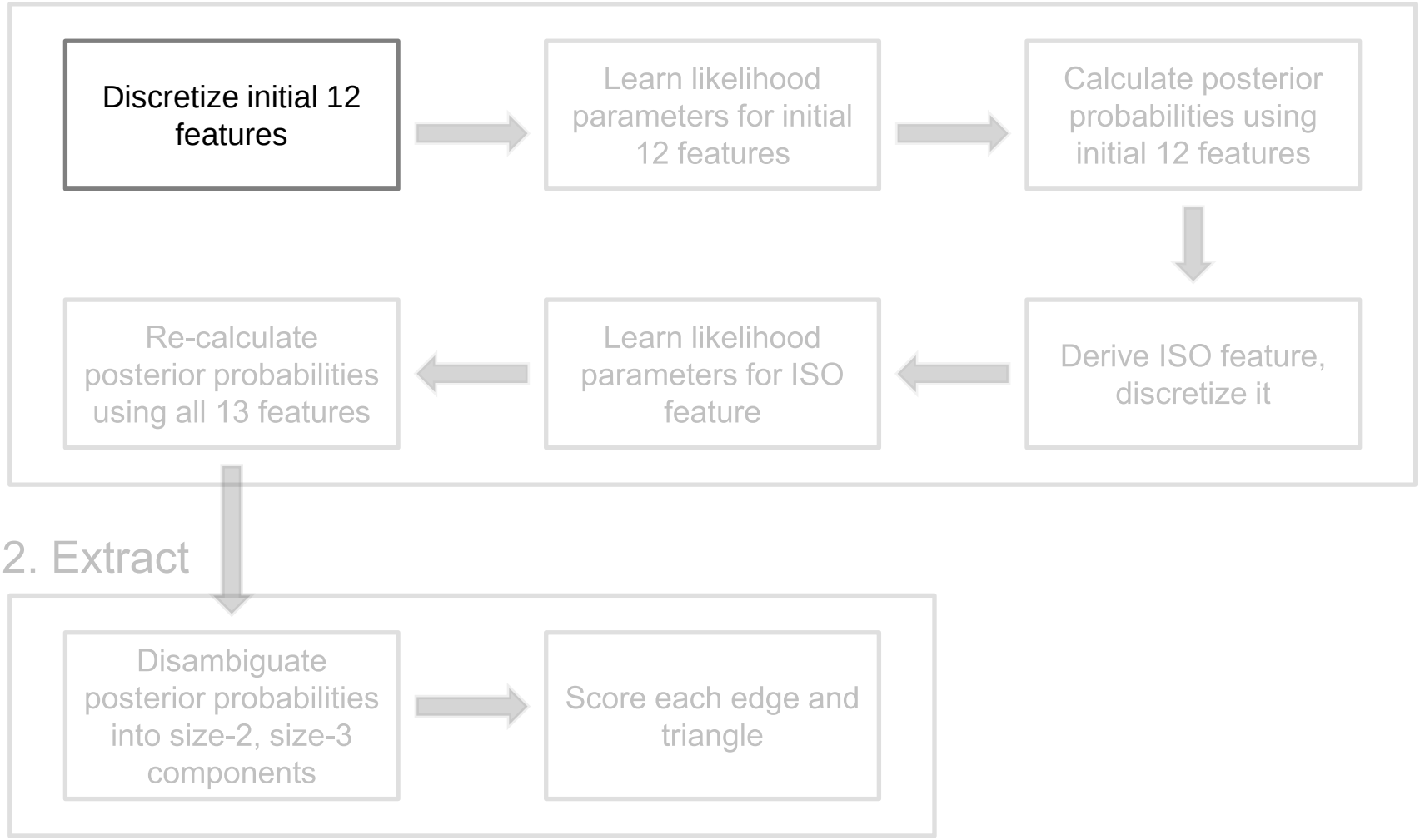
2. Extract



Yong et al., "Discovery of small protein complexes from PPI networks with size-specific supervised weighting". *BMC Systems Biology*, 8(Suppl 5):S3, 2014

Stage 1: SSS

1. Size-specific supervised weighting (SSS)



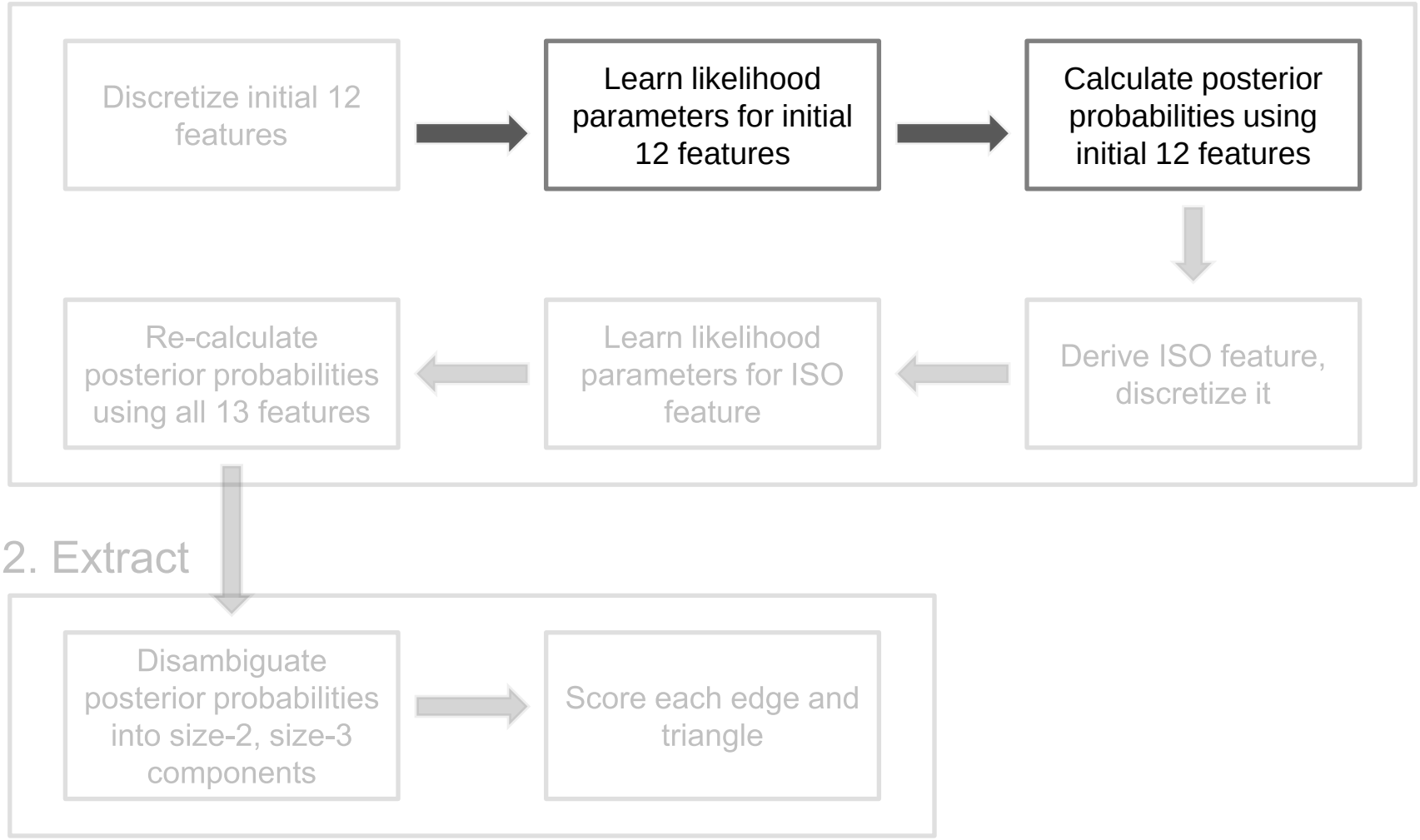
Discretize initial 12 features

- **Each edge in PPIN is cast as a data instance, with 12 initial features**
  - 3 data sources
    - **PPI (BioGrid + IntAct + MINT)**
    - **Functional associations (STRING)**
    - **Co-occurrence in literature (PUBMED)**
  - 3 topological characteristics for each data source
    - **Degree**
    - **Neighbourhood connectivity**
    - **Shared neighbours**

- **Discretize based on Minimum Description Length (MDL)**

Stage 1: SSS

1. Size-specific supervised weighting (SSS)



Learn likelihood parameters for initial 12 features

- **Likelihood models for 3 classes (small co-complex, large co-complex, non co-complex)**

$$P(F = f|sm\text{-}comp) = \frac{n_{sm,F=f}}{n_{sm}}$$

$$P(F = f|lg\text{-}comp) = \frac{n_{lg,F=f}}{n_{lg}}$$

$$P(F = f|non\text{-}comp) = \frac{n_{non,F=f}}{n_{non}}$$

Calculate posterior probabilities using initial 12 features

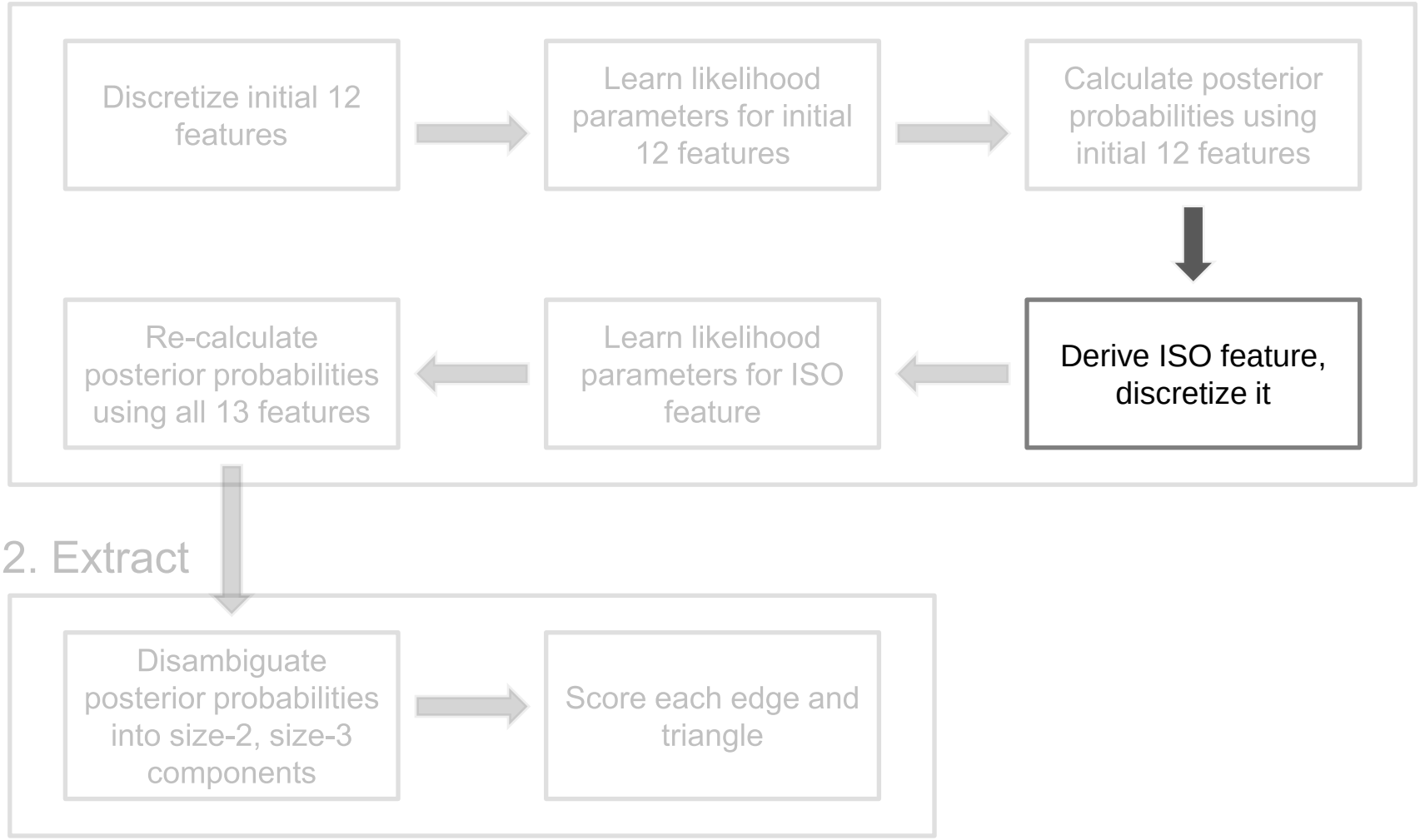
- **Weight each edge with its posterior probability of being small co-complex, large co-complex, or non co-complex, using the naïve-Bayes formulation**
  - Eg., probability that edge (a,b) is small co-complex

$$P((a,b)\ is\ sm\text{-}comp|F_1 = f_1, F_2 = f_2, \ldots) = \frac{\prod_i P(F_i = f_i|(a,b)\ is\ sm\text{-}comp)P(sm\text{-}comp)}{\sum_{class \in \{sm\text{-}comp, lg\text{-}comp, non\text{-}comp\}} \prod_i P(F_i = f_i|(a,b)\ is\ class)P(class)}$$

- **These three probabilities are abbreviated as**
  - $P_{(a,b),sm}$
  - $P_{(a,b),lg}$
  - $P_{(a,b),non}$

Stage 1: SSS

1. Size-specific supervised weighting (SSS)



Derive ISO feature

- **For each edge, derive a new feature, Isolatedness**
  - Prob that the edge is isolated, or is part of an isolated triangle
  - Uses posterior prob calculated previously

$$ISO(a,b) = ISO2(a,b) + ISO3(a,b)$$

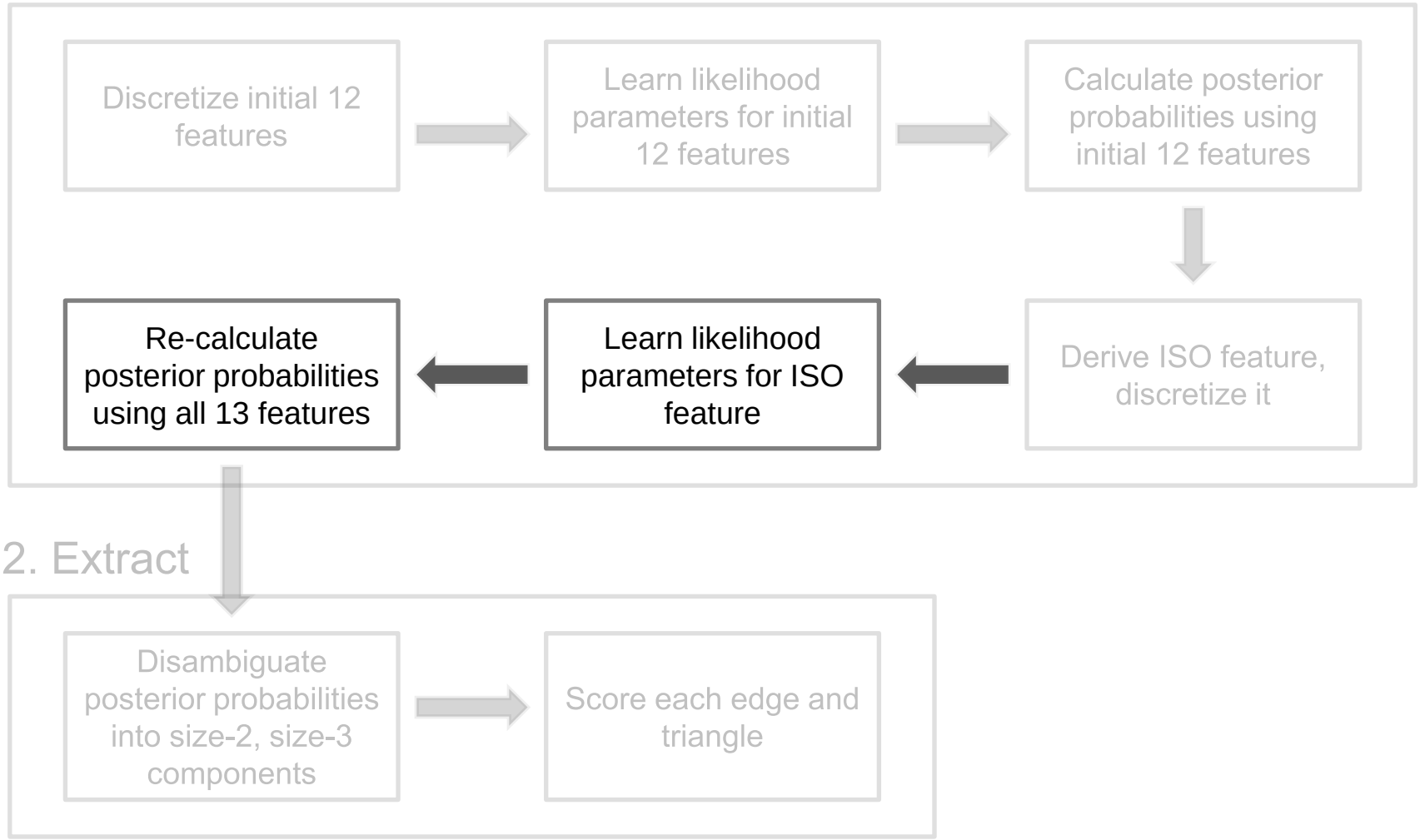
$$ISO2(a,b) = P_{(a,b),sm} \prod_{x \in \{a,b\}, y \in N_{a,b}} P_{(x,y),non}$$

$$ISO3(a,b) = \sum_{c \in N_a \cap N_b} \left(P_{(a,b),sm} P_{(a,c),sm} P_{(b,c),sm} \prod_{x \in \{a,b,c\}, y \in N_{a,b,c}} P_{(x,y),non} \right)$$

- **This feature is also discretized using MDL**

Stage 1: SSS

1. Size-specific supervised weighting (SSS)



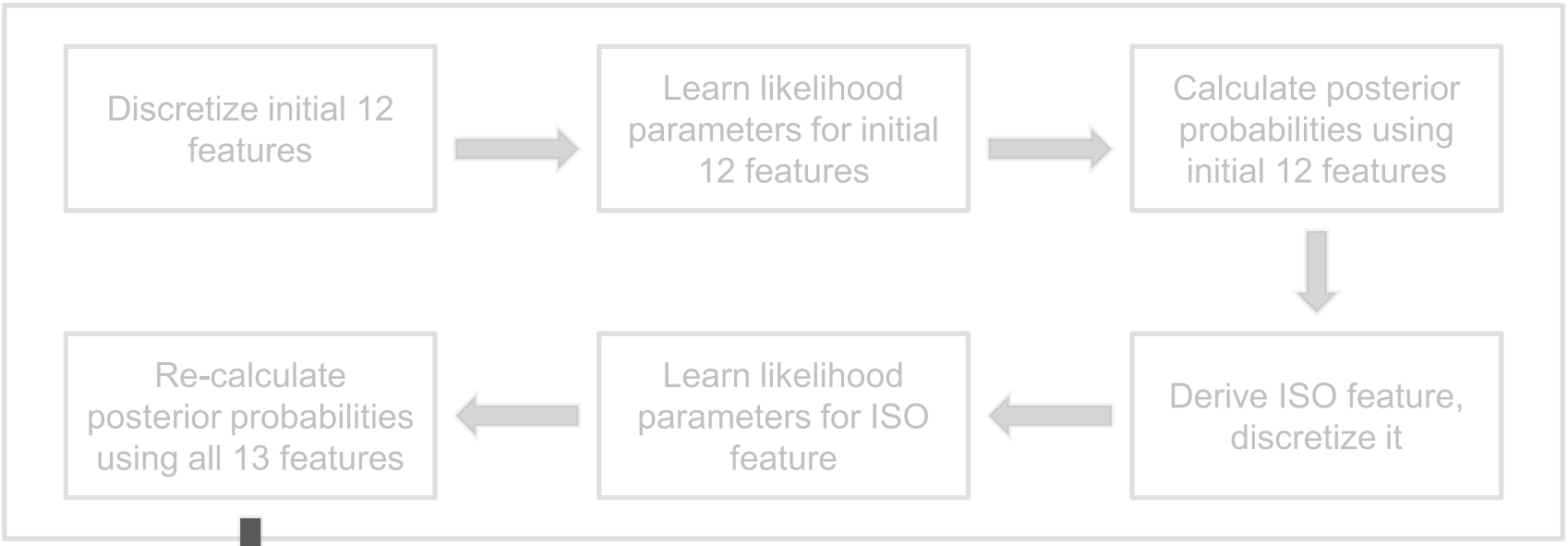
Learn likelihood parameters for ISO feature & Recalculate posterior prob using all 13 features

- **Likelihood parameters are learned for the ISO feature in the same way as with the previous features**

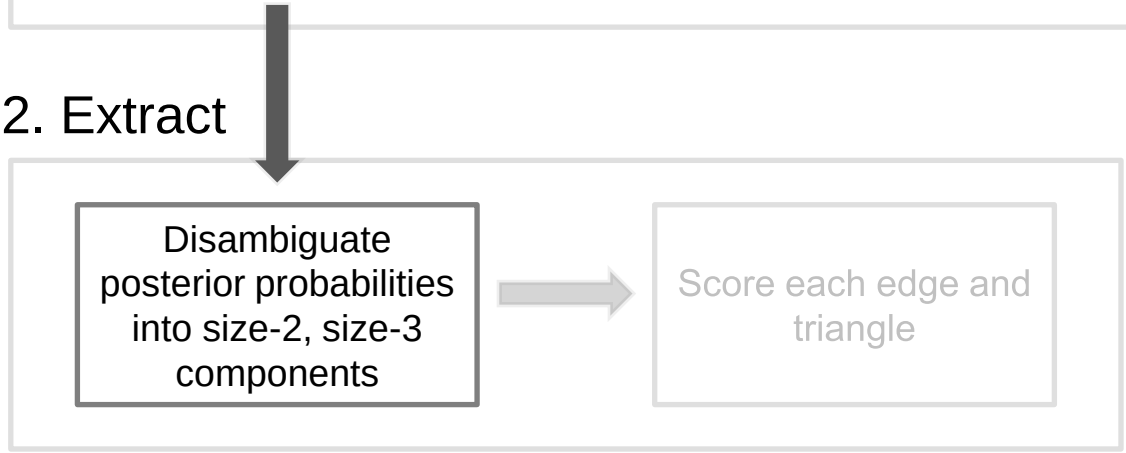
- **Posterior prob are re-calculated as before, this time incorporating the new ISO feature**
  - P(a,b),sm = prob that (a,b) is small co-complex
  - P(a,b),lg = prob that (a,b) is large co-complex
  - P(a,b),non = prob that (a,b) is non co-complex

Stage 2: Extract

1. Size-specific supervised weighting (SSS)



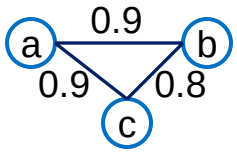
2. Extract



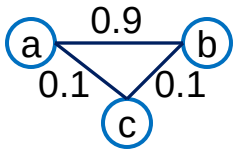
Disambiguate $P_{(a,b),sm}$, the prob that (a,b) is small co-complex, into size-2 and size-3 components

- **If (a,b) is part of a high-weighted triangle, then it is likelier to be part of a size-3 complex, so reduce its size-2 component**

$$P'_{(a,b),sm2} = P_{(a,b),sm} - \sum_{x \in N_a \cap N_b} P_{(a,b),sm} P_{(a,x),sm} P_{(b,x),sm}$$



(a,b) likelier to be part of a size-3 complex *abc* than a size-2 complex *ab*

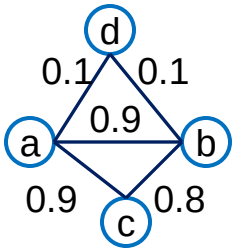


(a,b) likelier to be a size-2 complex than size-3 complex *abc*

Disambiguate $P_{(a,b),sm}$, the prob that (a,b) is small co-complex, into size-2 and size-3 components

- **If (a,b) is part of a high-weighted triangle, and is part of another low-weighted triangle, then it is likelier to be in a complex with the first triangle**

$$P'_{(a,b),sm3,abc} = P_{(a,b),sm} - \sum_{x \in N_a \cap N_b \setminus \{c\}} P_{(a,b),sm} P_{(a,x),sm} P_{(b,x),sm}$$



(a,b) likelier to be part of a size-3 complex *abc*, than complex *abd*

Stage 2: Extract

1. Size-specific supervised weighting (SSS)

Discretize initial 12 features → Learn likelihood parameters for initial 12 features → Calculate posterior probabilities using initial 12 features → Derive ISO feature, discretize it → Learn likelihood parameters for ISO feature → Re-calculate posterior probabilities using all 13 features

2. Extract

Disambiguate posterior probabilities into size-2, size-3 components → Score each edge and triangle

Score each edge and triangle

- **Every edge / triangle is taken as candidate size-2 / -3 complexes**

- **Score each candidate complex, using edges inside the complex, as well as outgoing edges from the complex**
  - For each candidate complex, its score is its cohesiveness multiplied by its weighted density

- **Cohesiveness:**

$$\frac{\sum edge\ weights\ inside\ cluster}{\sum edge\ weights\ inside\ cluster + \sum outdoing\ edge\ weights\ from\ cluster}$$

The cohesiveness of a size-2 cluster (a, b) and a size-3 cluster (a, b, c) respectively are:

$$Coh(a,b) = \frac{P'_{(a,b),sm2}}{P'_{(a,b),sm2} + \sum_{x \in \{a,b\}, y \in Na,b} \left(P_{(x,y),sm} + P_{(x,y),\lg}\right)}$$

$$Coh(a,b,c) = \frac{P'_{(a,b),sm3,abc} + P'_{(a,c),sm3,abc} + P'_{(b,c),sm3,abc}}{P'_{(a,b),sm3,abc} + P'_{(a,c),sm3,abc} + P'_{(b,c),sm3,abc} + \sum_{x \in \{a,b,c\}, y \in Na,b,c} \left(P_{(x,y),sm} + P_{(x,y),\lg}\right)}$$

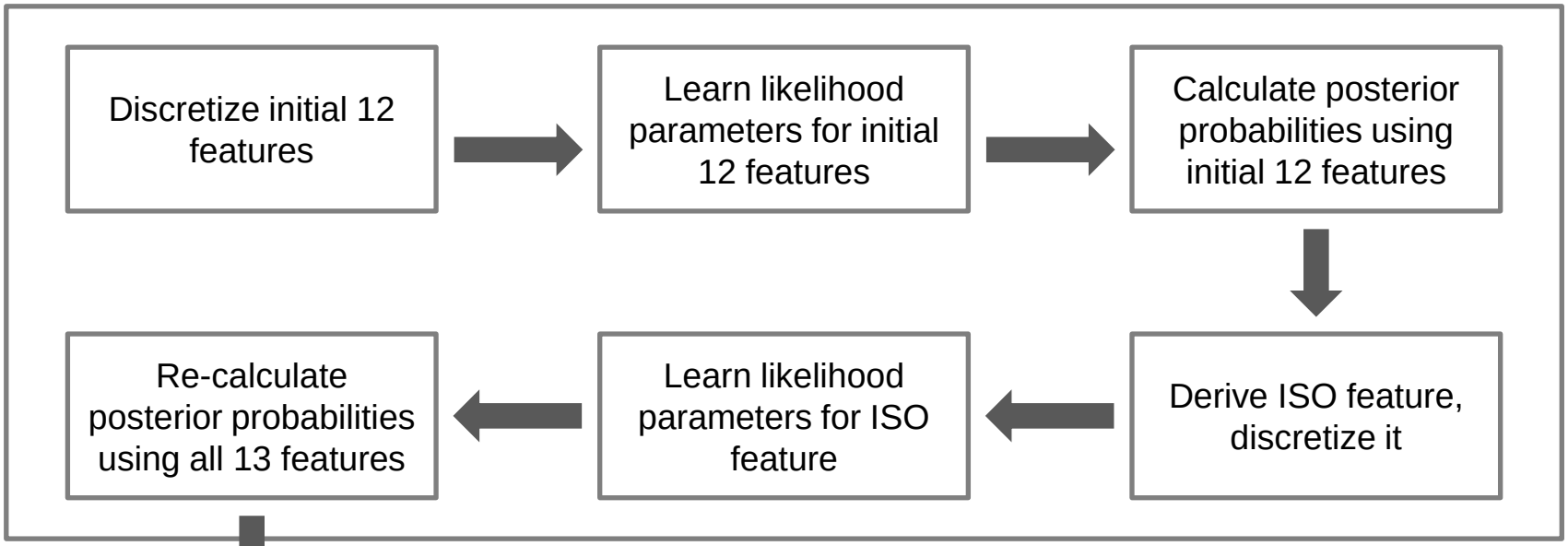
We then define the score of a cluster as its cohesiveness-weighted density, or the product of its weighted density and its cohesiveness. The score of a size-2 cluster (a, b), and a size-3 cluster (a, b, c) respectively are:

$$score(a,b) = Coh(a,b) P'_{(a,b),sm2}$$

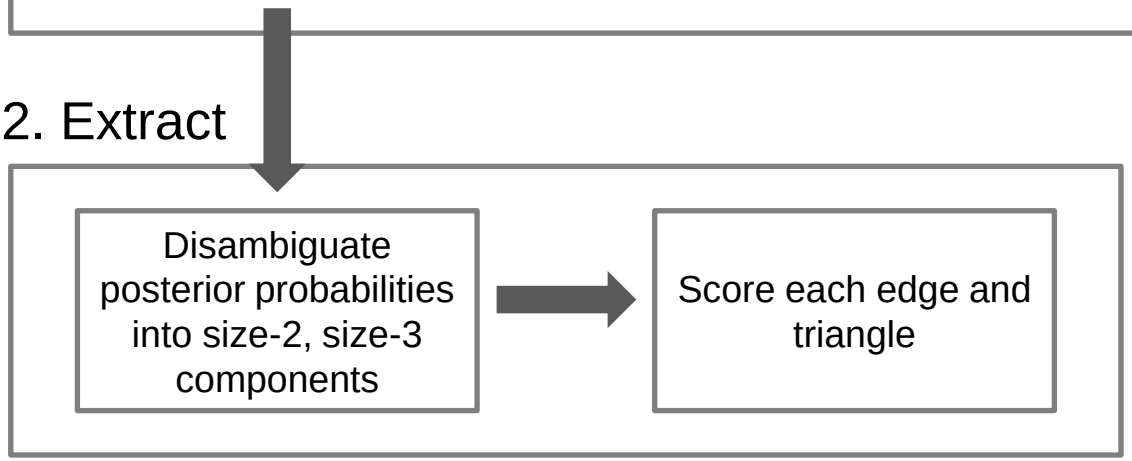
$$score(a,b,c) = Coh(a,b,c) \frac{(P'(a,b), sm3, abc + P'(a,c), sm3, abc + P'(b,c), sm3, abc)}{3}$$

Two-Stage Approach

1. Size-specific supervised weighting (SSS)



2. Extract



Benefits

- **Groups of proteins may take on small-complex topological characteristics in PPIN by chance**
  - ⇒ Use multiple data sources & their topological features
    - **Unlikely that all data sources share small-complex characteristics by chance**

- **Small-complex prediction is sensitive to noise in PPIN**
  - ⇒ Reduce noise by data integration with supervised learning

- **Other supervised-weighting complex-prediction approaches learn features of large complexes**
  - – Do not perform well for small complexes
  - ⇒ Size-specific weighting

- **Scoring candidate small complexes is sensitive to correct edge weights (very few edge weights used for scoring)**
  - ⇒ Use also outgoing edges from candidate complex during scoring

Experiment

- **Compare the following approaches:**
  - SSS + Extract: Proposed approach
  - Standard algo's with reliability-weighted PPI network (PPIREL)
  - Standard algo's with SSS-weighted network
- **10 rounds of cross-validation**
- **Prediction of yeast small complexes, with CYC2008 yeast reference complexes (human complexes also evaluated in manuscript)**
- **Exact-match evaluation: Predicted complexes have to match reference complexes exactly**

Yeast Small-Complex Prediction

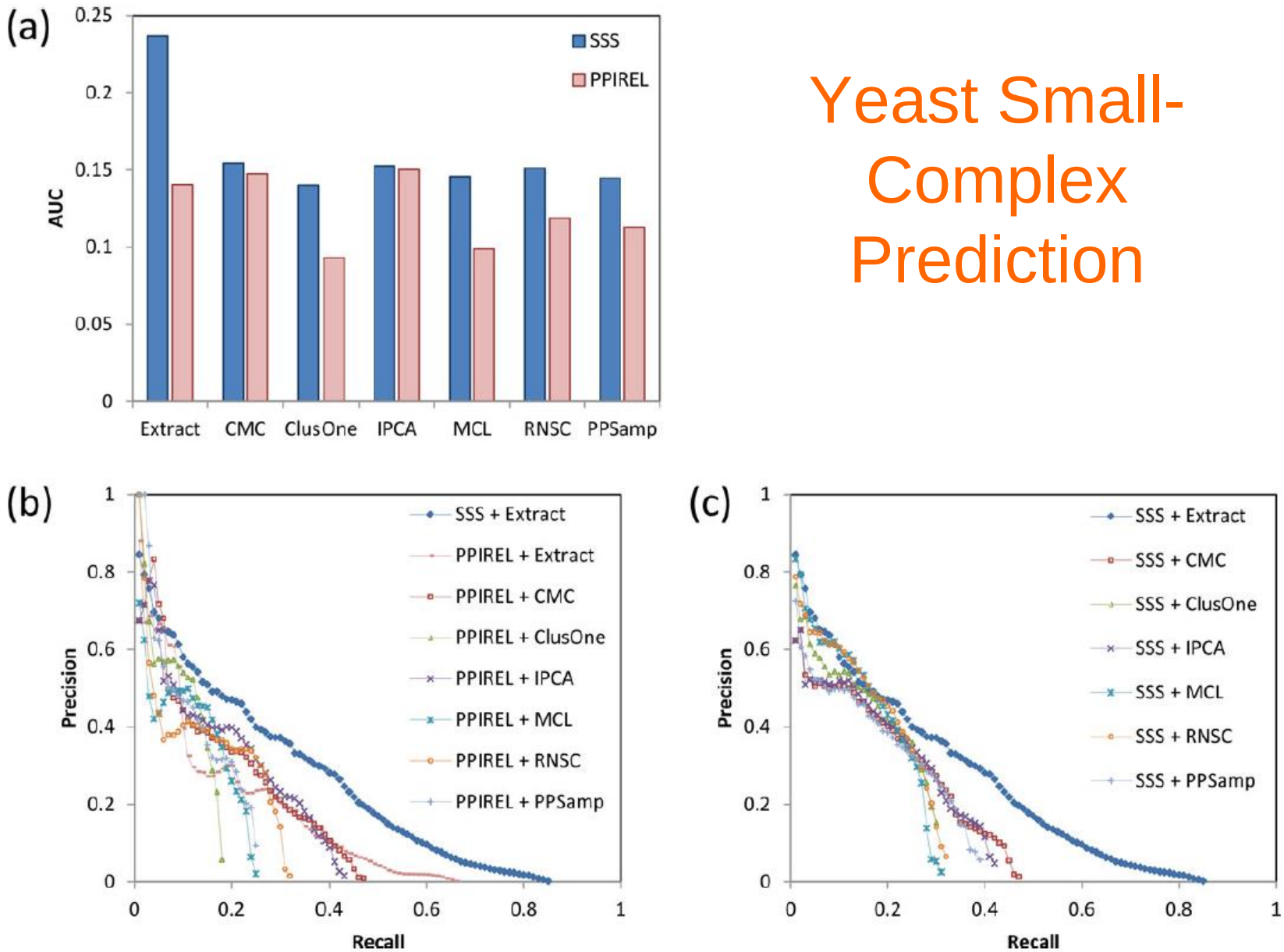
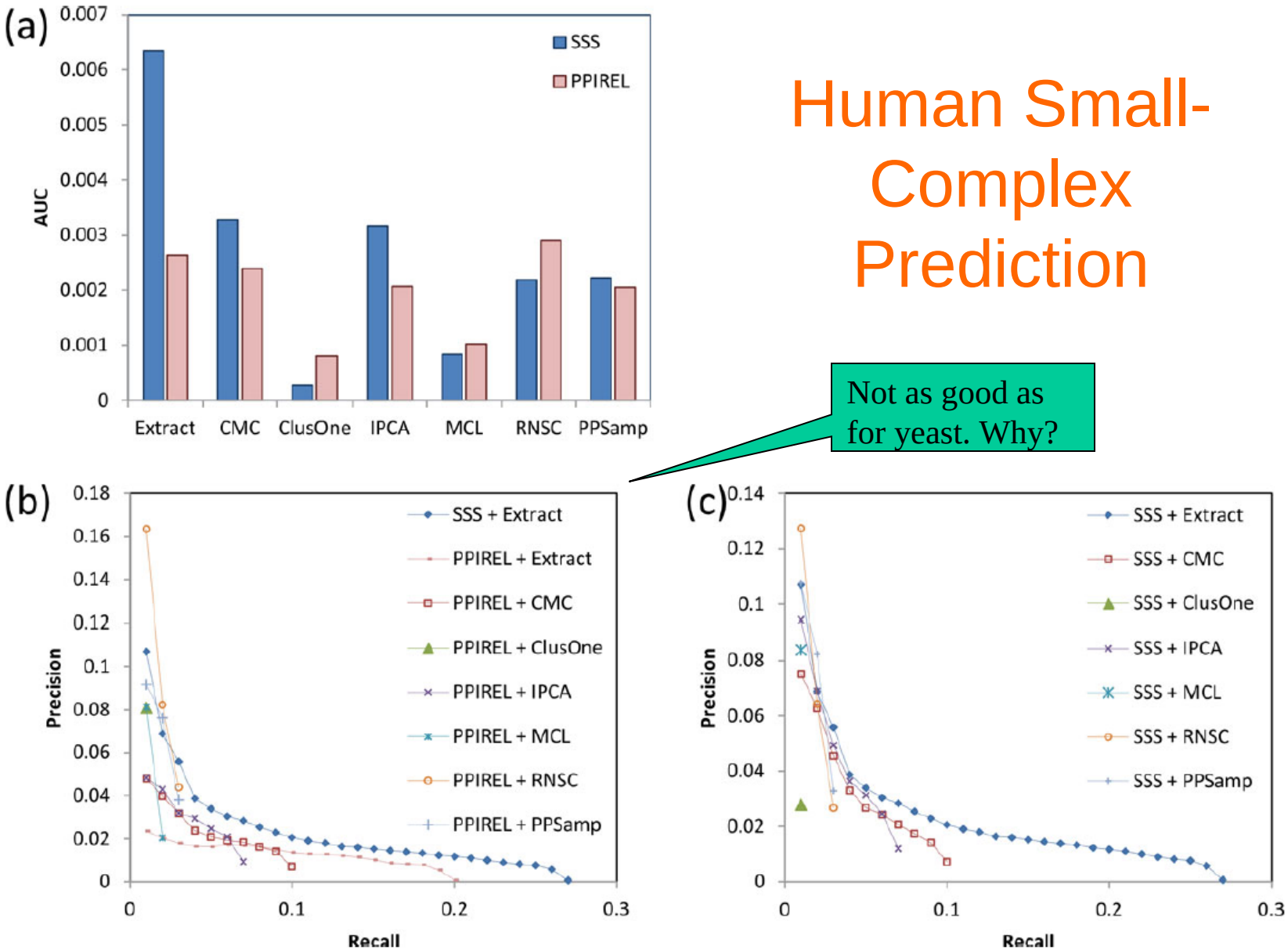


Figure 2 Performance of small complex prediction in yeast, (a) precision-recall AUC, (b) and (c) precision-recall graphs.

Human Small-Complex Prediction



Not as good as for yeast. Why?

Figure 3 Performance of small complex prediction in human, (a) precision-recall AUC, (b) and (c) precision-recall graphs.

Quality of Novel Complexes Predicted

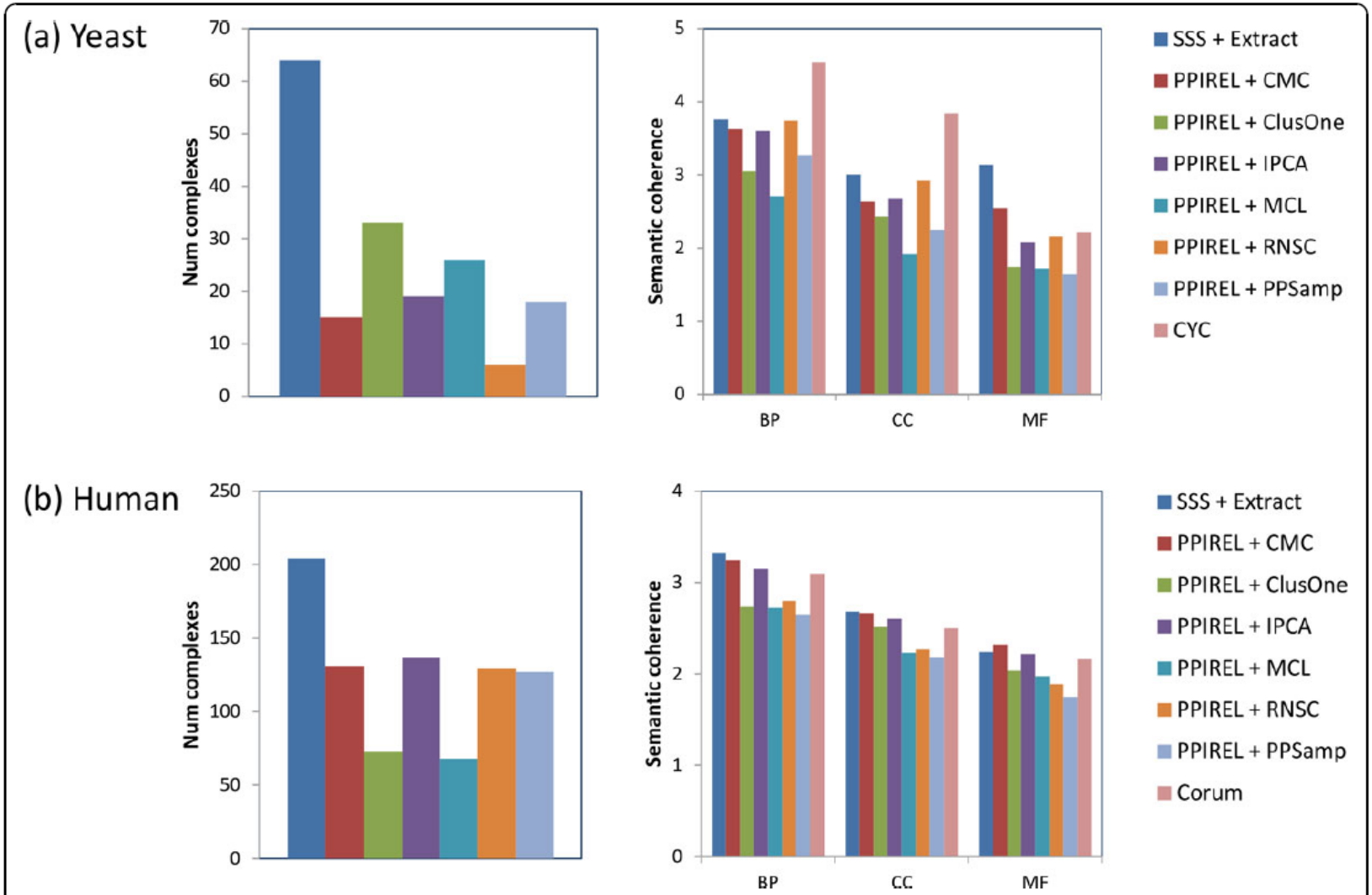
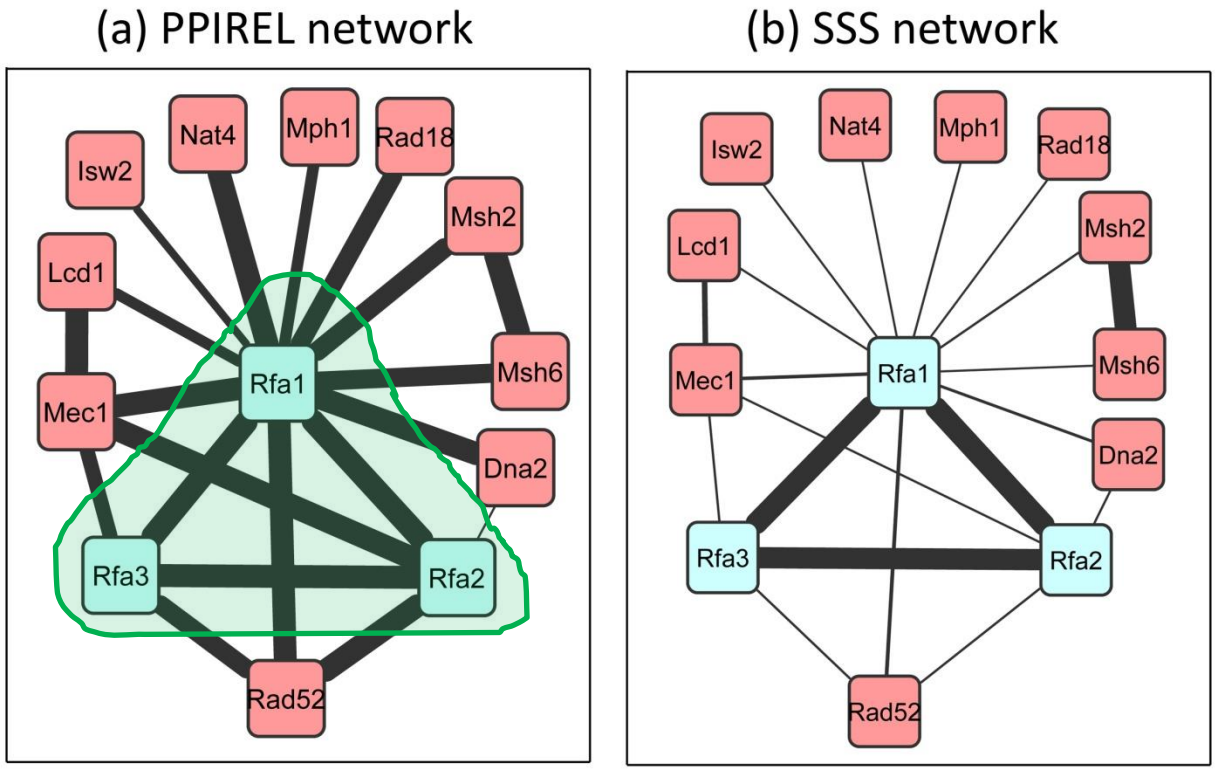


Figure 10 Number of high-confidence novel predictions, and their semantic coherences, in (a) yeast, (b) human.

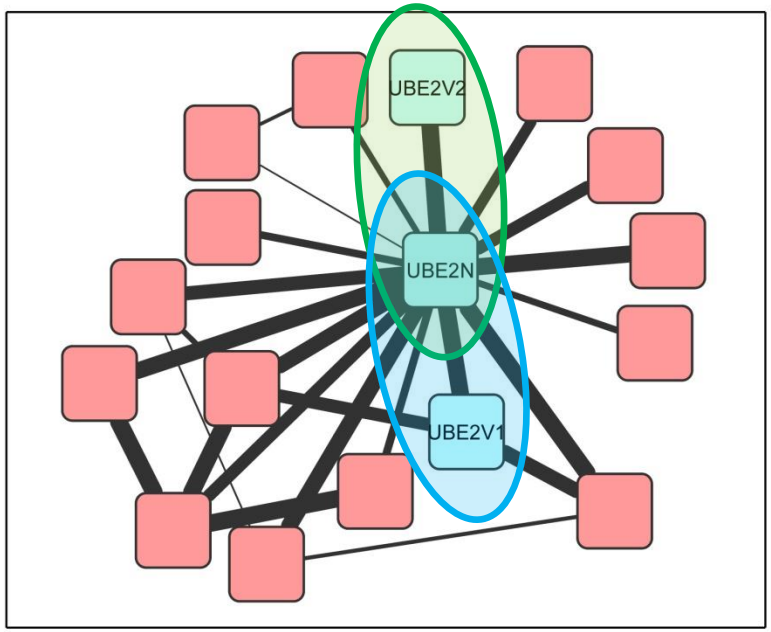
Yeast DNA Replication Factor A



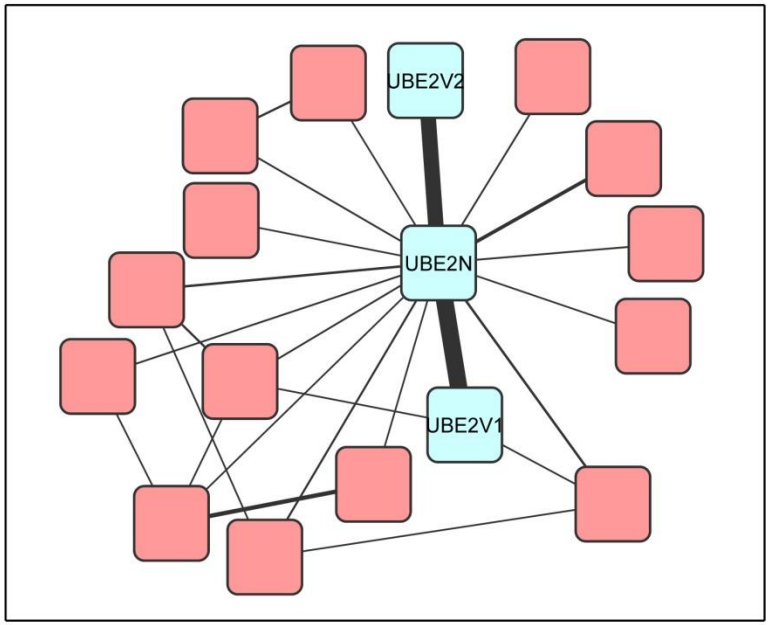
- **DNA replication factor A consists of 3 proteins**
- **Cannot be found by standard clustering algorithms on the PPI network**
  - Embedded within two size-4 cliques
  - Also part of many other size-3 cliques
- **After weighting by SSS, the internal weights of the complex remain high, while extraneous weights are lowered → Can be found in all cross-validation rounds**

Human Ubiquitin Ligase

(a) PPIREL network



(b) SSS network



- **Two human ubiquitin ligase complexes, which share 1 protein in common (UBE2N)**
- **Cannot be found by standard clustering algorithms on the PPI network**
  - Embedded within many larger cliques
  - Many extraneous edges
- **After weighting by SSS, the internal weights of the complex remain high, while extraneous weights are lowered**
  - UBE2V2-UBE2N can be found in all cross-validation rounds
  - UBE2V1-UBE2N can be found in 78% of cross-validation rounds

Conclusion

- **Most complexes are small, so small-complex prediction is an impt part of complex prediction**

- **Many challenges in small-complex prediction**
  - Searching for dense clusters is ineffectual
  - Sensitive to noise
  - Scoring candidate complexes is sensitive to edge weights

- **SSS + Extract**
  - Integrate 3 data sources w/ their topological features
  - Size-specific edge weighting by supervised learning
  - When scoring candidate complexes, incorporates outgoing edges from clusters as well

$\Rightarrow$ **Much improved performance in yeast and human**

Must Read

- Srihari et al. **Methods for protein complex prediction and their contributions towards understanding the organization, function and dynamics of complexes**. *FEBS Letters*, 589(19):2590--2602, 2015
- [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
- [SWC] Yong et al. **Supervised maximum-likelihood weighting of composite protein networks for complex prediction.** *BMC Systems Biology*, 6(Suppl 2):S13, 2012
- [SSS] Yong et al. **Discovery of small protein complexes from PPI networks with size-specific supervised weighting.** *BMC Systems Biology*, 8(Suppl 5):S3, 2014

Good to Read

- [MCODE] Bader & Hogue. **An automated method for finding molecular complexes in large protein interaction networks.** *BMC Bioinformatics*, 4:2, 2003
- [RNSC] King et al. **Protein complex prediction via cost-based clustering**. *Bioinformatics*, 20(17):3013-3020, 2004
- [MCL] Pereira-Leal et al. **Detection of functional modules from protein interaction networks.** *Proteins: Structure, Function, and Bioinformatics,*54:49-57, 2004
- Hirsh & Sharan. **Identification of conserved protein complexes based on a model of protein network evolution.** *Bioinformatics*, 23(2):e170-e176, 2007
- [RNSC] King et al. **Protein complex prediction via cost-based clustering.** *Bioinformatics*, 20(17):3013-3020, 2004
- [DECAFF] Li et al. **Discovering protein complexes in dense reliable neighbourhoods of protein interaction networks.** CSB, 2007, pp. 157-168
- [COACH] Wu et al. **A core-attachment based method to detect protein complexes in PPI networks.** *BMC Bioinformatics*, 10:169, 2009
- [SPIN] Jung et al. **Protein complex prediction based on simultaneous protein interaction network.** *Bioinformatics*, 26(3):385-391, 2010

Acknowledgements

- **A lot of the slides for this lecture were adapted from ppt files given to me by Sriganesh Srihari and Yong Chern Han**
- **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