

Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions

Limsoon Wong

Joint work with Hon Nian Chua & Wing-Kin Sung



Lecture at Yang Ming National University, Taipei, June 2006

2

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
- ...

Lecture at Yang Ming National University, Taipei, June 2006

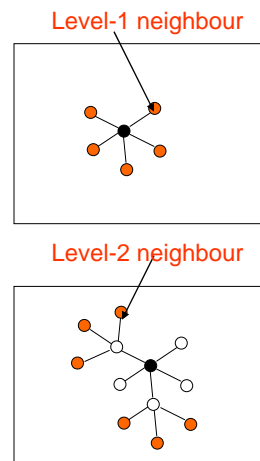
Copyright © 2006 by Limsoon Wong

Protein Interaction Based Approaches

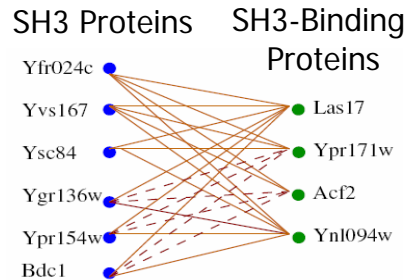
- **Neighbour counting** (Schwikowski et al, 2000)
 - Rank function based on freq in interaction partners
- **Chi-square** (Hishigaki et al, 2001)
 - Chi square statistics using expected freq of functions in interaction partners
- **Markov Random Fields** (Deng et al, 2003; Letovsky et al, 2003)
 - Belief 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



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

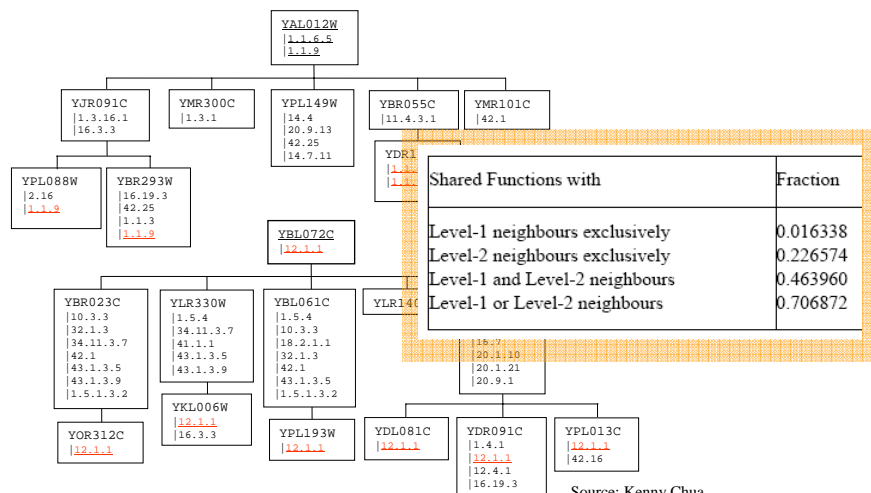


- **Protein interaction data from General Repository for Interaction Datasets (GRID)**
 - Data from published large-scale interaction datasets and curated interactions from literature
 - 13,830 unique and 21,839 total interactions
 - Includes most interactions from the Biomolecular Interaction Network (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



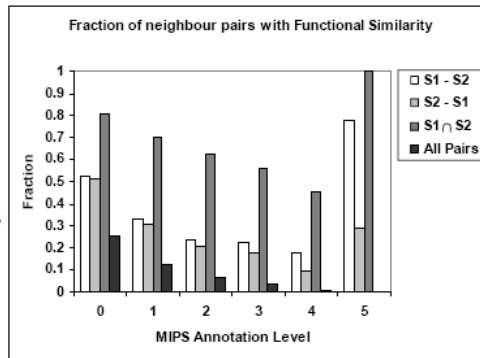
Over-Rep of Functions in Neighbours

- **Functional Similarity:**

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

- 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**

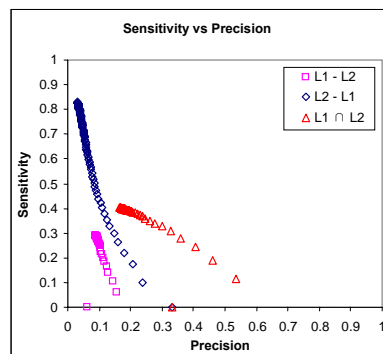


Prediction Power By Majority Voting

- **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**

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

- 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
- $X \Delta Y$ 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 ∪ 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:

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

- 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

Functional Similarity Estimate: Transitive FS Weighted Measure



- If protein u is similar to w , and w is similar to v , then proteins u and v may be similar also
- Transitive FS weighted measure

$$S_{TR}(u, v) = \max\left(S_R(u, v), \max_{w \in N_u} S_R(u, w)S_R(w, v)\right)$$

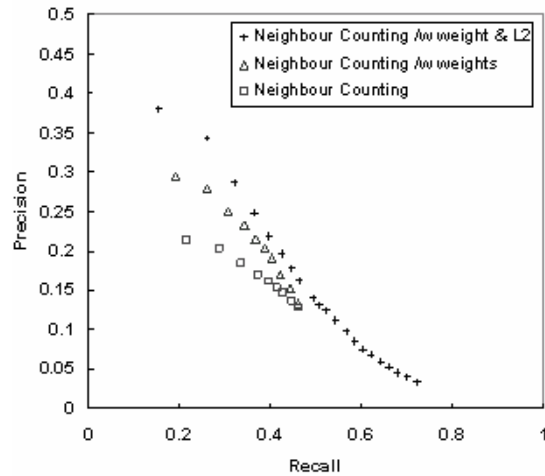
Integrating Transitivity



- Equiv measure shows improved correlation w/ functional similarity when transitivity is considered:

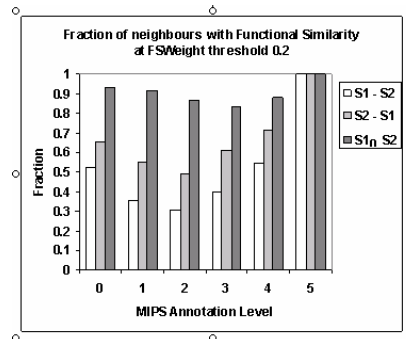
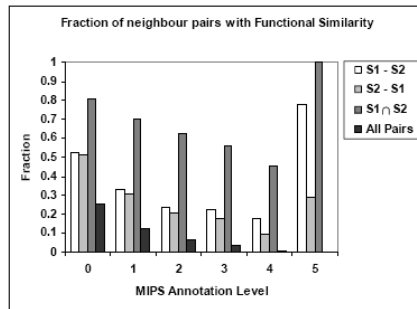
Neighbours	CD-Distance	FS-Weight	FS-Weight R	Transitive FS-Weight R
S_1	0.471810	0.498745	0.532596	0.532626
S_2	0.224705	0.298843	0.375317	0.381966
$S_1 \cup S_2$	0.224581	0.29629	0.363025	0.369378

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**

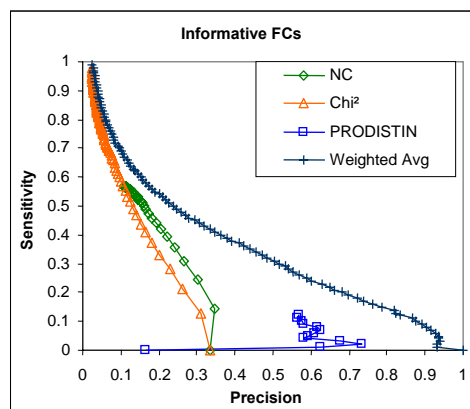
$$f_x(u) = \frac{1}{Z} \left[\lambda r_{int} \pi_x + \sum_{v \in N_u} \left(S_{TR}(u, v) \delta(v, x) + \sum_{w \in N_v} S_{TR}(u, w) \delta(w, x) \right) \right]$$

- 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

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

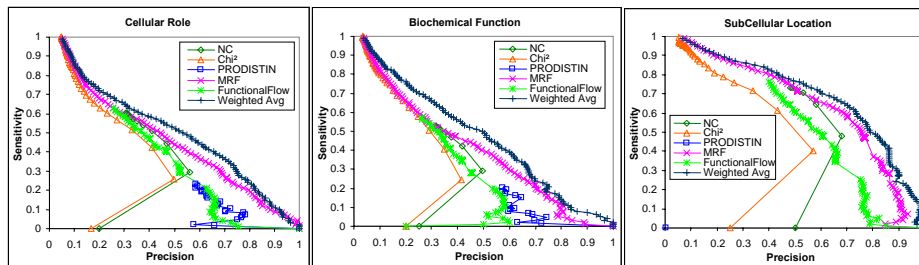
Performance of FS-Weighted Averaging

- **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

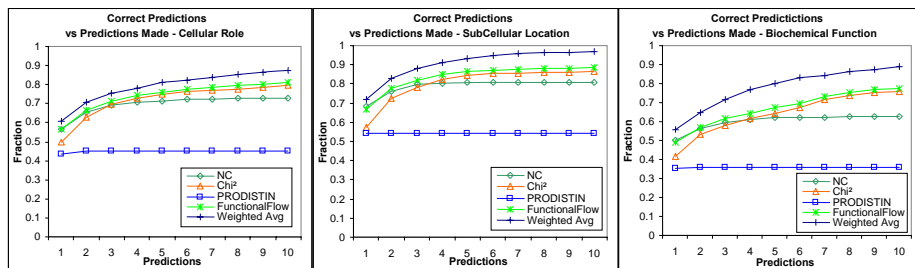


Lecture at Yang Ming National University, Taipei, June 2006

Copyright © 2006 by Limsoon Wong

Performance of FS-Weighted Averaging

- Correct Predictions made on at least 1 function vs Number of predictions made per protein

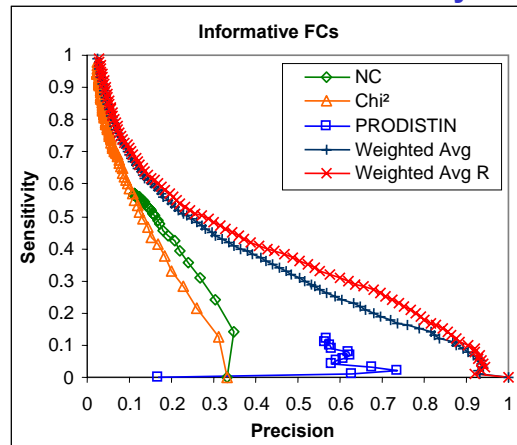


Lecture at Yang Ming National University, Taipei, June 2006

Copyright © 2006 by Limsoon Wong

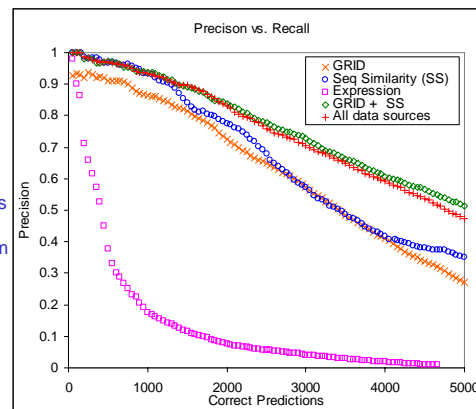
Performance of FS-Weighted Averaging

- Prediction performance further improves after incorporation of interaction reliability



Incorporating Other Info Sources

- **PPI Interaction Data**
 - General Rep of Interaction Data
 - 17815 Unique Pairs, 4914 Proteins
 - Reliability: 0.366 (Based on fraction with known functional similarity)
- **Sequence Similarity**
 - Smithwaterman betw seq of all proteins
 - For each seq, among all SW scores w/ all other seq, extract seq w/ SW score ≥ 3 standard deviations from mean
 - 32028 Unique Pairs, 6766 Proteins
 - Reliability: 0.659
- **Gene Expression**
 - Spellman w/ 77 timepoints
 - Extract all pairs w/ Pearson's > 0.7
 - 11586 Unique Pairs, 2082 Proteins
 - Reliability: 0.354



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

Acknowledgements

- Hon Nian Chua
- Wing Kin Sung

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

References

- Ruepp A., Zollner A., Maier D., Albermann K., Hani J., Moksrejs 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