

Guilt by Association: A Tutorial on Data Mining Techniques for Protein Function Inference

Limsoon Wong

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



PAKDD2007, Nanjing, 22 May 2007

2

Plan

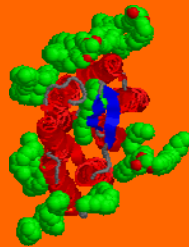


- Protein Function Prediction
- Guilt by Association of Seq Similarity
- Twists in the Tale
- Guilt by Association of Other Type of Info
- Guilt by Association of Multiple Types of Info

PAKDD2007, Nanjing, 2007

Copyright 2007 © Limsoon Wong

Protein Function Prediction: Motivation & Challenges



PAKDD2007, Nanjing, 22 May 2007

4



- 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



PAKDD2007, Nanjing, 2007

Copyright 2007 © Limsoon Wong

Function Assignment to Protein Seq

SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
 YVNILPYDHSRVHLTPVEGVPDSYINASFINGYQEKNFIAAQGPKEETVNDFWMIWE
 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
 VTNRKPQLITQFHFTSWPDFGVFPTPIGMLKFLKVKACNPQYAGAIVVHCSAGVGRGTG
 TFVVIDAMLDMMHSEKVDVYGFVSRIRAQRQMVQTDMQYVFIYQALLEHYLYGDTELE
 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 LARGKRLGSLVAEPAMIAECKTRTEVFEISRRLIDRTN 100
```

⇒ “Guilt by association” of sequence similarity!

Guilt by Association of Sequence Similarity

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



PAKDD2007, Nanjing, 22 May 2007

8

Guilt by Association: General Idea



- Compare the target sequence T with sequences S_1, \dots, S_n of known function in a database
- Determine which ones amongst S_1, \dots, S_n are the mostly likely homologs of T
- Then assign to T the same function as these homologs
- Finally, confirm with suitable wet experiments

PAKDD2007, Nanjing, 2007

Copyright 2007 © Limsoon Wong

Guilt by Association of Seq Similarity



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

```

Amicyanin      60 70 80 90 100
MFRVYVAVGLGSAALGPRKKGQATSLFFTEAGTDFICTYHFFRGRVVV
Ascorbate Oxidase  ILQRTWADGTASISQCAINPGEFFYFVDFVDFPFFRCHLQNRAGLVG
                    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

```

gi11347672|ref|NP_108301.1| unknown protein [Mesochorus loti]
gi11497493|ref|NP_076576.1| unknown protein [Mesochorus loti]
Length = 105

Score = 105 bits (262), Expect = 1e-22
Identities = 41/106 (39%), Positives = 73/106 (69%), Gaps = 1/106 (0%)
Query: 1  MKNELASLALAIPLPMTYAMAATIEITNRELFITSTYCAKVDITDFPQVYANT 60
           MKQLL  **  MAFAAATIE++LQSPVAKVDITDFPQVYANT 60
Sbjct: 1  MKNELASLALAIPLPMTYAMAATIEITNRELFITSTYCAKVDITDFPQVYANT 60
  
```

good match between Amicyanin and unknown M. loti protein

Assign to T same function as homologs

Confirm with suitable wet experiments

Seq Alignment



```

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

- A seq alignment maximizes the number of positions that are in agreement in two sequences
- Many implementations:
 - Global vs local alignment
 - Gapped vs ungapped
 - Filtered vs unfiltered, ...

Seq Alignment: Poor Example

- 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      MPHNVHVFVAGVLGEAALKGPFMMKKEQAYSLEFFTEAGTYDYHCTFHPFMRGKVVVE
                :...: . :...: :
Ascorbate Oxidase ILQRQTFWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYGSLI
                70      80      90      100      110      120

```

No obvious match between
Amicyanin and Ascorbate Oxidase

Seq Alignment: Good Example

- Good alignment usually has clusters of extensive matched positions
- ⇒ The two proteins are likely to be homologous

```

□ >gi113476732|ref|NP_108301.1| unknown protein [Mesorhizobium loti]
  gi114027493|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  MKPGRLASIALAIIFLPMVPAHAATIEITMENLVISPTVEVSAKVGDTIRWVNKDVFAHT 60
          MK G L  ++      MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1  MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVWNVVFAHT 60

```

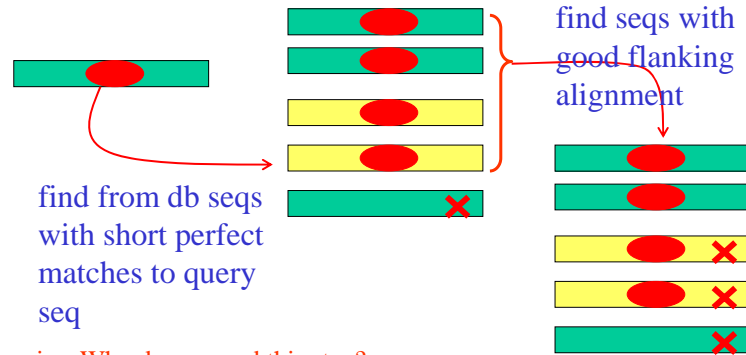
good match between
Amicyanin and unknown *M. loti* protein

BLAST: How It Works

Altschul et al., *JMB*, 215:403--410, 1990



- **BLAST is the most popular tool for “guilt by association” seq homology search**



Exercise: Why do we need this step?

NCBI *protein-protein* BLAST

Nucleotide Protein Translations Retrieve results for an RID

Search

NRYVMILFYDHSRVHLTPVEGVDPDSYINASFINGYQEKNFIAAQPKEETVNDVURMIUEQNTATIVMVTNLKERKECKCAQYMPDQGCUTYGNVRUSVEDVTVLVDYTVRKFCIQQVGDVTNRKFPRLITQFHFSTWDFGVFPFIPIGHLKFLKKVKACNPOYAGAIIVHCSAGVGRIGTFVVIDAMLDMHSEKVDVYGFVSRIRACRCQKVVQDMQYVFIYQALLEHVLVYGDTELE

Set subsequence From: To:

Choose database nr

Do CD-Search

Now: **BLAST!** or

Options for advanced blasting

Limit by [entrez query](#) or select from: All organisms



Homologs by BLAST

Sequences producing significant alignments:	Score (bits)	E Value
gi 14193729 cb AAK56109.1 AF332081.1 protein tyrosin phosph...	62.1	e-177
gi 126467 sp P18433 PTRA_HUMAN Protein-tyrosine phosphatase...	62.1	e-177
gi 4506303 ref NP_002827.1 protein tyrosine phosphatase, r...	62.1	e-176
gi 227294 prf I1701300A protein Tyr phosphatase	62.0	e-176
gi 18450369 ref NP_543030.1 protein tyrosine phosphatase, ...	62.1	e-176
gi 32067 emb CAA37447.1 tyrosine phosphatase precursor [Ho...	61.1	e-176
gi 285113 pir IJC1285 protein-tyrosine-phosphatase (EC 3.1....	61.9	e-176
gi 6981446 ref NP_036895.1 protein tyrosine phosphatase, r...	61.1	e-176
gi 2098414 pdb 1YFO A Chain A, Receptor Protein Tyrosine Ph...	61.1	e-174
gi 32313 emb CAA38662.1 protein-tyrosine phosphatase [Homo...	61.1	e-174
gi 450583 gb AA04150.1 protein tyrosine phosphatase >gi 4...	60.5	e-172
gi 6679557 ref NP_033006.1 protein tyrosine phosphatase, r...	60.1	e-172
gi 483922 gb AA17990.1 protein tyrosine phosphatase alpha	59.9	e-170

- Thus our example sequence could be a protein tyrosine phosphatase α (PTP α)



Example Alignment with PTP α

Score = 632 bits (1629), Expect = e-180
 Identities = 294/302 (97%), Positives = 294/302 (97%)

```

Query: 1  SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACP:QATCEAASXXXXXXXXR 60
          SPSTNRKYPPI.PVDKI.FEEINRRMADDNKI.FRRFFVAI.PACP:QATCEAAS      R
Sbjct: 202 SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACP:QATCEAASKEENKEKNR 261

Query: 61  YVNILPYDHSRVHLTPVEGVPSDYINASFINGYQEKKNFIAAQGPKEETVNDFWRMIWE 120
          YVNILPYDHSRVHLTPVEGVPSDYINASFINGYQEKKNFIAAQGPKEETVNDFWRMIWE
Sbjct: 262 YVNILPYDHSRVHLTPVEGVPSDYINASFINGYQEKKNFIAAQGPKEETVNDFWRMIWE 321

Query: 121 QNTATIYVMVTNLKERKECKCAQYWPDQGCWTYGNVRSVSDV:VLVDYTVRKFCIQQVGD 180
          QNTATIYVMVTNLKERKECKCAQYWPDQGCWTYGNVRSVSDV:VLVDYTVRKFCIQQVGD
Sbjct: 322 QNTATIYVMVTNLKERKECKCAQYWPDQGCWTYGNVRSVSDV:VLVDYTVRKFCIQQVGD 381

Query: 181 VTRKPKQLITQFHFTSWPFGVFPFTPIGMLKFLKXVKACNPQYAGAIYVHCSAGVGRGTG 240
          VTRKPKQLITQFHFTSWPFGVFPFTPIGMLKFLKXVKACNPQYAGAIYVHCSAGVGRGTG
Sbjct: 382 VTRKPKQLITQFHFTSWPFGVFPFTPIGMLKFLKXVKACNPQYAGAIYVHCSAGVGRGTG 441

Query: 241 TFVVIDAMLDDMHSEKVDVYGFVSRIRAQRCQMVQTDMQYVFIVQALLEHYLYGDTELE 300
          TFVVIDAMLDDMHSEKVDVYGFVSRIRAQRCQMVQTDMQYVFIVQALLEHYLYGDTELE
Sbjct: 442 TFVVIDAMLDDMHSEKVDVYGFVSRIRAQRCQMVQTDMQYVFIVQALLEHYLYGDTELE 501
  
```

References

- S.F. Altschul et al. "Basic local alignment search tool", *JMB*, 215:403--410, 1990
- S.F. Altschul et al. "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs", *NAR*, 25(17):3389--3402, 1997
- D. Brown et al. "Homology Search Methods", *The Practical Bioinformatician*, Chapter 10, pp 217—244, WSPC, 2004
- S.B. Needleman & C.D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins", *JMB*, 48:444—453, 1970
- J. Park et al. "Sequence comparisons using multiple sequences detect three times as many remote homologs as pairwise methods", *JMB*, 284(4):1201--1210, 1998
- T.F. Smith & M.S. Waterman. "Identification of common molecular subsequences", *JMB*, 147:195—197, 1981
- Z. Zhang et al. "Protein sequence similarity searches using patterns as seeds", *NAR*, 26(17):3986—3990, 1996

Twists in the Tale of Guilt by Association of Seq Similarity





Seq Similarity: Caveats

- Ensure that the effect of database size and other biases 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



Law of Large Numbers

- Suppose you are in a room with 365 other people
- Q: What is the prob that a specific person in the room has the same birthday as you?
- A: $1/365 = 0.3\%$
- Q: What is the prob that there is a person in the room having the same birthday as you?
- A: $1 - (364/365)^{365} = 63\%$
- Q: What is the prob that there are two persons in the room having the same birthday?
- A: 100%

Interpretation of P-value

- Seq. comparison progs, e.g. BLAST, often associate a P-value to each hit
 - P-value is interpreted as prob that a random seq has an equally good alignment
 - Suppose the P-value of an alignment is 10^{-6}
 - If database has 10^7 seqs, then you expect $10^7 * 10^{-6} = 10$ seqs in it that give an equally good alignment
- ⇒ Need to correct for database size if your seq comparison prog does not do that!

Lightning Does Strike Twice!

- **Roy Sullivan, a former park ranger from Virginia, was struck by lightning 7 times**
 - 1942 (lost big-toe nail)
 - 1969 (lost eyebrows)
 - 1970 (left shoulder seared)
 - 1972 (hair set on fire)
 - 1973 (hair set on fire & legs seared)
 - 1976 (ankle injured)
 - 1977 (chest & stomach burned)
- **September 1983, he committed suicide**



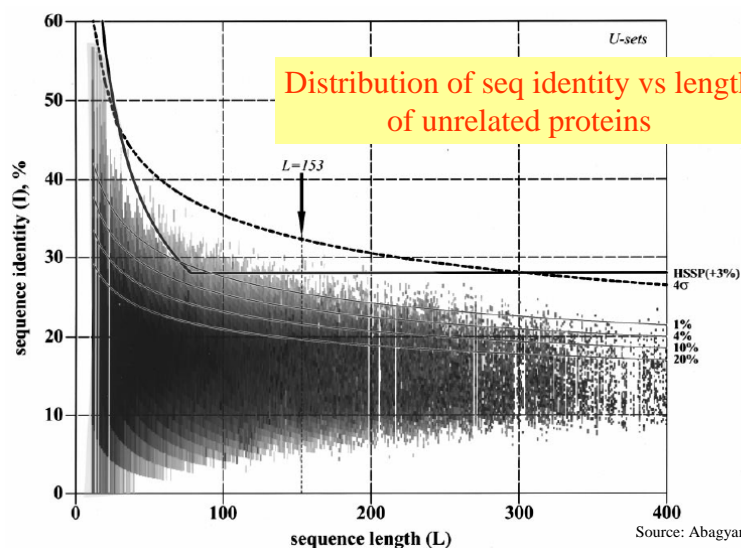
Cartoon: Ron Hipschman
Data: David Hand

Effect of Seq Compositional Bias

- One fourth of all residues in protein seqs occur in regions with biased amino acid composition
 - Alignments of two such regions achieves high score purely due to segment composition
- ⇒ While it is worth noting that two proteins contain similar low complexity regions, they are best excluded when constructing alignments
- E.g., by default, BLAST employs the SEG algo to filter low complexity regions from proteins before executing a search

Source: NCBI

Effect of Seq Length



Seq Similarity: Caveats

- Ensure that the effect of database size and other biases 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

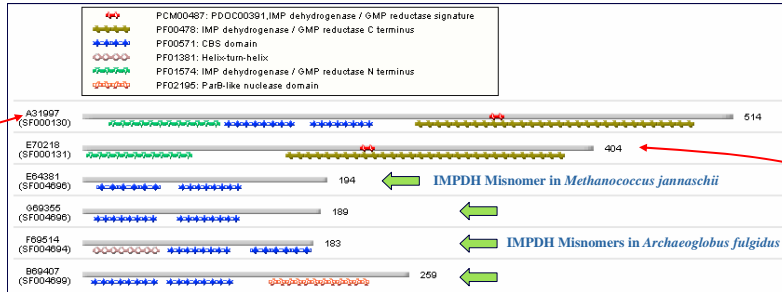
Examples of Invalid Function Assignment: The IMP Dehydrogenases (IMPDH)

18 entries were found

ID	Organism	PIR	Swiss-Prot/TrEMBL	RefSeq/GenPept
NF00181857	Methanococcus jannaschii	E64381 conserved hypothetical protein MJ0653	Y633_MET1A Hypothetical protein MJ0653	gi_1922300 inosine-5'-monophosphate dehydrogenase (guaf) NP_247627 inosine-5'-monophosphate dehydrogenase (guaf)
NF00187788	Archaeoglobus fulgidus	G69355 MJ0653 homolog AF0847 ALT_NAME5: inosine-monophosphate dehydrogenase (guaf-1) homolog [misnomer]	G28411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)	gi_2649754 inosine monophosphate dehydrogenase (guaf-1) NP_069631 inosine monophosphate dehydrogenase (guaf-1)
NF00188267	Archaeoglobus fulgidus	E29214 yhcV homolog 2 ALT_NAME5: inosine-monophosphate dehydrogenase (guaf-2) homolog [misnomer]	G28162 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-2)	gi_2649754 inosine monophosphate dehydrogenase (guaf-2) NP_070243 inosine monophosphate dehydrogenase (guaf-2)
NF00188697	Archaeo			inosophosphate rve monophosphate rve
NF00197776	Thermo			monophosphate d protein monophosphate d protein
NF00414709	Methanothermobacter thermoautotrophicus	G29025 conserved hypothetical protein ALT_NAME5: inosine-monophosphate dehydrogenase related protein V [misnomer]	G27204 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN V	dehydrogenase related protein V NP_276334 inosine-5'-monophosphate dehydrogenase related protein V
NF00414811	Methanothermobacter thermoautotrophicus	D39035 M11222 protein homolog MTH126 ALT_NAME5: inosine-5'-monophosphate dehydrogenase related protein VII [misnomer]	G26229 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN VII	gi_2621186 inosine-5'-monophosphate dehydrogenase related protein VII NP_275289 inosine-5'-monophosphate dehydrogenase related protein VII
NF00414837	Methanothermobacter thermoautotrophicus	H49212 M11225 related protein MTH992 ALT_NAME5: inosine-5'-monophosphate dehydrogenase related protein IX [misnomer]	G27073 INOSINE-5' MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN IX	gi_2622823 inosine-5' monophosphate dehydrogenase related protein IX NP_276121 inosine-5'-monophosphate dehydrogenase related protein IX
NF00414969	Methanothermobacter thermoautotrophicus	E69077 yhcV homolog 2 ALT_NAME5: inosine-monophosphate dehydrogenase related protein X [misnomer]	G27616 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN X	gi_2622822 inosine-5' monophosphate dehydrogenase related protein X NP_276687 inosine-5'-monophosphate dehydrogenase related protein X

A partial list of IMPdehydrogenase misnomers in complete genomes remaining in some public databases

IMPDH Domain Structure



- Typical IMPDHs have 2 IMPDH domains that form the catalytic core and 2 CBS domains.
- A less common but functional IMPDH (E70218) lacks the CBS domains.
- Misnomers show similarity to the CBS domains

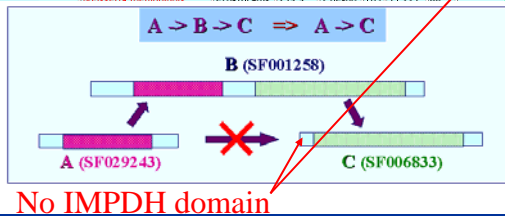
Invalid Transitive Assignment

Root of invalid transitive assignment

Accession	Gene ID	EC	Enzyme Name	Organism	Class	Length	Start	End	Start	End	Start	End
H70468	SF001258	051440	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Aquifex aeolicus	Prok/other	594.3	4.8e-26	205	39.086	197		
S76963	SF001258	039935	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Synechocystis sp.	Prok/gram-	557.0	5.7e-24	230	39.175	194		
T35073	SF029243	005738	probable phosphoribosyl-AMP cyclohydrolase	Streptomyces coelicolor	Prok/gram+	399.3	3.5e-15	128	42.157	102		
S53349	SF001257	001188	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) / histidinol dehydrogenase (EC 1.1.1.23)	Saccharomyces cerevisiae	Euk/fungi	384.1	2.5e-14	799	31.863	204		
E69493	SF029243	005738	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) [similarity]	Archaeoglobus fulgidus	Archae	396.8	4.8e-15	108	47.778	90		
G64337	SF006833	030827	phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Methanococcus jannaschii	Archae	246.9	1.1e-06	95	36.842	95		
D81178	SF006833	101491	phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) NMB0603 [similarity]	Naicocera meningitidis	Prok/gram-	239.0	7.4e-06	107	35.227	88		
G81925	SF006833	101491	phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) NMA0807 [similarity]									
S51513	SF001257	001188	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) / histidinol dehydrogenase (EC 1.1.1.23)									

B →
A →
C →

Mis-assignment of function



No IMPDH domain

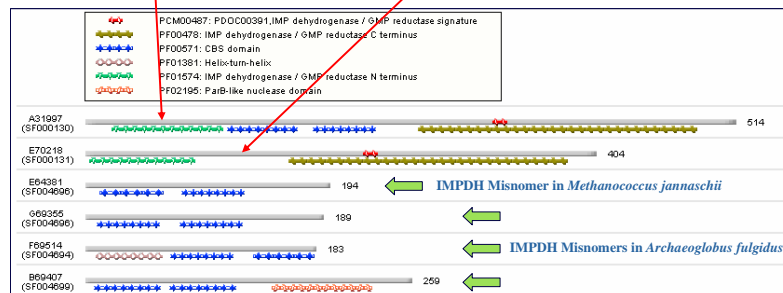
Seq Similarity: Caveats

- Ensure that the effect of database size and other biases 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

Emerging Pattern

Typical IMPDH

Functional IMPDH w/o CBS



- Most IMPDHs have 2 IMPDH and 2 CBS domains
 - Some IMPDH (E70218) lacks CBS domains
- ⇒ Alignment must preserve IMPDH domain to infer IMPDH

A more subtle twist ...

Identifying Key Mutation Sites

K.L.Lim et al., *JBC*, 273:28986--28993, 1998

Sequence from a typical PTP domain D2

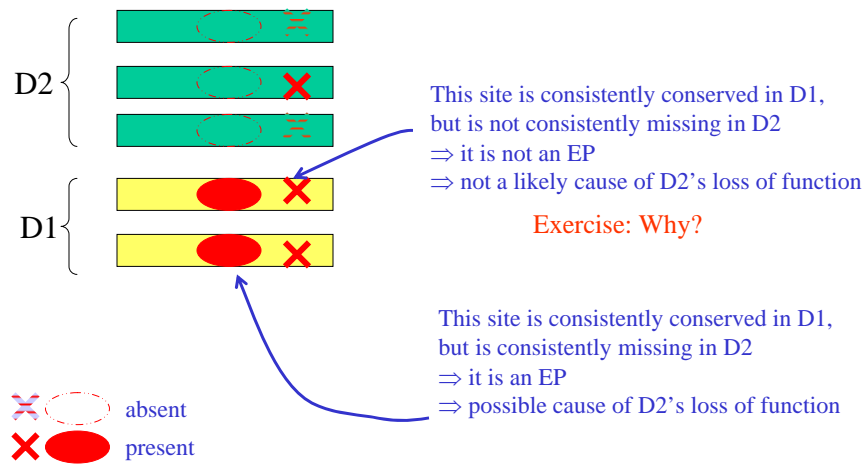
```
>g1|0000|PTPA-D2
EEEFKILTSIKIONDKIRTGMLPANEKIKNVLQIIPYEFHWIIPVAGEIDTDYHASF
IDGYRQKDSYIASQOPLEETIEDFURHIEWESCSIVELTELEERQQRCAQYPSDOLV
SYQDITVELKEEKECESTTVRDLVYNTREKESRQIQEFHFWPEYQIPSDGKQKLSII
AAVQRQQQSQNHPTVYECSSAQAGRTOTYFCALSTVLERVKAEQILDVYQTVIGLRLQRPE
EYQTLAQYEFCTYKVVQETIDAFSDYANFK
```

- Some PTPs have 2 PTP domains
- PTP domain D1 has much more activity than PTP domain D2
- Why? And how do you figure that out?

Emerging Patterns of PTP D1 vs D2

- Collect example PTP D1 sequences
- Collect example PTP D2 sequences
- Make multiple alignment A1 of PTP D1
- Make multiple alignment A2 of PTP D2
- Are there positions conserved in A1 that are violated in A2?
- These are candidate mutations that cause PTP activity to weaken
- Confirm by wet experiments

Emerging Patterns of PTP D1 vs D2



Key Mutation Site: PTP D1 vs D2

```

      ? ! ?           ?           ?           ? ??
gi|00000|P D2 QFHFGWPEVGIPSDGKMISIIAAVQKQQQ--SGNHPITVHCSAGAGRTGTFPCALSTVL
gi|126467| QFHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRGTGFVVIDAML
gi|2499753 QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCIYIVIDIML
gi|462550| QYHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRGTGTIYIVIDSML
gi|2499751 QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPEPILVHCSAGVGRGTGTFIAIDRLI
gi|1709906 D1 QFQFTAWPDHGVPYHATGLLSFIRRVKLSNP--PDAGPMVVHCSAGVGRGTGCFIVIDAML
gi|126471| QLHFTSWPDFGVPFTPIGMLKFLKKVKTLP--VHAGPIVVHCSAGVGRGTGTFIVIDAMM
gi|548626| QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCIYIVIDIML
gi|131570| QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PNAGPLVVHCSAGAGRTGCFIVIDIML
gi|2144715 QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPEPILVHCSAGVGRGTGTFIAIDRLI
      * ..  ** .*. * . . . . . ***** ** . . . . .

```

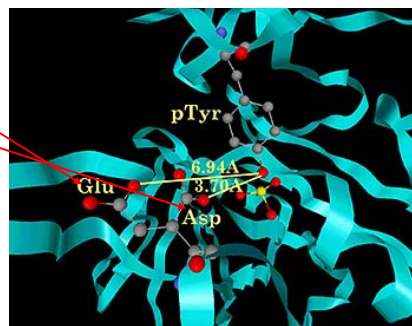
- Positions marked by “!” and “?” are likely places responsible for reduced PTP activity
 - All PTP D1 agree on them
 - All PTP D2 disagree on them

Key Mutation Site: PTP D1 vs D2

```

      ? ! ?
gi|00000|P D2 QFHFGWPEVGIPSDGK
gi|126467| QFHFTSWPDFGVPFTPI
gi|2499753 QFHFTGWPDHGVPYHAT
gi|462550| QYHYTQWPDMGVPEYAL
gi|2499751 QFHFTSWPDHGVPDTTD
gi|1709906 D1 QFQFTAWPDHGVPYHAT
gi|126471| QLHFTSWPDFGVPFTPI
gi|548626| QFHFTGWPDHGVPYHAT
gi|131570| QFHFTGWPDHGVPYHAT
gi|2144715 QFHFTSWPDHGVPDTTD
      * ..  ** .*. * . . . . .

```



- Positions marked by “!” are even more likely as 3D modeling predicts they induce large distortion to structure

Confirmation by Mutagenesis Expt

- **What wet experiments are needed to confirm the prediction?**
 - Mutate E → D in D2 and see if there is gain in PTP activity
 - Mutate D → E in D1 and see if there is loss in PTP activity

Exercise: Why do you need this 2-way expt?

Any Questions?



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



References

- S.E.Brenner. "Errors in genome annotation", *TIG*, 15:132--133, 1999
- D. Devos & A.Valencia. "Intrinsic errors in genome annotation", *TIG*, 17:429--431, 2001
- T.F.Smith & X.Zhang. "The challenges of genome sequence annotation or `The devil is in the details'", *Nature Biotech*, 15:1222--1223, 1997
- **C. Wu & W. Barker. "A Family Classification Approach to Functional Annotation of Proteins", *The Practical Bioinformatician*, Chapter 19, pages 401—416, WSPC, 2004**

Guilt by Association of Similarity of Dissimilarities



Image credit: www.comstock.com

PAKDD2007, Nanjing, 22 May 2007

42

Similarity of Dissimilarities



	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	...
apple ₂	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	..
...

PAKDD2007, Nanjing, 2007

Copyright 2007 © Limsoon Wong

SVM-Pairwise Framework

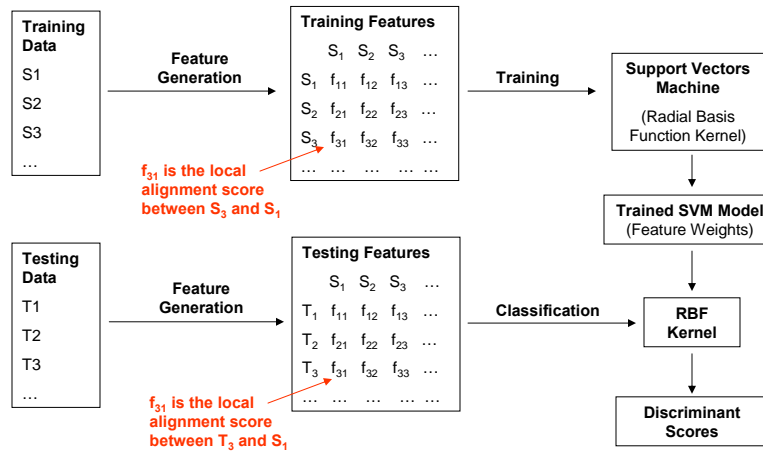
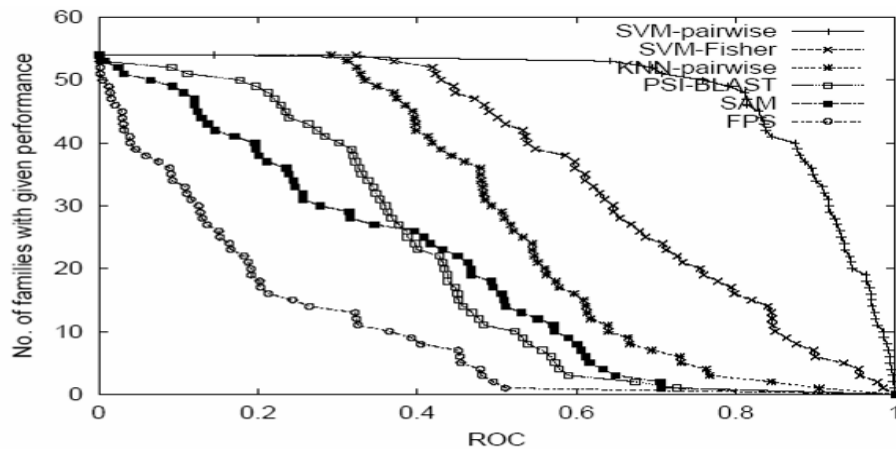


Image credit: Kenny Chua

PAKDD2007, Nanjing, 2007

Copyright 2007 © Limsoon Wong

Performance of SVM-Pairwise



- ROC: The area under the curve derived from plotting true positives as a function of false positives for various thresholds

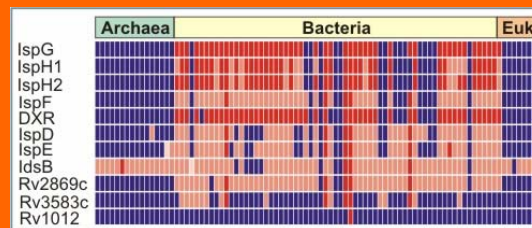
PAKDD2007, Nanjing, 2007

Copyright 2007 © Limsoon Wong

References

- Y.D. Cai & K.C. Chou. "Using functional domain composition to predict enzyme family classes". *J. Proteome Res.*, 4(1):109-111, 2005
- H.N. Chua & W.-K. Sung. "A better gap penalty for pairwise SVM". *Proc. APBC05*, pages 11-20
- T. Jaakkola, M. Diekhans, & D. Haussler. "A discriminative framework for detecting remote homologies". *JCB*, 7(1-2):95-11, 2000
- L. Liao & W.S. Noble. "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships". *JCB*, 10(6):857-868, 2003

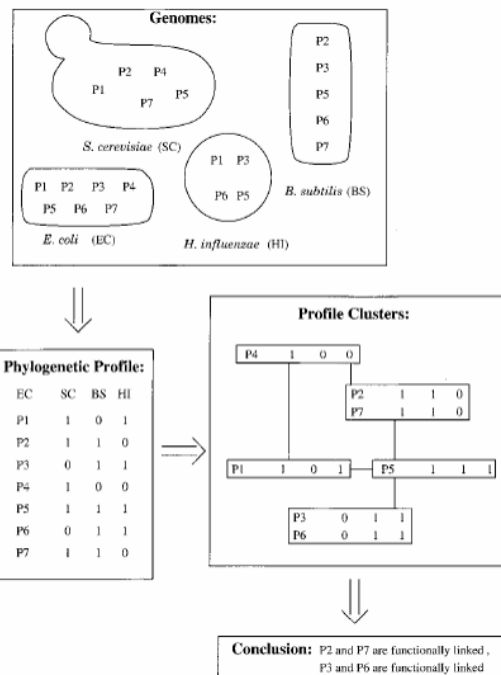
Guilt by Association of Genome Phylogenetic Profiles



Phylogenetic Profiling

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

- Gene (and hence proteins) with identical patterns of occurrence across phyla tend to function together
- ⇒ Even if no homolog with known function is available, it is still possible to infer function of a protein



Phylogenetic Profiling:
How It Works

Phylogenetic Profiling: P-value

The probability of observing by chance z occurrences of genes X and Y in a set of N lineages, given that X occurs in x lineages and Y in y lineages is

$$P(z|N, x, y) = \frac{w_z * \bar{w}_z}{W}$$

where

$$\begin{aligned}
 w_z &= \binom{N}{z} \\
 \bar{w}_z &= \binom{N-z}{x-z} * \binom{N-z}{y-z} \\
 W &= \binom{N}{x} * \binom{N}{y}
 \end{aligned}$$

No. of ways to distribute z co-occurrences over N lineage's
 No. of ways to distribute the remaining $x-z$ and $y-z$ occurrences over the remaining $N-z$ lineage's
 No. of ways of distributing X and Y over N lineage's without restriction

Phylogenetic Profiles: Evidence

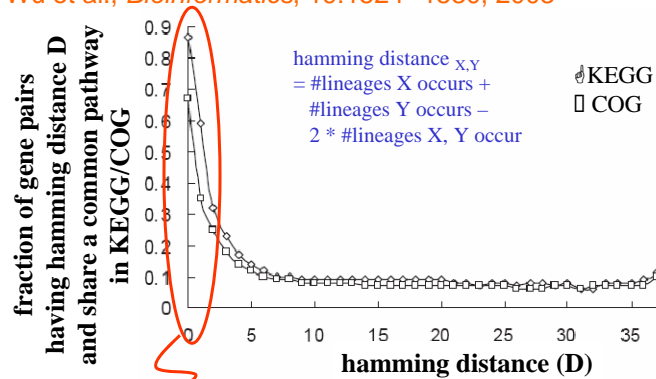
Pellegrini et al., *PNAS*, 96:4285--4288, 1999

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**

Phylogenetic Profiling: Evidence

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



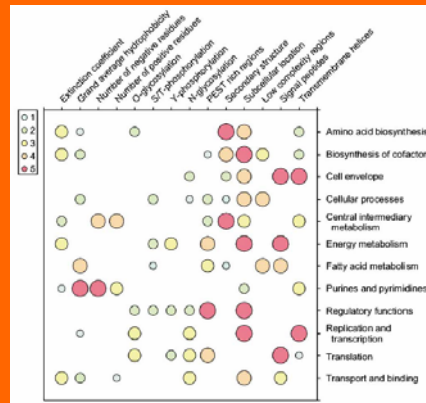
- Proteins having low hamming distance (thus highly similar phylogenetic profiles) tend to share common pathways
- Exercise: Why do proteins having high hamming distance also have this behaviour?

References



- M. Pellegrini et al. "Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles", *PNAS*, 96:4285--4288, 1999
- J. Wu et al. "Identification of functional links between genes using phylogenetic profiles", *Bioinformatics*, 19:1524--1530, 2003

Guilt by Association of Physico-Chemico Properties



PAKDD2007, Nanjing, 22 May 2007

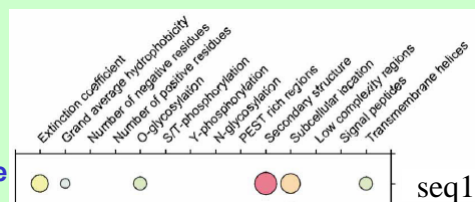
54

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, phosphorylation, signal peptide cleavage, ...
- These have associated consensus motifs, patterns, etc.

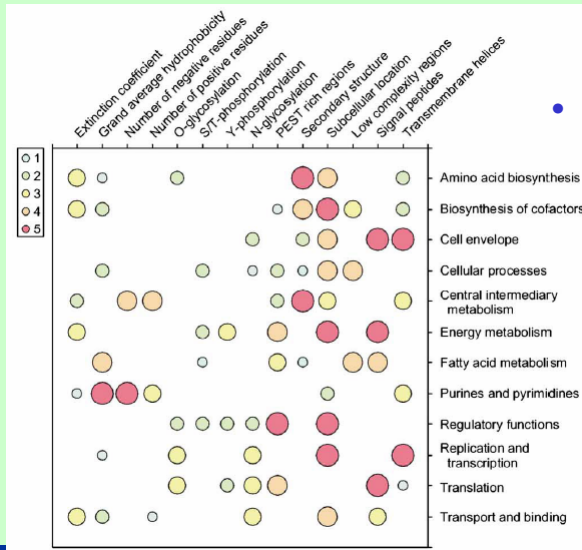


- Proteins performing similar functions should share some such "features"
- ⇒ Perhaps we can predict protein function by comparing its "feature" profile with other proteins?

PAKDD2007, Nanjing, 2007

Copyright 2007 © Limsoon Wong

ProtFun: Evidence



• Combinations of “features” seem to characterize some functional categories

ProtFun: How it Works

Abbreviation	Encoding	Description
ec	single value	Extinction coefficient predicted by ExPASy ProtParam
grav	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 phosphorylations predicted by NetPhos
phosY	potential in 10 bins	Tyrosine phosphorylations predicted by NetPhos
psipred	helix, sheet, coil in 5 bins	Predicted secondary structure from PSI-Pred
psort	20 probabilities	Subcellular location predictions 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	grav psipred psort
	30	oglyc psipred psort

Average the output of the 5 component ANNs

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

References

- L. Han et al. "Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity", *Proteomics*, 6(14):4023-4037, 2006
- L.J.Jensen et al. "Prediction of human protein function from post-translational modifications and localization features", *JMB*, 319:1257--1265, 2002

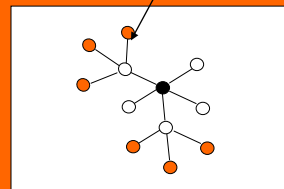
Any question?
Anyone needs a break?



PAKDD2007, Nanjing, 22 May 2007

Guilt by Association of
Common Interaction Partners:
Protein Function Prediction
from Protein Interactions

Level-2 neighbour



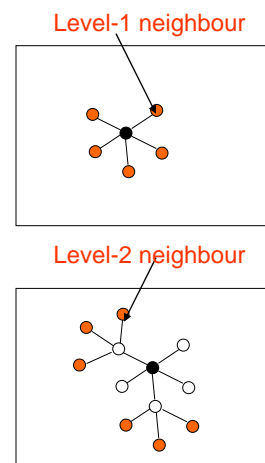
PAKDD2007, Nanjing, 22 May 2007

Protein Interaction Based Approaches

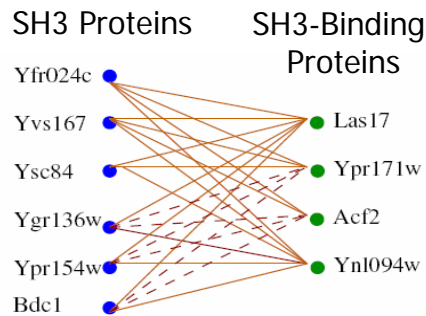
- **Neighbour counting** (Schwikowski et al, 2000)
 - Rank function based on freq in interaction partners
- **Chi-square** (Hishigaki et al, 2001)
 - Chi square statistics using expected freq of functions in interaction partners
- **Markov Random Fields** (Deng et al, 2003; Letovsky et al, 2003)
 - Belief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



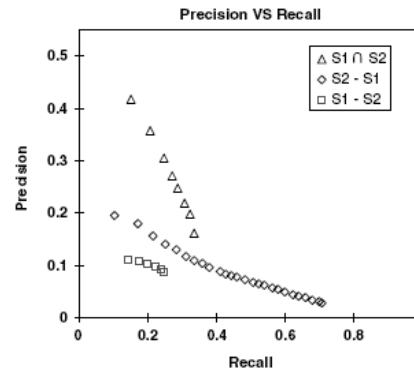
- **Protein interaction data from General Repository for Interaction Datasets (GRID)**
 - Data from published large-scale interaction datasets and curated interactions from literature
 - 13,830 unique and 21,839 total interactions
 - Includes most interactions from the Biomolecular Interaction Network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

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

⇒ Similarity can be defined as

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

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

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

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**

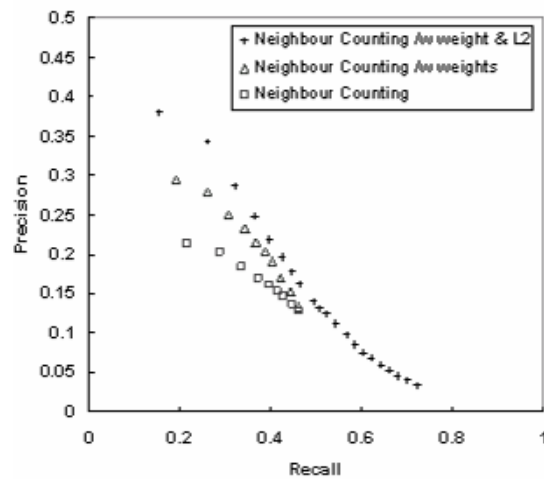
$$S(u, v) = \frac{2X}{2X + Y} \times \frac{2X}{2X + Z}$$

Integrating Reliability

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

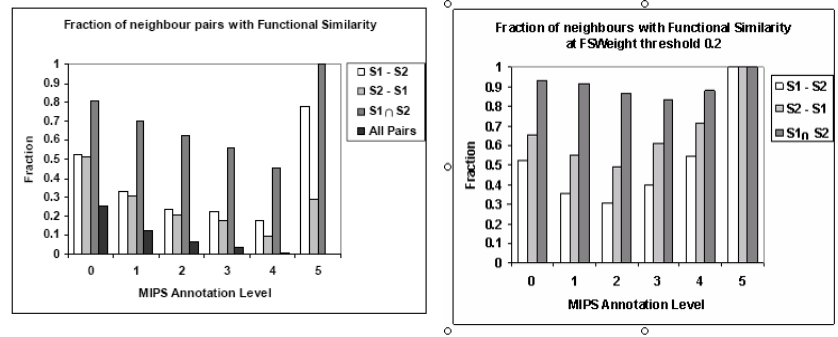
Neighbours	CD-Distance	FS-Weight	FS-Weight R
S_1	0.471810	0.498745	0.532596
S_2	0.224705	0.298843	0.375317
$S_1 \cup S_2$	0.224581	0.29629	0.363025

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

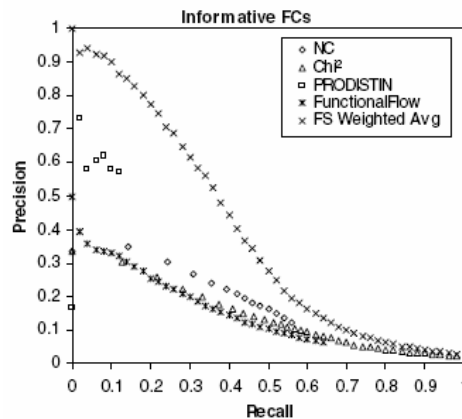
$$f_x(u) = \frac{1}{Z} \left[\lambda r_{int} \pi_x + \sum_{v \in N_u} \left(S_{TR}(u, v) \delta(v, x) + \sum_{w \in N_v} S_{TR}(u, w) \delta(w, x) \right) \right]$$

- 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

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

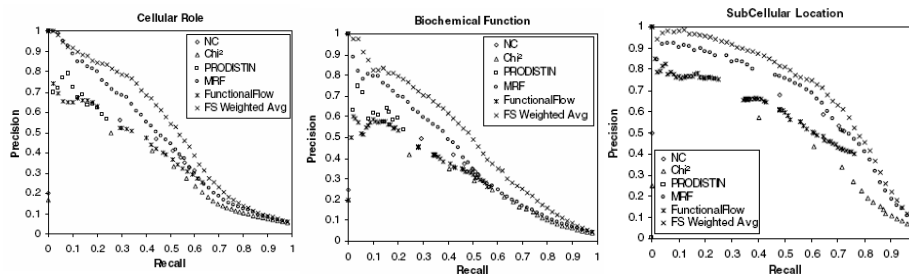
Performance of FS-Weighted Averaging

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



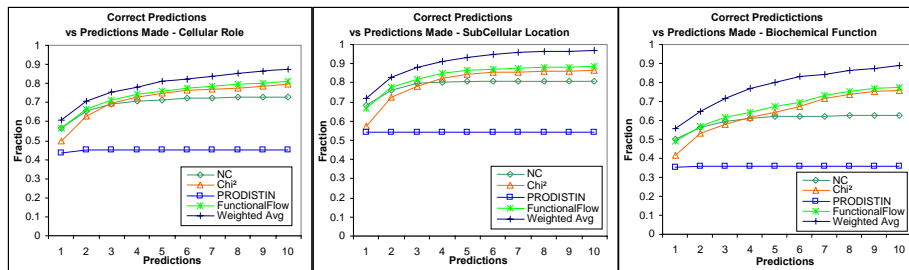
Performance of FS-Weighted Averaging

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



Performance of FS-Weighted Averaging

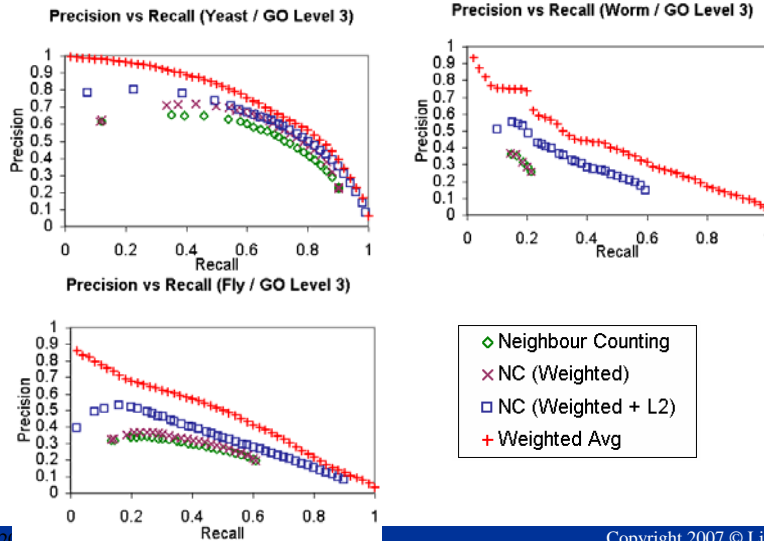
- Correct Predictions made on at least 1 function vs Number of predictions made per protein



Freq of Indirect Functional Association in Other Genomes

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

Effectiveness of FS Weighted Averaging in Other Genomes



PAKDD2

Copyright 2007 © Limsoon Wong

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

PAKDD2007, Nanjing, 2007

Copyright 2007 © Limsoon Wong



References

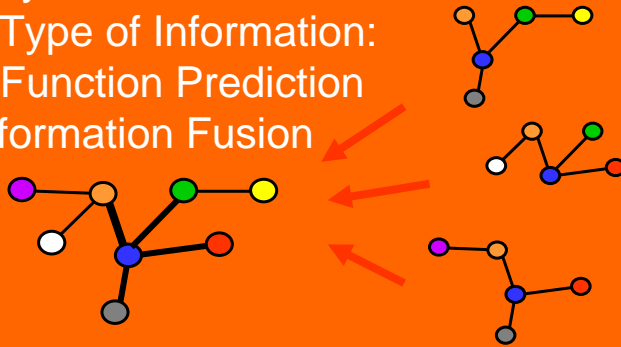
- C. Brun et al. "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network". *Genome Biol.* 5(1):R6, 2003
- **Chua H.N., Sung W.K., & Wong L. "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions". *Bioinformatics*, 22:1623-1630.**
- M. Deng et al. "Prediction of protein function using protein-protein interaction data". *JCB*, 10(6):947-960, 2003
- H. Hishigaki et al. "Assessment of prediction accuracy of protein function from protein-protein interaction data", *Yeast*, 18(6):523-531, 2001
- G.R.G. Lanckriet et al. "Kernel-based data fusion and its application to protein function prediction in yeast". *Proc. PSB 2004*, pp.300-311.



References

- S.Letovsky & S. Kasif. "Predicting protein function from protein/protein interaction data: a probabilistic approach". *Bioinformatics*. 19(Suppl.1):i197-i204, 2003
- M.P. Samanta & S. Liang. "Predicting protein functions from redundancies in large-scale protein interaction networks". *PNAS*, 100(22):12579-83, 2003
- A. Vazquez et al. "Global protein function prediction from protein-protein interaction networks". *Nature Biotechnology*. 21(6):697-670, 2003
- X. Zhou, M.C. Kao, & W.H. Wong. "Transitive functional annotation by shortest-path analysis of gene expression data". *PNAS*, 99(20):12783-88, 2002

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



PAKDD2007, Nanjing, 22 May 2007

86



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

PAKDD2007, Nanjing, 2007

Copyright 2007 © Limsoon Wong

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 metrics**
 - E.g., E-Score from BLAST vs Pearson correlation between expression profiles

Motivation

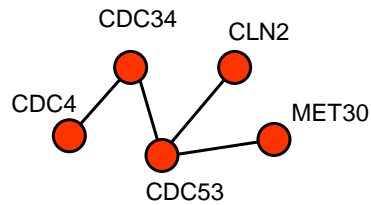
- **Problems:**
 - Complex models such as MRF and Kernel Fusion are computationally expensive
 - Difficult or not possible to identify contributing sources in a prediction
 - **Unified scoring of multiple sources has potential (Lee et al., *Science*, 2004)**
 - Simple scoring using Log Likelihood
 - Identified many functional clusters
- ⇒ **A simple, flexible, and effective way to integrate data sources that reports contributing sources in predictions to allow users to exercise judgment**

Strategy – Step 1

- Model a data source as undirected graph $G = \langle V, E \rangle$

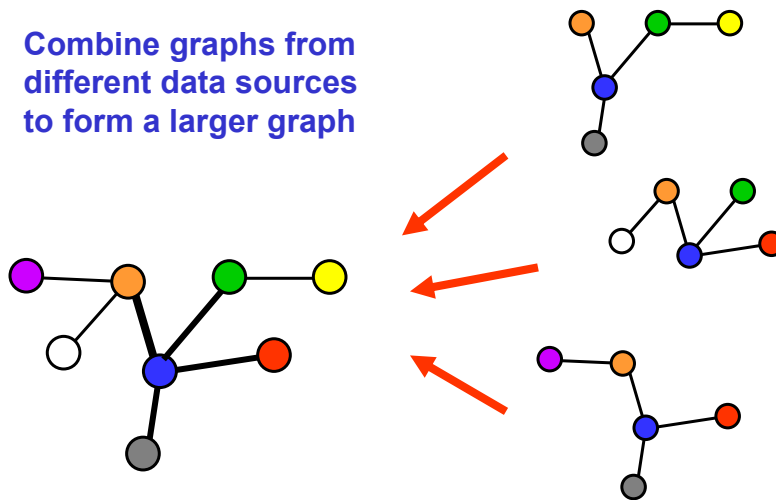
- V is a set of vertices; each vertex reps a protein

- E is a set of edges; each edge (u, v) reps a relationship (e.g. seq similarity, interaction) betw proteins u and v



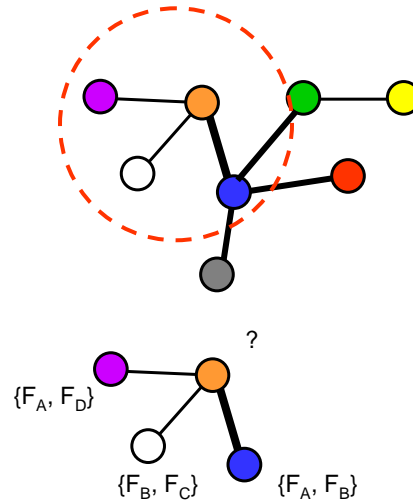
Strategy – Step 2

- 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

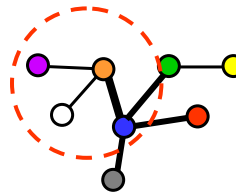
- **Scores may come in many forms**
 - E.g., Blast e-values, Pearson's correlation
- **A simple approach to discretization**
 - Split ranges into n equal intervals
 - Each interval becomes a new subtype
 - Assume linearity in range
 - Other strategies possible

Combination of Confidence

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

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

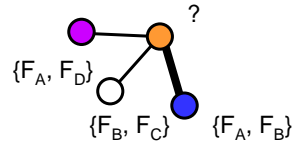
- $P(k.f)$ is confidence of edges of subtype k sharing function f
- $D_{u,v}$ is the set of subtypes of data sources which contains the edge (u,v)



Function Prediction

- **Weighted Average**

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

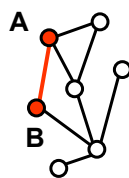


- $S_f(u)$ is score of function f for protein u
- $e_f(v)$ is 1 if protein v has function f , 0 otherwise
- N_u is set of neighbours of u
- $r_{u,v,f}$ is confidence of edge (u, v)

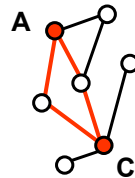
Level-2 Neighbours

- **Increase coverage of Protein-Protein interactions**

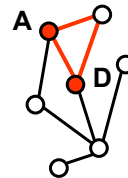
- Indirect function association (Chua et al. 2006)
- Topological weight applied to PPI
- Divide into 3 subtypes:



Level-1 Neighbours



Level-2 Neighbours



Level-1&2 Neighbours

- A threshold of 0.01 is applied on L2 neighbours to limit false positives

Topological Weight Applied to PPI: FS-Weighted Measure with Reliability



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

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

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

⇒ Rewriting

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

Comparison w/ Existing Approaches



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

Comparison w/ Existing Approaches

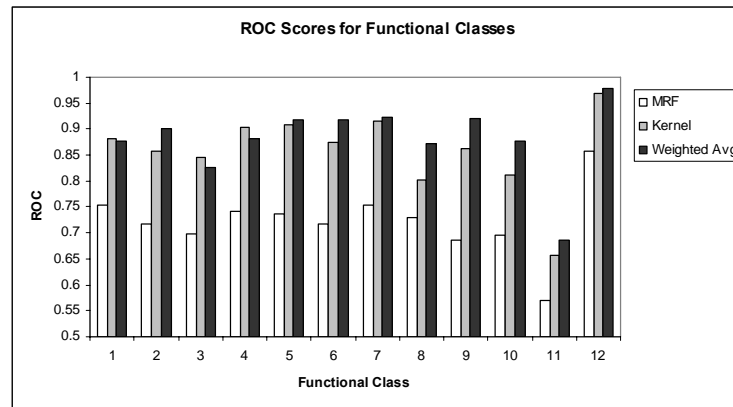
- **12 functional classes**

	Category	Size
1	Metabolism	1048
2	Energy	242
3	Cell cycle & DNA processing	600
4	Transcription	753
5	Protein synthesis	335
6	Protein fate	578
7	Cellular transport & transport mechanism	479
8	Cell rescue, defense & virulence	264
9	Interaction with the cellular environment	193
10	Cell fate	411
11	Control of cellular organization	192
12	Transport facilitation	306

Comparison w/ Existing Approaches

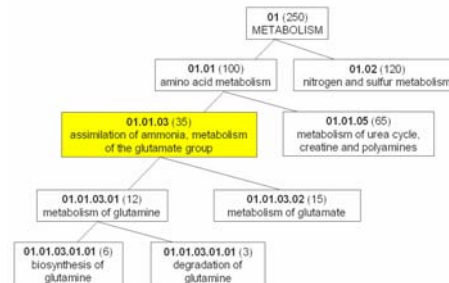
- **Validation Method (Lanckriet et al, 2004)**
 - Receiver Operating Characteristics (ROC)
 - True Positives vs False Positives
 - Area under ROC curve for each function
 - Averaged over 3 repetitions of 5-fold cross validation

Comparison w/ Existing Approaches



GO Terms Prediction for Yeast Proteins

- **Proteins from Saccharomyces Cerevisiae**
 - 5448 proteins from GO Annotation (SGD)
- **Functional Annotation**
 - Gene Ontology
 - Hierarchical
 - 3 Namespaces (molecular function, biological process, cellular component)



- **Informative GO Terms (for evaluation)**
 - Zhou et al. (2002)
 - FC associated with at least 30 proteins and no subclass associated with at least 30 proteins



Data Sources

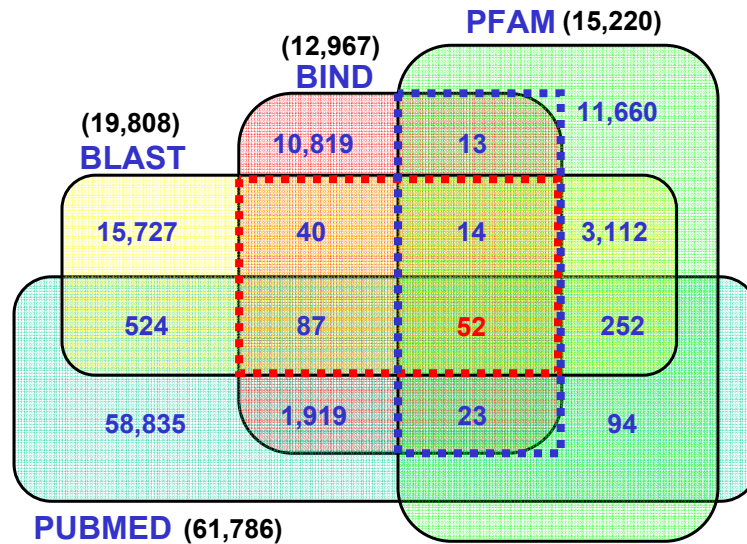
- **PPI**
 - BIND
 - 12,967 unique interactions betw yeast proteins
 - FS weight used as score
- **Protein Sequences**
 - Seqs from GO database (archive.godatabase.org)
 - Each yeast seq is aligned w/ rest using BLAST (cutoff E-Score = 1)
 - $-\log(\text{e-score})$ used as score
 - Top 5 results w/ known annotations
 - 19,808 unique pairs involving yeast proteins



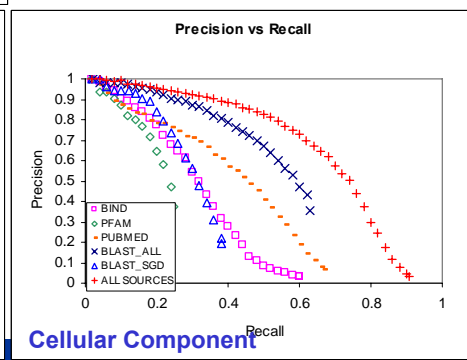
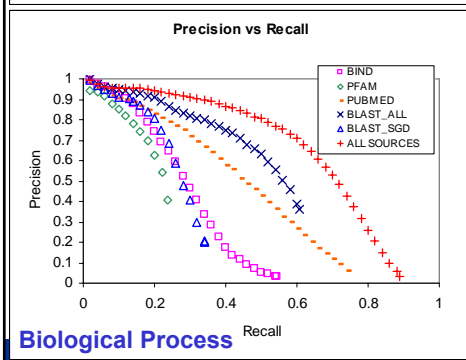
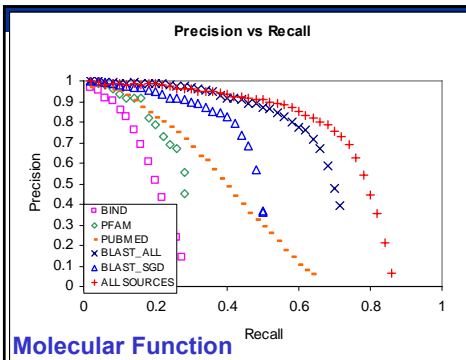
Data Sources

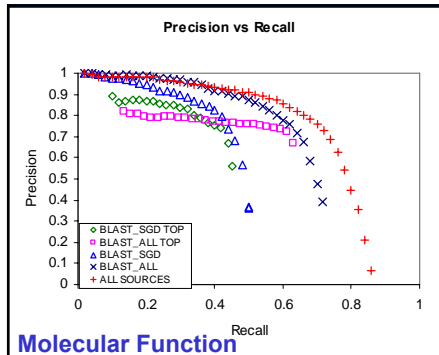
- **Pfam Domains**
 - SwissPfam database (<http://www.sanger.ac.uk/Software/Pfam/ftp.shtml>)
 - Precomputed Pfam domains for SwissProt and TrEMBL proteins w/ E-value threshold 0.01
 - Number of common domains used as score
 - 15,220 unique pairs involving yeast proteins
- **Pubmed Abstracts**
 - Pubmed abstracts obtained by searching protein's name and aliases on Pubmed
 - Limit to first 1000 abstracts returned
 - Fraction of abstracts w/ co-occurrence used as score
 - 61,786 unique pairs involving yeast proteins

Multiple Data Sources

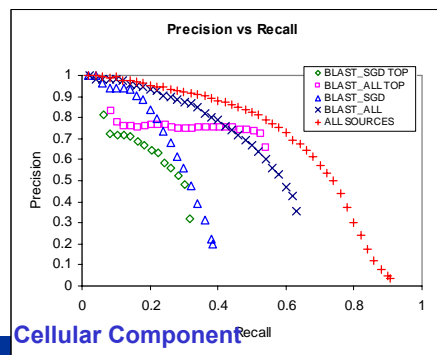
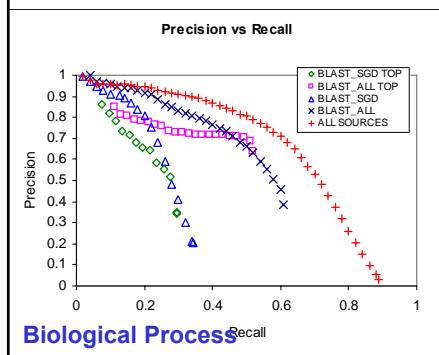


Combining all data sources outperforms any individual data source





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



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 ". In preparation, 2006
- M. Deng, T. Chen, & F. Sun. An integrated probabilistic model for functional prediction of proteins. *JCB*, 11(2-3):463-75, 2004.
- G.R. Lanckriet et al. "Kernel-based data fusion and its application to protein function prediction in yeast". *Proc. PSB 2004*, pp. 300-311.
- D.M. Martin, M. Berriman, G.J. Barton. "GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes". *BMC Bioinformatics*. 5:178, 2004
- G. Xiao, W. Pan. "Gene function prediction by a combined analysis of gene expression data and protein-protein interaction data". *JBCB*, 3(6):1371-89, 2005

Any Question?

