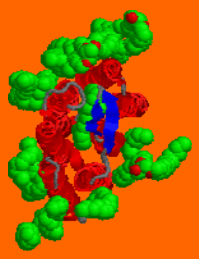
Guilt by Association

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

Plan

- **Protein Function Prediction**
  - Guilt by Association of Seq Similarity
- **Guilt by Association of Common Friends**
- **Guilt by Association of Multiple Types of Info**

Protein Function Prediction: Motivation & Challenges

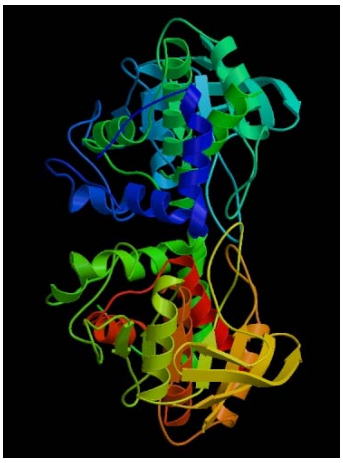


NUS National University of Singapore

Edinburgh, October 2007

NUS

- **A protein is a large complex molecule made up of one or more chains of amino acids**
- **Protein performs a wide variety of activities in the cell**



Function Assignment to Protein Seq

```
SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT
```

- **How do we attempt to assign a function to a new protein sequence?**

An Early Example of Seq Analysis

Source: Ken Sung

- **Doolittle et al. (*Science*, July 1983) searched for platelet-derived growth factor (PDGF) in his own DB. He found that PDGF is similar to v-sis oncogene**

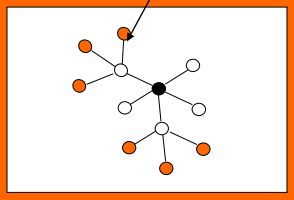
```
PDGF-2  1       SLGSLTIAEPAMIAECKTREEVFCICRRL?DR?? 34
p28sis 61 LARGKRSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTN 100
```

⇒ **"Guilt by association" of sequence similarity!**

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
  - Fusion of multiple types of info

Guilt by Association of Common Friends: Protein Function Prediction from Protein Interactions

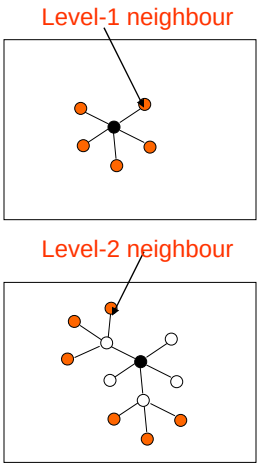


Protein Interaction Based Approaches

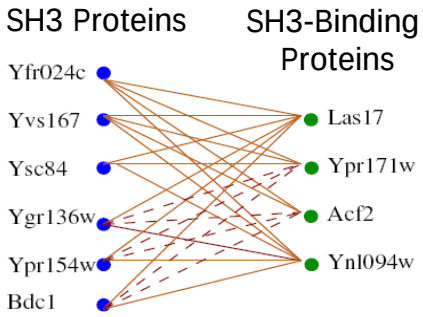
- **Neighbour counting** (Schwikowski et al, 2000)
  - Rank function based on freq in interaction partners
- **Chi-square** (Hishigaki et al, 2001)
  - Chi square statistics using expected freq of functions in interaction partners
- **Markov Random Fields** (Deng et al, 2003; Letovsky et al, 2003)
  - Belief propagation exploit unannotated proteins for prediction
- **Simulated Annealing** (Vazquez et al, 2003)
  - Global optimization by simulated annealing
  - Exploit unannotated proteins for prediction
- **Clustering** (Brun et al, 2003; Samanta et al, 2003)
  - Functional distance derived from shared interaction partners
  - Clusters based on functional distance represent proteins with similar functions
- **Functional Flow** (Nabieva et al, 2004)
  - Assign reliability to various expt sources
  - Function "flows" to neighbour based on reliability of interaction and "potential"
- **Indirect Functional Assoc** (Chua et al, 2006)
  - Identification of reliable common interaction partners

Functional Association Thru Interactions

- **Direct functional association:**
  - Interaction partners of a protein are likely to share functions w/ it
  - Proteins from the same pathways are likely to interact
- **Indirect functional association**
  - Proteins that share interaction partners with a protein may also likely to share functions w/ it
  - Proteins that have common biochemical, physical properties and/or subcellular localization are likely to bind to the same proteins



An Illustrative Case of Indirect Functional Association?



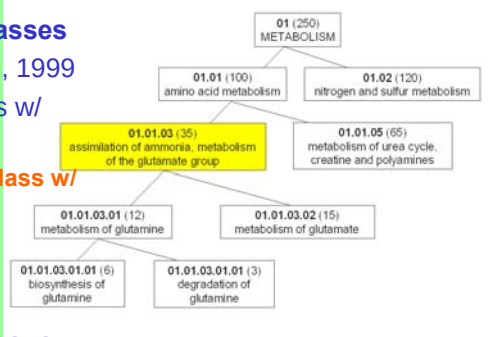
- **Is indirect functional association plausible?**
- **Is it found often in real interaction data?**
- **Can it be used to improve protein function prediction from protein interaction data?**

Materials

- **Protein interaction data from General Repository for Interaction Datasets (GRID)**
  - Data from published large-scale interaction datasets and curated interactions from literature
  - 13,830 unique and 21,839 total interactions
  - Includes most interactions from the Biomolecular Interaction Network (BIND) and the Munich Information Center for Protein Sequences (MIPS)
- **Functional annotation (FunCat 2.0) from Comprehensive Yeast Genome Database (CYGD) at MIPS**
  - 473 Functional Classes in hierarchical order

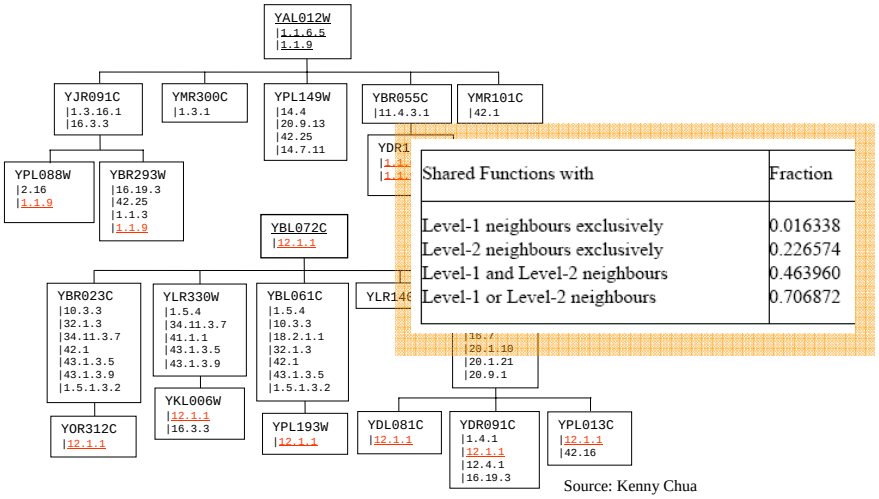
Validation Methods

- **Informative Functional Classes**
  - Adopted from Zhou et al, 1999
  - Select functional classes w/
    - **at least 30 members**
    - **no child functional class w/ at least 30 members**



- **Leave-One-Out Cross Validation**
  - Each protein with annotated function is predicted using all other proteins in the dataset

Freq of Indirect Functional Association



Shared Functions with	Fraction
Level-1 neighbours exclusively	0.016338
Level-2 neighbours exclusively	0.226574
Level-1 and Level-2 neighbours	0.463960
Level-1 or Level-2 neighbours	0.706872

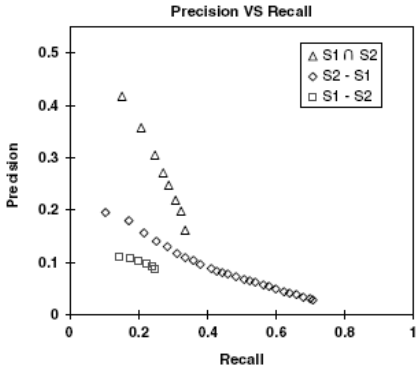
Source: Kenny Chua

Prediction Power By Majority Voting

- **Remove overlaps in level-1 and level-2 neighbours to study predictive power of "level-1 only" and "level-2 only" neighbours**
- **Sensitivity vs Precision analysis**

$$PR = \frac{\sum_i^K k_i}{\sum_i^K m_i} \qquad SN = \frac{\sum_i^K k_i}{\sum_i^K n_i}$$

- n_i is no. of fn of protein i
- m_i is no. of fn predicted for protein i
- k_i is no. of fn predicted correctly for protein i



⇒ **"level-2 only" neighbours performs better**

⇒ **L1 ∩ L2 neighbours has greatest prediction power**

Functional Similarity Estimate: Czekanowski-Dice Distance

- **Functional distance between two proteins** (Brun et al, 2003)

$$D(u,v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v| + |N_u \cap N_v|}$$

- N_k is the set of interacting partners of k
- X Δ Y is symmetric diff betw two sets X and Y
- Greater weight given to similarity

Is this a good measure if u and v have very diff number of neighbours?

⇒ **Similarity can be defined as**

$$S(u,v) = 1 - D(u,v) = \frac{2X}{2X + (Y + Z)}$$

Functional Similarity Estimate: FS-Weighted Measure

- **FS-weighted measure**

$$S(u,v) = \frac{2|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v|} \times \frac{2|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v|}$$

  - **N_k is the set of interacting partners of k**
  - **Greater weight given to similarity**

⇒ **Rewriting this as**

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

Correlation w/ Functional Similarity

- **Correlation betw functional similarity & estimates**

Neighbours	CD-Distance	FS-Weight
S_1	0.471810	0.498745
S_2	0.224705	0.298843
$S_1 \cup S_2$	0.224581	0.29629

- **Equiv measure slightly better in correlation w/ similarity for L1 & L2 neighbours**

Source: Kenny Chua

Reliability of Expt Sources

- **Diff Expt Sources have diff reliabilities**
  - Assign reliability to an interaction based on its expt sources (Nabieva et al, 2004)
- **Reliability betw u and v computed by:**

$$r_{u,v} = 1 - \prod_{i \in E_{u,v}} (1 - r_i)$$

  - **r_i is reliability of expt source i,**
  - **$E_{u,v}$ is the set of expt sources in which interaction betw u and v is observed**

Source	Reliability
Affinity Chromatography	0.823077
Affinity Precipitation	0.455904
Biochemical Assay	0.666667
Dosage Lethality	0.5
Purified Complex	0.891473
Reconstituted Complex	0.5
Synthetic Lethality	0.37386
Synthetic Rescue	1
Two Hybrid	0.265407

Functional Similarity Estimate: 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**

⇒ **Rewriting**

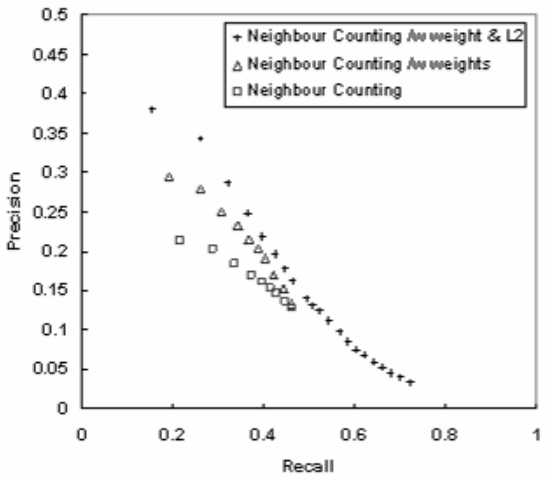
$$S(u,v) = \frac{2X}{2X+Y} \times \frac{2X}{2X+Z}$$

Integrating Reliability

- **Equiv measure shows improved correlation w/ functional similarity when reliability of interactions is considered:**

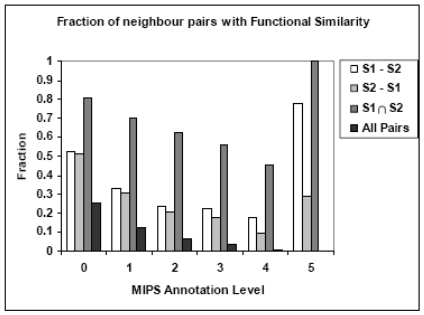
Neighbours	CD-Distance	FS-Weight	FS-Weight R
S_1	0.471810	0.498745	0.532596
S_2	0.224705	0.298843	0.375317
$S_1 \cup S_2$	0.224581	0.29629	0.363025

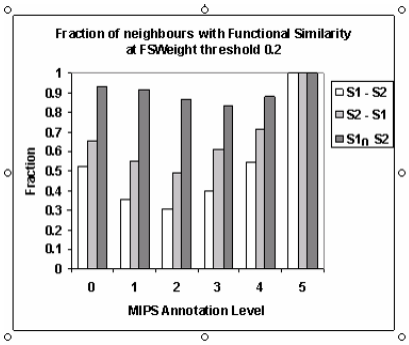
Improvement to Prediction Power by Majority Voting



Considering only neighbours w/ FS weight > 0.2

Improvement to Over-Rep of Functions in Neighbours





Source: Kenny Chua

Use L1 & L2 Neighbours for Prediction

- **FS-weighted Average**

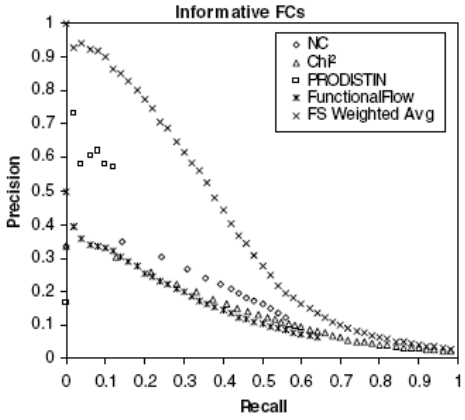
$$f_x(u) = \frac{1}{Z}\left[\lambda r_{\text{int}}\pi_x + \sum_{v\in N_u}\left(S_{TR}(u,v)\delta(v,x) + \sum_{w\in N_v} S_{TR}(u,w)\delta(w,x)\right)\right]$$

  - r_{int} is fraction of all interaction pairs sharing function
  - λ is weight of contribution of background freq
  - $\delta(k, x) = 1$ if k has function x, 0 otherwise
  - N_k is the set of interacting partners of k
  - π_x is freq of function x in the dataset
  - Z is sum of all weights

$$Z = 1 + \sum_{v\in N_u}\left(S_{TR}(u,v) + \sum_{w\in N_v} S_{TR}(u,w)\right)$$

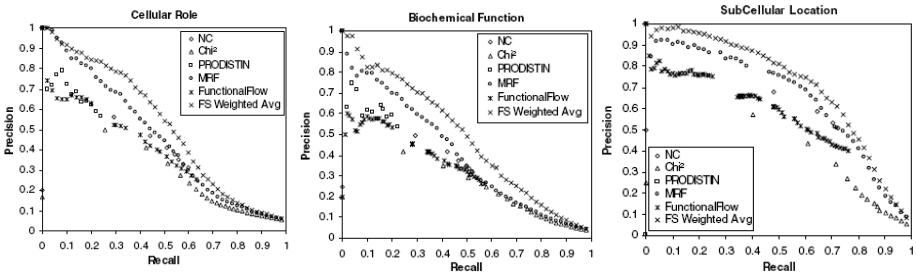
Performance of FS-Weighted Averaging

- **LOOCV comparison with Neighbour Counting, Chi-Square, PRODISTIN**



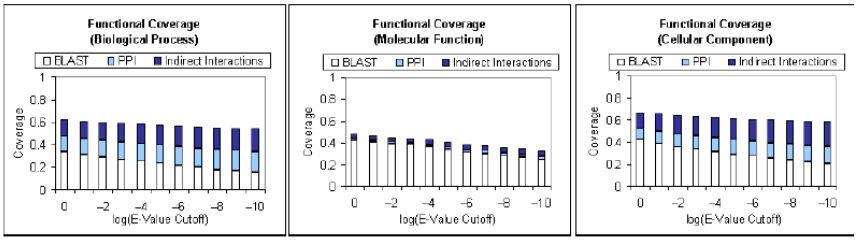
Performance of FS-Weighted Averaging

- **Dataset from Deng et al, 2003**
  - Gene Ontology (GO) Annotations
  - MIPS interaction dataset
- **Comparison w/ Neighbour Counting, Chi-Square, PRODISTIN, Markov Random Field, FunctionalFlow**



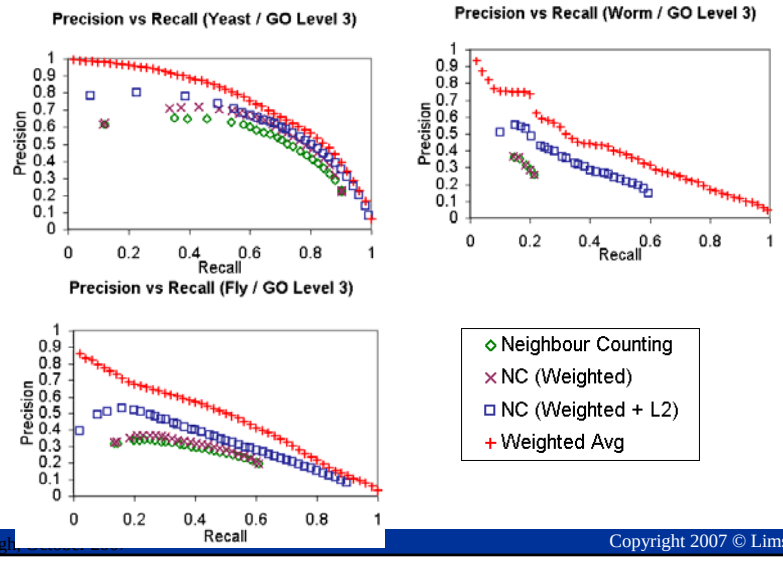
Freq of Indirect Functional Association in Other Genomes

D. melanogaster



Genome	Annotation	S_1-S_2	S_2-S_1	$S_1 \cap S_2$	$S_1 \cup S_2$
S. cerevisiae	MIPS	0.007193	0.226574	0.463960	0.706872
D. melanogaster	GO	0.008801	0.168622	0.138138	0.315561
C. elegans	GO	0.007193	0.051237	0.061080	0.119510

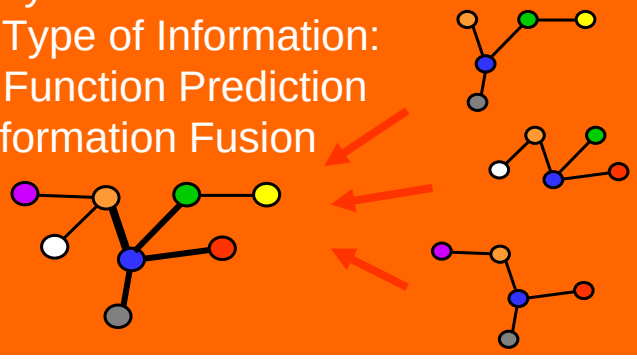
Effectiveness of FS Weighted Averaging in Other Genomes



Conclusions

- **Indirect functional association is plausible**
- **It is found often in real interaction data**
- **It can be used to improve protein function prediction from protein interaction data**
- **It should be possible to incorporate interaction networks extracted by literature in the inference process within our framework for good benefit**

Guilt by Association of Multiple Type of Information: Protein Function Prediction by 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

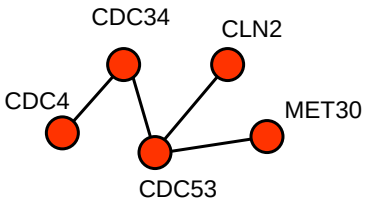
Motivation

- **Problems:**
  - Complex models such as MRF and Kernel Fusion are computationally expensive
  - Difficult or not possible to identify contributing sources in a prediction

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

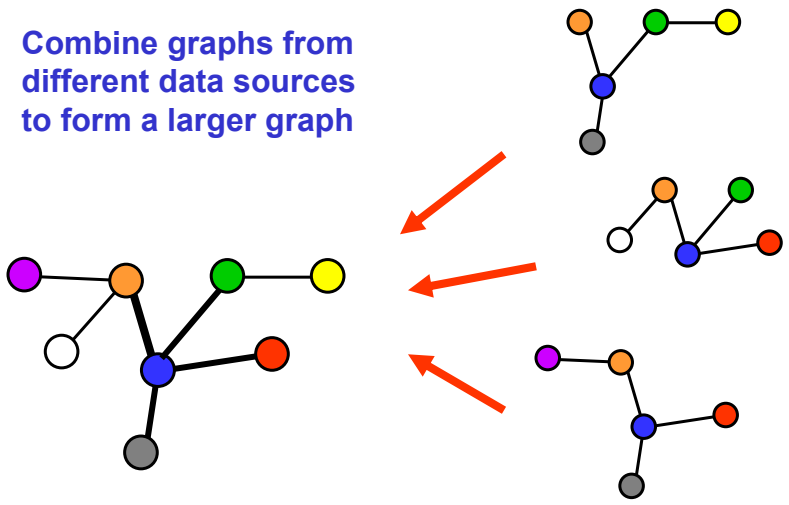
Strategy – Step 1

- **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



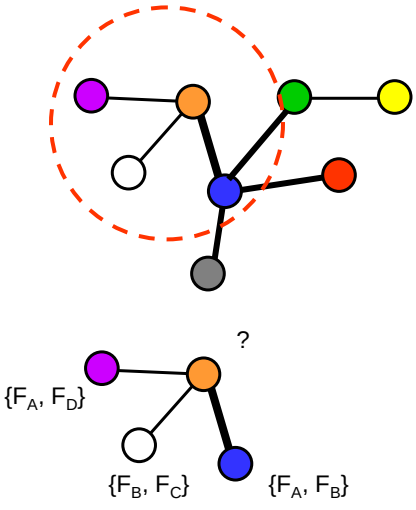
Strategy – Step 2

- **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

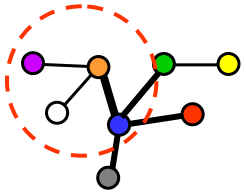
- **Scores may come in many forms**
  - E.g., Blast e-values, Pearson's correlation
- **A simple approach to discretization**
  - Split ranges into n equal intervals
  - Each interval becomes a new subtype
  - Assume linearity in range
  - Other strategies possible

Combination of Confidence

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

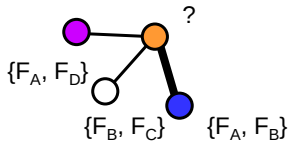
$$r_{u,v,f} = 1 - \prod_{k \in D_{u,v}} (1 - p(k, f))$$

  - 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

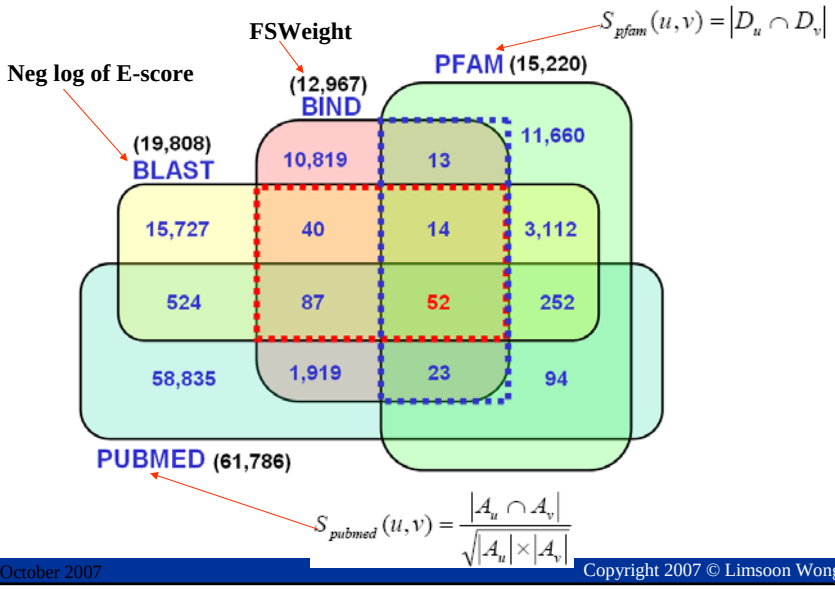
- **Weighted Average**



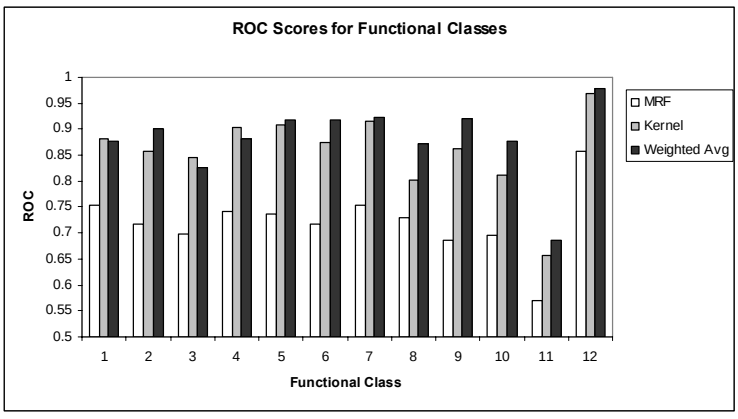
$$S_f(u) = \frac{\sum_{v \in N_u} (e_f(v) \times r_{u,v,f})}{1 + \sum_{v \in N_u} r_{u,v,f}}$$

  - $S_f(u)$ 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)

An Expt on Multiple Data Sources

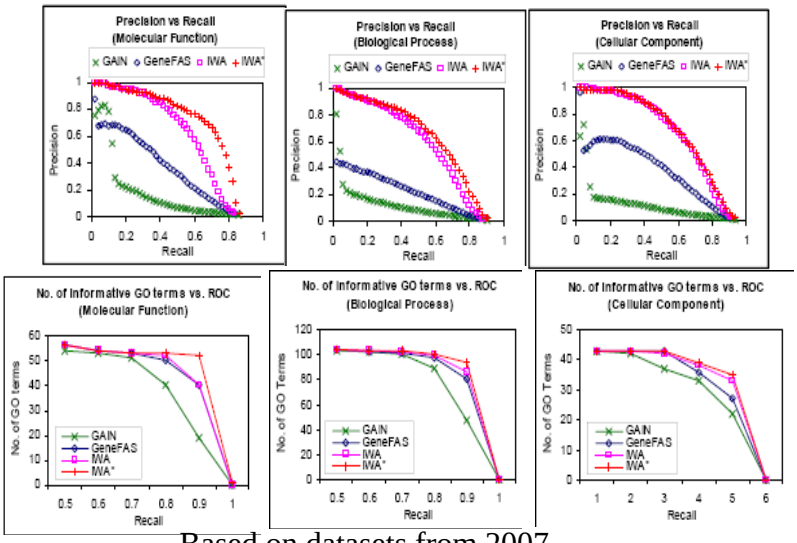


Comparison w/ Existing Approaches

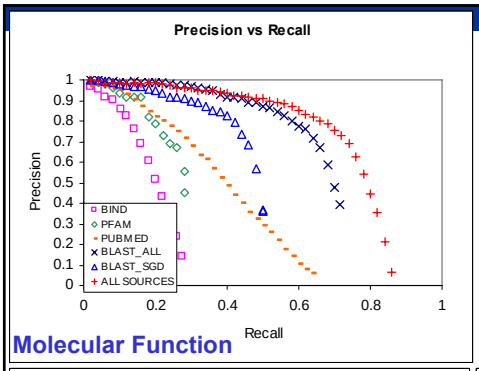


Based on datasets from 2004

Comparison w/ Existing Approaches

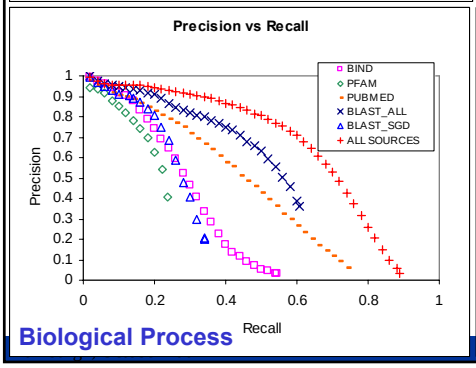


Based on datasets from 2007

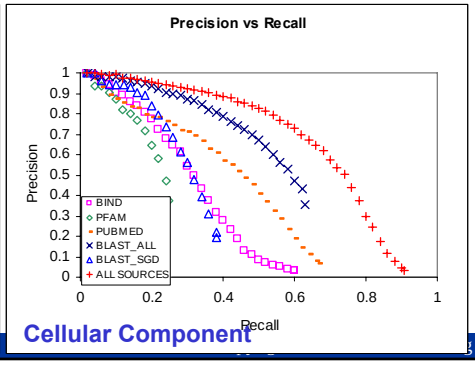


Molecular Function

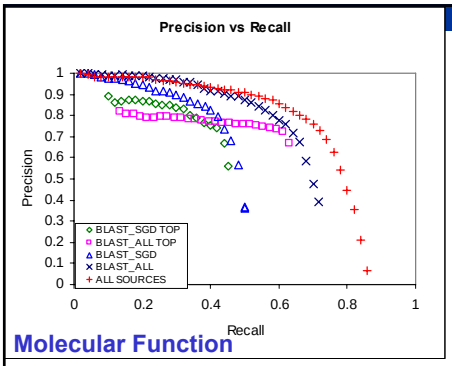
Combining all data sources outperforms any individual data source



Biological Process

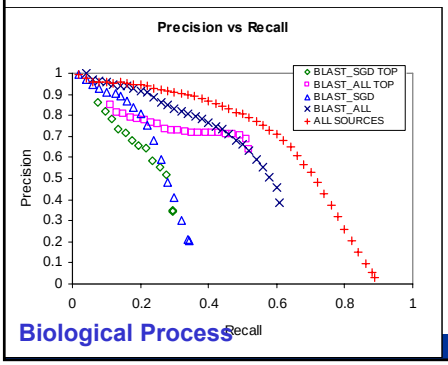


Cellular Component

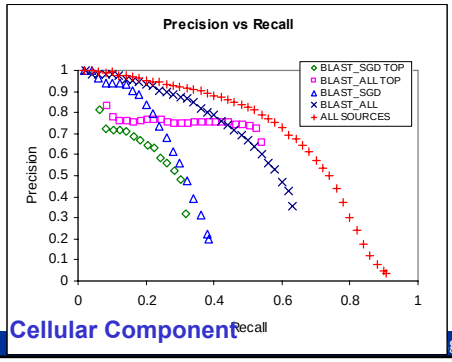


Molecular Function

- 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

- 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

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

- **H. N. Chua, W.K. Sung, & L. Wong. "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions". *Bioinformatics*, 22:1623-1630, 2006**

Any Question?

Edinburgh, October 2007