For written notes on this lecture, please read chapter 19 of *The Practical Bioinformatician*

CS2220: Introduction to Computational Biology

Lecture 7: Sequence Homology Interpretation

Limsoon Wong

9 March 2007

NUS
National University of Singapore

Plan

- **Recap of sequence alignment**
- **Guilt by association**
- **Active site/domain discovery**
- **What if no homology of known function is found?**
  - Genome phylogenetic profiling
  - Protfun
  - SVM-Pairwise
  - Protein-protein interactions
- **Key mutation site discovery**

Very Brief Recap of Sequence Comparison/Alignment

National University of Singapore

Motivations for Sequence Comparison

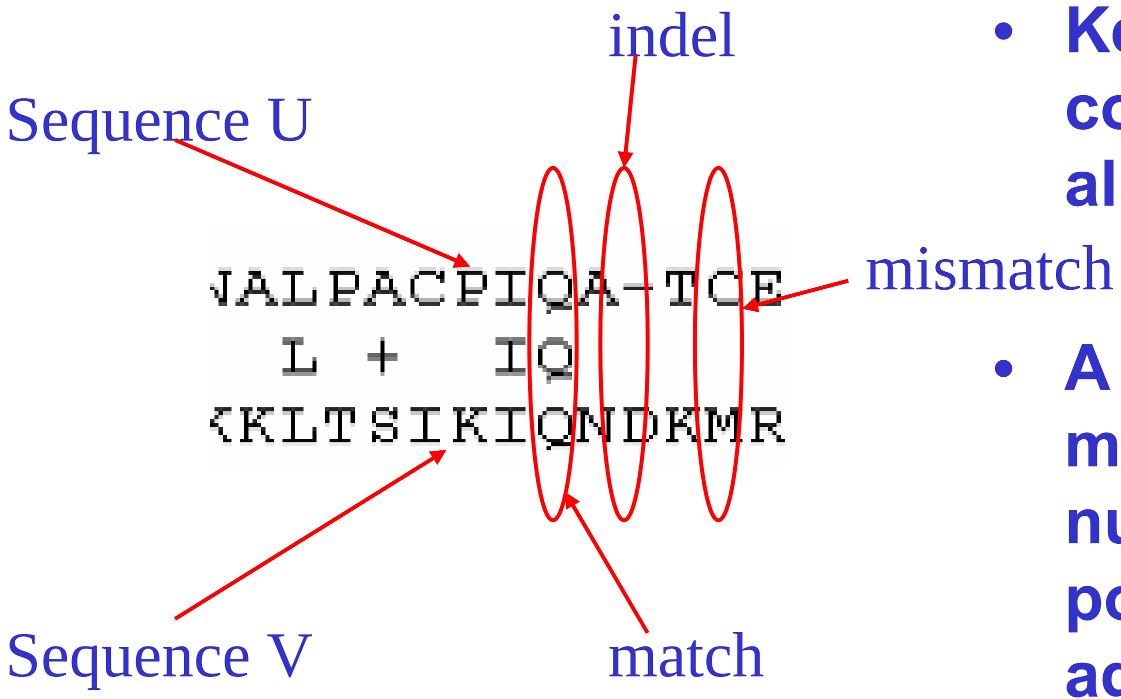
- DNA is blue print for living organisms

⇒ Evolution is related to changes in DNA

⇒ By comparing DNA sequences we can infer evolutionary relationships between the sequences w/o knowledge of the evolutionary events themselves

- Foundation for inferring function, active site, and key mutations

Sequence Alignment

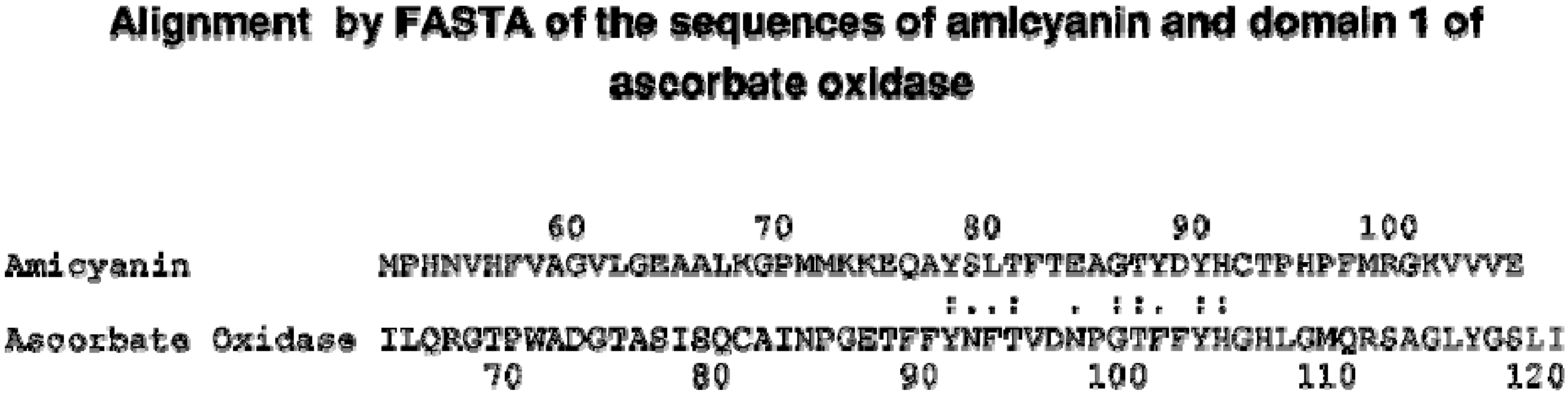


- **Key aspect of seq comparison is seq alignment**
- **A seq alignment maximizes the number of positions that are in agreement in two sequences**

Sequence 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

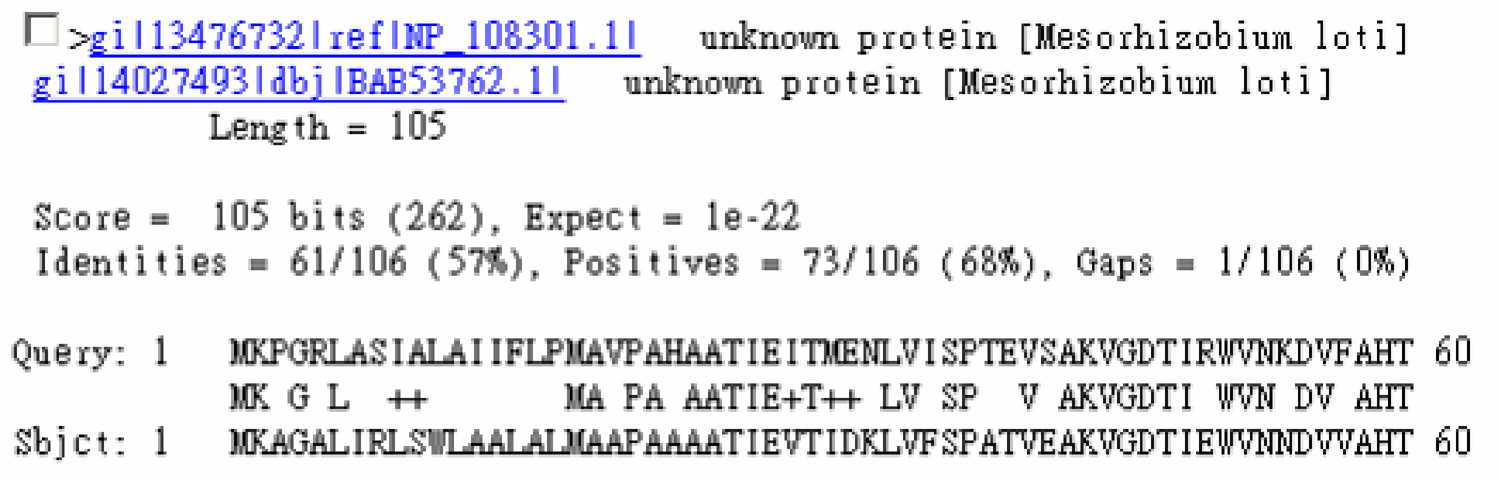
```
                          60        70        80        90       100
Amicyanin         MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVVE
                                             :..:  . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYGSLI
                       70        80        90       100       110       120
```

No obvious match between
Amicyanin and Ascorbate Oxidase

Sequence Alignment: Good Example

- **Good alignment usually has clusters of extensive matched positions**

⇒ The two proteins are likely to be homologous

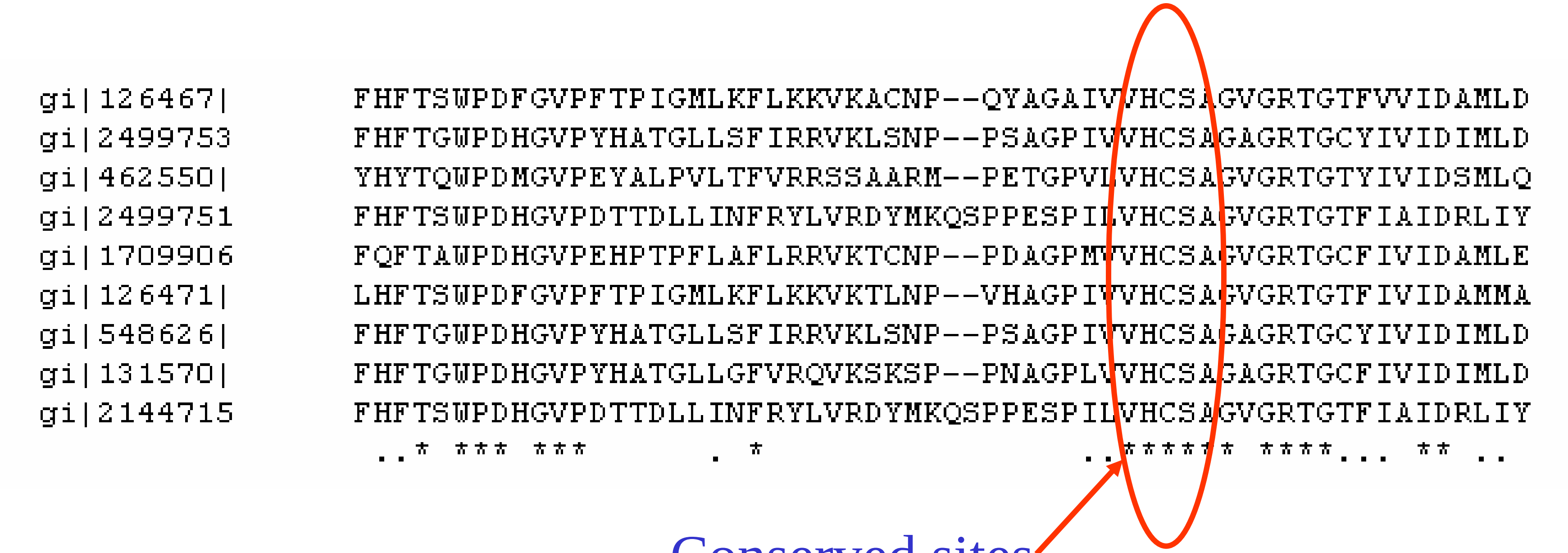


```
□ >gi|13476732|ref|NP_108301.1|   unknown protein [Mesorhizobium loti]
 gi|14027493|dbj|BAB53762.1|   unknown protein [Mesorhizobium loti]
          Length = 105

 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

Multiple Alignment: An Example

- **Multiple seq alignment maximizes number of positions in agreement across several seqs**
- **seqs belonging to same "family" usually have more conserved positions in a multiple seq alignment**



```
gi|126467|     FHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAMLD
gi|2499753     FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|462550|     YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSMLQ
gi|2499751     FHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLIY
gi|1709906     FQFTAWPDHGVPEHPTPFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAMLE
gi|126471|     LHFTSWPDFGVPFTPIGMLKFLKKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDAMMA
gi|548626|     FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|131570|     FHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPLVVHCSAGAGRTGCFIVIDIMLD
gi|2144715     FHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLIY
                ..* *** ***      . *                ..****** ****... ** ..
```

Conserved sites

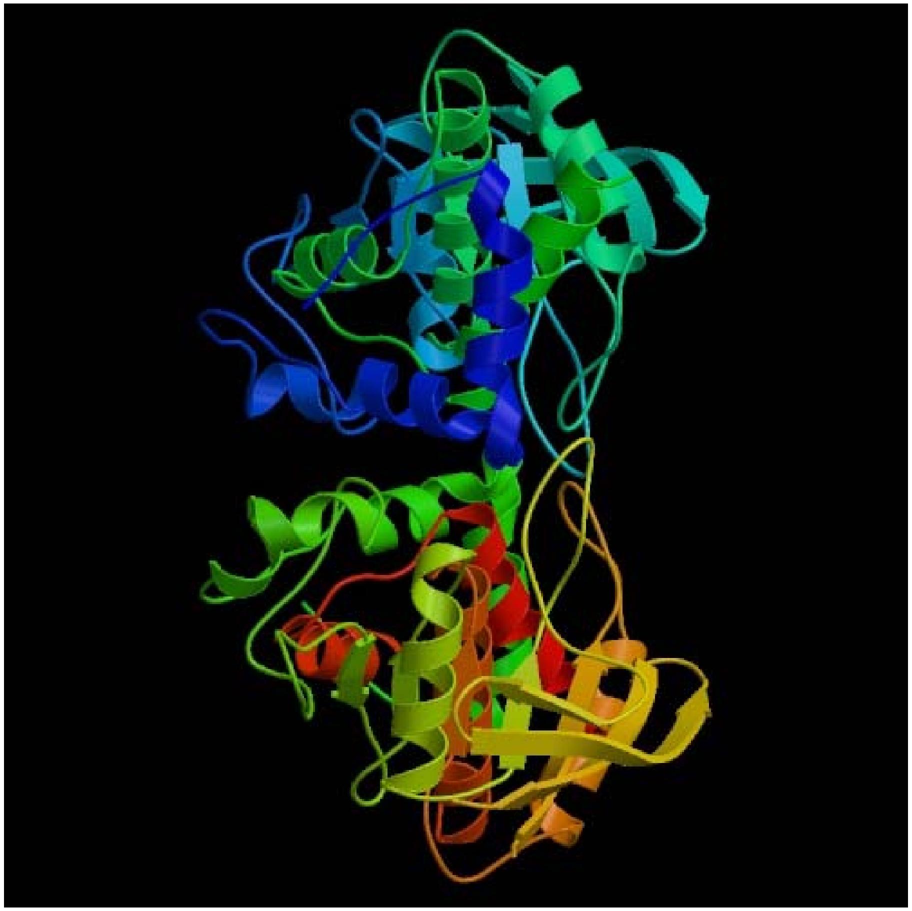
Application of Sequence Comparison: Guilt-by-Association

NUS

National University of Singapore

A protein is a ...

- **A protein is a large complex molecule made up of one or more chains of amino acids**
- **Protein performs a wide variety of activities in the cell**



Function Assignment to Protein Sequence

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

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

Guilt-by-Association

- Compare the target sequence T with sequences S_1, …, S_n of known function in a database

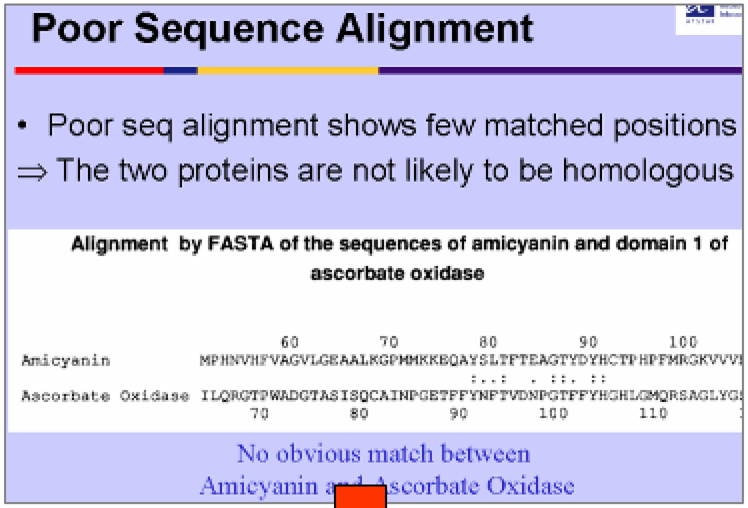
- Determine which ones amongst S_1, …, S_n are the mostly likely homologs of T

- Then assign to T the same function as these homologs

- Finally, confirm with suitable wet experiments

Guilt-by-Association

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



Poor Sequence Alignment

- Poor seq alignment shows few matched positions

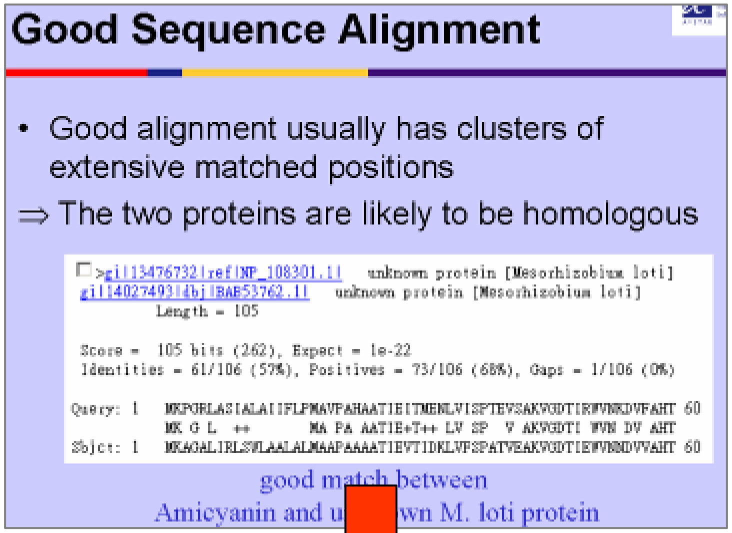
⇒ The two proteins are not likely to be homologous

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```
                        60        70        80        90        100
Amicyanin          MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVVE
                                              :..:   . ::. ::
Ascorbate Oxidase  ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYGS
                      70        80        90        100       110
```

No obvious match between Amicyanin and Ascorbate Oxidase

Discard this function as a candidate



Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

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

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

Query: 1   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

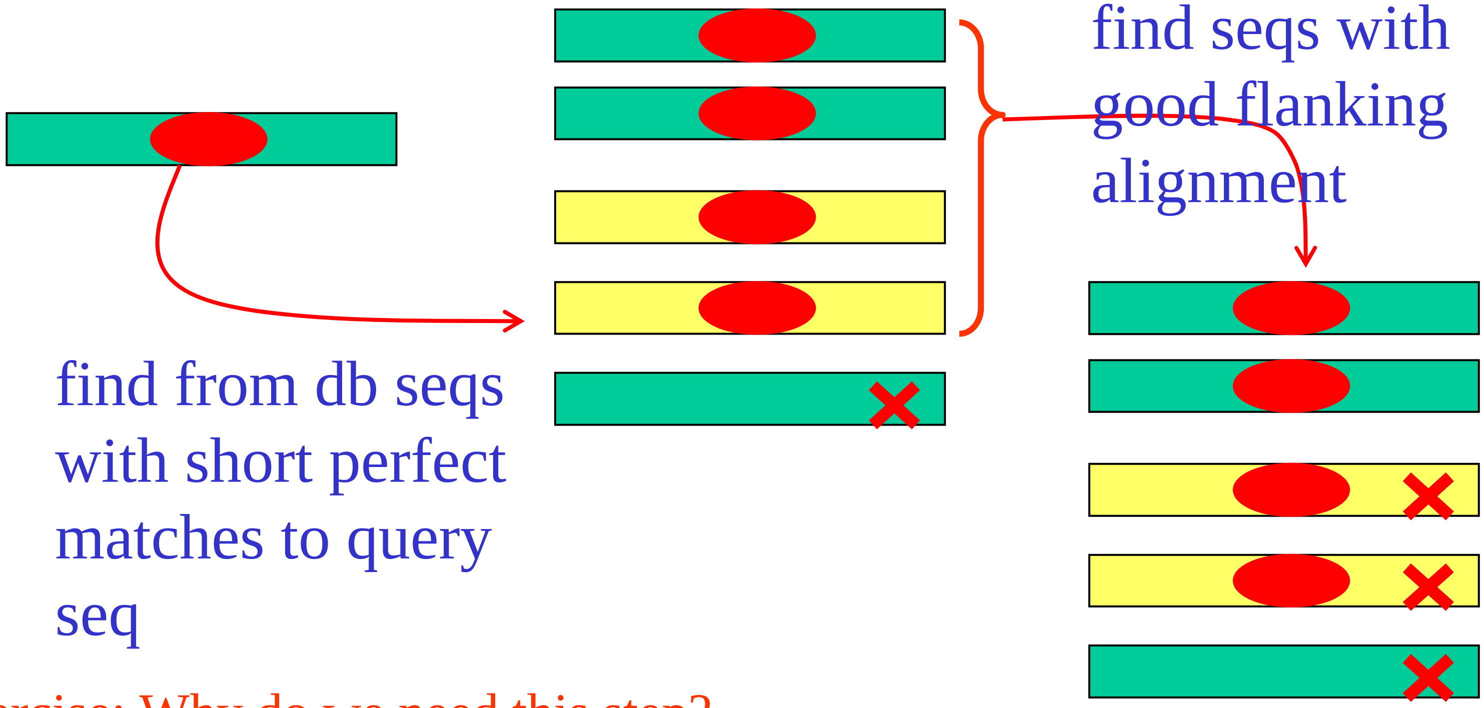
Assign to *T* same function as homologs

Confirm with suitable wet experiments

BLAST: How It Works

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

- **BLAST is one of the most popular tool for doing "guilt-by-association" sequence homology search**

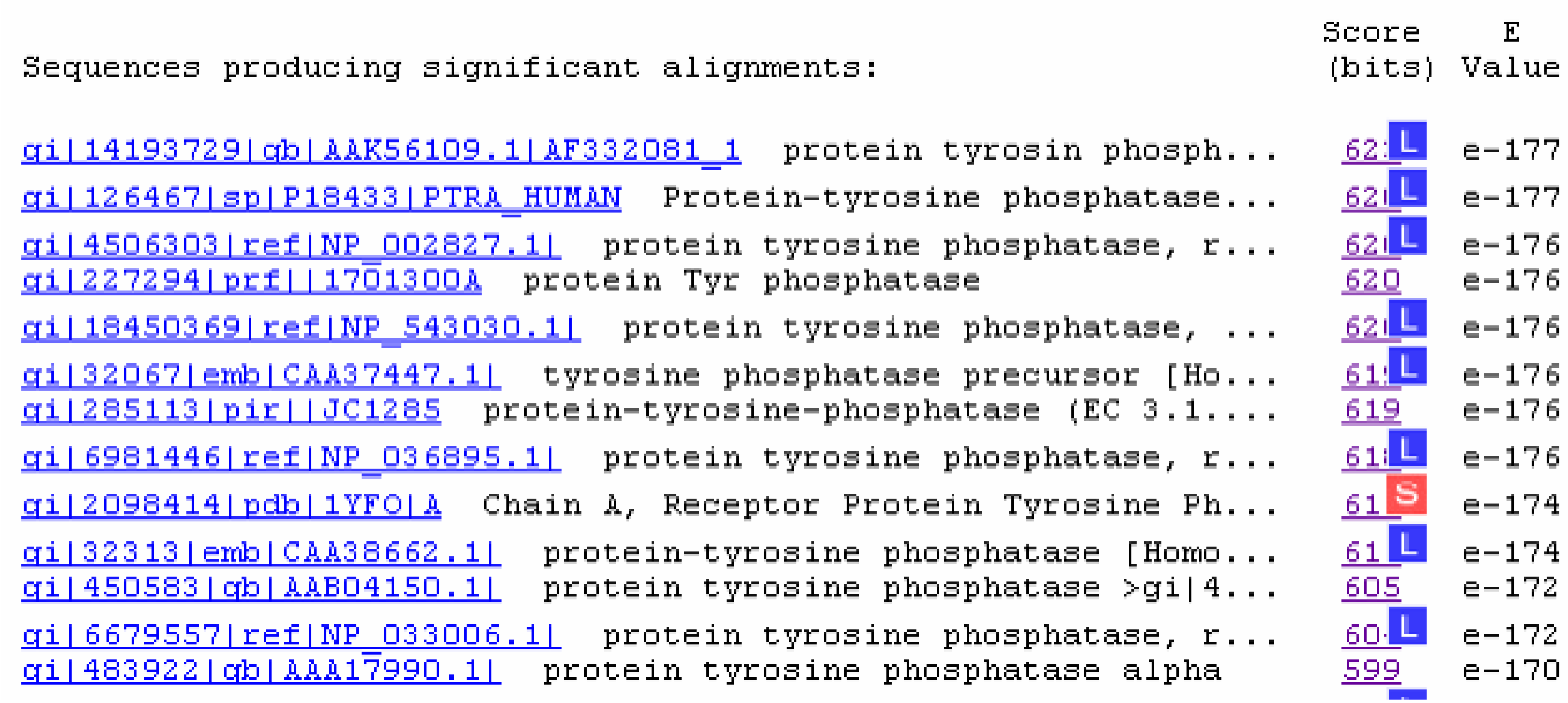


find from db seqs with short perfect matches to query seq

find seqs with good flanking alignment

Exercise: Why do we need this step?

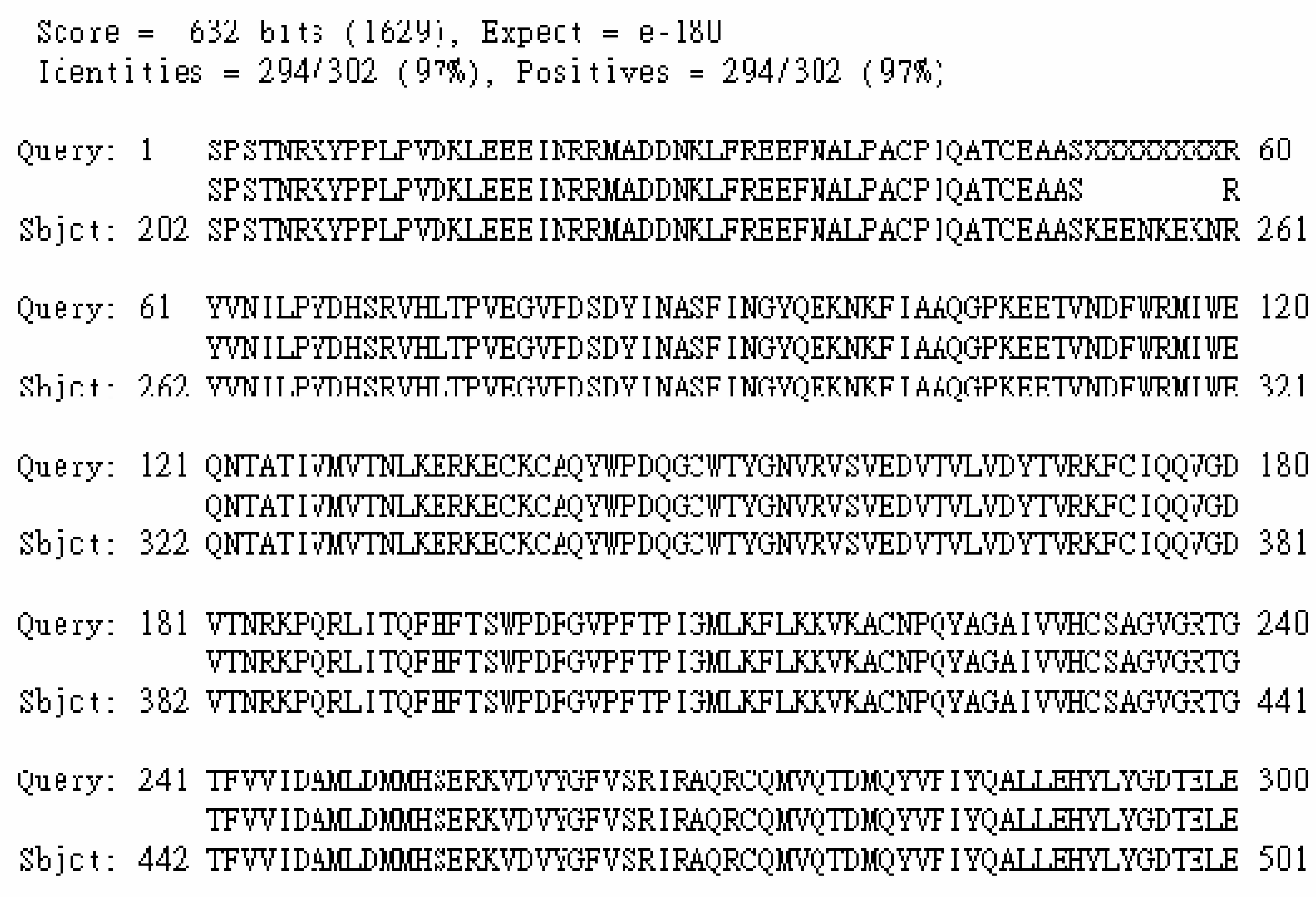
Homologs obtained by BLAST



```
                                                                   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  YVNILPYDHSRVHLTPVEGVFDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE 120
            YVNILPYDHSRVHLTPVEGVFDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
Sbjct: 262 YVNILPYDHSRVHLTPVEGVFDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE 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
```

Guilt-by-Association: Caveats

- Ensure that the effect of database size has been accounted for

- Ensure that the function of the homology is not derived via invalid “transitive assignment”

- Ensure that the target sequence has all the key features associated with the function, e.g., active site and/or domain

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!**

Exercise: Name a commonly used method for correcting p-value for a situation like this

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**

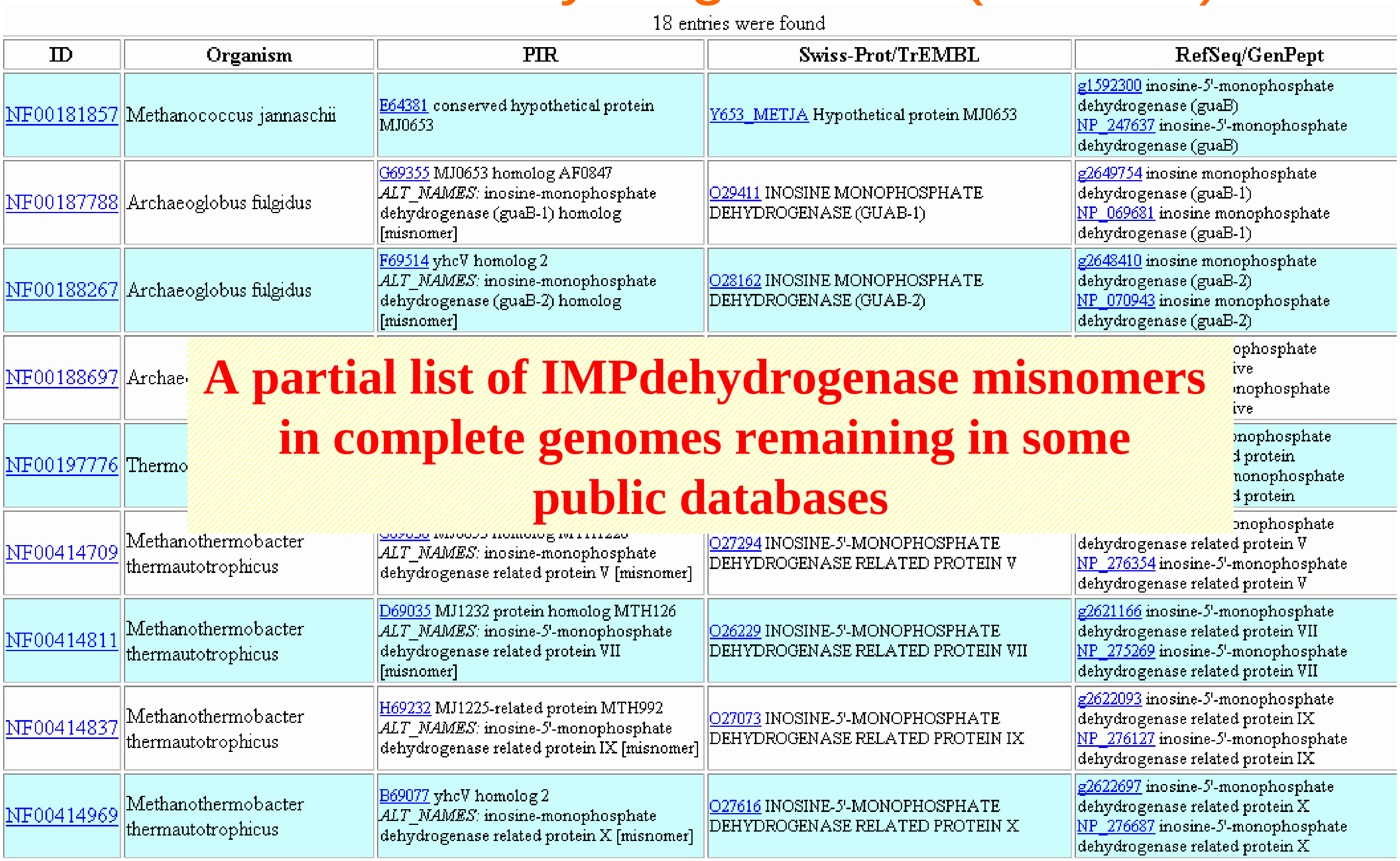
⇒ **While it is worth noting that two proteins contain similar low complexity regions, they are best excluded when constructing alignments**

- **E.g., by default, BLAST employs the SEG algo to filter low complexity regions from proteins before executing a search**

Source: NCBI

Examples of Invalid Function Assignment:

The IMP Dehydrogenases (IMPDH)

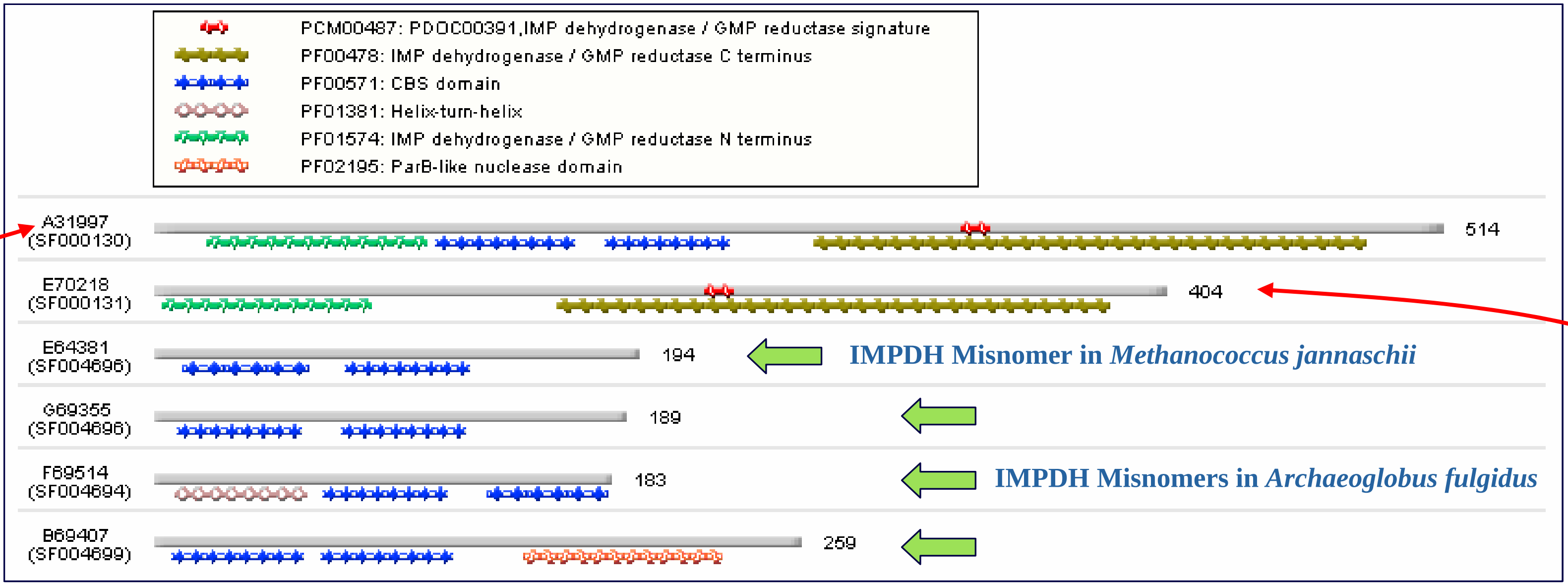


18 entries were found

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 nonophosphate d protein
NF00414709	Methanothermobacter thermautotrophicus	[illegible] *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
G69355 (SF004696)
189
F69514 (SF004694)
183
B69407 (SF004699)
259

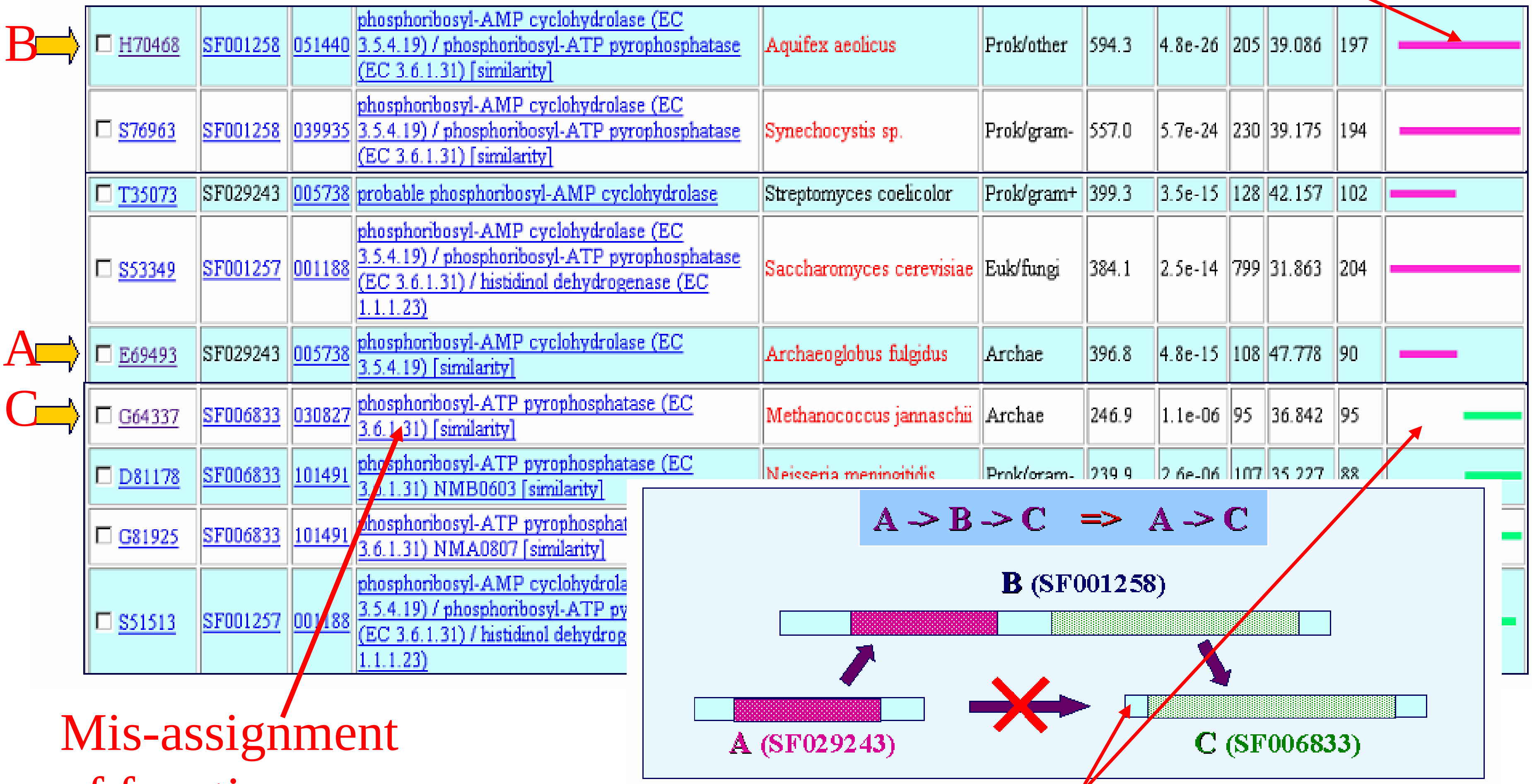
IMPDH Misnomer in *Methanococcus jannaschii*

IMPDH Misnomers in *Archaeoglobus fulgidus*

- 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)								━

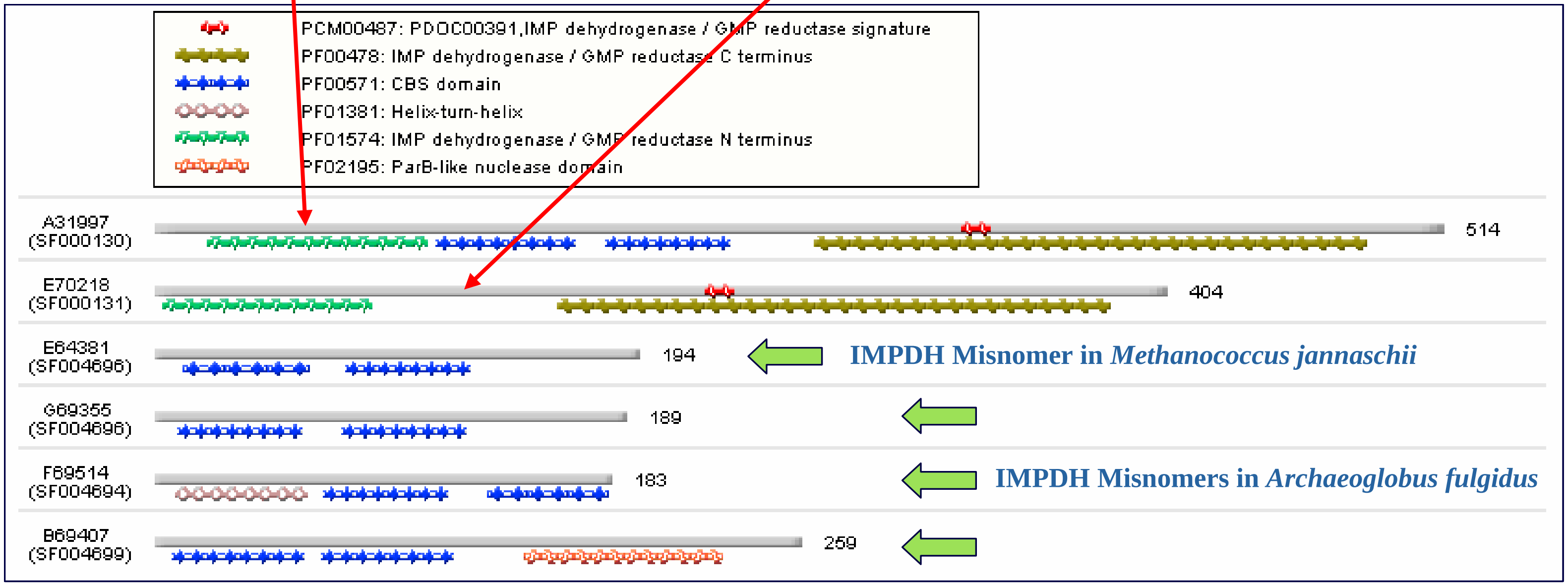
Mis-assignment of function

No IMPDH domain

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
G69355 (SF004696)
189
F69514 (SF004694)
183
B69407 (SF004699)
259

IMPDH Misnomer in *Methanococcus jannaschii*

IMPDH Misnomers in *Archaeoglobus fulgidus*

- **Most IMPDHs have 2 IMPDH and 2 CBS domains**
- **Some IMPDH (E70218) lacks CBS domains**

⇒ **IMPDH domain is the emerging pattern**

Application of Sequence Comparison: Active Site/Domain Discovery

National University of Singapore

Discover Active Site and/or Domain

- **How to discover the active site and/or domain of a function in the first place?**
  - Multiple alignment of homologous seqs
  - Determine conserved positions

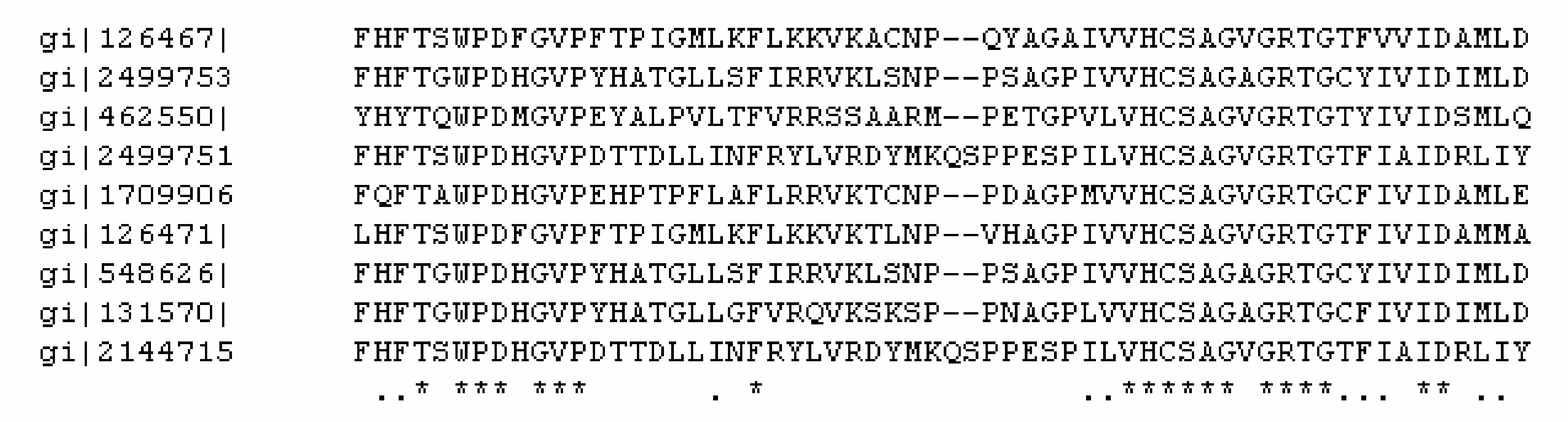
  ⇒ Emerging patterns relative to background

  ⇒ Candidate active sites and/or domains

- **Easier if sequences of distance homologs are used**

Exercise: Why?

Multiple Alignment of PTPs



```
gi|126467|       FHFTSWPDFGVPFTPIGMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAMLD
gi|2499753       FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|462550|       YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSMLQ
gi|2499751       FHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLIY
gi|1709906       FQFTAWPDHGVPEHPTPFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAMLE
gi|126471|       LHFTSWPDFGVPFTPIGMLKFLKKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDAMMA
gi|548626|       FHFTGWPDHGVPYHATGLLSFIRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|131570|       FHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPLVVHCSAGAGRTGCFIVIDIMLD
gi|2144715       FHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPESPILVHCSAGVGRTGTFIAIDRLIY
                  ..* *** ***      . *                ..****** ****... ** ..
```

- **Notice the PTPs agree with each other on some positions more than other positions**
- **These positions are more impt wrt PTPs**
- **Else they wouldn’t be conserved by evolution**

⇒ **They are candidate active sites**

Guilt-by-Association:
What if no homolog of known function is found?

genome phylogenetic profiles

protfun's feature profiles

Similarity of dissimilarities

NUS
National University of Singapore

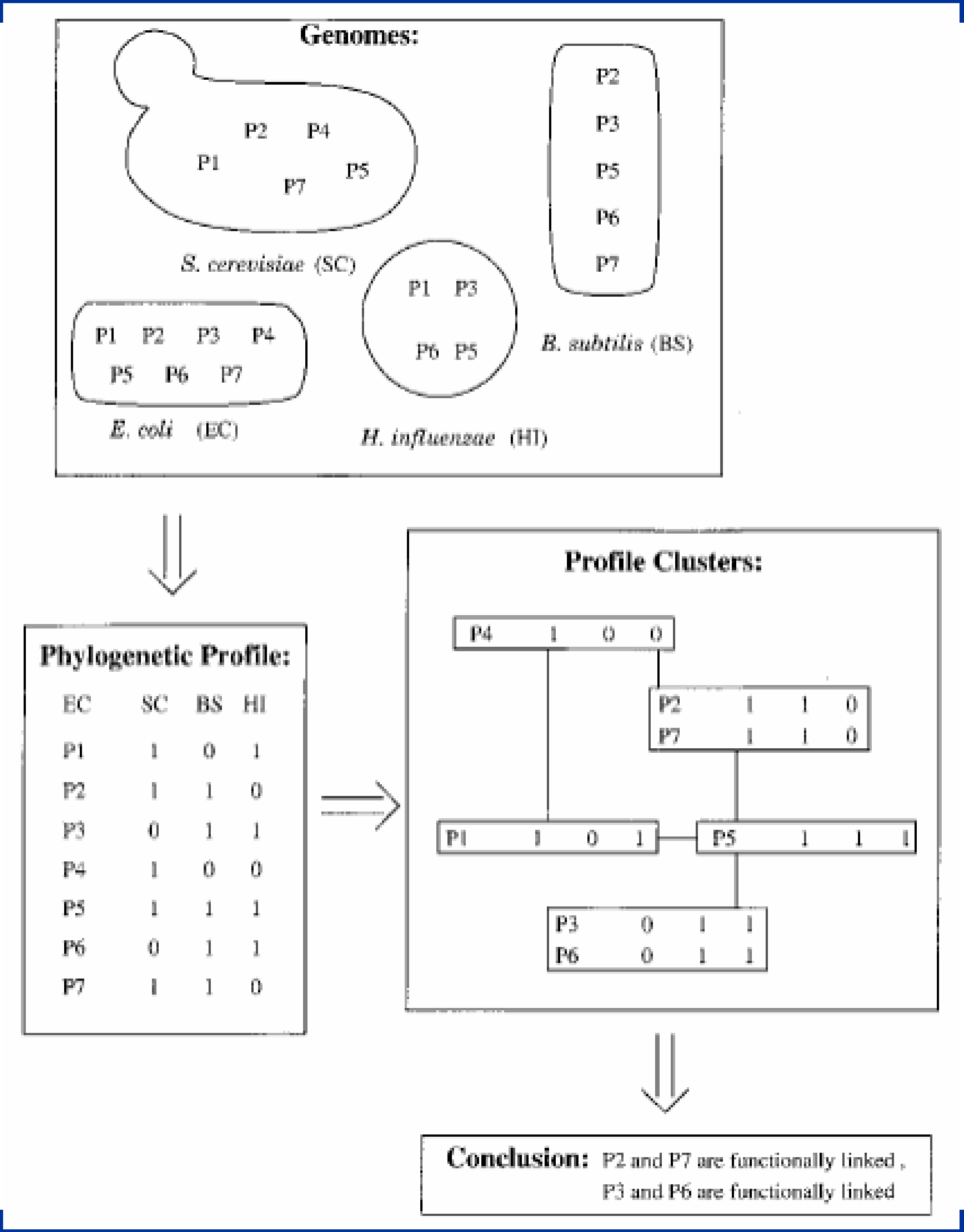
Phylogenetic Profiling

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 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

Profile Clusters:

P4	1	0	0
P2	1	1	0
P7	1	1	0
P1	1	0	1
P5	1	1	1
P3	0	1	1
P6	0	1	1

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

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

$$w_z = \binom{N}{z}$$

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

$$\overline{w_z} = \binom{N-z}{x-z} * \binom{N-Z}{y-z}$$

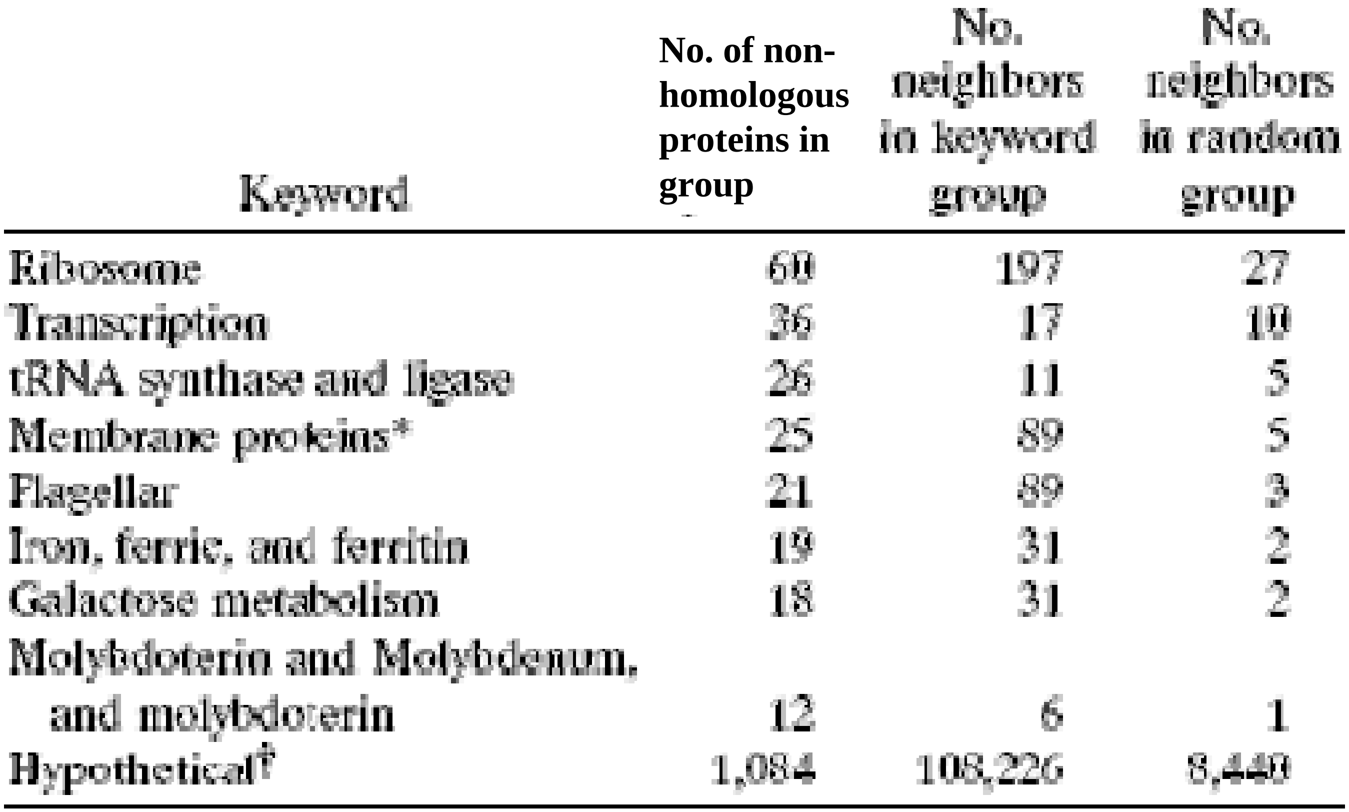
No. of ways to distribute the remaining *x – z* and *y – z* occurrences over the remaining *N – z* lineage's

$$W = \binom{N}{x} * \binom{N}{y}$$

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

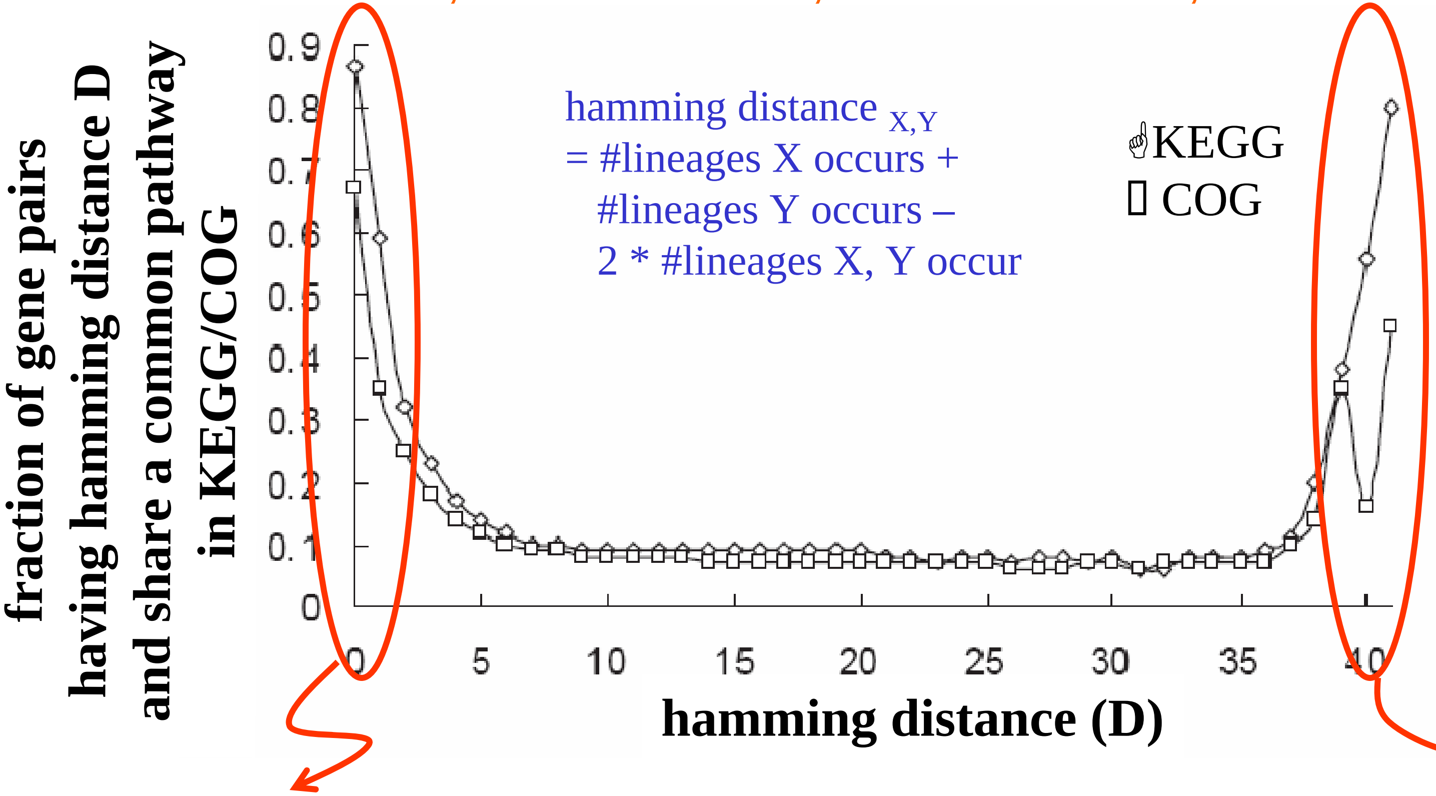


Keyword	No. of non-homologous proteins in group	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins*	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdoterin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

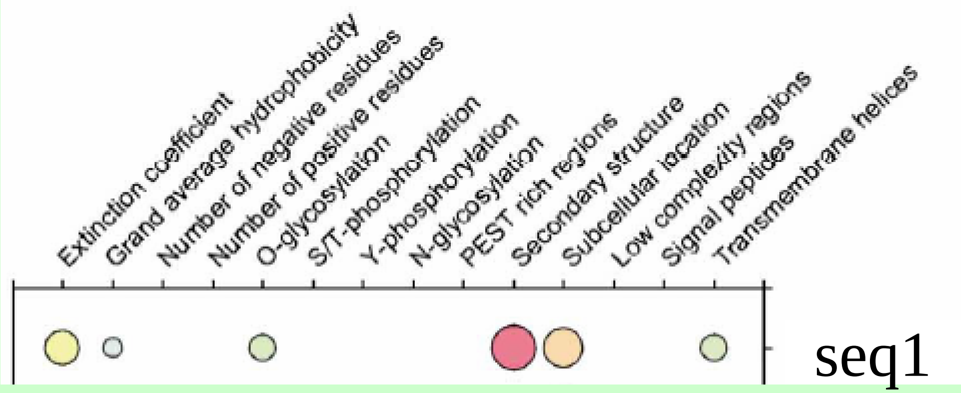
- **Proteins having low hamming distance (thus highly similar phylogenetic profiles) tend to share common pathways**

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

The ProtFun Approach

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

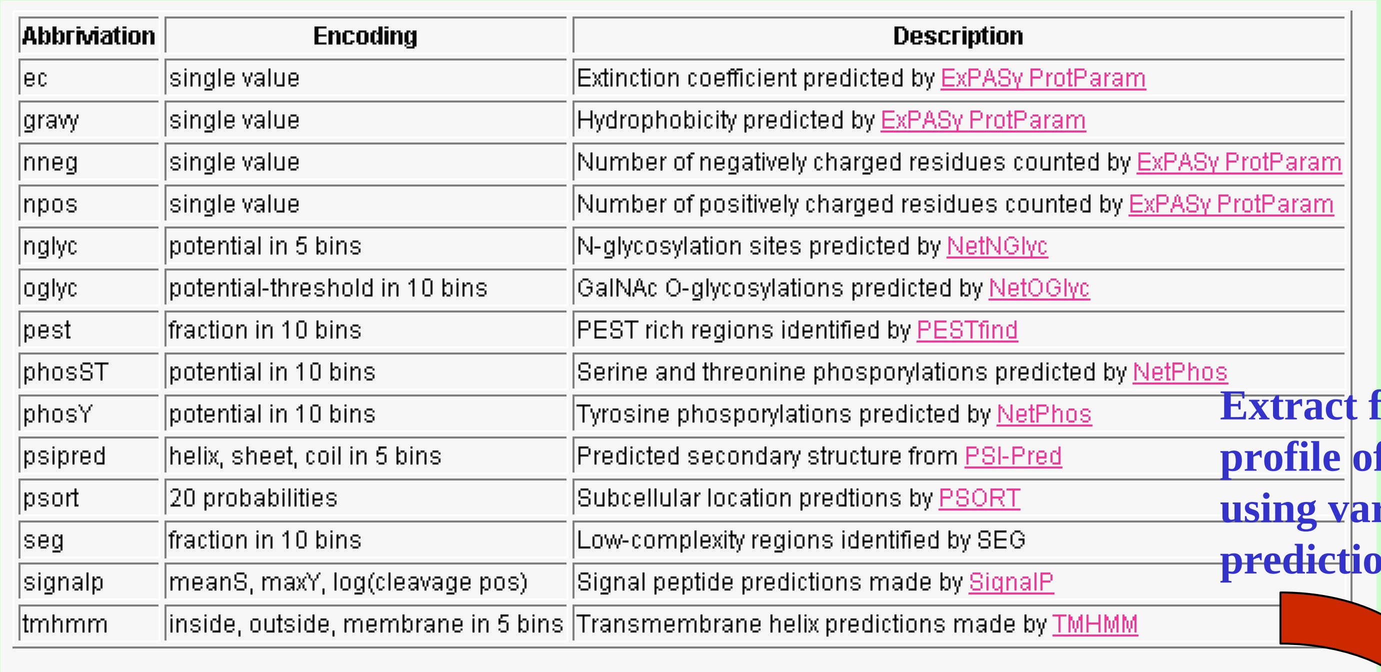
- **A protein is not alone when performing its biological function**
- **It operates using the same cellular machinery for modification and sorting as all other proteins do, such as glycosylation, phospharylation, signal peptide cleavage, …**
- **These have associated consensus motifs, patterns, etc.**



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

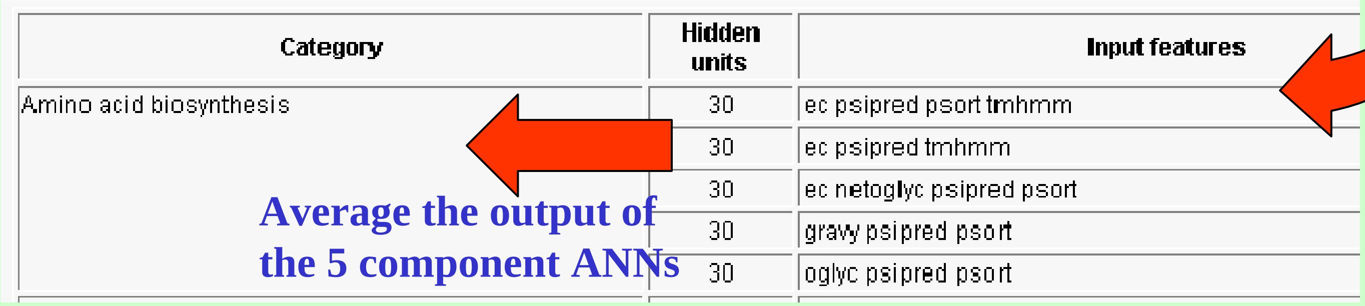
⇒ **Perhaps we can predict protein function by comparing its "feature" profile with other proteins?**

ProtFun: How it Works



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

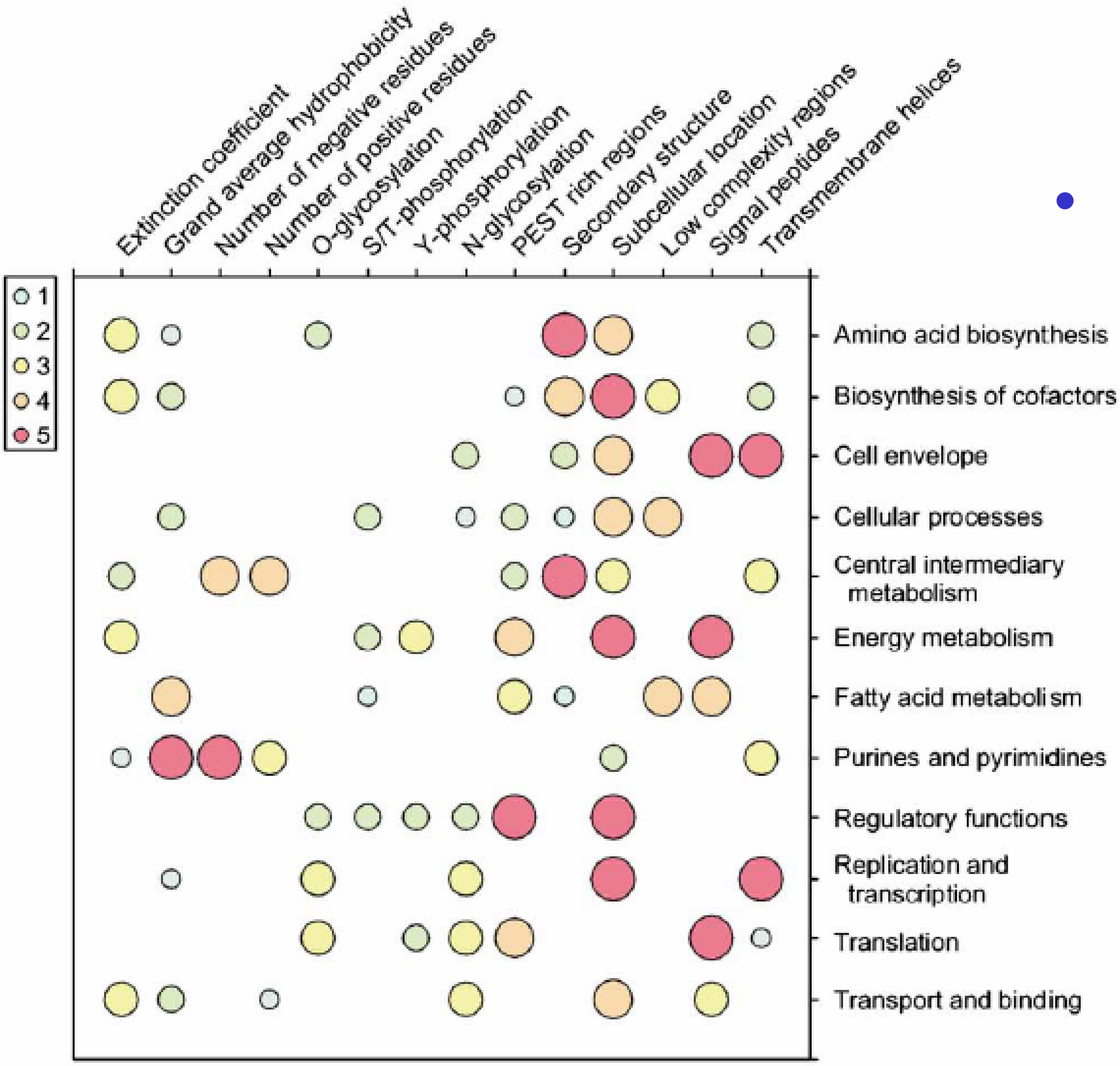
Extract feature profile of protein using various prediction methods



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

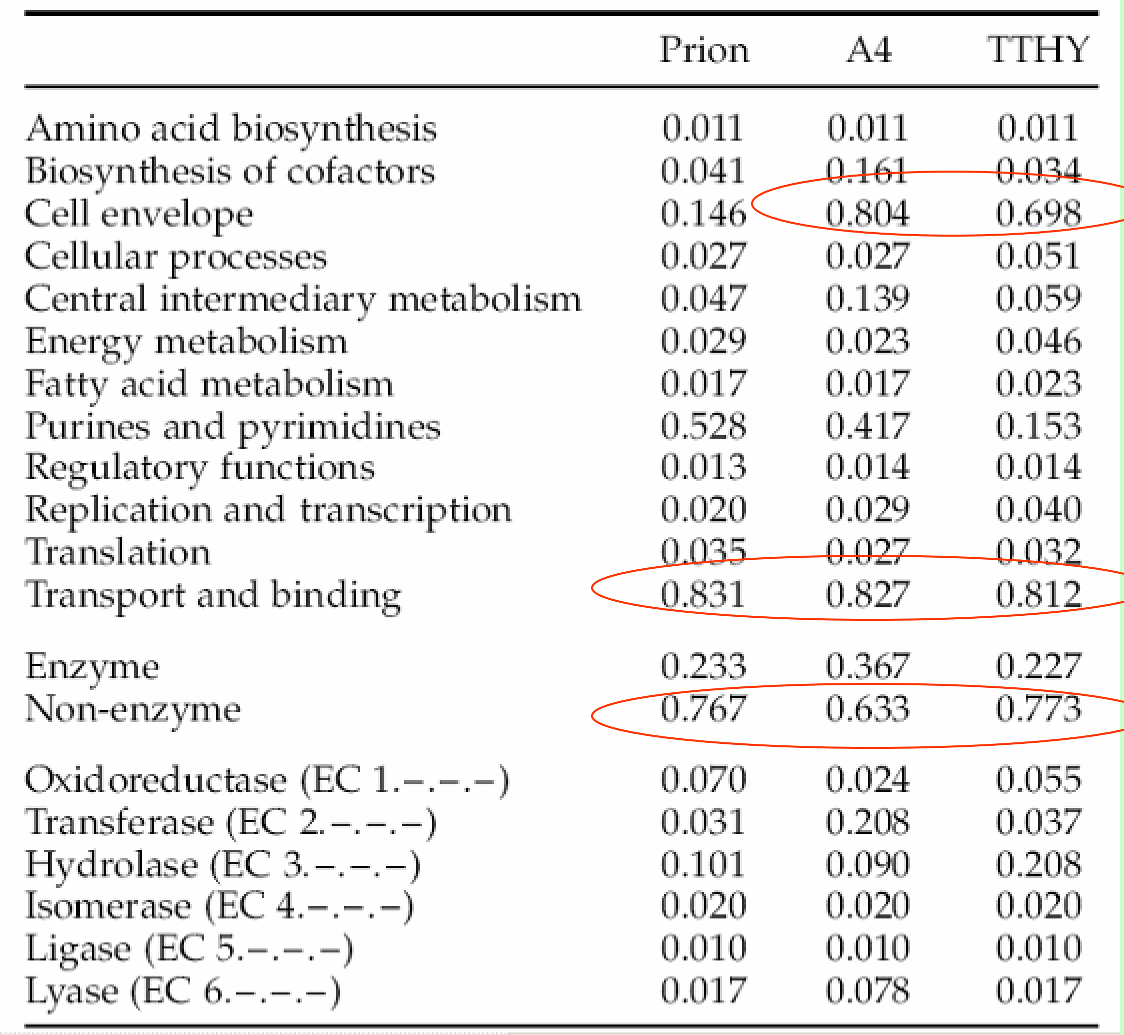
Average the output of the 5 component ANNs

ProtFun: Evidence



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

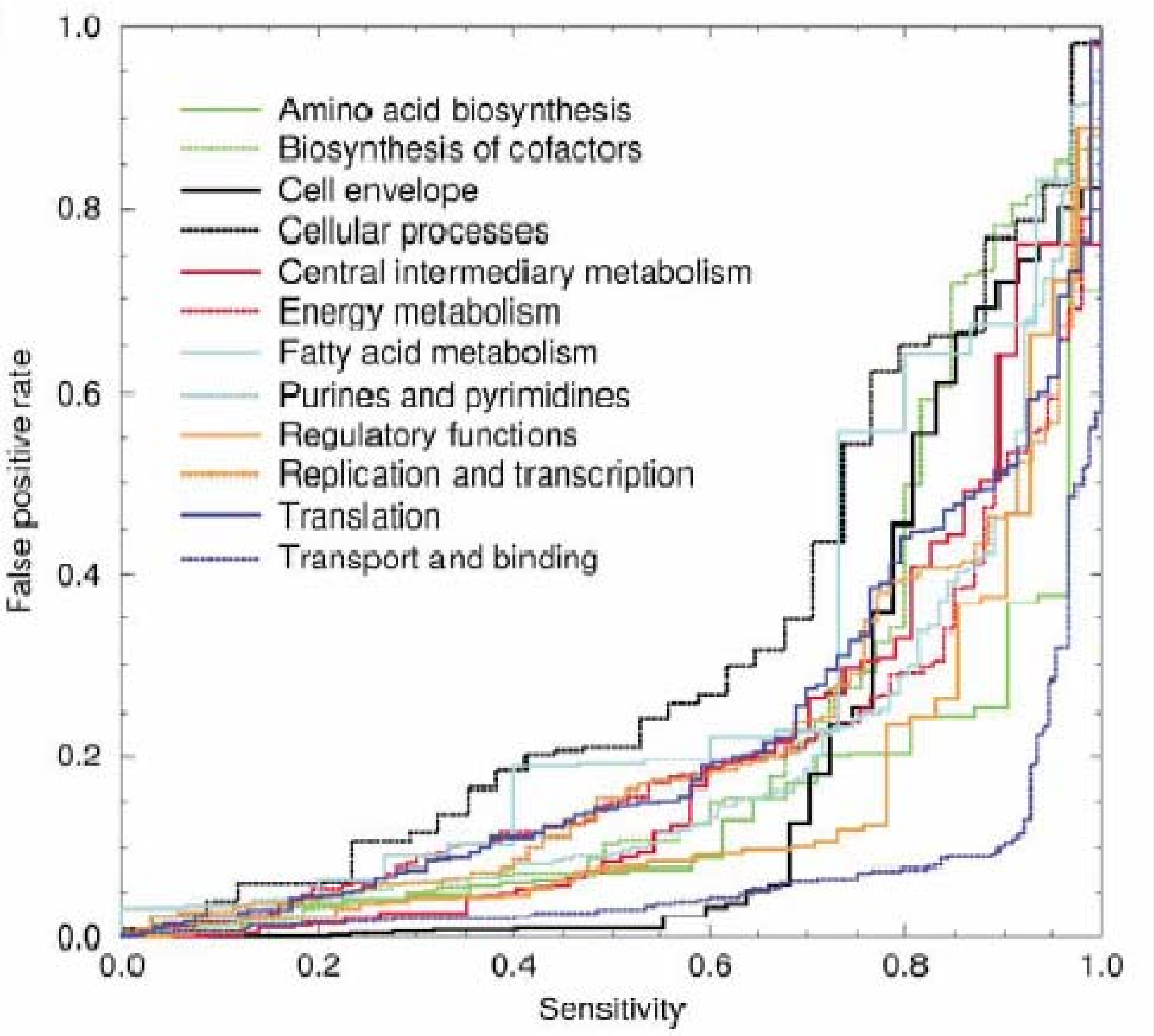
ProtFun: Example Output



	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**

ProtFun: Performance



Similarity of Dissimilarities

	orange$_1$	banana$_1$	…

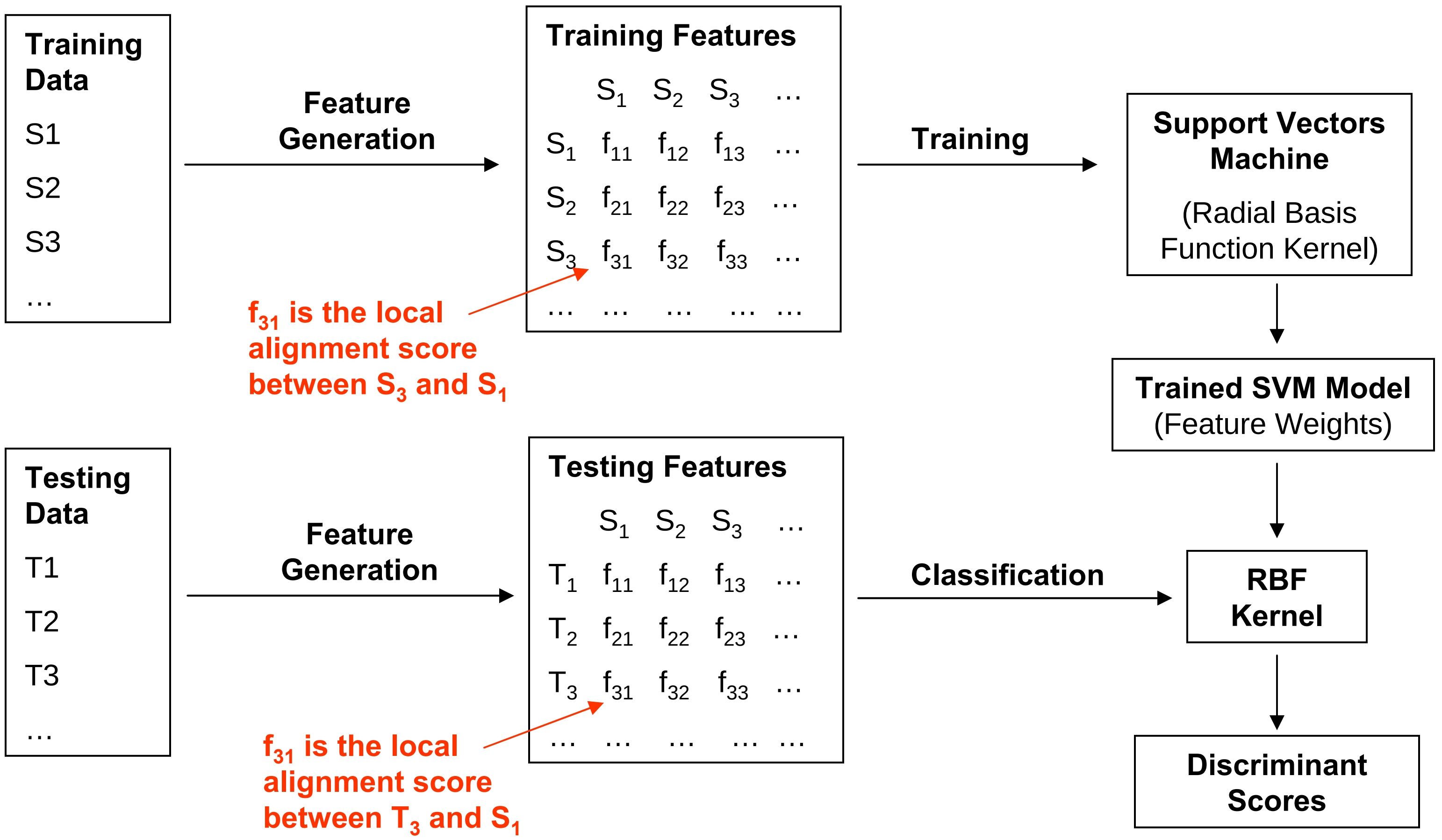
apple$_1$	Color = red vs orange
	Skin = smooth vs rough
	Size = small vs small
	Shape = round vs round
		Color = red vs yellow
		Skin = smooth vs smooth
		Size = small vs small
		Shape = round vs oblong
			…

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
			..

…	…	…	…

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}	…
…	…	…	…	…

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

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}	…
…	…	…	…	…

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

Classification

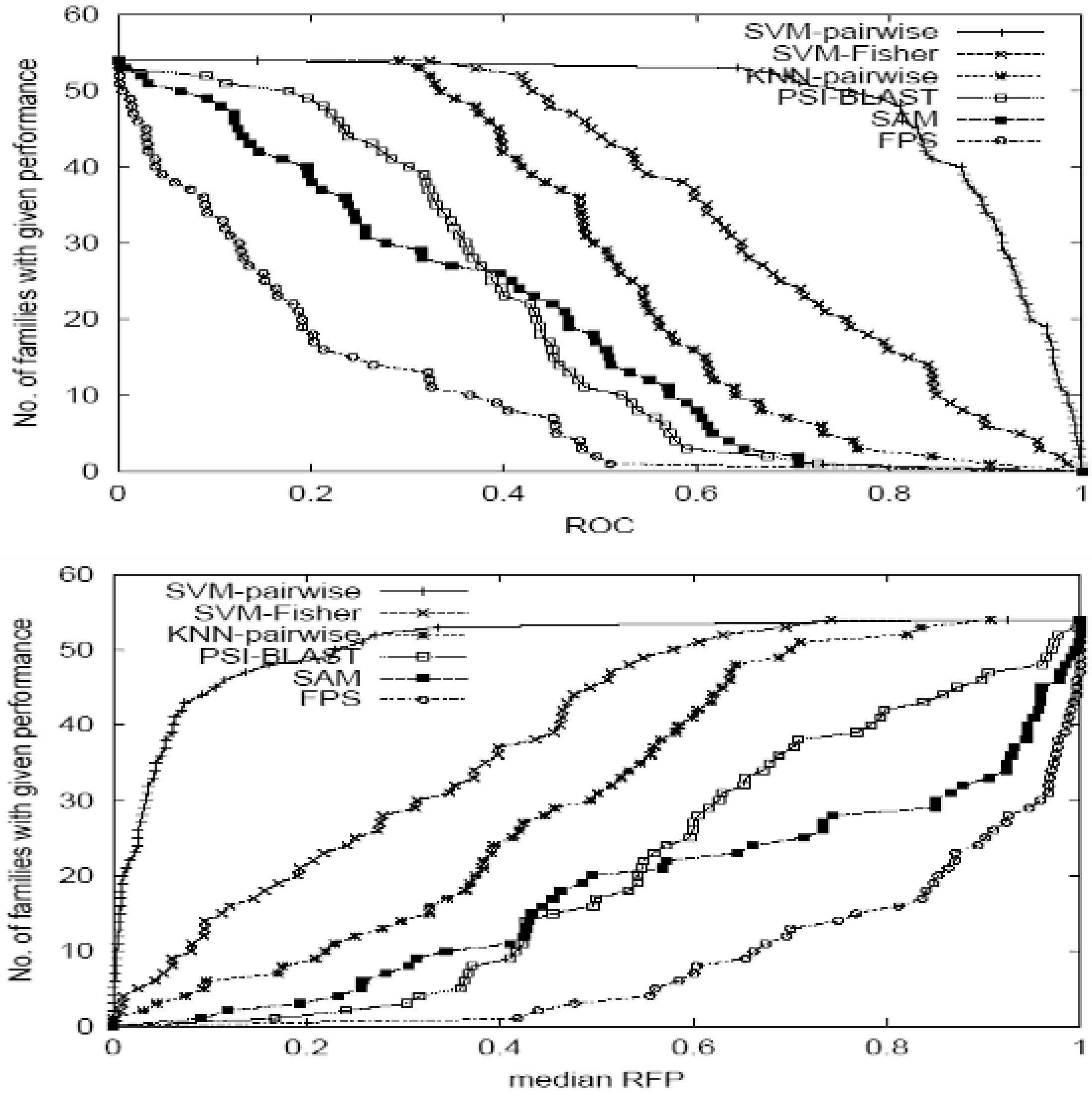
RBF Kernel

Discriminant Scores

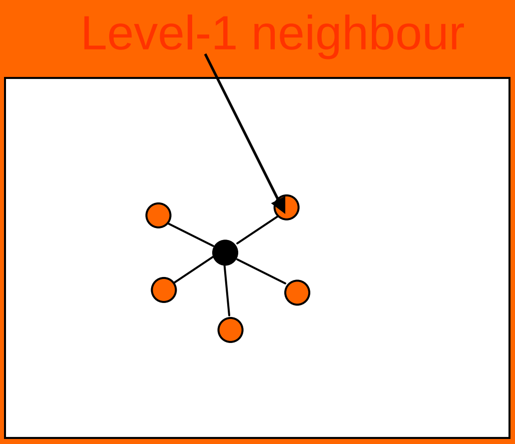
Image credit: Kenny Chua

Performance of SVM-Pairwise

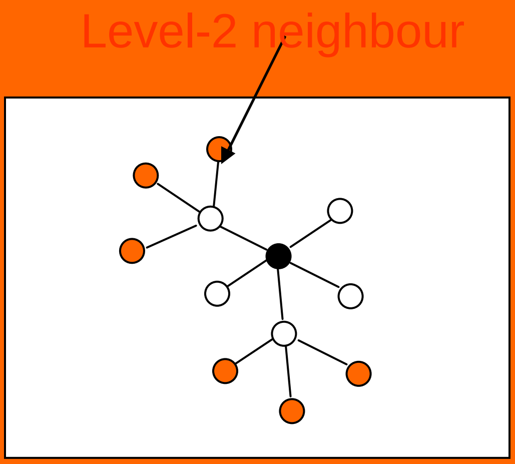
- **Receiver Operating Characteristic (ROC)**
  - The area under the curve derived from plotting true positives as a function of false positives for various thresholds.
- **Rate of median False Positives (RFP)**
  - The fraction of negative test examples with a score better or equals to the median of the scores of positive test examples.



Protein Function Prediction from Protein Interactions



Level-1 neighbour

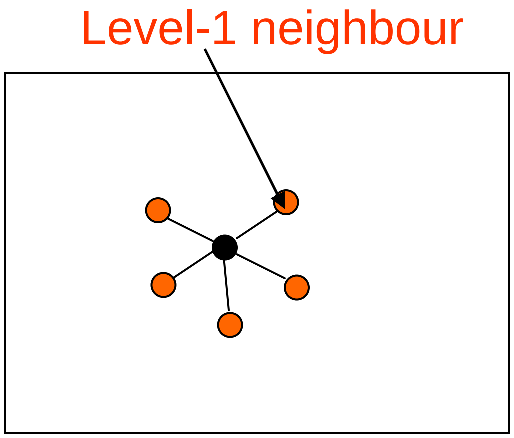


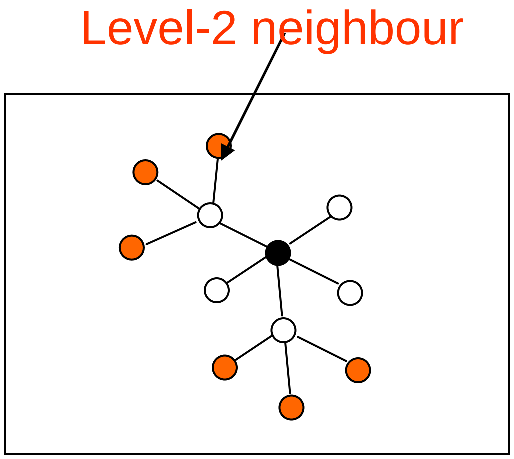
Level-2 neighbour

NUS
National University
of Singapore

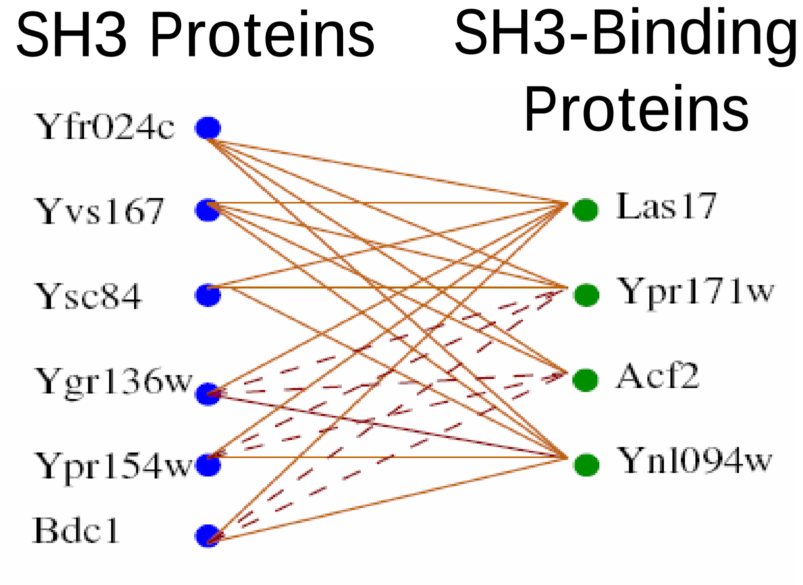
Functional Association Thru Interactions

- **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



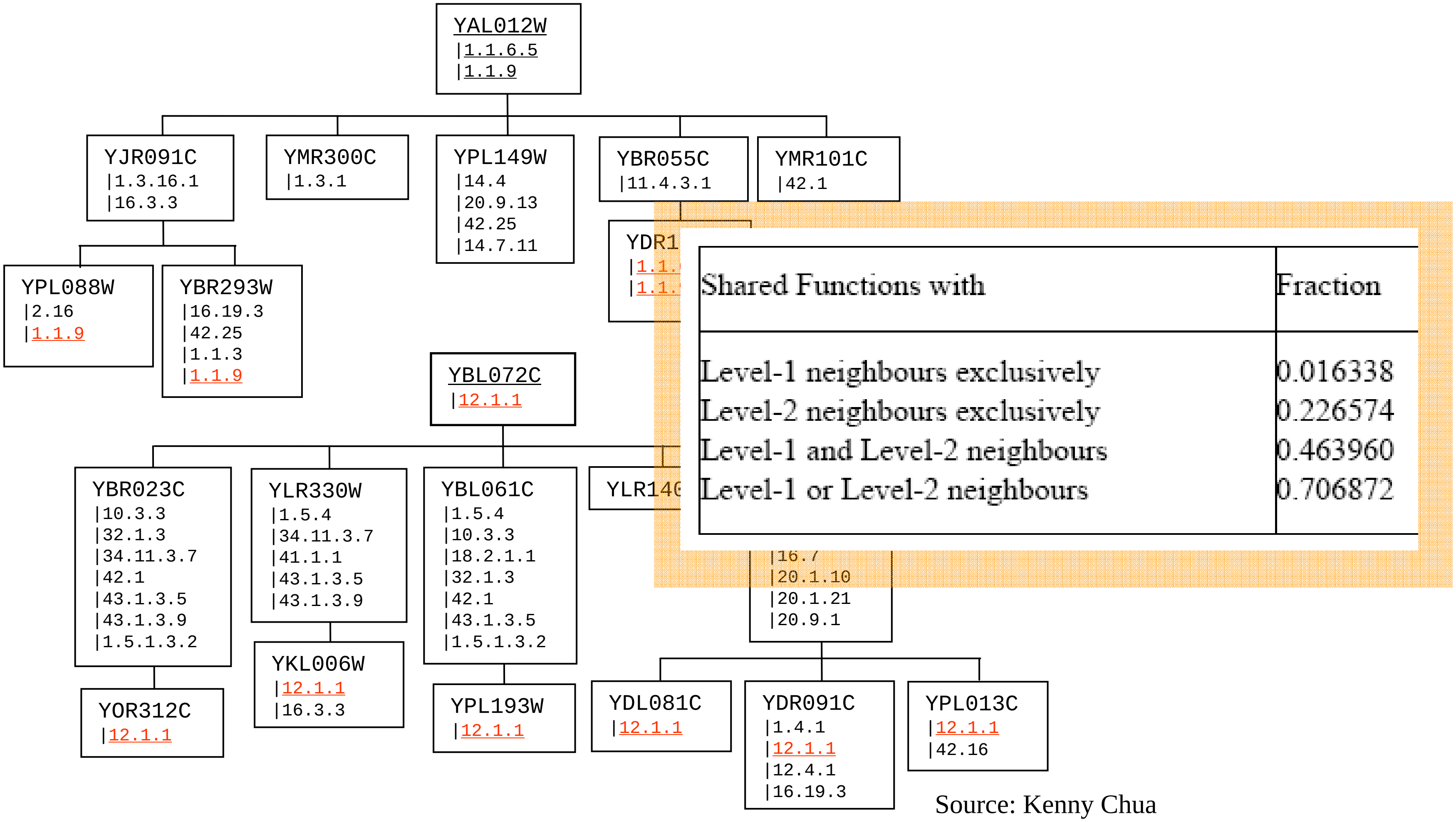


An illustrative Case of Indirect Functional Association?



- 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?

Freq of Indirect Functional Association



YAL012W |1.1.6.5 |1.1.9
YJR091C |1.3.16.1 |16.3.3
YMR300C |1.3.1
YPL149W |14.4 |20.9.13 |42.25 |14.7.11
YBR055C |11.4.3.1
YMR101C |42.1
YDR1 |1.1. |1.1.
YPL088W |2.16 |1.1.9
YBR293W |16.19.3 |42.25 |1.1.3 |1.1.9
YBL072C |12.1.1
YBR023C |10.3.3 |32.1.3 |34.11.3.7 |42.1 |43.1.3.5 |43.1.3.9 |1.5.1.3.2
YLR330W |1.5.4 |34.11.3.7 |41.1.1 |43.1.3.5 |43.1.3.9
YBL061C |1.5.4 |10.3.3 |18.2.1.1 |32.1.3 |42.1 |43.1.3.5 |1.5.1.3.2
YLR140
|16.7 |20.1.10 |20.1.21 |20.9.1
YOR312C |12.1.1
YKL006W |12.1.1 |16.3.3
YPL193W |12.1.1
YDL081C |12.1.1
YDR091C |1.4.1 |12.1.1 |12.4.1 |16.19.3
YPL013C |12.1.1 |42.16

Shared Functions with	Fraction
Level-1 neighbours exclusively	0.016338
Level-2 neighbours exclusively	0.226574
Level-1 and Level-2 neighbours	0.463960
Level-1 or Level-2 neighbours	0.706872

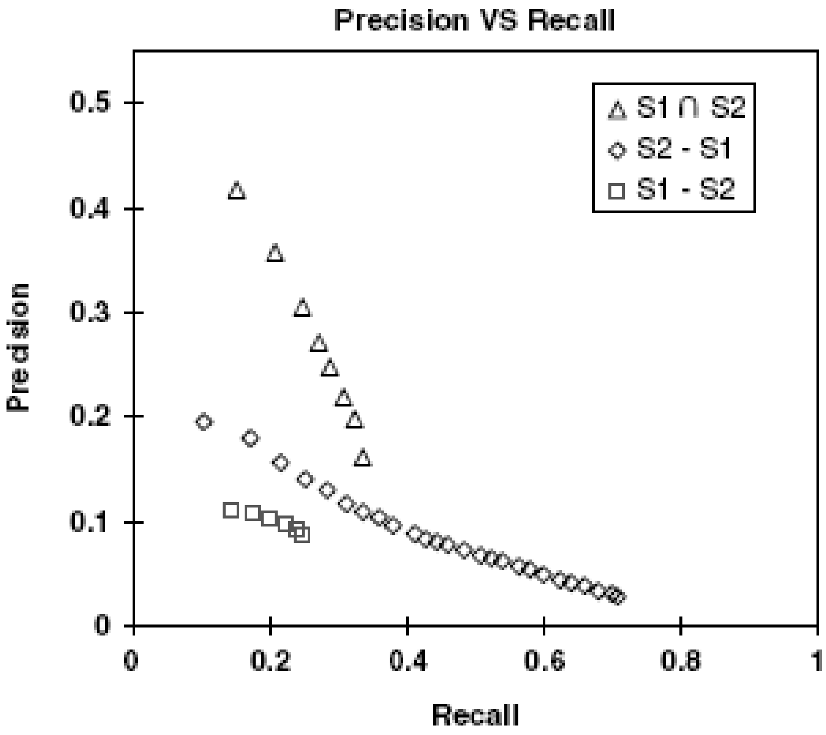
Source: Kenny Chua

Prediction Power By Majority Voting

- Remove overlaps in level-1 and level-2 neighbours to study predictive power of "level-1 only" and "level-2 only" neighbours
- Sensitivity vs Precision analysis

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

  - n_i is no. of fn of protein i
  - m_i is no. of fn predicted for protein i
  - k_i is no. of fn predicted correctly for protein i



⇒ "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 Δ Y is symmetric diff betw two sets X and Y**
  - **Greater weight given to similarity**

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

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

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

Functional Similarity Estimate: FS-Weighted Measure

- **FS-weighted measure**

$$S(u,v)=\frac{2|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v|} \times \frac{2|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v|}$$

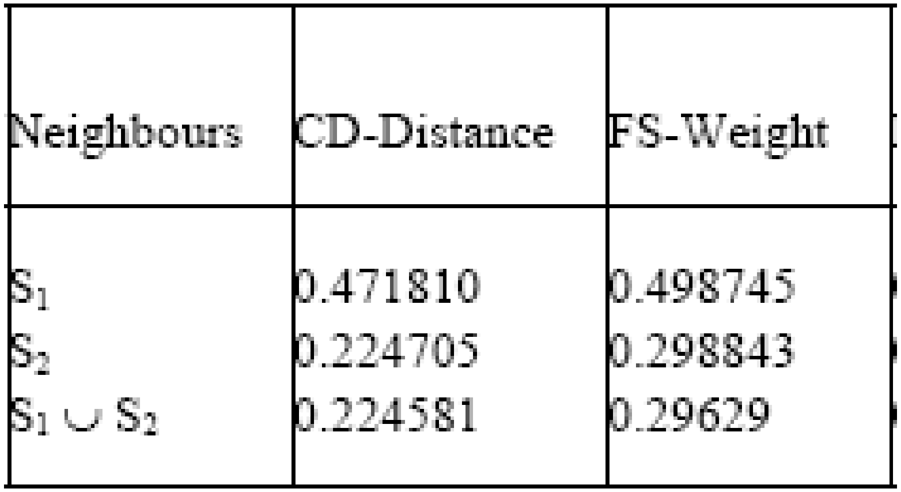
  - **N_k is the set of interacting partners of k**
  - **Greater weight given to similarity**

$\Rightarrow$ **Rewriting this as**

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

Correlation w/ Functional Similarity

- **Correlation betw functional similarity & estimates**



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

- **Equiv measure slightly better in correlation w/ similarity for L1 & L2 neighbours**

Reliability of Expt Sources

- **Diff Expt Sources have diff reliabilities**
  - Assign reliability to an interaction based on its expt sources (Nabieva et al, 2004)
- **Reliability betw u and v computed by:**

$$r_{u,v} = 1 - \prod_{i \in E_{u,v}} (1 - r_i)$$

  - **r_i is reliability of expt source i,**
  - **$E_{u,v}$ is the set of expt sources in which interaction betw u and v is observed**

Source	Reliability
Affinity Chromatography	**0.823077**
Affinity Precipitation	**0.455904**
Biochemical Assay	**0.666667**
Dosage Lethality	**0.5**
Purified Complex	**0.891473**
Reconstituted Complex	**0.5**
Synthetic Lethality	**0.37386**
Synthetic Rescue	**1**
Two Hybrid	**0.265407**

Functional Similarity Estimate: FS-Weighted Measure with Reliability

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

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

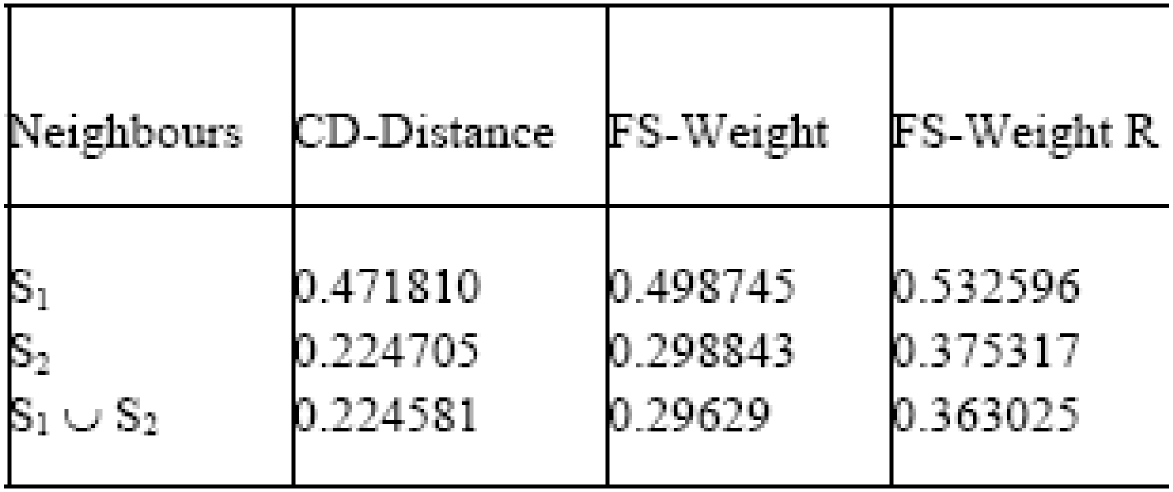
  - **N_k is the set of interacting partners of k**
  - **$r_{u,w}$ is reliability weight of interaction betw u and v**

$\Rightarrow$ **Rewriting**

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

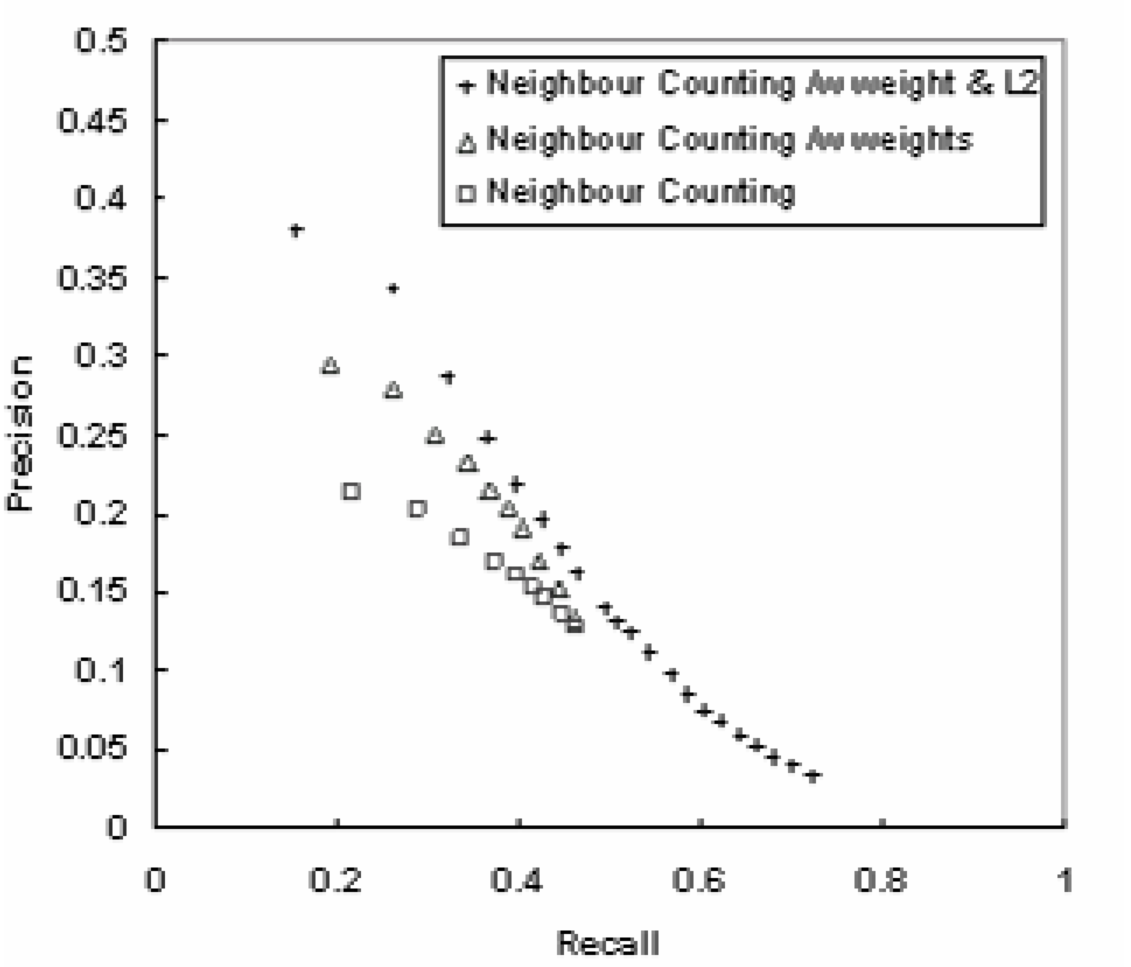
Integrating Reliability

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



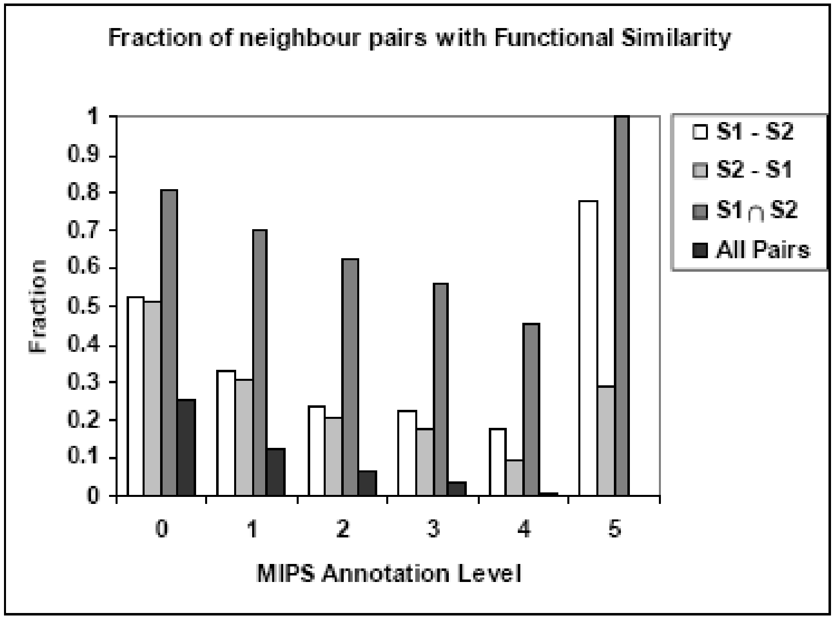
Neighbours	CD-Distance	FS-Weight	FS-Weight R
S_1	0.471810	0.498745	0.532596
S_2	0.224705	0.298843	0.375317
$S_1 \cup S_2$	0.224581	0.29629	0.363025

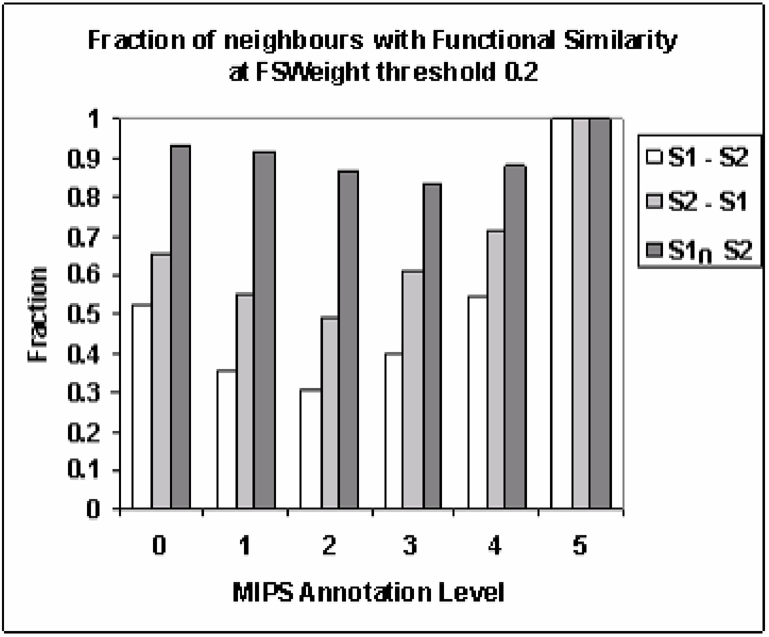
Improvement to Prediction Power by Majority Voting



Considering only neighbours w/ FS weight > 0.2

Improvement to
Over-Rep of Functions in Neighbours





Use L1 & L2 Neighbours for Prediction

- **FS-weighted Average**

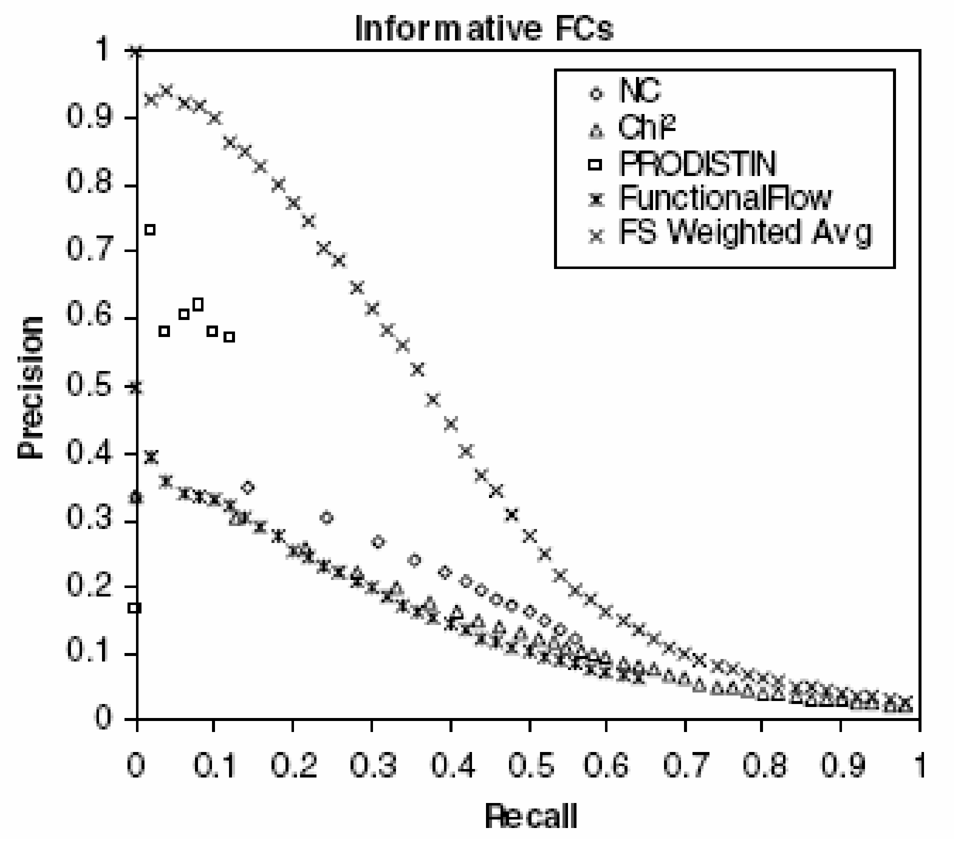
$$f_x(u) = \frac{1}{Z}\left[\lambda r_{\mathrm{int}} \pi_x + \sum_{v \in N_u}\left(S_{TR}(u,v)\delta(v,x) + \sum_{w \in N_v} S_{TR}(u,w)\delta(w,x)\right)\right]$$

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

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

Performance of FS-Weighted Averaging

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



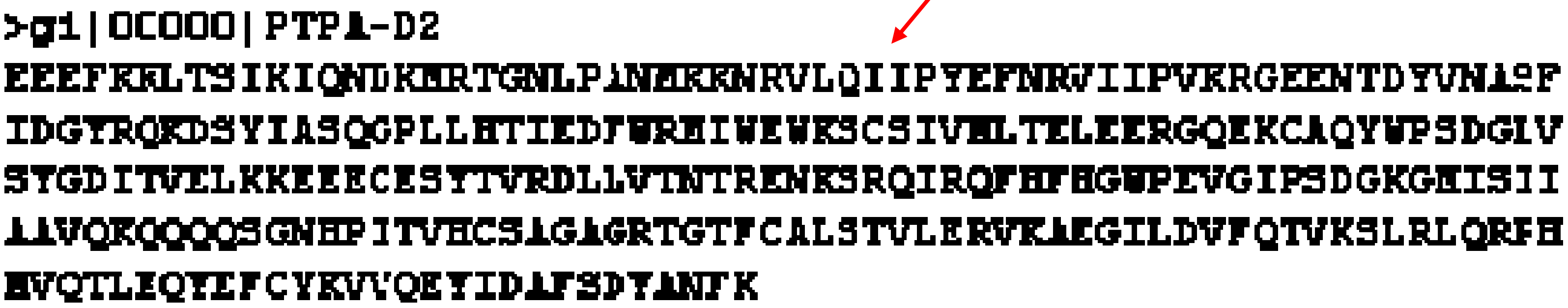
Application of
Sequence Comparison:
Key Mutation Site Discovery

National University of Singapore

Identifying Key Mutation Sites

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

Sequence from a typical PTP domain D2



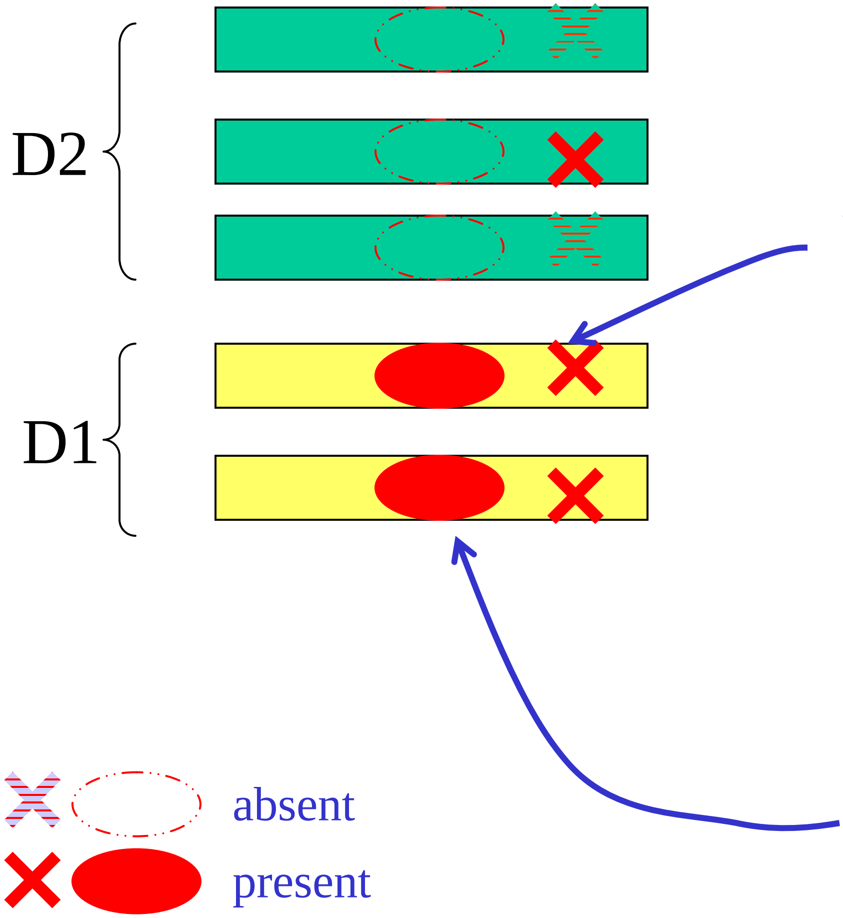
```
>gi|0C000|PTPA-D2
EEEFKKLTSIKIQNDKMRTGNLPANMKKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASF
IDGYRQKDSYIASQGPLLHTIEDFWRMIWEWKSCSIVMLTELEERGQEKCAQYWPSDGLV
SYGDITVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFHGWPEVGIPSDGKGMISII
AAVQKQQQQSGNHPITVHCSAGAGRTGTFCALSTVLERVKAEGILDVFQTVKSLRLQRPH
MVQTLEQYEFCYKVVQEYIDAFSDYANFK
```

- **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

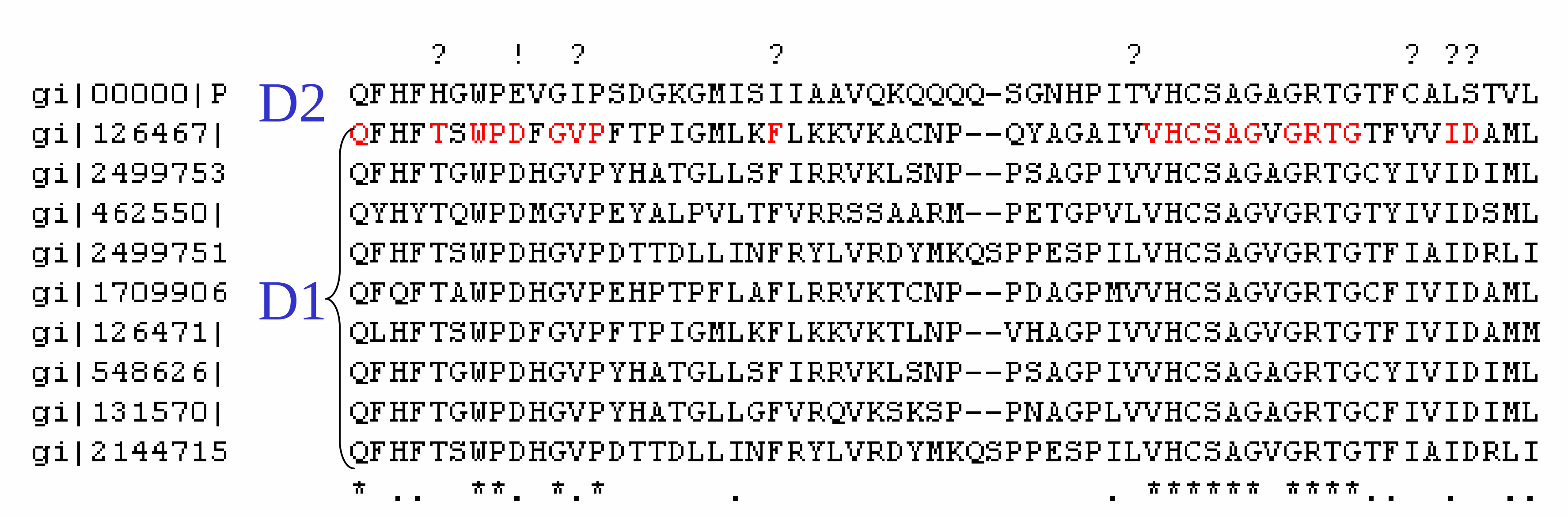


This site is consistently conserved in D1,
but is not consistently missing in D2
⇒ it is not an EP
⇒ not a likely cause of D2's loss of function

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

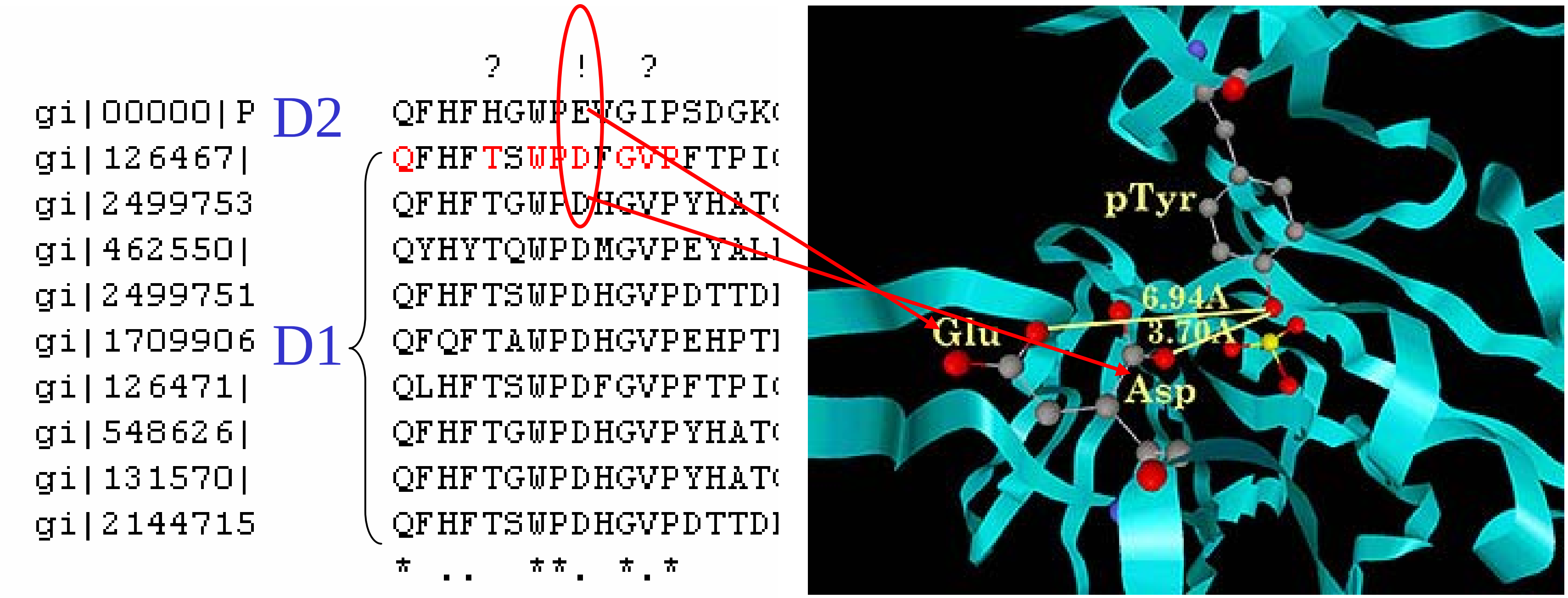
Key Mutation Site: PTP D1 vs D2



```
                    ?   !  ?         ?                 ?             ? ??
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 Questions?

NUS
National University of Singapore

Acknowledgements

- **Some of the slides are based on slides given to me by Kenny Chua**

References

- T.F.Smith & X.Zhang. “The challenges of genome sequence annotation or `The devil is in the details’”, *Nature Biotech*, 15:1222--1223, 1997
- D. Devos & A.Valencia. “Intrinsic errors in genome annotation”, *TIG*, 17:429--431, 2001
- K.L.Lim et al. “Interconversion of kinetic identities of the tandem catalytic domains of receptor-like protein tyrosine phosphatase PTP-alpha by two point mutations is synergist and substrate dependent”, *JBC*, 273:28986--28993, 1998
- 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

References

- S.E.Brenner. “Errors in genome annotation”, *TIG*, 15:132--133, 1999
- 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
- L.J.Jensen et al. “Prediction of human protein function from post-translational modifications and localization features”, *JMB*, 319:1257--1265, 2002
- C. Wu, W. Barker. “A Family Classification Approach to Functional Annotation of Proteins”, *The Practical Bioinformatician*, Chapter 19, pages 401—416, WSPC, 2004

References

- H.N. Chua, W.-K. Sung. A better gap penalty for pairwise SVM. Proc. APBC05, pages 11-20
- Hon Nian Chua, Wing Kin Sung, Limsoon Wong. Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions. *Bioinformatics*, 22:1623-1630, 2006.
- T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote homologies. *JCB*, 7(1-2):95—11, 2000