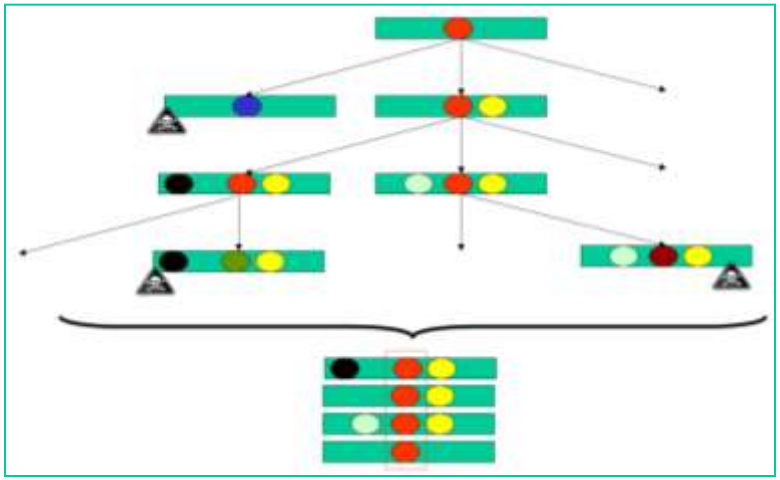
CS4220: Knowledge Discovery Methods for Bioinformatics

Unit 7: Protein Function Prediction

Limsoon Wong

NUS
National University
of Singapore

Protein function prediction w/o informative sequence homologs



- **Basic protein function prediction**
- **"Guilt by association" of other properties**
- **Protein function prediction from PPIs**
- **"Guilt by association" of multiple types of information**

Basic Protein Function Prediction

Limsoon Wong

NUS
National University
of Singapore

A protein is a ...

- **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 Sequence

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

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

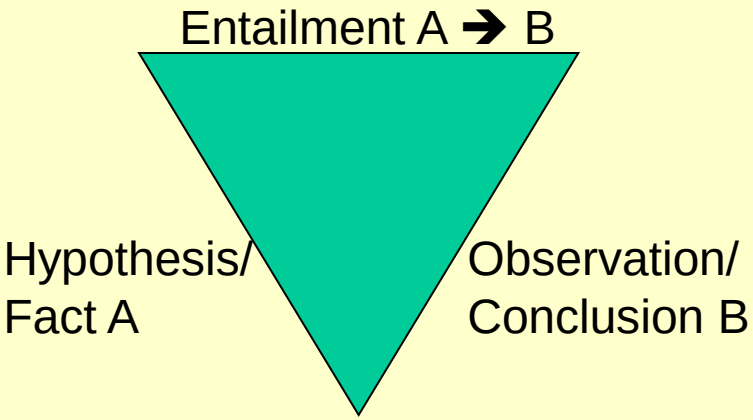
I hope you remember most of what I am going to tell you in the next ~10 slides. If not, dig out your old CS2220 lecture notes or slides!

Invariant and Abductive Reasoning

- **Function is determined by 3D struct of protein & environment protein is in**
- **Constraints imposed by 3D struct & environment give rise to "invariant" properties observed in proteins having the ancestor with that function**

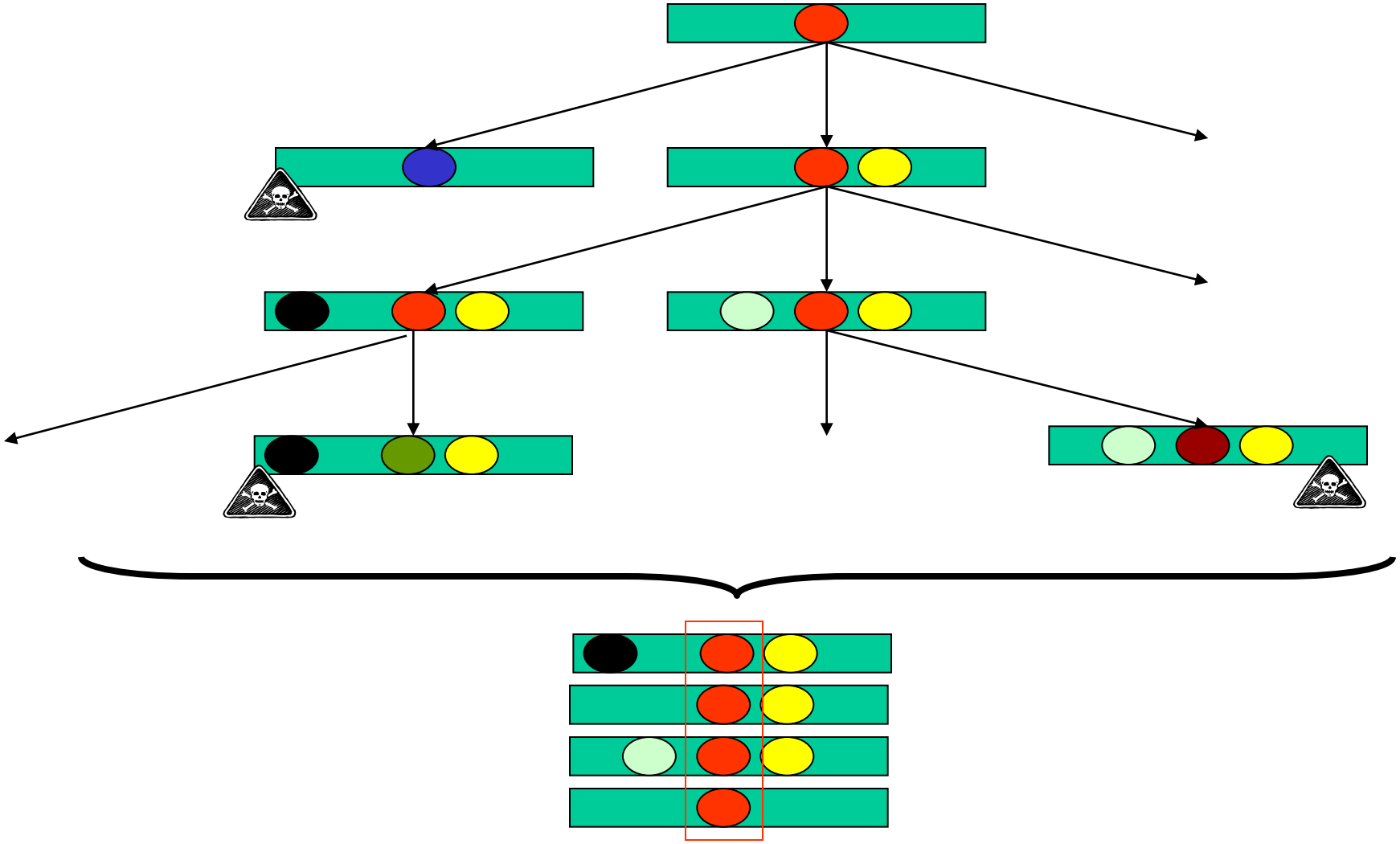
⇒ **Abductive reasoning**

– If those invariant properties are seen in a protein, then the protein is homolog of this protein



⇒ **"Guilt by association"**

In the course of evolution…



Guilt-by-Association

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

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

```
                          60        70        80        90       100
Amicyanin            MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTPTEAGTYDYHCTPHPFMRGKVVV
                                                   :.:  . ::. ::
Ascorbate Oxidase    ILQRGTPWADGTASISQCAINPGETFFYNPTVDNPGTFFYHGHLGMQRSAGLYG
                          70        80        90       100       110
```

No obvious match between Amicyanin and Ascorbate Oxidase

Discard this function as a candidate

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

```
>gi|13476732|ref|NP_108301.1|   unknown protein [Mesorhizobium loti]
 gi|14027493|dbj|BAB53762.1|    unknown protein [Mesorhizobium loti]
          Length = 105

 Score =  105 bits (262), Expect = 1e-22
 Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1   MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIKWVNKDVFAHT 60
           MK G L  ++        MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSVLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

good match between Amicyanin and unknown M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments

Guilt-by-Association: Caveats

- **Ensure that the effect of database size has been accounted for**
- **Ensure that the function of the homology is not derived via invalid "transitive assignment"**
- **Ensure that the target sequence has all the key features associated with the function, e.g., active site and/or domain**

Guilt by Association of Other Properties

Limsoon Wong

NUS
National University
of Singapore

What if there is no useful seq homolog?

- **Guilt by other types of association!**
  - Domain modeling (e.g., HMMPFAM)
  - Similarity of phylogenetic profiles
  - Similarity of dissimilarities (e.g., SVM-PAIRWISE)
  - 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

Forslund & Sonnhammer. "Predicting protein function from domain content". Bioinformatics, 24(15):1681-1687, 2008

Domain Modeling

- **Annotate known proteins in a database with their domains using, e.g., HMMPFAM**
- **Association-rule approach**
  - Do association rule mining to get high-confidence rules $D_1, \ldots, D_k \Rightarrow F$
  - Predict unknown protein to have function F if domains $D_1, \ldots, D_k$ are found in the protein
- **Probabilistic approach**
  - Prob of protein having D will have F, $P(F|D)$
  - Prob of protein having D will not have F, $P(\sim F|D)$
  - Odds ratio, $\alpha = P(F|D)/P(\sim F|D)$

$\Rightarrow P(F|D) = \alpha/(1 + \alpha)$

Similarity of Phylogenetic Profiles

- **Proteins carry out their function within the context of biological pathways**
- **Genes coding for proteins participating in the same pathway are present together in genomes where the pathway is functional**

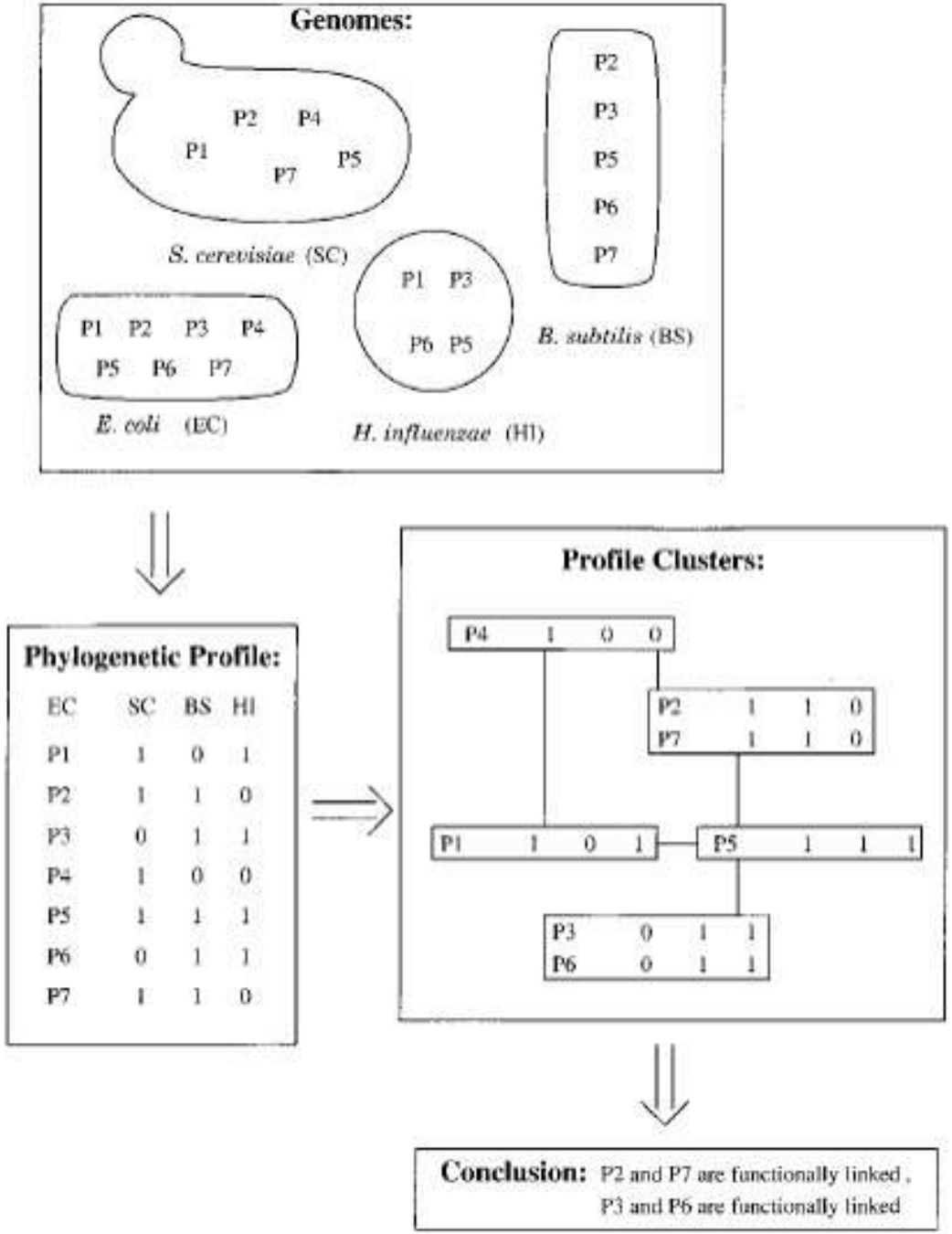
By abduction,

- **Genes (and hence proteins) with identical patterns of occurrence across phyla participate in the same pathway and function together**

⇒ **Phylogenetic profiling**

Pellegrini et al., *PNAS*, 96:4285--4288, 1999

Phylogenetic Profiling: How it Works



Pellegrini et al., *PNAS*, 96:4285--4288, 1999

Phylogenetic Profiles: Evidence

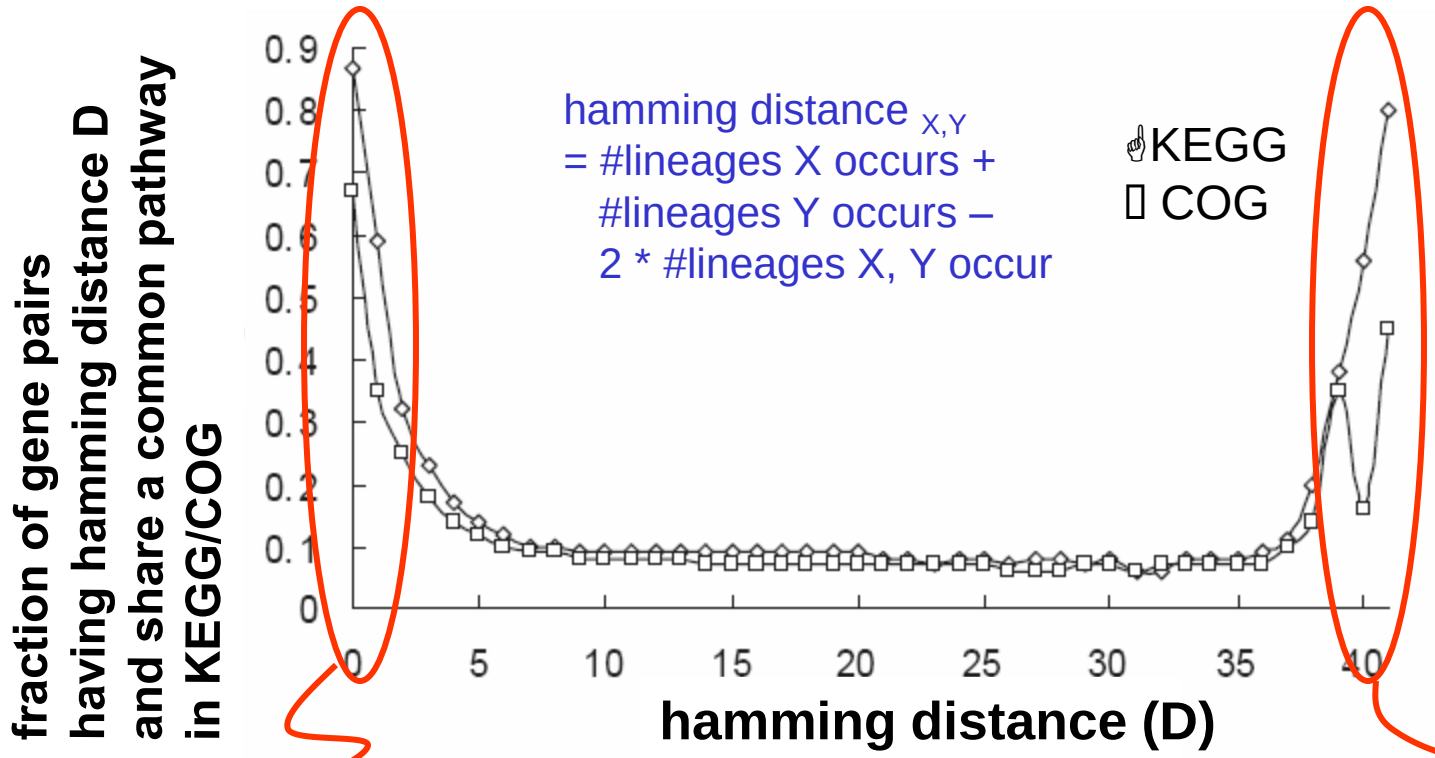
Keyword	No. of non-homologous proteins in group	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins*	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdoterin and Molybdenum, and molybdoterin	12	6	1
Hypothetical†	1,084	108,226	8,440

- **E. coli proteins grouped based on similar keywords in SWISS-PROT have similar phylogenetic profiles**

Pellegrini et al., *PNAS*, 96:4285--4288, 1999

Wu et al., *Bioinformatics*, 19:1524--1530, 2003

Phylogenetic Profiling: Evidence



- **Proteins having low hamming distance (thus highly similar phylogenetic profiles) tend to share common pathways**

Why do proteins having high hamming distance also have this behaviour?

Similarity of Dissimilarities

Differences of "unknown" to other fruits are same as "apple" to other fruits

"unknown" is an "apple"!

	Orange$_1$	Banana$_1$	...
Apple$_1$	Color = red vs orange Skin = smooth vs rough **Size = small vs small** **Shape = round vs round**	Color = red vs yellow **Skin = smooth vs smooth** **Size = small vs small** Shape = round vs oblong	...
Orange$_2$	**Color = orange vs orange** **Skin = rough vs rough** **Size = small vs small** **Shape = round vs round**	Color = orange vs yellow **Skin = rough vs smooth** **Size = small vs small** Shape = round vs oblong	...
Unknown$_1$	**Color = red vs orange** **Skin = smooth vs rough** Size = small vs small Shape = round vs round	**Color = red vs yellow** Skin = smooth vs smooth Size = small vs small **Shape = round vs oblong**	...
...	...	...	...

Li & Noble. *JCB*, 10(6):857-868, 2003

SVM-Pairwise Framework

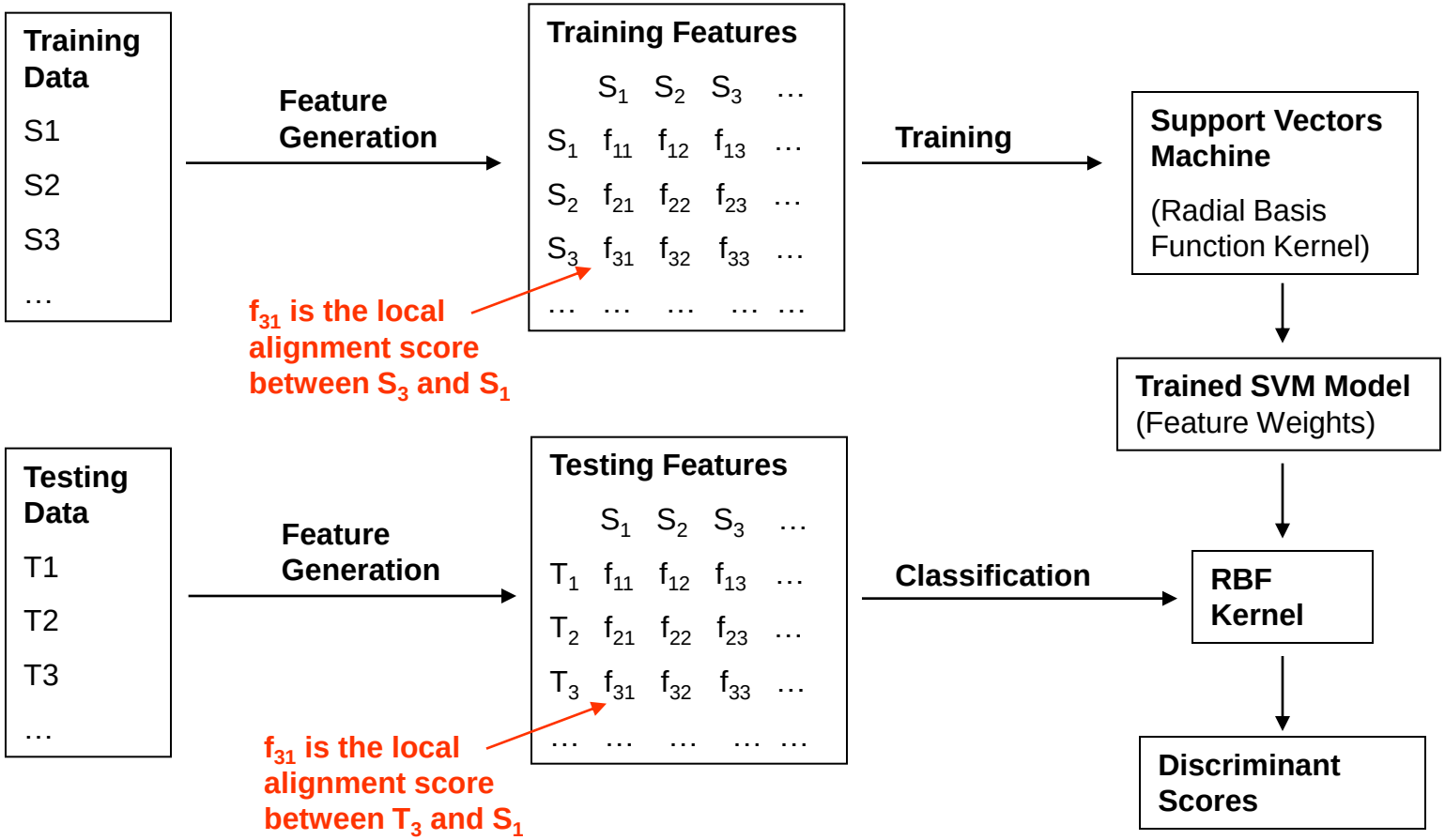
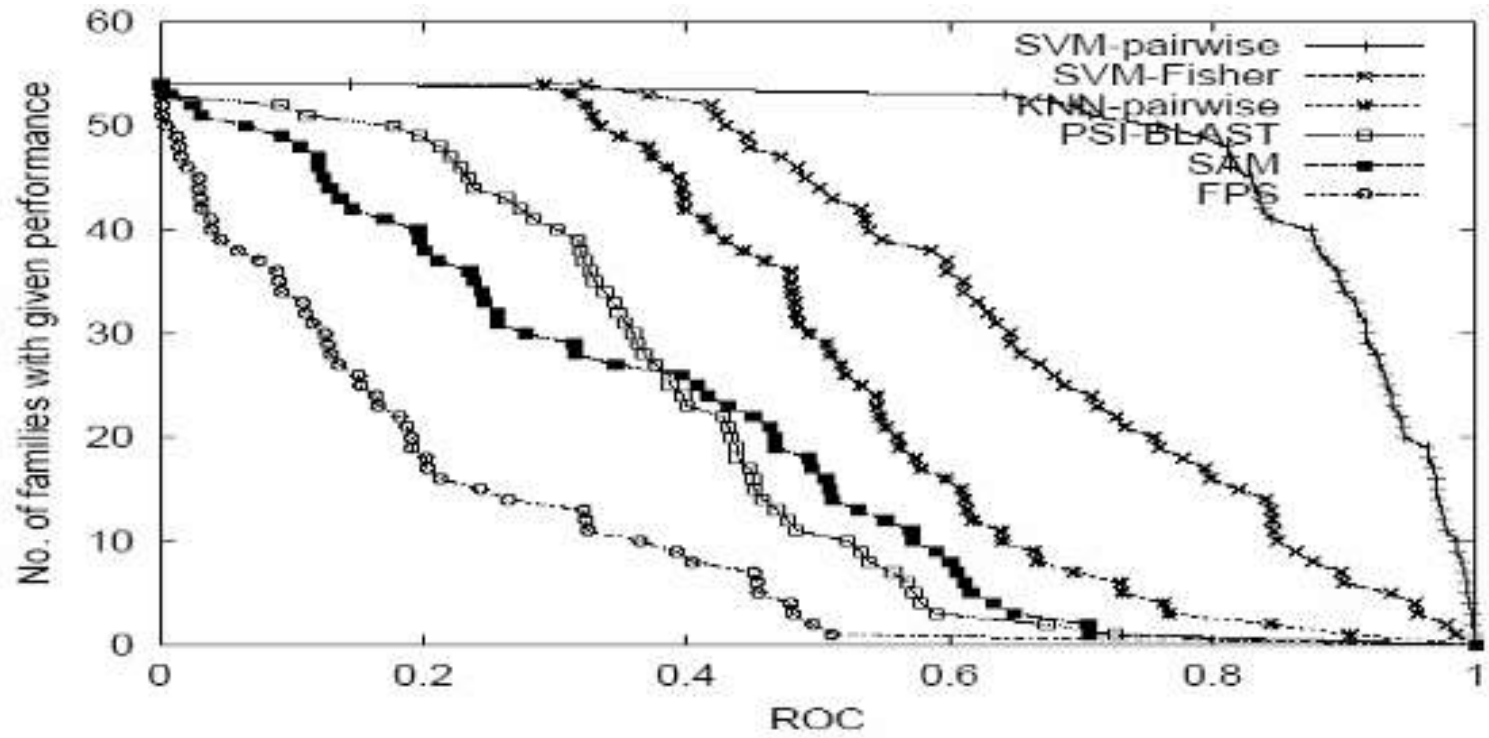


Image credit: Kenny Chua

Li & Noble. *JCB*, 10(6):857-868, 2003

Performance of SVM-Pairwise



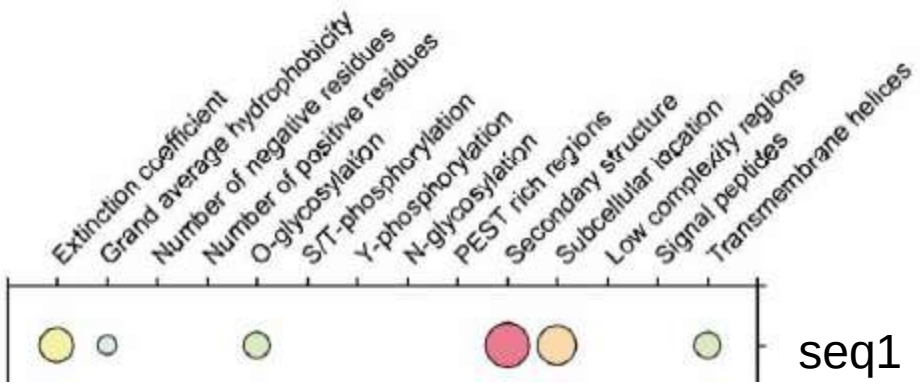
- **Receiver Operating Characteristic (ROC)**
  - The area under the curve derived from plotting true positives as a function of false positives for various thresholds.

Jensen, *JMB*, 319:1257--1265, 2002

The ProtFun Approach

Jensen, *JMB*, 319:1257--1265, 2002

- **A protein is not alone when performing its biological function**
- **It operates using the same cellular machinery for modification and sorting as all other proteins do, such as glycosylation, phospharylation, signal peptide cleavage, …**
- **These have associated consensus motifs, patterns, etc.**



- **Proteins performing similar functions should share some such "features"**

⇒ **Perhaps we can predict protein function by comparing its "feature" profile with other proteins?**

Jensen, *JMB*, 319:1257--1265, 2002

ProtFun: How it Works

Abbriviation	Encoding	Description
ec	single value	Extinction coefficient predicted by ExPASy ProtParam
gravy	single value	Hydrophobicity predicted by ExPASy ProtParam
nneg	single value	Number of negatively charged residues counted by ExPASy ProtParam
npos	single value	Number of positively charged residues counted by ExPASy ProtParam
nglyc	potential in 5 bins	N-glycosylation sites predicted by NetNGlyc
oglyc	potential-threshold in 10 bins	GalNAc O-glycosylations predicted by NetOGlyc
pest	fraction in 10 bins	PEST rich regions identified by PESTfind
phosST	potential in 10 bins	Serine and threonine phosporylations predicted by NetPhos
phosY	potential in 10 bins	Tyrosine phosporylations predicted by NetPhos
psipred	helix, sheet, coil in 5 bins	Predicted secondary structure from PSI-Pred
psort	20 probabilities	Subcellular location predtions by PSORT
seg	fraction in 10 bins	Low-complexity regions identified by SEG
signalp	meanS, maxY, log(cleavage pos)	Signal peptide predictions made by SignalP
tmhmm	inside, outside, membrane in 5 bins	Transmembrane helix predictions made by TMHMM

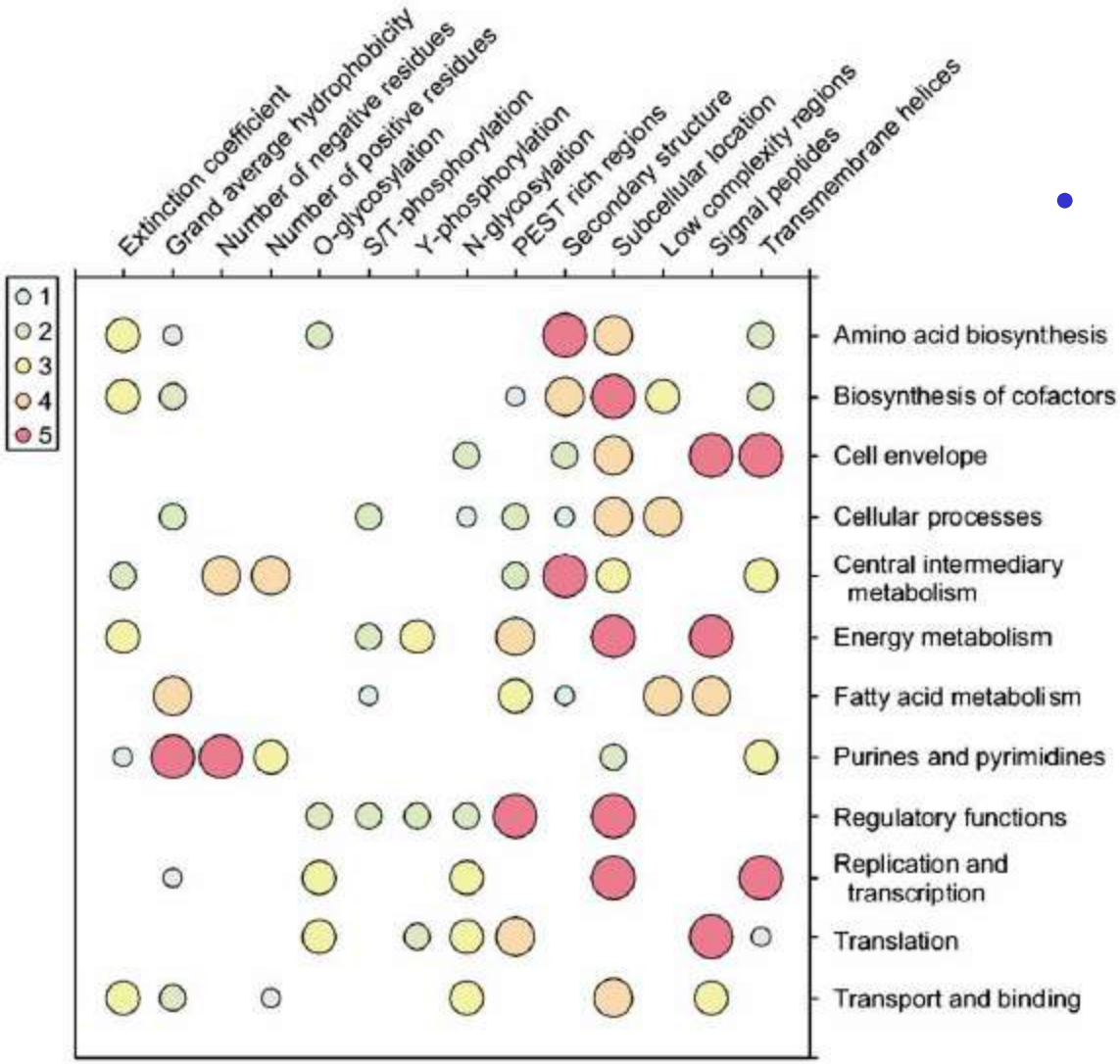
Extract feature profile of protein using various prediction methods

Category	Hidden units	Input features
Amino acid biosynthesis	30	ec psipred psort tmhmm
	30	ec psipred tmhmm
	30	ec netoglyc psipred psort
	30	gravy psipred psort
	30	oglyc psipred psort

Average the output of the 5 component ANNs

Jensen, *JMB*, 319:1257--1265, 2002

ProtFun: Evidence



- **Combinations of "features" seem to characterize some functional categories**

Jensen, *JMB*, 319:1257--1265, 2002

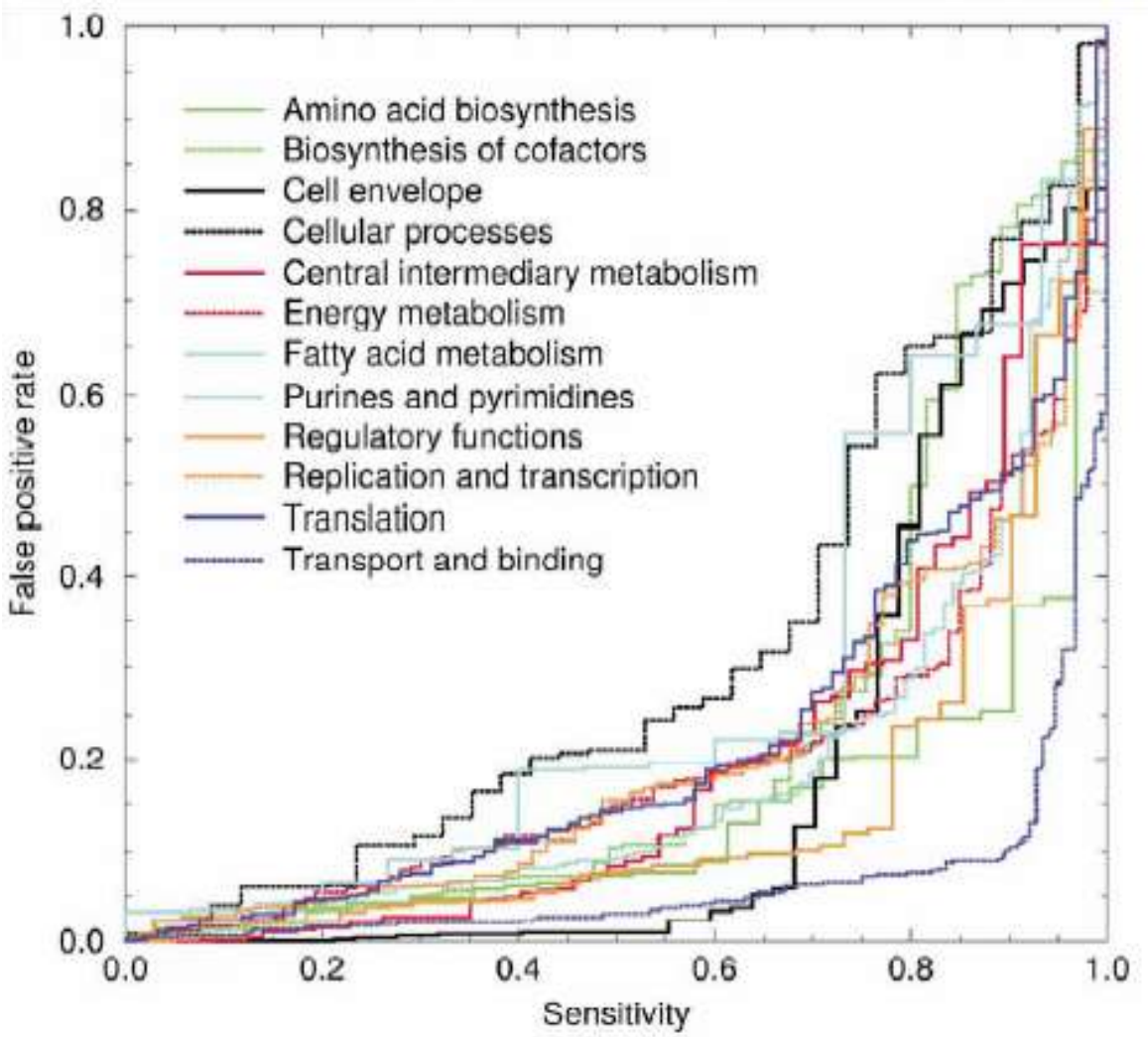
ProtFun: Example Output

	Prion	A4	TTHY
Amino acid biosynthesis	0.011	0.011	0.011
Biosynthesis of cofactors	0.041	0.161	0.034
Cell envelope	0.146	0.804	0.698
Cellular processes	0.027	0.027	0.051
Central intermediary metabolism	0.047	0.139	0.059
Energy metabolism	0.029	0.023	0.046
Fatty acid metabolism	0.017	0.017	0.023
Purines and pyrimidines	0.528	0.417	0.153
Regulatory functions	0.013	0.014	0.014
Replication and transcription	0.020	0.029	0.040
Translation	0.035	0.027	0.032
Transport and binding	0.831	0.827	0.812
Enzyme	0.233	0.367	0.227
Non-enzyme	0.767	0.633	0.773
Oxidoreductase (EC 1.–.–.–)	0.070	0.024	0.055
Transferase (EC 2.–.–.–)	0.031	0.208	0.037
Hydrolase (EC 3.–.–.–)	0.101	0.090	0.208
Isomerase (EC 4.–.–.–)	0.020	0.020	0.020
Ligase (EC 5.–.–.–)	0.010	0.010	0.010
Lyase (EC 6.–.–.–)	0.017	0.078	0.017

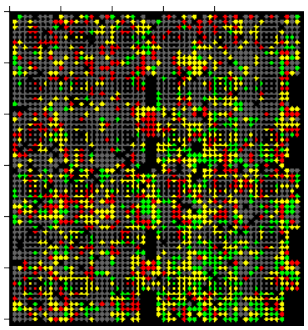
- **At the seq level, Prion, A4, & TTHY are dissimilar**
- **ProtFun predicts them to be cell envelope-related, tranport & binding**
- **This is in agreement w/ known functionality of these proteins**

Jensen, *JMB*, 319:1257--1265, 2002

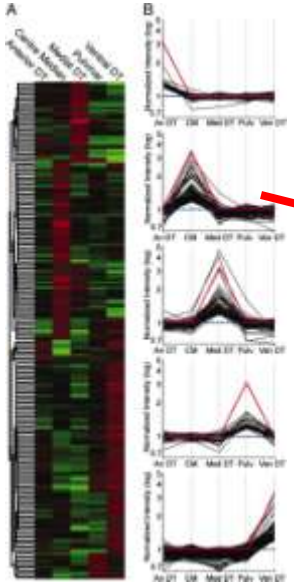
ProtFun: Performance



Similarity of Gene Expression Profiles



Cluster the genes by expression profiles



Prob of ≥k genes with function F within a cluster C by random chance

$$P(C,F) = 1 - \sum_{i=0}^{k-1} \frac{\binom{n_F}{i}\binom{N-n_F}{n_C-i}}{\binom{N}{n_C}}$$

N = # of genes in genome,
nF = # of genes having F,
nC = # of genes in C

- **P-value of gene G having function F is thus**

$$P(G,F) = \min_{C:G\in C} P(C,F).$$

$\Rightarrow$ **Predict G has function F when P(G, F) is small**

Xiao & Pan. *JBCB*, 3(6):1371-89, 2005

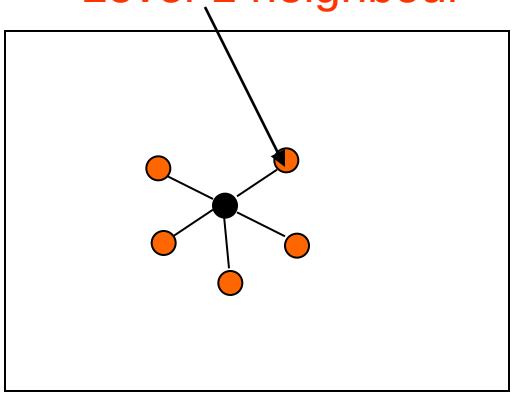
Direction Functional Association in PPIN

- **Prob of k genes with function F interacting with unknown gene G by random chance**

$$P_I(G,F) = 1 - \sum_{i=0}^{k} \frac{\binom{n_F}{i}\binom{N-1-n_F}{I_G-i}}{\binom{N-1}{I_G}}$$

N = # of genes in genome,
nF = # of genes having F,
IG = # of genes interacting with G

Level-1 neighbour

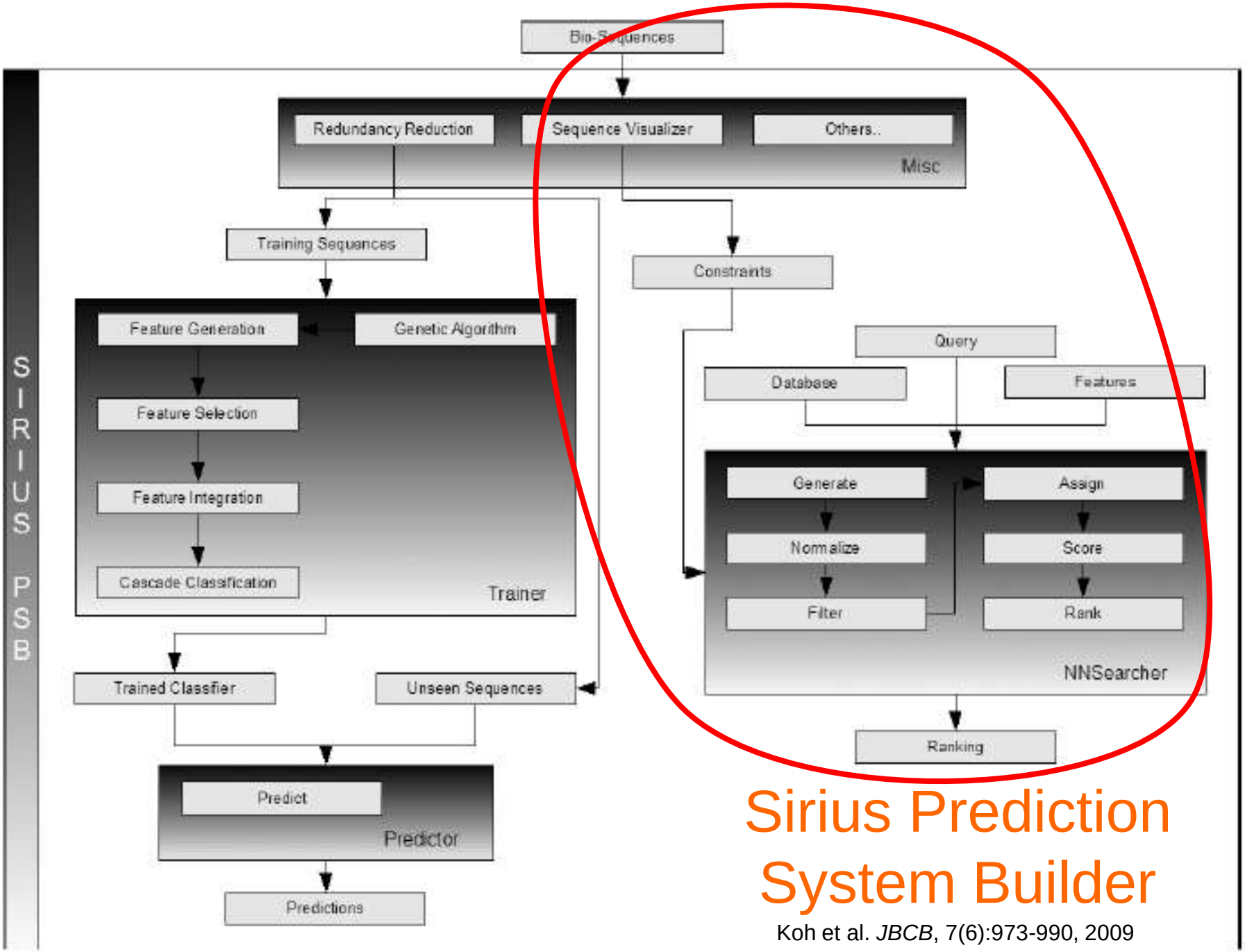


$\Rightarrow$ **Predict G has function F when $P_i(G,F)$ is small**

Xiao & Pan. *JBCB*, 3(6):1371-89, 2005

The approaches described earlier assume you have lots of training data.

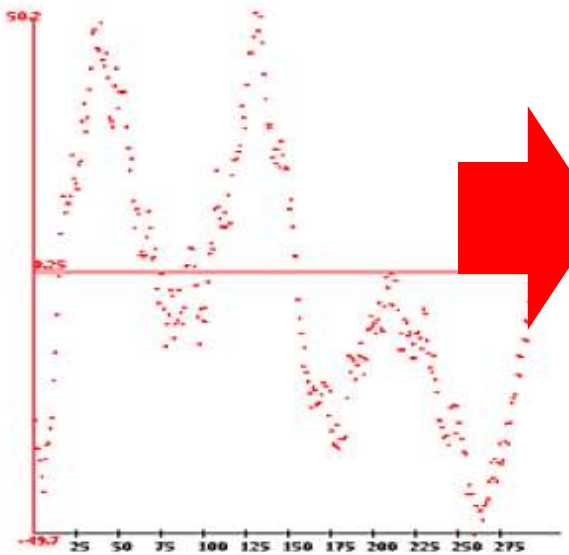
What if you have only a few training samples?

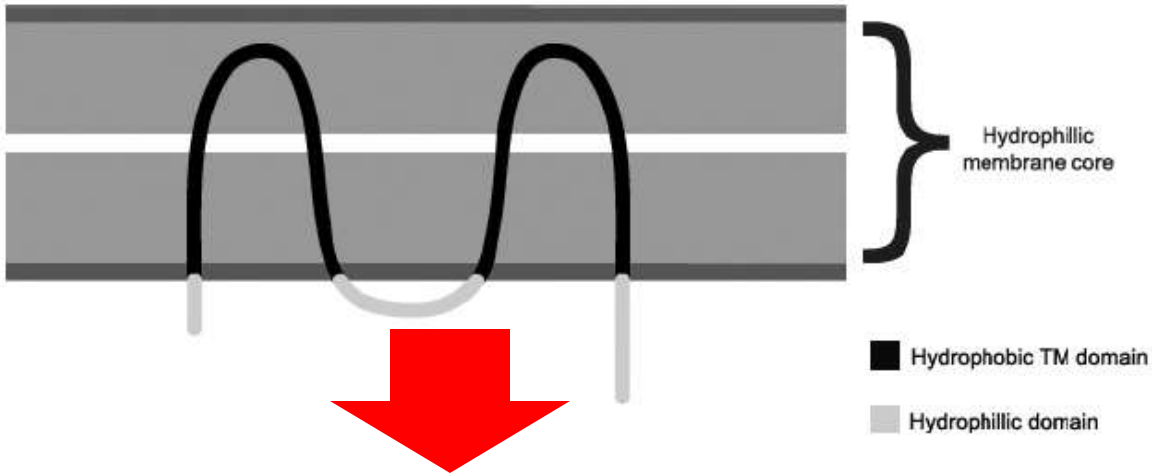


Sirius Prediction System Builder

Koh et al. *JBCB*, 7(6):973-990, 2009

Koh et al. *JBCB*, 7(6):973-990, 2009





Sirius PSB

- **Visualize & specify seq features to search for related proteins w/ low seq similarity**

Table 4 Top 10 hits of NNSearch with RTN1 & RTN2 as query and

No.	Sequence header
1	sp|Q04947|RTN1_YEAST Reticulon-like protein 1 OS=Sacch GN=RTN1
2	sp|Q12440|RTN2_YEAST Reticulon-like protein 2 OS=Sacch GN=RTN2
3	sp|P20641|MPCP_YEAST Mitochondrial phosphate carrier p cerevisiae GN=MIR1
4	sp|Q12402|YOP1_YEAST Protein YOP1 OS=Saccharomyces
5	sp|P50600|PRA1_YEAST Prenylated Rab acceptor 1 OS=Sa GN=YIP3
6	sp|P00410|COX2_YEAST Cytochrome c oxidase subunit 2 O GN=COX2
7	sp|P09692|MET10_YEAST Sulfite reductase [NADPH] flavop OS=Saccharomyces cerevisiae GN = MET10
8	sp|Q05142|IMB1_YEAST Importin subunit beta-1 OS=Sacch GN=KAP95
9	sp|P40069|IMB4_YEAST Importin subunit beta-4 OS=Sacch GN=KAP120
10	sp|P38029|YB85_YEAST Uncharacterized membrane protein OS=Saccharomyces cerevisiae GN=YBR205W

References

- **Must Read**
  - Friedberg. "Automated protein function prediction---the genomic challenge". *Briefings in Bioinformatics*, 7(3):225-242, 2006
- **Good to Read**
  - [Phylogenetic Profile] Pellegrini et al. "Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles", *PNAS*, 96:4285-4288, 1999
  - [Domain Content] Forslund & Sonnhammer. "Predicting protein function from domain content". *Bioinformatics*, 24(15):1681-1687, 2008
  - [ProtFun] Jensen et al. "Prediction of human protein function from post-translational modifications and localization features", *JMB*, 319:1257--1265, 2002
  - [SVM-Pairwise] Li & Noble. "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships". *JCB*, 10(6):857-868, 2003
  - [Sirius PSB] Koh et al. "Sirius PSB: A Generic System for Analysis of Biological Sequences". *JBCB*, 7(6):973-990, 2009

Protein Function Prediction from PPIN

Limsoon Wong

NUS
National University
of Singapore

Main Hypotheses of PPIN-Based Function Prediction

- **Proteins with similar function are topologically close in PPIN**
  - Direct functional association
  - Indirect functional association

Abduction!

A pair of proteins that participate in the same cellular processes or localize to the same cellular compartment are many times more likely to interact than a random pair of proteins

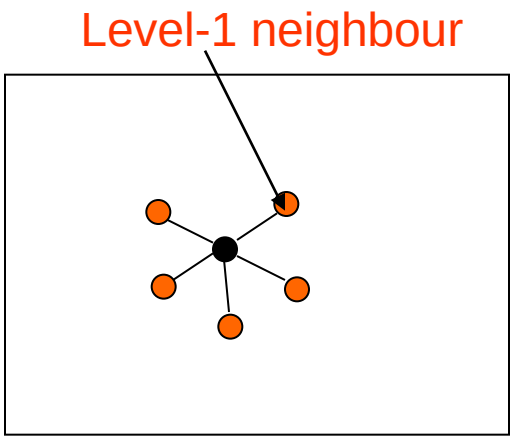
- **Proteins with similar function have interaction neighborhoods that are similar**

What do you get if you apply abduction here?

When proteins in the neighborhood of a protein X have similar functions to proteins in the neighborhood of a protein Y, then proteins X & Y likely operate in similar environment

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

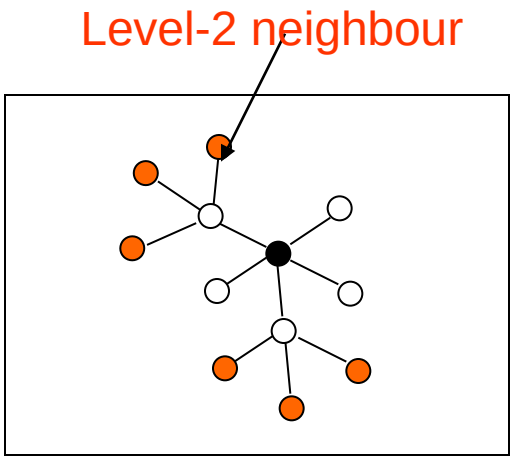
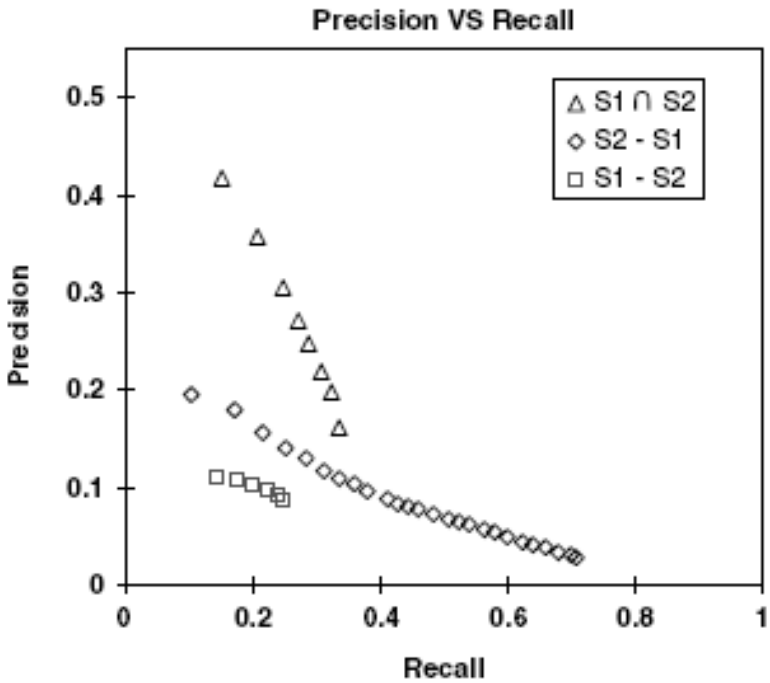


Image credit: Kenny Chua

Majority Voting

- **Proteins with similar function are topologically close in PPIN**

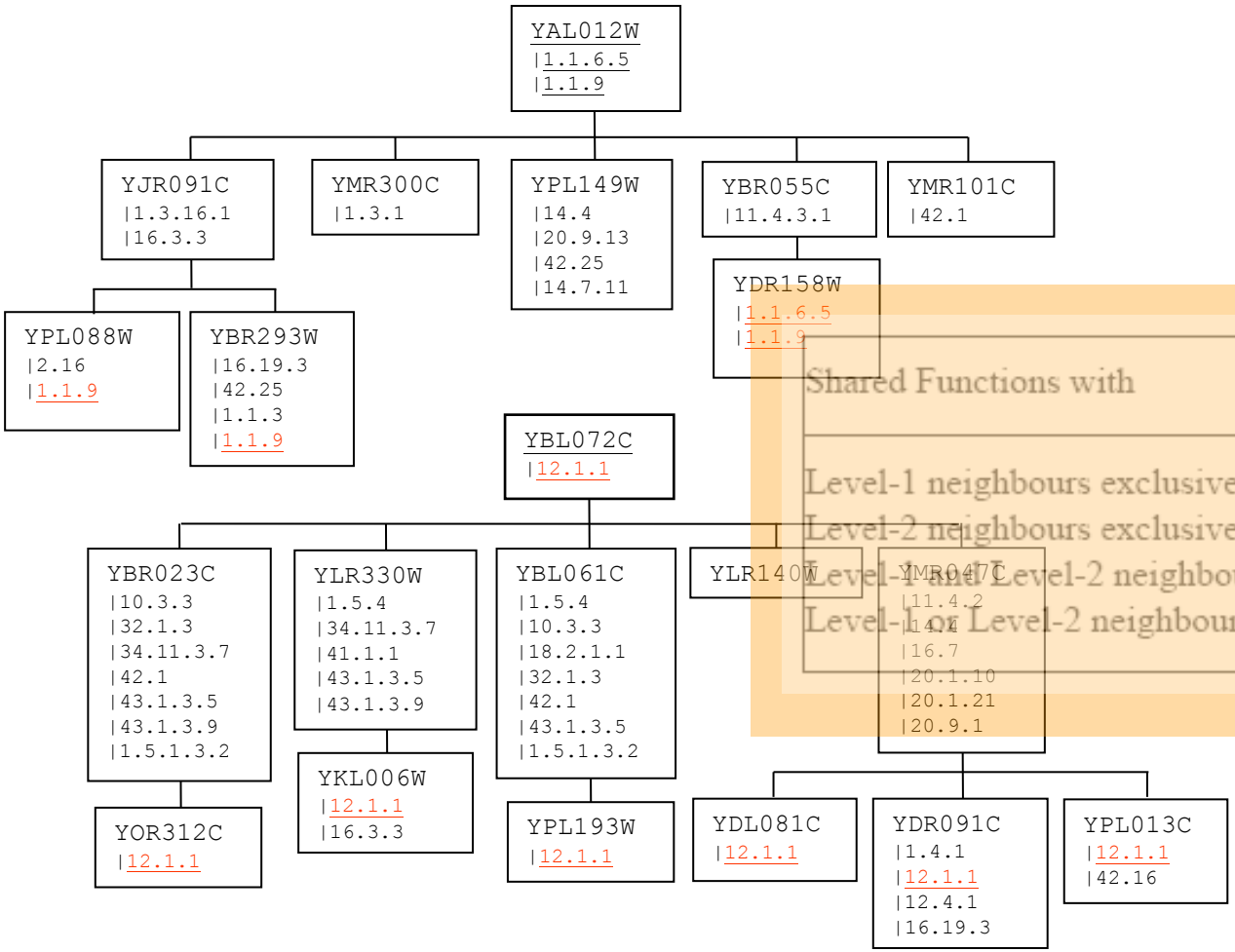
- **Assign a protein a function that is over represented among its interaction partners**



- **Shortcomings**
  - L1 is not sensitive
  - L2 is noisy

Hishigaki et al. *Yeast*, 18:523-531, 2001

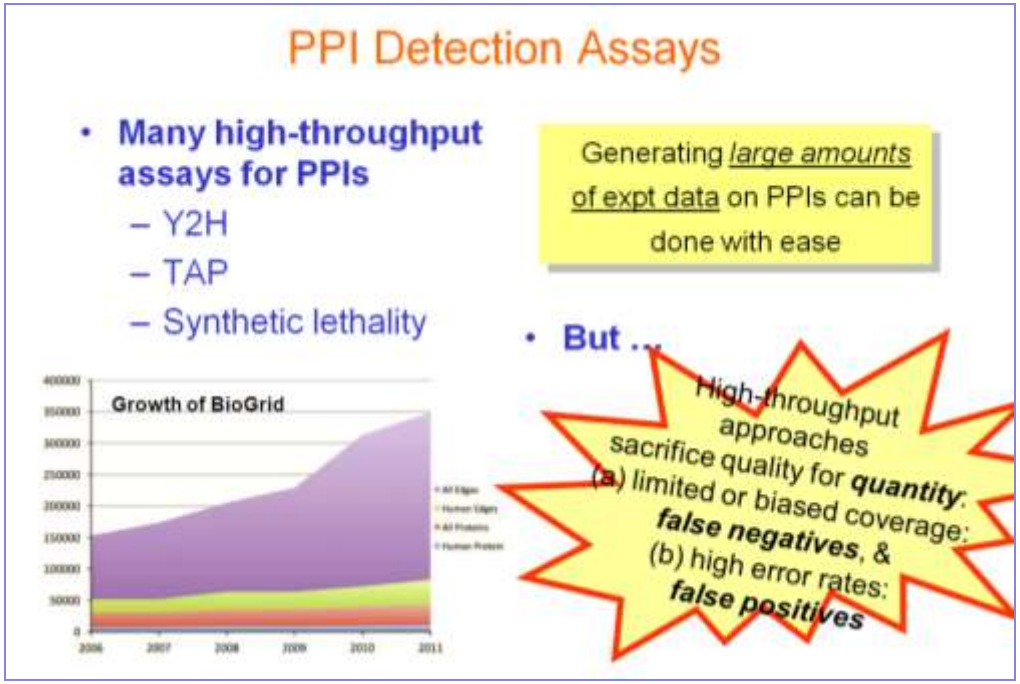
Why is L1 not sensitive?



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

Chua et al. *Bioinformatics*, 22:1623-1630, 2006.

Why is L2 noisy?



Sprinzak et al., *JMB*, 327:919-923, 2003

Experimental method category[a]	Number of interacting pairs	Co-localization[b] (%)	Co-cellular-role[b] (%)
All: All methods	9347	64	49
A: Small scale Y2H	1861	73	62
A0: GY2H Uetz *et al.* (published results)	956	66	45
A1: GY2H Uetz *et al.* (unpublished results	516	53	33
A2: GY2H Ito *et al.* (core)	798	64	40
A3: GY2H Ito *et al.* (all)	3655	41	15
B: Physical methods	71	98	95
C: Genetic methods	1052	77	75
D1: Biochemical, *in vitro*	614	87	79
D2: Biochemical, chromatography	648	93	88
E1: Immunological, direct	1025	90	90
E2: Immunological, indirect	34	100	93
2M: Two different methods	2360	87	85
3M: Three different methods	1212	92	94
4M: Four different methods	570	95	93

2360
1212
570

Large disagreement between experiments!

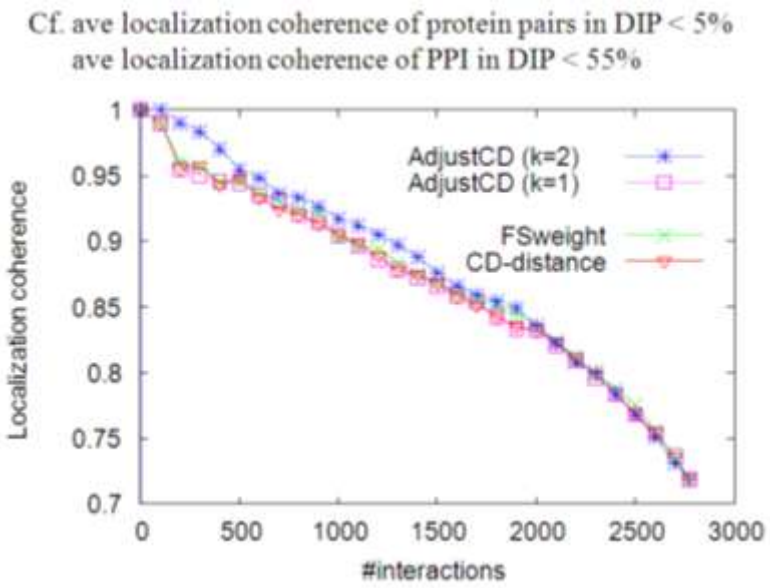
Chua & Wong. Increasing the Reliability of Protein Interactomes.
Drug Discovery Today, 13(15/16):652--658, 2008

Dealing with noise in PPIN

- **Two proteins participating in same biological process are more likely to interact**
- **Two proteins in the same cellular compartments are more likely to interact**

- **CD-distance**
- **FS-Weight**

CD-distance & FS-Weight: Based on concept that two proteins with many interaction partners in common are likely to be in same biological process & localize to the same compartment



Czekanowski-Dice Distance

- **Functional distance between two proteins**

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

- **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?

$\Rightarrow$ **Similarity can be defined as**

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

Brun, et al. *Genome Biology,* 5(1):R6, 2003

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**

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

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

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

- **FS-Weight is slightly better in correlation w/ similarity for L1 & L2 neighbours**

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

Reliability of Expt Sources

- **Diff expt sources have diff reliabilities**
  - Assign reliability to an interaction based on its expt sources
- **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**

FS-Weighted Measure with Reliability

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

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

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

$\Rightarrow$ **Rewriting**

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

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

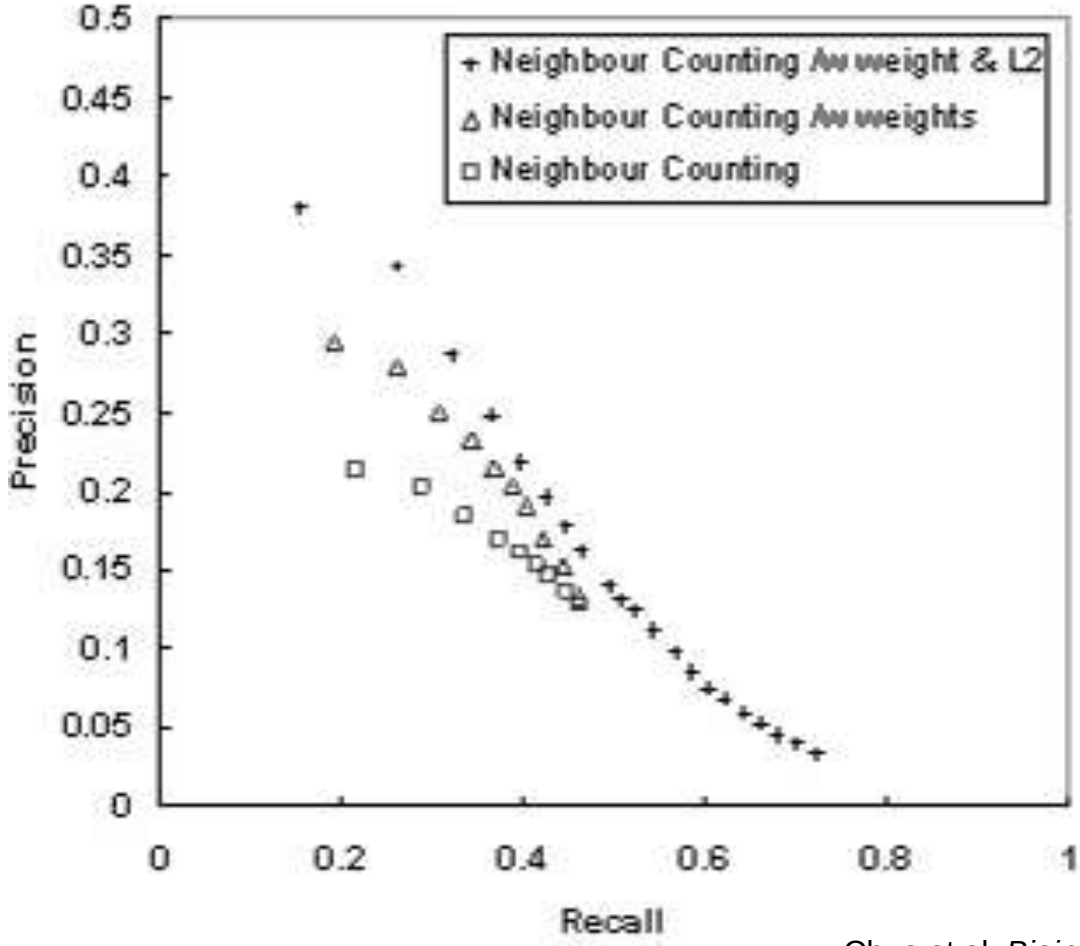
Integrating Reliability

- **FS-Weight 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

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

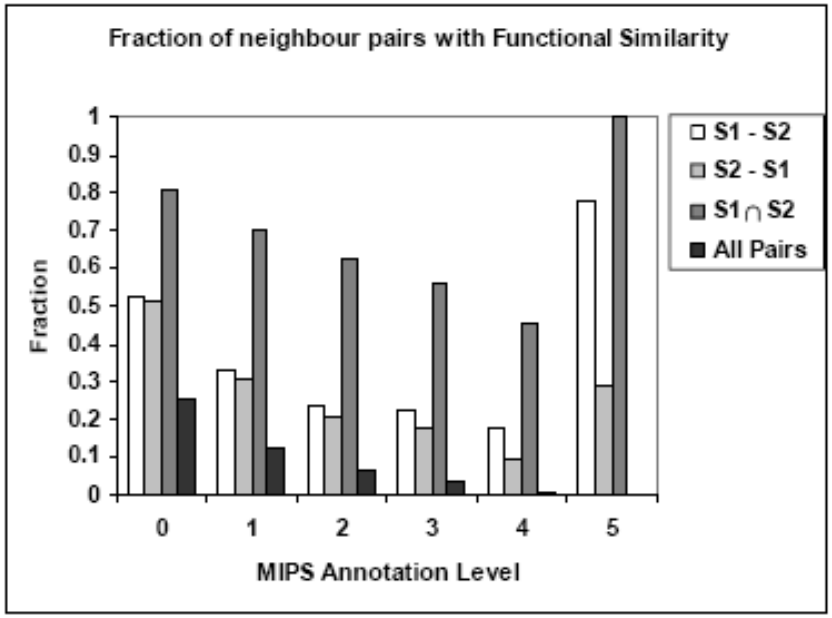
Improvement to Prediction Power by Majority Voting

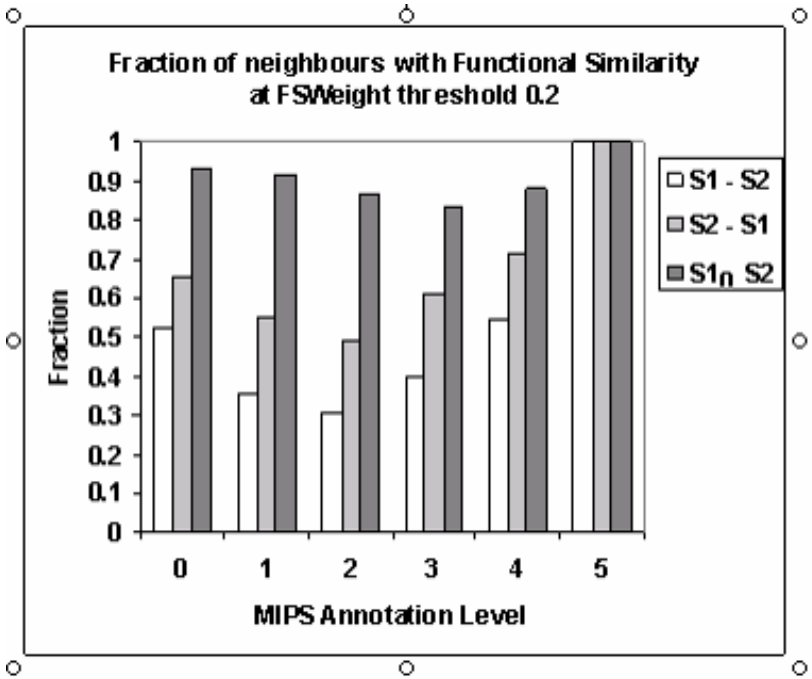


Considering only neighbours w/ FS weight > 0.2

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

Improvement to Over-Rep of Functions in Neighbours





Chua et al. *Bioinformatics*, 22:1623-1630, 2006

Use L1 & L2 Neighbours for Prediction

- **FS-weighted Averaging (FWA)**

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

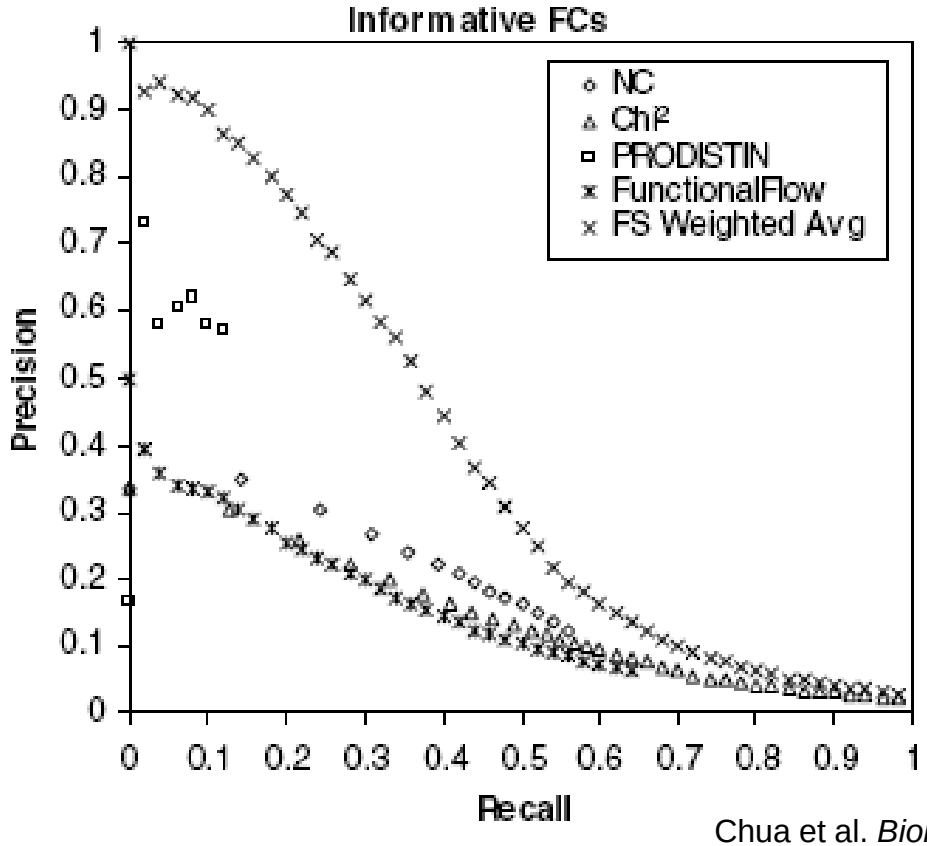
  - **r_{int} is fraction of all interaction pairs sharing function**
  - **λ is weight of contribution of background freq**
  - **δ(k, x) = 1 if k has function x, 0 otherwise**
  - **N_k is the set of interacting partners of k**
  - **π_x is freq of function x in the dataset**
  - **Z is sum of all weights**

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

Chua et al. *Bioinformatics*, 22:1623-1630, 2006

Performance of FS-Weighted Averaging

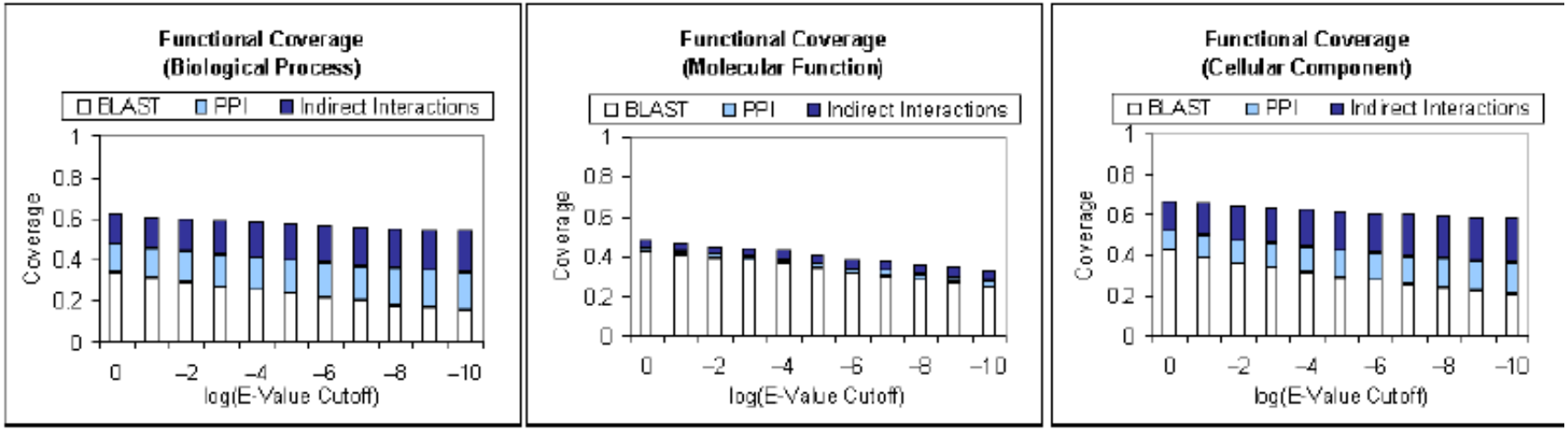
- **LOOCV comparison with Neighbour Counting, Chi-Square, PRODISTIN**



Chua et al. *Bioinformatics*, 22:1623-1630, 2006

Freq of indirect functional association in other genomes

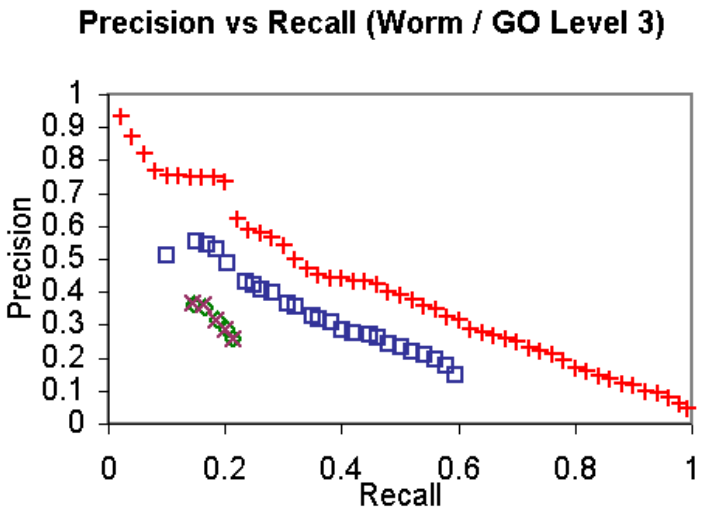
D. melanogaster

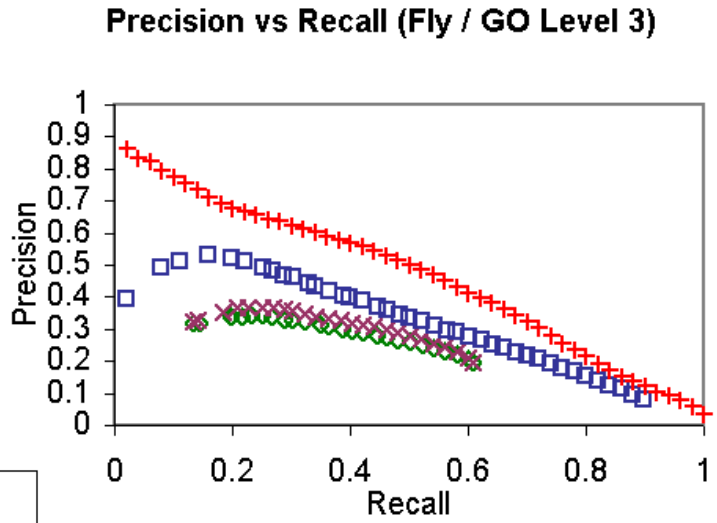


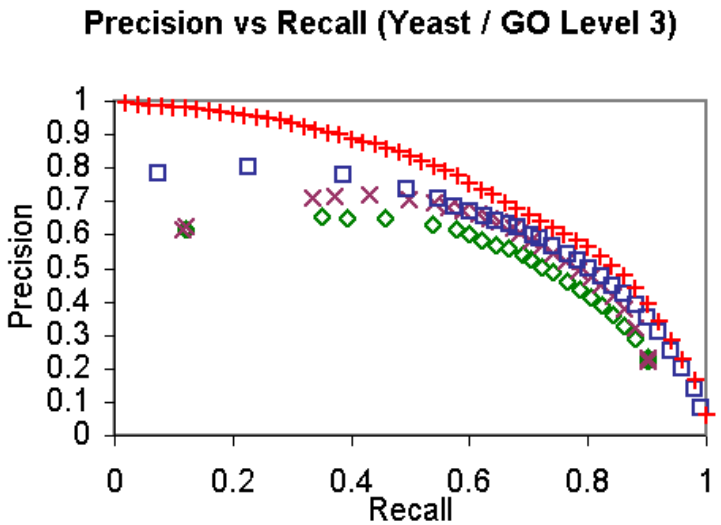
Genome	Annotation	S_1-S_2	S_2-S_1	$S_1 \cap S_2$	$S_1 \cup S_2$
S. cerevisiae	MIPS	0.007193	0.226574	0.463960	0.706872
D. melanogaster	GO	0.008801	0.168622	0.138138	0.315561
C. elegans	GO	0.007193	0.051237	0.061080	0.119510

Chua et al. Using Indirect Protein Interactions for the Prediction of Gene Ontology Functions. *BMC Bioinformatics*, 8(Suppl 4):S8, 2007

Effectiveness of FSWeighted Averaging in other genomes







- ◇ Neighbour Counting
- × NC (Weighted)
- □ NC (Weighted + L2)
- + Weighted Avg

Chua et al. Using Indirect Protein Interactions for the Prediction of Gene Ontology Functions. *BMC Bioinformatics*, 8(Suppl 4):S8, 2007

What have we learned?

- **Proteins with similar function are topologically close in PPIN**

⇒ **Assign protein to a function that is over represented in its neighborhood**

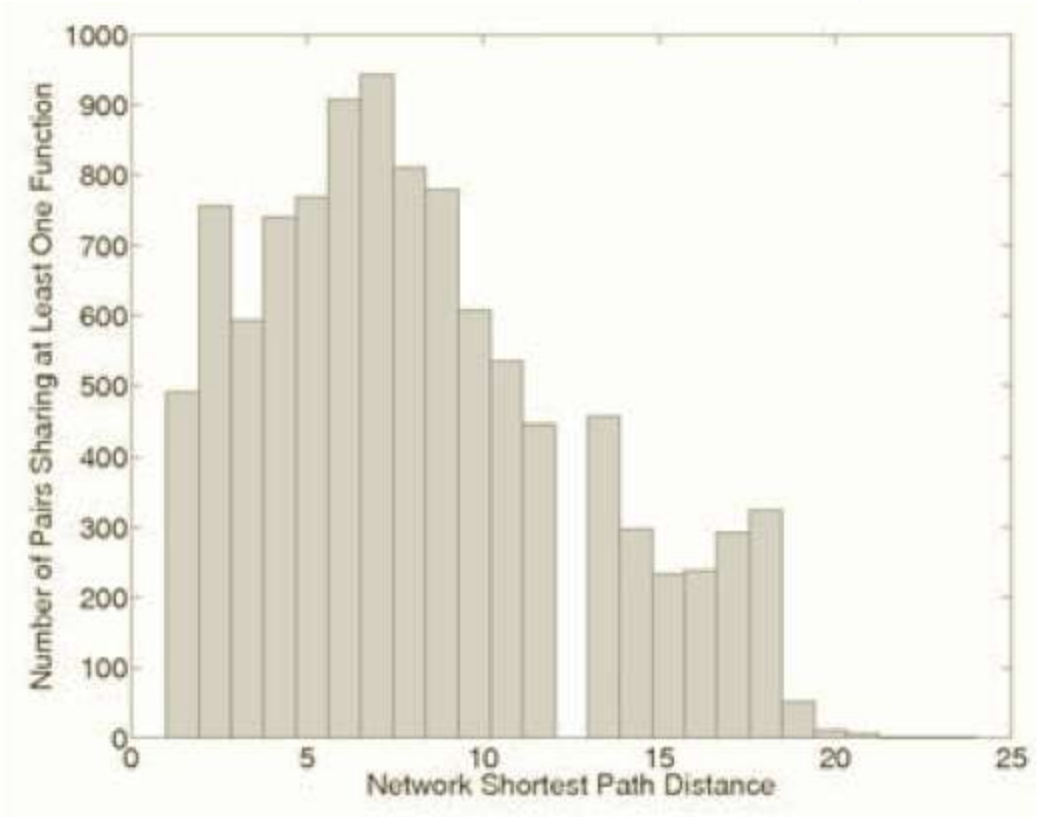
  – Indirect neighbors are useful

- **PPIN is noisy**

  – Not are neighbors are "real"

⇒ **Need to clean up the PPIN before "voting"**

But genes sharing annotations do not always interact…

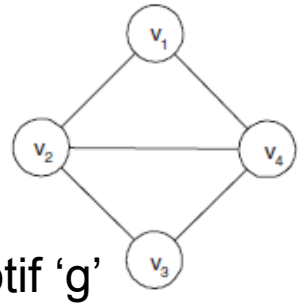


- **Similar functions are sometimes at large network distances**

Source: Bogdanov & Singh. *TCBB,* 7:208–217, 2010

Labeled Motifs

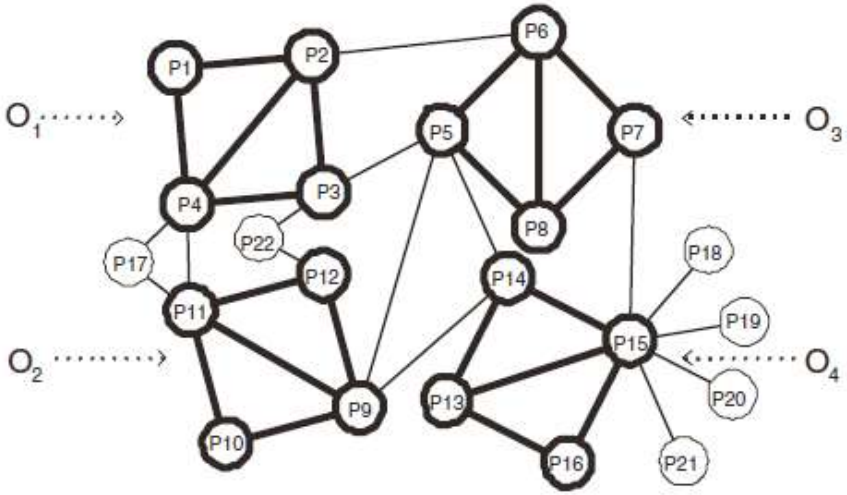
- **Proteins with similar function have interaction neighborhoods that are similar**



Network motif 'g'

g

- **Assign a protein a function based on "network motif" that its neighborhood matches**

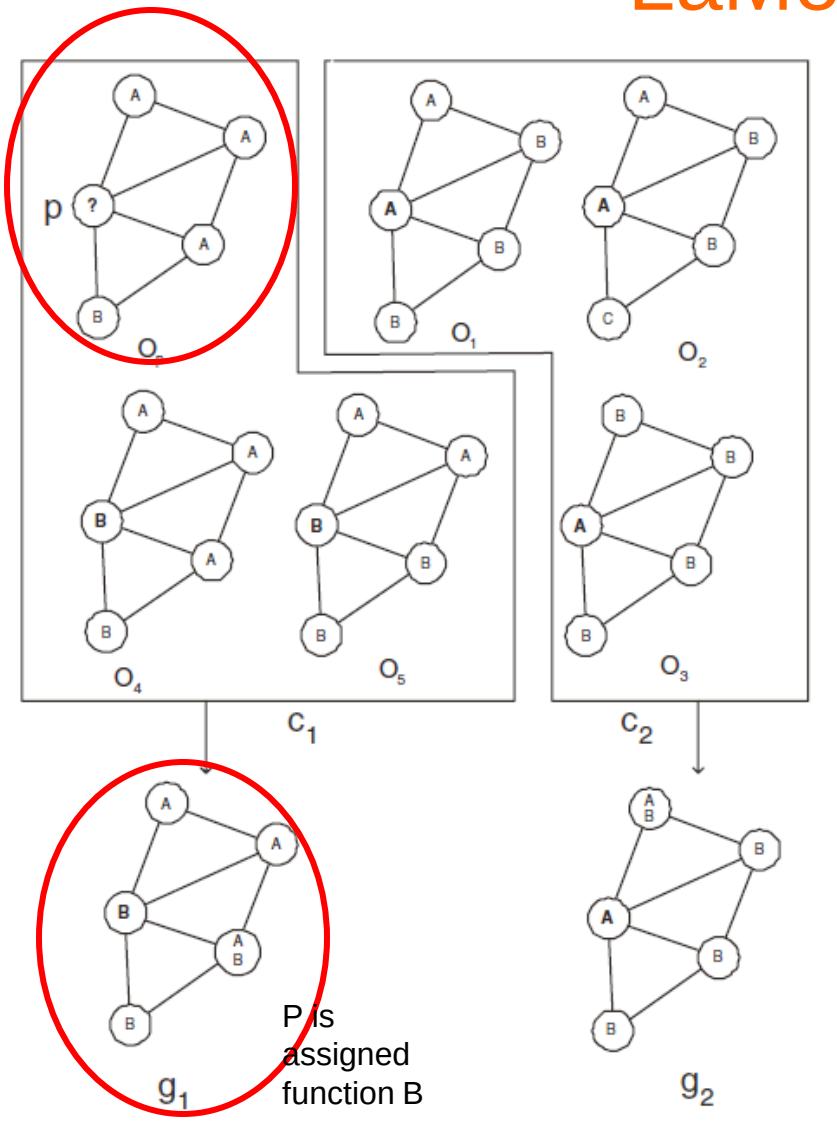


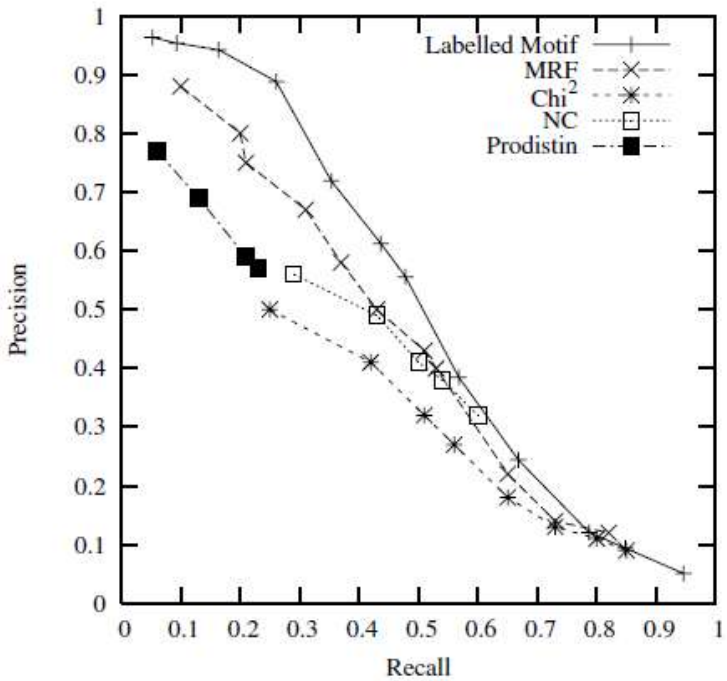
4 occurrences of 'g' in this PPIN

Image credit: Chen et al. *ICDE2007*, pp. *546–555*

Chen et al. *ICDE2007*, pp. *546–555*

LaMoFinder





- **Shortcoming**
  - Works only for proteins in subnets that can be mapped to network motifs

Pattern-Based Annotation Prediction (PAP)

- Kirac & Ozsoyoglu, *RECOMB2008*, pp 197-213
- **Find the best pairwise graph alignment of the functionally labeled subgraph rooted at the unknown protein to functionally labeled subgraphs rooted at other nodes in the protein interaction network**

- **Shortcoming**
  - Rely on topological matching of subnetworks

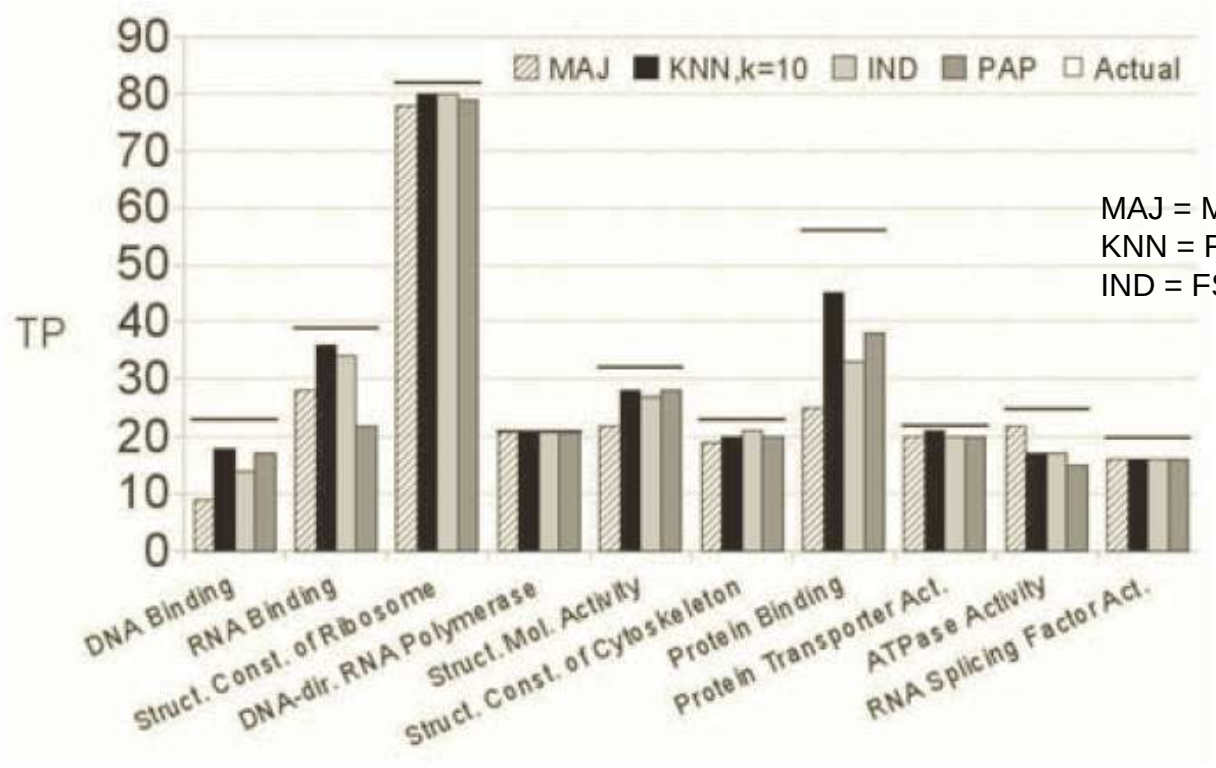
  ⇒Sensitive to noise & missing edges in PPIN

Functional Neighborhood Features

- Bogdanov & Singh. *TCBB,* 7:208–217, 2010
- **Predict function of an unknown protein v by weighted voting of the k proteins having most similar functional profiles to v**

- **Affinity of protein u to protein v**
  - $P_{u,v}$ = Prob of random walks from u to v
- **Affinity of protein v to function a**
  - $Sf_v(a) = \Sigma P_{u,v}$, over all proteins u having function a
- **Functional profile of a protein v**
  - $[Sf_v(a_1), \ldots, Sf_v(a_k)]$, normalized

Comparisons



MAJ = Majority Voting
KNN = Functional Neighborhood Features
IND = FSWeight

Fig. 10. Number of TP per GO molecular function (*FYI*, $T = 20$). The top two functions are considered as predictions for each of the methods. The horizontal bars represent the total number of TPs for each GO term.

- **Functional neighborhood features is slightly better than FSWeight**

Bogdanov & Singh. *TCBB,* 7:208–217, 2010

What have we learned?

- **Proteins with similar function can be far apart**

- **If the functional neighborhood features of two proteins are similar, they may have similar function**

$\Rightarrow$ **Assign protein to a function based on network motif (and generalizations thereof) that it matches**

References

- **Must Read**
  - Wong. "Using biological networks in protein function prediction and gene expression analysis". *Internet Math*, 7(4):274--298, 2011
  - [FSWeight] Chua et al. "Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions". *Bioinformatics*, 22:1623-1630, 2006
- **Good to Read**
  - [Majority Voting, $\chi 2$] Hishigaki et al. "Assessment of prediction accuracy of protein function from protein-protein interaction data". *Yeast*, 18:523-531, 2001
  - [LaMoFinder] Chen et al. "Labeling Network Motifs in Protein Interactomes for Protein Function Prediction". *ICDE2007*, 546–555
  - [PAP] Kirac & Ozsoyoglu. "Protein Function Prediction based on Patterns in Biological Networks". *RECOMB2008*, 197–213
  - [Functional Neighborhood Features] Bogdanov & Singh. "Molecular Function Prediction Using Neighborhood Features". *TCBB*, 7:208–217, 2010

Guilt by Association of Multiple Types of Information

Limsoon Wong

NUS
National University of Singapore

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

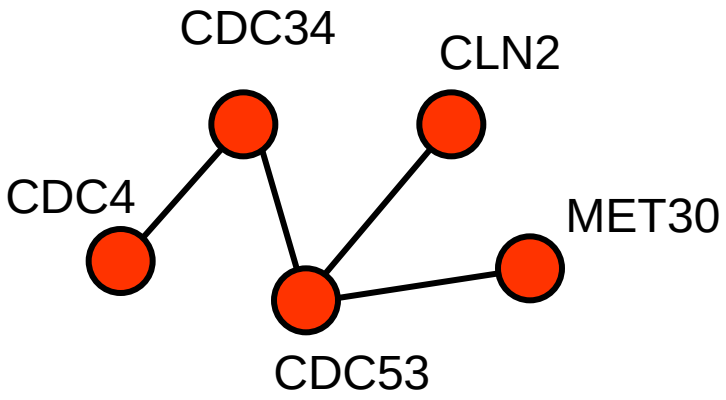
Motivation

- **Unified scoring of multiple sources has potential**
  - Lee et al., "Probabilistic functional network of yeast genes". *Science*, 306:1555–1558, 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

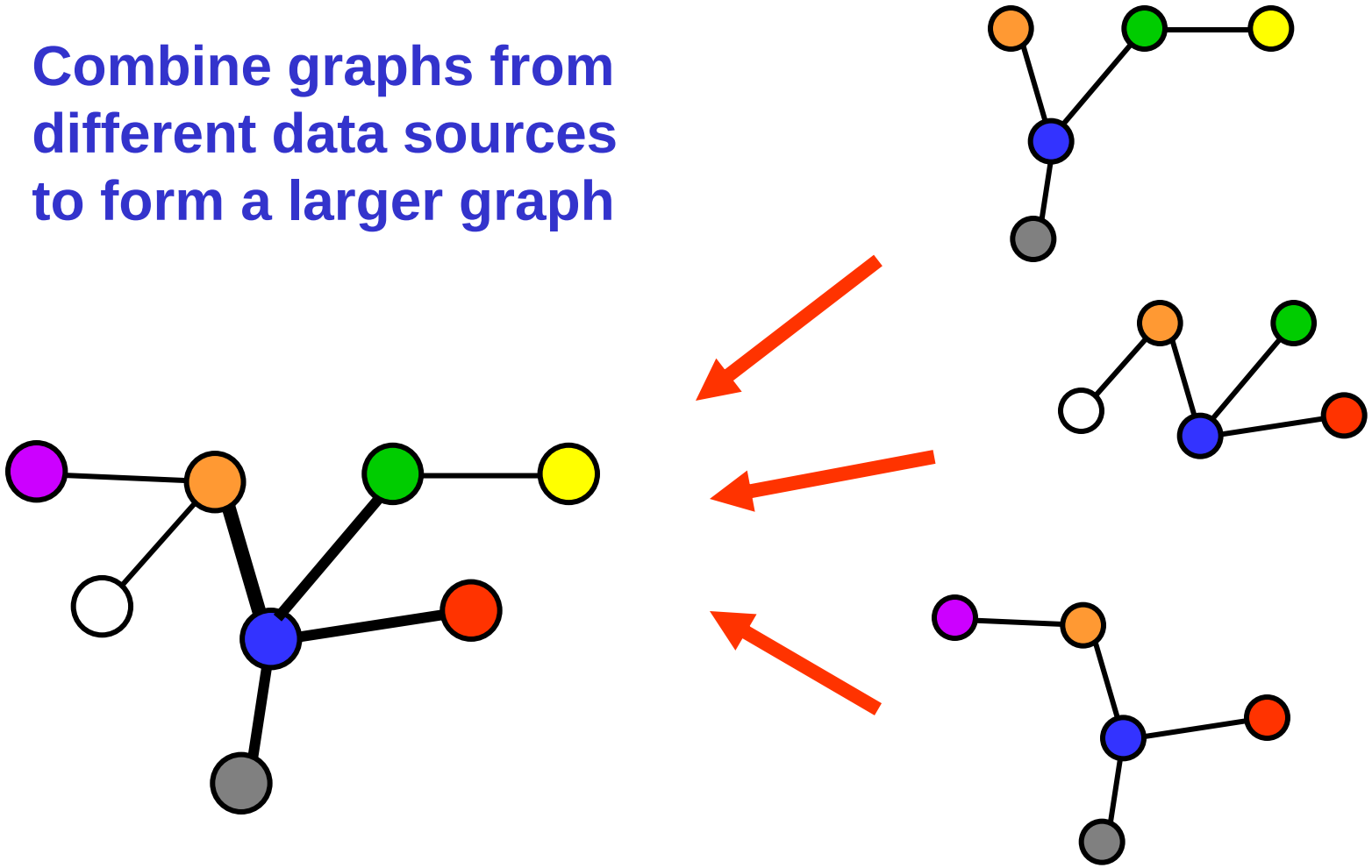
- **Model data source as undirected graph G = ⟨V,E⟩**
  - V is a set of vertices; each vertex reps a protein
  - E is a set of edges; each edge (u , v) reps a relationship (e.g. seq similarity, interaction) betw proteins u and v



Chua et al. *Bioinformatics*, 23(24):3364-3373, 2007

Strategy – Step 2

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

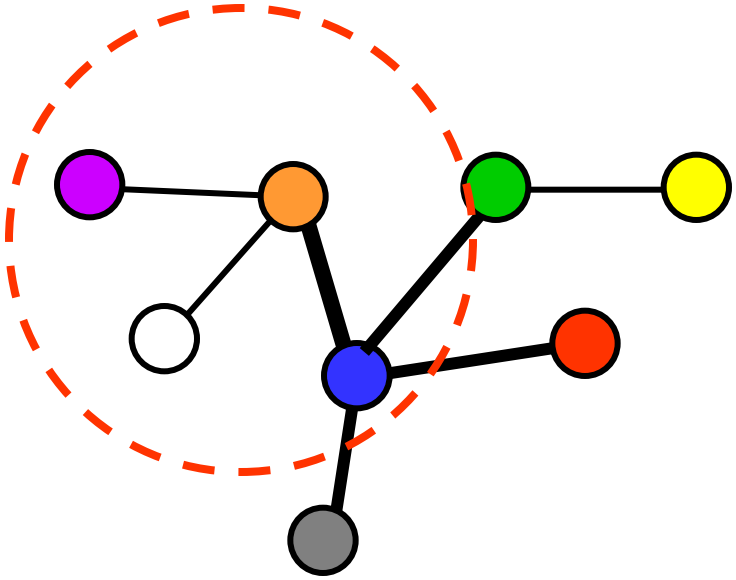


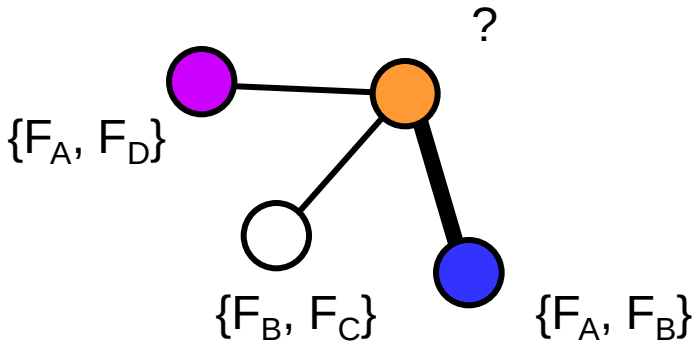
Chua et al. *Bioinformatics*, 23(24):3364-3373, 2007

Strategy – Step 3

- **Estimate edge confidence from contributing data sources**

- **Predict function by observing which functions occur frequently in the high-confidence neighbours**





Chua et al. *Bioinformatics*, 23(24):3364-3373, 2007

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:**

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

  - **$E_{k,f}$ 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**

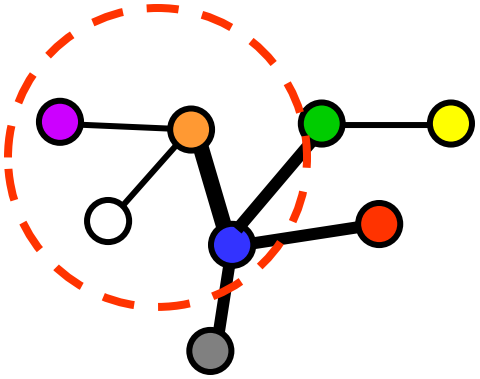
Chua et al. *Bioinformatics*, 23(24):3364-3373, 2007

Combination of Confidence

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

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

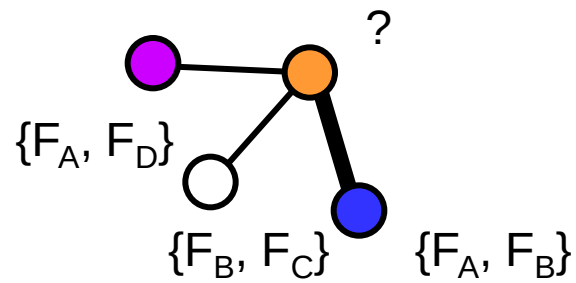
  - **P(k.f) is confidence of edges of subtype k sharing function f**
  - **$D_{u,v}$ is the set of subtypes of data sources which contains the edge (u,v)**



Chua et al. *Bioinformatics*, 23(24):3364-3373, 2007

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)**

Chua et al. *Bioinformatics*, 23(24):3364-3373, 2007

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

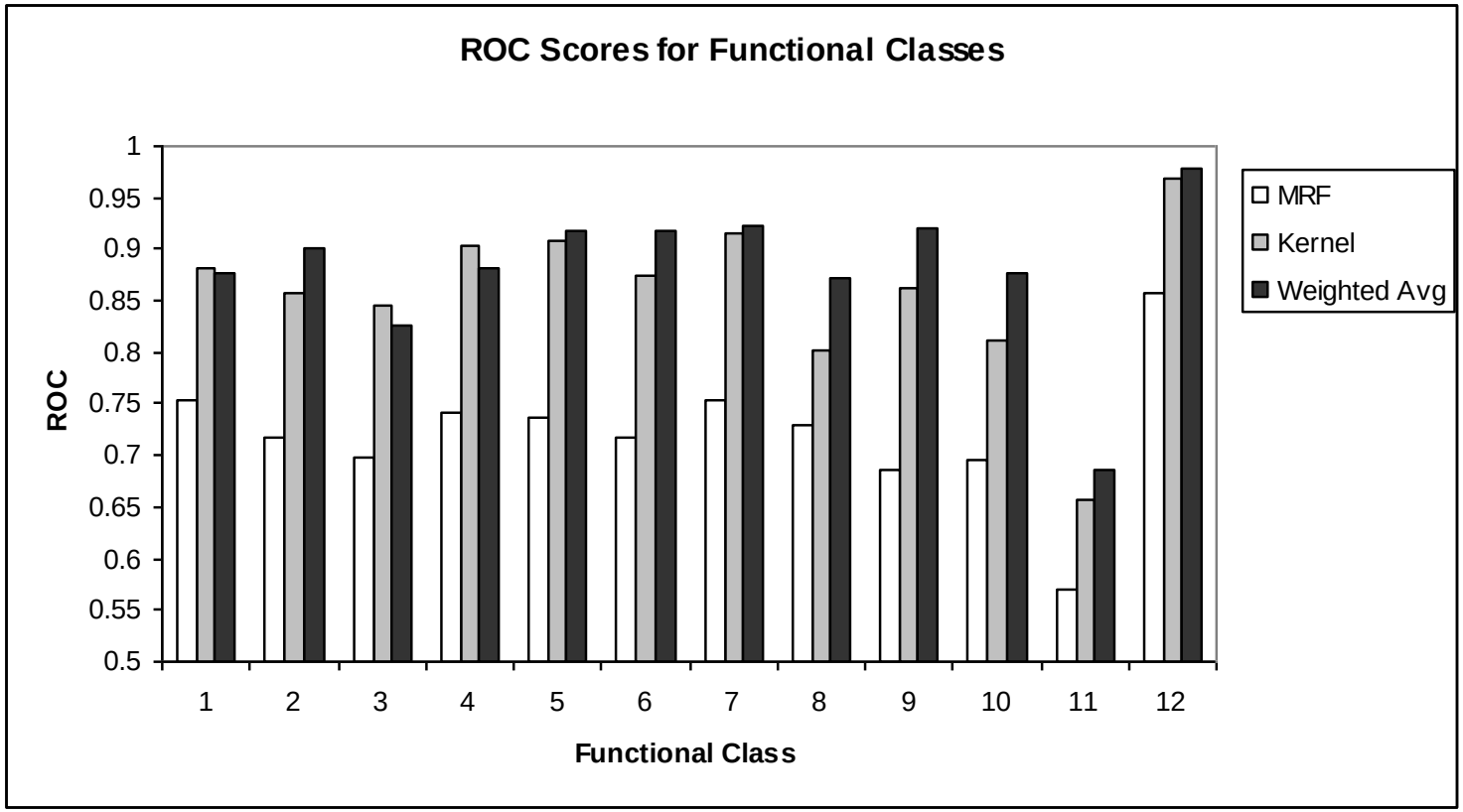
- **12 functional classes**

	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, *PSB 2004*, pp. 300-311
  - Area under ROC curve for each function
  - Averaged over 3 repetitions of 5-fold cross validation

Comparison w/ Existing Approaches

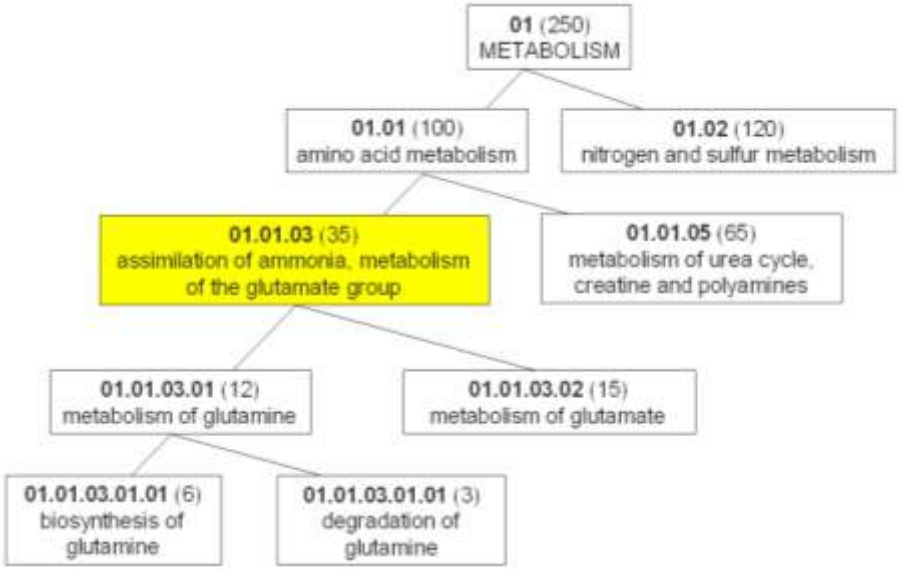


Chua et al. *Bioinformatics*, 23(24):3364-3373, 2007

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

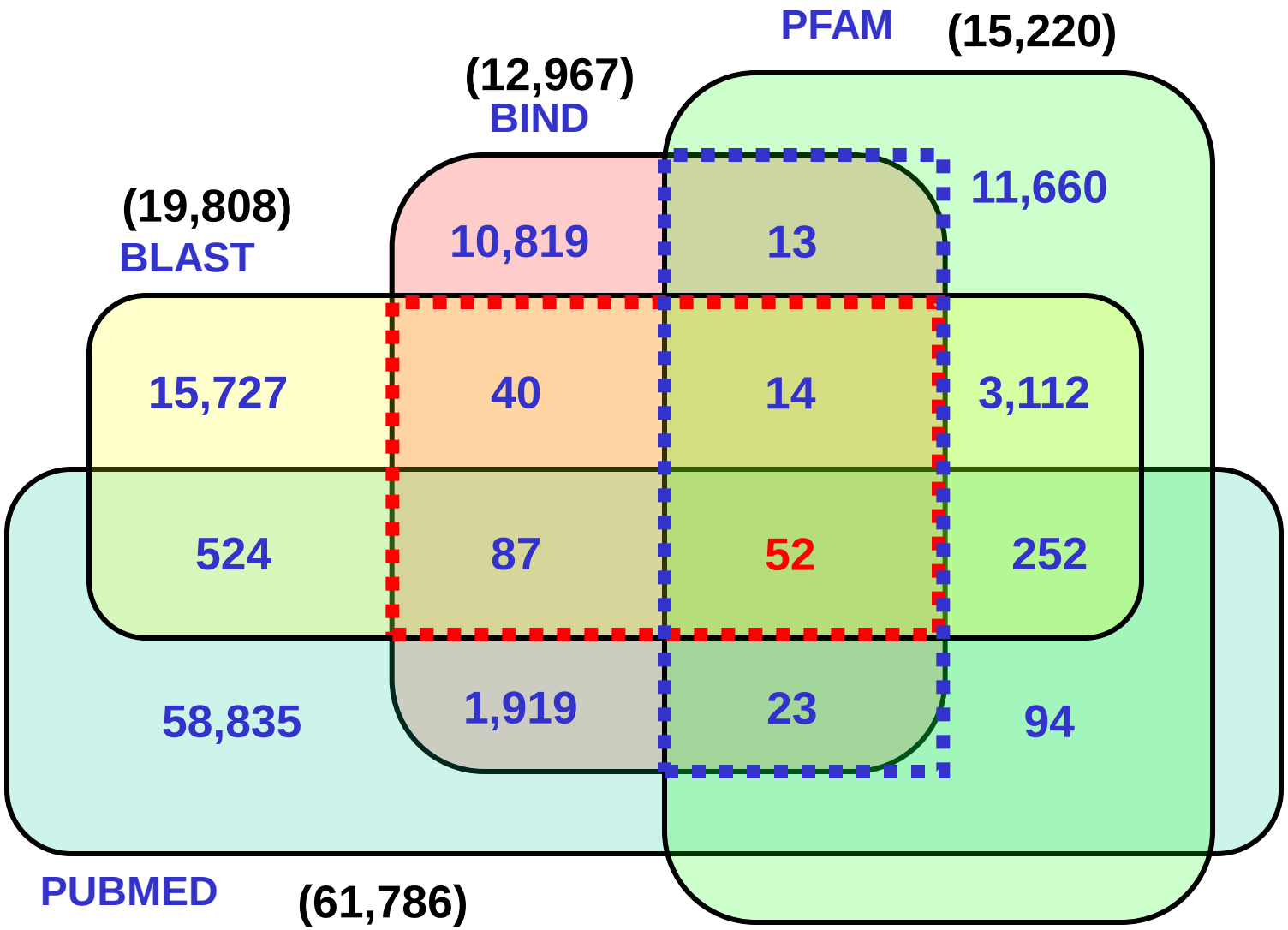
- **PPI**
  - BIND
  - 12,967 unique interactions betw yeast proteins
  - Score = FS weight
- **Protein Sequences**
  - Seqs from GO database
  - Each yeast seq is aligned w/ rest using BLAST
  - Score = –log(e_score)
  - Top 5 results w/ known annotations
  - 19,808 unique pairs involving yeast proteins

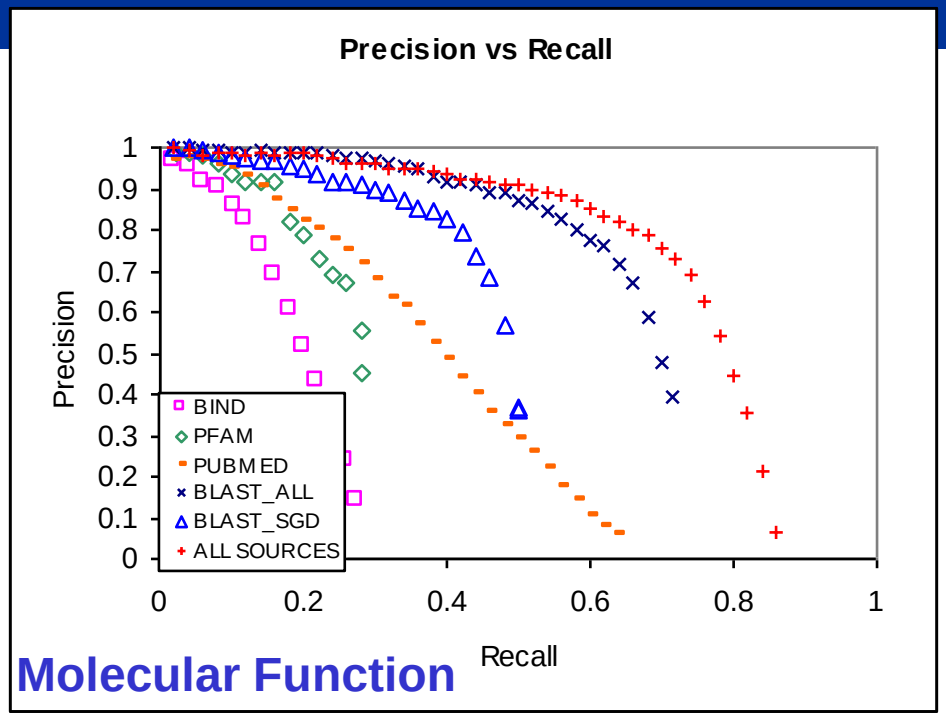
Data Sources

- **Pfam Domains**
  - SwissPfam database
  - Pfam domains for SwissProt & TrEMBL proteins w/ E-value threshold 0.01
  - Score = # of common domains
  - 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
  - Score = Fraction of abstracts w/ co-occurrence
  - 61,786 unique pairs involving yeast proteins

Chua et al. *Bioinformatics*, 23(24):3364-3373, 2007

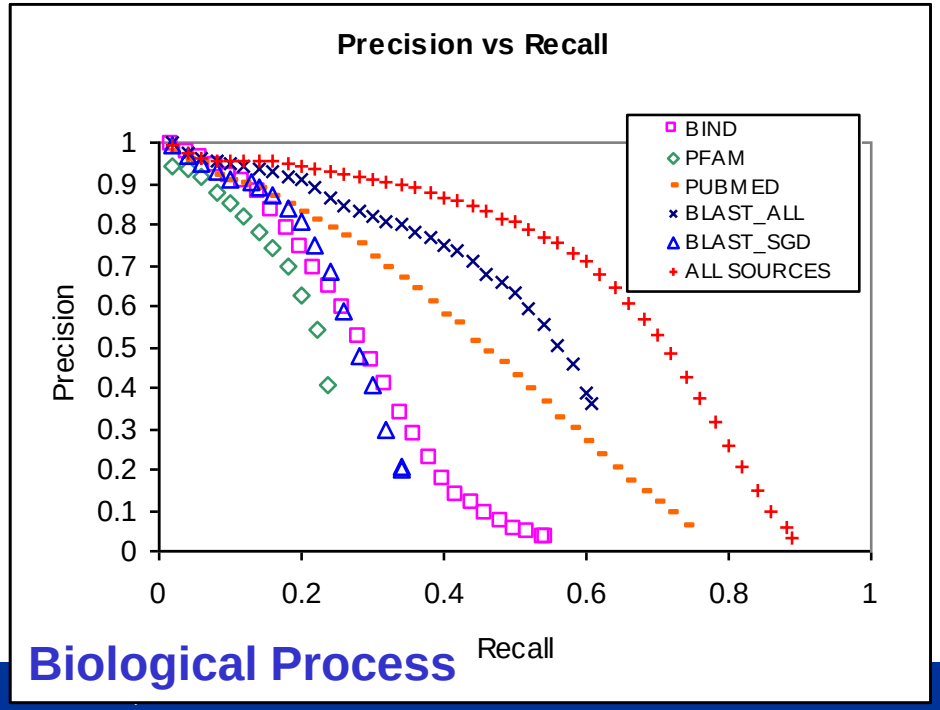
Multiple Data Sources

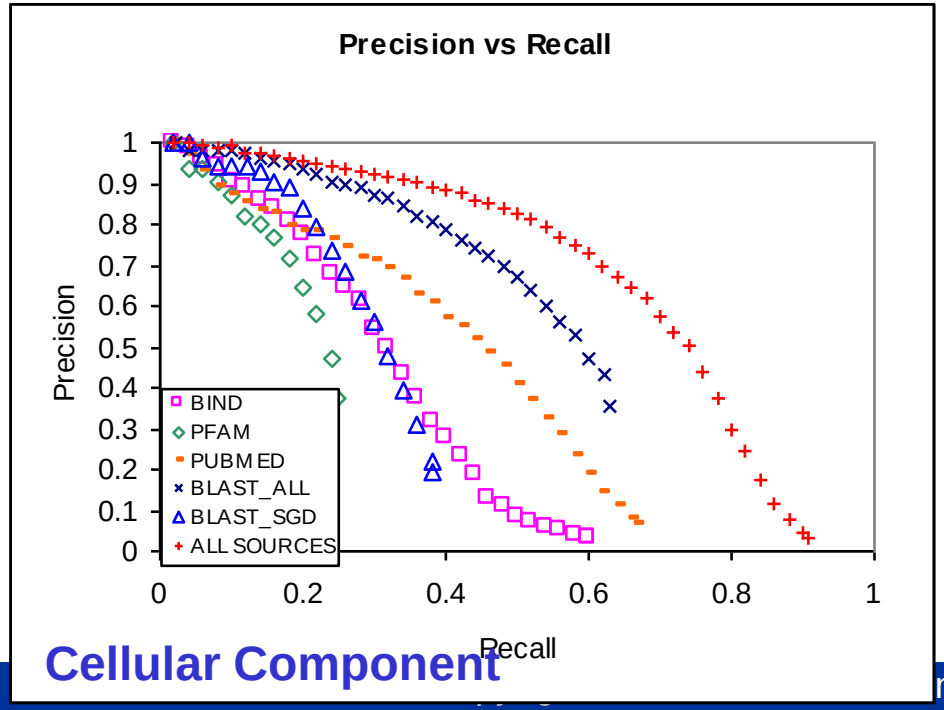


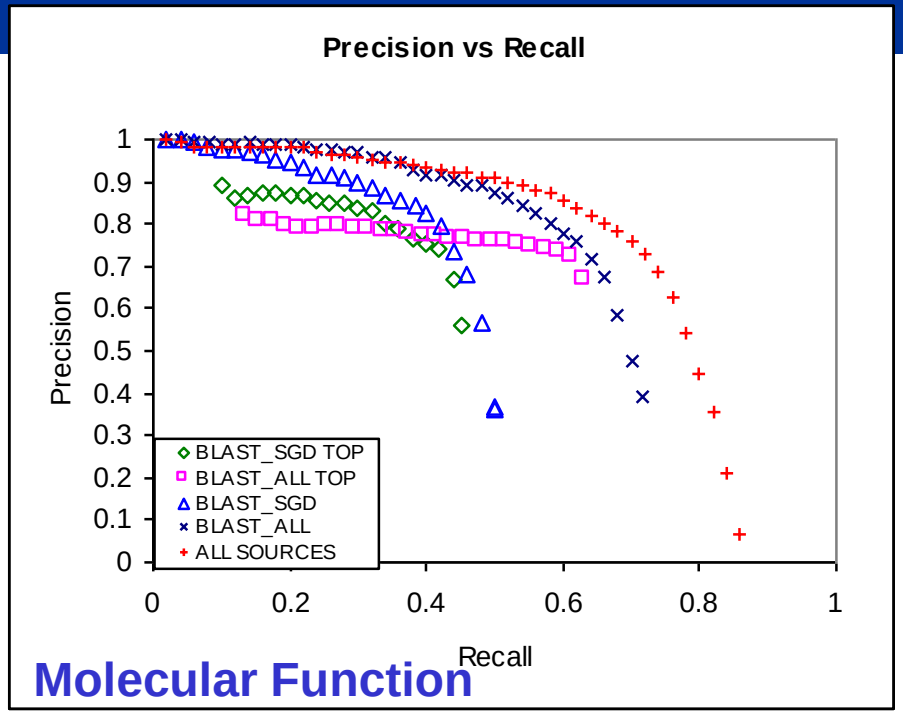


Combining all data sources outperforms any single data source

Chua et al. *Bioinformatics*, 23(24):3364-3373, 2007



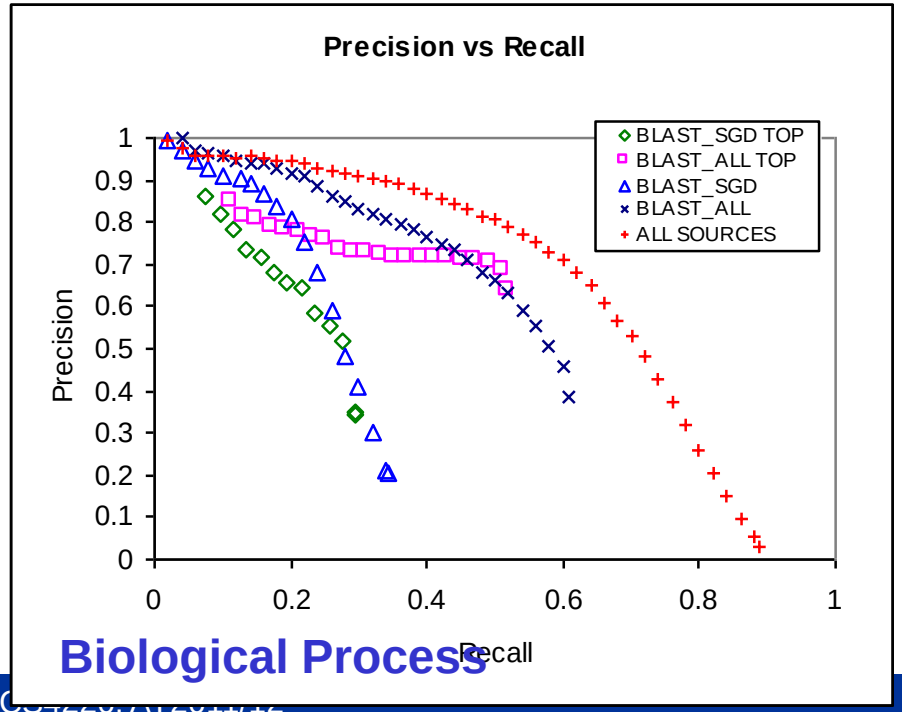




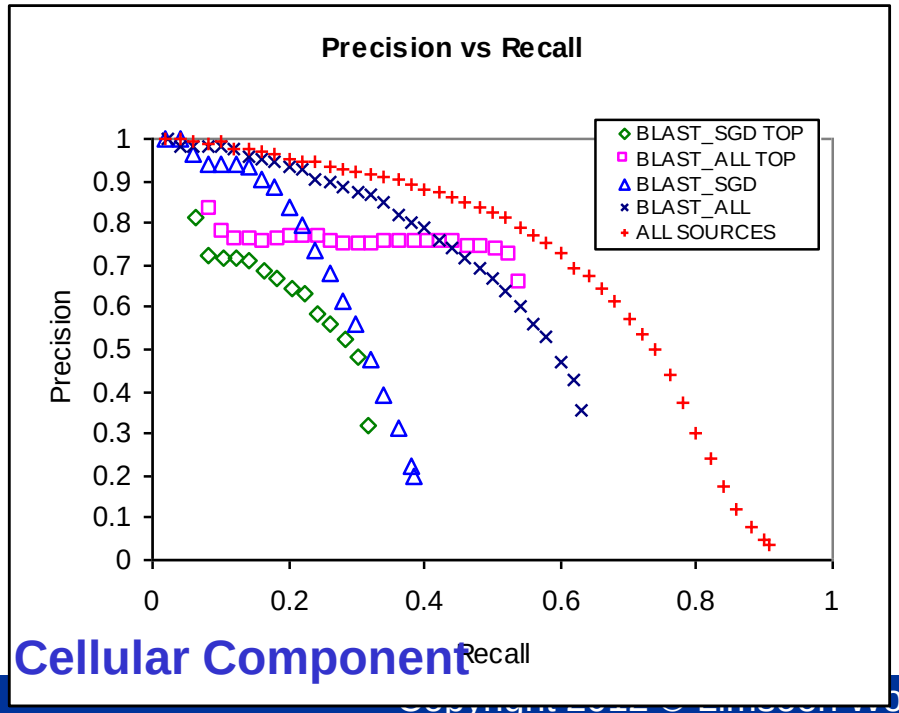
Molecular Function

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

Chua et al. *Bioinformatics*, 23(24):3364-3373, 2007



Biological Process

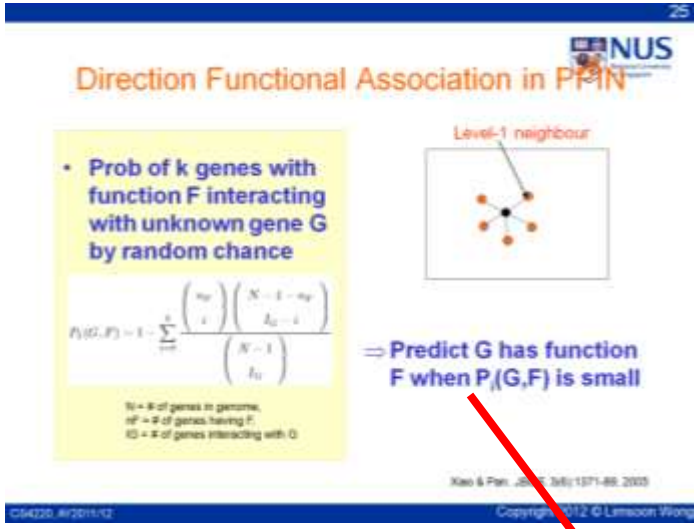


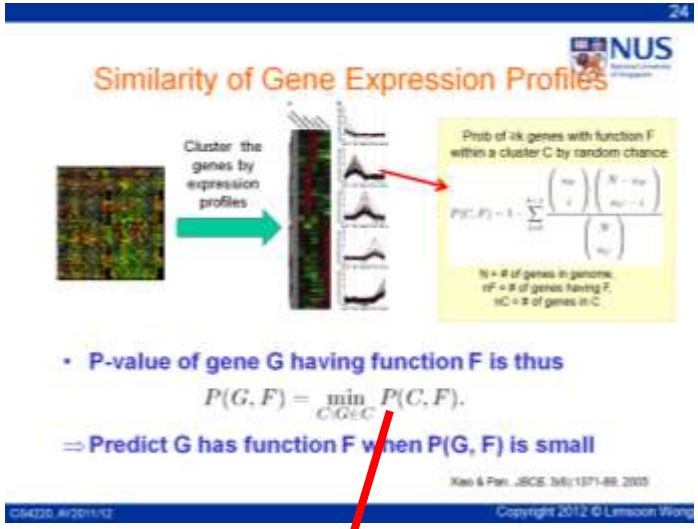
Cellular Component

Conclusions

- **A graph-based method that combines multiple sources of data sources for function prediction**
- **It is simple, flexible and can report data sources contributing to each prediction**
- **It performs comparable, if not better, than existing approaches**

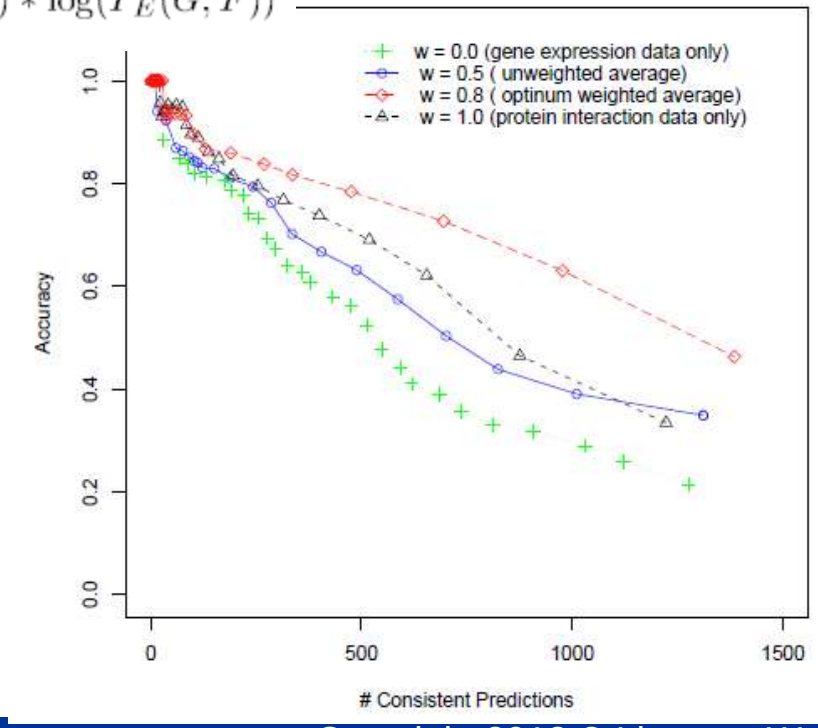
There are many other ways to integrate multiple types of information for protein function prediction…





$$\log(P(G,F)) = w * \log(P_I(G,F)) + (1-w) * \log(P_E(G,F))$$

Combining GE & PPI Data



Xiao & Pan. *JBCB*, 3(6):1371-89, 2005

General Information Fusion Methods

- **Markov Random Fields**
  - Deng et al., *JCB*, 11(2-3):463-75, 2004
  - Maximum Likelihood
  - Model data sources as binary relation betw proteins

- **Kernel Fusion**
  - Lanckriet et al., *PSB 2004*, pp. 300-311
  - Discriminative approach
  - Models each data source w/ diff feature vectors
  - Weighted linear combination of kernels via semi-definite programming

References

- **Must Read**
  - Chua et al. “An efficient strategy for extensive integration of diverse biological data for protein function prediction”. *Bioinformatics*, 23(24):3364-3373, 2007
- **Good to Read**
  - Deng et al. “An integrated probabilistic model for functional prediction of proteins”. *JCB*, 11(2-3):463-75, 2004.
  - Lanckriet et al. “Kernel-based data fusion and its application to protein function prediction in yeast”. *PSB 2004*, pp. 300-311.
  - Martin et al. “GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes”. *BMC Bioinformatics*. 5:178, 2004
  - Xiao & Pan. “Gene function prediction by a combined analysis of gene expression data and protein-protein interaction data”. *JBCB*, 3(6):1371-89, 2005

Acknowledgements

Kenny Chua

- **A large part of this lecture is based on work done by my past student, Kenny Chua**