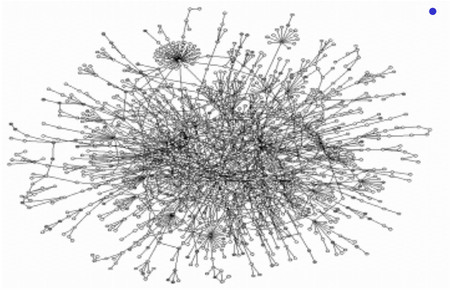
Identifying Protein Complexes from Protein Interactome Maps

Limsoon Wong

(Joint work with Kenny Chua & Guimei Liu)

Motivation

- **Nature of high-throughput PPI expts**
  - Proteins are taken out of their natural context!

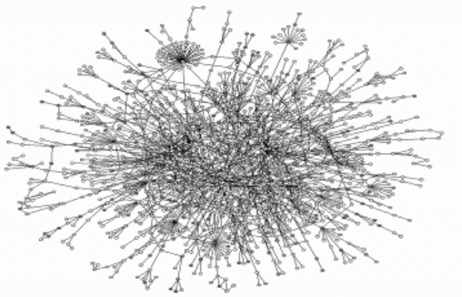


- **Can a protein interact with so many proteins simultaneously?**
- **A big "hub" and its "spokes" should probably be decomposed into subclusters**
  - Each subcluster is a set proteins that interact in the same space and time
  - Viz., a protein complex

Plan

- Motivation and Approaches
- **PPI Network Cleansing based on PPI Topology**
- **Impact of Cleansing on PPI-based Protein Complex Prediction Methods**

Approaches



Approaches to PPI-Based Protein Complex Prediction

	RNSC	MCODE	MCL
Type	Clustering, local search cost based	Local neighborhood density search	Flow simulation
Multiple assignment of protein	No	Yes	No
Weighted edge	No	No	Yes

- And several other methods….
- Recall vs precision is poor

Cause of Low Recall/Precision

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A1: GY2H Uetz *et al.* (unpublished results)	516	53	33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

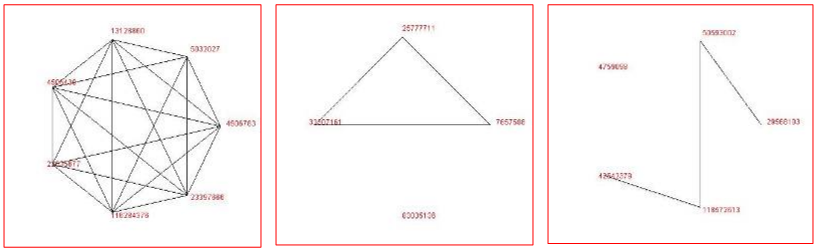
Sprinzak et al., *JMB*, 327:919-923, 2003

Large disagreement betw methods

- High level of noise

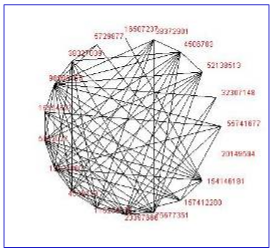
⇒ Need to clean up before protein complex prediction

Cause of Low Recall/Precision



- **Trouble with “non-ball-like” complexes**

⇒ **Clique merging? Relative density? Core-n-attachment?**



PPI Network Cleansing based on PPI Topology

Measures that correlate with function homogeneity and localization coherence

- Two proteins participating in same biological process are more likely to interact
- Two proteins in the same cellular compartments are more likely to interact

→

- **CD-distance**
- **FS-Weight**

CD-distance & FS-Weight: Based on concept that two proteins with many interaction partners in common are likely to be in same biological process & localize to the same compartment

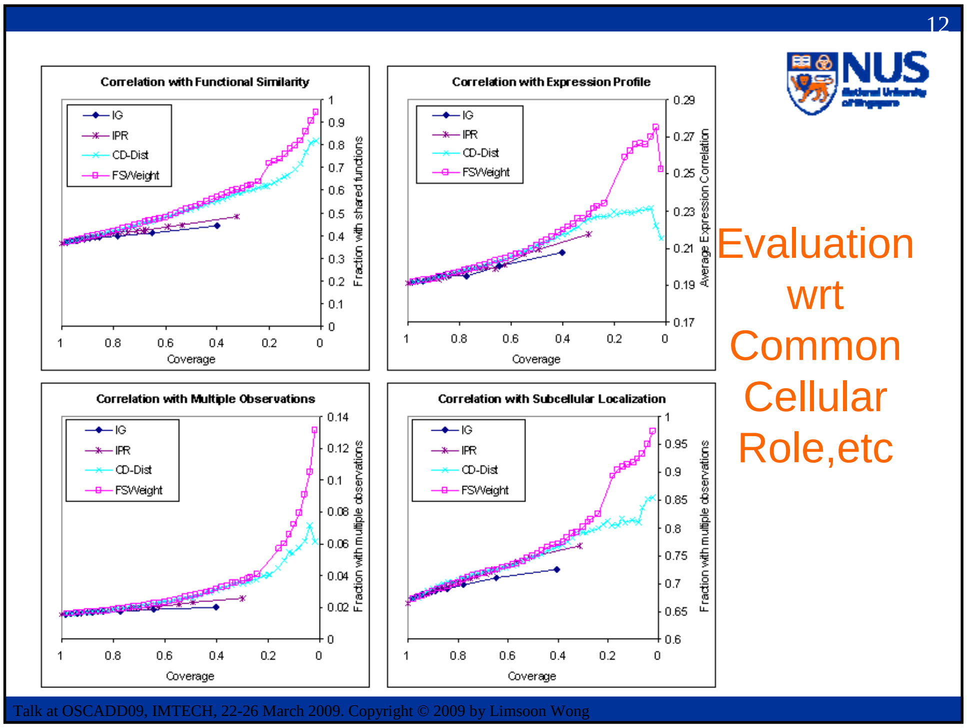
---

Czekanowski-Dice Distance (Brun et al, 2003)

- **Given a pair of proteins (u, v) in a PPI network**
  - N_u = the set of neighbors of u
  - N_v = the set of neighbors of v
- **CD(u,v) =** $$\frac{2\,|N_u \cap N_v|}{|N_u| + |N_v|}$$
- **Consider relative intersection size of the two neighbor sets, not absolute intersection size**
  - Case 1: $|N_u| = 1$, $|N_v| = 1$, $|N_u \cap N_v| = 1$, CD(u,v)=1
  - Case 2: $|N_u| = 10$, $|N_v| = 10$, $|N_u \cap N_v| = 10$, CD(u,v)=1

FSWeight (Chua et al, 2006)

- **Try to overcome weakness of CD-distance**
- **FS(u,v) =** $$\frac{2|N_u \cap N_v|}{|N_u| + |N_u \cap N_v| + \lambda_u} \times \frac{2|N_u \cap N_v|}{|N_v| + |N_u \cap N_v| + \lambda_v}$$
- **λ_u and λ_v penalize proteins with few neighbors**
  - $\lambda u = \max\{0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_u|\}$, $\lambda v = \max\{0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_v|\}$
- **Suppose average degree is 4, then**
  - Case 1: $|N_u| = 1$, $|N_v| = 1$, $|N_u \cap N_v| = 1$, FS(u,v)=4/25=0.16
  - Case 2: $|N_u| = 10$, $|N_v| = 10$, $|N_u \cap N_v| = 10$, FS(u,v)=1



Evaluation wrt Common Cellular Role,etc

A simpler formulation seems to work too…

Iterated CD-Distance (Liu et al, 2008)

- **Variant of CD-distance that penalizes proteins with few neighbors**

$$\textbf{wL(u,v)} = \frac{2|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

$$\lambda_u = \max\{0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_u|\},\ \lambda_v = \max\{0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_v|\}$$

- **Suppose average degree is 4, then**
  - Case 1: $|N_u| = 1$, $|N_v| = 1$, $|N_u \cap N_v| = 1$, wL(u,v)=0.25
  - Case 2: $|N_u| = 10$, $|N_v| = 10$, $|N_u \cap N_v| = 10$, wL(u,v)=1

A thought…

wL(u,v) = $\dfrac{2\,|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$

- **Weight of interaction reflects its reliability**

⇒ Can we get better results if we use this weight to re-calculate the score of other interactions?

Iterated CD-Distance (Liu et al, 2006)

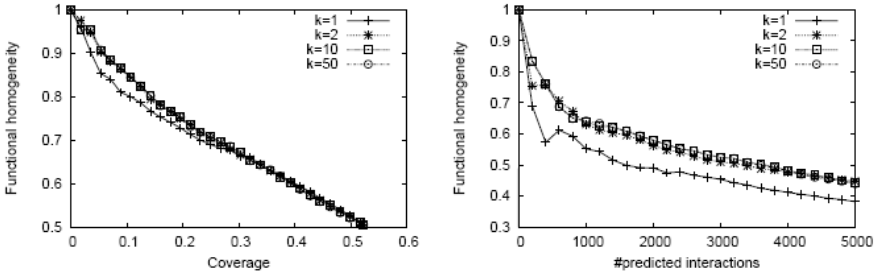
- **$wL^0(u,v)$ = 1 if $(u,v) \in G$, otherwise $wL^0(u,v)=0$**
- **$wL^1(u,v)$ =** $\dfrac{|N_u \cap N_v| + |N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$
- **$wL^k(u,v)$ =** $\dfrac{\sum_{x \in Nu \cap Nv} w_L^{k-1}(u,x) + \sum_{x \in Nu \cap Nv} w_L^{k-1}(v,x)}{\sum_{x \in Nu} w_L^{k-1}(u,x) + \lambda^k{}_u + \sum_{x \in Nv} w_L^{k-1}(v,x) + \lambda^k{}_v}$
- **$\lambda^k{}_u$ = max{0,** $\dfrac{\sum_{x \in V} \sum_{y \in Nx} w_L^{k-1}(x,y)}{|V|} - \sum_{x \in Nu} w_L^{k-1}(u,x)$ **}**
- **$\lambda^k{}_v$ = max{0,** $\dfrac{\sum_{x \in V} \sum_{y \in Nx} w_L^{k-1}(x,y)}{|V|} - \sum_{x \in Nv} w_L^{k-1}(v,x)$ **}**

Validation

- **DIP yeast dataset**
  - Functional homogeneity is 32.6% for PPIs where both proteins have functional annotations and 3.4% over all possible PPIs
  - Localization coherence is 54.7% for PPIs where both proteins have localization annotations and 4.9% over all possible PPIs
- **Let’s see how much better iterated CD-distance is over the baseline above, as well as over the original CD-distance/FS-weight**

How many iteration is enough?

Cf. ave functional homogeneity of protein pairs in DIP < 4%
ave functional homogeneity of PPI in DIP < 33%



- **Iterated CD-distance achieves best performance wrt functional homogeneity at k=2**
- **Ditto wrt localization coherence (not shown)**

How many iteration is enough?

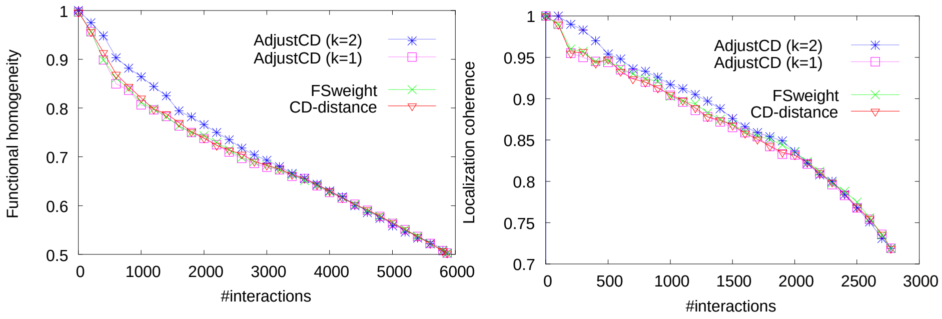
noise level	k	#common PPIs	avg_rank_diff	avg_score_diff
100%	1	5669	540.21	0.10
	2	5870	144.86	0.02
	20	5849	67.00	0.01
300%	1	5322	881.77	0.18
	2	5664	367.45	0.06
	20	5007	249.85	0.02
500%	1	5081	1013.14	0.23
	2	5502	625.46	0.12
	20	5008	317.33	0.05
1000%	k=1	4472	1187.10	0.28
	k=2	5101	1021.69	0.27
	k=20	5264	614.66	0.13

- **Iterative CD-distance at diff k values on noisy network**

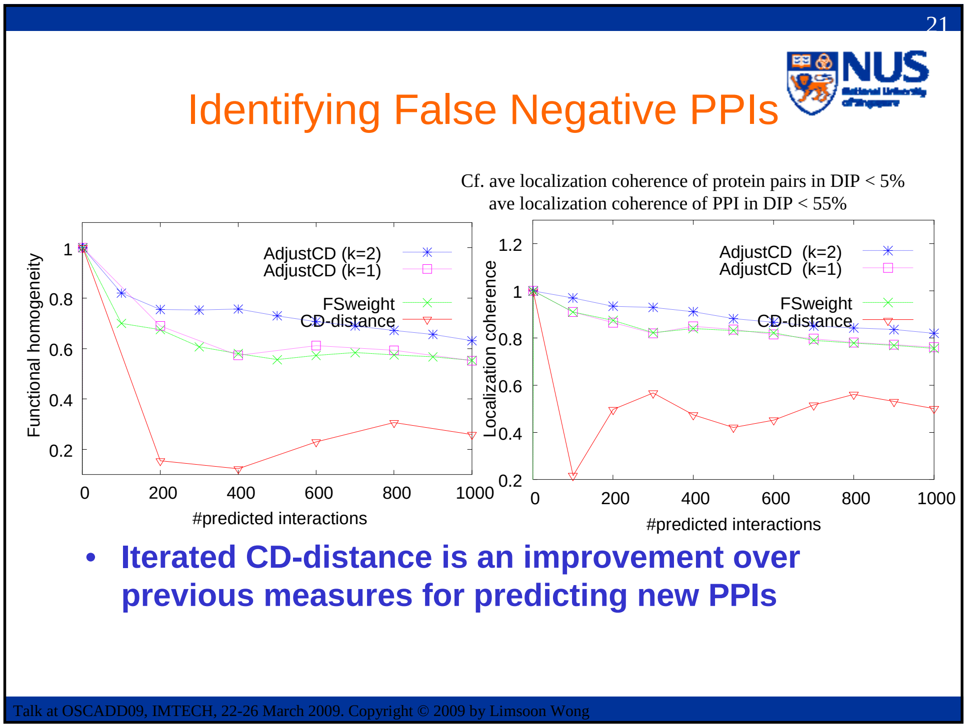
⇒ # of iterations depends on amt of noise

Identifying False Positive PPIs

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%



- **Iterated CD-distance is an improvement over previous measures for assessing PPI reliability**



Identifying False Negative PPIs

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%

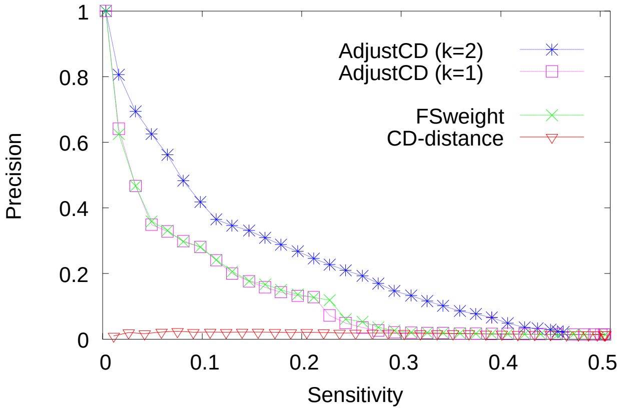
- **Iterated CD-distance is an improvement over previous measures for predicting new PPIs**

5-Fold Cross-Validation

- **DIP core dataset**
  - Ave # of proteins in 5 groups: 986
  - Ave # of interactions in 5 training datasets: 16723
  - Ave # of interactions in 5 testing datasets: 486591
  - Ave # of correct answer interactions: 307

- **Measures:**
  - sensitivity =TP/(TP + FN)
  - specificity =TN/(TN + FP)
    - **#negatives >> #positives, specificity is always high**
    - **>97.8% for all scoring methods**
  - precision =TP/(TP + FP)

5-Fold X-Validation



- **Iterated CD-distance is an improvement over previous measures for identifying false positive & false negative PPIs**

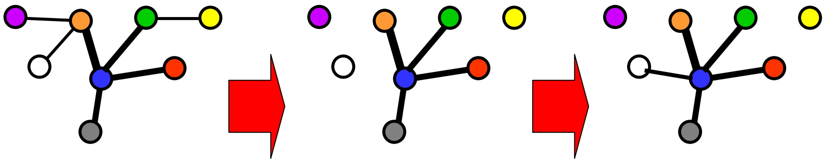
Impact of Cleansing on PPI-Based Protein Complex Prediction Methods

PPI-Based Complex Prediction Algo

	RNSC	MCODE	MCL
Type	Clustering, local search cost based	Local neighborhood density search	Flow simulation
Multiple assignment of protein	No	Yes	No
Weighted edge	No	No	Yes

- **Issue: recall vs precision has to be improved**
- **Does a “cleaner” PPI network help?**
- **How to capture non-ball-like complexes?**

Cleaning PPI Network



- **Modify existing PPI network as follow**
  - – Remove level-1 interactions with low weight
  - – Add level-2 interactions with high weight

- **Then run RNSC, MCODE, MCL, …, as well as our own method CMC**

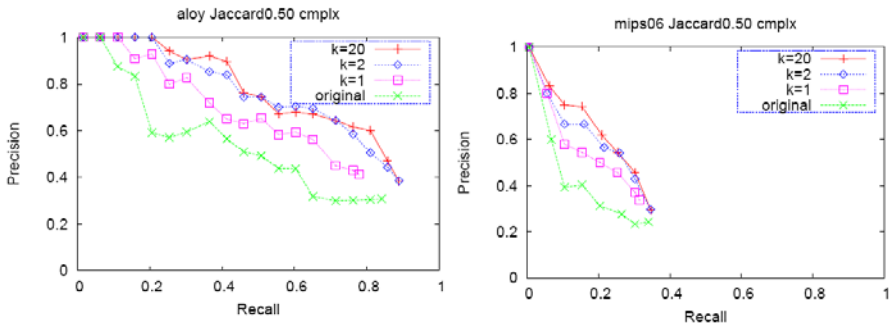
Clustering based on Maximal Cliques

- **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**

Validation Experiments

- **Matching a predicted complex S with a true complex C**
  - Vs: set of proteins in S
  - Vc: set of proteins in C
  - $\text{Overlap}(S, C) = |Vs \cap Vc| / |Vs \cup Vc|$
  - $\text{Overlap}(S, C) \geq 0.5$
- **Evaluation**
  - Precision = matched predictions / total predictions
  - Recall = matched complexes / total complexes
- **Datasets: combined info from 6 yeast PPI expts**
  - #interactions: 20461 PPI from 4671 proteins
  - #interactions with >0 common neighbor: 11487

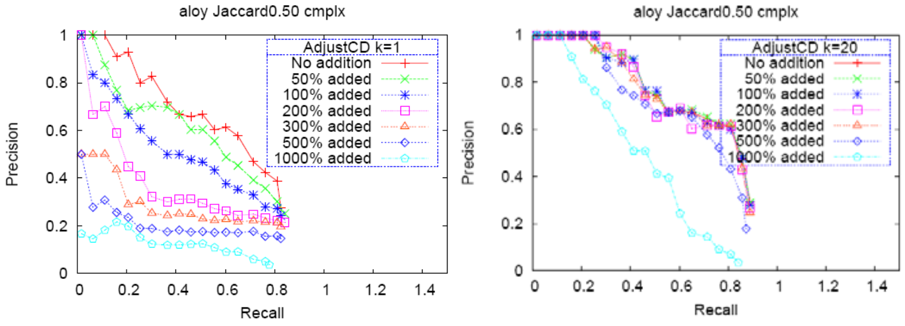
Effecting of Cleaning on CMC



(a) Aloy, *match_thres*=0.50 (b) MIPS, *match_thres*=0.50

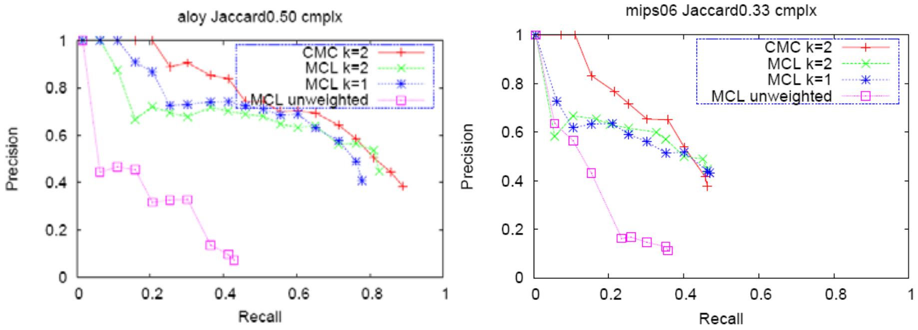
- **Cleaning by Iterated CD-distance improves recall & precision of CMC**

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 protein complex prediction methods**

CMC vs Others

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

What have we learned?

- **Guilt by association of common interaction partners is useful for predicting**
  - PPI cellular localization
  - Missing PPIs
  - Protein complexes

- **Acknowledgement**
  - Kenny Chua, Guimei Liu

Readings

- H.N. Chua, et al. “Using Indirect Protein-Protein Interactions for Protein Complex Prediction”, *Journal of Bioinformatics and Computational Biology*, 6(3):435--466, 2008
- H. N. Chua, L. Wong. “Increasing the Reliability of Protein Interactomes”, *Drug Discovery Today*, 13(15/16):652--658, 2008
- G. Liu, J. Li, L. Wong. “Assessing and predicting protein interactions using both local and global network topological metrics”, *Proc GIW2008*
- G. Liu, L. Wong, H. N. Chua. “Complex Discovery from Weighted PPI Networks”, submitted.

Any Question?