Protein Complex Prediction from PPIN

Limsoon Wong

NUS National University of Singapore

Lecture Outline

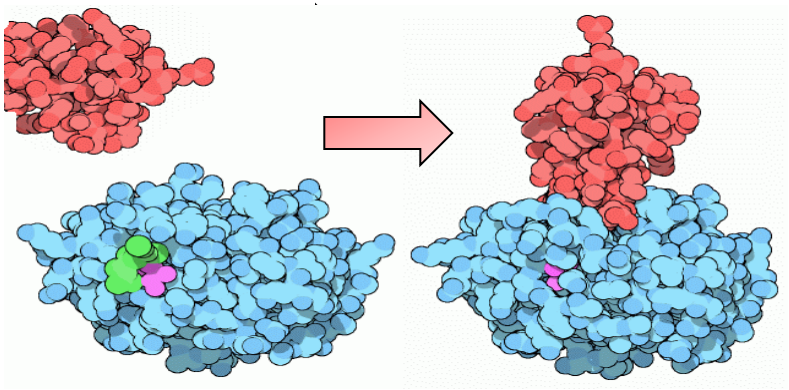
- **Overview of protein complex prediction**
- **Impact of PPIN cleansing**
- **Detecting overlapping complexes**
- **Detecting sparse complexes**
- **Detecting small complexes**

Overview of
Protein Complex Detection from PPIN

Source: Sriganesh Srihari

NUS National University of Singapore

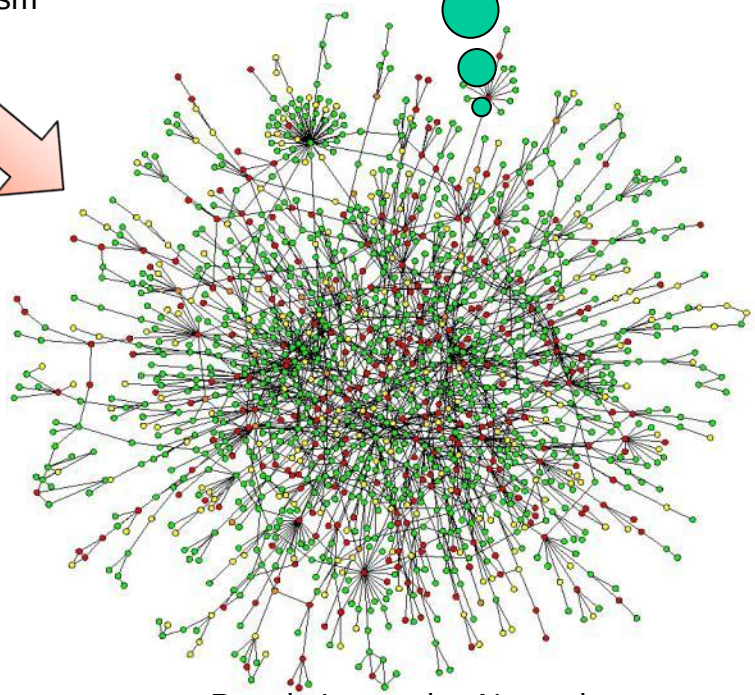
Protein Interaction Networks



Individual proteins come together and interact

Collection of such interactions in an organism

PPIN
Valuable source of knowledge

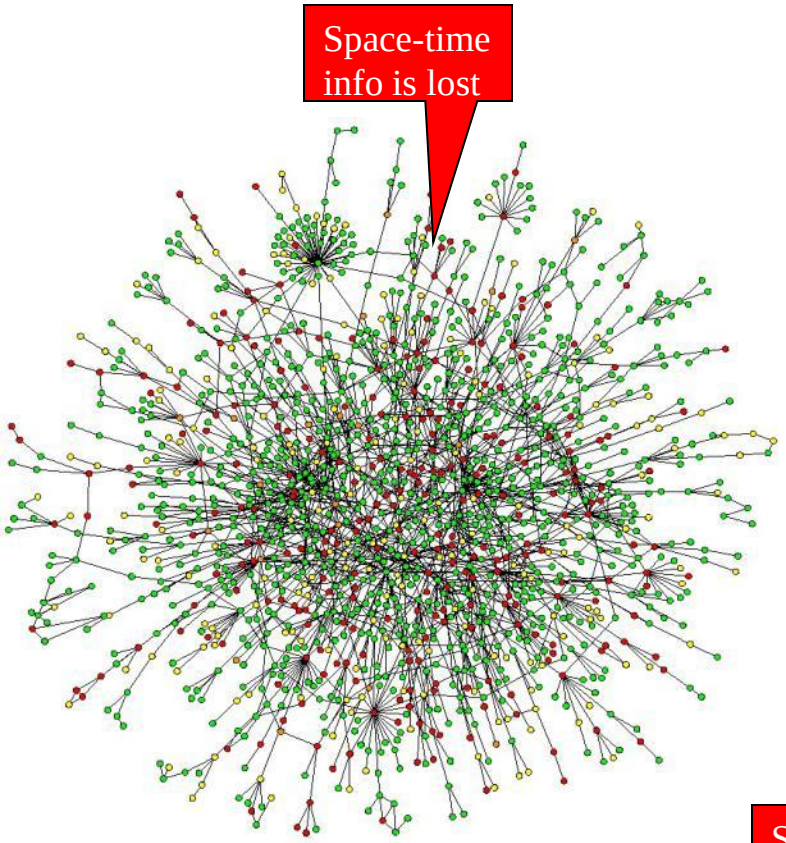


Protein Interaction Network

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

Source: Sriganesh Srihari

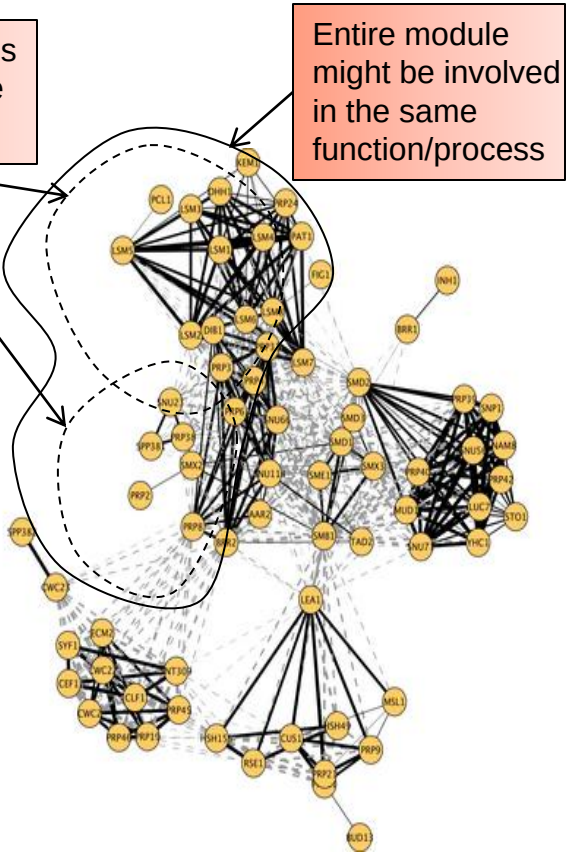
Detection & Analysis of Protein Complexes in PPIN



PPIN derived from several high-throughput expt

Identifying embedded complexes

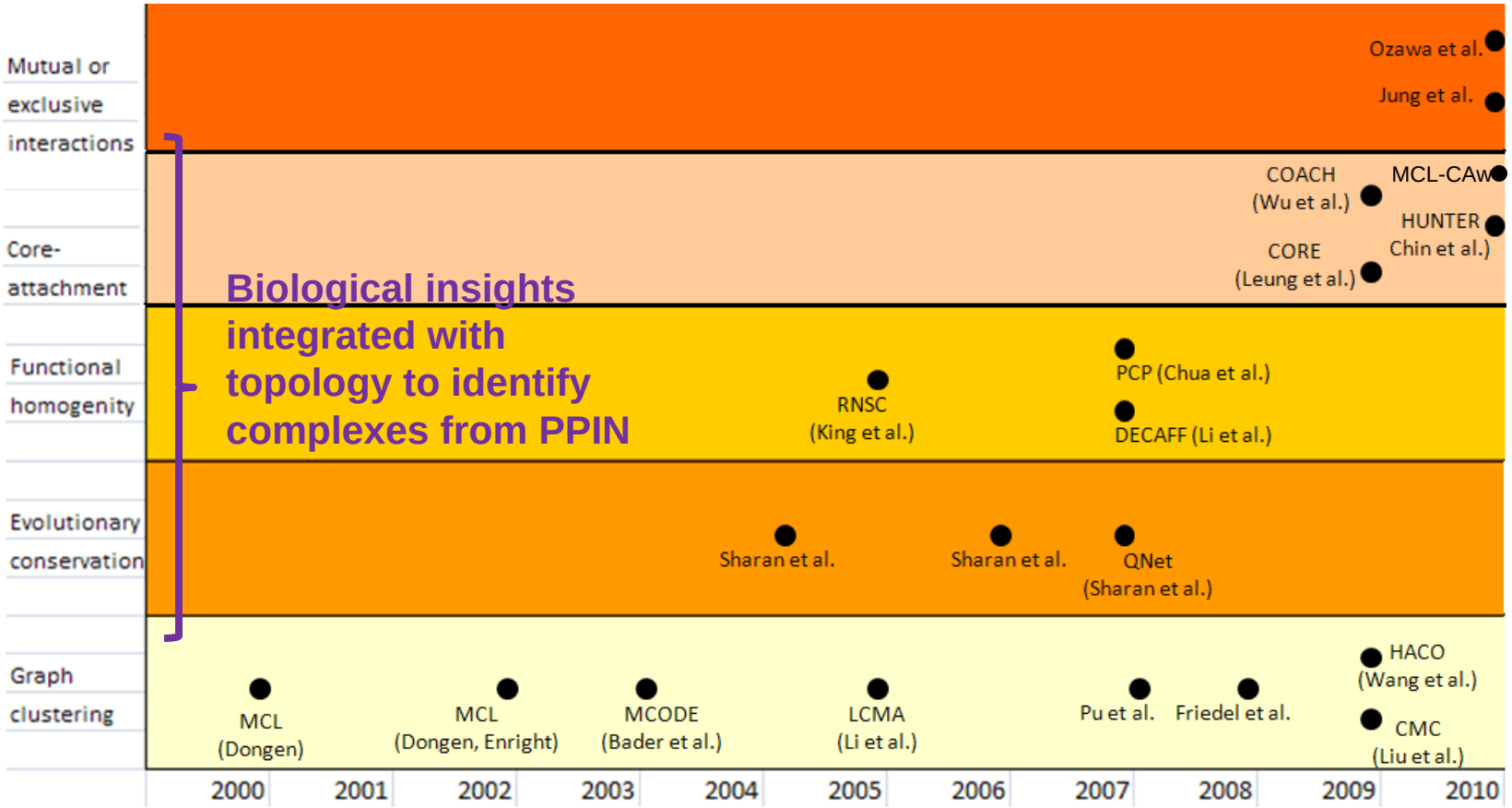
Space-time info is "recovered"



Embedded **complexes** identified from PPIN

Source: Sriganesh Srihari

Chronology of Protein Complex Prediction Methods

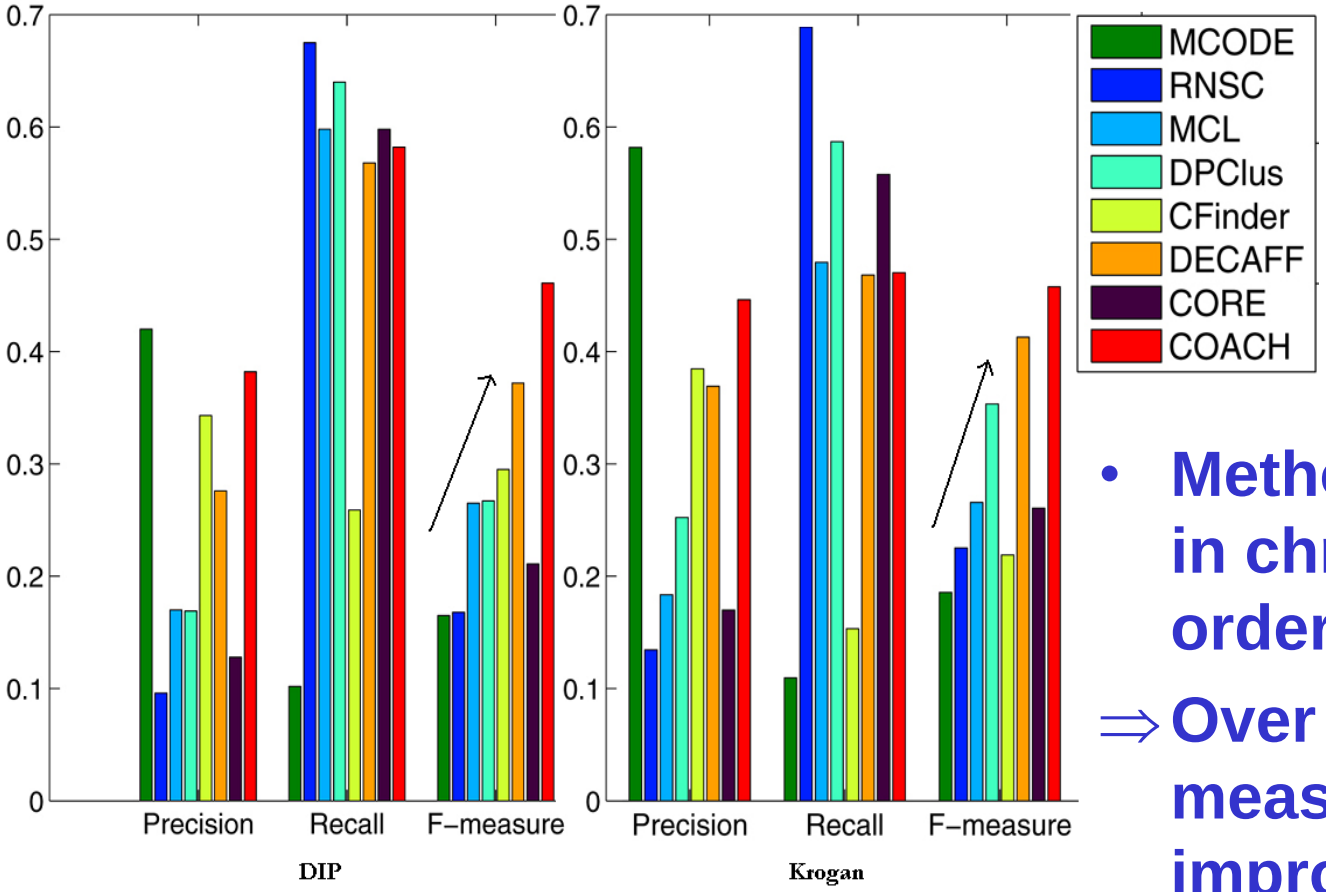


- **As researchers try to improve basic graph clustering techs, they also incorporate bio insights into the methods**

Li et al., *BMC Genomics*, 11(Suppl 1):S3, 2010

NUS National University of Singapore

Comparative Assessment



(Noisy, sparse, old)

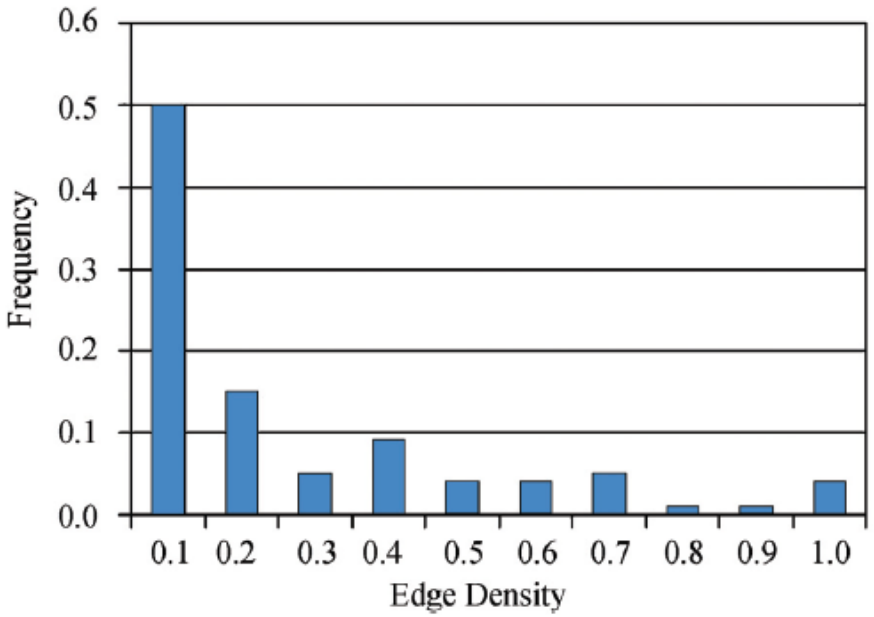
(Good quality, dense, new)

- **Methods arranged in chronological order**

⇒ **Over the years, F1-measure have improved!**

Challenges

- **Recall & precision of protein complex prediction algo's have lots to be improved**



- **Does a "cleaner" PPI network help?**
- **How to capture "high edge density" complexes that overlap each other?**
- **How to capture "low edge density" complexes?**
- **How to capture small complexes?**

Impact of PPIN Cleansing on Protein Complex Prediction

NUS National University of Singapore

Noise in PPI Networks

Experimental method category[a]	Number of interacting pairs	Co-localization[b] (%)	Co-cellular-role[b] (%)
All: All methods	9347	64	49
A: Small scale Y2H	1861	73	62
A0: GY2H Uetz *et al.* (published results)	956	66	45
A1: GY2H Uetz *et al.* (unpublished results)	516	53	33
A2: GY2H Ito *et al.* (core)	798	64	40
A3: GY2H Ito *et al.* (all)	3655	41	15
B: Physical methods	71	98	95
C: Genetic methods	1052	77	75
D1: Biochemical, *in vitro*	614	87	79
D2: Biochemical, chromatography	648	93	88
E1: Immunological, direct	1025	90	90
E2: Immunological, indirect	34	100	93
2M: Two different methods	2360	87	85
3M: Three different methods	1212	92	94
4M: Four different methods	570	95	93

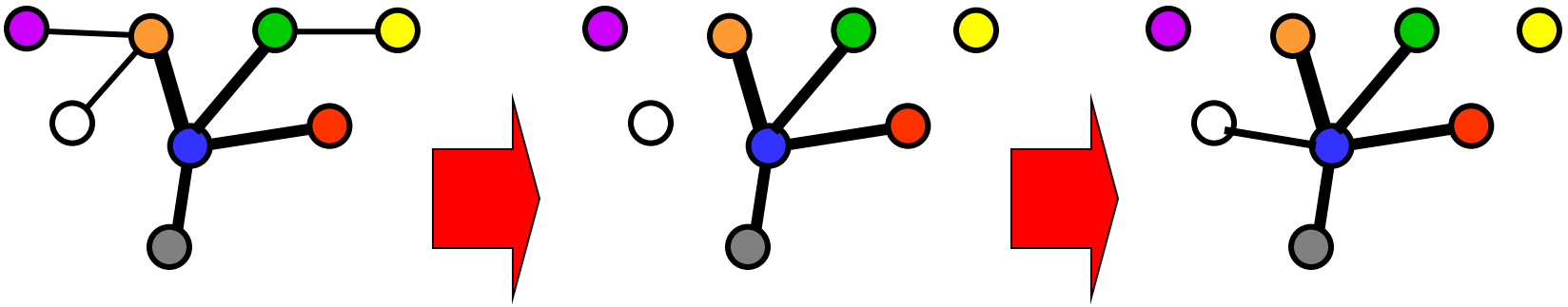
Sprinzak et al., *JMB*, 327:919-923, 2003

Large disagreement betw methods

- **High level of noise**

⇒ **Need to clean up before making inference on PPI networks**

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**

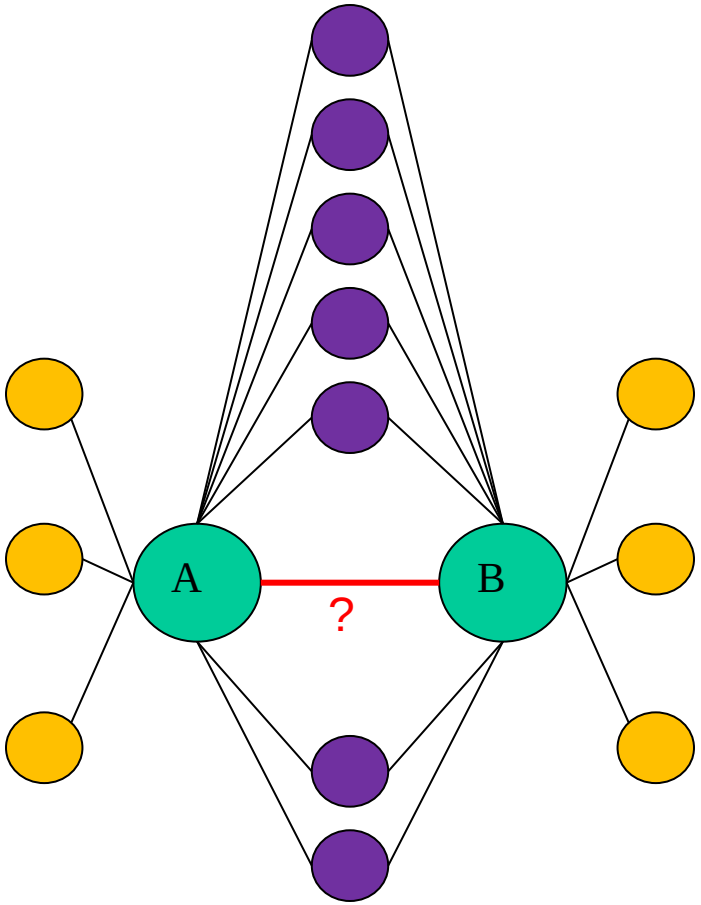
Liu, et al. “Complex Discovery from Weighted PPI Networks”, *Bioinformatics*, 25(15):1891-1897, 2009

NUS National University of Singapore

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**

CD Distance



- **Suppose 20% noise in the PPIN**

⇒ **≥ 3 purple proteins are real partners of both A and B**

⇒ **A and B are likely localized to the same cellular compartment (Why?)**

- **Fact: Proteins in the same cellular compartment are 10x more likely to interact than other proteins**

⇒ **A and B are likely to interact**

- **CD distance measures the proportion of A and B's neighbours that are common between them**

Liu, et al. *Bioinformatics*, 25(15):1891-1897, 2009

NUS National University of Singapore

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**

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

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

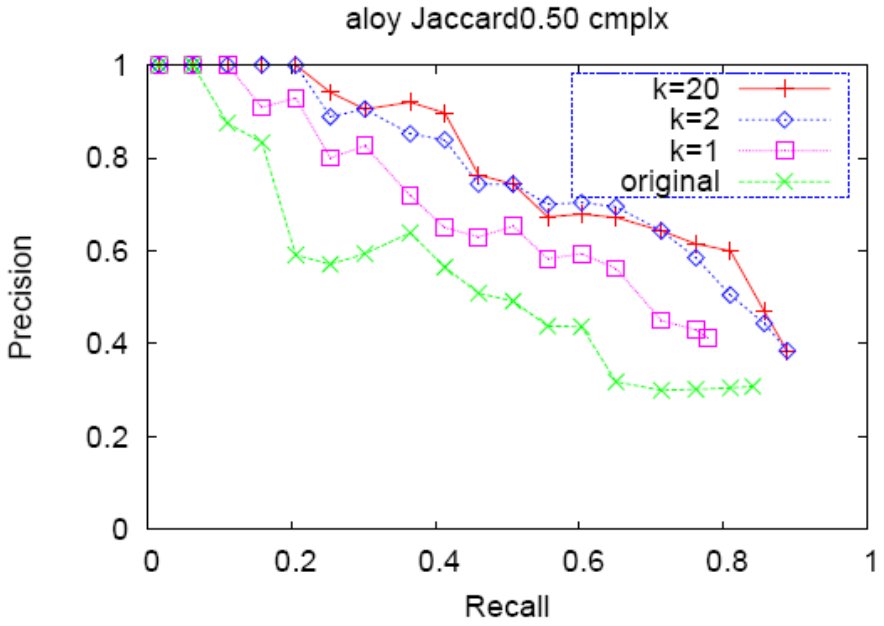
NUS National University of Singapore

Validation Experiments

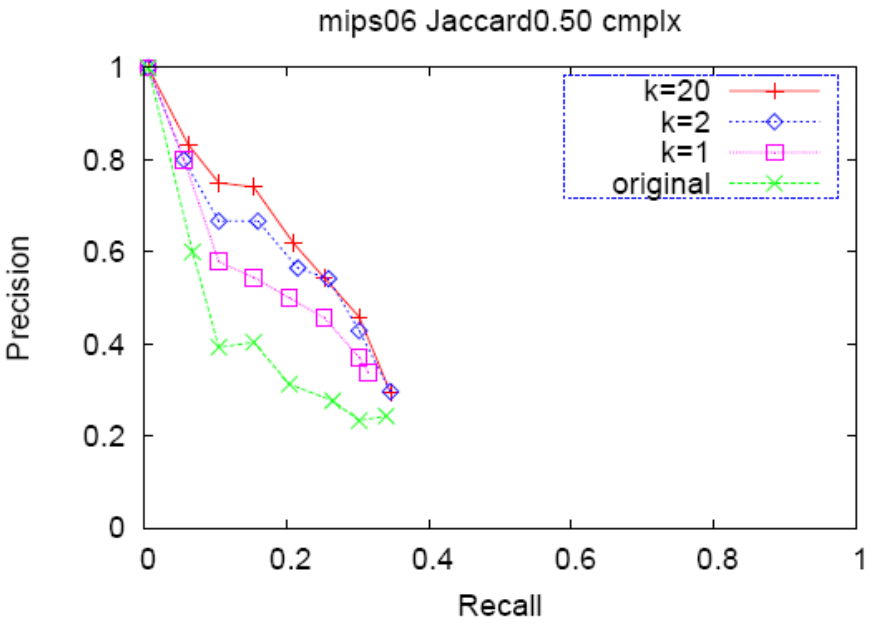
- **Matching a predicted complex S with a true complex C**
  - Vs: set of proteins in S
  - Vc: set of proteins in C
  - Overlap(S, C) = $|Vs \cap Vc| / |Vs \cup Vc|$, Overlap(S, C) ≥ 0.5
- **Evaluation**
  - Precision = matched predictions / total predictions
  - Recall = matched complexes / total complexes
- **Datasets: combined info from 6 yeast PPI expts**
  - #interactions: 20,461 PPI from 4,671 proteins
  - #interactions with >0 common neighbor: 11,487

Liu, et al. *Bioinformatics*, 25(15):1891-1897, 2009

Effecting of Cleaning on CMC



(a) Aloy, *match_thres*=0.50

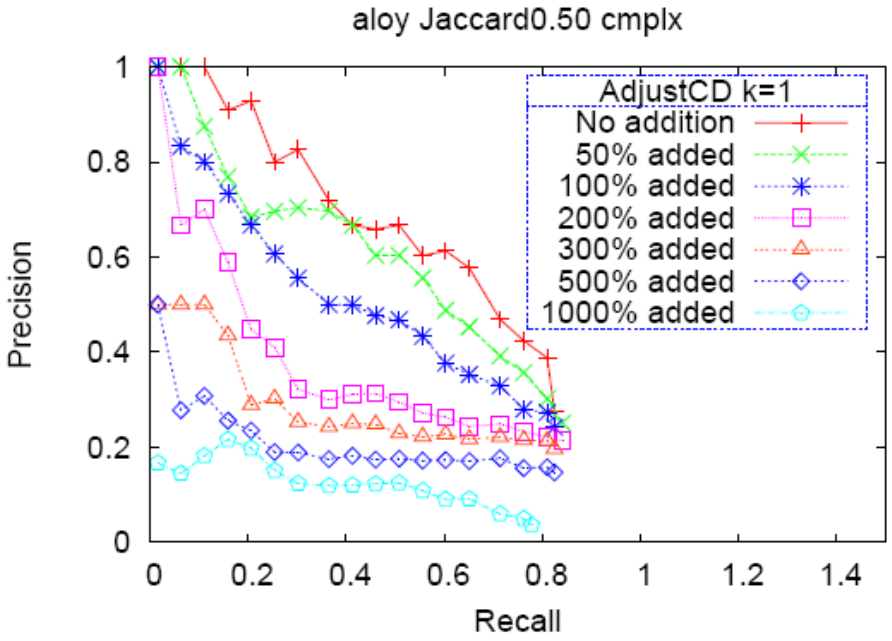


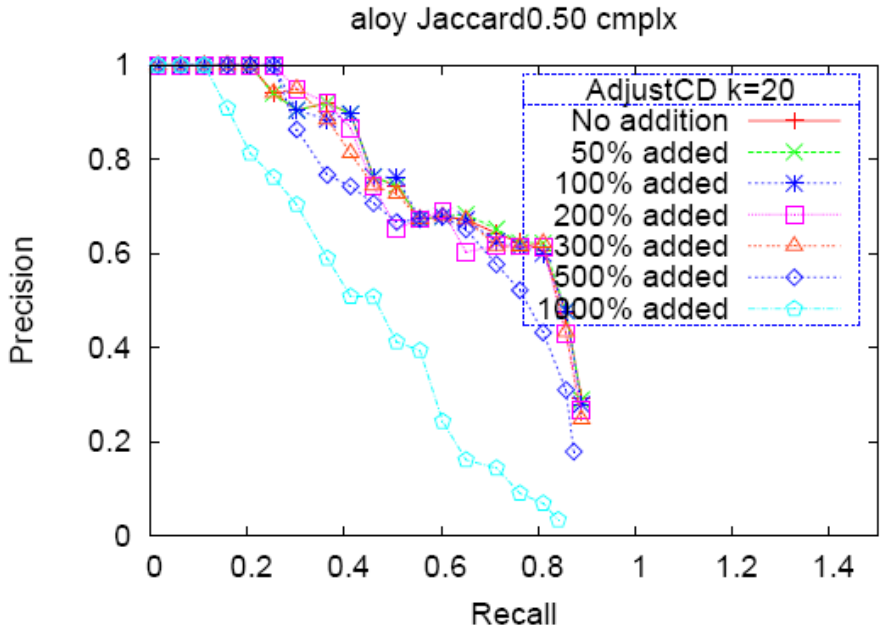
(b) MIPS, *match_thres*=0.50

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

Liu, et al. *Bioinformatics*, 25(15):1891-1897, 2009

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!**

Liu, et al. *Bioinformatics*, 25(15):1891-1897, 2009

NUS National University of Singapore

Other methods benefit too..

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.

Detecting Overlapping Protein Complexes from Dense Regions of PPIN

Overlapping Complexes in Dense Regions of PPIN

- **Dense regions of PPIN often contain multiple overlapping protein complexes**
- **These complexes often get clustered together and cannot be corrected detected**

- **Two ideas to solve this problem**
  - 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

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**

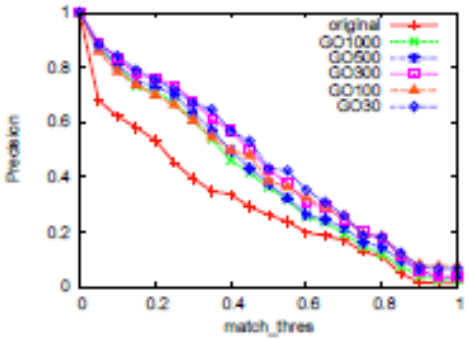
⇒ **Use general cellular component (CC) GO terms to decompose a given PPI network into several smaller PPI networks**

- **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”?

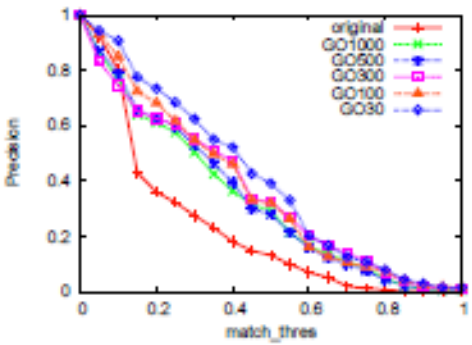
Liu, et al. “Decomposing PPI Networks for Complex Discovery”. *Proteome Science*, 9(Suppl. 1):S15, 2011

NUS National University of Singapore

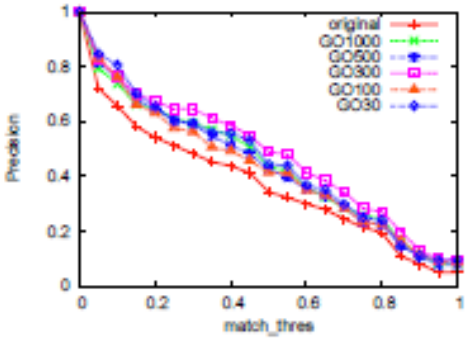
Effect of N_{GO} on Precision



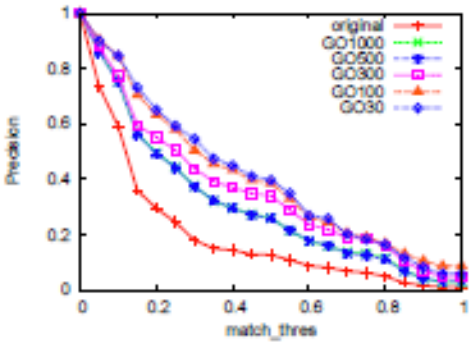
(b) MCL precision



(f) IPCA precision



(d) RNSC precision



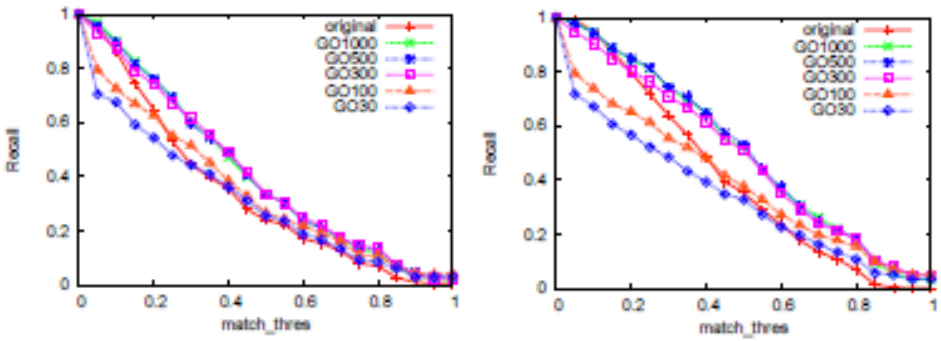
(h) CMC precision

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

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

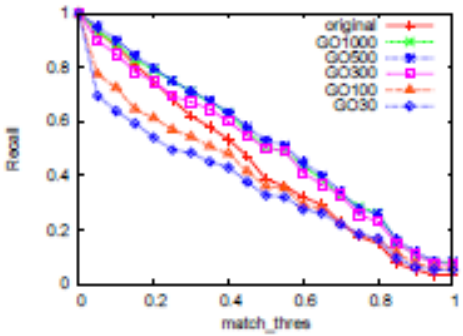
NUS National University of Singapore

Effect of N_{GO} on Recall

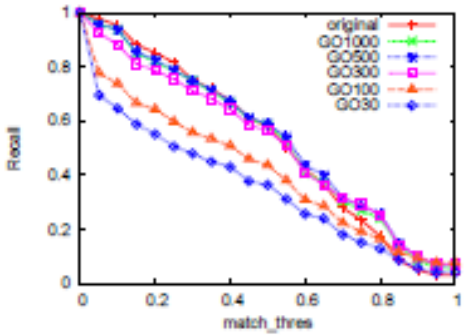


(a) MCL recall

(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**

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

NUS National University of Singapore

Idea II: Remove Big Hubs

- **Hub proteins are those proteins that have many neighbors in the PPI network**

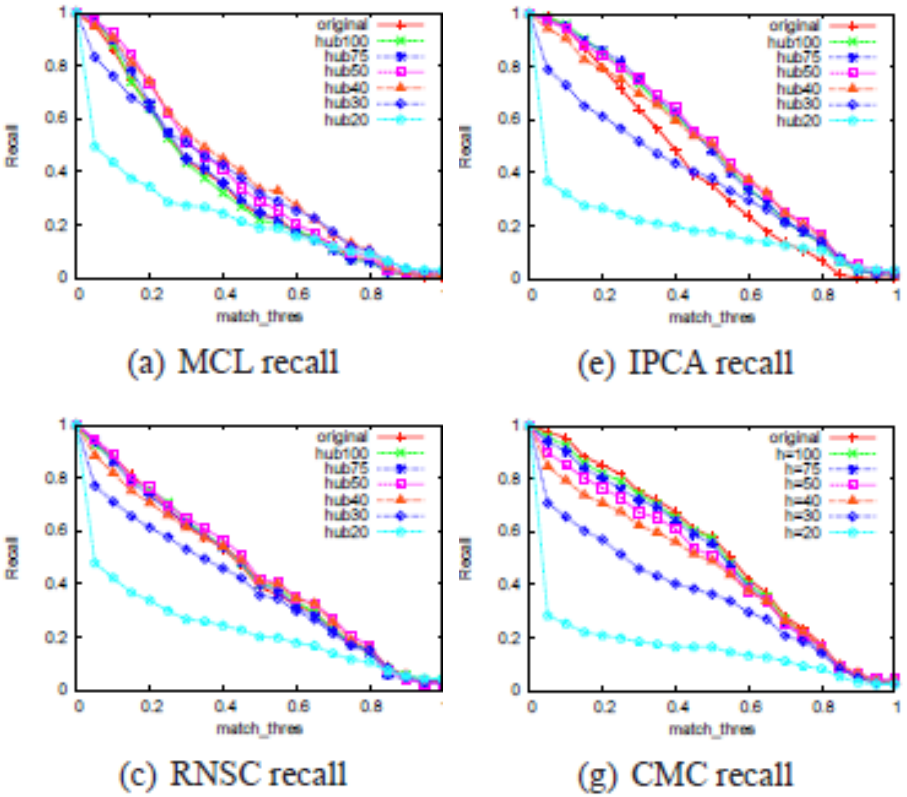
- **Large hubs are likely to be “date hubs”; i.e., proteins that participate in many complexes**
  - Likely to confuse protein complex prediction algo

⇒ **Remove large hubs before protein complex prediction**
  - How to choose threshold N_{hub} to decide whether a hub is “large”?

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

NUS National University of Singapore

Effect of N_{hub} on 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

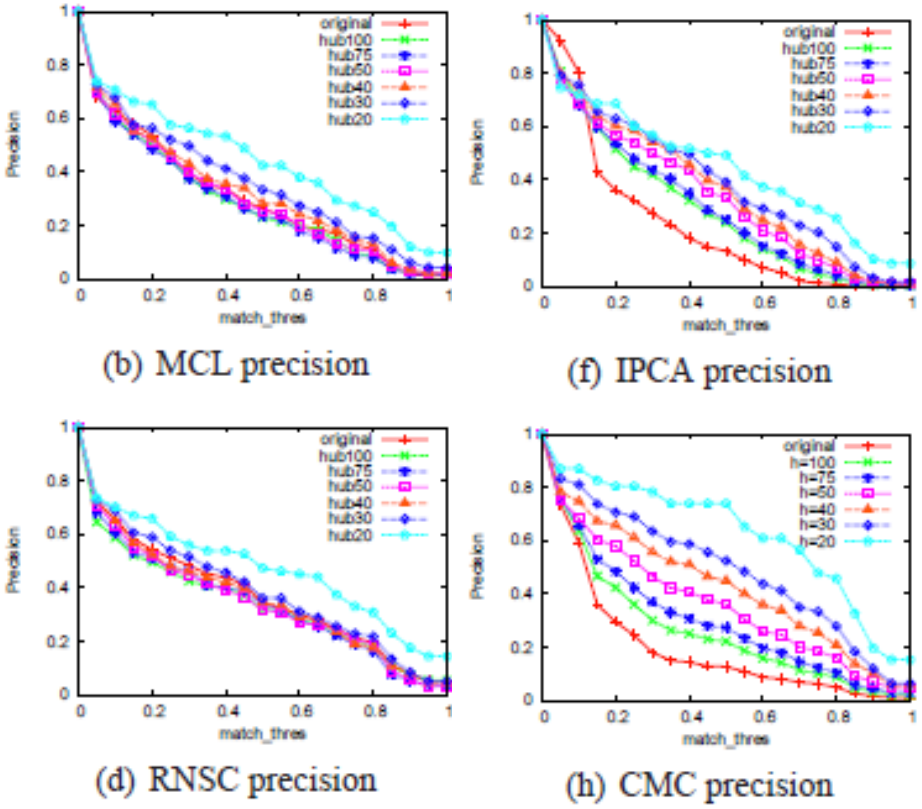
Table 4. Number of hub proteins and PPIs removed under different N_{hub}.

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

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

NUS National University of Singapore

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.

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

NUS National University of Singapore

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 $\mathcal{C}$.

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

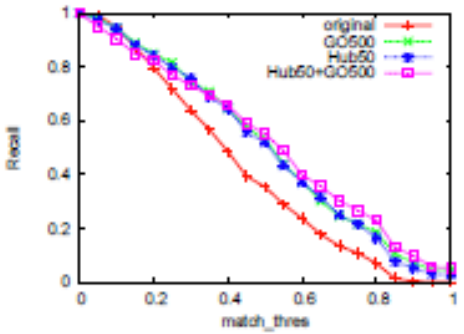
NUS National University of Singapore

Effect of Combining N_{GO} & N_{hub}

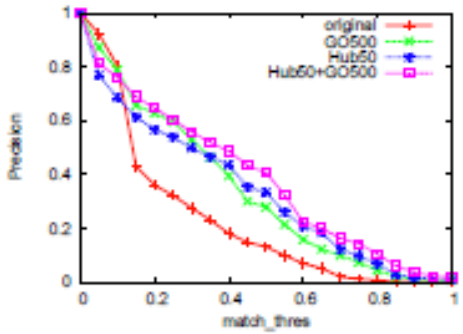
Table 5 - F1-measure of the four algorithms when $match\_thres$=0.5

	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

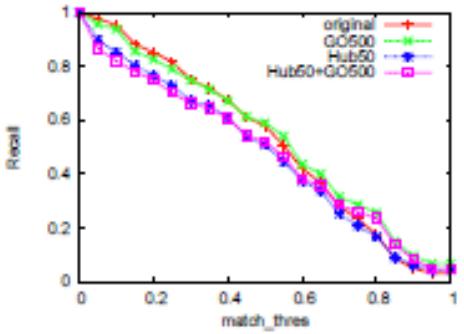
- **RNSC doesn’t benefit further**
- **MCL, IPCA & CMC all gain further**



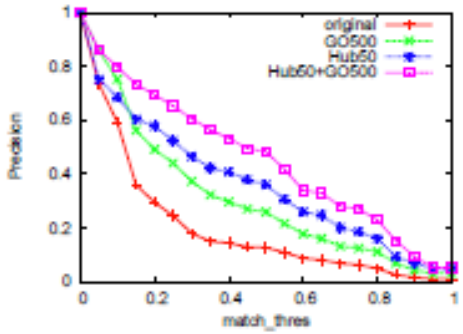
(e) IPCA recall



(f) IPCA precision



(g) CMC recall



(h) CMC precision

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

NUS National University of Singapore

Conclusions

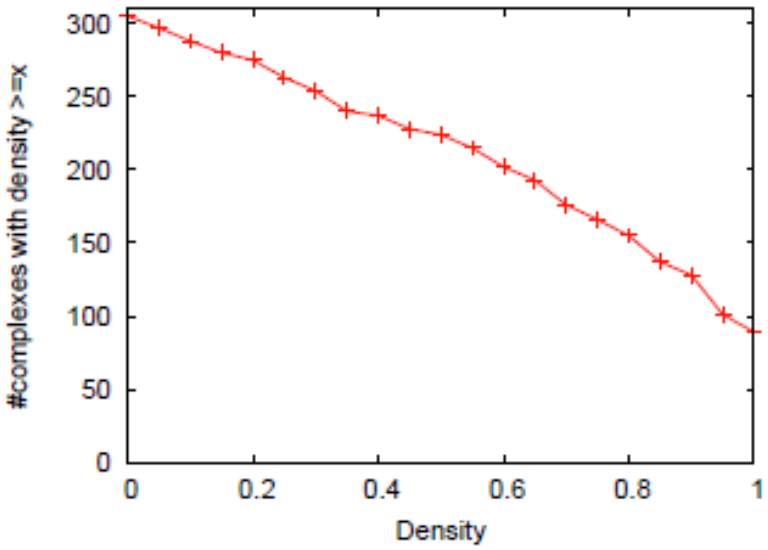
Table 5 - F1-measure of the four algorithms when $match\_thres$=0.5

	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…**

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

Many complexes not detectable. Why?

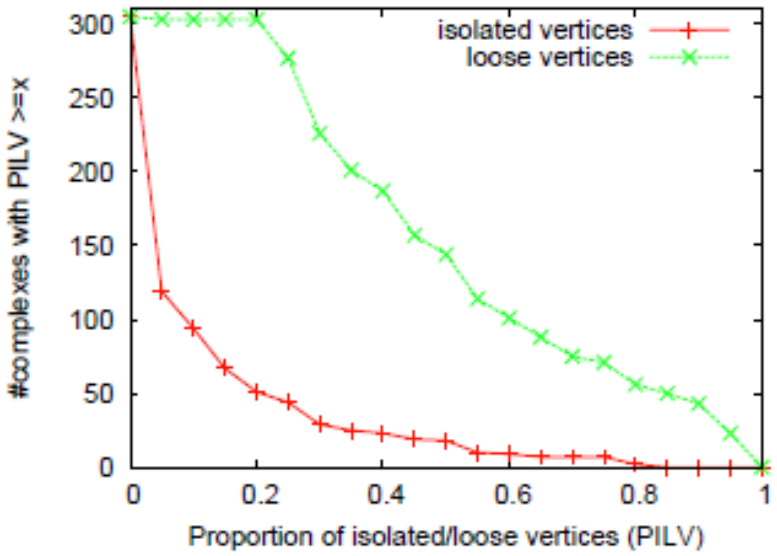


(a) density

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

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

Many complexes not detectable. Why?

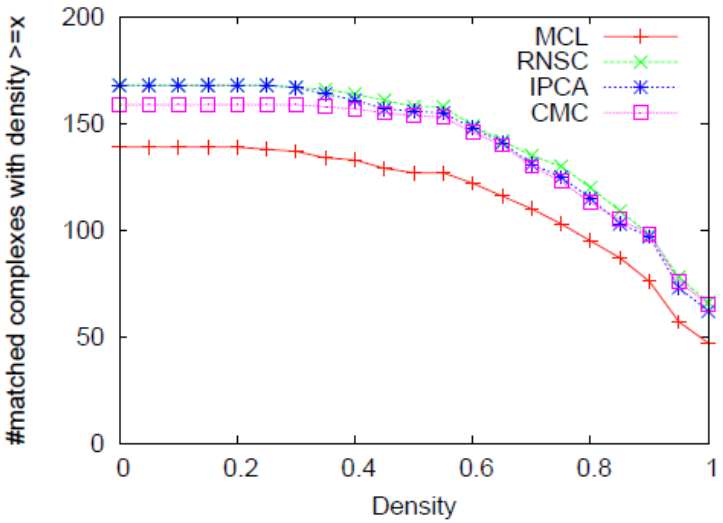


(b) connectivity

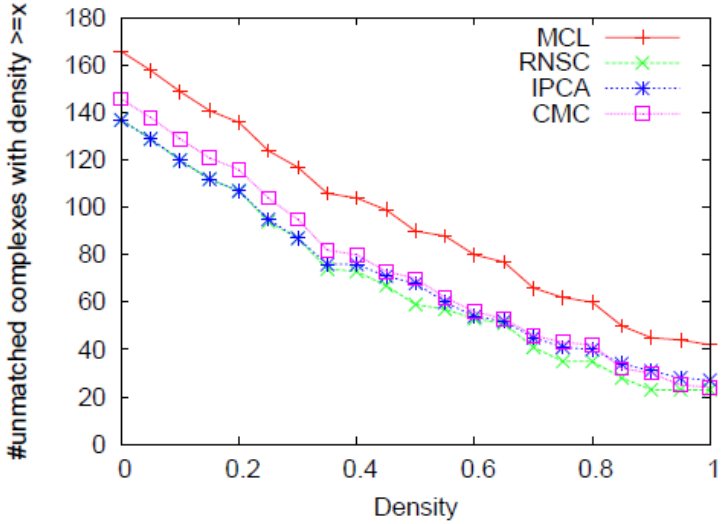
- **18 complexes w/ more than half of their proteins being isolated**
  - *Isolated vertex* connects to no other vertices in the complex
- **144 complexes w/ more than half of their proteins being loose**
  - *Loose vertex connects to < 50% of* other vertices in the complex

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

Many complexes not detectable. Why?



(a) detected complexes



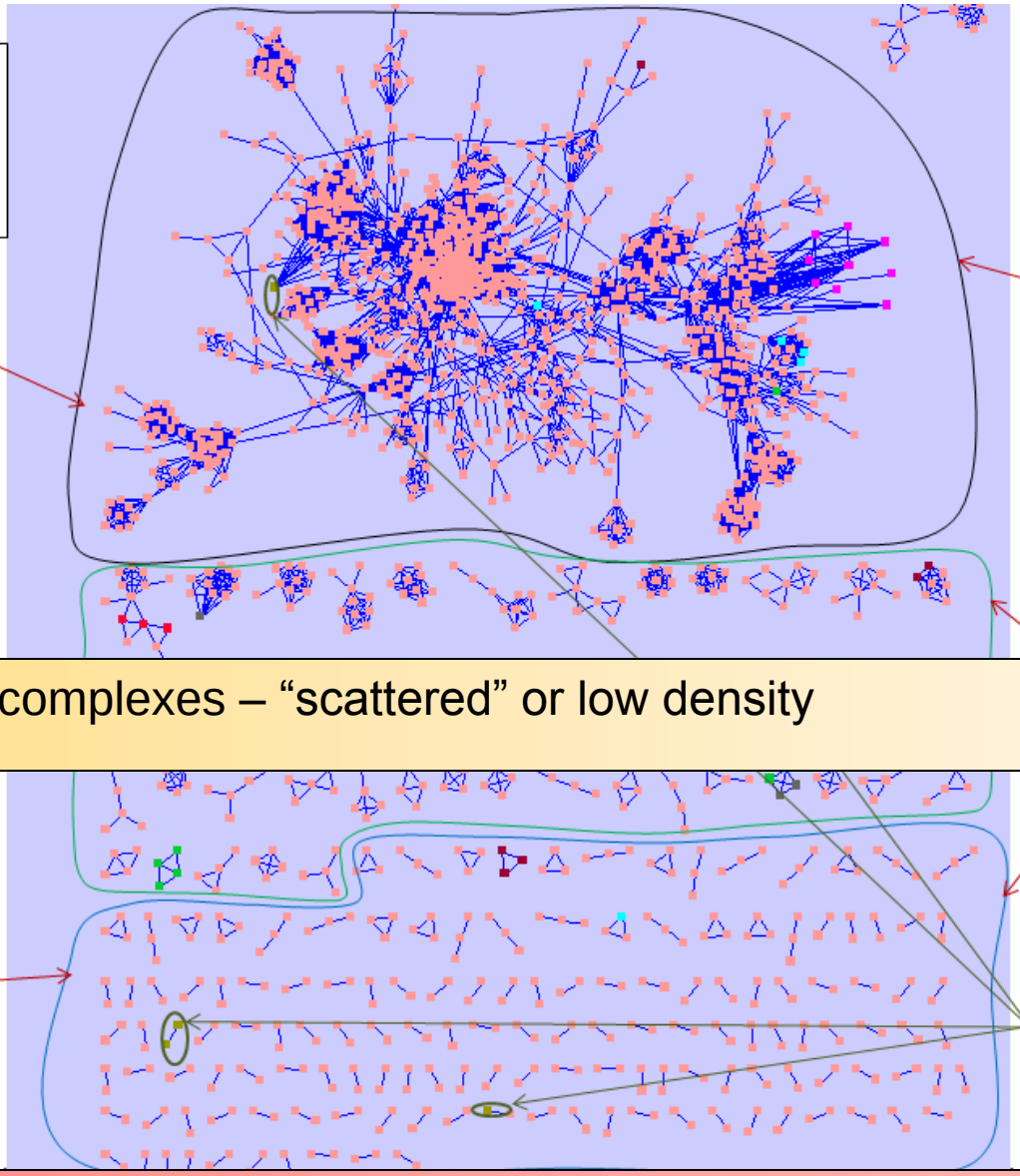
(b) undetected complexes

- **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

Source: Sriganesh Srihari

Sparse Complexes



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

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

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.**

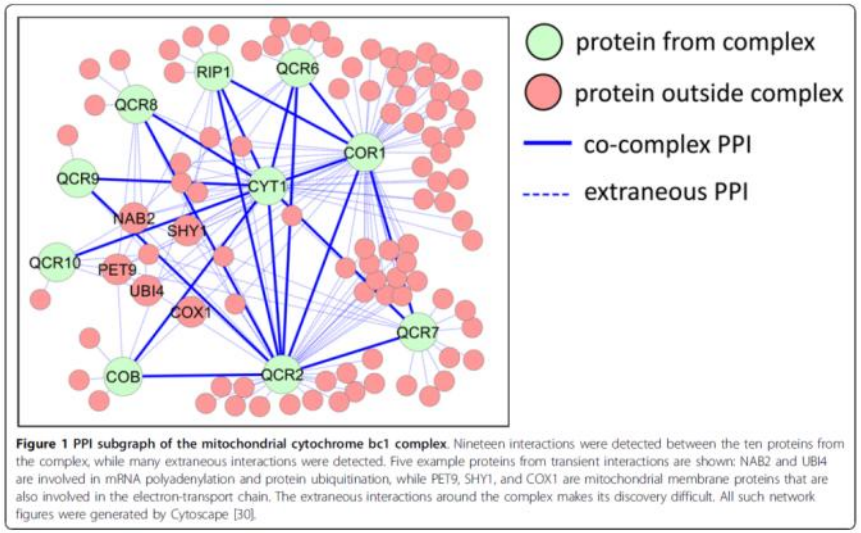
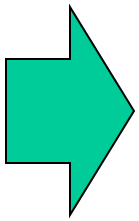


Figure 1 PPI subgraph of the mitochondrial cytochrome bc1 complex. Nineteen interactions were detected between the ten proteins from the complex, while many extraneous interactions were detected. Five example proteins from transient interactions are shown: NAB2 and UBI4 are involved in mRNA polyadenylation 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. “upervised 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**

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

NUS National University of Singapore

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**

NUS National University of Singapore

1. Integrate multiple data 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:

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

NUS National University of Singapore

3. Complex discovery

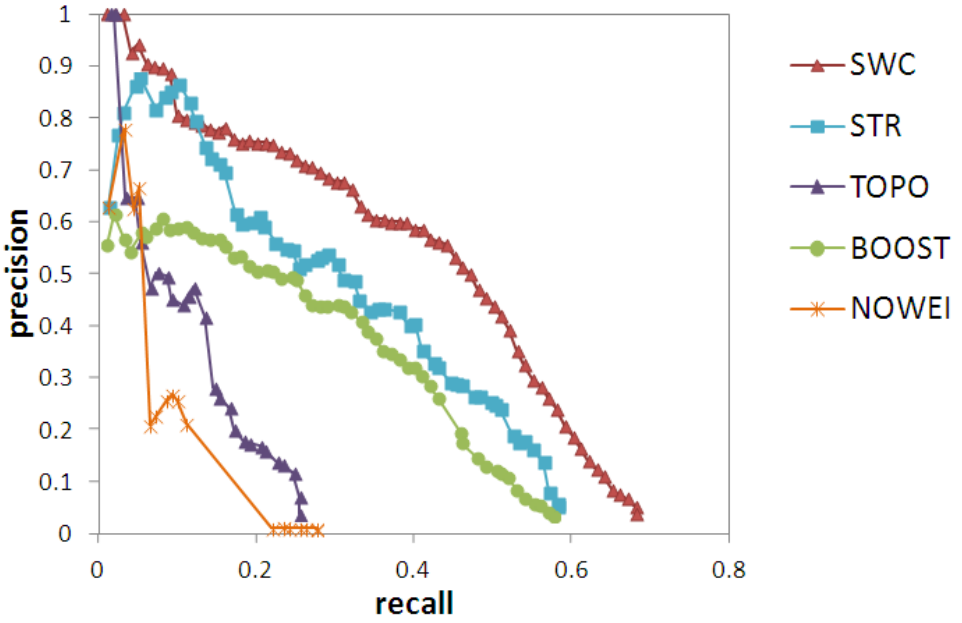
- **Weighted composite network used as input to clustering algorithms**
  - CMC, ClusterONE, IPCA, MCL, RNSC, HACO
- **Predicted complexes scored by weighted density**

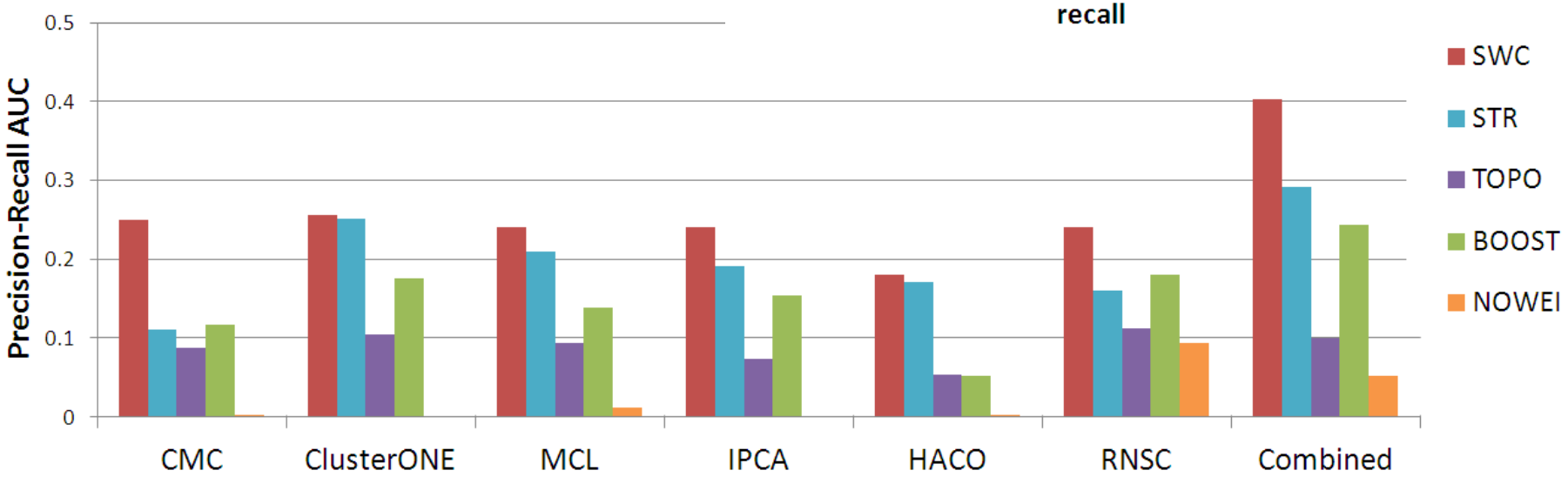
- **The clustering algo's generate clusters with low overlap**
  - Only 15% of clusters are generated by two or more algo's

⇒ **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**

NUS National University of Singapore

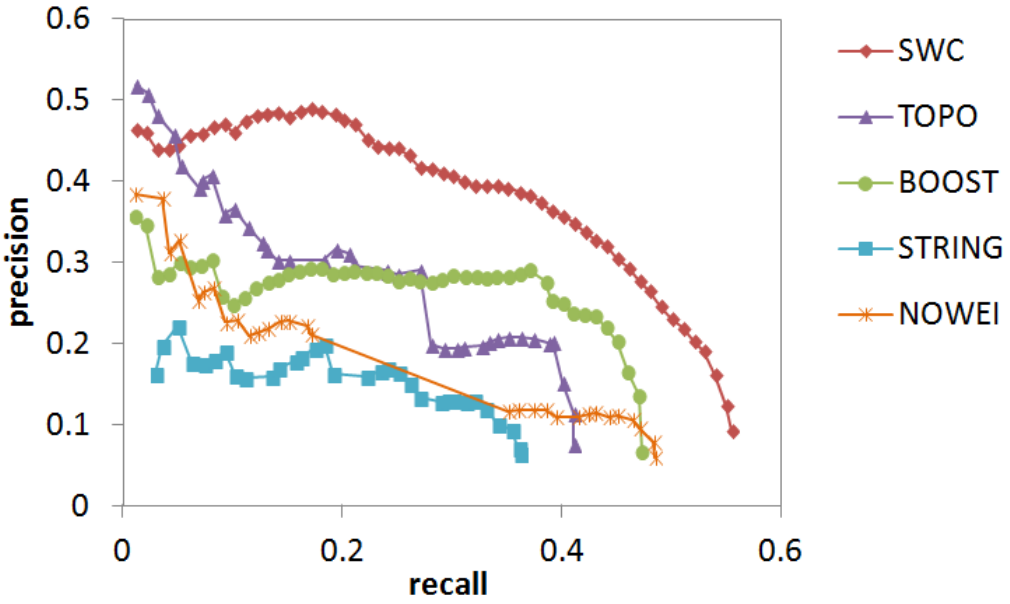
Evaluation wrt Yeast Complex Prediction

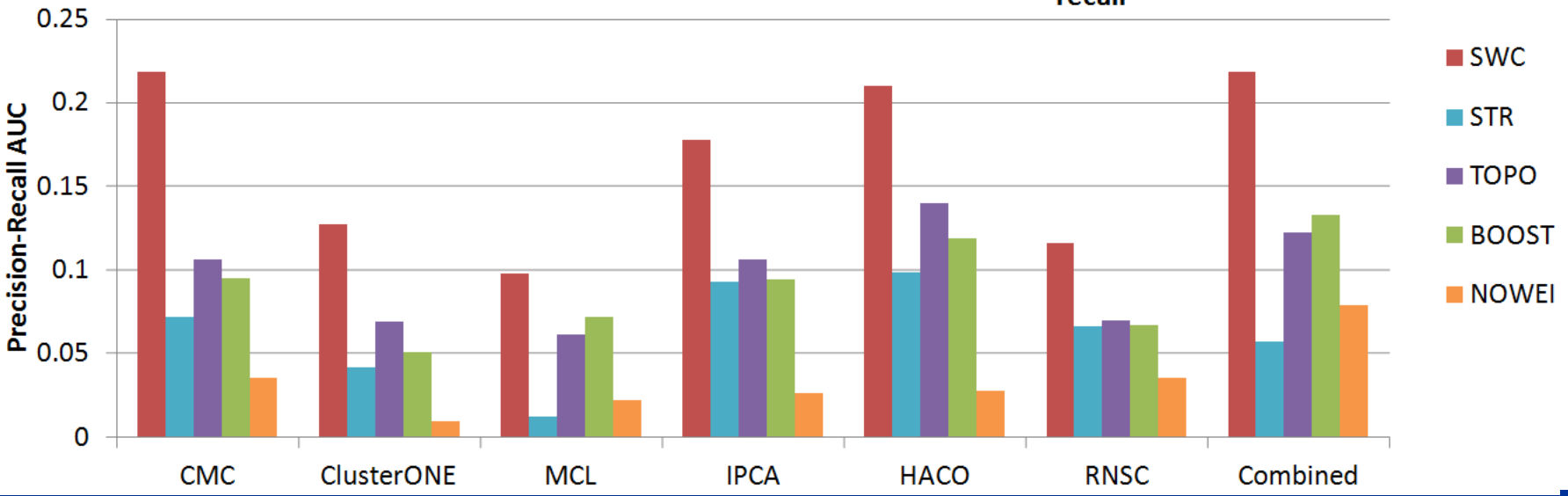




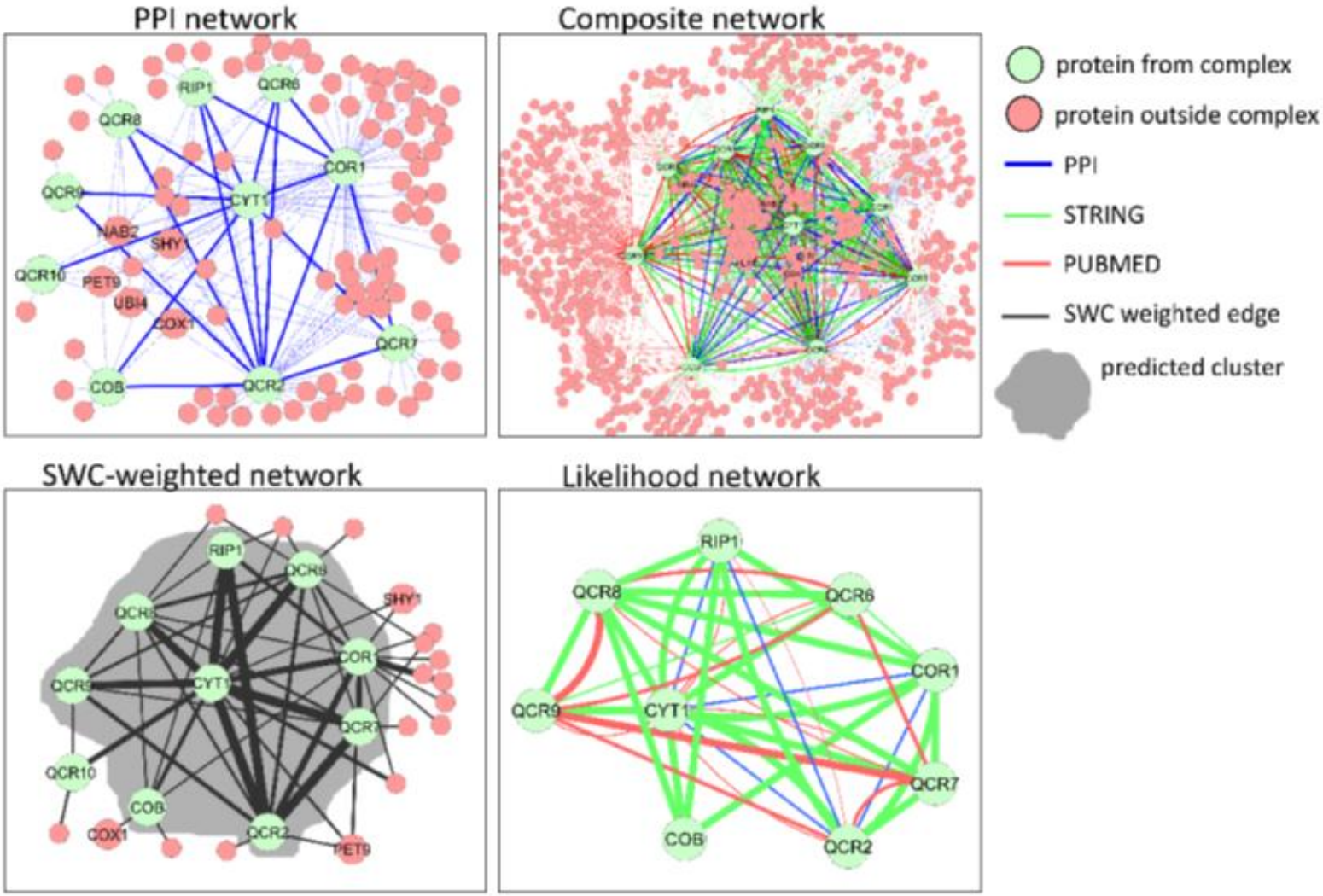
NUS National University of Singapore

Evaluation wrt Human Complex Prediction



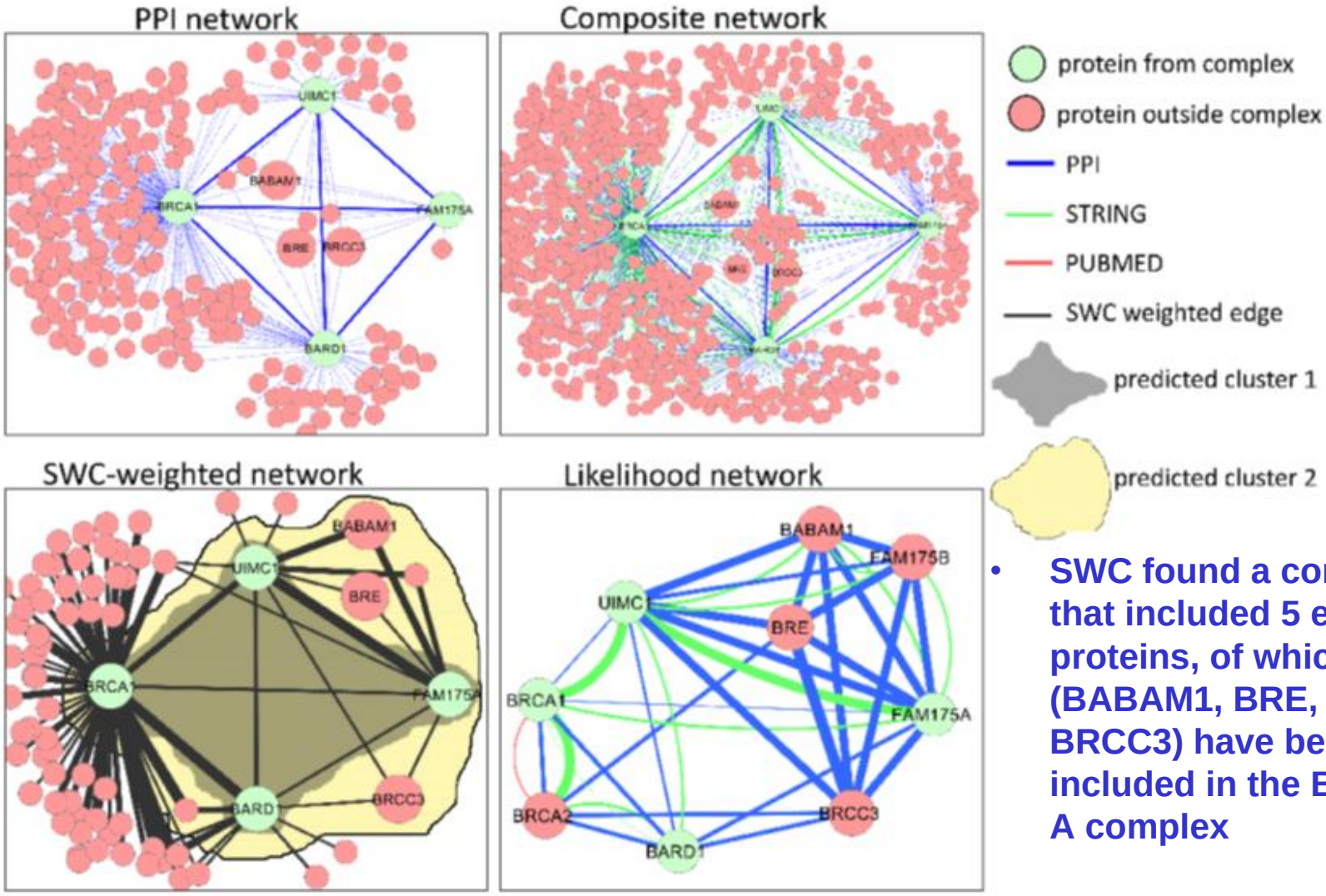


Example: Yeast BC1 Complex



NUS National University of Singapore

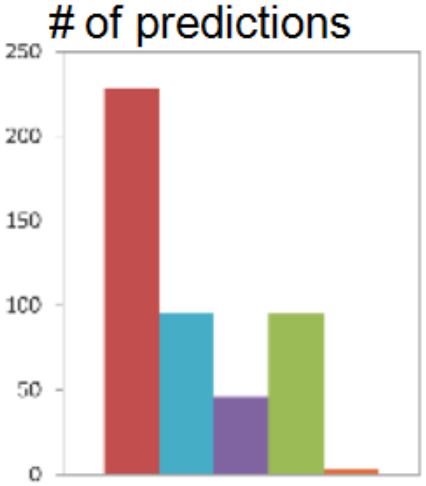
Example: Human BRCA1-A complex

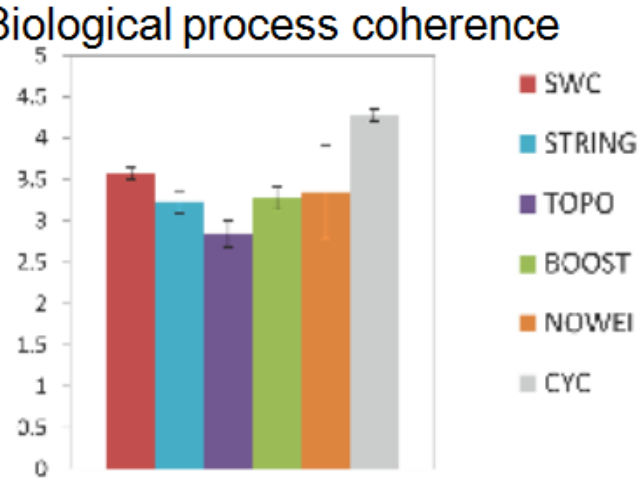


- **SWC found a complex that included 5 extra proteins, of which 3 (BABAM1, BRE, BRCC3) have been included in the BRCA1-A complex**

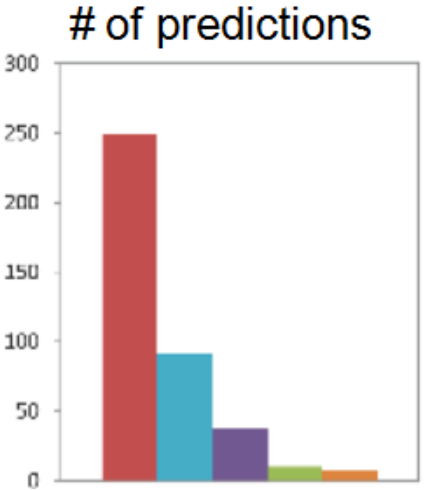
High-Confidence Predicted Complexes

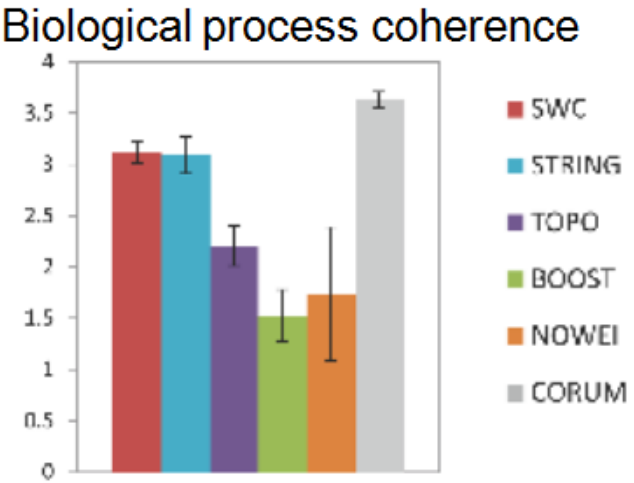
Yeast



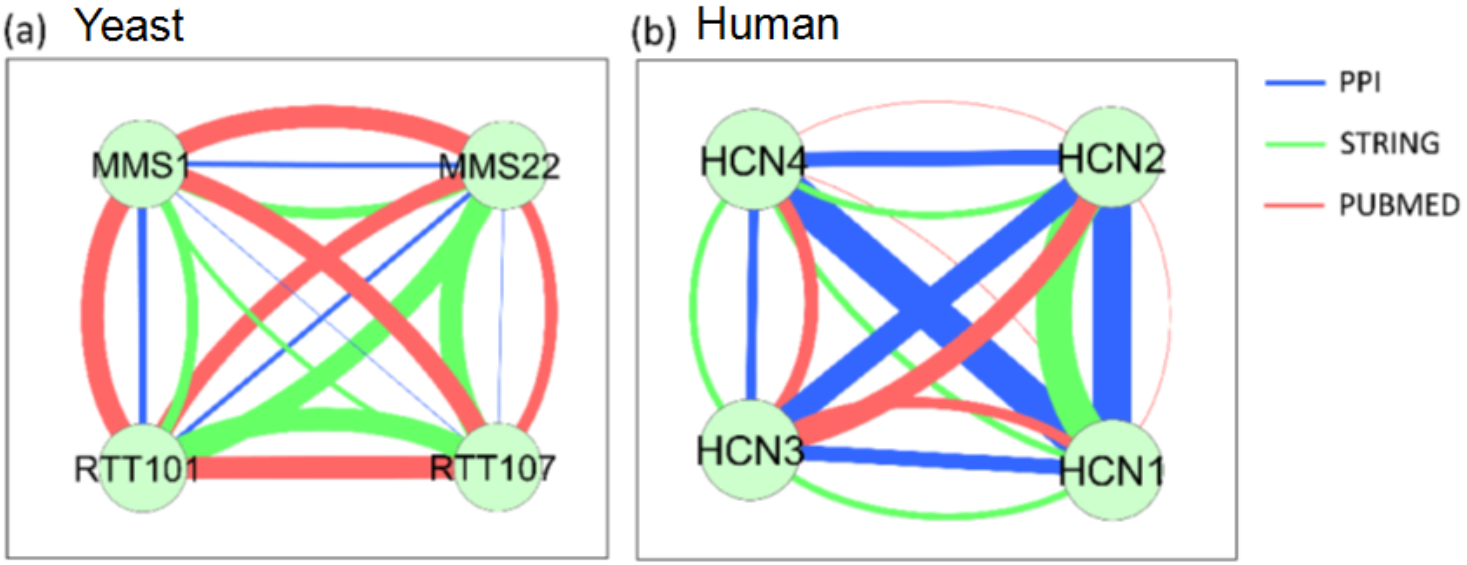


Human





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**

NUS National University of Singapore

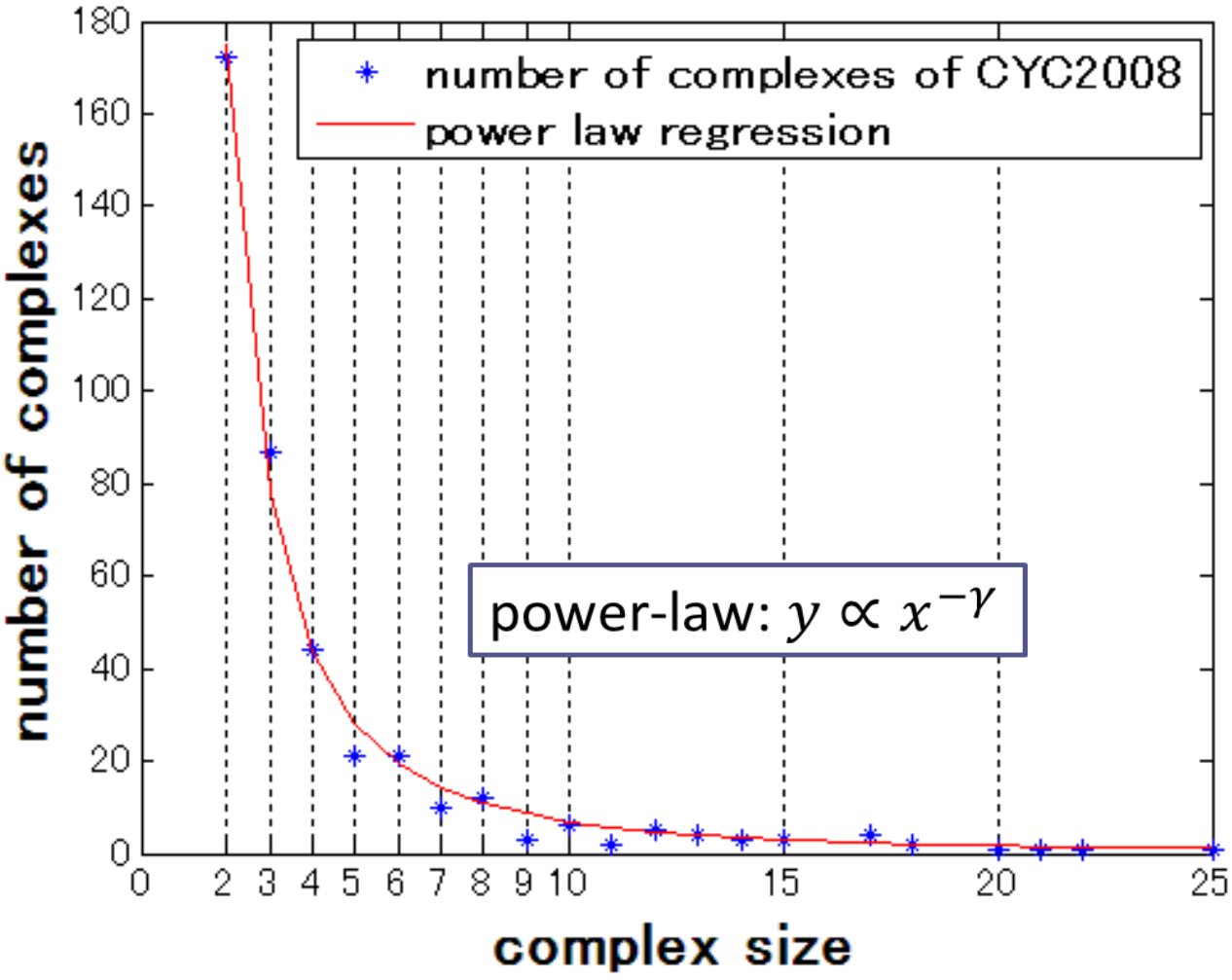
Conclusions

- **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

Remaining Challenge: Detecting Small Protein Complexes

There are many small complexes. Density-based methods cannot predict them from PPI networks



Source: Osamu Maruyama

Acknowledgements

Lui Guimei

Yong Chern Han

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