Protein Function Inference Enhanced by Text Mining

Limsoon Wong

(Based on work w/ Kenny Chua & Ken Sung)

Advanced NLP and Text Mining Forum, Tokyo, 11-13 March 2007

Plan

- **Motivation**
  - Can text mining association help?
  - Can fusion of multiple types of info help?
- **Info fusion framework**
- **Effect of co-occurences of protein Names in MEDLINE abstracts**

Motivation

Protein Function Prediction

- **Protein function prediction is a key problem**

- **It is solved using "guilt by association"**
  - Compare the target sequence T with sequences S_1, …, S_n of known function in a database
  - Determine which ones amongst S_1, …, S_n are the mostly likely homologs of T
  - Then assign to T the same function as these homologs
  - Finally, confirm with suitable wet experiments

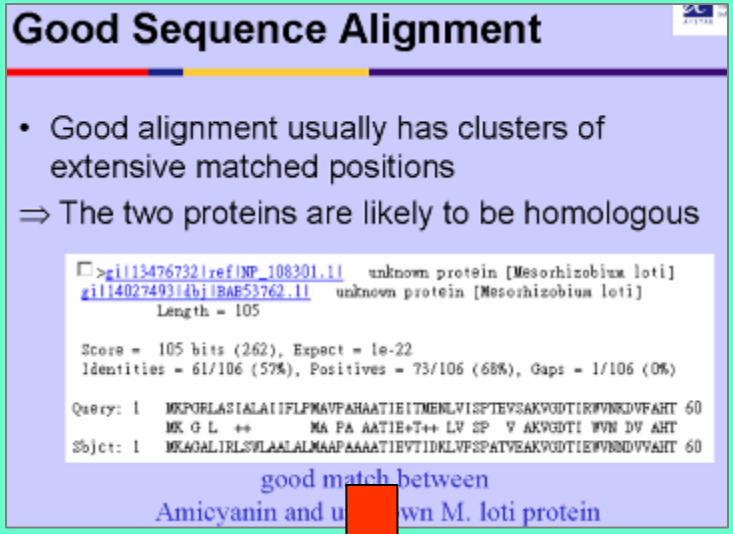
Guilt by Association of Seq Similarity

Compare *T* with seqs of known function in a db

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous



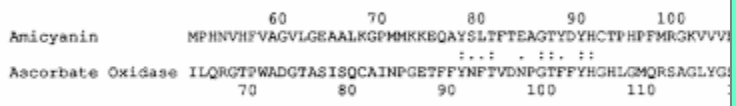
good match between
Amicyanin and u[illegible]wn M. loti protein

Poor Sequence Alignment

- Poor seq alignment shows few matched positions

⇒ The two proteins are not likely to be homologous

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase



No obvious match between
Amicyanin and Ascorbate Oxidase

Discard this function as a candidate

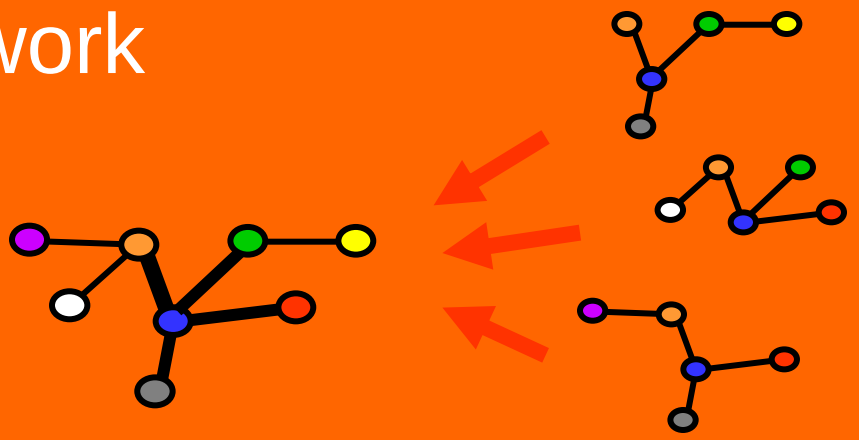
Assign to *T* same function as homologs

Confirm with suitable wet experiments

Important Unsolved Challenges

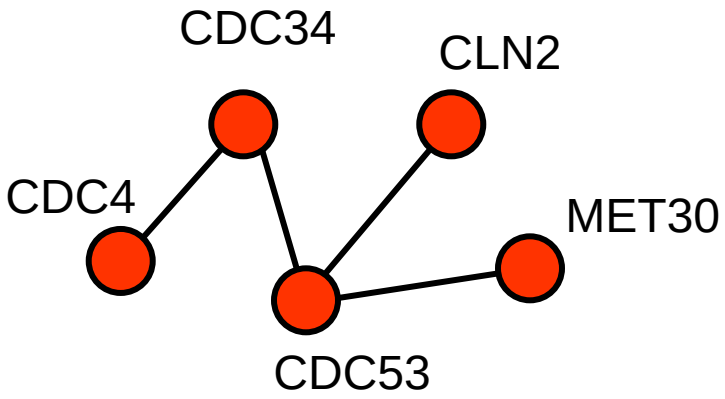
- **What if there is no useful seq homolog?**
- **Guilt by other types of association!**
  - Domain modeling (e.g., HMMPFAM)
  - Similarity of dissimilarities (e.g., SVM-PAIRWISE)
  - Similarity of phylogenetic profiles
  - Similarity of subcellular co-localization & other physico-chemico properties(e.g., PROTFUN)
  - Similarity of gene expression profiles
  - Similarity of protein-protein interaction partners
- **Can text mining association help?**
- **Can fusion of multiple types of info help?**

Information Fusion Framework



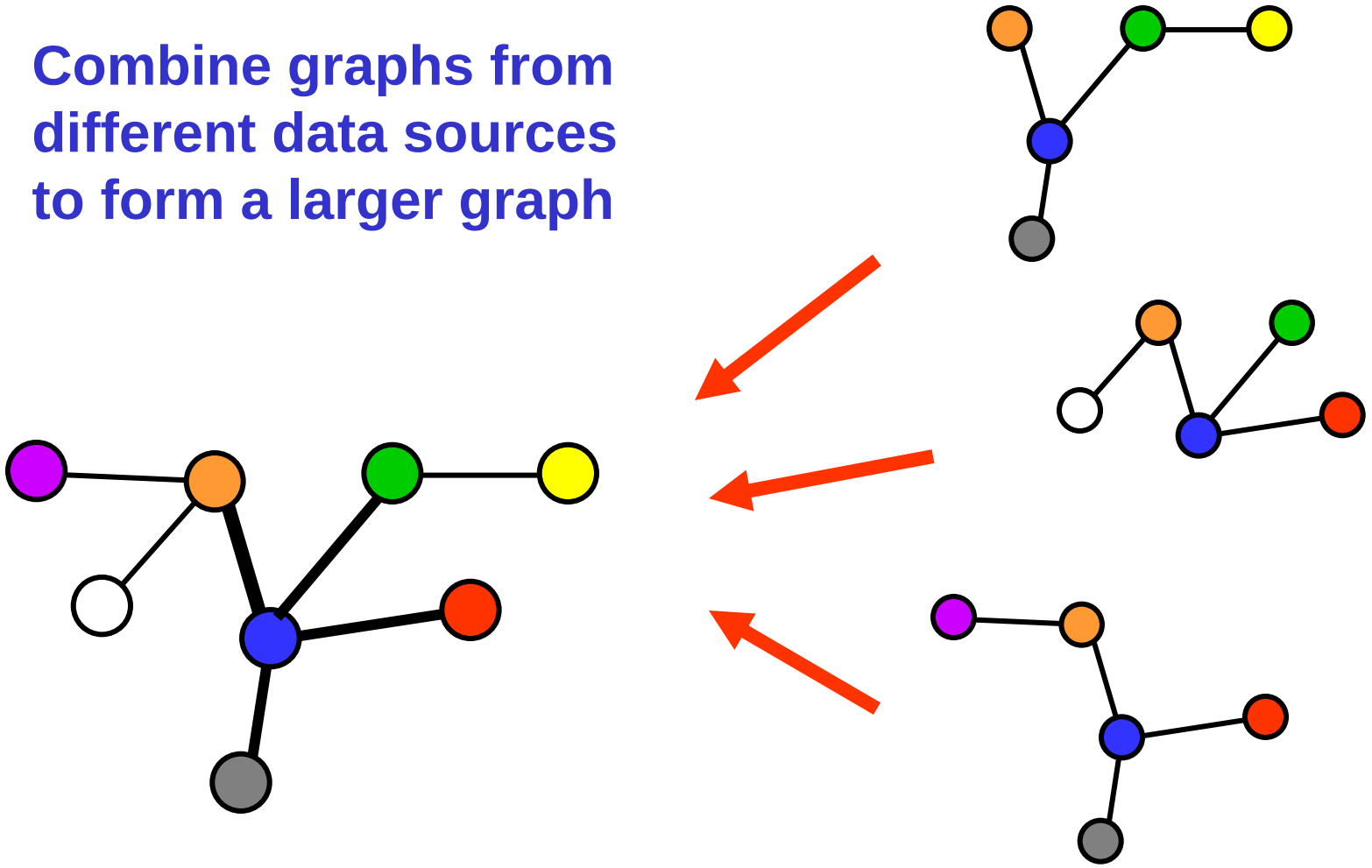
Strategy – Step 1

- **Model a data source as undirected graph G = ⟨V,E⟩**
  - V is a set of vertices; each vertex reps a protein
  - E is a set of edges; each edge (u , v) reps a relationship (e.g. seq similarity, interaction) betw proteins u and v



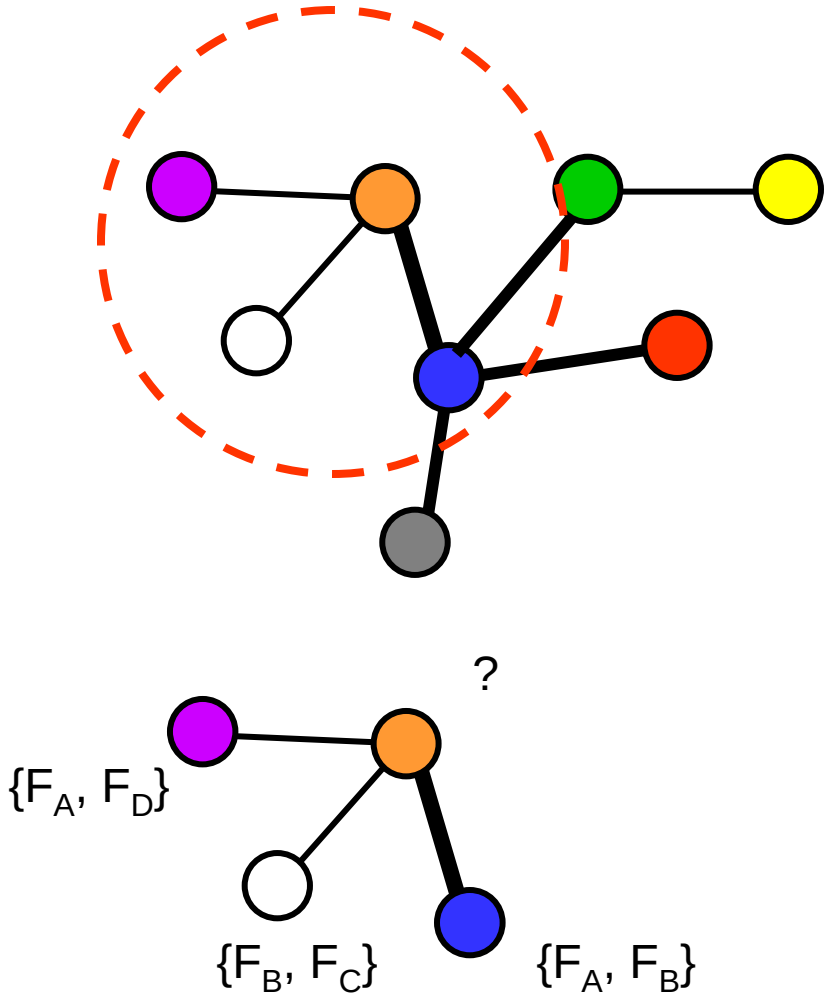
Strategy – Step 2

- **Combine graphs from different data sources to form a larger graph**



Strategy – Step 3

- **Estimate edge confidence from contributing data sources**
- **Predict function by observing which functions occur frequently in the high-confidence neighbours**



Unified Confidence Evaluation

- **Subdivide each data source into subtypes to improve precision (e.g., expt sources, sub-ranges of existing scores like E-scores)**

- **In general, estimate confidence of subtype k for sharing function f by:**

$$p(k, f) = \frac{\sum_{(u,v) \in E_{k,f}} S_f(u,v)}{\left| E_{k,f} \right| + 1}$$

  - $E_{k,f}$ is subset of edges of subtype k where each edge has either one or both of its vertices annotated with function f
  - $S_f(u,v) = 1$ if u and v shares function f, 0 otherwise

Discretization of Existing Scores

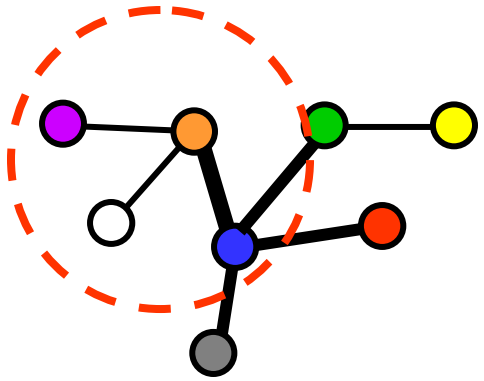
- **Scores may come in many forms**
  - E.g., Blast e-values, Pearson's correlation
- **A simple approach to discretization**
  - Split ranges into n equal intervals
  - Each interval becomes a new subtype
  - Assume linearity in range
  - Other strategies possible

Combination of Confidence

- **Combine confidence of data sources contributing to each edge:**

$$r_{u,v,f} = 1 - \prod_{k \in D_{u,v}} \left(1 - p\left(k, f\right)\right)$$

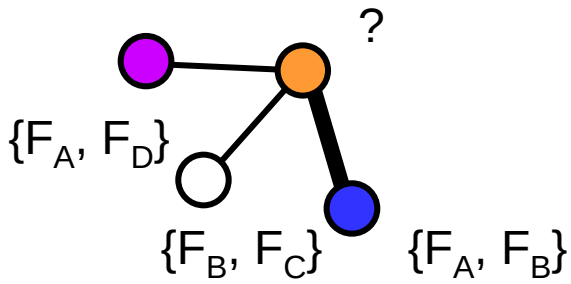
  - P(k.f) is confidence of edges of subtype k sharing function f
  - $D_{u,v}$ is the set of subtypes of data sources which contains the edge (u,v)



Function Prediction

- **Weighted Average**

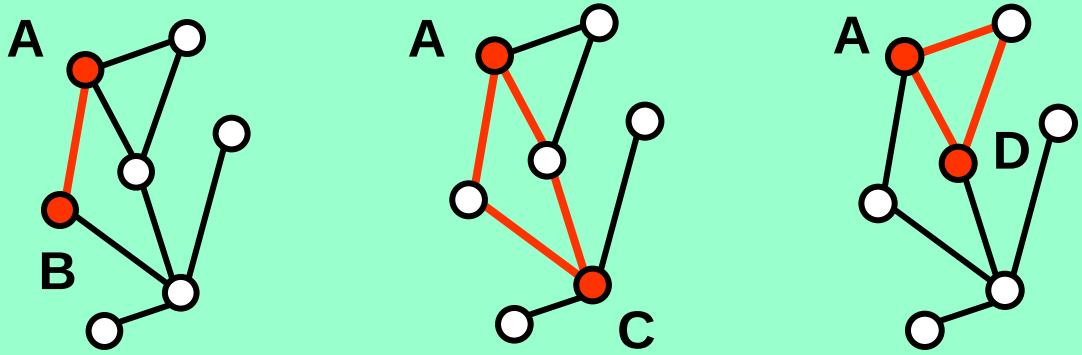
$$S_f(u) = \frac{\sum_{v \in N_u} \left(e_f(v) \times r_{u,v,f} \right)}{1 + \sum_{v \in N_u} r_{u,v,f}}$$



- $S_f(u)$ is score of function f for protein u
- $e_f(v)$ is 1 if protein v has function f, 0 otherwise
- N_u is set of neighbours of u
- $r_{u,v,f}$ is confidence of edge (u, v)

Level-2 Neighbours

- **Increase coverage of Protein-Protein interactions**
  - Indirect function association (Chua et al. 2006)
  - Topological weight applied to PPI
  - Divide into 3 subtypes:



Level-1 Neighbours **Level-2 Neighbours** **Level-1&2 Neighbours**

  - A threshold of 0.01 is applied on L2 neighbours to limit false positives

Topological Weight Applied to PPI: FS-Weighted Measure with Reliability

- **Take reliability into consideration when computing FS-weighted measure:**

$$S_R(u,v)=\frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in N_u} r_{u,w}+\sum_{w\in(N_u\cap N_v)} r_{u,w}(1-r_{v,w})\right)+2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}\times\frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}{\left(\sum_{w\in N_v} r_{v,w}+\sum_{w\in(N_u\cap N_v)} r_{v,w}(1-r_{u,w})\right)+2\sum_{w\in(N_u\cap N_v)} r_{u,w}r_{v,w}}$$

  - **N_k is the set of interacting partners of k**
  - **$r_{u,w}$ is reliability weight of interaction betw u and v**

$\Rightarrow$ **Rewriting**

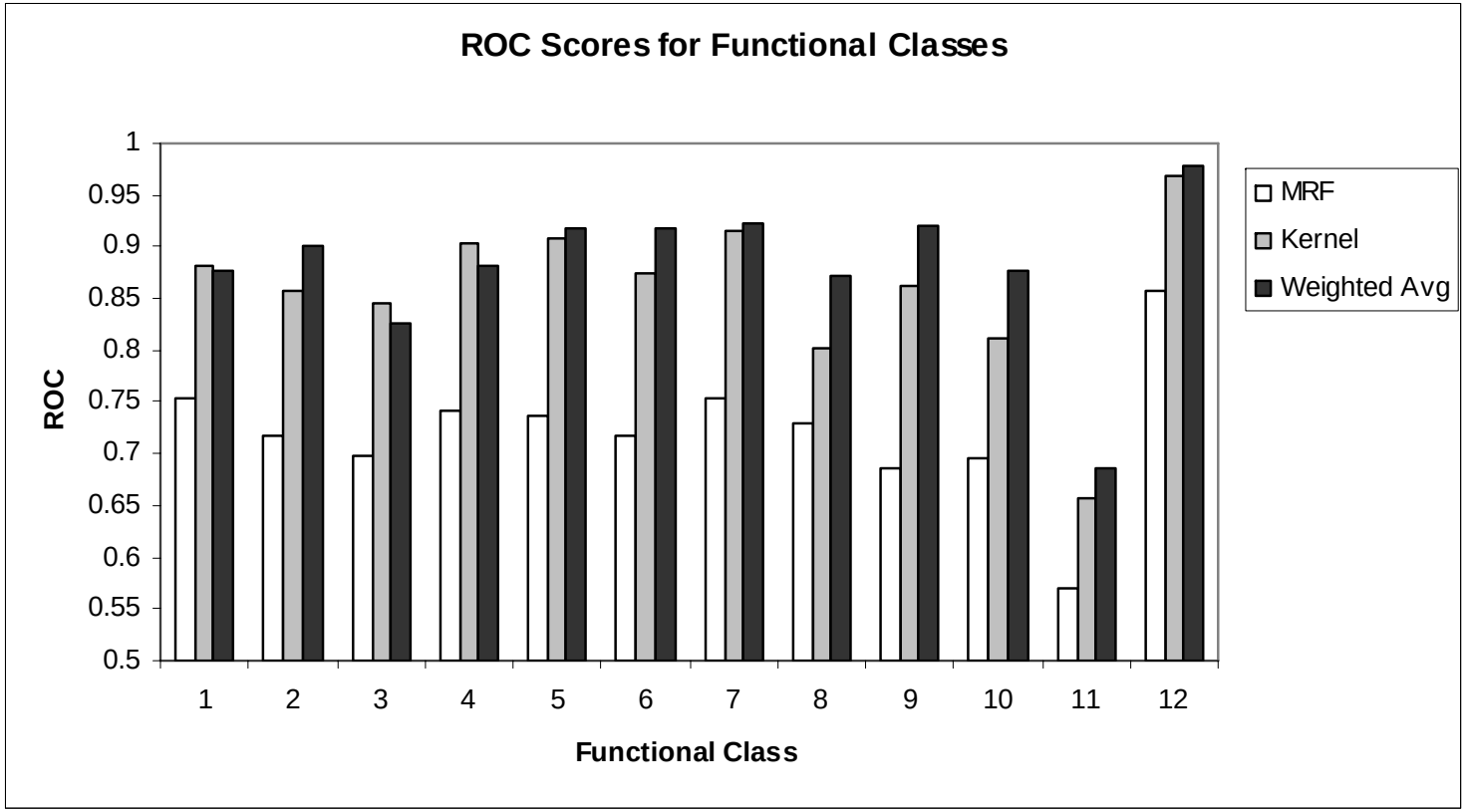
$$S(u,v)=\frac{2X}{2X+Y}\times\frac{2X}{2X+Z}$$

Comparison w/ Existing Approaches

- **Datasets of Deng et al, '04**
- **4 data sets (S. cerevisiae)**
  - Protein-Protein Interactions
    - **2,448 edges**
  - Protein Complexes
    - **30,731 edges**
  - Pfam Domains
    - **28,616 edges**
  - Expression Correlation
    - **1,366 edges**
- **12 functional classes**

	Category	Size
1	Metabolism	1048
2	Energy	242
3	Cell cycle & DNA processing	600
4	Transcription	753
5	Protein synthesis	335
6	Protein fate	578
7	Cellular transport & transport mechanism	479
8	Cell rescue, defense & virulence	264
9	Interaction with cellular env	193
10	Cell fate	411
11	Control of cellular organization	192
12	Transport facilitation	306

Comparison w/ Existing Approaches

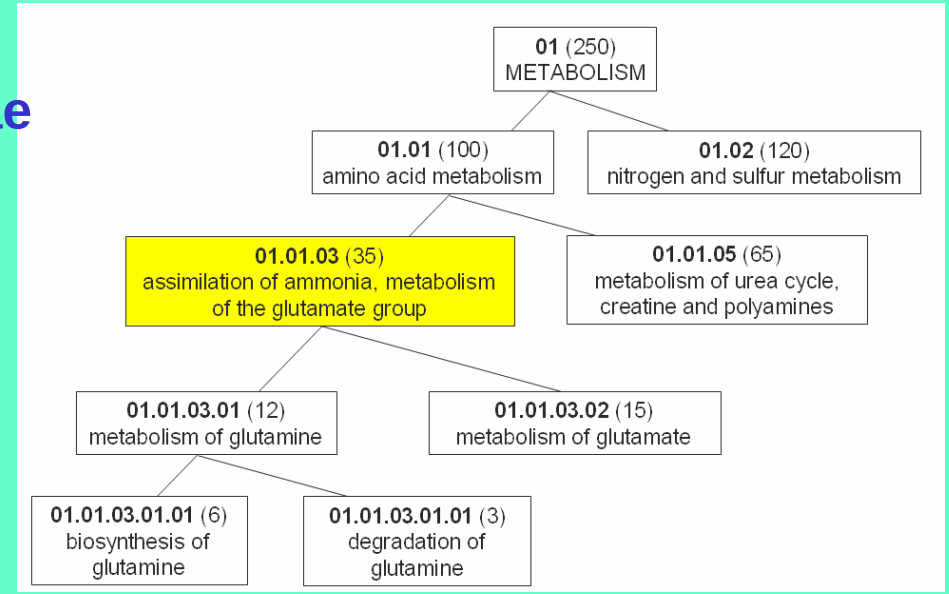


- **Validation Method (Lanckriet et al, 2004)**
  - Receiver Operating Characteristics (ROC)
  - True Positives vs False Positives
  - Area under ROC curve for each function
  - Averaged over 3 repetitions of 5-fold cross validation

Effect of Co-occurences of Protein Names in MEDLINE Abstracts

GO Terms Prediction for Yeast Proteins

- **Proteins from S. Cerevesiae**
  - 5448 proteins from GO Annotation (SGD)
- **Functional Annotation**
  - Gene Ontology
  - Hierarchical
  - 3 Namespaces (molecular function, biological process, cellular component)



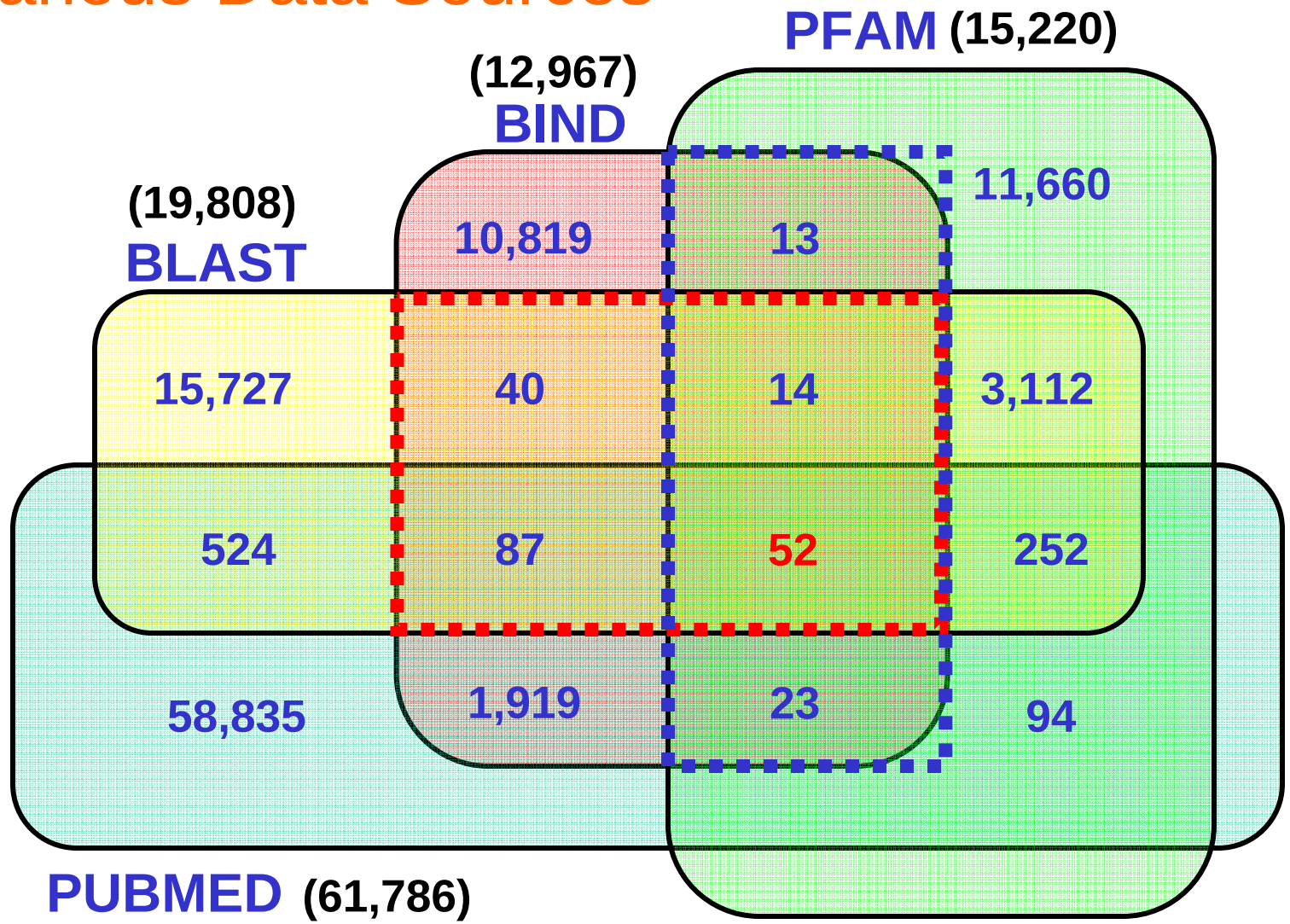
- **Informative GO Terms (for evaluation)**
  - Zhou et al. (2002)
  - FC associated with at least 30 proteins and no subclass associated with at least 30 proteins

Data Sources

- **Protein Sequences**
  - Seqs from GO database
  - Each yeast seq is aligned w/ rest using BLAST (cutoff E-Score = 1)
  - -log(e-score) used as score
  - Top 5 results w/ known annotations
  - 19,808 unique pairs involving yeast proteins
- **Pfam Domains (SwissPfam)**
  - Precomputed Pfam domains for SwissProt and TrEMBL proteins w/ E-value threshold 0.01
  - No. of common domains as score
  - 15,220 unique pairs involving yeast proteins
- **PPI (BIND)**
  - 12,967 unique interactions betw yeast proteins
  - FS weight used as score

- **Pubmed Abstracts**
  - Pubmed abstracts obtained by searching protein's name and aliases on Pubmed
  - Limit to first 1000 abstracts returned
  - Fraction of abstracts w/ co-occurrence used as score
  - 61,786 unique pairs involving yeast proteins

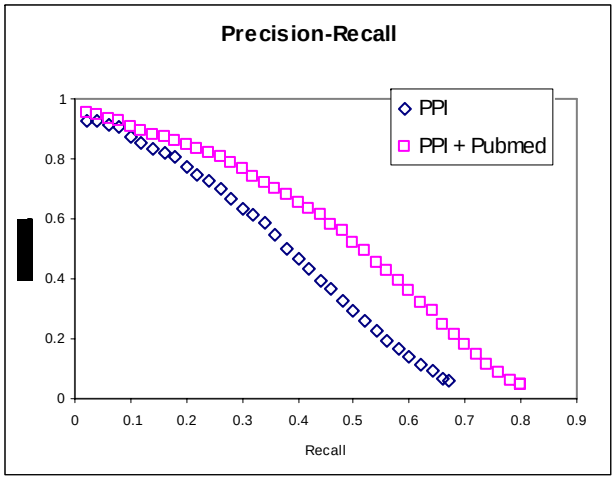
Pairs Involving Yeast Proteins in Various Data Sources

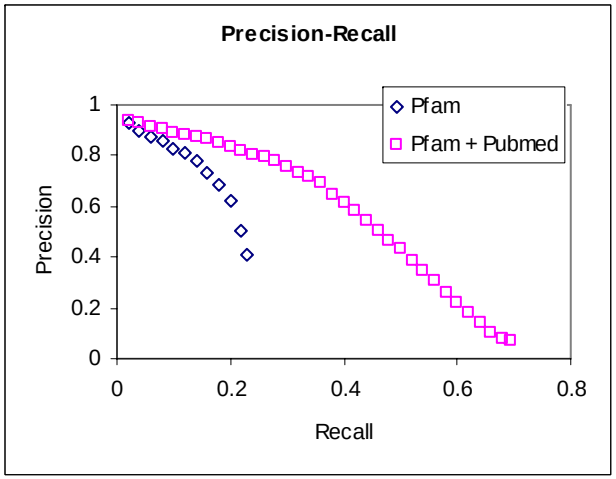


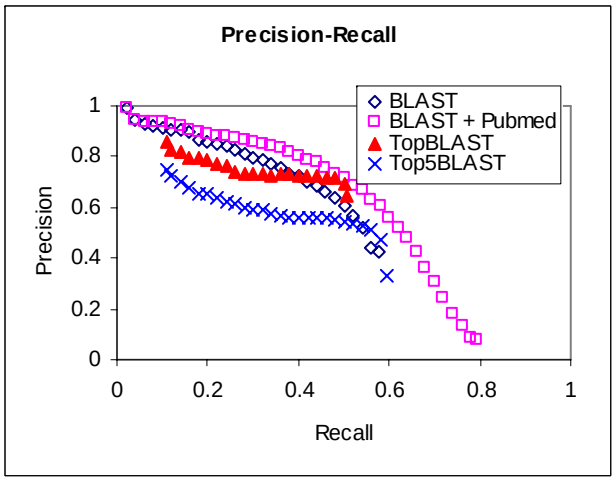
Can literature co-occurrence info help?

- **Need comparisons of**
  - PPI info w/ & w/o literature occurrence info,
  - BLAST info w/ & w/o literature occurrence info,
  - Pfam info w/ & w/o literature occurrence info,
  - "combined" w/ & w/o literature occurrence info,
  - Top-blast info w/ & w/o literature occurrence info

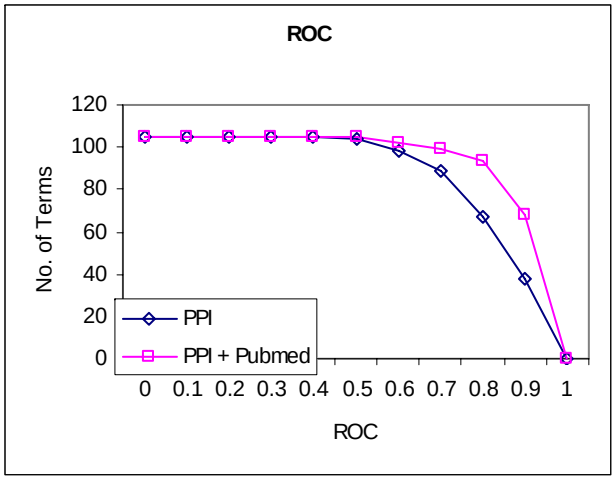
Diff in Recall-Precision by Literature Co-Occurrence

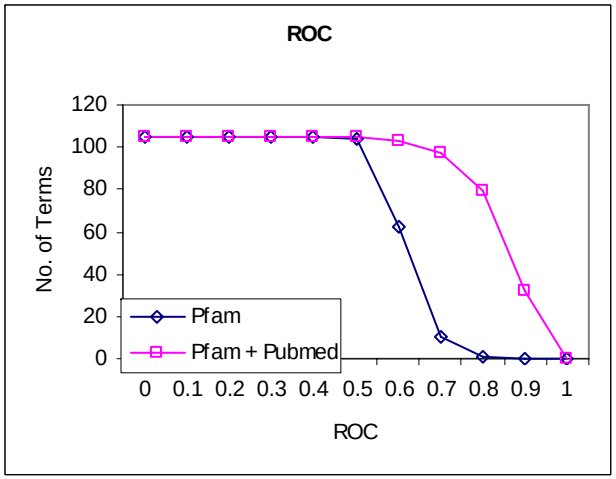


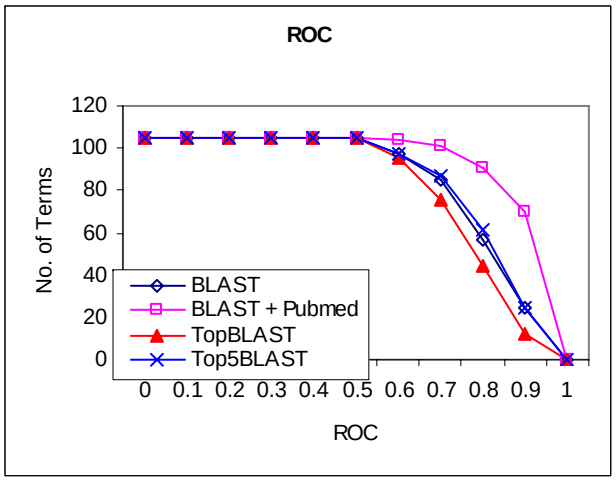


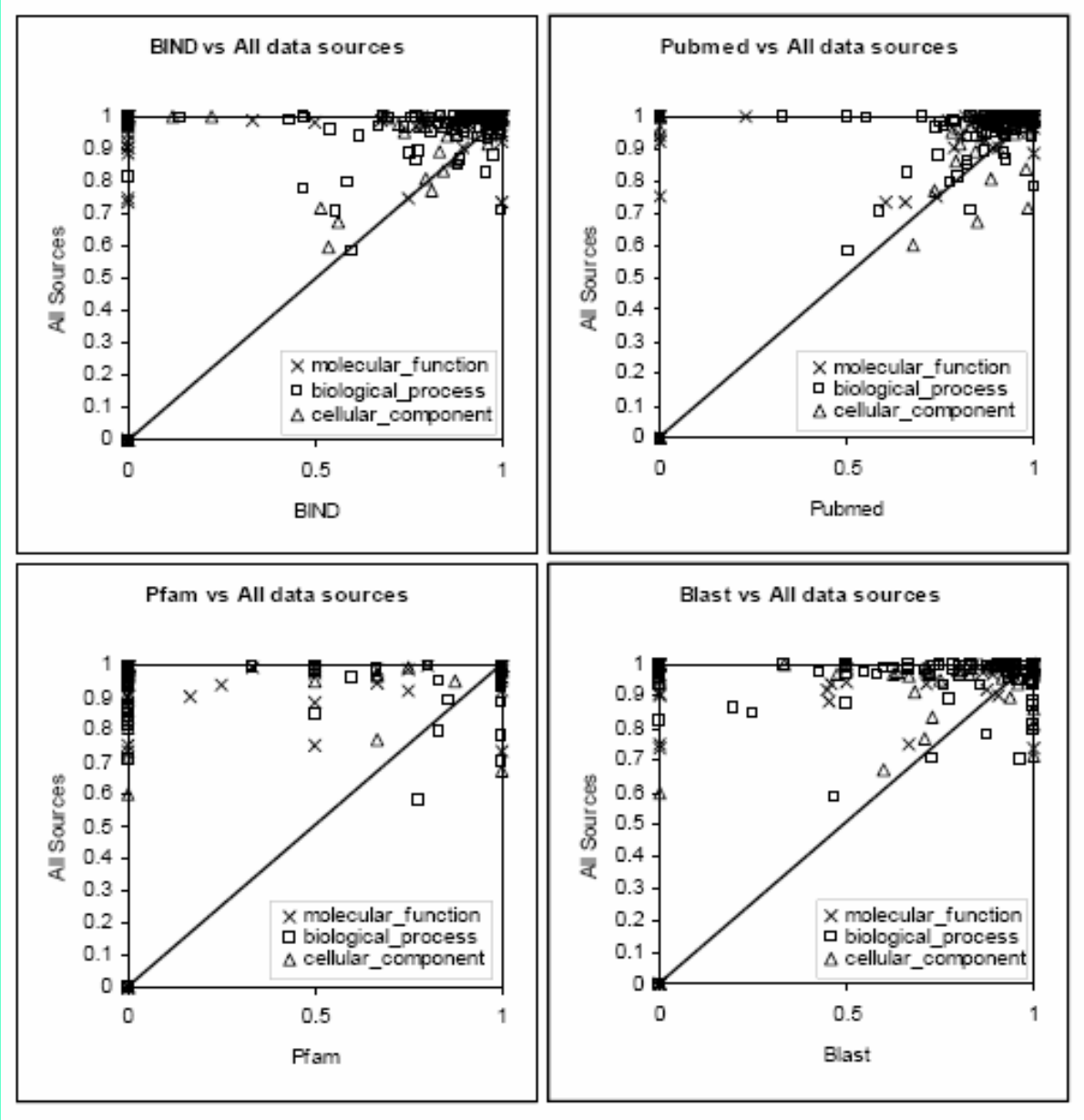


Diff in No. of Terms w/ Better ROC by Literature Co-Occurrence

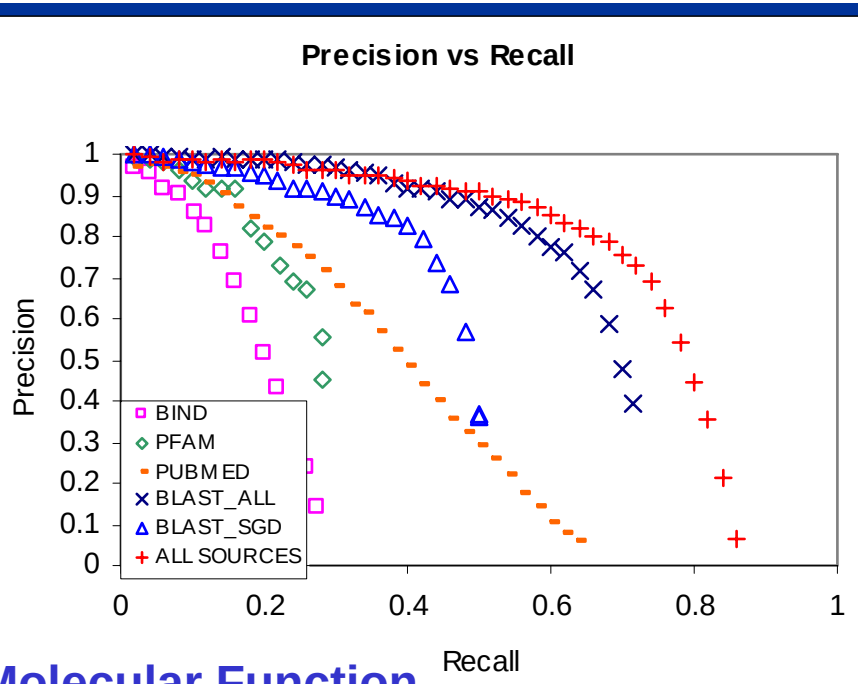








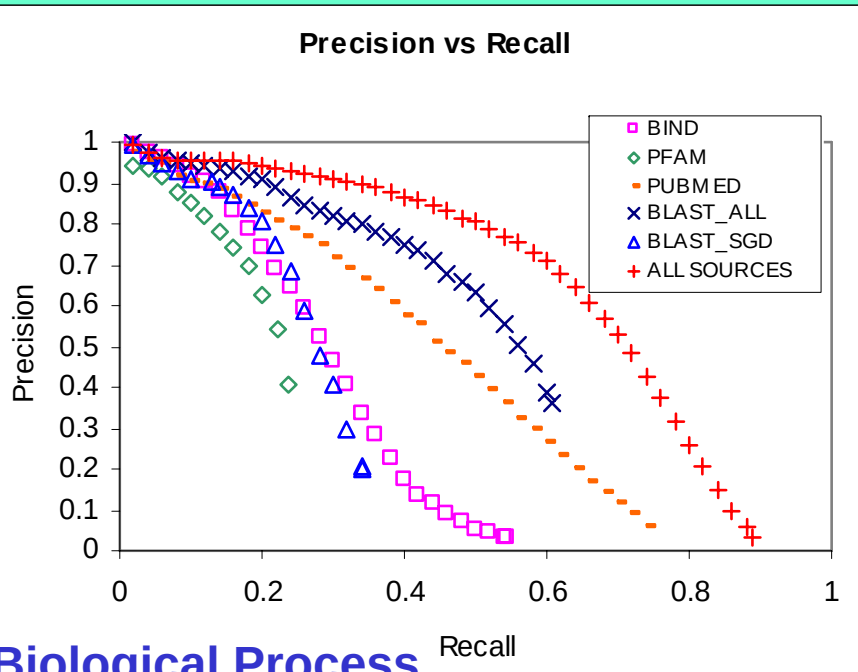
Literature co-occurrence seems to contribute especially well to cellular component



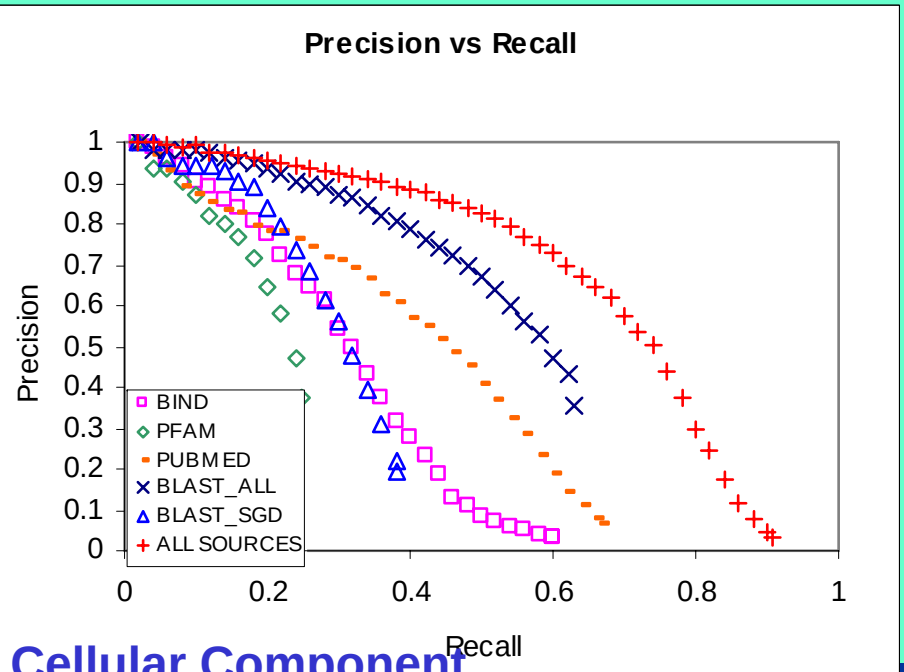
Molecular Function

NUS National University of Singapore

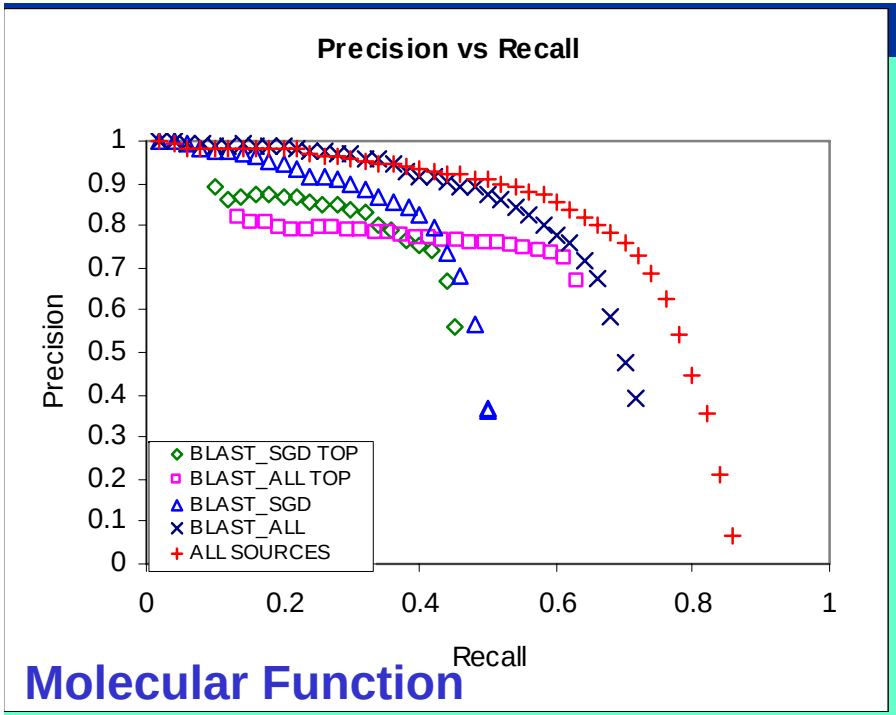
Combining all data sources outperforms any individual data source



Biological Process

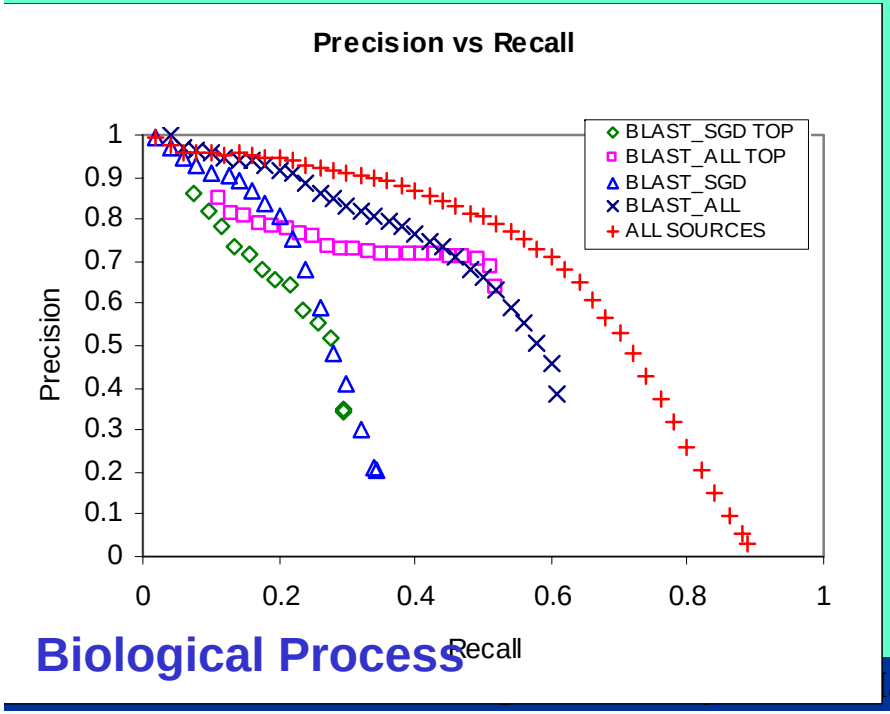


Cellular Component

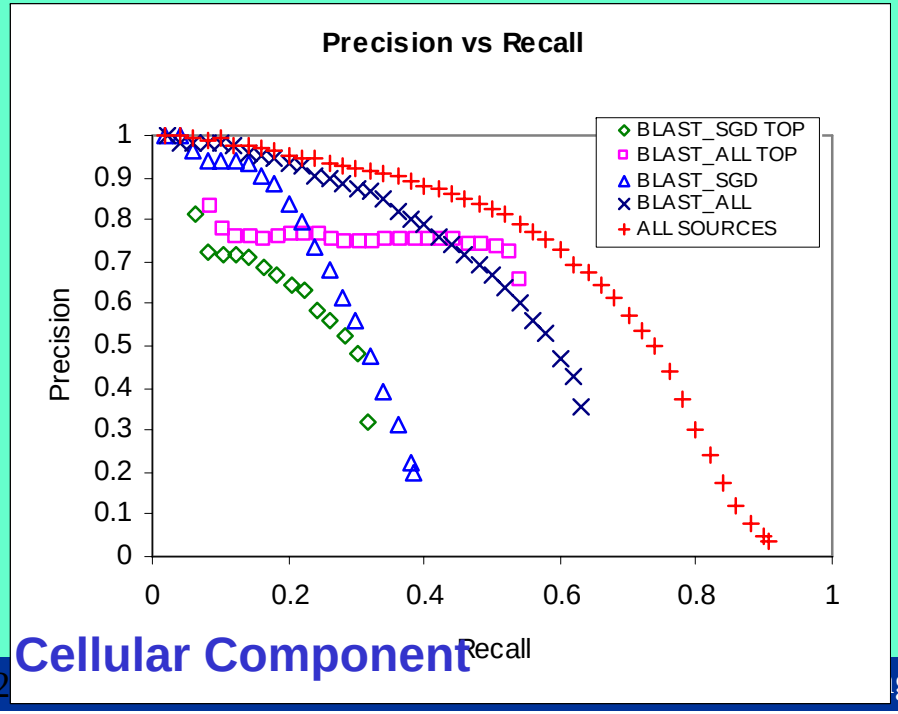


Molecular Function

- Weighted Averaging predicts w/ better precision than transferring function from top blast hit
- Using all data sources outperforms topblast in both sensitivity and precision



Biological Process



Cellular Component

Conclusions

- **A simple graph-based method that combines multiple sources of data sources for function prediction**
- **Even simple co-occurrence count can give reasonable sensitivity & precision for function prediction**
- **Combining multiple info sources outperforms any single info source**

References

- Hon Nian Chua, Wing-Kin Sung, Limsoon Wong. **Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions**. *Bioinformatics*, 22:1623--1630, 2006 [PPI]

- H.N. Chua, W.K. Sung, & L. Wong. **A graph-based approach to integrating multiple data sources for protein function prediction**. *In preparation*, 2007

- M. Deng, T. Chen, & F. Sun. **An integrated probabilistic model for functional prediction of proteins**. *JCB*, 11(2-3):463-75, 2004 [MRF]

- G.R. Lanckriet et al. **Kernel-based data fusion and its application to protein function prediction in yeast**. *Proc. PSB 2004*, pp. 300-311 [Kernel]

Any Question?