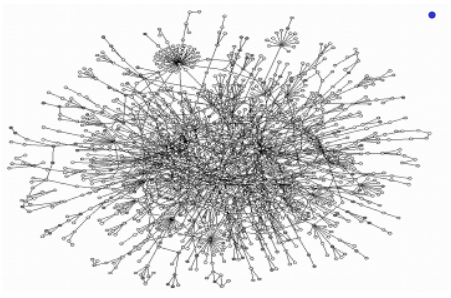
Identifying Protein Complexes from Protein Interactome Maps

Limsoon Wong

(Joint work with Kenny Chua & Guimei Liu)

Motivation

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

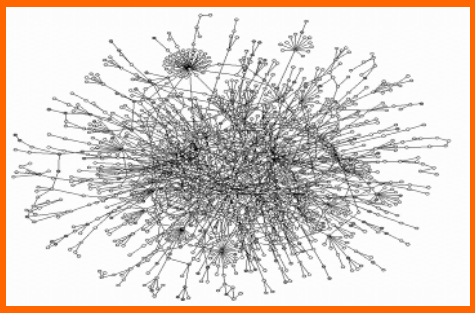


- **Can a protein interact with so many proteins simultaneously?**
- **A big "hub" and its "spokes" should probably be decomposed into subclusters**
  - Each subcluster is a set proteins that interact in the same space and time
  - Viz., a protein complex

Plan

- **Motivation and Approaches**
- **PPI Network Cleansing based on PPI Topology**
  - CD-Distance, FS-Weight
- **Impact of Cleansing on PPI-based Protein Complex Prediction Methods**
- **Recent Improvement to PPI Network Cleansing and PPI-Based Protein Complex Prediction**

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

- **Recall vs precision is poor**
  - Noise in PPI network?
  - Non-ball-like complexes?

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A1: GY2H Uetz *et al.* (unpublished results)	516	53	33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

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**

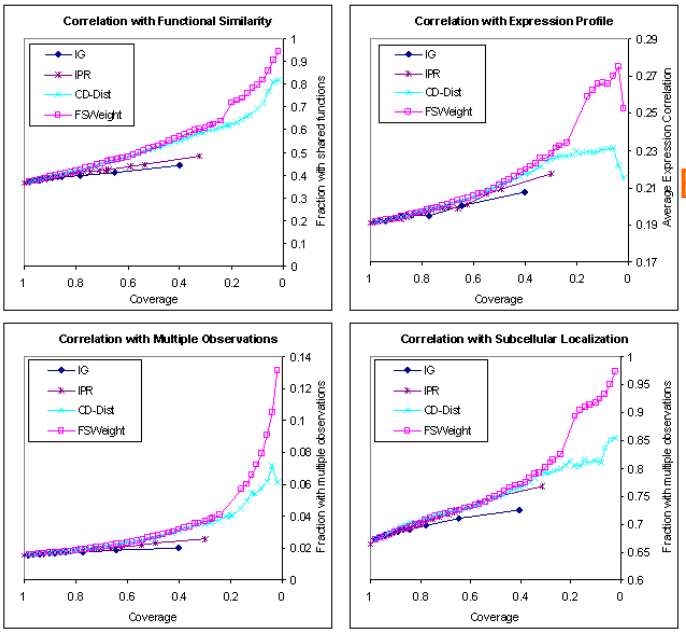
CD-distance & FS-Weight: Based on concept that two proteins with many interaction partners in common are likely to be in same biological process & localize to the same compartment

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) =** $\dfrac{2 \mid N_u \cap N_v \mid}{\mid N_u \mid + \mid N_v \mid}$
- **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

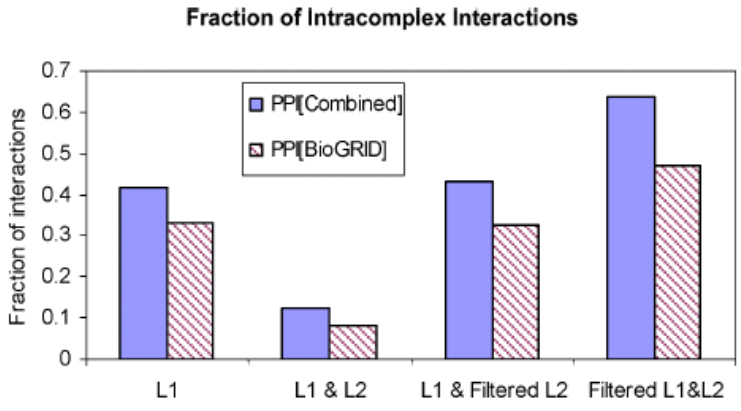
FSWeight (Chua et al, 2006)

- **Try to overcome weakness of CD-distance**
- **FS(u,v) =** $\dfrac{2 \mid N_u \cap N_v \mid}{\mid N_u \mid + \mid N_u \cap N_v \mid + \lambda_u} \times \dfrac{2 \mid N_u \cap N_v \mid}{\mid N_v \mid + \mid N_u \cap N_v \mid + \lambda_v}$
- **λ_u and λ_v penalize proteins with few neighbors**
  - $\lambda u = \max\{0, \dfrac{\sum_{x \in G} \mid N_x \mid}{\mid V \mid} - \mid N_u \mid\}$, $\lambda v = \max\{0, \dfrac{\sum_{x \in G} \mid N_x \mid}{\mid V \mid} - \mid N_v \mid\}$
- **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

Evaluation wrt Intracomplex Interactions



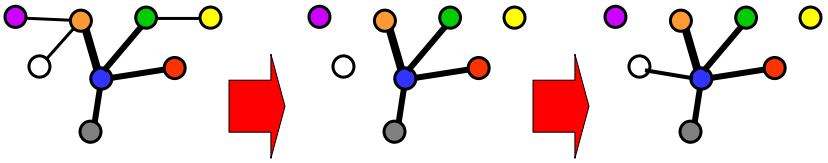
Impact of Cleansing on PPI-Based Protein Complex Prediction Methods

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 by FS-Weight



- **Modify existing PPI network as follow**
  - Remove level-1 interactions with low FS-weight
  - Add level-2 interactions with high FS-weight
- **Then run RNSC, MCODE, MCL, & PCP**

PCP Algorithm: Dealing w/ Non-Ball-like Complexes?

- **Find all max cliques in the modified PPI network**
- **Merge resulting (partial) cliques with good inter-cluster density**

$$ICD(S_a, S_b) = \frac{\sum \mathrm{FS}(i,j) \mid i \in (V_a - V_b), j \in (V_b - V_a), (i,j) \in E}{|V_a - V_b| \cdot |V_b - V_a|}$$

- **Modify the PPI network by treating the merged partial cliques as vertices**
- **Iterate the steps above**

Chua et al, *JBCB*, 6:435-466, 2008

Experiments

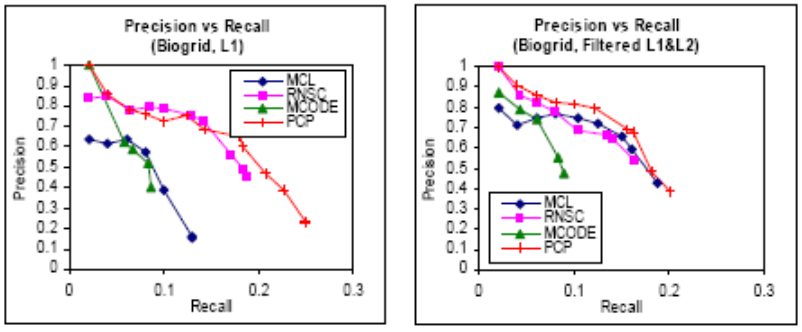
- **PPI datasets**
  - PPI[BioGRID], BioGRID db from Stark et al., 2006
- **Gold standards**
  - PC_{2004}, Protein complexes from MIPS 03/30/2004
  - PC_{2006}, Protein complexes from MIPS 05/18/2006
- **Validation criteria**

$$overlap(S,C) = \frac{|V_S \cap V_C|^2}{|V_S| \cdot |V_C|}$$

where

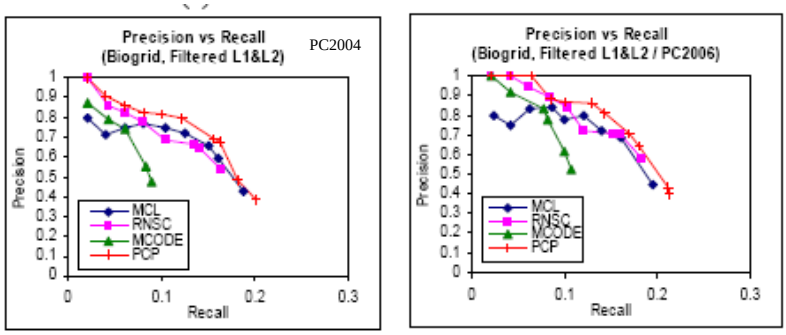
  - S = predicted cluster
  - C = true complex
  - V_x = vertices of subgraph defined by X
- **Overlap(S,C) ≥ 0.25 is considered a correct prediction**

Validation on PC_{2004}

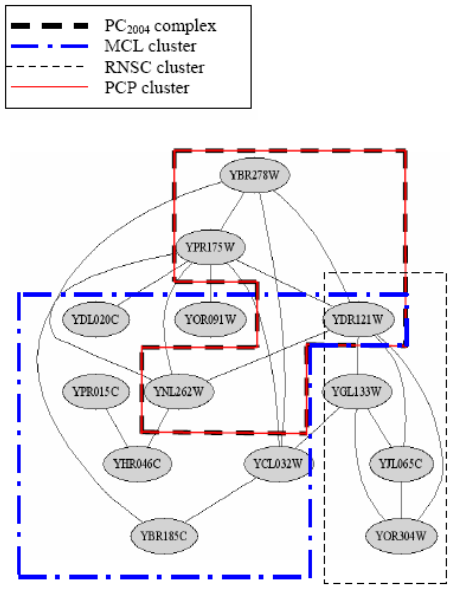


- **Precision is improved in all methods**

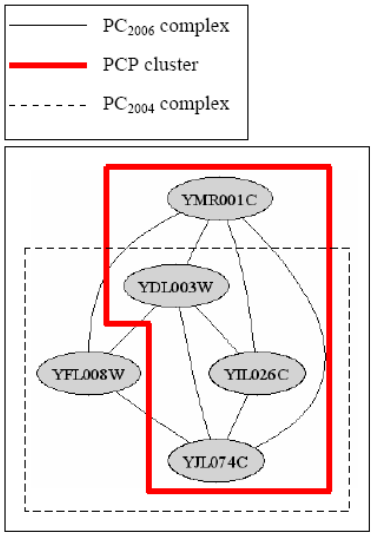
Validation on PC_{2006}



- **When predictions are validated against PC_{2006}, precision of all algo improved**
- **Many "false positives" wrt PC_{2004} are actually real**



PCP Prediction Example 1



PCP Prediction Example 2

Conclusions

- **Precision of protein complex prediction can be improved by**
  - PPI network augmented with level-2 interactions
  - PPI network cleansed by FS-weight
- **Clique merging may capture more possibly non-ball-like complexes**

Recent Improvement to PPI Network Cleansing & PPI-Based Protein Complex Prediction

Local Topological Metric

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

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

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

- **Iterate local topological metric**
  - Weight of interaction reflects its reliability
  - Can get better results if we use this weight to re-calculate the score of other interactions?

Iterate Local Topological Metric

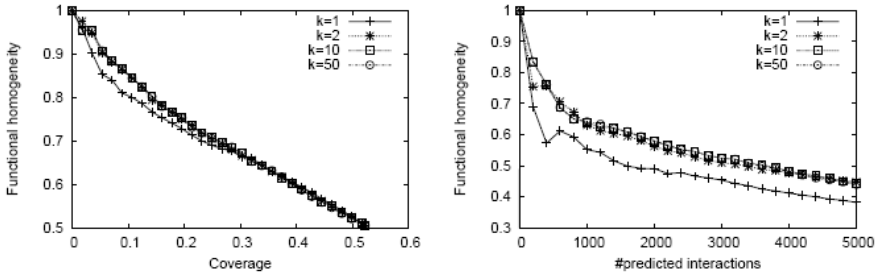
- $wL^0(u,v) = 1$ if $(u,v) \in G$, otherwise $wL^0(u,v)=0$
- $wL^1(u,v) = \dfrac{|N_u \cap N_v| + |N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$
- $wL^k(u,v) = \dfrac{\sum_{x \in Nu \cap Nv} w_L^{k-1}(u,x) + \sum_{x \in Nu \cap Nv} w_L^{k-1}(v,x)}{\sum_{x \in Nu} w_L^{k-1}(u,x) + \lambda^k{}_u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

Validation of Iterated CD-Distance

- **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?

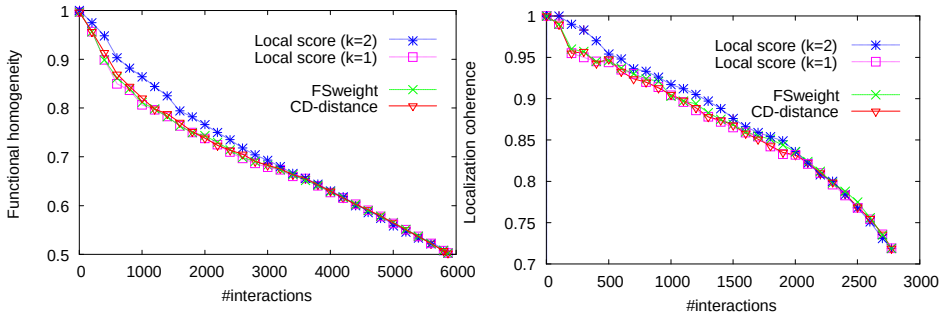
Cf. ave functional homogeneity of protein pairs in DIP < 4%
ave functional homogeneity of PPI in DIP < 33%



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

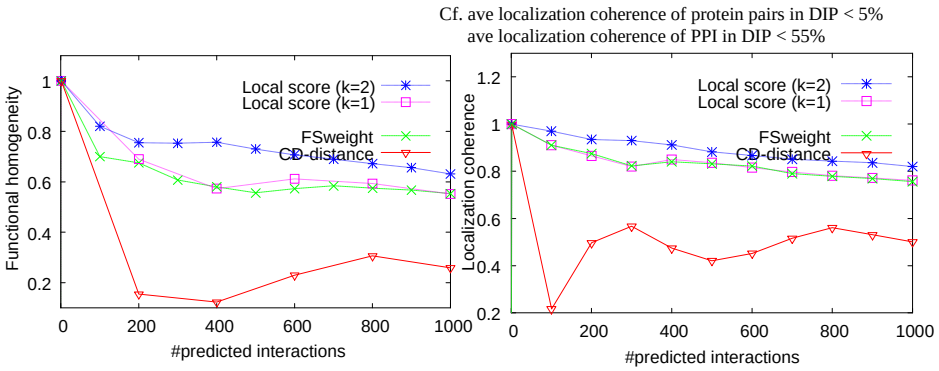
Identifying False Positive PPIs

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

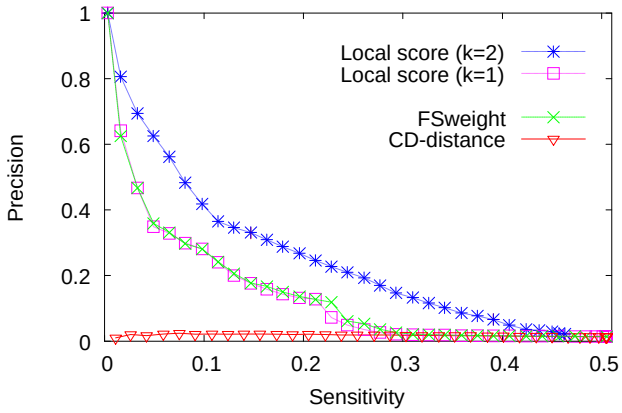


- **Iterated CD-distance is an improvement over previous measures for predicting new PPIs**

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**

Now we can make protein complex prediction as follows…

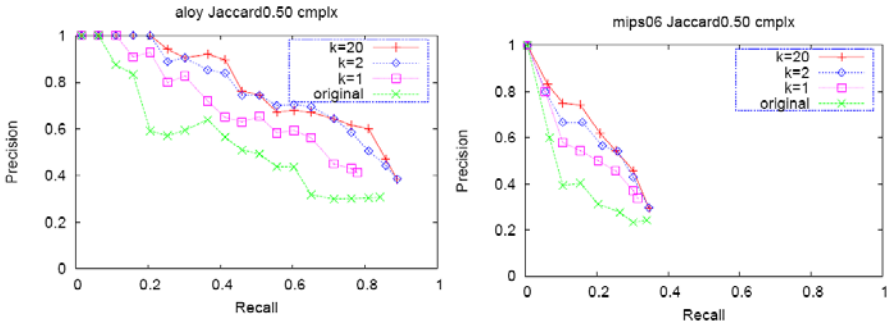
- **Remove noise edges in the input PPI network by discarding edges having low iterated CD-distance score**
- **Augment the input PPI network by addition of missing edges having high iterated CD-distance score**
- **Predict protein complex by finding and merging maximal cliques**

CMC --- Clustering based on Maximal Cliques

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 ∩Vc| /|Vs∪Vc|
  - Overlap(S, C) ≥ 0. 5
- **Evaluation**
  - Precision = matched predictions / total predictions
  - Recall = matched complexes / total complexes
- **Datasets: BioGrid yeast**
  - #interactions: 38555
  - #interactions with >0 common neighbor: 27940

Effecting of Cleaning on CMC

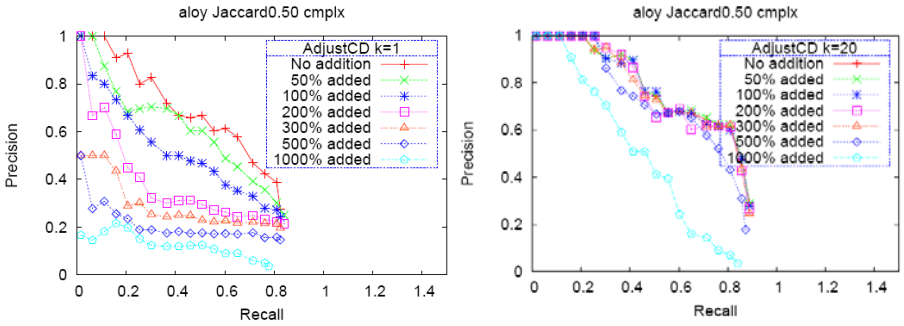


(a) Aloy, *match_thres*=0.50

(b) MIPS, *match_thres*=0.50

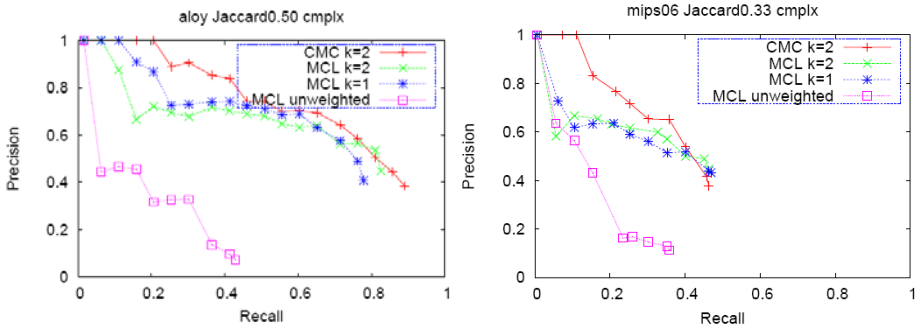
- **Evaluated based on protein complexes from Aloy and MIPS**

Noise Tolerance of CMC



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

CMC vs MCL



- **MCL benefits significantly from cleaning too**
- **MCL is not as good as CMC**

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**

- **G. Liu, J. Li, L. Wong. "Assessing and predicting protein interactions using both local and global network topological metrics", *Proc GIW2008***

Any Question?