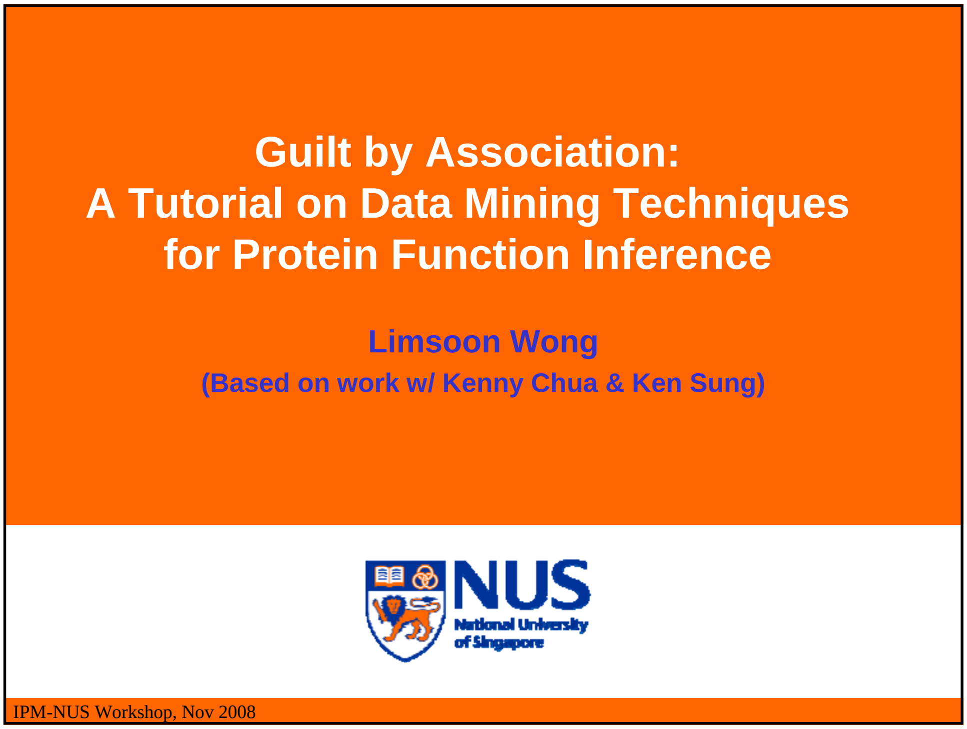
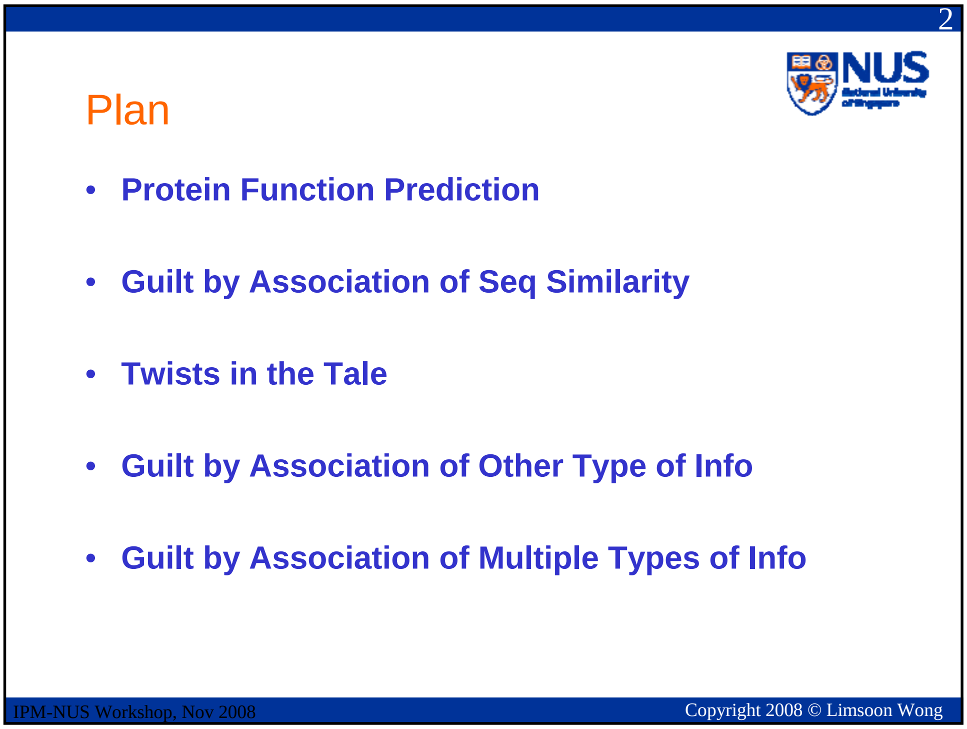


Guilt by Association: A Tutorial on Data Mining Techniques for Protein Function Inference

Limsoon Wong

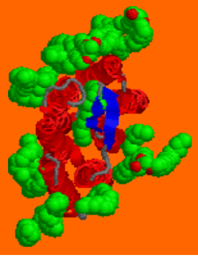
(Based on work w/ Kenny Chua & Ken Sung)



Plan

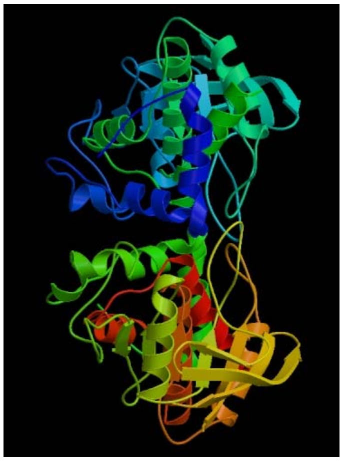
- **Protein Function Prediction**
- **Guilt by Association of Seq Similarity**
- **Twists in the Tale**
- **Guilt by Association of Other Type of Info**
- **Guilt by Association of Multiple Types of Info**

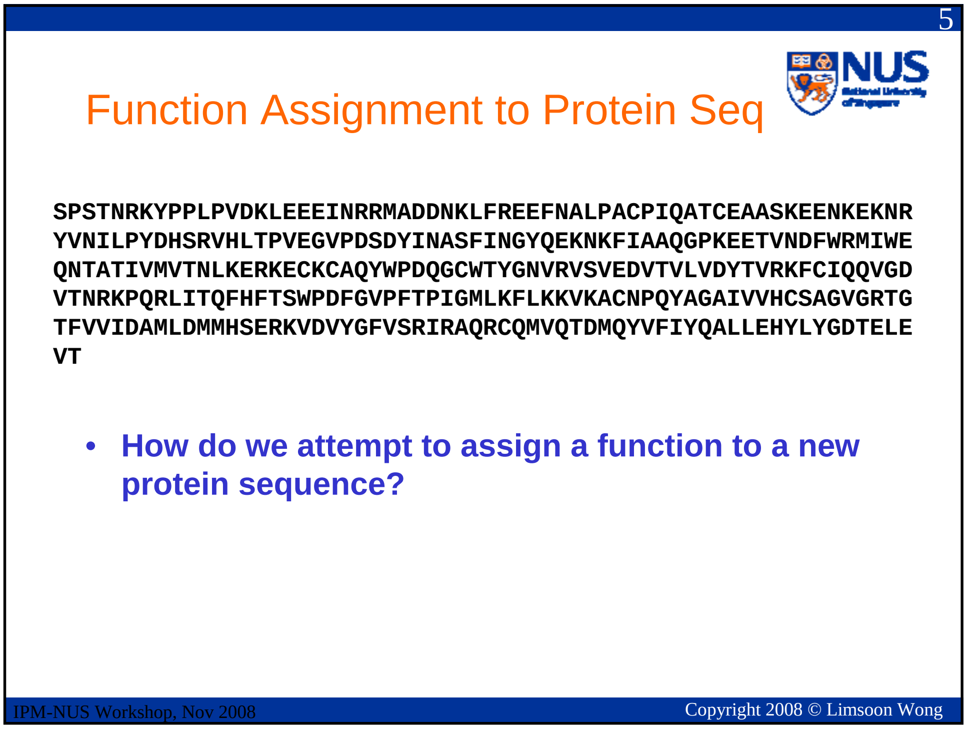
Protein Function Prediction: Motivation & Challenges



- A protein is a large complex molecule made up of one or more chains of amino acids

- Protein performs a wide variety of activities in the cell

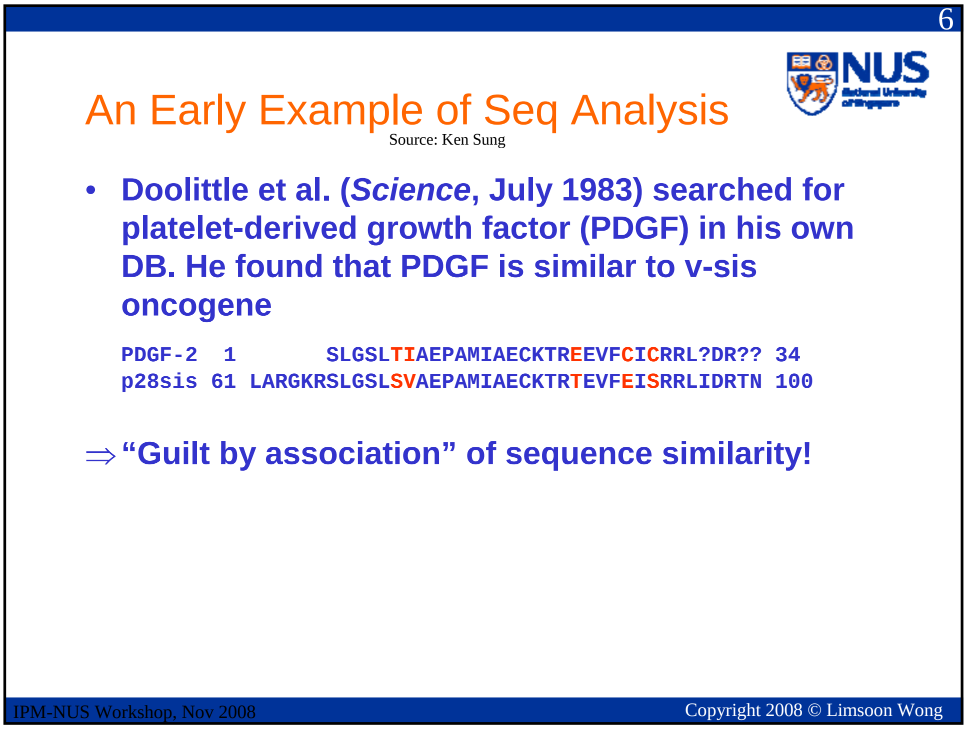




Function Assignment to Protein Seq

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

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



An Early Example of Seq Analysis

Source: Ken Sung

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

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

⇒ “Guilt by association” of sequence similarity!

Guilt by Association of Sequence Similarity

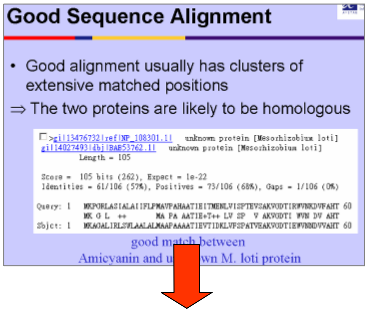
```
PDGF-2  1       SLGSLTIAEPAMIAECKTREEVFCICRRL?DR?? 34
p28sis 61 LARGKRSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTN 100
```

Guilt by Association: General Idea

- **Compare the target sequence T with sequences S_1, …, S_n of known function in a database**
- **Determine which ones amongst S_1, …, 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**

Guilt by Association of Seq Similarity

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



Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

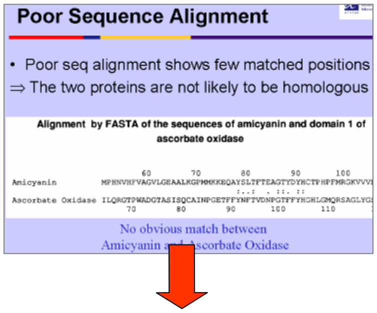
⇒ The two proteins are likely to be homologous

```
□ >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   MKPORLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
            MK G L  ++       MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

good match between Amicyanin and unknown M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments



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              MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVV
                                                   :..:  . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYG
                          70        80        90       100       110
```

No obvious match between Amicyanin and Ascorbate Oxidase

Discard this function as a candidate

Seq Alignment

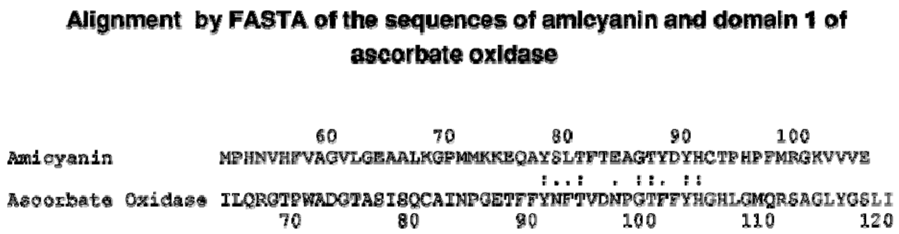
```
PDGF-2  1       SLGSLTIAEPAMIAECKTREEVFCICRRL?DR?? 34
p28sis 61 LARGKRSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTN 100
```

- **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

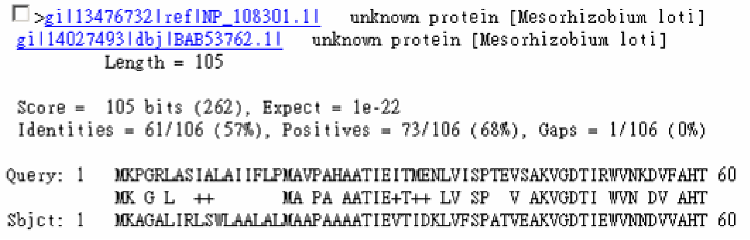
```
                          60        70        80        90       100
Amicyanin         MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVVE
                                             :..:  . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYGSLI
                        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



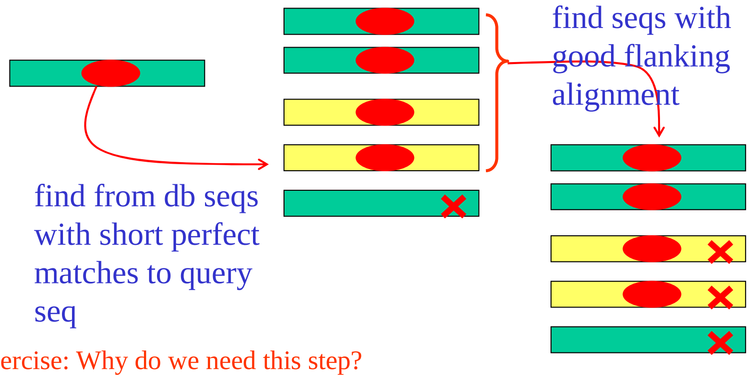
```
□ >gi|13476732|ref|NP_108301.1|   unknown protein [Mesorhizobium loti]
 gi|14027493|dbj|BAB53762.1|   unknown protein [Mesorhizobium loti]
          Length = 105

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

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

good match between
Amicyanin and unknown M. loti protein

BLAST: How It Works

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

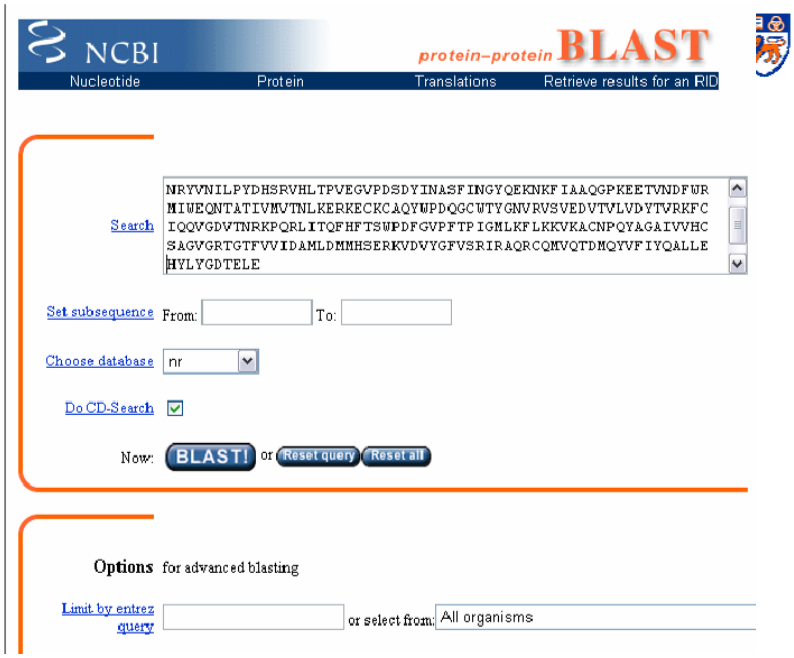
- **BLAST is the most popular tool for “guilt by association” seq homology search**



find from db seqs with short perfect matches to query seq

find seqs with good flanking alignment

Exercise: Why do we need this step?



Search

```
NRYVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWR
MIWEQNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFC
IQQVGDVTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHC
SAGVGRTGTFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLE
HYLYGDTELE
```

Set subsequence From: To:

Choose database nr

Do CD-Search ☑

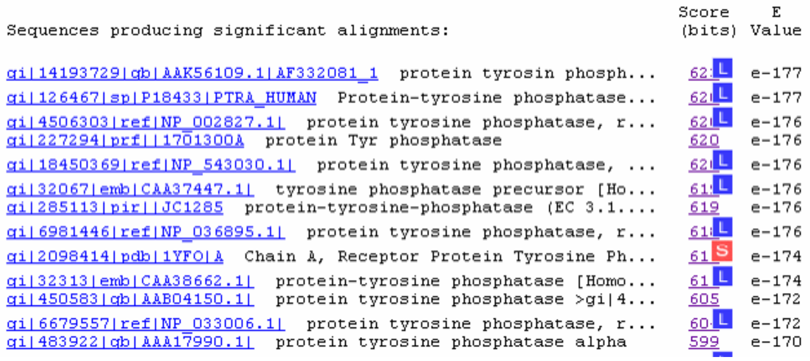
Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query

or select from: All organisms

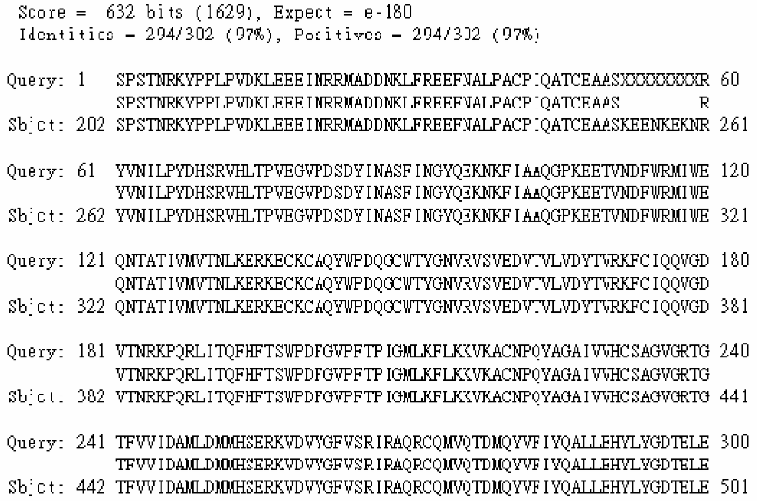
Homologs by BLAST



```
                                                                   Score    E
Sequences producing significant alignments:                        (bits) Value

gi|14193729|gb|AAK56109.1|AF332081_1  protein tyrosin phosph...    62 L   e-177
gi|126467|sp|P18433|PTRA_HUMAN  Protein-tyrosine phosphatase...    62 L   e-177
gi|4506303|ref|NP_002827.1|  protein tyrosine phosphatase, r...    62 L   e-176
gi|227294|prf||1701300A  protein Tyr phosphatase                   620    e-176
gi|18450369|ref|NP_543030.1|  protein tyrosine phosphatase, ...    62 L   e-176
gi|32067|emb|CAA37447.1|  tyrosine phosphatase precursor [Ho...    61 L   e-176
gi|285113|pir||JC1285  protein-tyrosine-phosphatase (EC 3.1....    619    e-176
gi|6981446|ref|NP_036895.1|  protein tyrosine phosphatase, r...    61 L   e-176
gi|2098414|pdb|1YFO|A  Chain A, Receptor Protein Tyrosine Ph...    61 S   e-174
gi|32313|emb|CAA38662.1|  protein-tyrosine phosphatase [Homo...    61 L   e-174
gi|450583|gb|AAB04150.1|  protein tyrosine phosphatase >gi|4...    605    e-172
gi|6679557|ref|NP_033006.1|  protein tyrosine phosphatase, r...    60 L   e-172
gi|483922|gb|AAA17990.1|  protein tyrosine phosphatase alpha       599    e-170
```

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

Example Alignment with PTPα



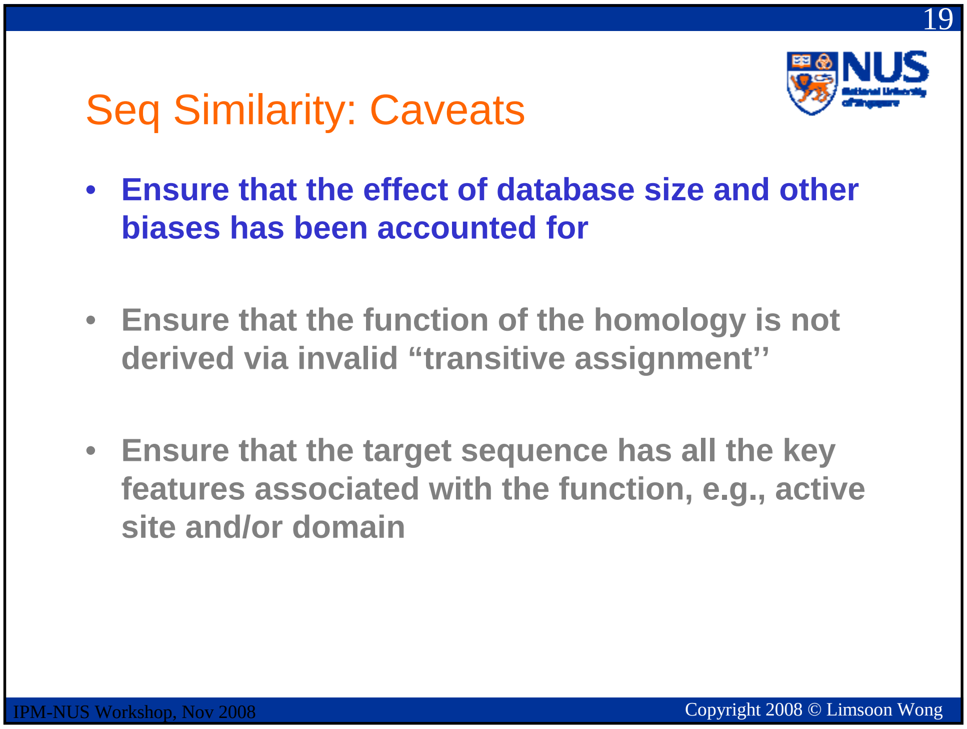
```
 Score =  632 bits (1629), Expect = e-180
 Identities = 294/302 (97%), Positives = 294/302 (97%)

Query: 1   SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASXXXXXXXXR 60
            SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAAS        R
Sbjct: 202 SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR 261

Query: 61  YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE 120
            YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
Sbjct: 262 YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE 321

Query: 121 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD 180
            QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
Sbjct: 322 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD 381

Query: 181 VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG 240
            VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
Sbjct: 382 VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG 441

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

References

- S.F.Altshcul 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

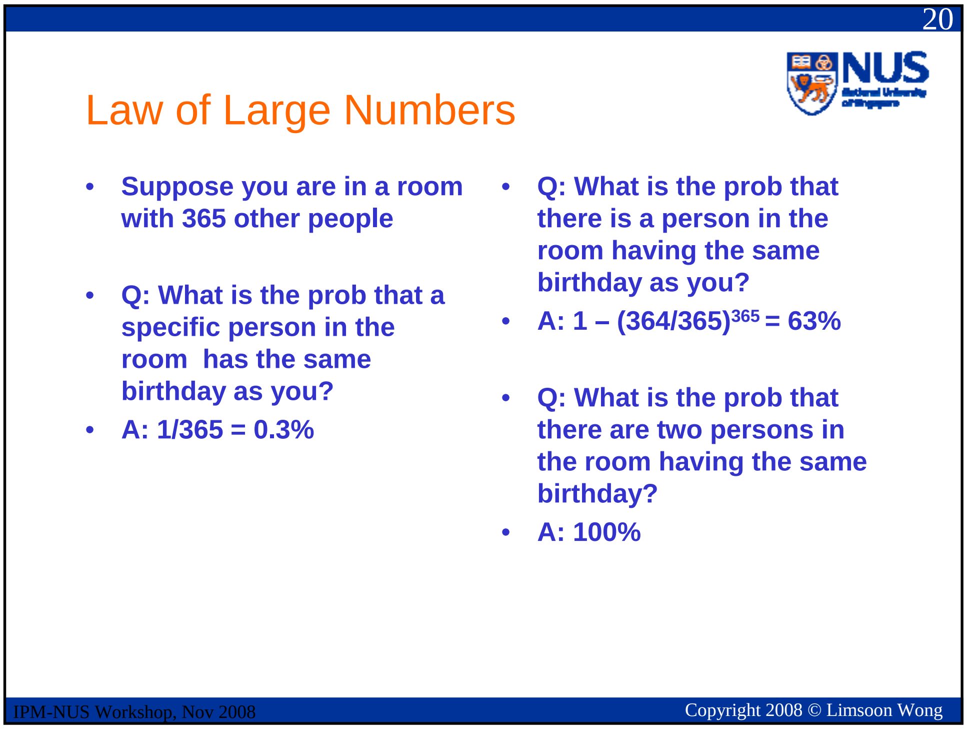
Twists in the Tale of Guilt by Association of Seq Similarity

Image credit: Shanti Christensen, http:// static.flickr.com/46/148437681_7f2dfa977e_m.jpg



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 Virgina, 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

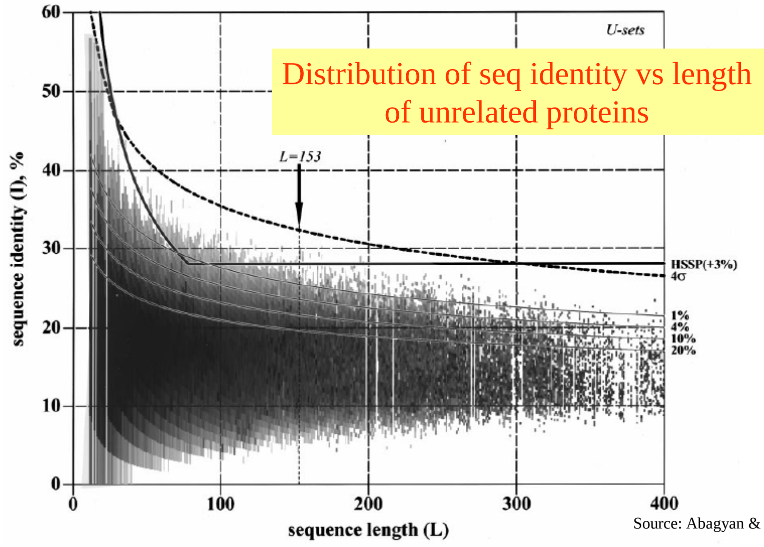
- One fourth of all residues in protein seqs occur in regions with biased amino acid composition
- Alignments of two such regions achieves high score purely due to segment composition

⇒ 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



Distribution of seq identity vs length of unrelated proteins

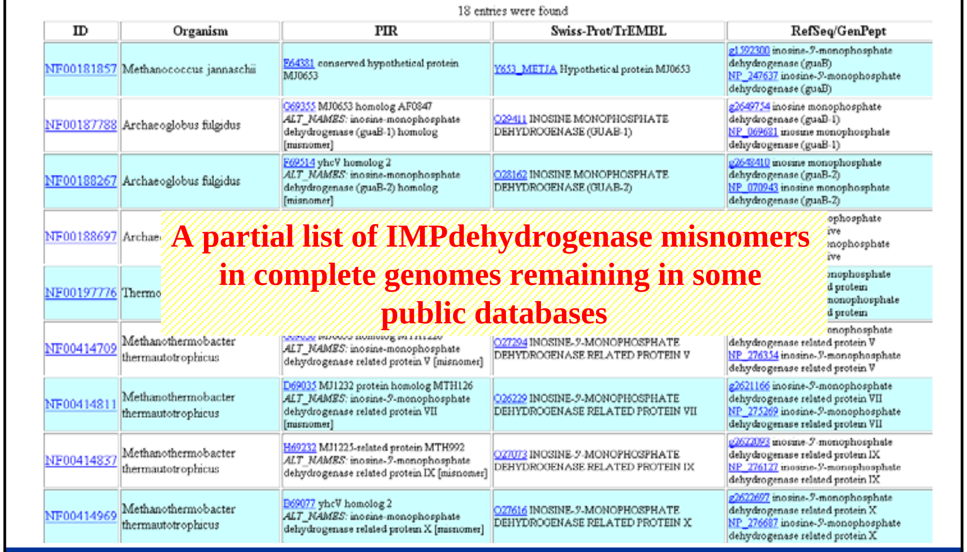
Source: Abagyan & Batalov

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)

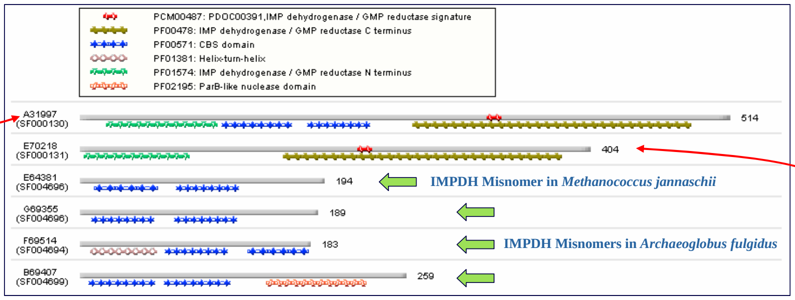


18 entries were found

ID	Organism	PIR	Swiss-Prot/TrEMBL	RefSeq/GenPept
NF00181857	Methanococcus jannaschii	E64381 conserved hypothetical protein MJ0653	Y653_METJA Hypothetical protein MJ0653	g1592300 inosine-5'-monophosphate dehydrogenase (guaB) NP_247637 inosine-5'-monophosphate dehydrogenase (guaB)
NF00187788	Archaeoglobus fulgidus	G69355 MJ0653 homolog AF0847 *ALT_NAMES:* inosine-monophosphate dehydrogenase (guaB-1) homolog [misnomer]	O29411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)	g2649754 inosine monophosphate dehydrogenase (guaB-1) NP_069681 inosine monophosphate dehydrogenase (guaB-1)
NF00188267	Archaeoglobus fulgidus	F69514 yhcV homolog 2 *ALT_NAMES:* inosine-monophosphate dehydrogenase (guaB-2) homolog [misnomer]	O28162 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-2)	g2648410 inosine monophosphate dehydrogenase (guaB-2) NP_070943 inosine monophosphate dehydrogenase (guaB-2)
NF00188697	Archae			ophosphate ive nophosphate ive
NF00197776	Thermo			nophosphate d protein monophosphate d protein
NF00414709	Methanothermobacter thermautotrophicus	*ALT_NAMES:* inosine-monophosphate dehydrogenase related protein V [misnomer]	O27294 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN V	onophosphate dehydrogenase related protein V NP_276354 inosine-5'-monophosphate dehydrogenase related protein V
NF00414811	Methanothermobacter thermautotrophicus	D69035 MJ1232 protein homolog MTH126 *ALT_NAMES:* inosine-5'-monophosphate dehydrogenase related protein VII [misnomer]	O26229 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN VII	g2621166 inosine-5'-monophosphate dehydrogenase related protein VII NP_275269 inosine-5'-monophosphate dehydrogenase related protein VII
NF00414837	Methanothermobacter thermautotrophicus	H69232 MJ1225-related protein MTH992 *ALT_NAMES:* inosine-5'-monophosphate dehydrogenase related protein IX [misnomer]	O27073 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN IX	g2622093 inosine-5'-monophosphate dehydrogenase related protein IX NP_276127 inosine-5'-monophosphate dehydrogenase related protein IX
NF00414969	Methanothermobacter thermautotrophicus	B69077 yhcV homolog 2 *ALT_NAMES:* inosine-monophosphate dehydrogenase related protein X [misnomer]	O27616 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN X	g2622697 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

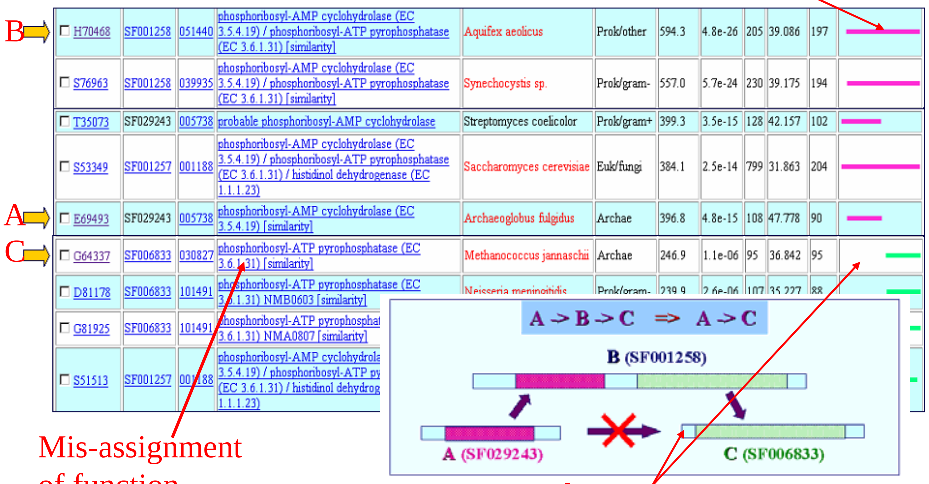
IMPDH Domain Structure


PCM00487: PDOC00391,IMP dehydrogenase / GMP reductase signature
PF00478: IMP dehydrogenase / GMP reductase C terminus
PF00571: CBS domain
PF01381: Helix-turn-helix
PF01574: IMP dehydrogenase / GMP reductase N terminus
PF02195: ParB-like nuclease domain
A31997 (SF000130)
514
E70218 (SF000131)
404
E64381 (SF004696)
194
IMPDH Misnomer in Methanococcus jannaschii
G69355 (SF004696)
189
F69514 (SF004694)
183
IMPDH Misnomers in Archaeoglobus fulgidus
B69407 (SF004699)
259

- **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



B	□ 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
A	□ E69493	SF029243	005738	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) [similarity]	Archaeoglobus fulgidus	Archae	396.8	4.8e-15	108	47.778	90
C	□ 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]	Neisseria meningitidis	Prok/gram-	239.9	2.6e-06	107	35.227	88
	□ G81925	SF006833	101491	phosphoribosyl-ATP pyrophosphat 3.6.1.31) NMA0807 [similarity]							
	□ S51513	SF001257	001188	phosphoribosyl-AMP cyclohydrola 3.5.4.19) / phosphoribosyl-ATP py (EC 3.6.1.31) / histidinol dehydrog 1.1.1.23)							

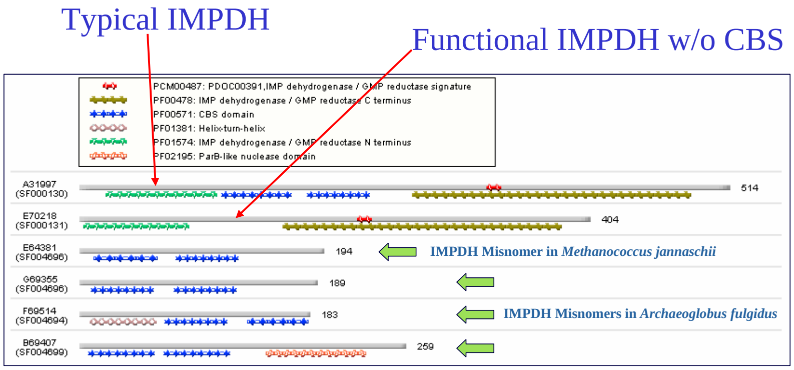
Mis-assignment of function

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
PCM00487: PDOC00391,IMP dehydrogenase / GMP reductase signature
PF00478: IMP dehydrogenase / GMP reductase C terminus
PF00571: CBS domain
PF01381: Helix-turn-helix
PF01574: IMP dehydrogenase / GMP reductase N terminus
PF02195: ParB-like nuclease domain
A31997 (SF000130)
514
E70218 (SF000131)
404
E64381 (SF004696)
194
IMPDH Misnomer in *Methanococcus jannaschii*
G69355 (SF004696)
189
F69514 (SF004694)
183
IMPDH Misnomers in *Archaeoglobus fulgidus*
B69407 (SF004699)
259

- **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 …

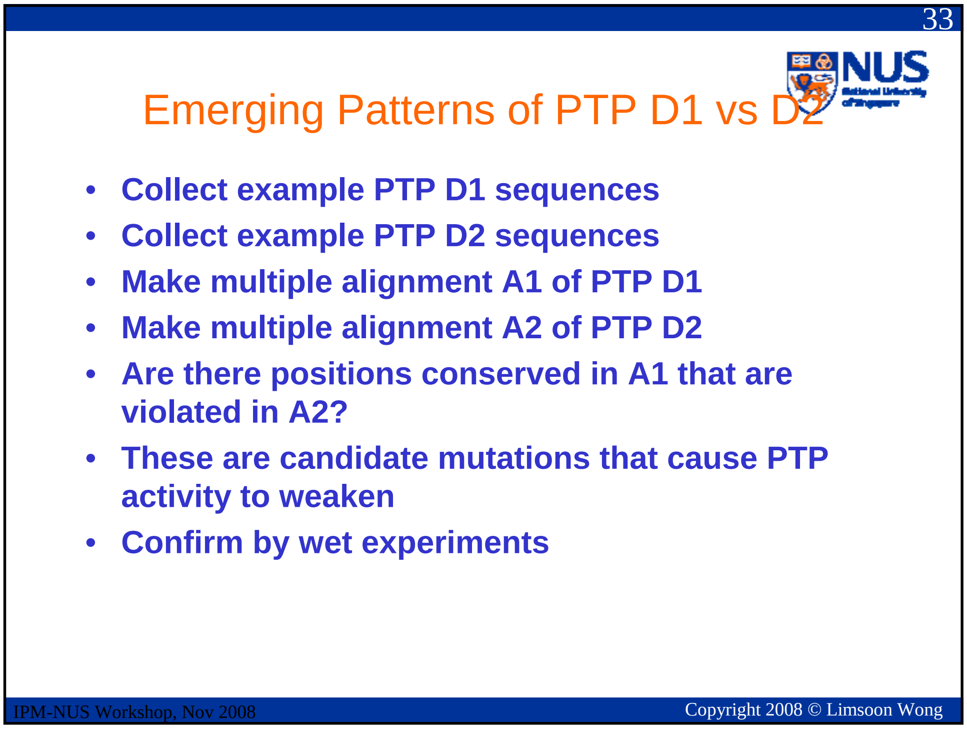
Identifying Key Mutation Sites

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

Sequence from a typical PTP domain D2

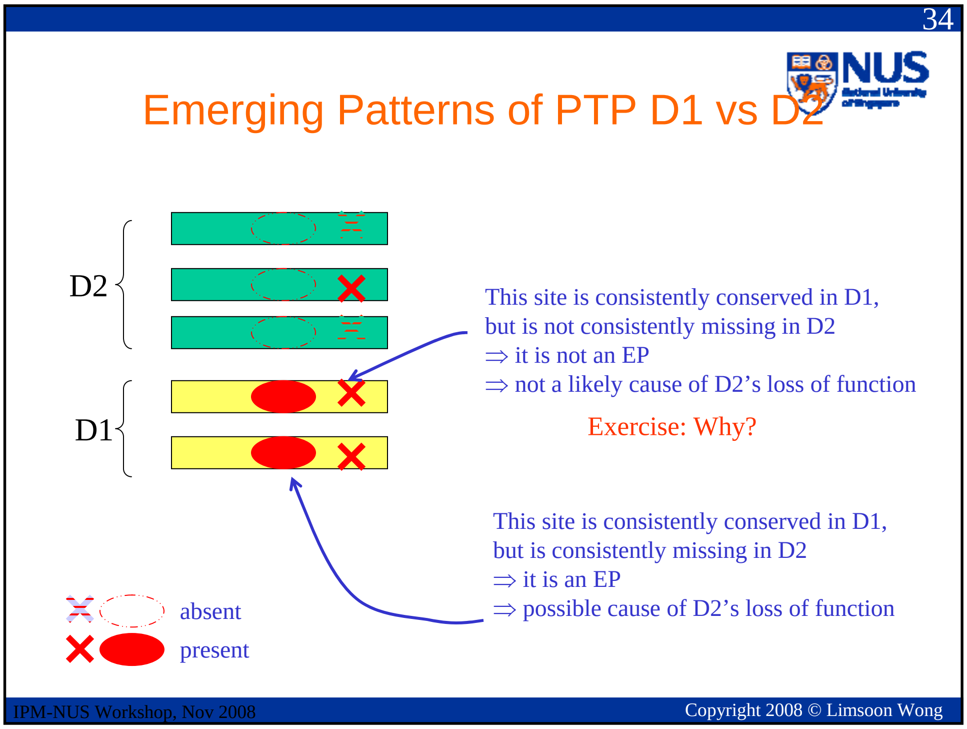
```
>gi|00000|PTPA-D2
EEEFKKLTSIKIQNDKMRTGNLPANMKKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASF
IDGYRQKDSYIASQGPLLHTIEDFWRMIWEWKSCSIVMLTELEERGQEKCAQYWPSDGLV
SYGDITVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFHGWPEVGIPSDGKGMISII
AAVQKQQQQSGNHPITVHCSAGAGRTGTFCALSTVLERVKAEGILDVFQTVKSLRLQRPH
MVQTLEQYEFCYKVVQEYIDAFSDYANFK
```

- Some PTPs have 2 PTP domains
- PTP domain D1 is 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

This site is consistently conserved in D1, but is not consistently missing in D2
⇒ it is not an EP
⇒ not a likely cause of D2’s loss of function

Exercise: Why?

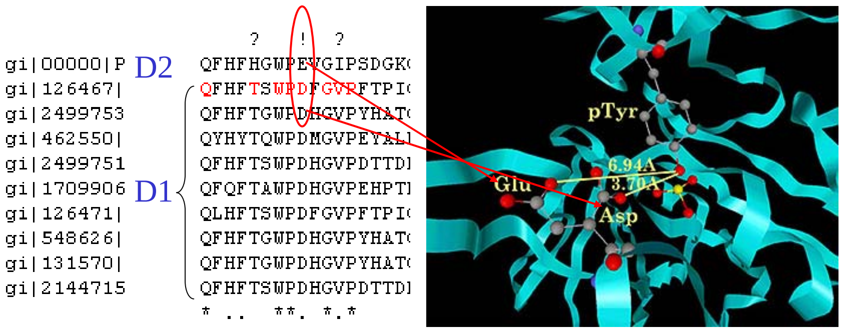
This site is consistently conserved in D1, but is consistently missing in D2
⇒ it is an EP
⇒ possible cause of D2’s loss of function

Key Mutation Site: PTP D1 vs D2

```
                     ?   !  ?         ?                 ?             ? ??
gi|00000|P  D2   QFHFHGWPEVGIPSDGKGMISIIAAVQKQQQQ-SGNHPITVHCSAGAGRTGTFCALSTVL
gi|126467|      ⎧QFHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAML
gi|2499753      ⎪QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIML
gi|462550|      ⎪QYHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSML
gi|2499751      ⎪QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLI
gi|1709906  D1  ⎨QFQFTAWPDHGVPEHPTPFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAML
gi|126471|      ⎪QLHFTSWPDFGVPFTPIGMLKFLKKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDAMM
gi|548626|      ⎪QFHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIML
gi|131570|      ⎪QFHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPLVVHCSAGAGRTGCFIVIDIML
gi|2144715      ⎩QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLI
                 * ..  **. *.*      .                  . ****** ****..  .  ..
```

- **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   QFHFHGWPEVGIPSDGK
gi|126467|      ⎧QFHFTSWPDFGVPFTPI
gi|2499753      ⎪QFHFTGWPDHGVPYHAT
gi|462550|      ⎪QYHYTQWPDMGVPEYAL
gi|2499751      ⎪QFHFTSWPDHGVPDTTD
gi|1709906  D1  ⎨QFQFTAWPDHGVPEHPT
gi|126471|      ⎪QLHFTSWPDFGVPFTPI
gi|548626|      ⎪QFHFTGWPDHGVPYHAT
gi|131570|      ⎪QFHFTGWPDHGVPYHAT
gi|2144715      ⎩QFHFTSWPDHGVPDTTD
                 * ..  **. *.*
```

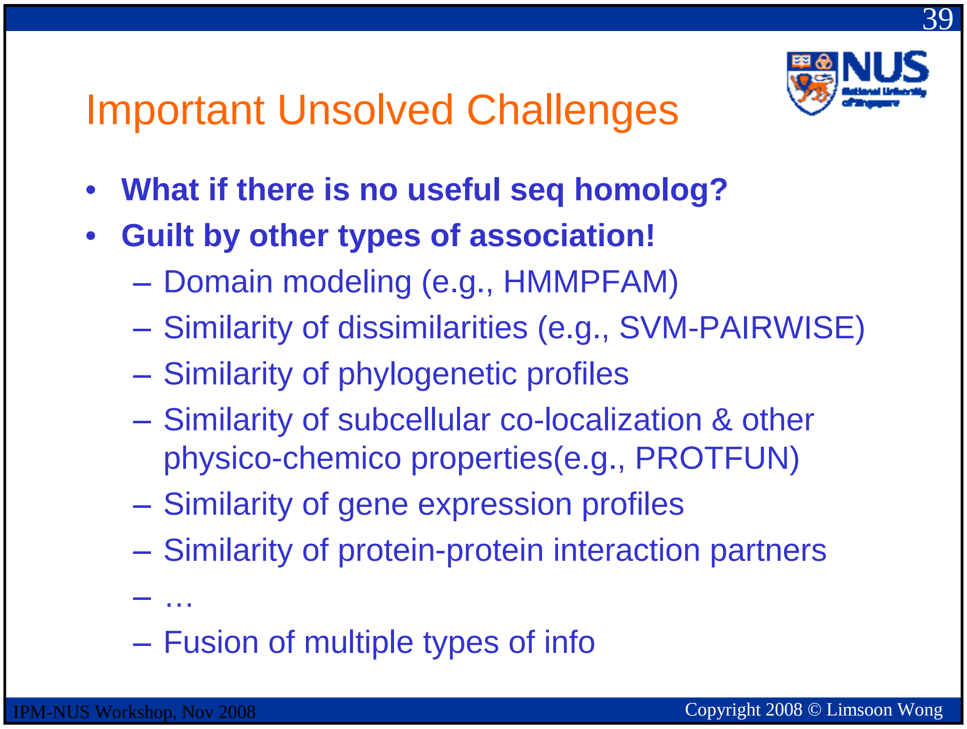
- **Positions marked by “!” are even more likely as 3D modeling predicts they induce large distortion to structure**

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

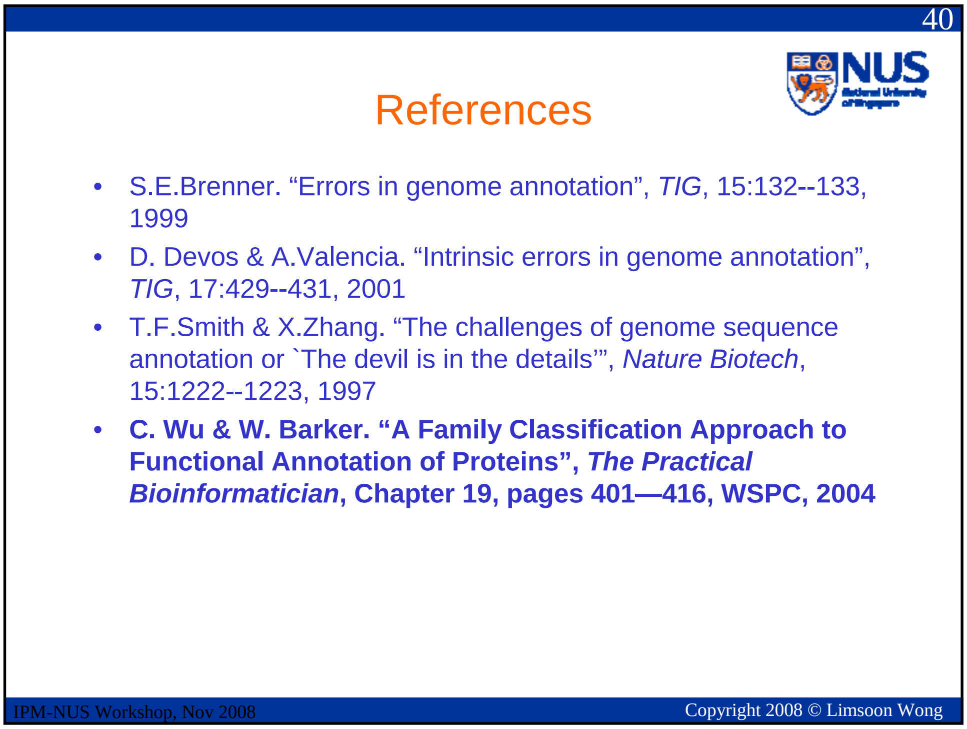
Exercise: Why do you need this 2-way expt?

Any Question?



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

Similarity of Dissimilarities

Differences of “unknown” to other fruits are same as “apple” to other fruits

“unknown” is an “apple”!

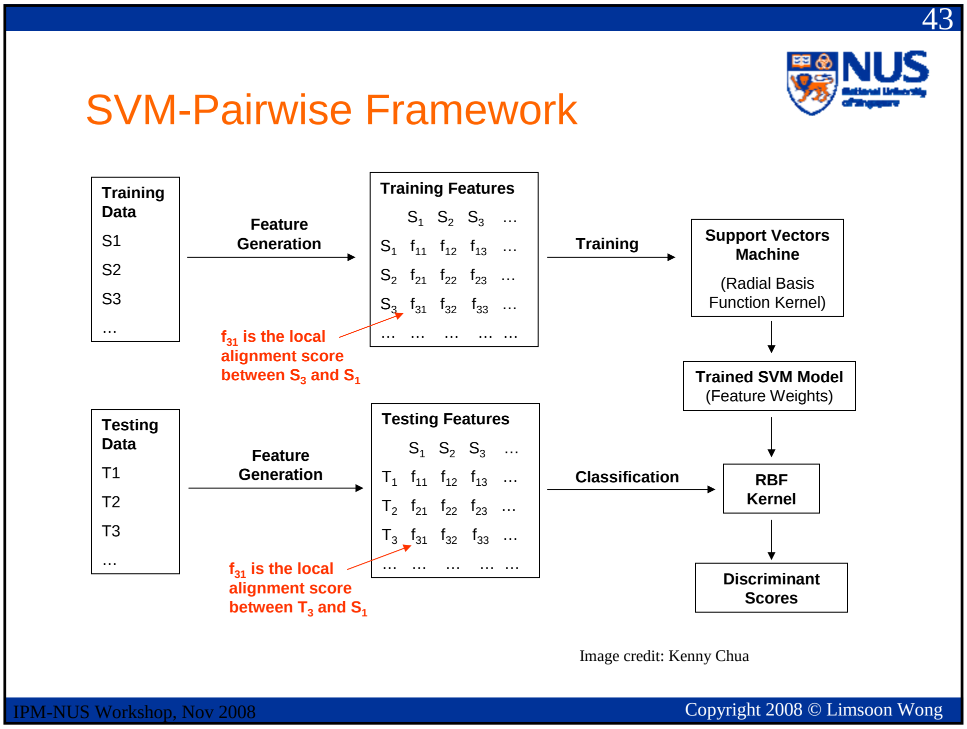
	Orange$_1$	Banana$_1$	…

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
			…

…	…	…	…



SVM-Pairwise Framework

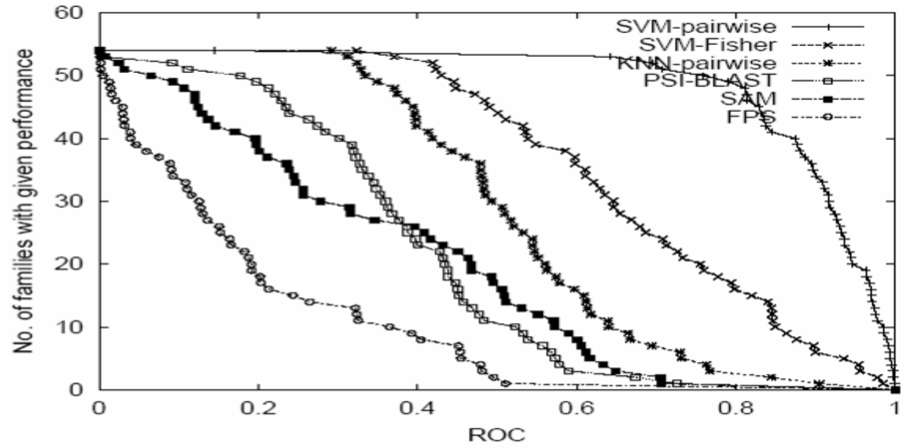
Training Data
S1
S2
S3
…
Feature Generation
Training Features
S_1 S_2 S_3 …
S_1 f_{11} f_{12} f_{13} …
S_2 f_{21} f_{22} f_{23} …
S_3 f_{31} f_{32} f_{33} …
… … … … …
Training
Support Vectors Machine
(Radial Basis Function Kernel)
Trained SVM Model (Feature Weights)
Testing Data
T1
T2
T3
…
Feature Generation
Testing Features
S_1 S_2 S_3 …
T_1 f_{11} f_{12} f_{13} …
T_2 f_{21} f_{22} f_{23} …
T_3 f_{31} f_{32} f_{33} …
… … … … …
Classification
RBF Kernel
Discriminant Scores

f_{31} is the local alignment score between S_3 and S_1

f_{31} is the local alignment score between T_3 and S_1

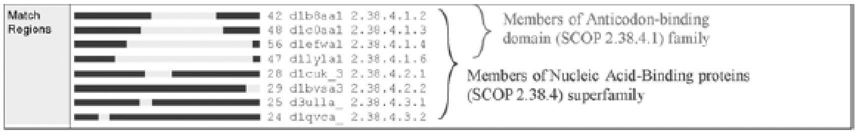
Image credit: Kenny Chua

Performance of SVM-Pairwise



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

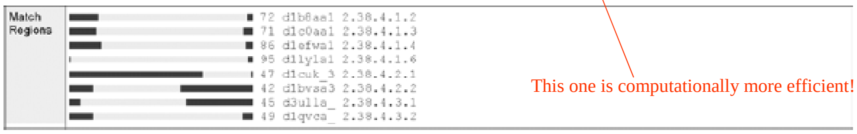
Simple Refinement to Capture Multiple Local Similarities



Members of Anticodon-binding domain (SCOP 2.38.4.1) family

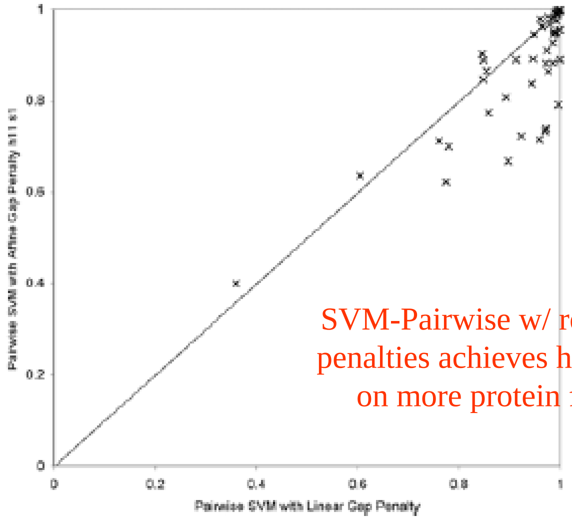
Members of Nucleic Acid-Binding proteins (SCOP 2.38.4) superfamily

relax gap penalties
(affine gap penality of open = -4, extend = -1)
(or linear gap penalty of -4)



This one is computationally more efficient!

ROC 2-D Plot of SVM Pairwise w/ vs w/o Relaxed Penalties



SVM-Pairwise w/ relaxed gap penalties achieves higher ROC on more protein families

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

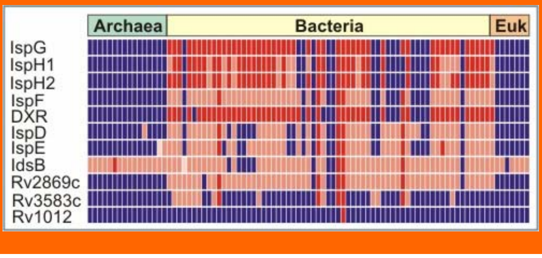


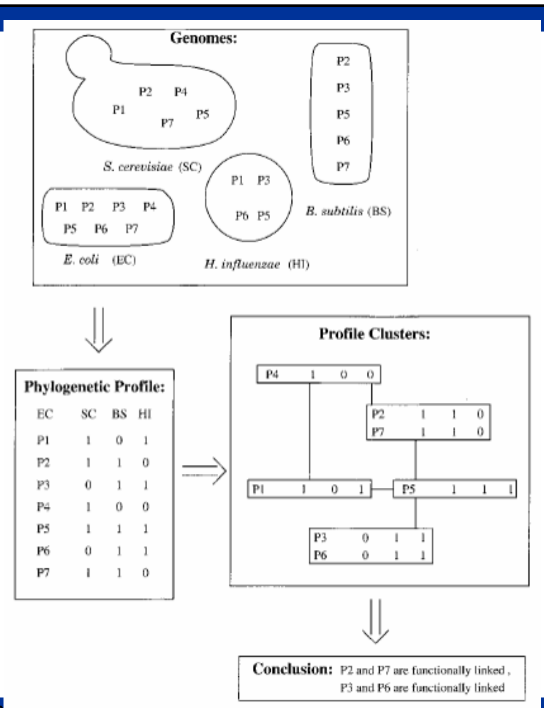
Image credit: Ed Marcotte, http://apropos.icmb.utexas.edu/plex/tour/isoprenoid.jpg

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 Profile:

EC	SC	BS	HI
P1	1	0	1
P2	1	1	0
P3	0	1	1
P4	1	0	0
P5	1	1	1
P6	0	1	1
P7	1	1	0

Conclusion: P2 and P7 are functionally linked, P3 and P6 are functionally linked

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

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

where

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

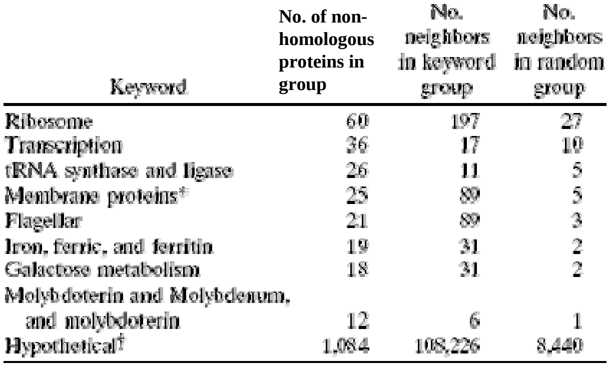
No. of ways to distribute *z* co-occurrences over *N* lineage's

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

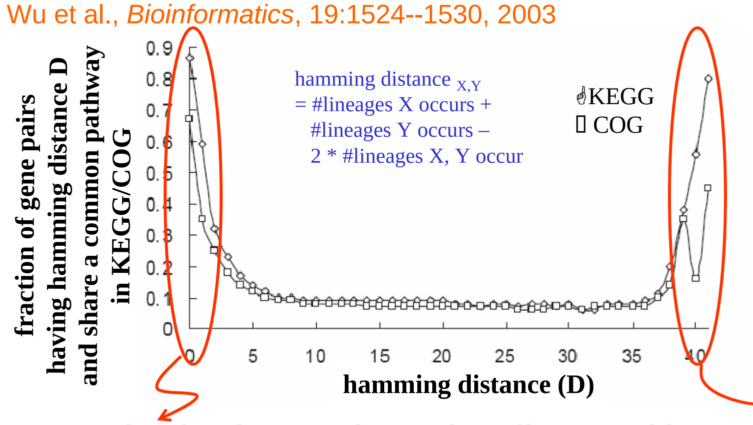


Keyword	No. of non-homologous proteins in group	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins*	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdoterin 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



hamming distance $_{X,Y}$
= #lineages X occurs +
#lineages Y occurs –
2 * #lineages X, Y occur

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

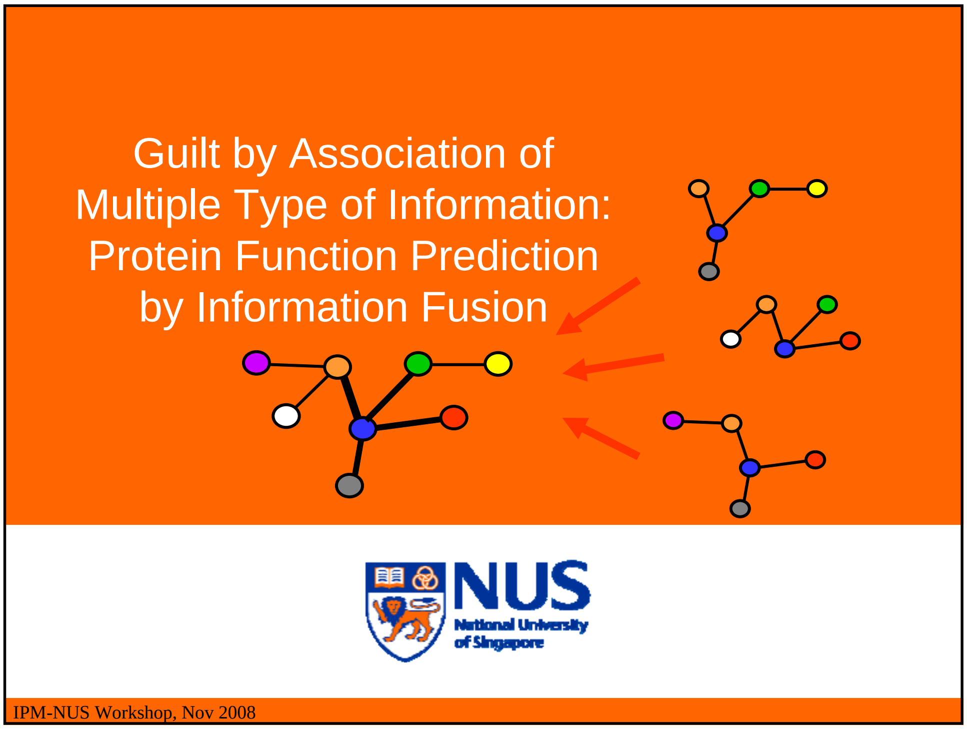
Exercise: Why do proteins having high hamming distance also have this behaviour?

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

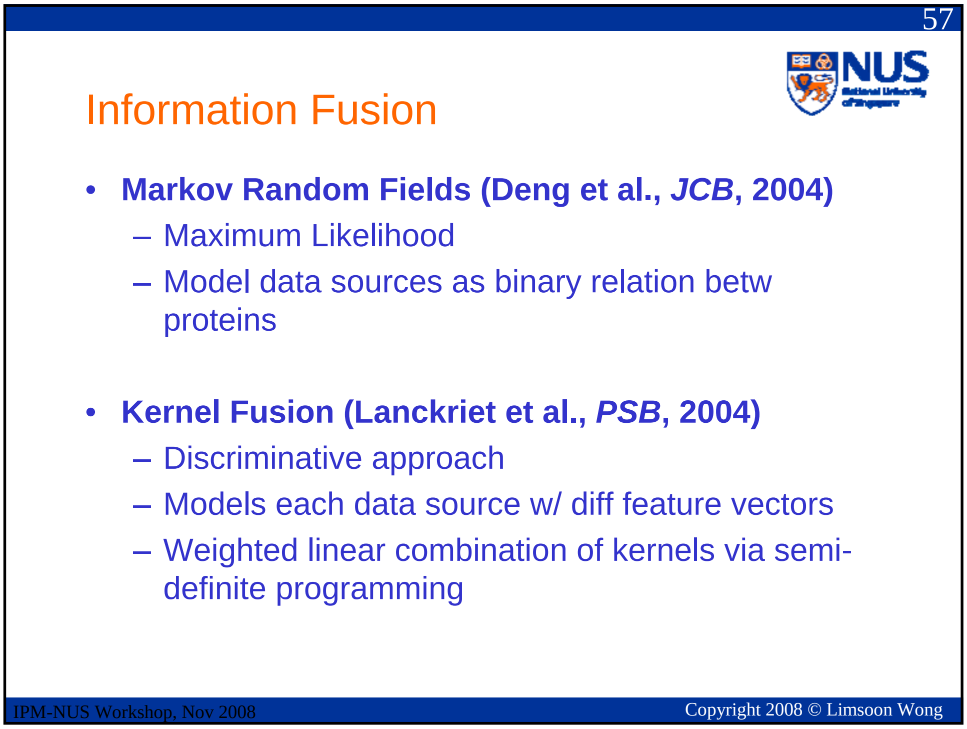
Any question?
Anyone needs a break?

National University of Singapore



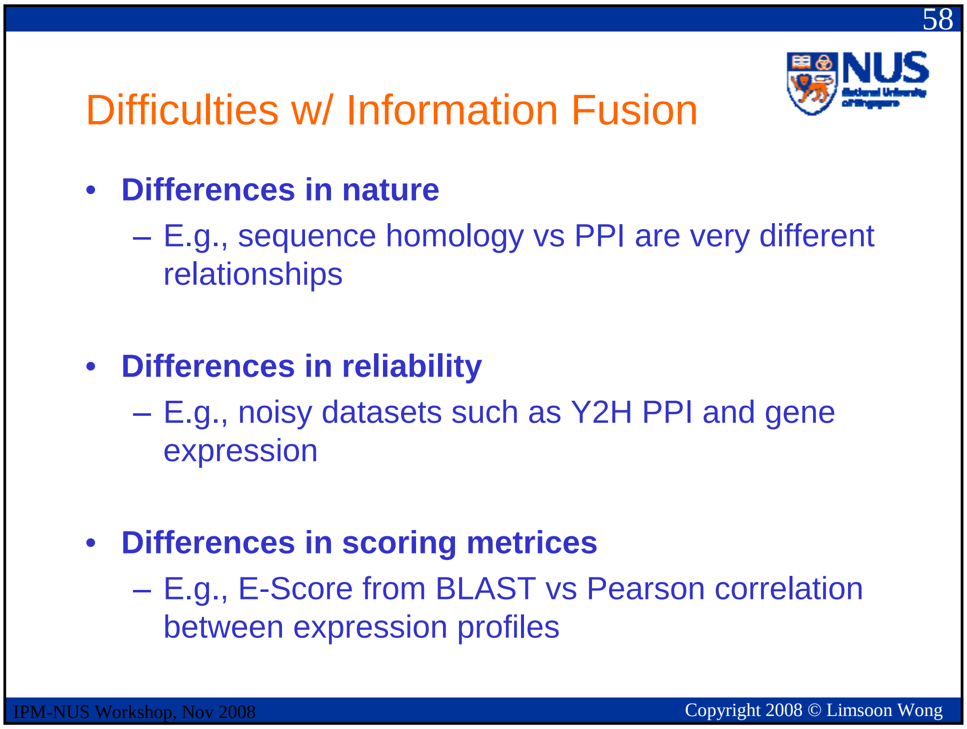
Guilt by Association of Multiple Type of Information: Protein Function Prediction by Information Fusion

National University of Singapore



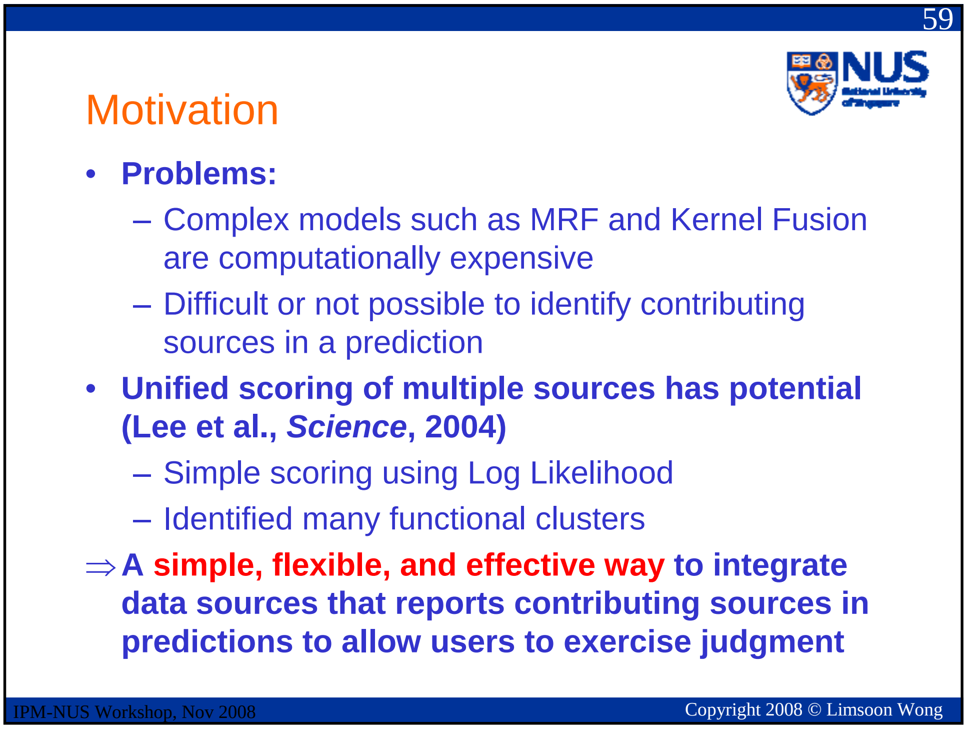
Information Fusion

- **Markov Random Fields (Deng et al., *JCB*, 2004)**
  - Maximum Likelihood
  - Model data sources as binary relation betw proteins
- **Kernel Fusion (Lanckriet et al., *PSB*, 2004)**
  - Discriminative approach
  - Models each data source w/ diff feature vectors
  - Weighted linear combination of kernels via semi-definite programming



Difficulties w/ Information Fusion

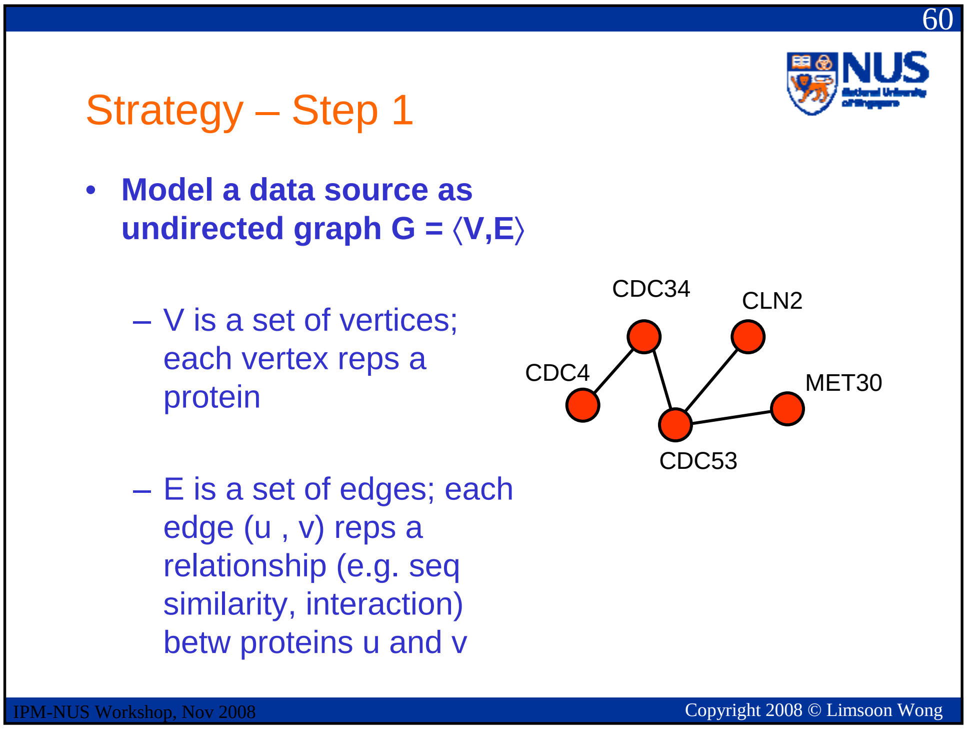
- **Differences in nature**
  - E.g., sequence homology vs PPI are very different relationships
- **Differences in reliability**
  - E.g., noisy datasets such as Y2H PPI and gene expression
- **Differences in scoring metrices**
  - E.g., E-Score from BLAST vs Pearson correlation between expression profiles



Motivation

- **Problems:**
  - Complex models such as MRF and Kernel Fusion are computationally expensive
  - Difficult or not possible to identify contributing sources in a prediction
- **Unified scoring of multiple sources has potential (Lee et al., *Science*, 2004)**
  - Simple scoring using Log Likelihood
  - Identified many functional clusters

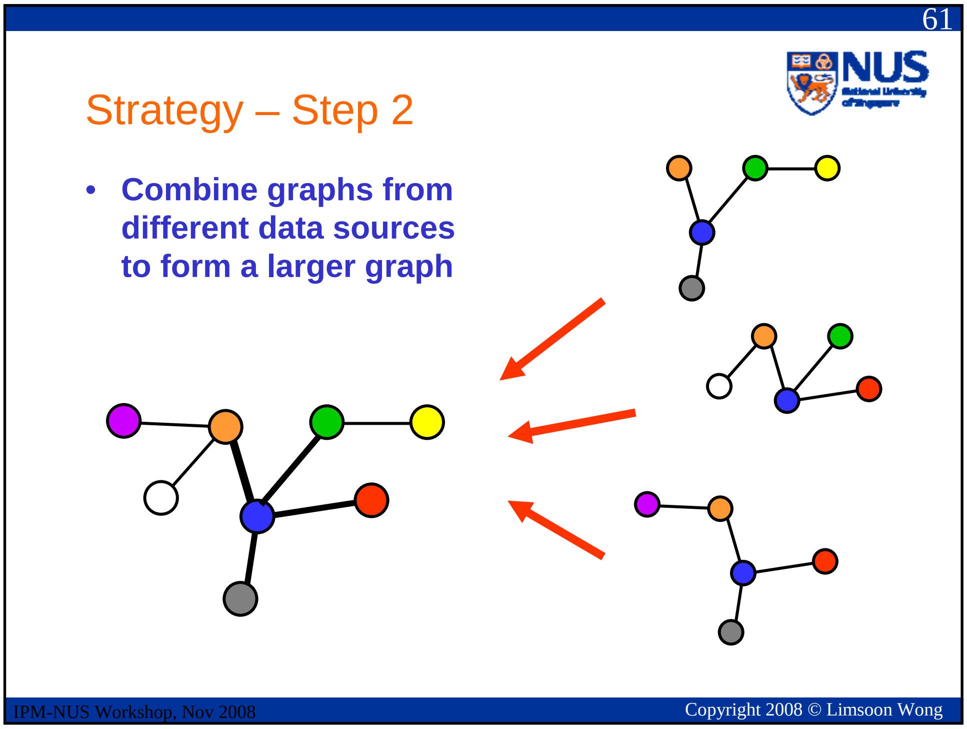
⇒**A simple, flexible, and effective way to integrate data sources that reports contributing sources in predictions to allow users to exercise judgment**



Strategy – Step 1

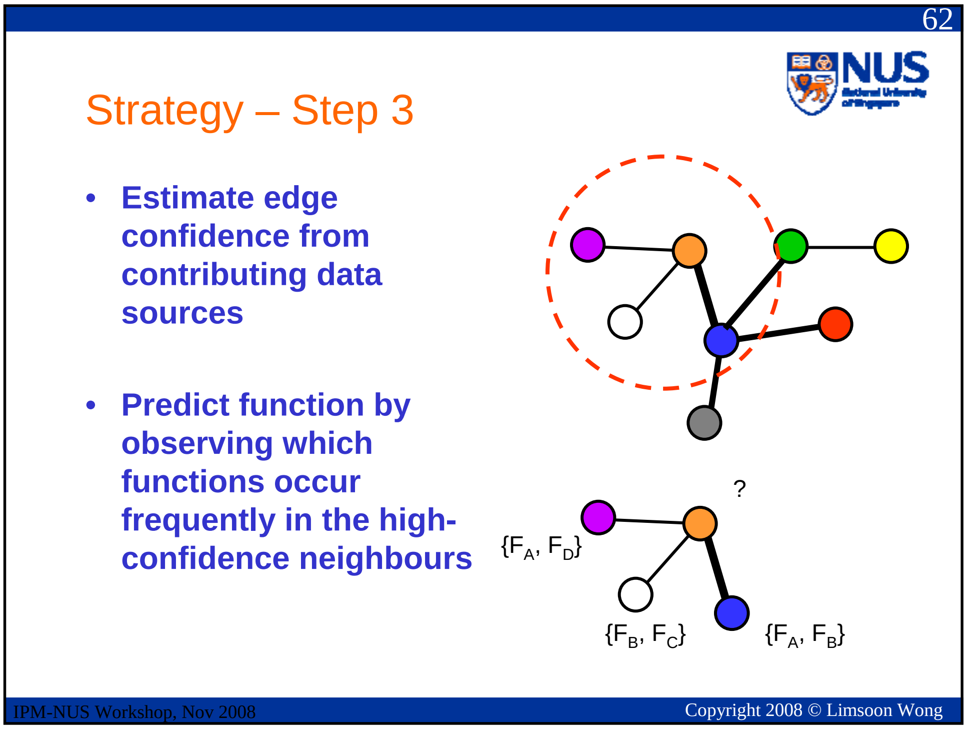
- **Model a data source as undirected graph G = ⟨V,E⟩**
  - V is a set of vertices; each vertex reps a protein
  - E is a set of edges; each edge (u , v) reps a relationship (e.g. seq similarity, interaction) betw proteins u and v

CDC34
CLN2
CDC4
MET30
CDC53



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:**

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

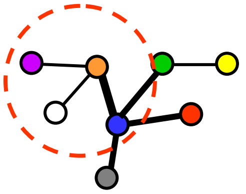
  - $E_{k,f}$ is subset of edges of subtype k where each edge has either one or both of its vertices annotated with function f
  - $S_f(u,v)$ = 1 if u and v shares function f, 0 otherwise

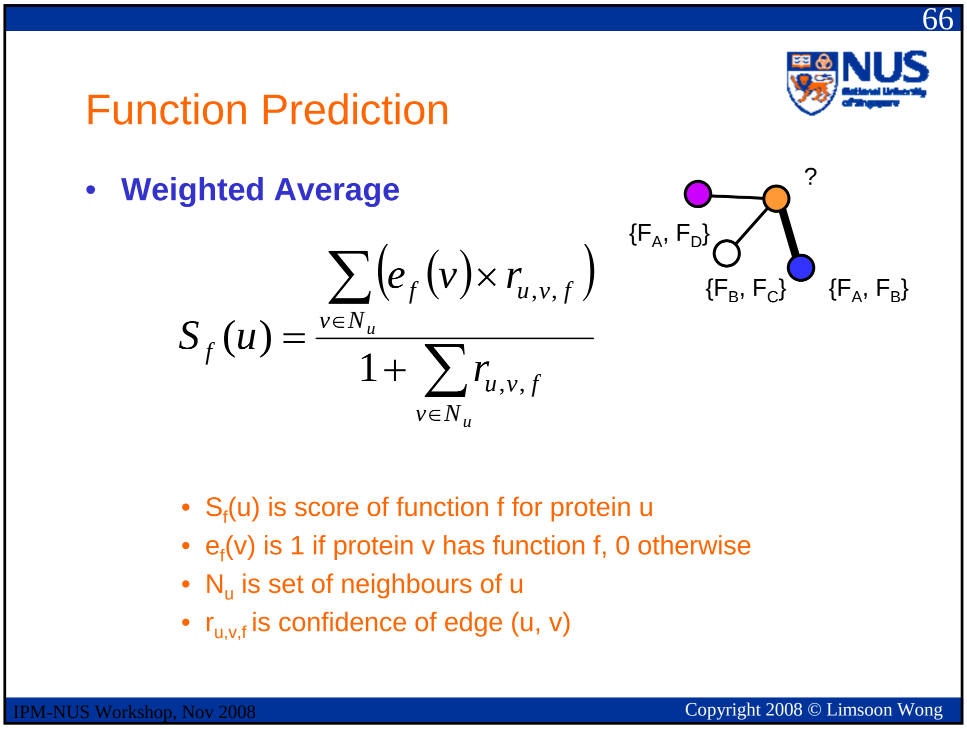
Combination of Confidence

- **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)





66

Function Prediction

- **Weighted Average**

{F$_A$, F$_D$}

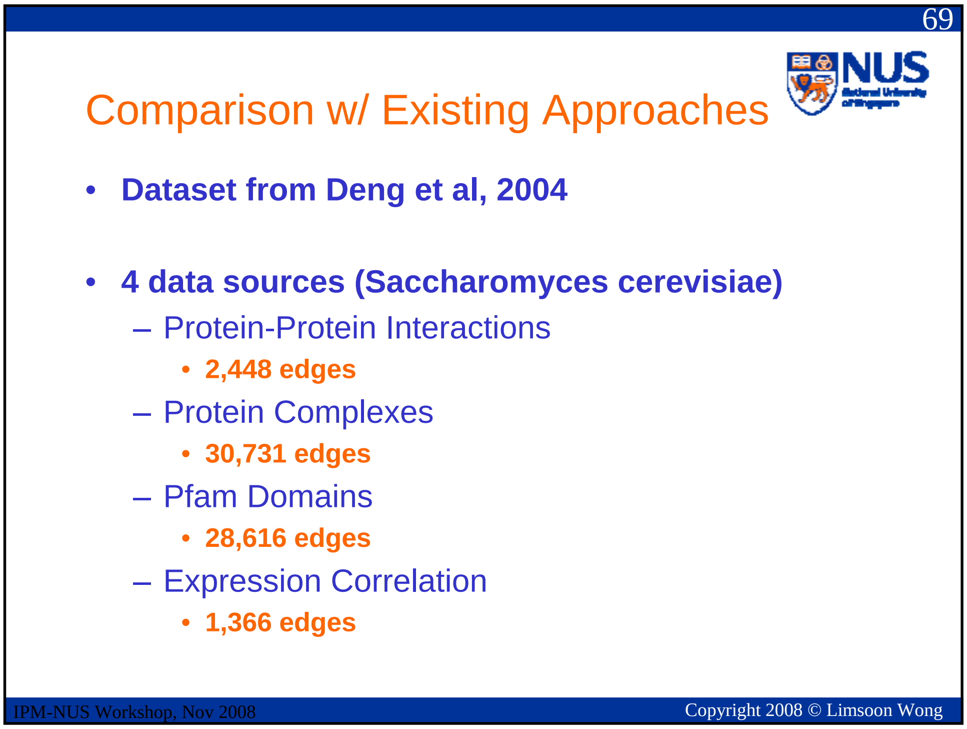
?

{F$_B$, F$_C$}

{F$_A$, F$_B$}

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

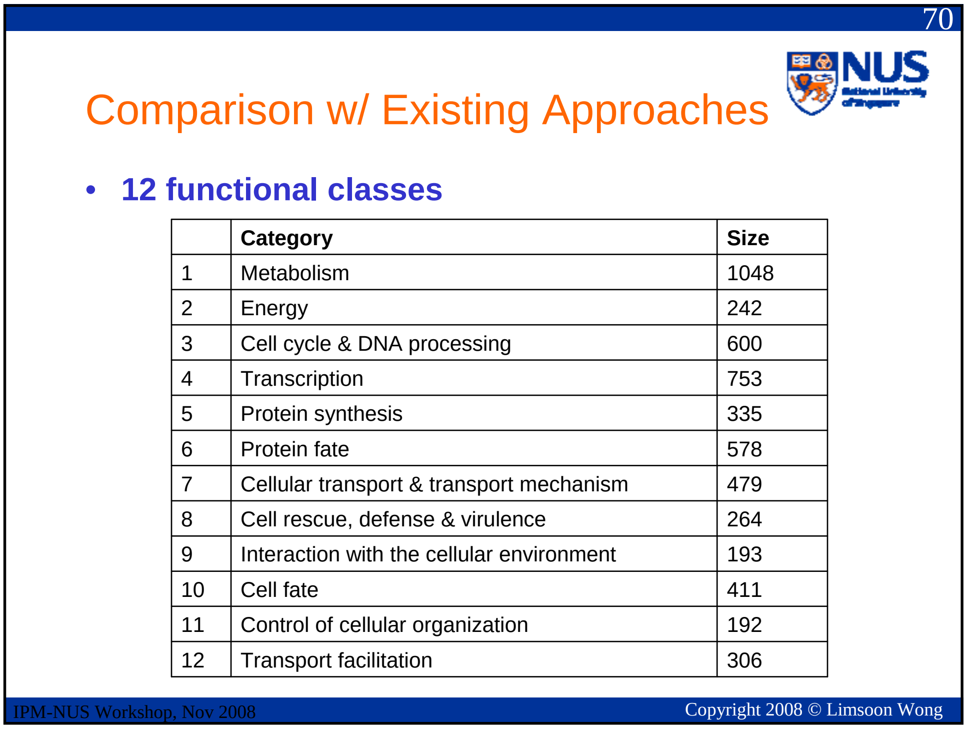
- $S_f(u)$ is score of function f for protein u
- $e_f(v)$ is 1 if protein v has function f, 0 otherwise
- N_u is set of neighbours of u
- $r_{u,v,f}$ is confidence of edge (u, v)



69

Comparison w/ Existing Approaches

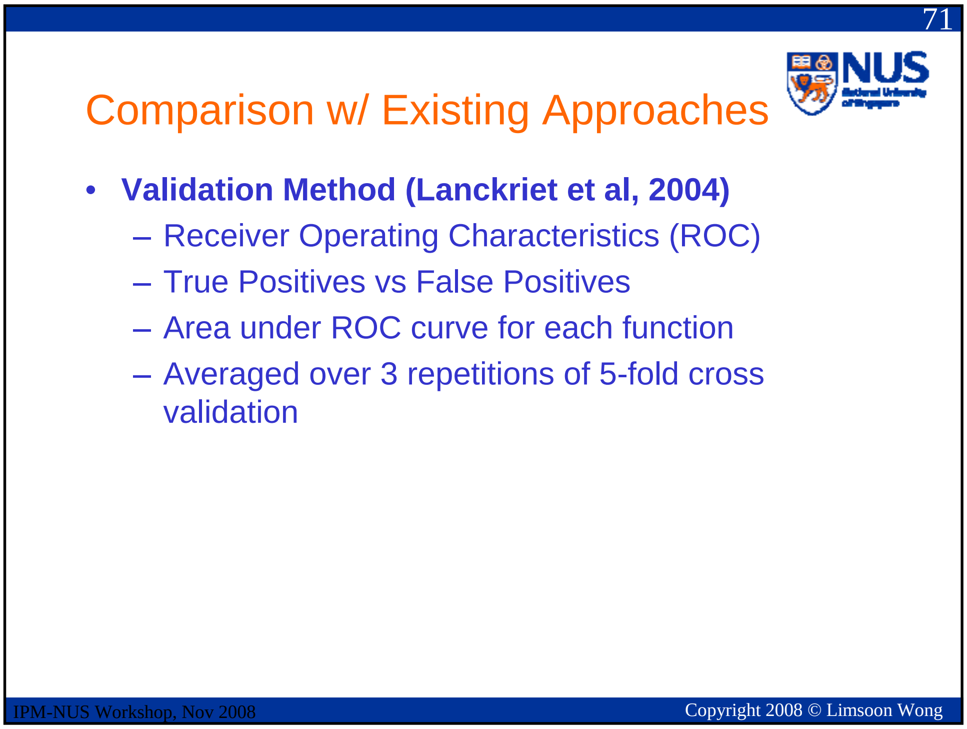
- **Dataset from Deng et al, 2004**
- **4 data sources (Saccharomyces cerevisiae)**
  - Protein-Protein Interactions
    - **2,448 edges**
  - Protein Complexes
    - **30,731 edges**
  - Pfam Domains
    - **28,616 edges**
  - Expression Correlation
    - **1,366 edges**



Comparison w/ Existing Approaches

- **12 functional classes**

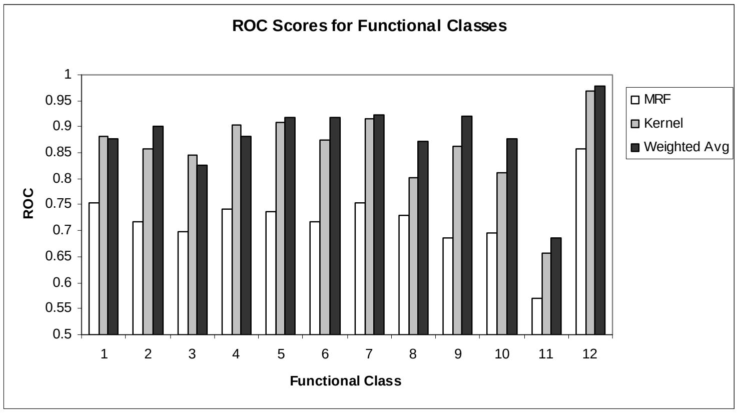
	Category	**Size**
1	Metabolism	1048
2	Energy	242
3	Cell cycle & DNA processing	600
4	Transcription	753
5	Protein synthesis	335
6	Protein fate	578
7	Cellular transport & transport mechanism	479
8	Cell rescue, defense & virulence	264
9	Interaction with the cellular environment	193
10	Cell fate	411
11	Control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



Comparison w/ Existing Approaches

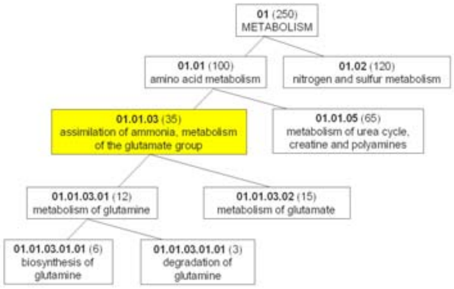
- **Validation Method (Lanckriet et al, 2004)**
  - Receiver Operating Characteristics (ROC)
  - True Positives vs False Positives
  - Area under ROC curve for each function
  - Averaged over 3 repetitions of 5-fold cross validation

Comparison w/ Existing Approaches



GO Terms Prediction for Yeast Proteins

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



01 (250) METABOLISM

01.01 (100) amino acid metabolism

01.02 (120) nitrogen and sulfur metabolism

01.01.03 (35) assimilation of ammonia, metabolism of the glutamate group

01.01.05 (65) metabolism of urea cycle, creatine and polyamines

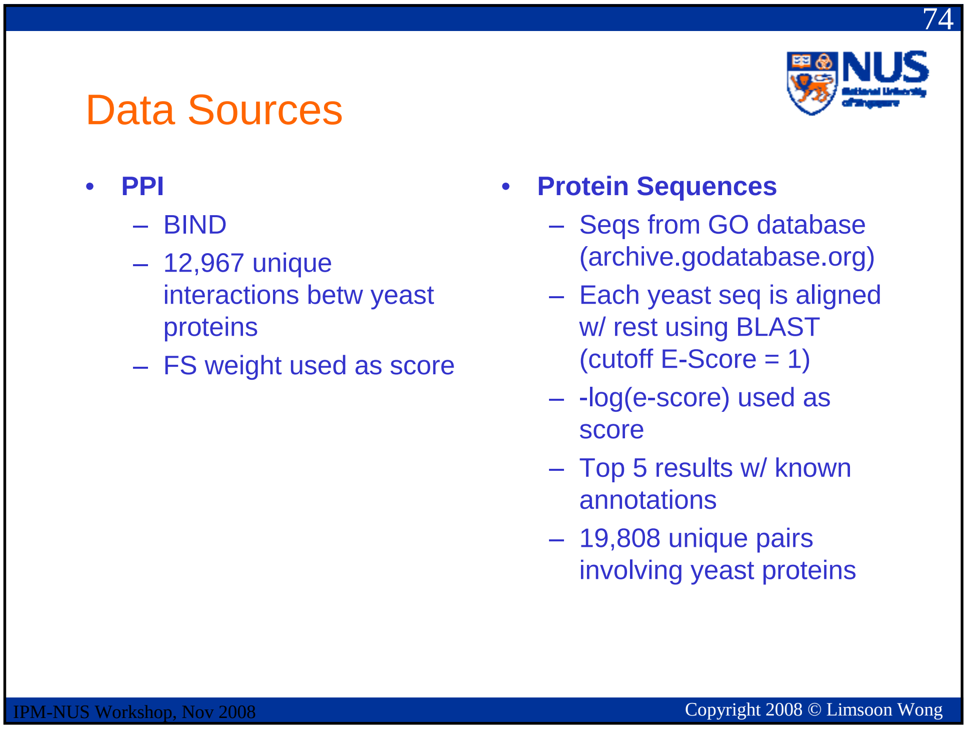
01.01.03.01 (12) metabolism of glutamine

01.01.03.02 (15) metabolism of glutamate

01.01.03.01.01 (6) biosynthesis of glutamine

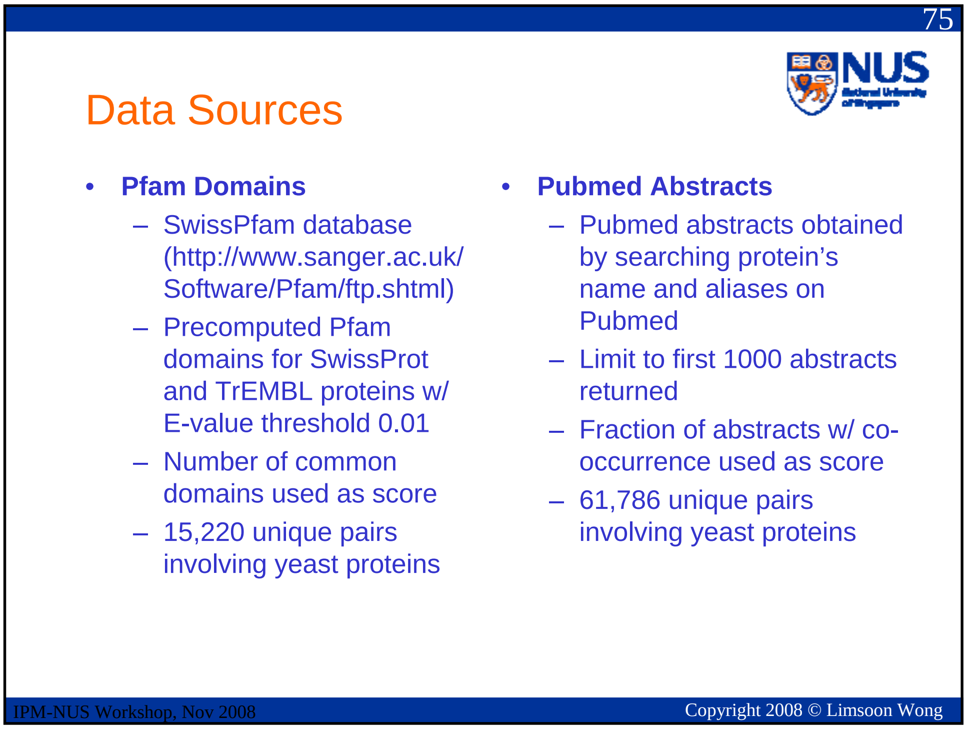
01.01.03.01.01 (3) degradation of glutamine

- **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

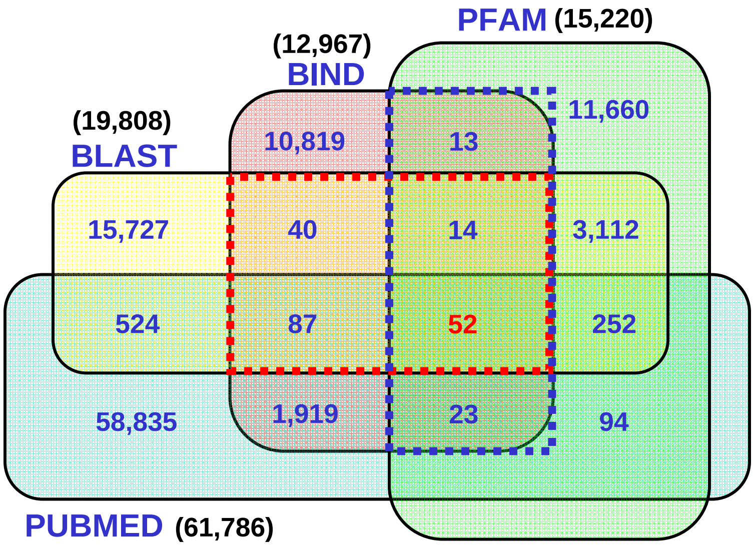
- **PPI**
  - BIND
  - 12,967 unique interactions betw yeast proteins
  - FS weight used as score
- **Protein Sequences**
  - Seqs from GO database (archive.godatabase.org)
  - Each yeast seq is aligned w/ rest using BLAST (cutoff E-Score = 1)
  - -log(e-score) used as score
  - Top 5 results w/ known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

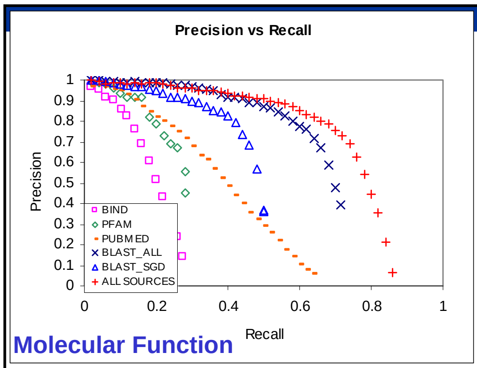


Data Sources

- **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

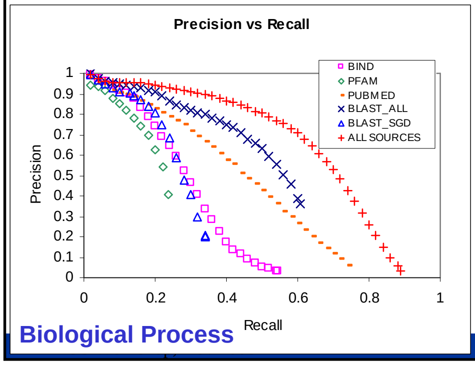
Multiple Data Sources



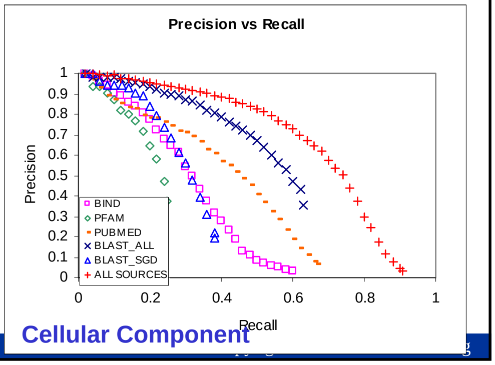


Molecular Function

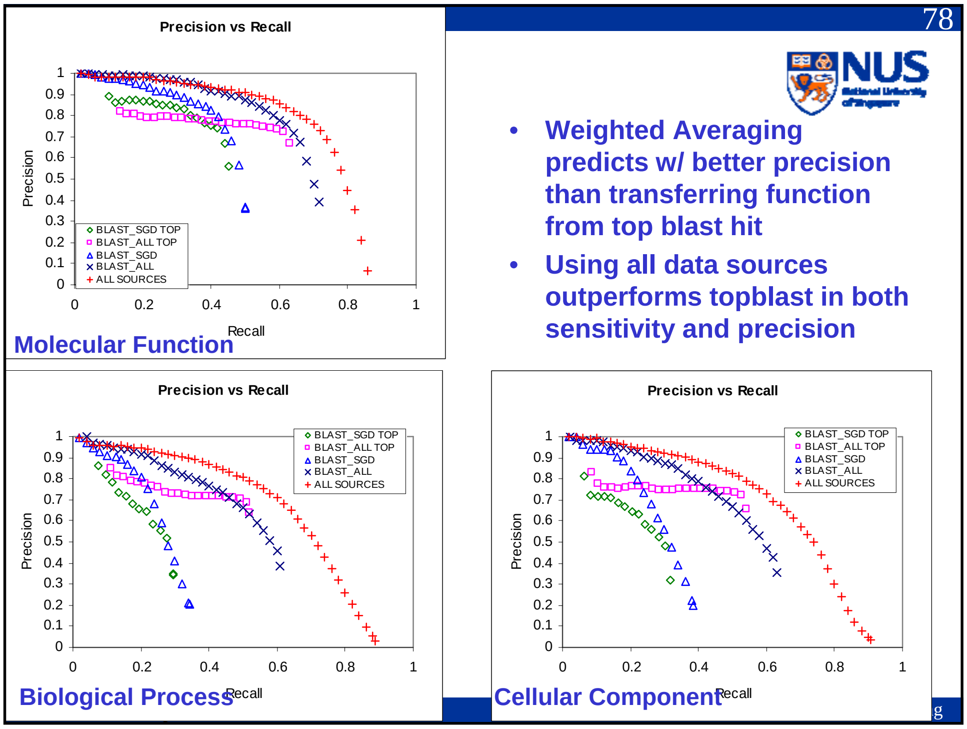
Combining all data sources outperforms any individual data source



Biological Process



Cellular Component



Molecular Function

Biological Process

Cellular Component

- 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?