Guilt by Association of Common Interaction Partners

Limsoon Wong

Joint work with 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**
- **…**

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"**

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

Level-1 neighbour
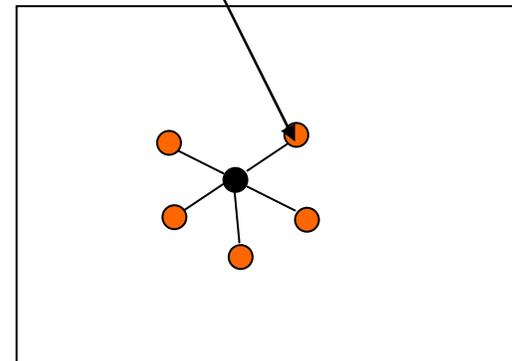
Level-2 neighbour
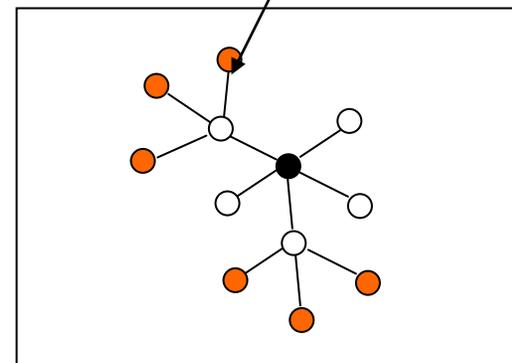
An illustrative Case of Indirect Functional Association?

SH3 Proteins   SH3-Binding Proteins
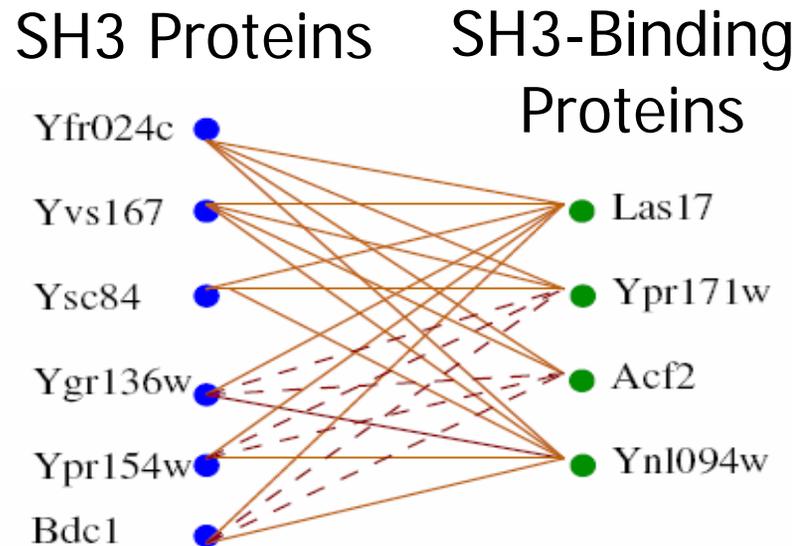
- **Is *indirect functional association* plausible?**
- **Is it found often in real interaction data?**
- **Can it be used to improve protein function prediction from protein interaction data?**

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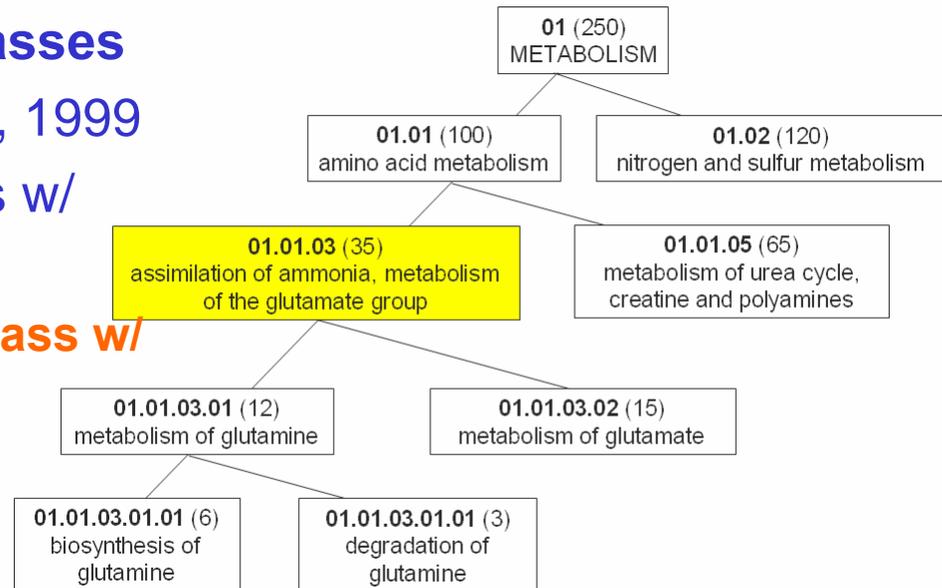
- **Leave-One-Out Cross Validation**
  - Each protein with annotated function is predicted using all other proteins in the dataset

Freq of Indirect Functional Association

```
                          ┌─────────────┐
                          │ YAL012W     │
                          │|1.1.6.5     │
                          │|1.1.9       │
                          └─────────────┘
```

YJR091C	YMR300C	YPL149W	YBR055C	YMR101C
|1.3.16.1	|1.3.1	|14.4	|11.4.3.1	|42.1
|16.3.3			|20.9.13		
			|42.25		
			|14.7.11		

YDR1...
|1.1.
|1.1.

YPL088W	YBR293W
|2.16	|16.19.3
|1.1.9	|42.25
	|1.1.3
	|1.1.9

```
                ┌─────────────┐
                │ YBL072C     │
                │|12.1.1      │
                └─────────────┘
```

YBR023C	YLR330W	YBL061C	YLR140
|10.3.3	|1.5.4	|1.5.4	
|32.1.3	|34.11.3.7	|10.3.3	
|34.11.3.7	|41.1.1	|18.2.1.1	
|42.1	|43.1.3.5	|32.1.3	
|43.1.3.5	|43.1.3.9	|42.1	
|43.1.3.9		|43.1.3.5	
|1.5.1.3.2		|1.5.1.3.2	

|16.7
|20.1.10
|20.1.21
|20.9.1

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

YKL006W
|12.1.1
|16.3.3

YOR312C
|12.1.1

YPL193W
|12.1.1

YDL081C
|12.1.1

| YDR091C	YPL013C
|1.4.1	|12.1.1
|12.1.1	|42.16
|12.4.1	
|16.19.3	

Source: Kenny Chua

Over-Rep of Functions in Neighbours

- **Functional Similarity:**

$$S(i, j) = \frac{\left| F_i \cap F_j \right|}{\left| F_i \cup F_j \right|}$$

- **where F_k is the set of functions of protein k**

- **L1 ∩ L2 neighbours show greatest over-rep**

- **L3 neighbours show little observable over-rep**

Fraction of neighbour pairs with Functional Similarity
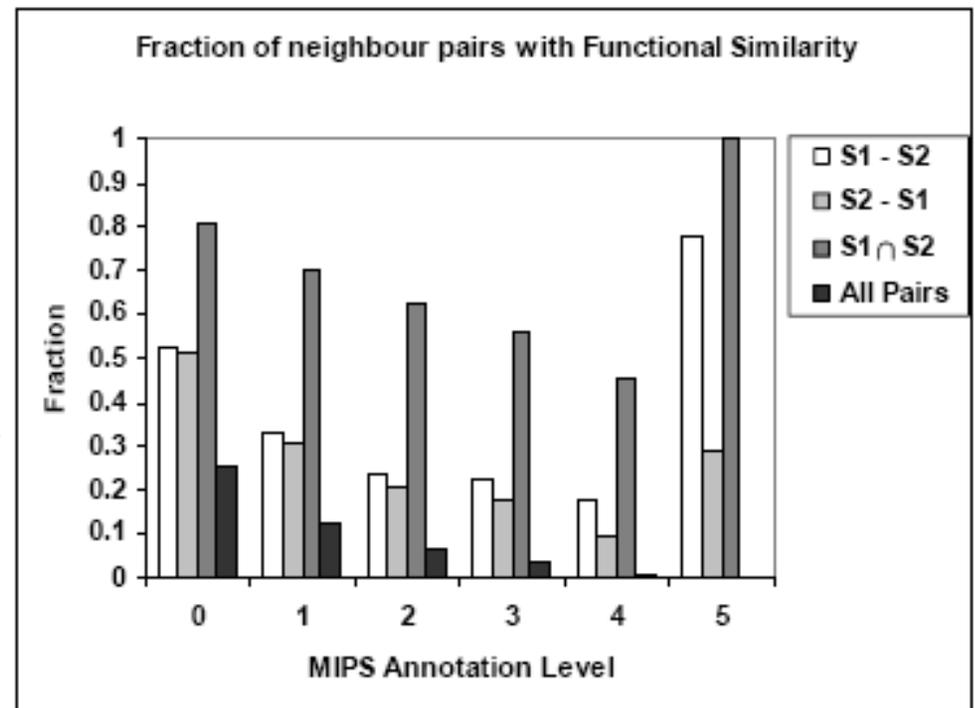
Prediction Power By Majority Voting

- **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**

Precision VS Recall

Legend:
- △ S1 ∩ S2
- ◇ S2 - S1
- □ S1 - S2
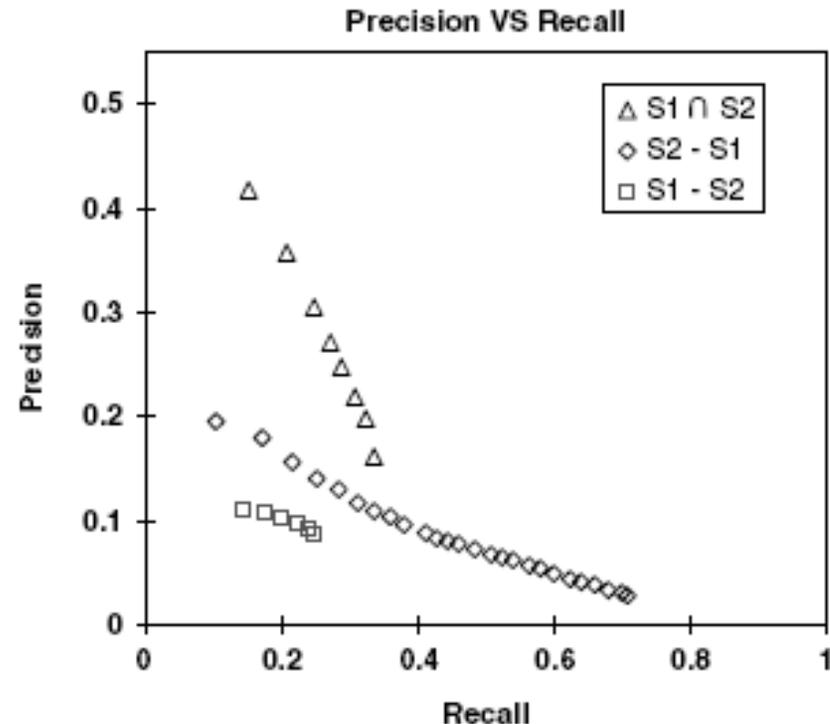
$\Rightarrow$ **"level-2 only" neighbours performs better**

$\Rightarrow$ **L1 ∩ L2 neighbours has greatest prediction power**

Functional Similarity Estimate: Czekanowski-Dice Distance

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

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

- **N_k is the set of interacting partners of k**
- **X $\triangle$ 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?

$\Rightarrow$ **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**

$\Rightarrow$ **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**

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:**

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

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

$\Rightarrow$ **Rewriting**

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

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
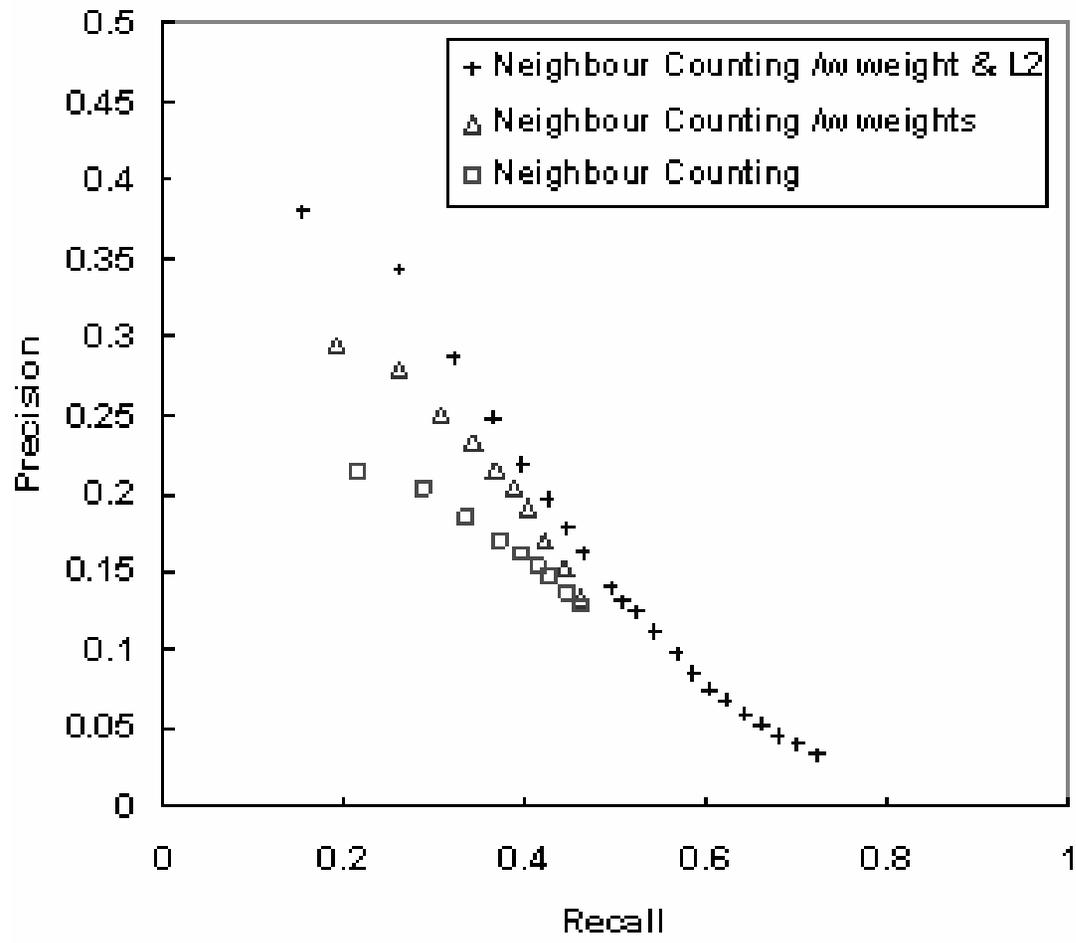
Considering only neighbours w/ FS weight > 0.2

Improvement to
Over-Rep of Functions in Neighbours
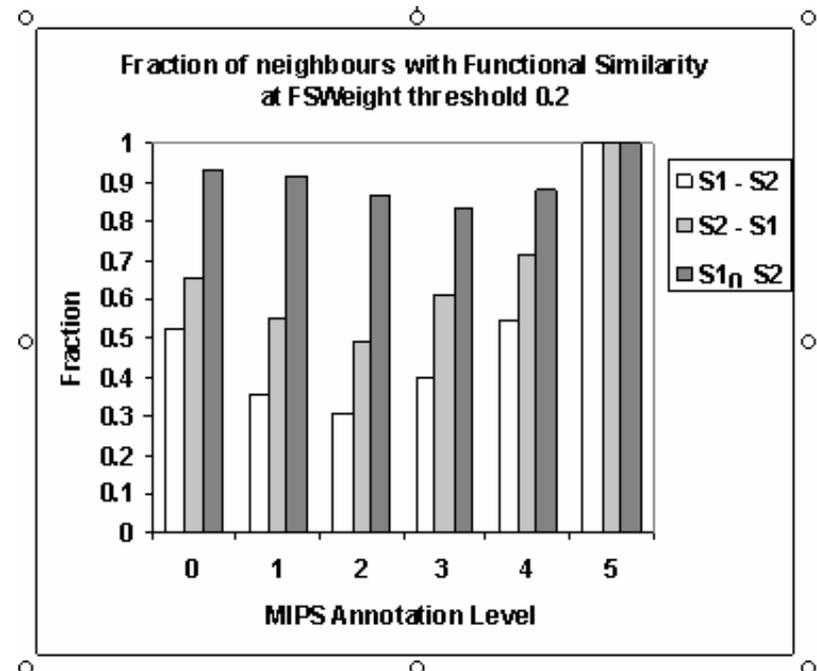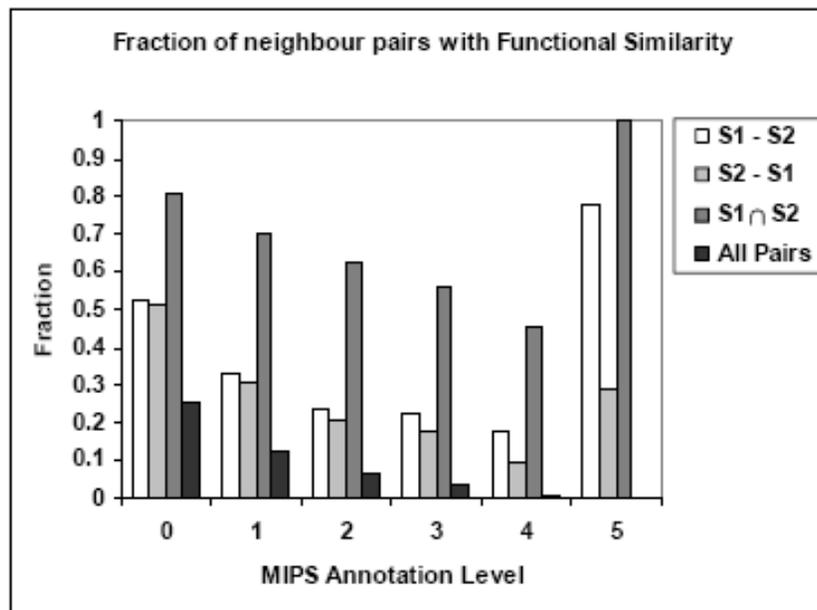
Use L1 & L2 Neighbours for Prediction

- **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**
- δ**(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**
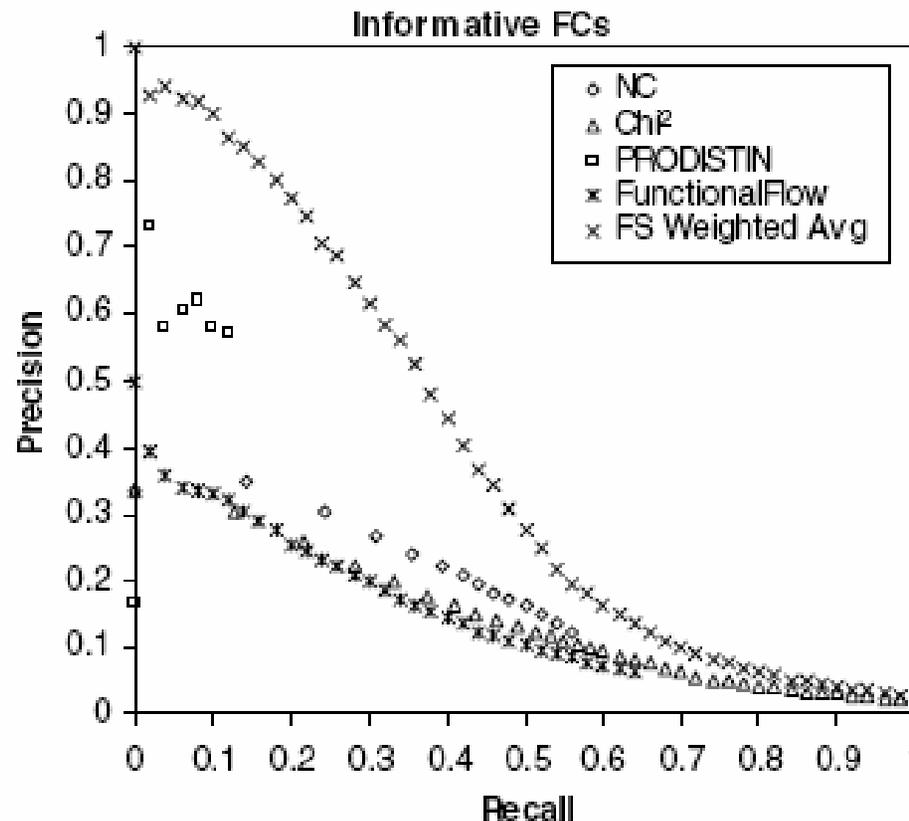
Performance of FS-Weighted Averaging

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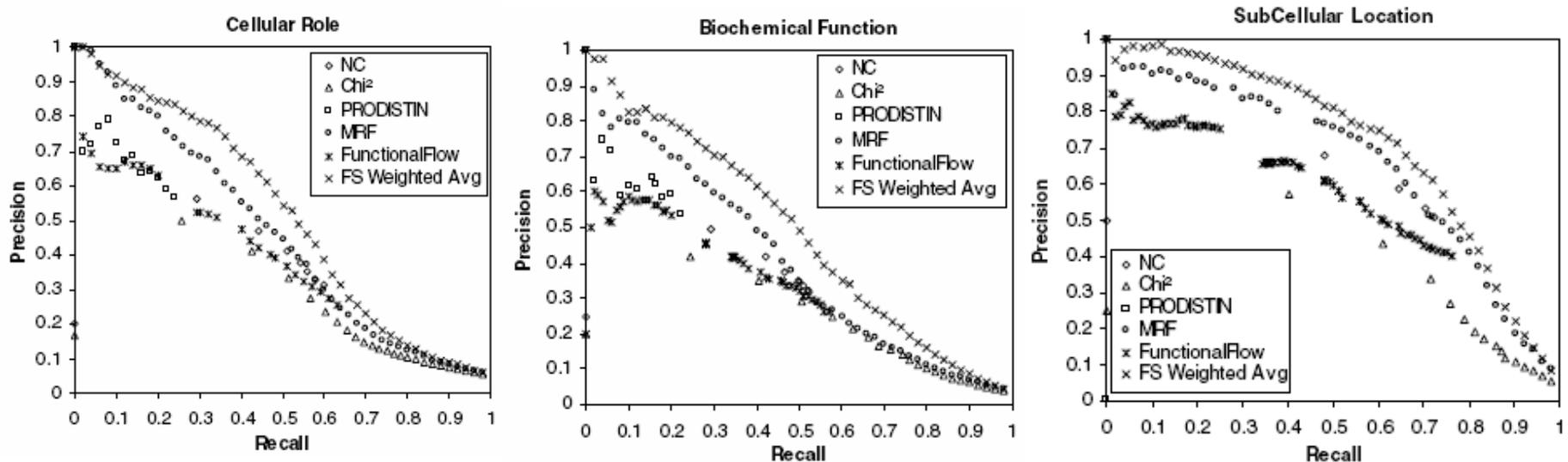
Performance of FS-Weighted Averaging

- **Correct Predictions made on at least 1 function vs Number of predictions made per protein**
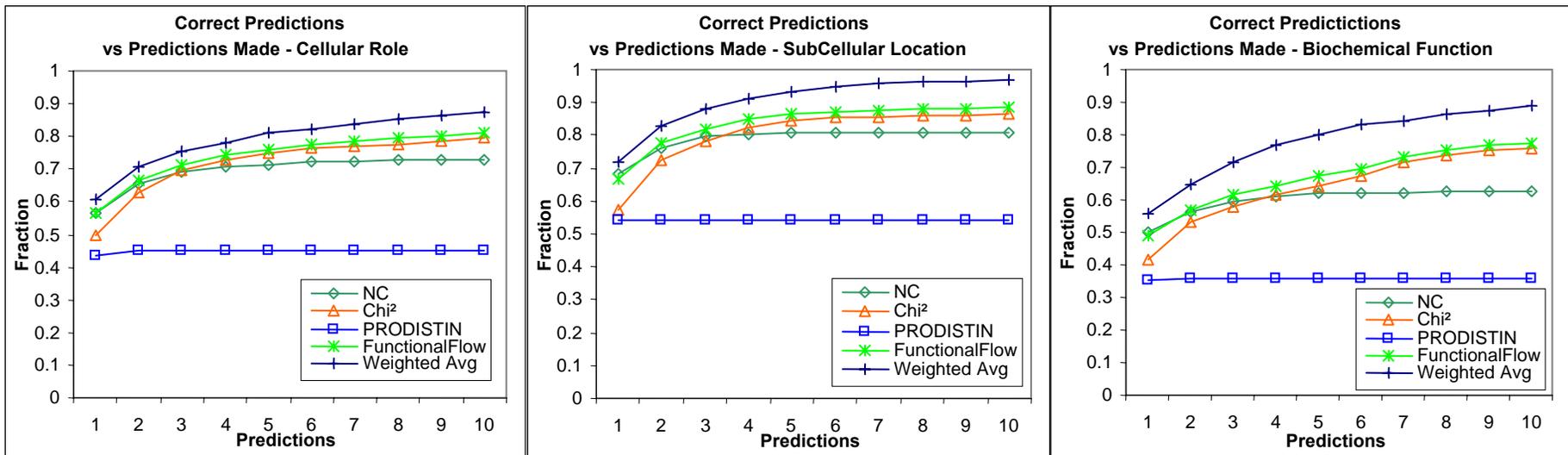
# Freq of Indirect Functional Association in Other Genomes

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
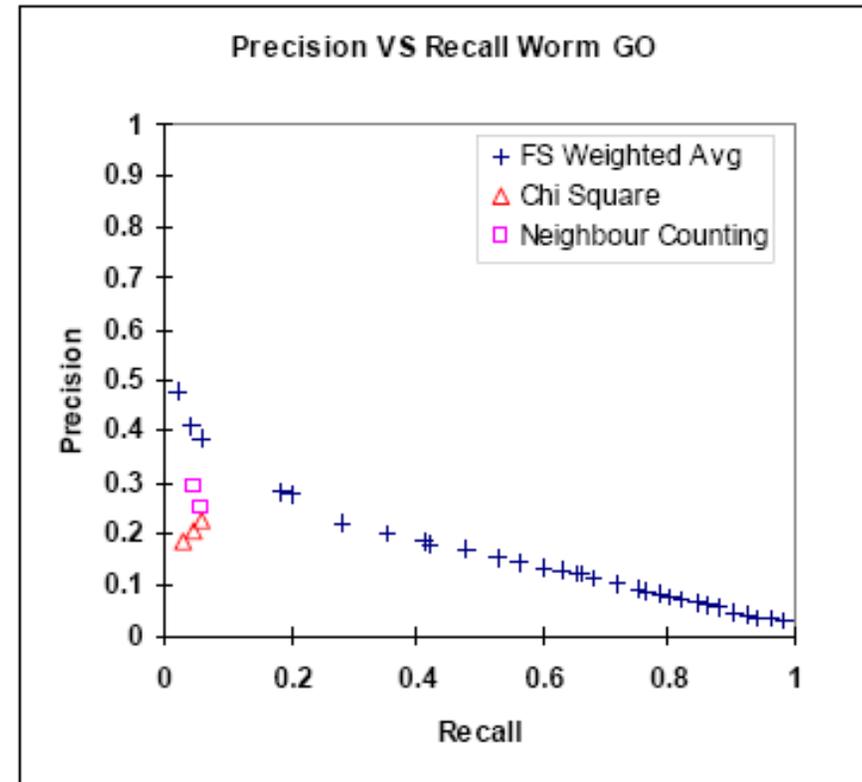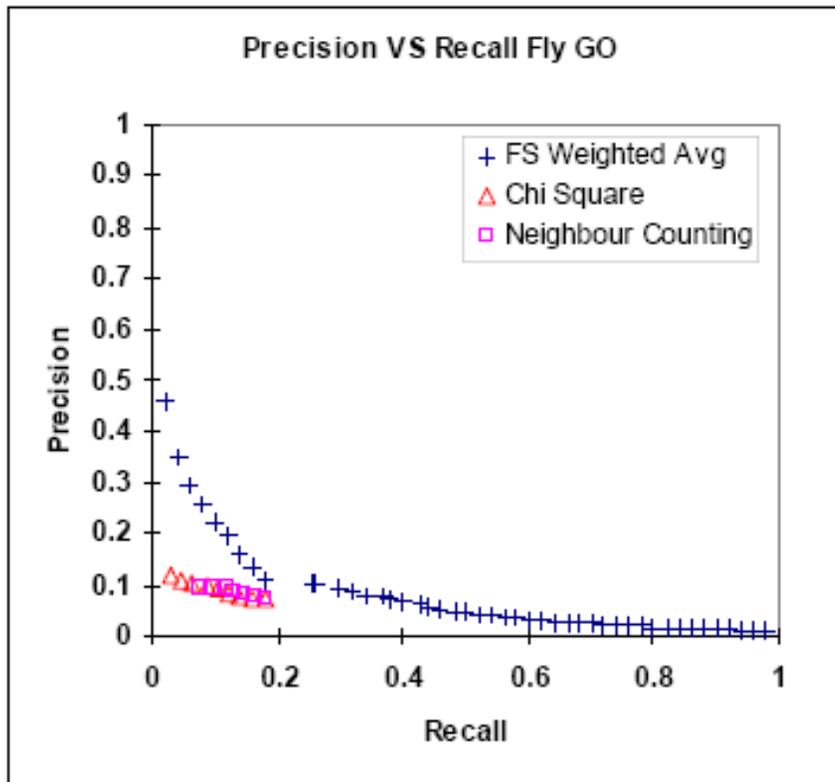
Conclusions

- **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**

Main 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.**

- **Hon Nian Chua, Wing Kin Sung, Limsoon Wong. Guilt by Indirect Functional Association, *AFP 2006*. To appear.**

Other References

- **Breitkreutz, B. J., Stark, C. and Tyers, N. (2003) The GRID: The General Repository for Interaction Datasets. Genome Biology, 4:R23**
- **Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., Jacq, B. (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol. 5(1):R6**
- **Deng, M., Zhang, K., Mehta, S.Chen, T. and Sun, F. Z. (2003) Prediction of protein function using protein-protein interaction data. J. Comp. Biol. 10(6):947-960**
- **Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T. (2001) Assessment of prediction accuracy of protein function from protein-protein interaction data, Yeast, 18(6):523-531**
- **Lanckriet, G. R. G., Deng, M., Cristianini,, N., Jordan, M. I. and Noble, W. S. (2004) Kernel-based data fusion and its application to protein function prediction in yeast. Proc. Pacific Symposium on Biocomputing 2004. pp.300-311.**
- **Letovsky, S. and Kasif, S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics. 19(Suppl.1):i197–i204**

Other References

- **Ruepp A., Zollner A., Maier D., Albermann K., Hani J., Mokrejs M., Tetko I., Guldener U., Mannhaupt G., Munsterkotter M., Mewes H.W. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 14:32(18):5539-45**
- **Samanta, M. P., Liang, S. (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl. Acad. Sci. U S A. 100(22):12579-83**
- **Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of interacting proteins in yeast. Nature Biotechnology 18(12):1257-1261**
- **Titz B., Schlesner M. and Uetz P. (2004) What do we learn from high-throughput protein interaction data? Expert Rev.Proteomics 1(1):111–121**
- **Vazquez, A., Flammi, A., Maritan, A. and Vespignani, A. (2003) Global protein function prediction from protein-protein interaction networks. Nature Biotechnology. 21(6):697-670**
- **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**