Protein Function Prediction By Information Fusion

Limsoon Wong

(joint work w/ Hon Nian Chua & Wing Kin Sung)

NUS
National University
of Singapore

Protein Function Prediction Approaches

- Sequence alignment (e.g., BLAST)
- Generative domain modeling (e.g., HMMPFAM)
- Discriminative approaches (e.g., SVM-PAIRWISE)
- Phylogenetic profiling
- Subcellular co-localization (e.g., PROTFUN)
- Gene expression co-relation
- Protein-protein interaction
- Information fusion, …

Information Fusion

- **Markov Random Fields (Deng et al., *JCB*, 2004)**
  - Maximum Likelihood
  - Model data sources as binary relation betw proteins
- **Kernel Fusion (Lanckriet et al., *PSB*, 2004)**
  - Discriminative approach
  - Models each data source w/ diff feature vectors
  - Weighted linear combination of kernels via semi-definite programming

Difficulties w/ Information Fusion

- **Differences in nature**
  - E.g., sequence homology vs PPI are very different relationships
- **Differences in reliability**
  - E.g., noisy datasets such as Y2H PPI and gene expression
- **Differences in scoring metrices**
  - E.g., E-Score from BLAST vs Pearson correlation between expression profiles

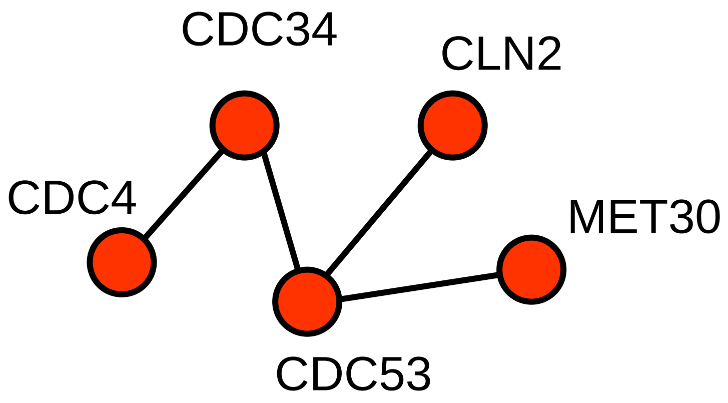
Motivation

- **Problems:**
  - Complex models such as MRF and Kernel Fusion are computationally expensive
  - Difficult or not possible to identify contributing sources in a prediction
- **Unified scoring of multiple sources has potential (Lee et al., *Science*, 2004)**
  - Simple scoring using Log Likelihood
  - Identified many functional clusters

⇒ **A simple, flexible, and effective way to integrate data sources that reports contributing sources in predictions to allow users to exercise judgment**

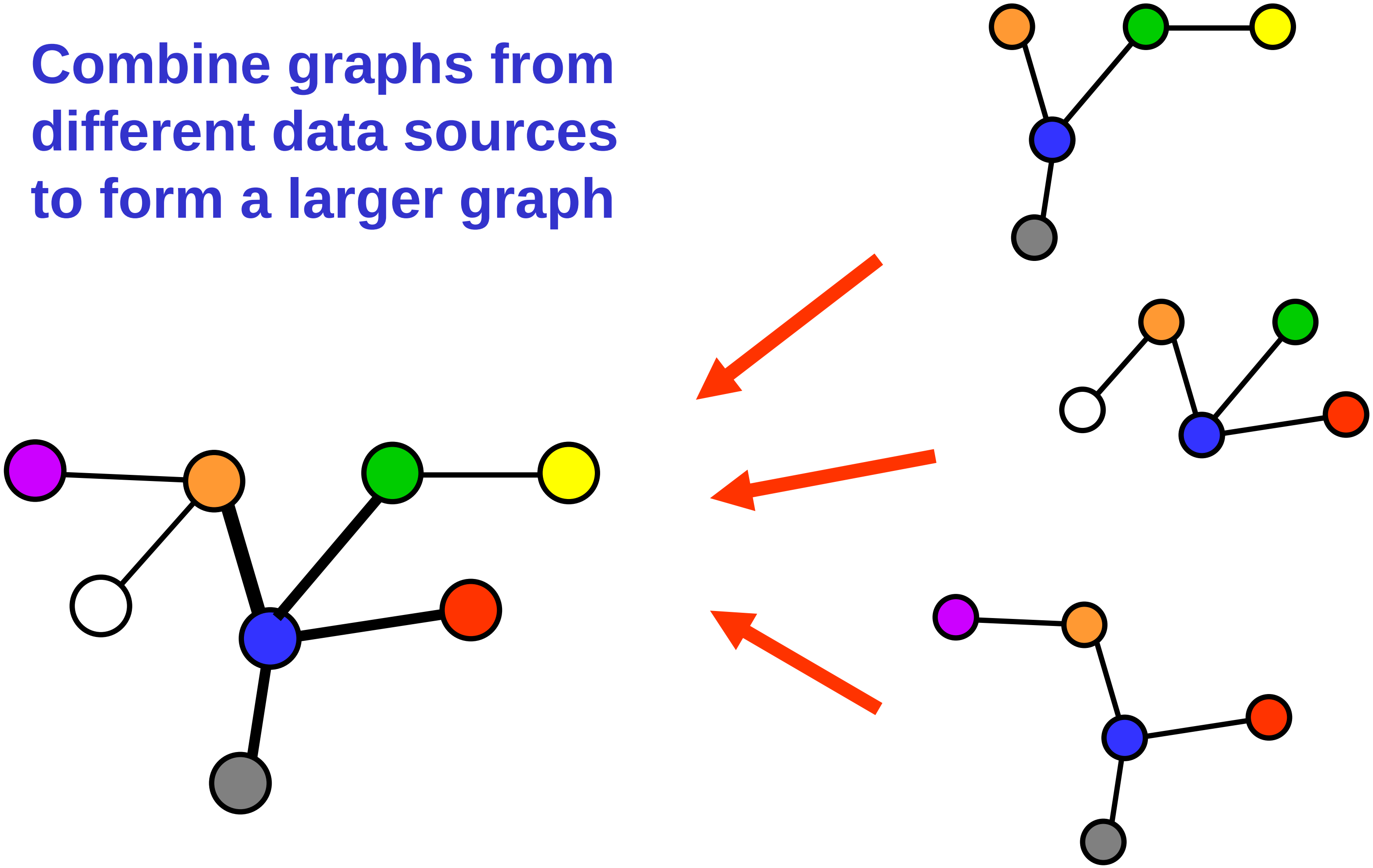
Strategy – Step 1

- **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

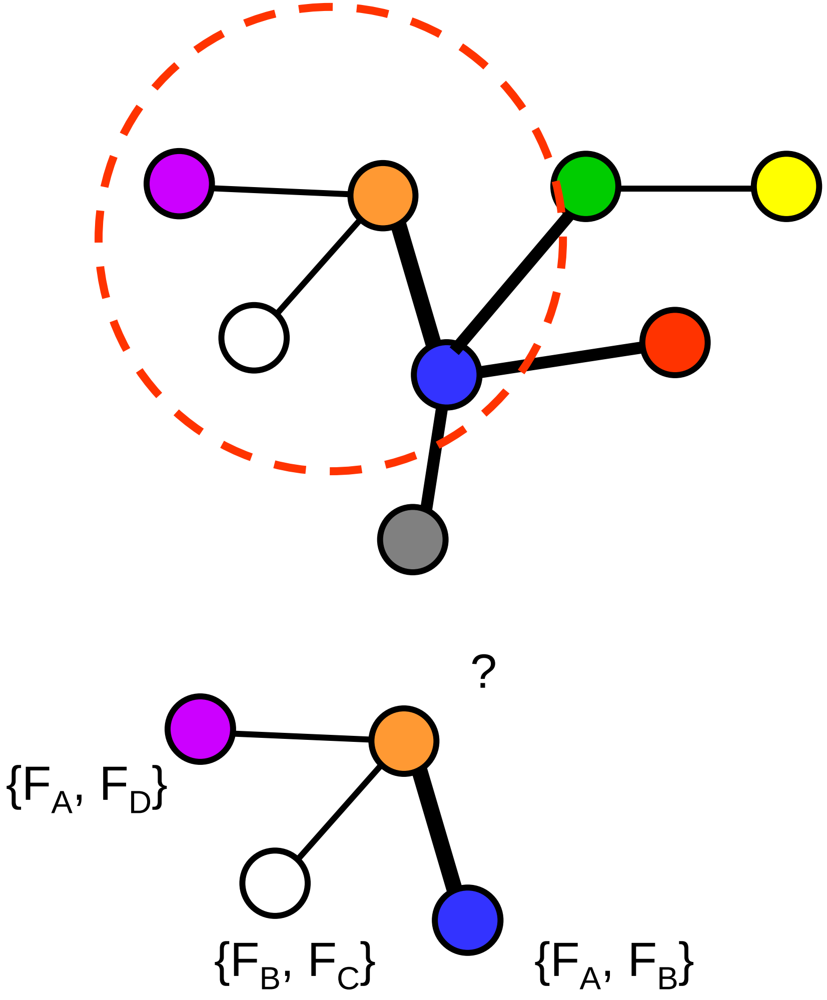
- **Combine graphs from different data sources to form a larger graph**



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)**

- **Estimate confidence of subtype k for sharing function f by:**

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

  - $E_{k,f}$ 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

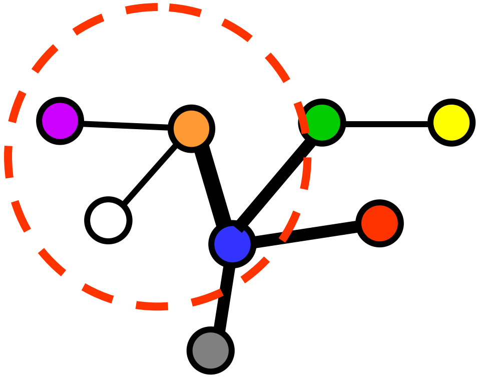
- **A simple approach to discretization**
  - Split ranges into n equal intervals
  - Each interval becomes a new subtype
  - Assume linearity in range
  - Other strategies possible

Combination of Confidence

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

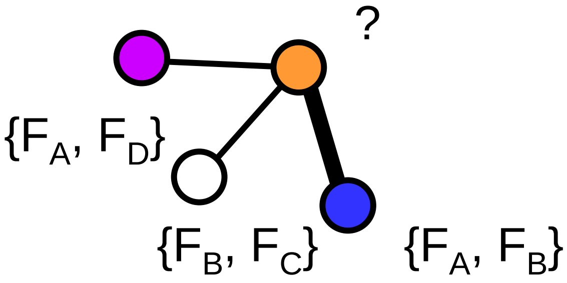
$$r_{u,v,f} = 1 - \prod_{k \in D_{u,v}} \left(1 - p\left(k, f\right)\right)$$

  - 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**

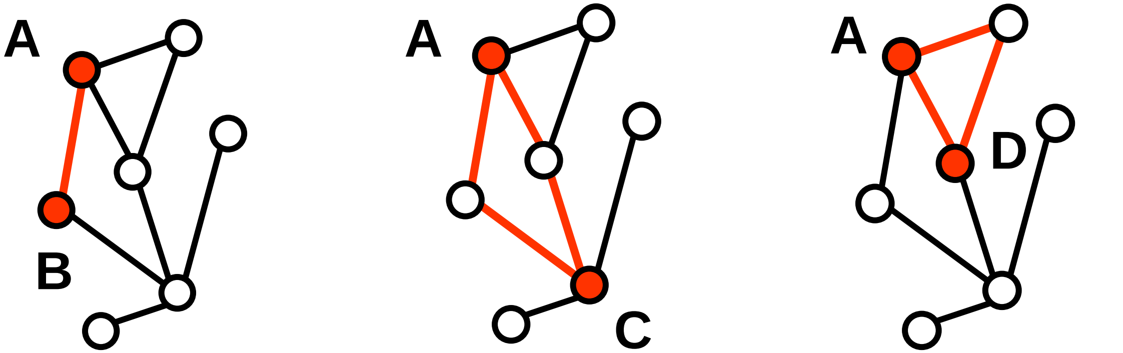


$$S_f(u) = \frac{\sum_{v \in N_u} \left(e_f(v) \times r_{u,v,f} \right)}{1 + \sum_{v \in N_u} r_{u,v,f}}$$

- $S_f(u)$ 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 theshold 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

- **Dataset from Deng et al, 2004**
- **4 data sources (Saccharomyces cerevisiae)**
  - Protein-Protein Interactions
    - **2,448 edges**
  - Protein Complexes
    - **30,731 edges**
  - Pfam Domains
    - **28,616 edges**
  - Expression Correlation
    - **1,366 edges**

Comparison w/ Existing Approaches

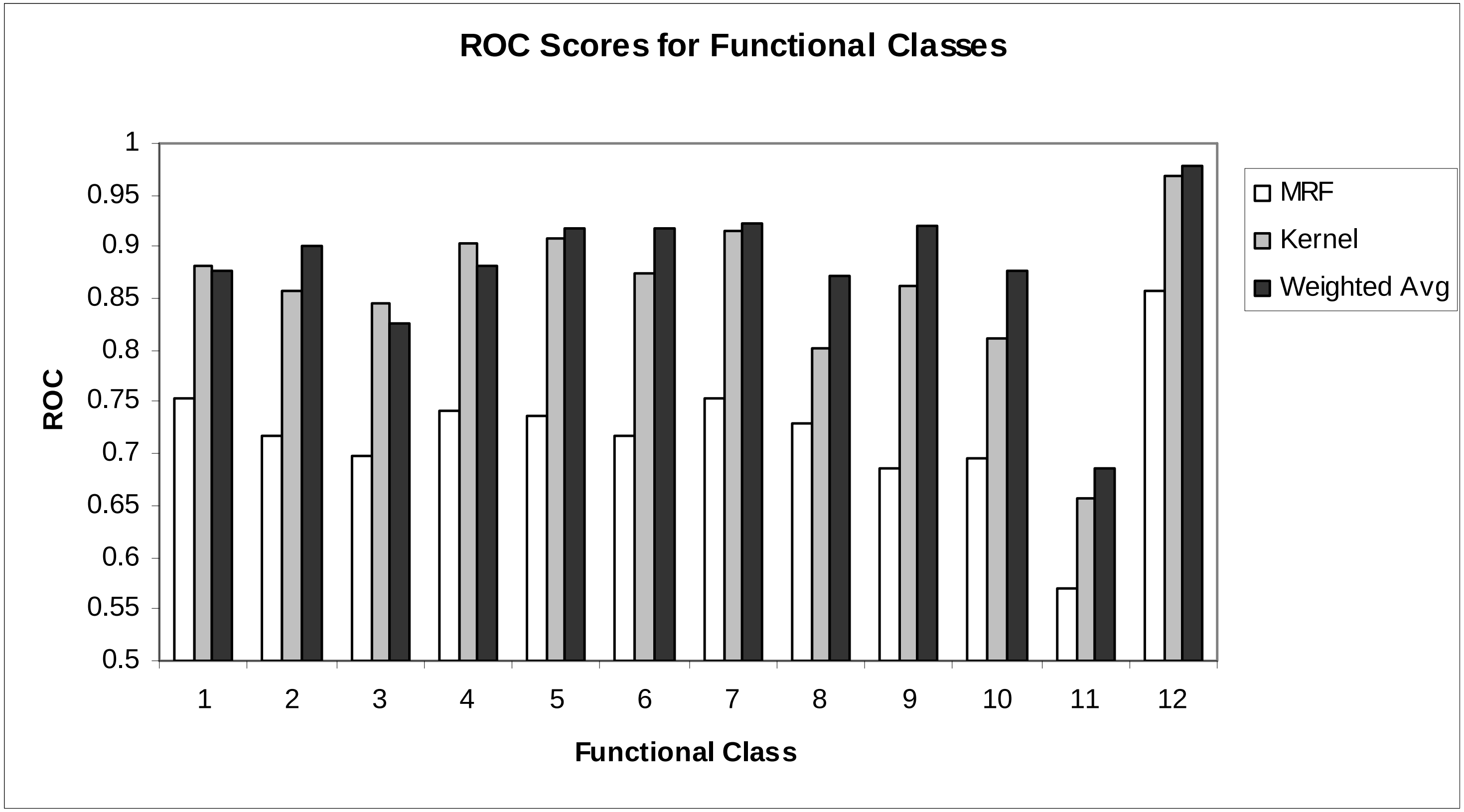
- **12 functional classes**

	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 the cellular environment	193
10	Cell fate	411
11	Control of cellular organization	192
12	Transport facilitation	306

Comparison w/ Existing Approaches

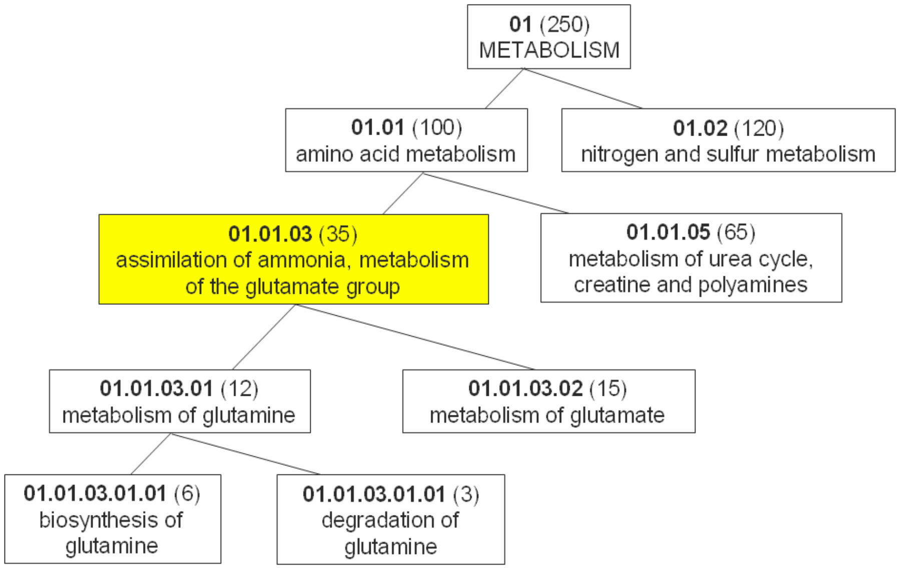
- **Validation Method (Lanckriet et al, 2004)**
  - Receiver Operating Characteristics (ROC)
  - True Positives vs False Positives
  - Area under ROC curve for each function
  - Averaged over 3 repetitions of 5-fold cross validation

Comparison w/ Existing Approaches



GO Terms Prediction for Yeast Proteins

- **Proteins from Saccharomyces 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

- **PPI**
  - BIND
  - 12,967 unique interactions betw yeast proteins
  - FS weight used as score

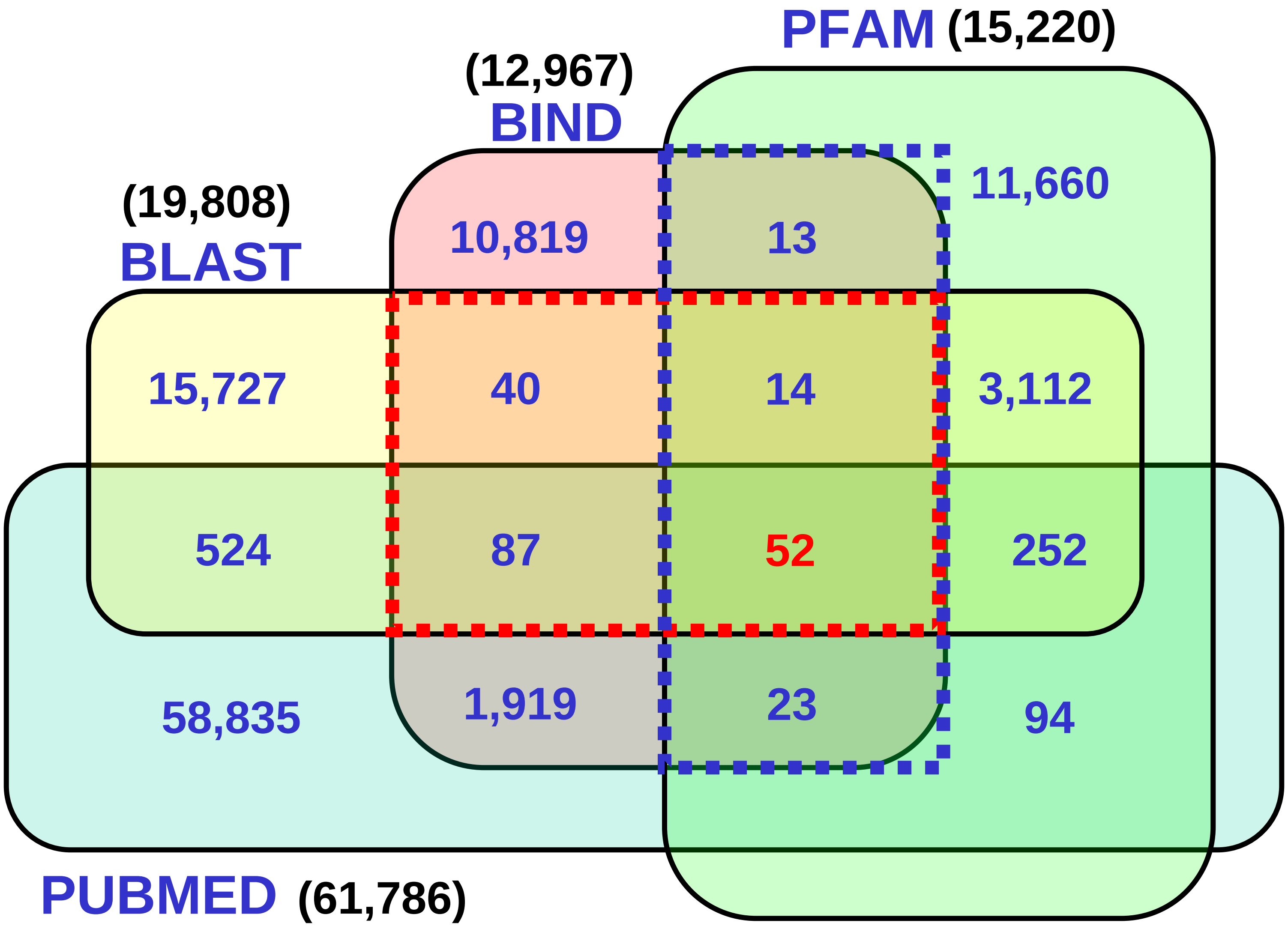
- **Protein Sequences**
  - Seqs from GO database (archive.godatabase.org)
  - Each yeast seq is aligned w/ rest using BLAST (cutoff E-Score = 1)
  - -log(e-score) used as score
  - Top 5 results w/ known annotations
  - 19,808 unique pairs involving yeast proteins

Data Sources

- **Pfam Domains**
  - SwissPfam database (http://www.sanger.ac.uk/Software/Pfam/ftp.shtml)
  - Precomputed Pfam domains for SwissProt and TrEMBL proteins w/ E-value threshold 0.01
  - Number of common domains used as score
  - 15,220 unique pairs involving yeast proteins

- **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

Multiple Data Sources



Validation

- **Precision vs Recall**
  - Precision

$$\frac{\sum_i^K k_i}{\sum_i^K m_i}$$

k_i is the number of functions correctly predicted for protein i

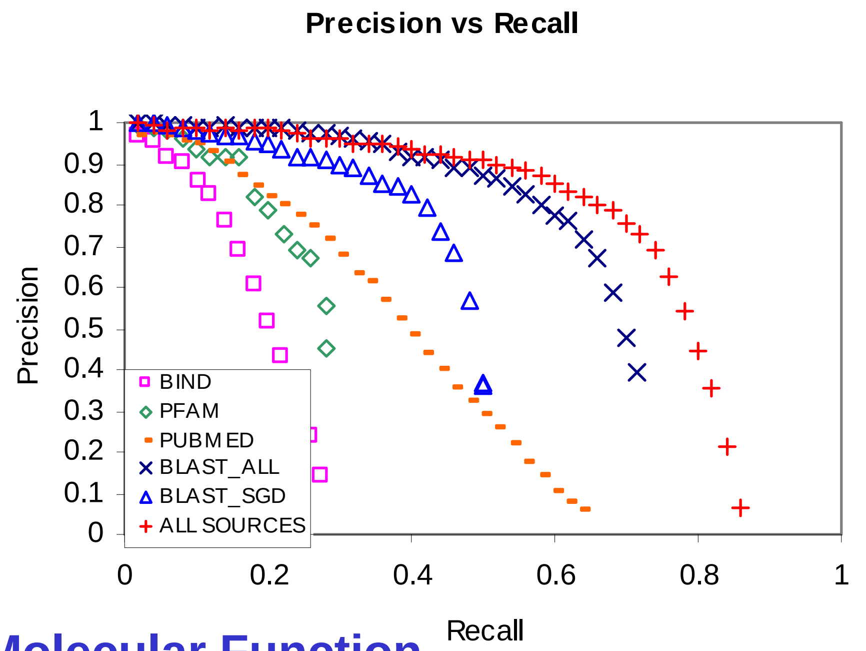
m_i is the number of functions predicted for protein i

  - Recall

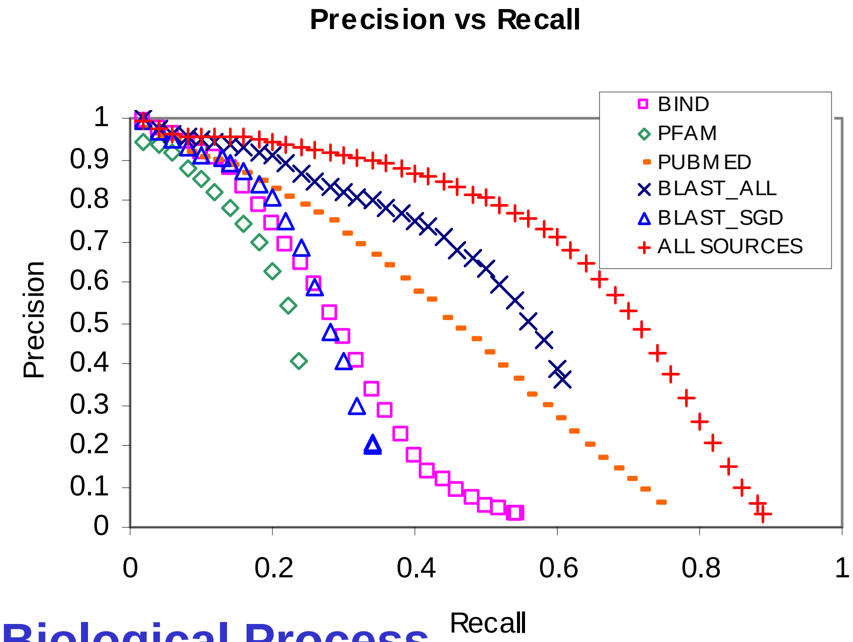
$$\frac{\sum_i^K k_i}{\sum_i^K n_i}$$

n_i is the number of functions annotated for protein i

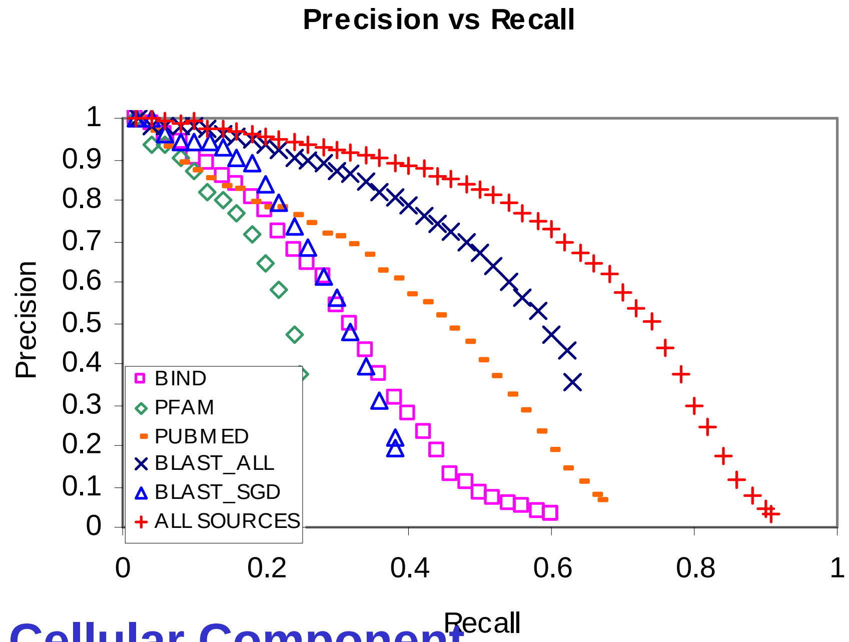
Combining all data sources outperforms any individual data source



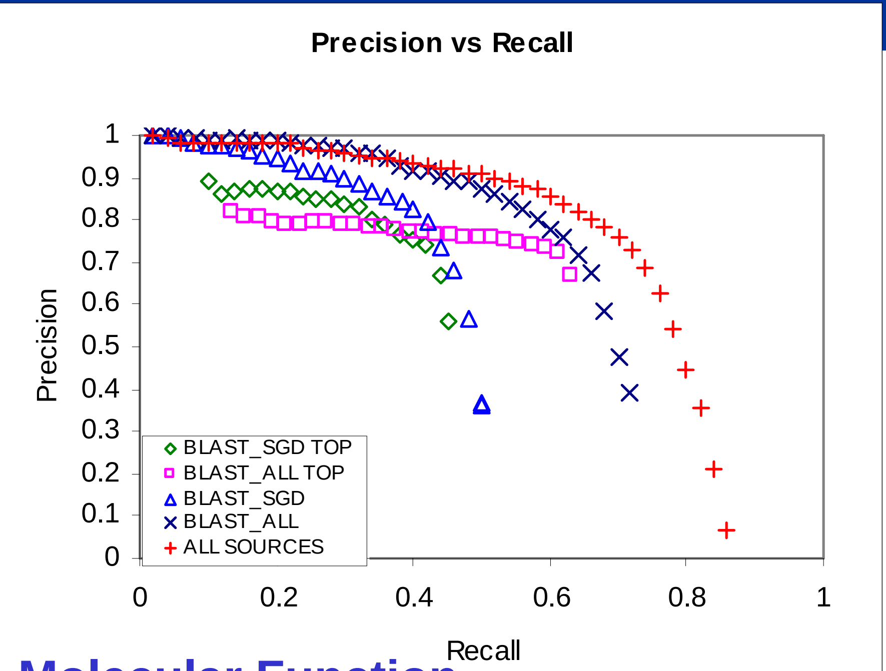
Molecular Function



Biological Process

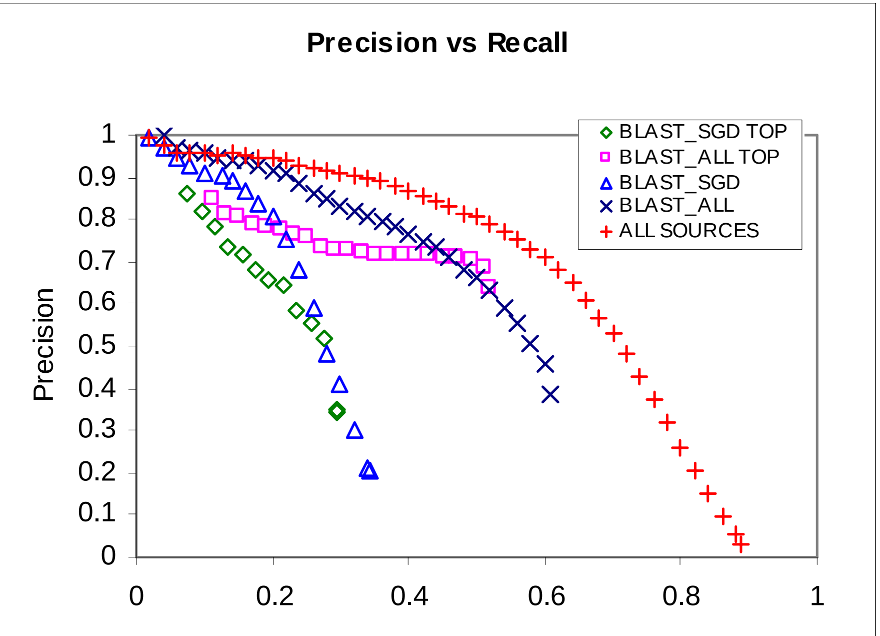


Cellular Component

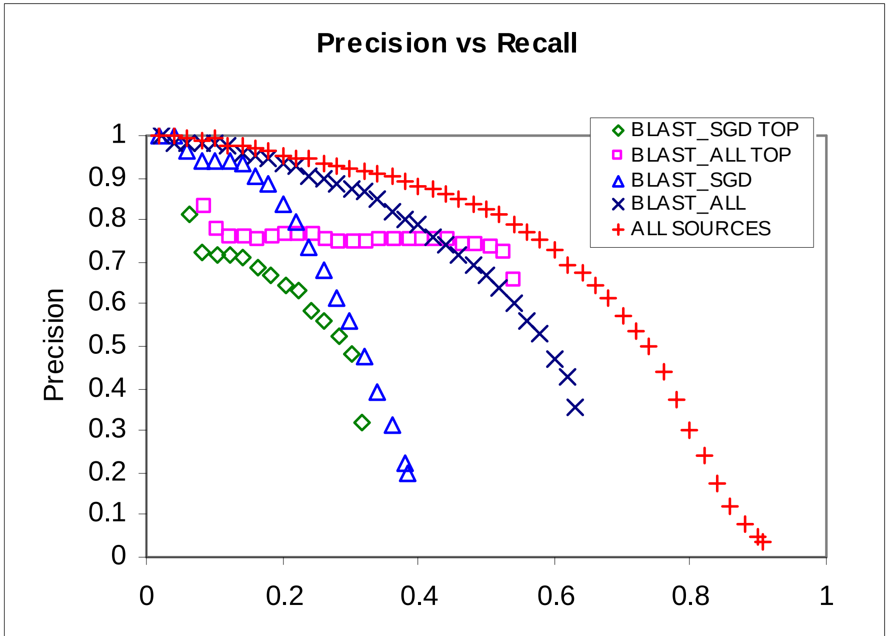


Molecular Function

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

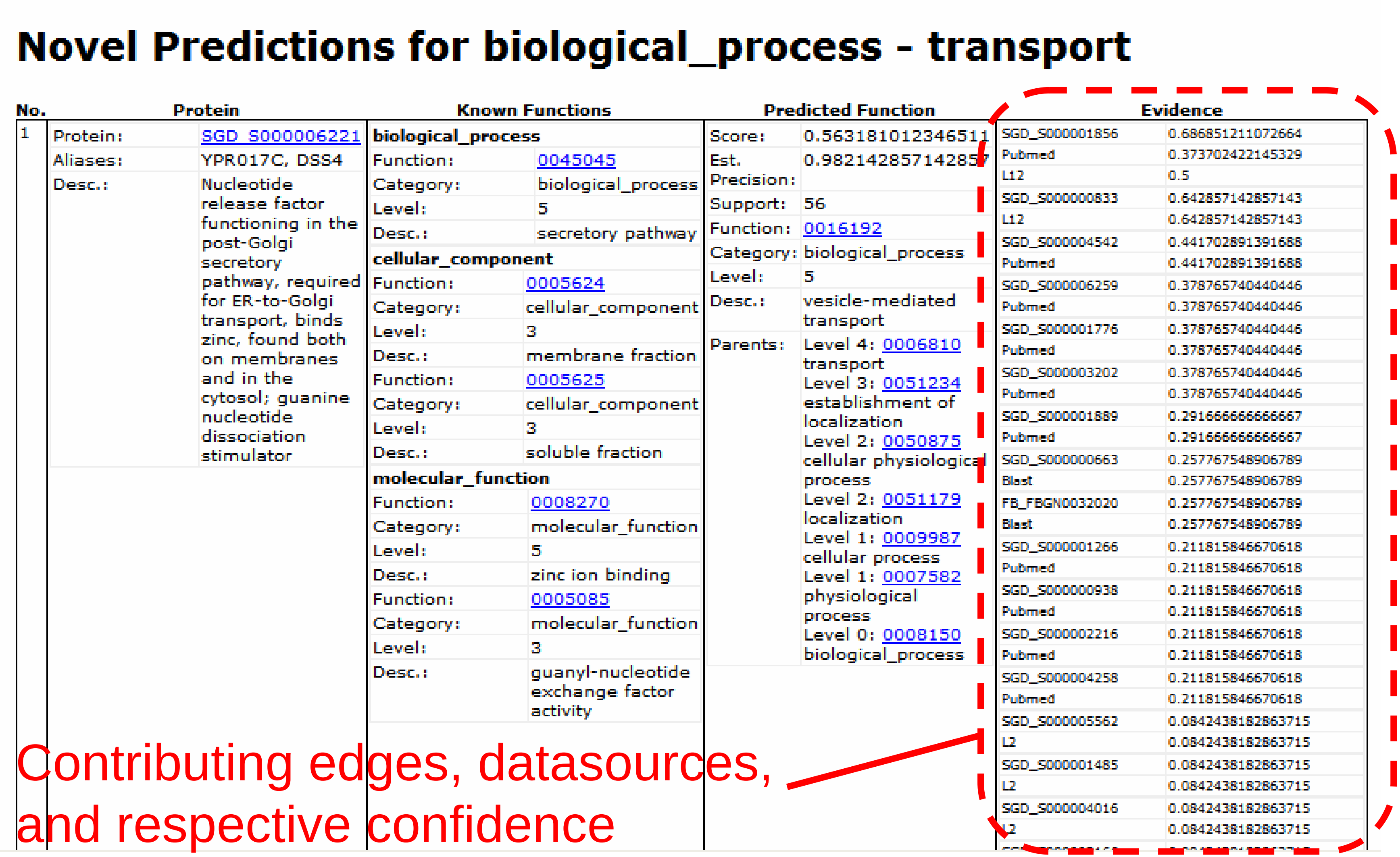


Biological Process



Cellular Component

Predictions



Novel Predictions for biological_process - transport

No.	Protein	Known Functions	Predicted Function	Evidence
1	Protein: SGD_S000006221 Aliases: YPR017C, DSS4 Desc.: Nucleotide release factor functioning in the post-Golgi secretory pathway, required for ER-to-Golgi transport, binds zinc, found both on membranes and in the cytosol; guanine nucleotide dissociation stimulator	**biological_process** Function: 0045045 Category: biological_process Level: 5 Desc.: secretory pathway **cellular_component** Function: 0005624 Category: cellular_component Level: 3 Desc.: membrane fraction Function: 0005625 Category: cellular_component Level: 3 Desc.: soluble fraction **molecular_function** Function: 0008270 Category: molecular_function Level: 5 Desc.: zinc ion binding Function: 0005085 Category: molecular_function Level: 3 Desc.: guanyl-nucleotide exchange factor activity	Score: 0.563181012346511 Est. Precision: 0.982142857142857 Support: 56 Function: 0016192 Category: biological_process Level: 5 Desc.: vesicle-mediated transport Parents: Level 4: 0006810 transport Level 3: 0051234 establishment of localization Level 2: 0050875 cellular physiological process Level 2: 0051179 localization Level 1: 0009987 cellular process Level 1: 0007582 physiological process Level 0: 0008150 biological_process	SGD_S000001856 0.686851211072664 Pubmed 0.373702422145329 L12 0.5 SGD_S000000833 0.642857142857143 L12 0.642857142857143 SGD_S000004542 0.441702891391688 Pubmed 0.441702891391688 SGD_S000006259 0.378765740440446 Pubmed 0.378765740440446 SGD_S000001776 0.378765740440446 Pubmed 0.378765740440446 SGD_S000003202 0.378765740440446 Pubmed 0.378765740440446 SGD_S000001889 0.291666666666667 Pubmed 0.291666666666667 SGD_S000000663 0.257767548906789 Blast 0.257767548906789 FB_FBGN0032020 0.257767548906789 Blast 0.257767548906789 SGD_S000001266 0.211815846670618 Pubmed 0.211815846670618 SGD_S000000938 0.211815846670618 Pubmed 0.211815846670618 SGD_S000002216 0.211815846670618 Pubmed 0.211815846670618 SGD_S000004258 0.211815846670618 Pubmed 0.211815846670618 SGD_S000005562 0.0842438182863715 L2 0.0842438182863715 SGD_S000001485 0.0842438182863715 L2 0.0842438182863715 SGD_S000004016 0.0842438182863715 L2 0.0842438182863715

Contributing edges, datasources, and respective confidence

Conclusions

- **We developed a simple graph-based method that combines multiple sources of data sources for function prediction**

- **Our method is simple, flexible and can report datasources contributing to each prediction**

- **We have shown that our method performs comparable, if not better, than existing approaches**

References

- Ashburner M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*. 25(1):25-29.
- Zhou, X., Kao, M.C., Wong, W.H. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U S A*. 99(20), 12783-88.
- Chua H.N., Sung W.K., Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics, 22:1623-1630.
- Deng M., Chen T., Sun F. (2004) An integrated probabilistic model for functional prediction of proteins*. J. Comp. Biol.* 11(2-3):463-75.
- Lanckriet G.R., et al. (2004) Kernel-based data fusion and its application to protein function prediction in yeast. *Proceedings of the Pacific Symposium on Biocomputing*, January 3-8, 2004. pp. 300-311.
- Cherry J.M., et al. (1997) Genetic and physical maps of Saccharomyces cerevisiae..*Nature*, 387(6632 Suppl):67-73.
- Lee I.,et al. Probabilistic functional network of yeast genes. *Science*. 306(5701):1555-8.
- Martin D.M., Berriman M., Barton G.J. (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes.
  BMC Bioinformatics. 5:178
- Cohen A.M., et al. (2005) Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. *BMC Bioinformatics*. 6:103
- Xiao G., Pan W. (2005) Gene function prediction by a combined analysis of gene expression data and protein-protein interaction data.
  J. Bioinform. Comp. Biol., 3(6):1371-89