Supervised Maximum Likelihood Weighting of Composite Protein Networks for Protein Complex Prediction

Limsoon Wong

NUS
National University
of Singapore

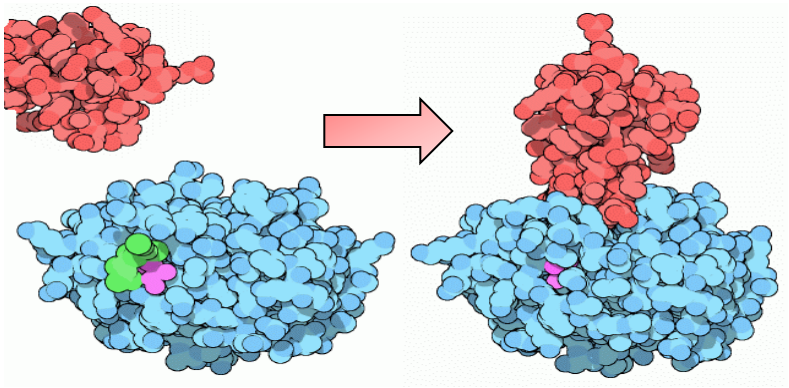
Outline

- **Background**
- **Supervised Weighting of Composite Networks**
  - Data integration
  - Supervised edge weighting
  - Clustering
- **Results**
  - Prediction accuracy
  - Semantic coherence
  - Examples
- **Conclusion**

Copyright 2013 © Limsoon Wong

Source: Sriganesh Srihari

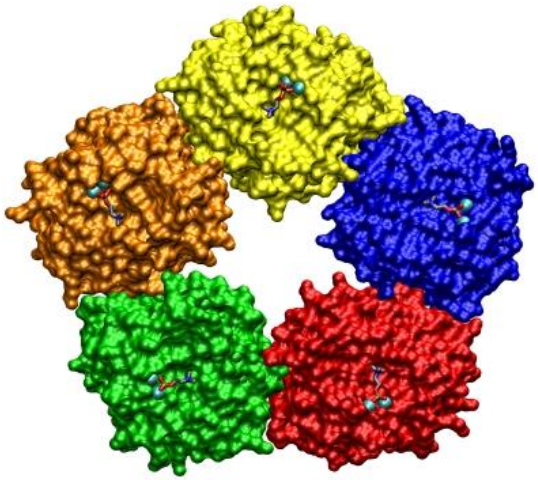
"Assemblies" of Interacting Proteins



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

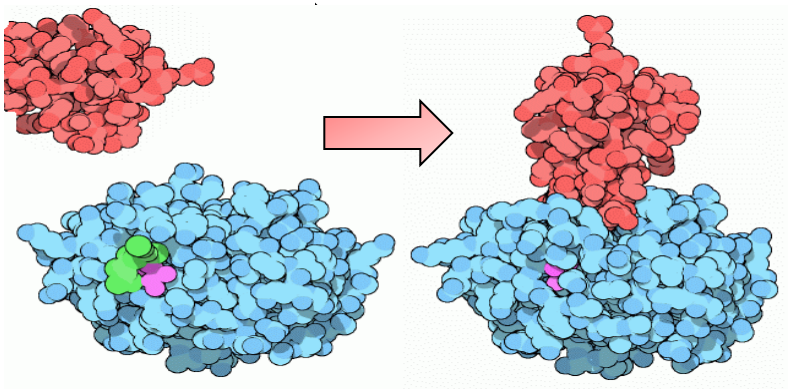


Protein assembly of multiple proteins

Copyright 2013 © Limsoon Wong

Source: Sriganesh Srihari

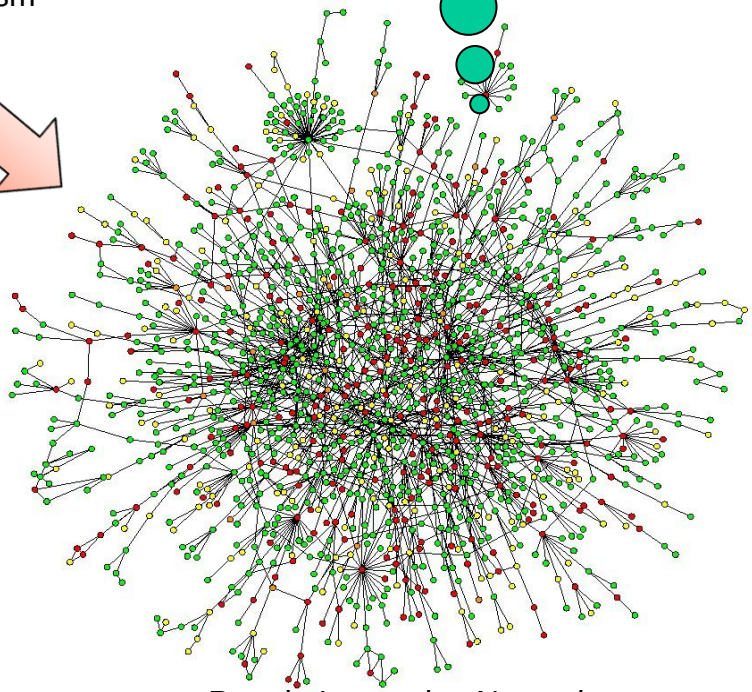
Protein Interaction Networks



Individual proteins come together and interact

Collection of such interactions in an organism

PPIN
Valuable source of knowledge



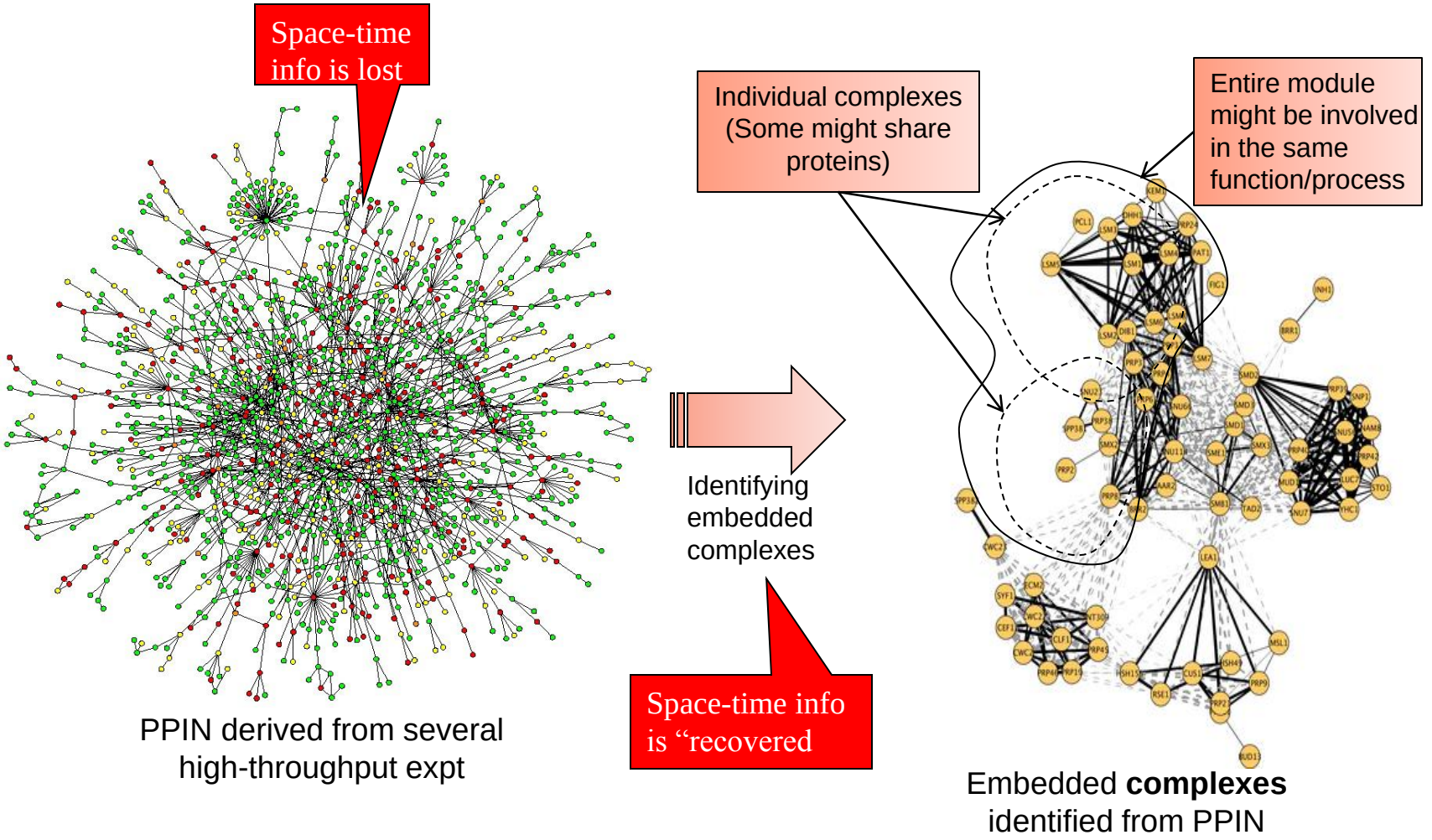
Protein Interaction Network

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

Copyright 2013 © Limsoon Wong

Source: Sriganesh Srihari

Detection & Analysis of Protein Complexes in PPIN



Copyright 2013 © Limsoon Wong

Difficulties

- **Typical complex discovery method: Predict dense subgraphs in PPIN as complexes**

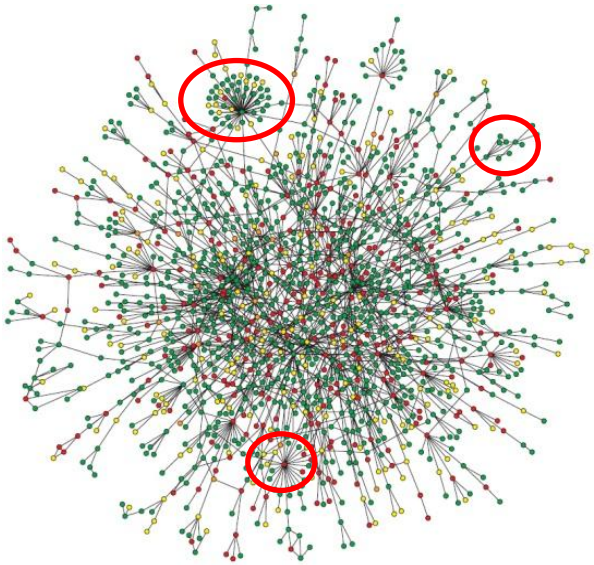


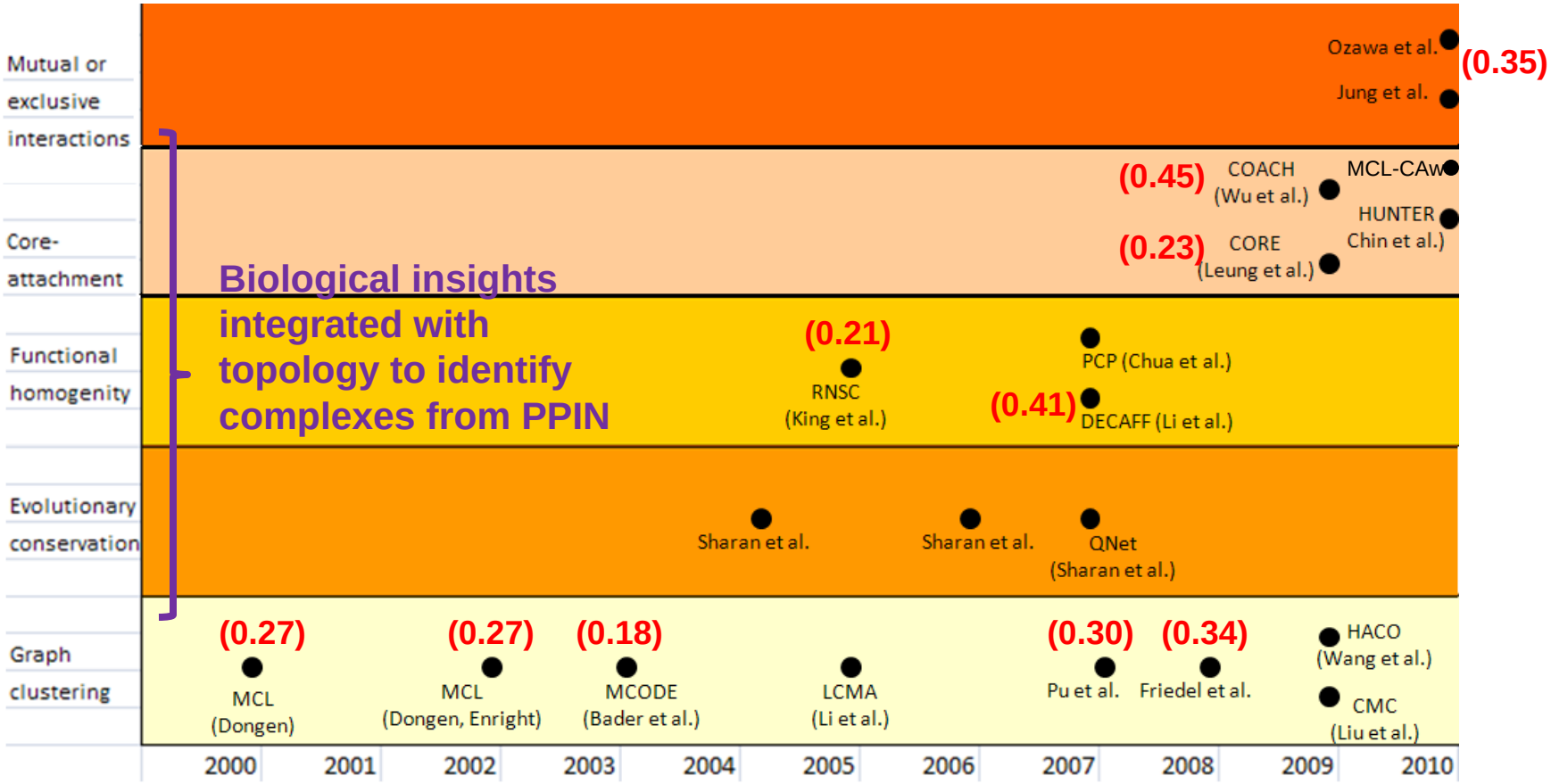
Image Source: Barabási and Oltvai, 2004

- **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
- **Not all proteins in the same complex may actually interact with each other**

Copyright 2013 © Limsoon Wong

Source: Sriganesh Srihari

Protein Complex Prediction Methods

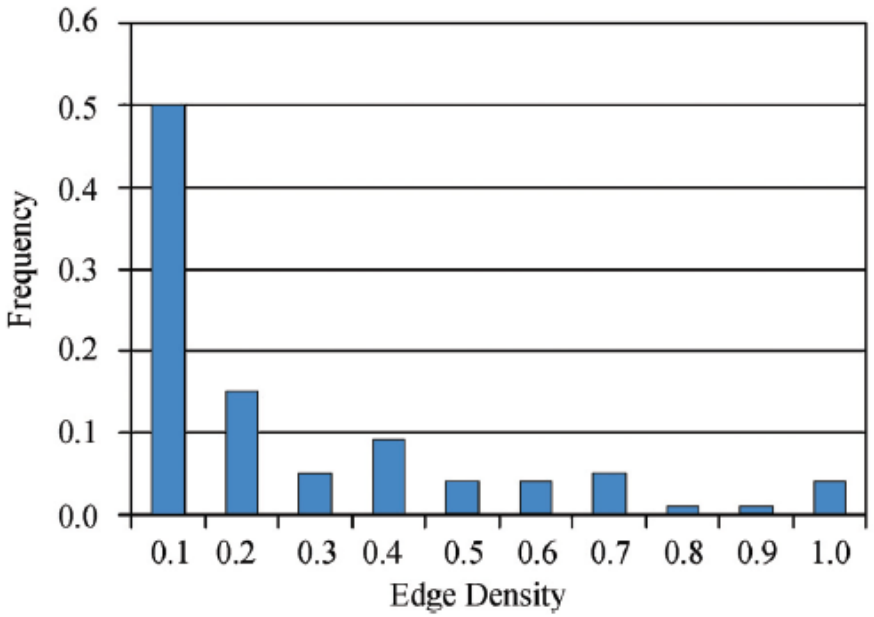


- **Adding biological info improves F1**

Copyright 2013 © Limsoon Wong

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?**

Copyright 2013 © Limsoon Wong

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.**

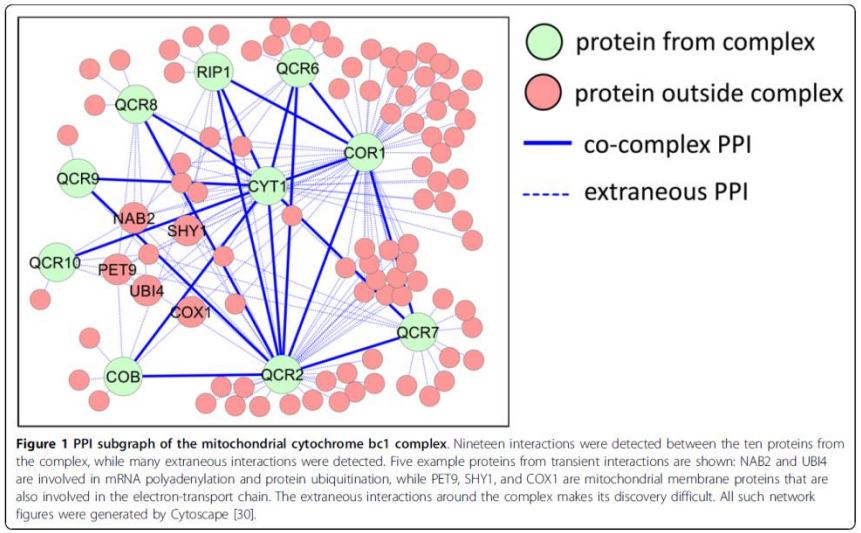


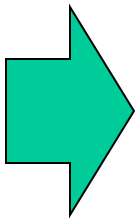
Figure 1 PPI subgraph of the mitochondrial cytochrome bc1 complex. Nineteen interactions were detected between the ten proteins from the complex, while many extraneous interactions were detected. Five example proteins from transient interactions are shown: NAB2 and UBI4 are involved in mRNA polyadenylation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

Copyright 2013 © Limsoon Wong

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**

Copyright 2013 © Limsoon Wong

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

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**

Copyright 2013 © Limsoon Wong

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%

Copyright 2013 © Limsoon Wong

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:

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

Copyright 2013 © Limsoon Wong

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**

Copyright 2013 © Limsoon Wong

NUS National University of Singapore

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

Copyright 2013 © Limsoon Wong

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

Evaluation wrt Co-Complex Prediction

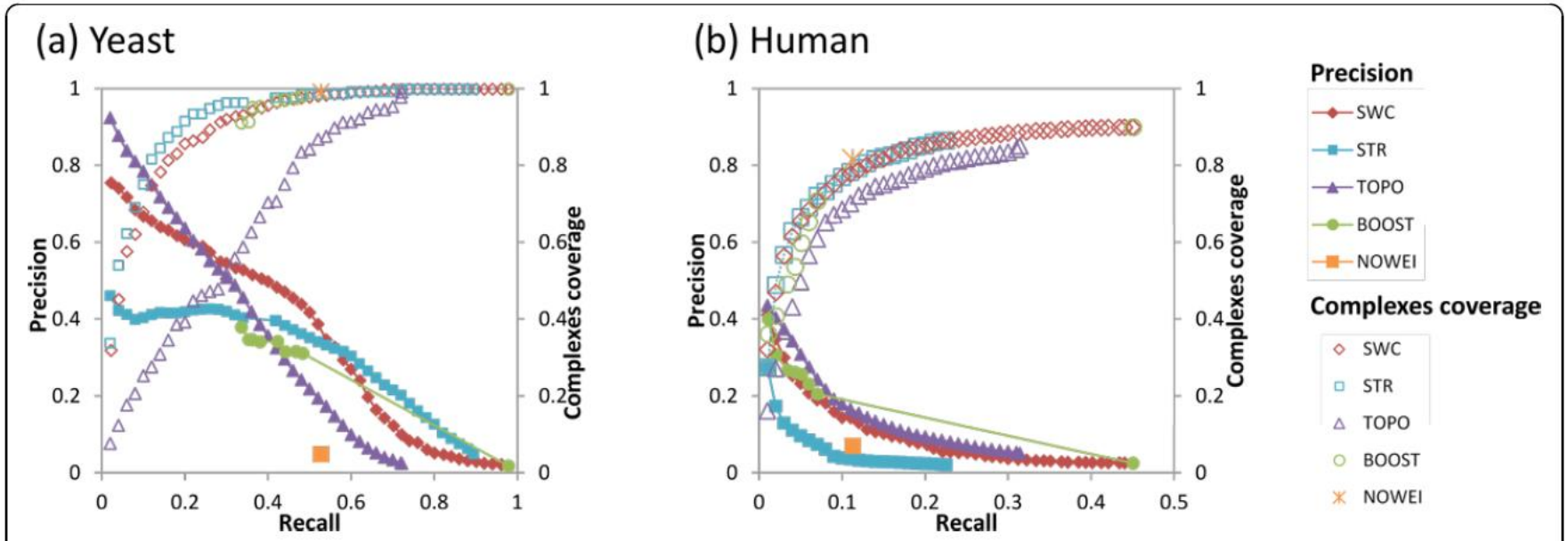
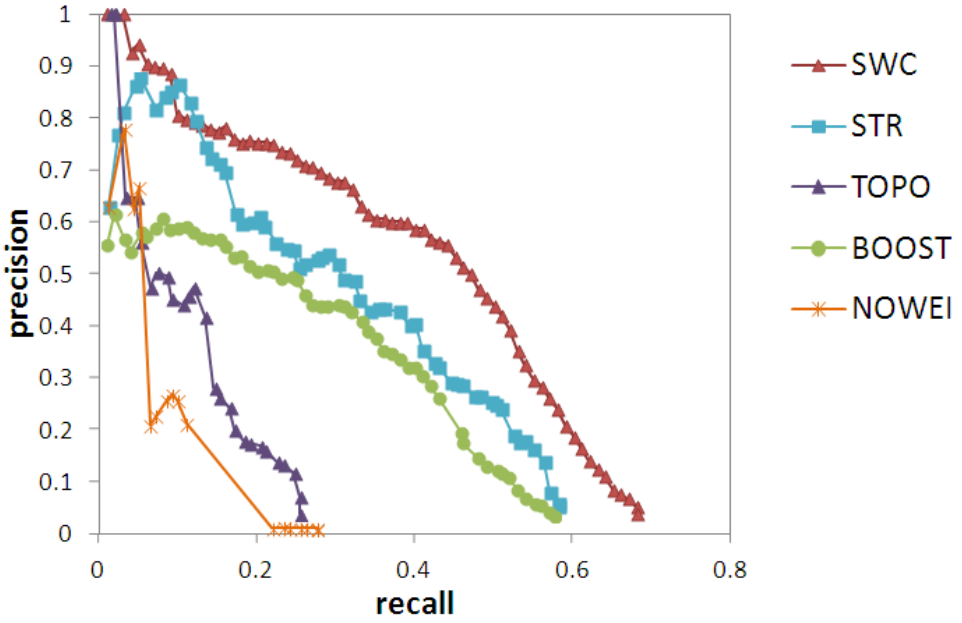
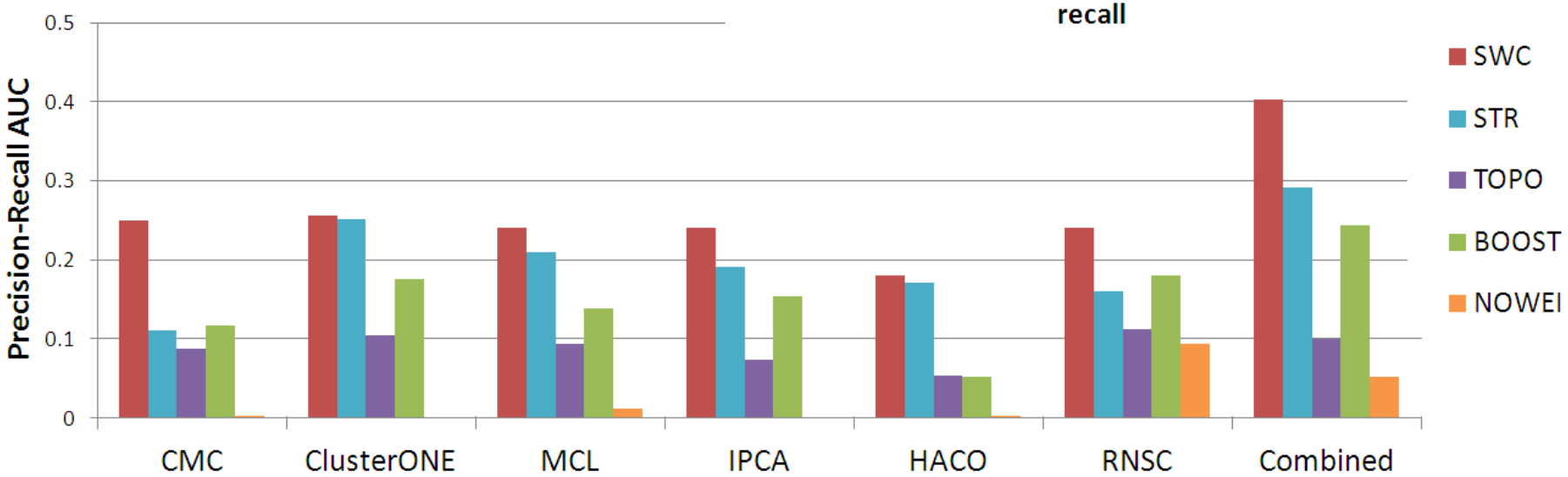


Figure 2 Precision-recall graph for classification of co-complex edges using the five weighting schemes. (a) Classification of yeast co-complex edges. SWC and BOOST achieve the highest recall through data integration. TOPO has high precision for its top-scoring edges, but these are clustered in a few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

Copyright 2013 © Limsoon Wong

Evaluation wrt Yeast Complex Prediction

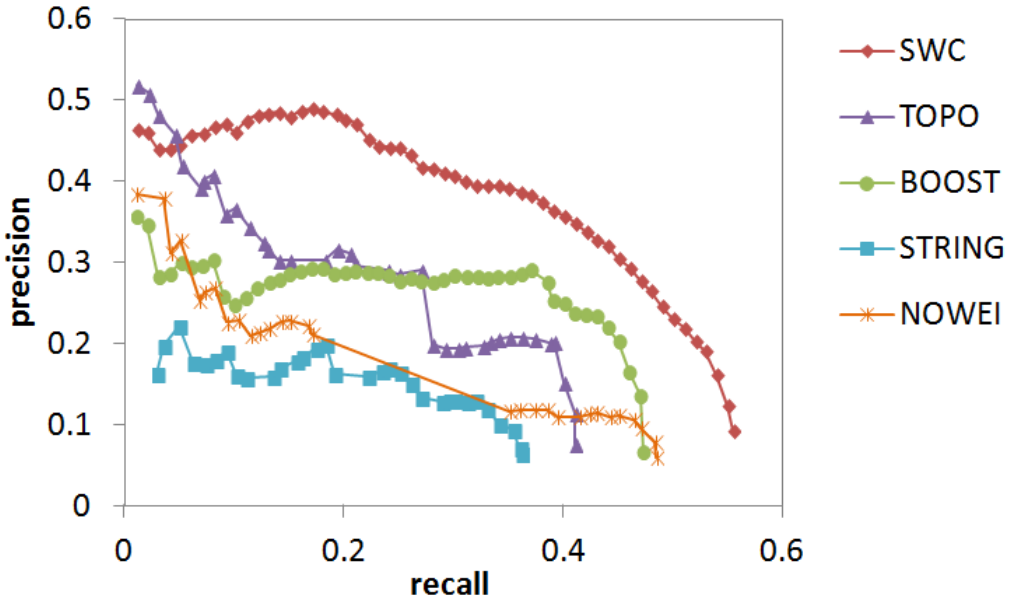


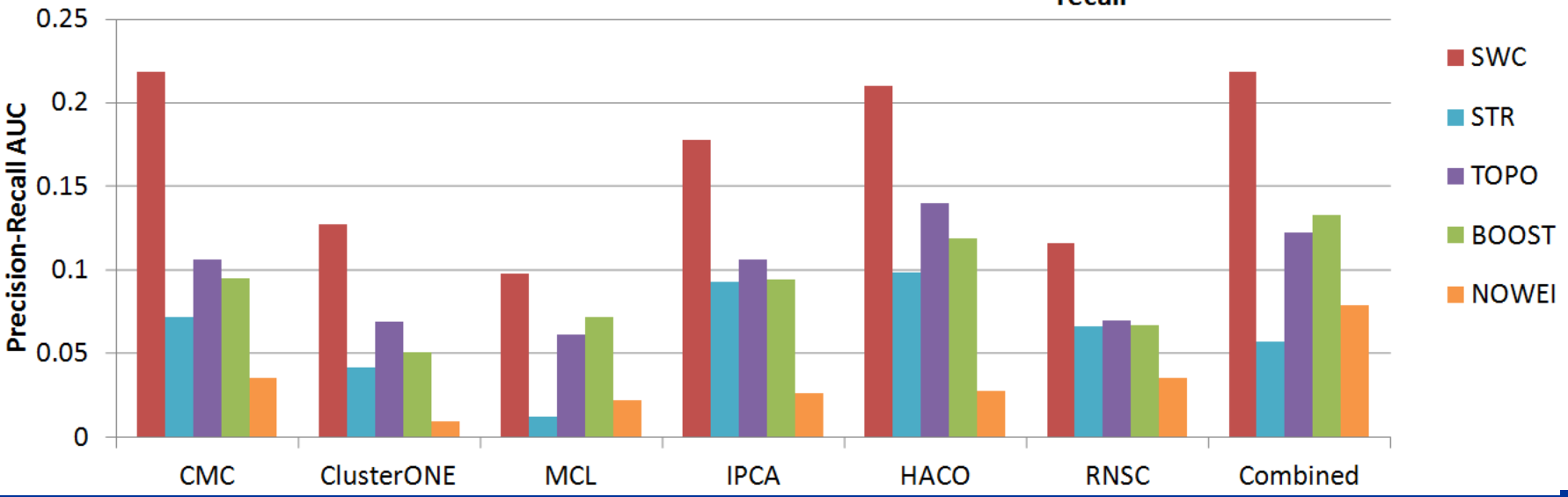


Copyright 2013 © Limsoon Wong

NUS National University of Singapore

Evaluation wrt Human Complex Prediction





Copyright 2013 © Limsoon Wong

Why the "COMBINED" Strategy?

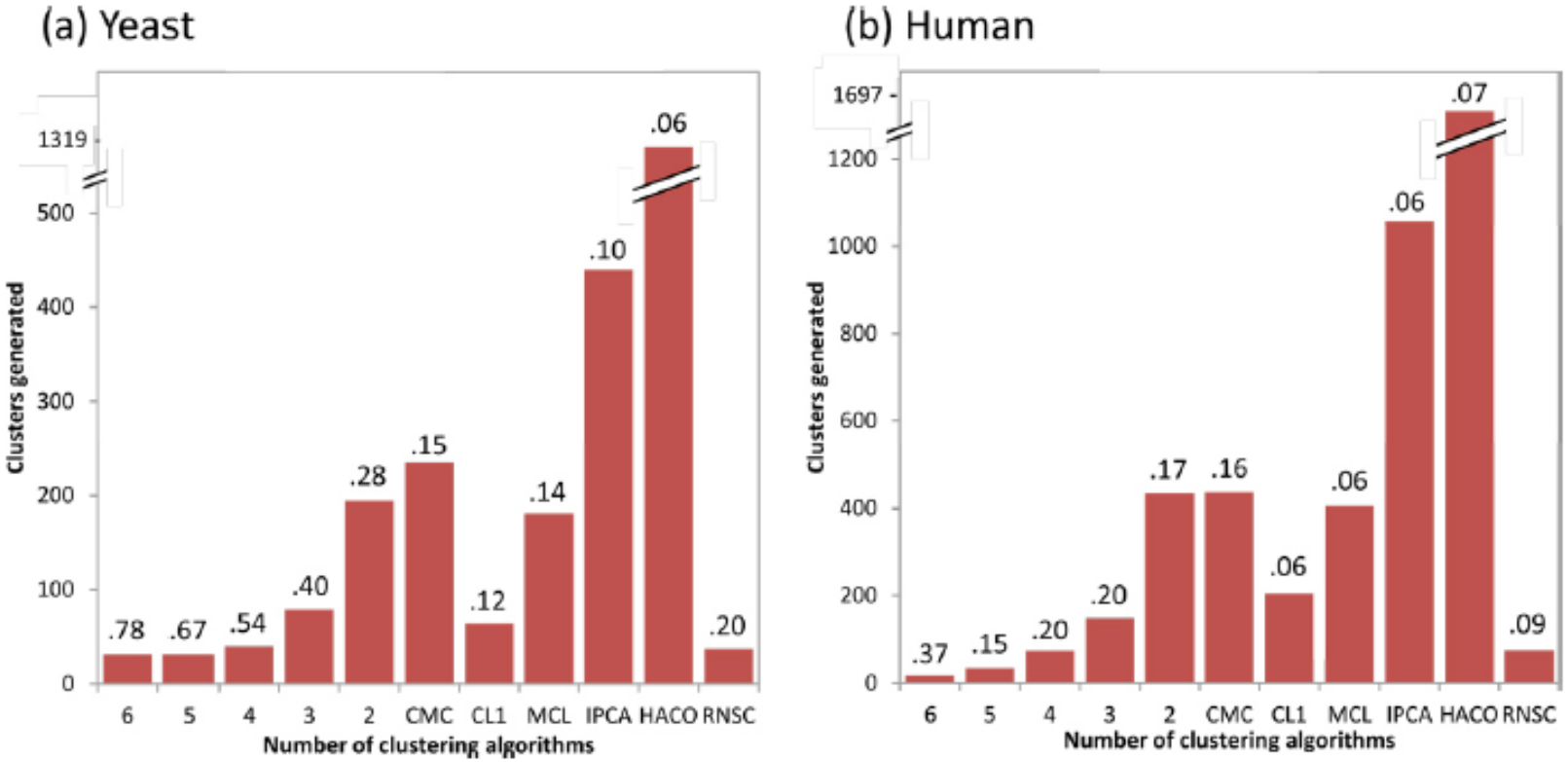


Figure 4 Distribution of clusters from the COMBINED strategy among any single or multiple number of clustering algorithms that generated them using the SWC network, and their precision (proportion of clusters that match test complexes), in (a) yeast, (b) human. Different clustering algorithms produce different sets of clusters: in either yeast or human, about 85% of clusters are generated by a single unique algorithm, while less than 7% of clusters are generated by three or more algorithms. Thus aggregating clusters from different algorithms increases the recall of complex prediction. Furthermore, precision increases as clusters are generated by a greater number of algorithms: the highest precision of clusters generated by a single algorithm is 16%, increasing to 78% for clusters generated by all algorithms in yeast, and 37% in human.

Copyright 2013 © Limsoon Wong

Power of the "COMBINED" Strategy

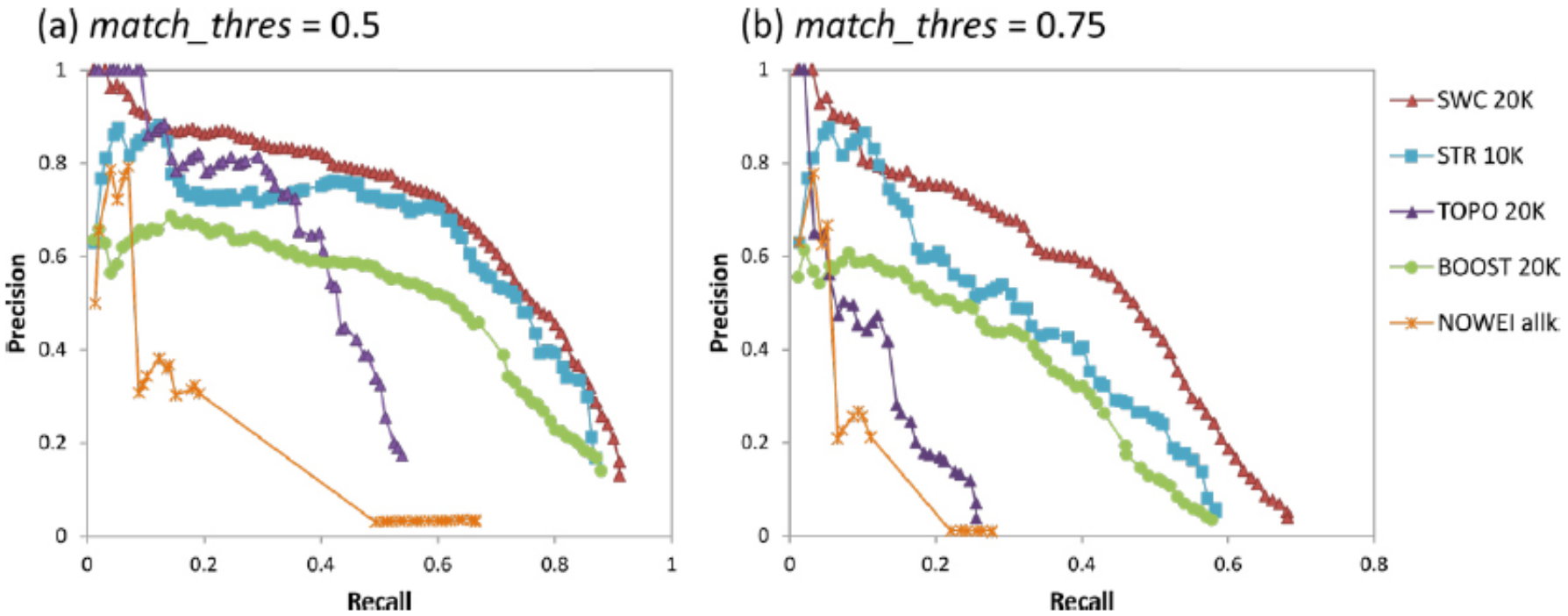


Figure 5 Precision-recall graphs for yeast complex prediction using the five weighting approaches with the COMBINED clustering strategy, using k = 20000 for SWC, TOPO, and BOOST, k = 10000 for STR, and k = *all edges* for NOWEI. (a) *match_thres* = 0.5, (b) *match_thres* = 0.75. SWC achieves the highest recall, with the highest precision at almost all recall levels, especially with the stricter *match_thres* = 0.75. Thus it outperforms all other weighting approaches, especially at predicting complexes with fine granularity.

Copyright 2013 © Limsoon Wong

Power of the "COMBINED" Strategy

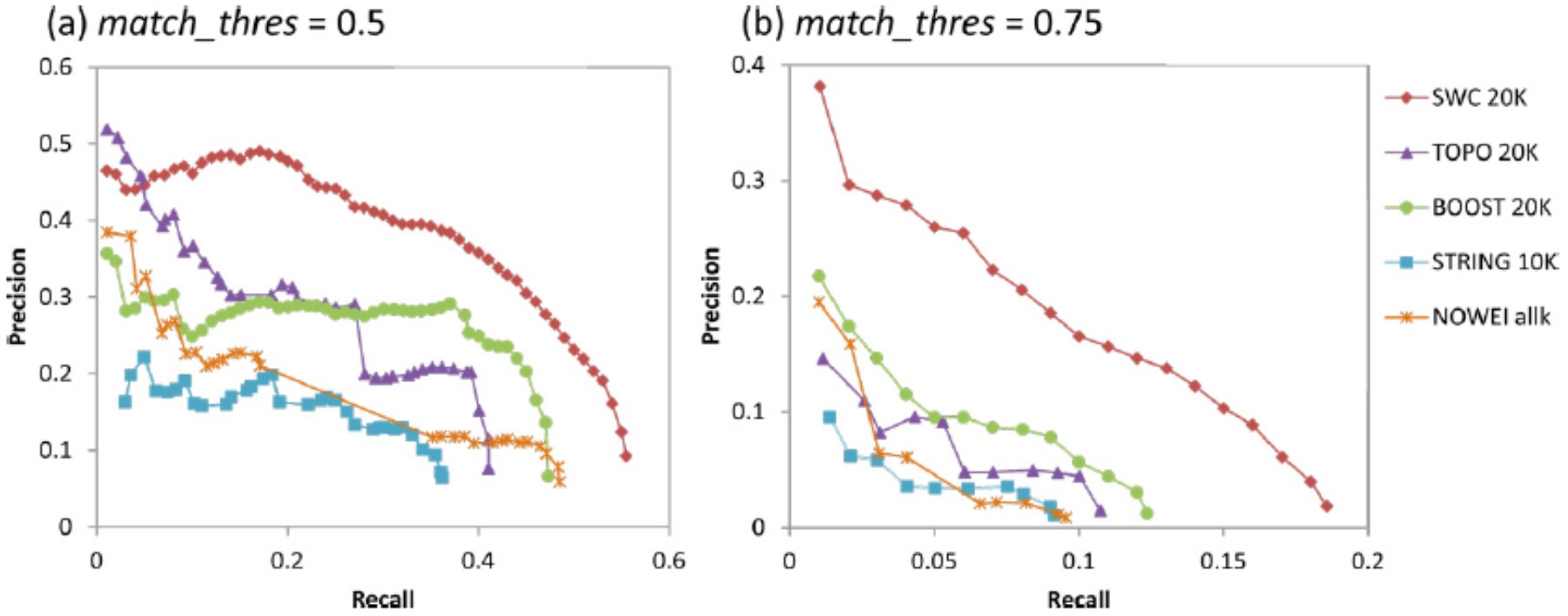


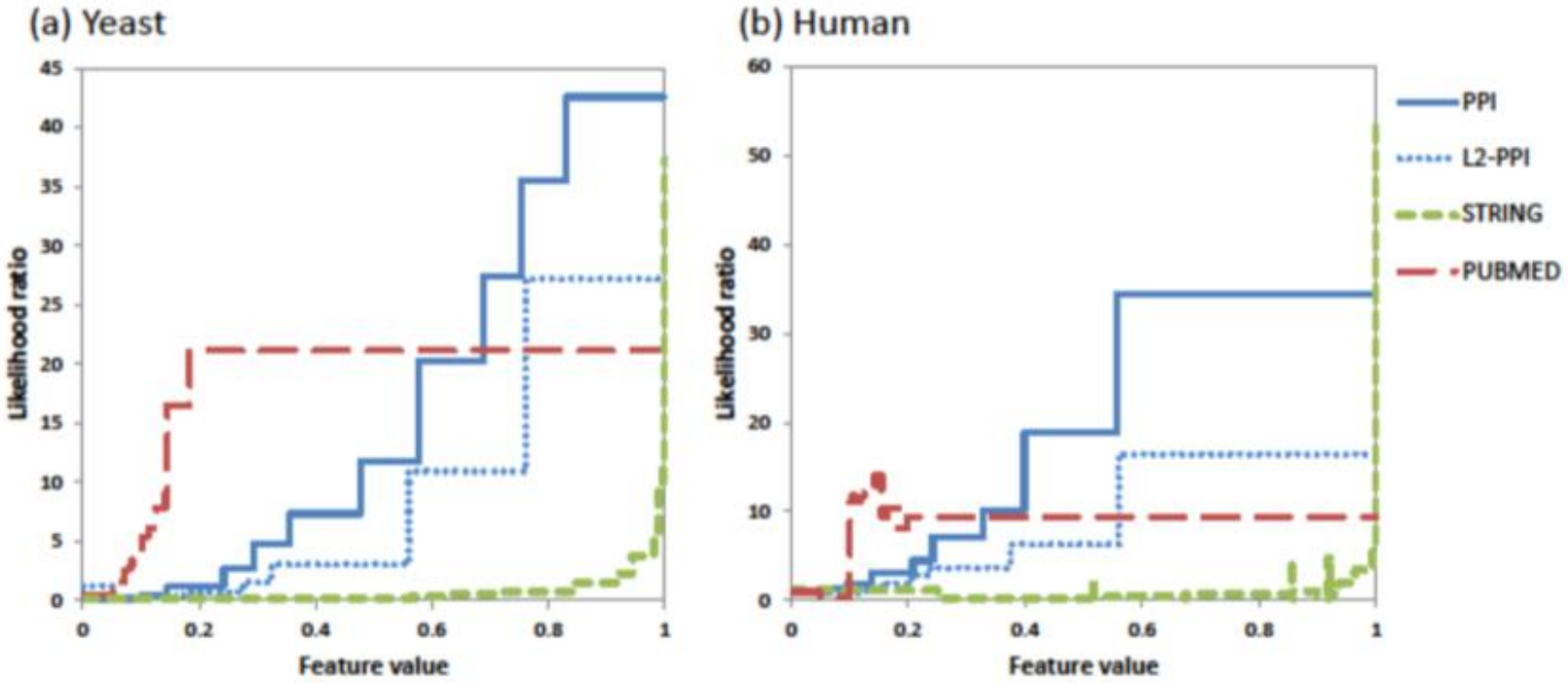
Figure 7 Precision-recall graphs for human complex prediction using the five weighting approaches for the COMBINED clustering strategy. SWC achieves the highest recall with the highest precision at almost all recall levels, especially with the stricter *match_thres* = 0.75, where SWC recalls at least 50% more test complexes compared to the other approaches and maintains almost twice the precision throughout its recall range. Thus it outperforms all other weighting approaches, especially at predicting complexes with ne granularity.

Copyright 2013 © Limsoon Wong

Evidence: Co-complexness Likelihoods

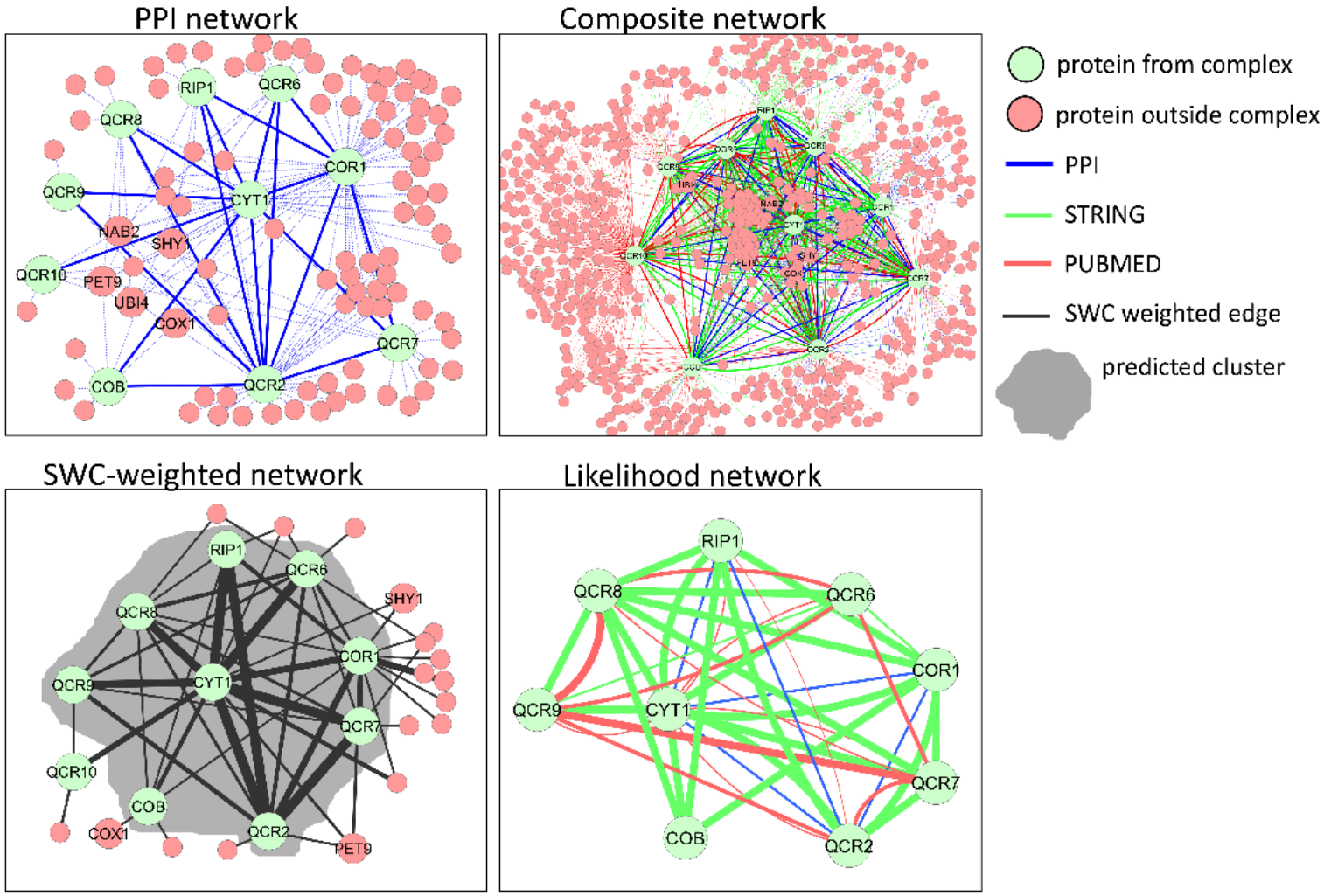
- **"Co-complexness strength" of a feature F with score f can be expressed as:**

$$likelihood\ ratio = \frac{P(F = f|co\text{-}complex)}{P(F = f|non\text{-}co\text{-}complex)}$$



Copyright 2013 © Limsoon Wong

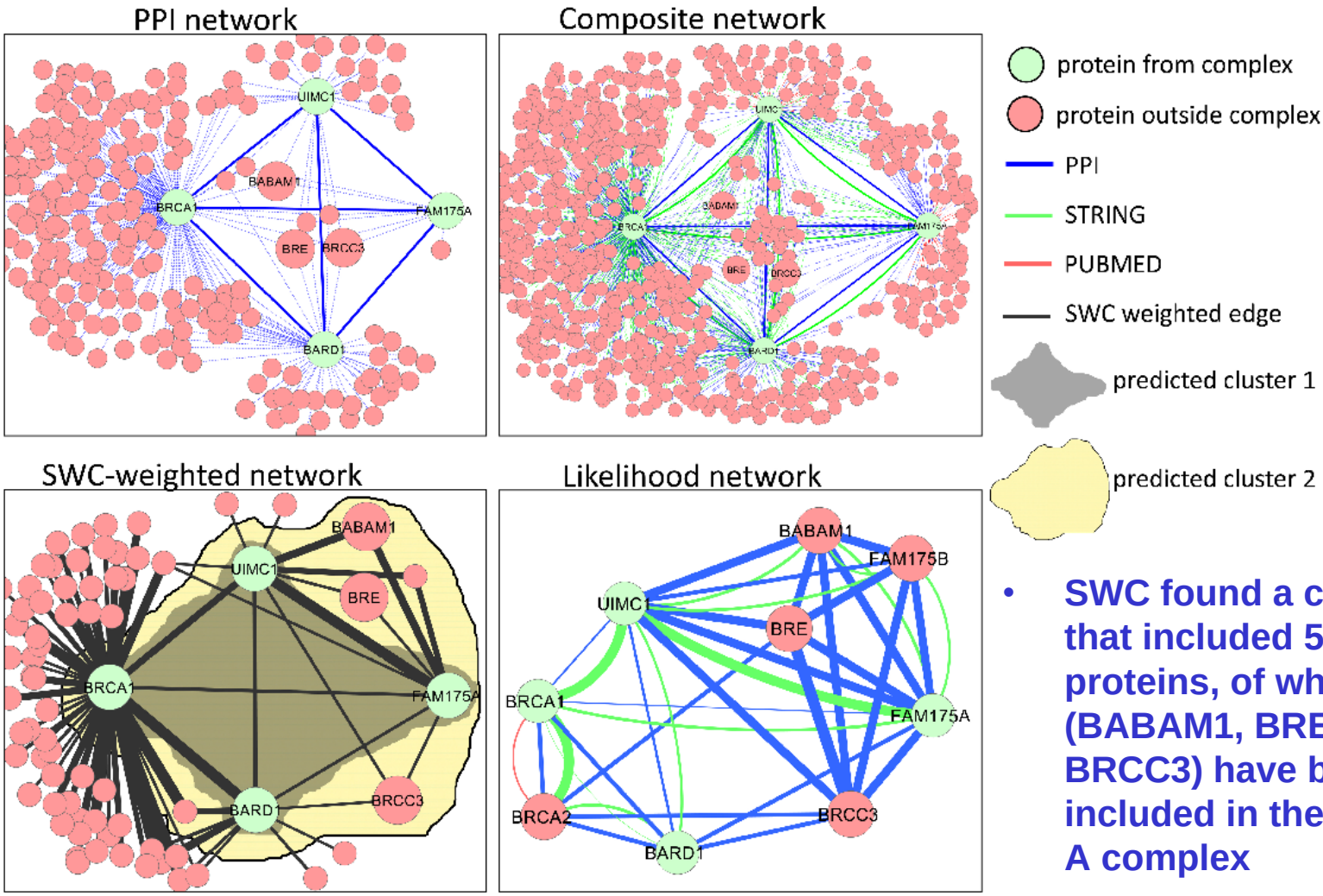
Example: Yeast BC1 Complex



Copyright 2013 © Limsoon Wong

NUS National University of Singapore

Example: Human BRCA1-A complex



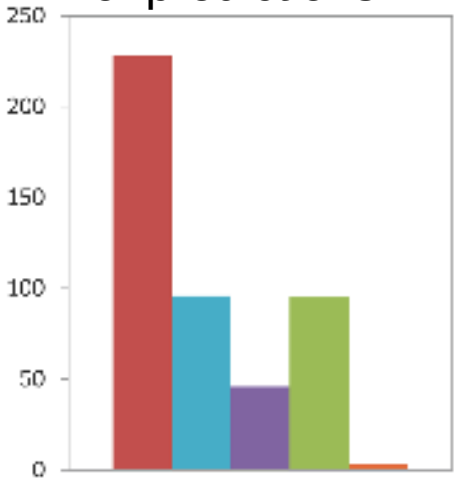
- **SWC found a complex that included 5 extra proteins, of which 3 (BABAM1, BRE, BRCC3) have been included in the BRCA1-A complex**

Copyright 2013 © Limsoon Wong

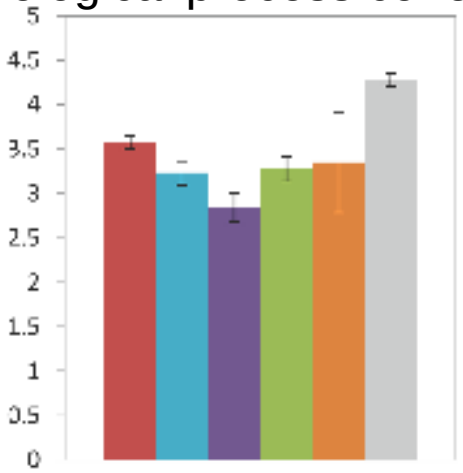
High-Confidence Predicted Complexes

Yeast

# of predictions



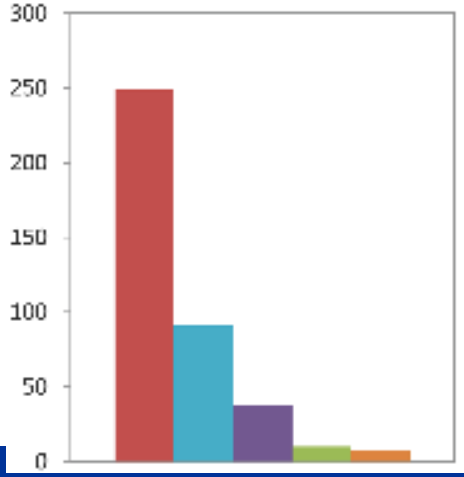
Biological process coherence



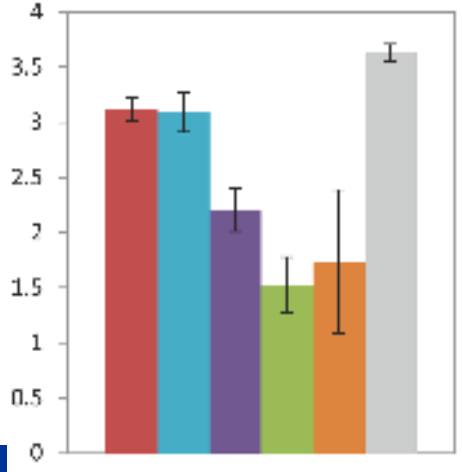
[27]. Briefly, the semantic similarity of two GO terms is first defined as the information content of their most informative common ancestor. Next, the BP semantic similarity of two proteins is defined as the highest semantic similarity between their two sets of annotated BP terms. Then, we define the BP semantic coherence of a predicted complex as the average BP semantic similarity between every pair of proteins in that complex (likewise for CC and MF).

Human

# of predictions

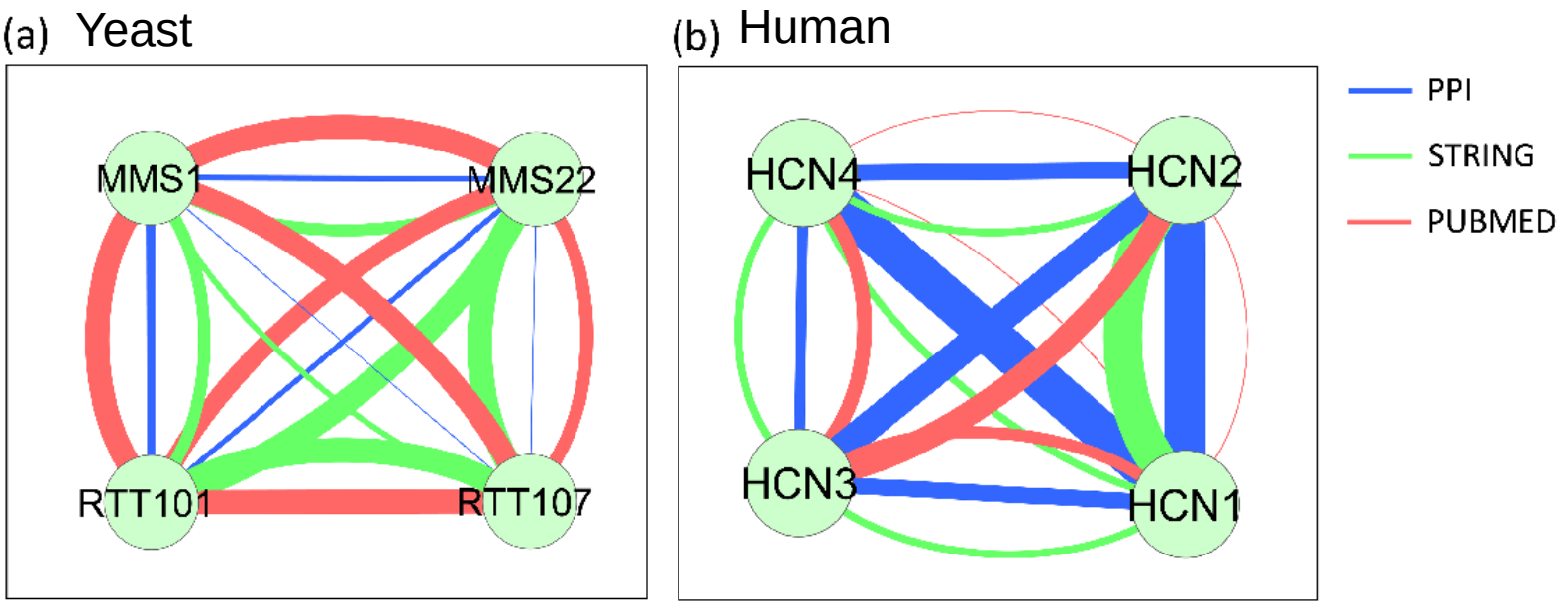


Biological process coherence



Copyright 2013 © Limsoon Wong

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**

Copyright 2013 © Limsoon Wong

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

Copyright 2013 © Limsoon Wong

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

Copyright 2013 © Limsoon Wong

Acknowledgement

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

Yong Chern Han

Copyright 2013 © Limsoon Wong