

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

# KI1972: Applied Bioinformatics & Computational Biology

# Sequence Homology Interpretation

**Limsoon Wong**

**June 2006**



**Limsoon  
Wong**  
Professor of  
Computing and  
Medicine  
NUS



<http://www.comp.nus.edu.sg/~wongls>

- **Education**

- BSc (Eng) (Computing) Imperial College, 1988
- PhD (Comp & Info Sci), Univ of Penn, 1994

- **Research**

- Query languages, knowledge discovery, bioinformatics
- 5 books, ~100 articles, ~100 keynote & invited lectures, 2 patents, 4 techs commercialised

- **Professional Activities**

- Chairman, Molecular Connections, India
- International Panel, National Research Program on Genomic Medicine, Taiwan
- Managing Editor, *Journal of Bioinformatics and Computational Biology*, ICP
- Editor, *Bioinformatics*, OUP
- Editor, *Drug Discovery Today*, Elsevier

- **Honours**

- 2004 Ranked as 40<sup>th</sup> Best Nurturer of Computer Science Research, among >50k Comp Sci researchers worldwide indexed by DBLP
- 2003 Far Eastern Economic Review Asian Innovation Gold Award for “a simple test for childhood ALL that promises safer treatment and higher cure rates for kids in the developing world”



# Plan

- **Recap of sequence alignment**
- **Guilt by association**
- **Active site/domain discovery**
- **Key mutation site discovery**
  
- **Guilt by other types of association**
  - Genome phylogenetic profiling
  - Protfun
  - SVM-Pairwise

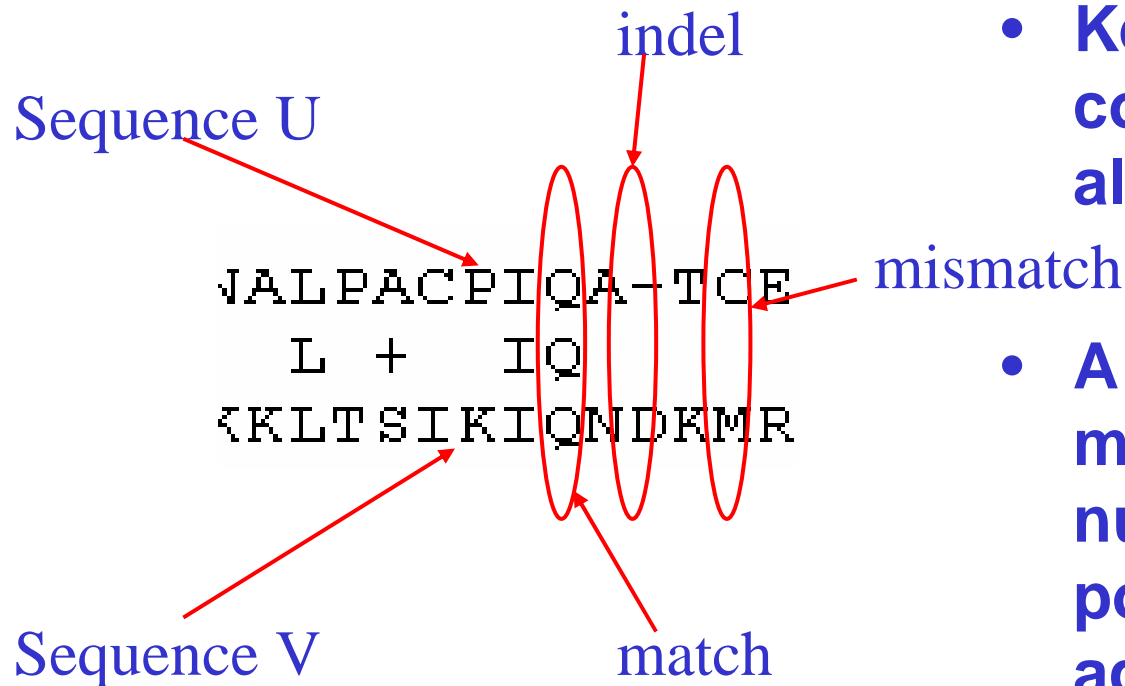
# Very Brief Recap of Sequence Comparison/Alignment



# Motivations for Seq 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	MPHNVHFVAGVLGEAALKGPMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVVE				
			...	.	...
Ascorbate Oxidase	ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLMQRSAGLYGSLI				
	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**

```

 >gil13476732|ref|NP\_108301.1| unknown protein [Mesorhizobium loti]
gil14027493|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  MKPGRLASIALAIIFLPMVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 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|      FHFTSWPDFGVPFTP I GMLKFLK KVKACNP--QYAGAI VVHCSAGVGR TGTFVVIDAMLD
gi|2499753     FHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPI VVHCSAGAGRTGCYIVIDIMLD
gi|462550|     YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVI VVHCSAGVGR TGTYIVIDSMLQ
gi|2499751     FHFTSWPDHGVPD TTDLLINFRYLVRDYMKQSPPE SPILVHCSAGVGR TGTFIAIDRLIY
gi|1709906     FQFTA WPDHGVP EHP T PFLAFLRRVKTCNP--PDAGPM VVHCSAGVGR TGCFIVIDAMLE
gi|126471|     LHFTSWPDFGVPFTP I GMLKFLK KVKTLNP--VHAGPI VVHCSAGVGR TGTFIVIDAMMA
gi|548626|     FHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPI VVHCSAGAGRTGCYIVIDIMLD
gi|131570|     FHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPL VVHCSAGAGRTGCFIVIDIMLD
gi|2144715     FHFTSWPDHGVPD TTDLLINFRYLVRDYMKQSPPE SPILVHCSAGVGR TGTFIAIDRLIY
                ..*  ***  ***          .  *          ..*****  *****  ** ..

```

Conserved sites



# Application of Sequence Comparison: Guilt-by-Association



# Function Assignment to Protein Seq

SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR  
YVNILPYDHSRVHLTPVEGVPDSYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE  
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD  
VTNRKPQRLITQFHFTSWPDFGVPFTP I GMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG  
TFVVIDAMLDMMSERKVDVYGFVSRIRAQRCQMVQTD MQYVFIYQALLEHYLYGDTELE  
VT

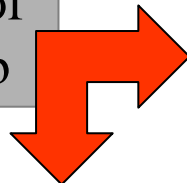
- How do we attempt to assign a function to a new protein sequence?

# Guilt-by-Association

- Compare the target sequence  $T$  with sequences  $S_1, \dots, S_n$  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

# Guilt-by-Association

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



## Poor Sequence Alignment

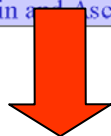
- Poor seq alignment shows few matched positions  
⇒ The two proteins are not likely to be homologous

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

```

Amicyanin      60      70      80      90     100
MPHNVHFVAGVLGEAALKGPMMKKEQAYSLSLTFTEAGTYDYHCTPHPFMRGKVVV
                . . . . .
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFYVDNPGTFFYHGHLMQRSAGLYG
                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

```

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

good match between Amicyanin and unknown M. loti protein



Assign to  $T$  same function as homologs

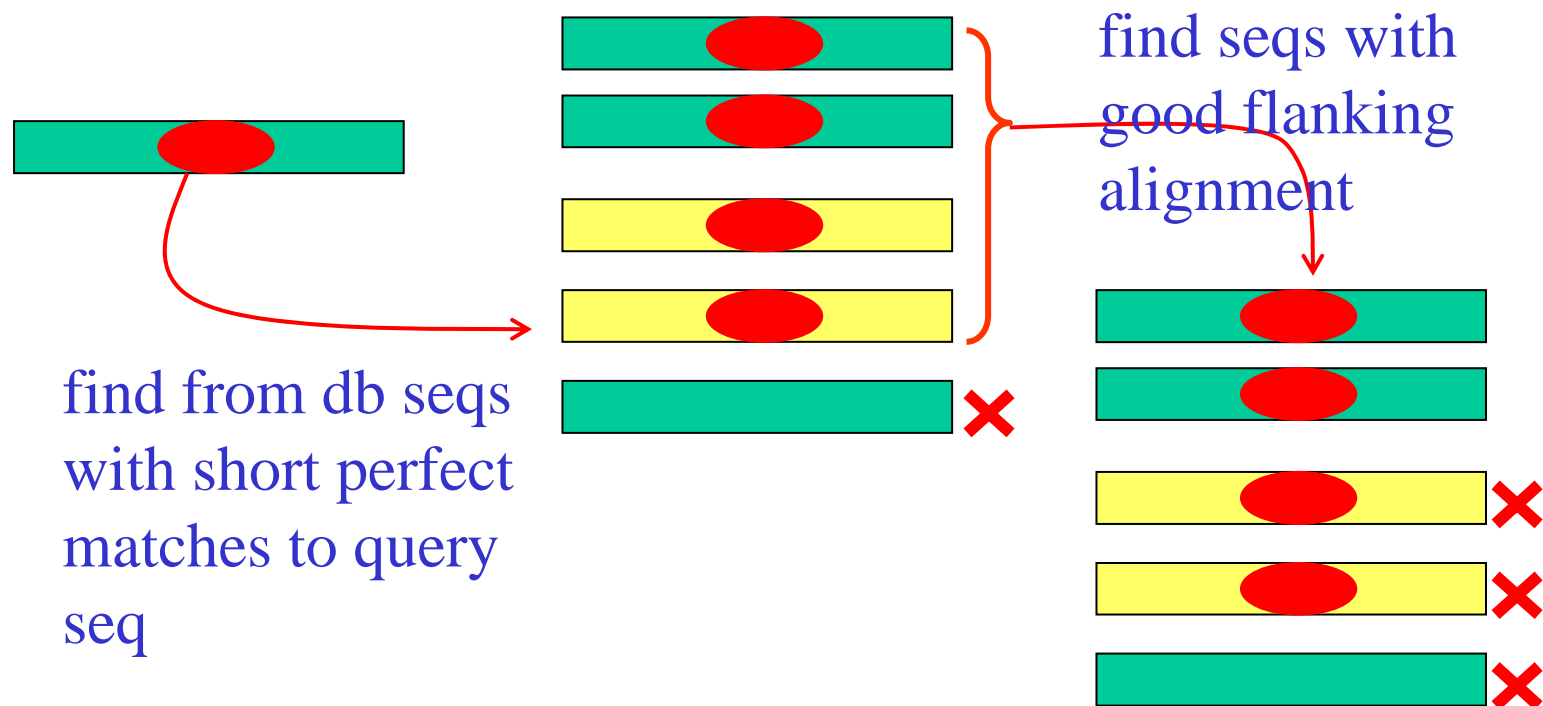


Confirm with suitable wet experiments

# 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



NCBI BLAST - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Home Mail Print Mail News RSS AutoLink

Address <http://www.ncbi.nlm.nih.gov/BLAST/>

Google Search PageRank 6 blocked Check AutoLink

NCBI → Latest news: 6 December 2005 : BLAST 2.2.13 released

## BLAST

About

- Getting started
- News
- FAQs

More info

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software

- Downloads
- Developer info

Other resources

- References
- NCBI Contributors
- Mailing list
- Contact us

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

<p><b>Nucleotide</b></p> <ul style="list-style-type: none"> <li>Quickly search for highly similar sequences (<a href="#">megablast</a>)</li> <li>Quickly search for divergent sequences (<a href="#">discontiguous megablast</a>)</li> <li>Nucleotide-nucleotide BLAST (<a href="#">blastn</a>)</li> <li>Search for short, nearly exact matches</li> <li>Search trace archives with <a href="#">megablast</a> or <a href="#">discontiguous megablast</a></li> </ul>	<p><b>Protein</b></p> <ul style="list-style-type: none"> <li>Protein-protein BLAST (<a href="#">blastp</a>)</li> <li>Position-specific iterated and pattern-hit initiated BLAST (<a href="#">PSI- and PHI-BLAST</a>)</li> <li>Search for short, nearly exact matches</li> <li>Search the conserved domain database (<a href="#">rpsblast</a>)</li> <li>Protein homology by domain architecture (<a href="#">cdart</a>)</li> </ul>
<p><b>Translated</b></p> <ul style="list-style-type: none"> <li>Translated query vs. protein database (<a href="#">blastx</a>)</li> <li>Protein query vs. translated database (<a href="#">tblastn</a>)</li> <li>Translated query vs. translated database (<a href="#">tblastx</a>)</li> </ul>	<p><b>Genomes</b></p> <ul style="list-style-type: none"> <li>Human, mouse, rat, chimp, cow, pig, dog, sheep, cat</li> <li>Chicken, puffer fish, zebrafish</li> <li>Fly, honey bee, other insects</li> <li>Microbes, environmental samples</li> <li>Plants, nematodes</li> <li>Fungi, protozoa, other eukaryotes</li> </ul>

start 2 Window... nus-ki-nov0... Microsoft P... NCBI BLAST... Type to search

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Mail Print Mailbox Address Book

Address [http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO\\_FORMAT=Semiauto&ALIGNMENT=](http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENT=)

Google Search PageRank 6 blocked Check AutoLink



Nucleotide

Protein

Translations

Retrieve results for an RID

*protein-protein* **BLAST**

[Search](#)

```

NRYVNIILPYDHSRVHLTPVEGVPSDYINASFINGYQEKKNFIAAQGPKEETVNDFWR
MIWEQNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFC
IQQVGDVITNRKPQLITQFHFTSWPDFGVPFTP I GMLKFLKKVKACNPQYAGAIVVHC
SAGVGRGTGFVVVIDAMLDMMHSEKVDVYGFVSRIRAQRCQMVTDMQYVFIYQALLE
HYLYGDTELE

```

[Set subsequence](#)

 From:  To: 
[Choose database](#)

[Do CD-Search](#)


Now:

or



### Options for advanced blasting

[Limit by entrez query](#)


or select from:

[Compositional](#)



Back Search Favorites PageRank 6 blocked Check AutoLink AutoFill Options

Address <http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi>



Nucleotide

Protein

Translations

Retrieve results for an RID

## formatting BLAST

Your request has been successfully submitted and put into the Blast Queue.

Query = (302 letters)

Putative conserved domains have been detected, click on the image below for detailed results.



The request ID is

or

The results are estimated to be ready in 9 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again request results of a different search by entering any other valid request ID to see other recent jobs.

# Homologs obtained by BLAST

Sequences producing significant alignments:	Score (bits)	E Value
<a href="#">gi 14193729 gb AAK56109.1 AF332081_1</a> protein tyrosin phosph...	<a href="#">62</a> <sup>L</sup>	e-177
<a href="#">gi 126467 sp P18433 PTRA_HUMAN</a> Protein-tyrosine phosphatase...	<a href="#">62</a> <sup>L</sup>	e-177
<a href="#">gi 4506303 ref NP_002827.1</a> protein tyrosine phosphatase, r...	<a href="#">62</a> <sup>L</sup>	e-176
<a href="#">gi 227294 prf  1701300A</a> protein Tyr phosphatase	<a href="#">620</a>	e-176
<a href="#">gi 18450369 ref NP_543030.1</a> protein tyrosine phosphatase, ...	<a href="#">62</a> <sup>L</sup>	e-176
<a href="#">gi 32067 emb CAA37447.1</a> tyrosine phosphatase precursor [Ho...	<a href="#">61</a> <sup>L</sup>	e-176
<a href="#">gi 285113 pir  JC1285</a> protein-tyrosine-phosphatase (EC 3.1....	<a href="#">619</a>	e-176
<a href="#">gi 6981446 ref NP_036895.1</a> protein tyrosine phosphatase, r...	<a href="#">61</a> <sup>L</sup>	e-176
<a href="#">gi 2098414 pdb 1YFO A</a> Chain A, Receptor Protein Tyrosine Ph...	<a href="#">61</a> <sup>S</sup>	e-174
<a href="#">gi 32313 emb CAA38662.1</a> protein-tyrosine phosphatase [Homo...	<a href="#">61</a> <sup>L</sup>	e-174
<a href="#">gi 450583 gb AAB04150.1</a> protein tyrosine phosphatase >gi 4...	<a href="#">605</a>	e-172
<a href="#">gi 6679557 ref NP_033006.1</a> protein tyrosine phosphatase, r...	<a href="#">60</a> <sup>L</sup>	e-172
<a href="#">gi 483922 gb AAA17990.1</a> protein tyrosine phosphatase alpha	<a href="#">599</a>	e-170

- Thus our example sequence could be a protein tyrosine phosphatase  $\alpha$  (PTP $\alpha$ )

## Example Alignment with PTP $\alpha$

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

```

Query: 1   SPSTNRKYPPLPVDKLEEE INRRMADDNKLFREEFNALPACPIQATCEAASXXXXXXXXXR 60
          SPSTNRKYPPLPVDKLEEE INRRMADDNKLFREEFNALPACPIQATCEAAS      R
Sbjct: 202 SPSTNRKYPPLPVDKLEEE INRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR 261

Query: 61  YVNILPYDHSRVHLTPVEGVPSDYINASF INGYQEKNKF IAAQGPKEETVNDFWRMIWE 120
          YVNILPYDHSRVHLTPVEGVPSDYINASF INGYQEKNKF IAAQGPKEETVNDFWRMIWE
Sbjct: 262 YVNILPYDHSRVHLTPVEGVPSDYINASF INGYQEKNKF IAAQGPKEETVNDFWRMIWE 321

Query: 121 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD 180
          QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
Sbjct: 322 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD 381

Query: 181 VTNRKPQRLITQFHFTSWPDFGVPFITP IGMLKFLKKVKACNPQYAGAI VVHCSAGVGRTG 240
          VTNRKPQRLITQFHFTSWPDFGVPFITP IGMLKFLKKVKACNPQYAGAI VVHCSAGVGRTG
Sbjct: 382 VTNRKPQRLITQFHFTSWPDFGVPFITP IGMLKFLKKVKACNPQYAGAI VVHCSAGVGRTG 441

Query: 241 TFWVIDAMLDMHSEKVDVYGFVSRIRAQRCQMVQTD MQYVF IYQALLEHYLYGDTELE 300
          TFWVIDAMLDMHSEKVDVYGFVSRIRAQRCQMVQTD MQYVF IYQALLEHYLYGDTELE
Sbjct: 442 TFWVIDAMLDMHSEKVDVYGFVSRIRAQRCQMVQTD MQYVF IYQALLEHYLYGDTELE 501
  
```

# HSPs, E-Value, Bits, & P-Value

- **HSPs**

- A local alignment without gaps consists simply of a pair of equal length segments, one from each of the two sequences being compared.
- A segment pair whose score cannot be improved by extension or trimming is called high-scoring segment pairs or HSPs

- **E-Value**

- For large seq lengths  $m$  and  $n$ , the stats of HSP scores are characterized by two params,  $K$  and  $\lambda$
- Expected number of HSPs with score  $> S$  is given by  $E = Kmne^{-\lambda S}$

Source: NCBI

# HSPs, E-Value, Bit Score, & P-Value

- **Bit Score**

- “Citing a raw score alone is like citing a distance without specifying feet, meters, or light years”
- Normalize raw score to  $S' = (\lambda S - \ln K) / \ln 2$  to get "bit score“, which has a standard set of units
- E-value corresponds to bit score as  $E = mn2^{-S'}$

- **P-Value**

- Number of random HSPs with score  $\geq S$  is described by a Poisson distribution
- $\Rightarrow$  Chance of finding no HSPs with score  $\geq S$  is  $e^{-E}$
- $\Rightarrow$  **Prob of finding  $\geq 1$  such HSP is  $P = 1 - e^{-E}$**

Source: NCBI

# Guilt-by-Association: Caveats

- **Ensure that the effects of database size and composition have 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 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**



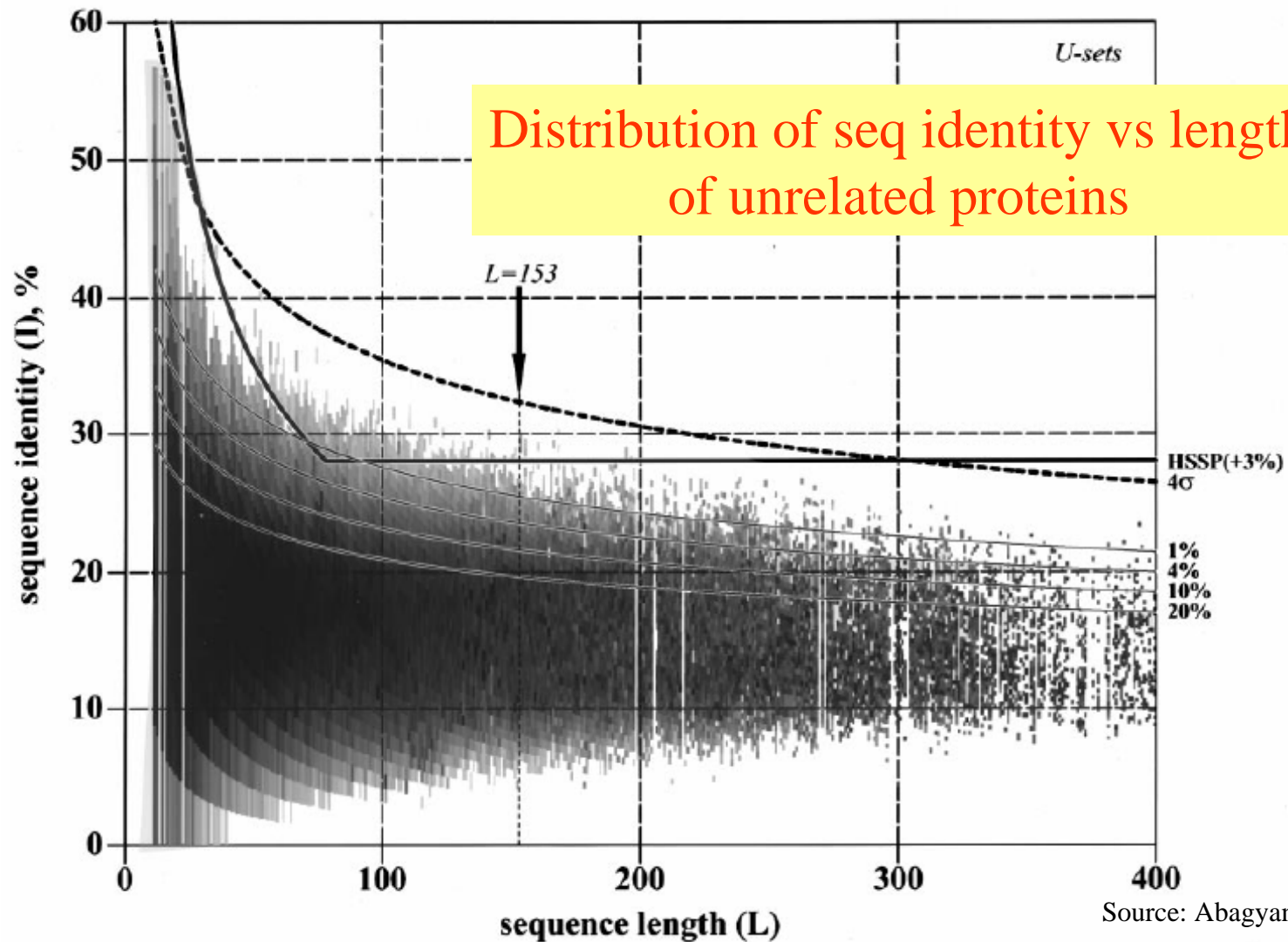
Cartoon: Ron Hipschman  
Data: David Hand

# 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**
- **BLAST employs the SEG algorithm to filter low complexity regions from proteins before executing a search**

Source: NCBI

# Effect of Sequence Length



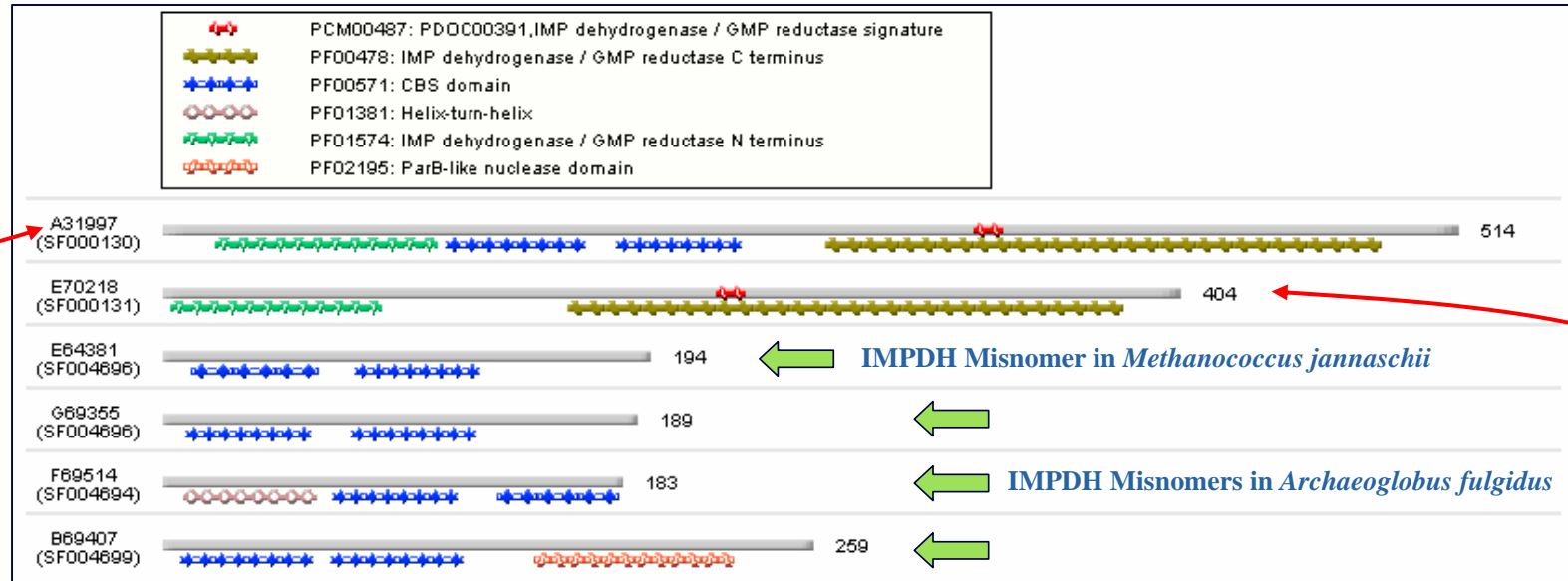
# Examples of Invalid Function Assignment: The IMP Dehydrogenases (IMPDH)

18 entries were found

ID	Organism	PIR	Swiss-Prot/TrEMBL	RefSeq/GenPept	
<a href="#">NF00181857</a>	Methanococcus jannaschii	<a href="#">E64381</a> conserved hypothetical protein MJ0653	<a href="#">Y653_METJA</a> Hypothetical protein MJ0653	<a href="#">g1592300</a> inosine-5'-monophosphate dehydrogenase (guaB) <a href="#">NP_247637</a> inosine-5'-monophosphate dehydrogenase (guaB)	
<a href="#">NF00187788</a>	Archaeoglobus fulgidus	<a href="#">G69355</a> MJ0653 homolog AF0847 <i>ALT_NAMES</i> : inosine-monophosphate dehydrogenase (guaB-1) homolog [misnomer]	<a href="#">O29411</a> INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)	<a href="#">g2649754</a> inosine monophosphate dehydrogenase (guaB-1) <a href="#">NP_069681</a> inosine monophosphate dehydrogenase (guaB-1)	
<a href="#">NF00188267</a>	Archaeoglobus fulgidus	<a href="#">F69514</a> yhcV homolog 2 <i>ALT_NAMES</i> : inosine-monophosphate dehydrogenase (guaB-2) homolog [misnomer]	<a href="#">O28162</a> INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-2)	<a href="#">g2648410</a> inosine monophosphate dehydrogenase (guaB-2) <a href="#">NP_070943</a> inosine monophosphate dehydrogenase (guaB-2)	
<a href="#">NF00188697</a>	Archaeo	<p style="text-align: center;"><b>A partial list of IMP dehydrogenase misnomers in complete genomes remaining in some public databases</b></p>			osphate ive nophosphate ive
<a href="#">NF00197776</a>	Thermo				nophosphate d protein nonophosphate d protein
<a href="#">NF00414709</a>	Methanothermobacter thermautotrophicus	<a href="#">G07630</a> M10055 homolog M111220 <i>ALT_NAMES</i> : inosine-monophosphate dehydrogenase related protein V [misnomer]	<a href="#">O27294</a> INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN V	dehydrogenase related protein V <a href="#">NP_276354</a> inosine-5'-monophosphate dehydrogenase related protein V	
<a href="#">NF00414811</a>	Methanothermobacter thermautotrophicus	<a href="#">D69035</a> MJ1232 protein homolog MTH126 <i>ALT_NAMES</i> : inosine-5'-monophosphate dehydrogenase related protein VII [misnomer]	<a href="#">O26229</a> INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN VII	<a href="#">g2621166</a> inosine-5'-monophosphate dehydrogenase related protein VII <a href="#">NP_275269</a> inosine-5'-monophosphate dehydrogenase related protein VII	
<a href="#">NF00414837</a>	Methanothermobacter thermautotrophicus	<a href="#">H69232</a> MJ1225-related protein MTH992 <i>ALT_NAMES</i> : inosine-5'-monophosphate dehydrogenase related protein IX [misnomer]	<a href="#">O27073</a> INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN IX	<a href="#">g2622093</a> inosine-5'-monophosphate dehydrogenase related protein IX <a href="#">NP_276127</a> inosine-5'-monophosphate dehydrogenase related protein IX	
<a href="#">NF00414969</a>	Methanothermobacter thermautotrophicus	<a href="#">B69077</a> yhcV homolog 2 <i>ALT_NAMES</i> : inosine-monophosphate dehydrogenase related protein X [misnomer]	<a href="#">O27616</a> INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN X	<a href="#">g2622697</a> inosine-5'-monophosphate dehydrogenase related protein X <a href="#">NP_276687</a> inosine-5'-monophosphate dehydrogenase related protein X	

Source: Cathy Wu

# IMPDH Domain Structure





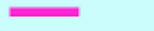

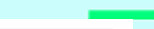




- Typical IMPDHs have 2 IMPDH domains that form the catalytic core and 2 CBS domains.
- A less common but functional IMPDH (E70218) lacks the CBS domains.
- Misnomers show similarity to the CBS domains

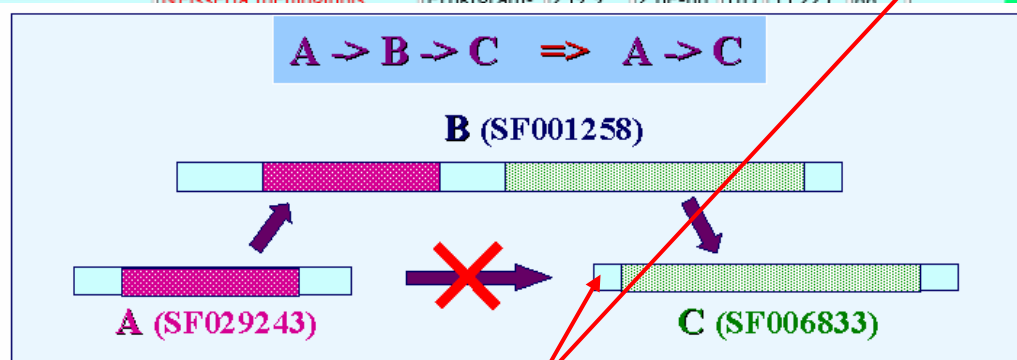
Source: Cathy Wu

# Invalid Transitive Assignment

## Root of invalid transitive assignment

<b>B</b> →	<input type="checkbox"/> H70468	SF001258	051440	<a href="#">phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]</a>	<i>Aquifex aeolicus</i>	Prok/other	594.3	4.8e-26	205	39.086	197	
	<input type="checkbox"/> S76963	SF001258	039935	<a href="#">phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]</a>	<i>Synechocystis sp.</i>	Prok/gram-	557.0	5.7e-24	230	39.175	194	
	<input type="checkbox"/> T35073	SF029243	005738	<a href="#">probable phosphoribosyl-AMP cyclohydrolase</a>	<i>Streptomyces coelicolor</i>	Prok/gram+	399.3	3.5e-15	128	42.157	102	
	<input type="checkbox"/> S53349	SF001257	001188	<a href="#">phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) / histidinol dehydrogenase (EC 1.1.1.23)</a>	<i>Saccharomyces cerevisiae</i>	Euk/fungi	384.1	2.5e-14	799	31.863	204	
<b>A</b> →	<input type="checkbox"/> E69493	SF029243	005738	<a href="#">phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) [similarity]</a>	<i>Archaeoglobus fulgidus</i>	Archae	396.8	4.8e-15	108	47.778	90	
<b>C</b> →	<input type="checkbox"/> G64337	SF006833	030827	<a href="#">phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]</a>	<i>Methanococcus jannaschii</i>	Archae	246.9	1.1e-06	95	36.842	95	
	<input type="checkbox"/> D81178	SF006833	101491	<a href="#">phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) NMB0603 [similarity]</a>	<i>Neisseria meningitidis</i>	Prok/oram-	239.9	2.6e-06	107	35.227	88	
	<input type="checkbox"/> G81925	SF006833	101491	<a href="#">phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) NMA0807 [similarity]</a>								
	<input type="checkbox"/> S51513	SF001257	001188	<a href="#">phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) / histidinol dehydrogenase (EC 1.1.1.23)</a>								

Mis-assignment  
of function



No IMPDH domain

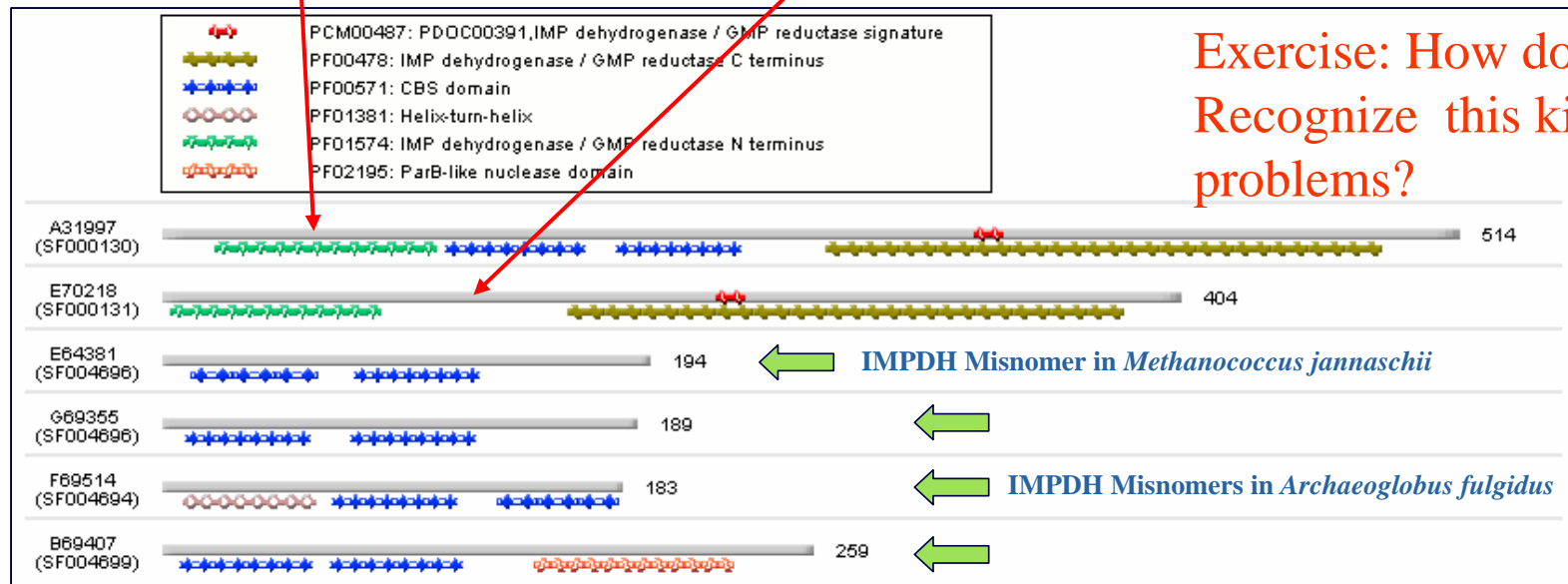
Source: Cathy Wu

# Emerging Pattern

Typical IMPDH

Functional IMPDH w/o CBS

Exercise: How do you  
Recognize this kind of  
problems?



- Most IMPDHs have 2 IMPDH and 2 CBS domains
  - Some IMPDH (E70218) lacks CBS domains
- ⇒ IMPDH domain is the emerging pattern

Source: Cathy Wu

# Application of Sequence Comparison: Active Site/Domain Discovery





## What is a domain

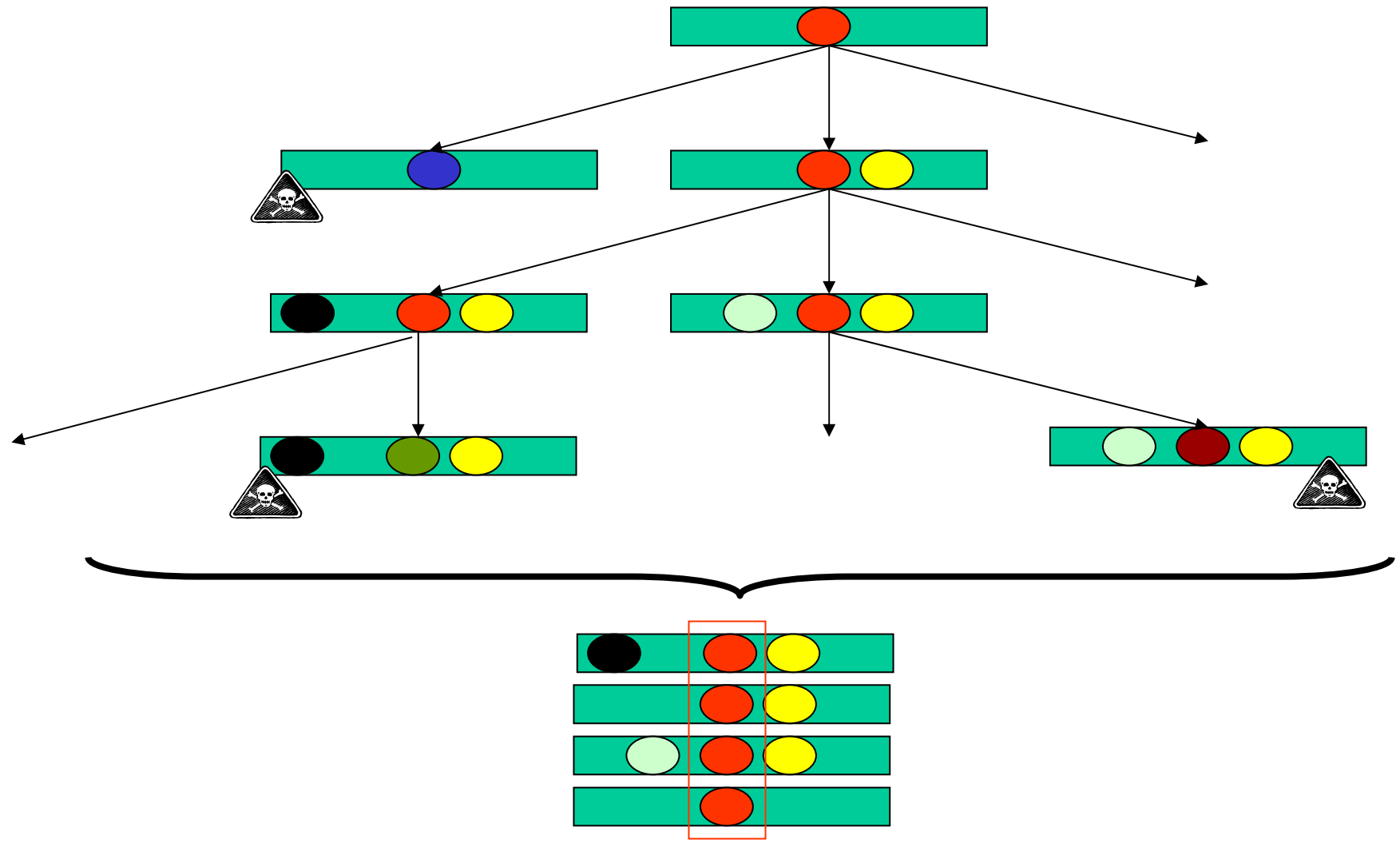
- A **domain** is a component of a protein that is self-stabilizing and folds independently of the rest of the protein chain
  - Not unique to protein products of one gene; can appear in a variety of proteins
  - Play key role in the biological function of proteins
  - Can be "swapped" by genetic engineering between one protein and another to make chimeras
- May be composed of one, more than one, or not any **structural motifs** (often corresponding to **active sites**)

# Discovering Domain and Active Sites

```
>gi|475902|emb|CAA83657.1| protein-tyrosine-phosphatase alpha
MDLWFFVLLLGSGLISVGATNVTTEPPTTVPTSTRIPTKAPTAAPDGGTTPRVSSLNVSSPMTTSAPASE
PPTTTATSISPNATTASLNASTPGTSVPTSAPVAISLPPSATPSALLTALPSTEAEEMTERNVSATVTTQE
TSSASHNGNSDRRDETPIIAVMVALSLLVIVFIIIVLYMLRFKKYKQAGSHSNSFRLPNGRTDDAEPQS
MPLLARSPSTNRKYPPPLPVDKLEEEINRRIGDDNKLFFREEFNALPACPIQATCEAASKEENKEKNRYVNI
LPYDHSRVHLTPVEGV PDSHYINTSFINSYQEKNKFI AAQGPKEETVND FWRMIWEQNTATIVMVTNLKE
RKECKCAQYWPDQGCWTYGNIRVSVEDVTVLVDYTVRKFCIQQVGDVTNKKPQRLVTQFHFTSWPDFGVP
FTP I GMLKFLKKVKTCNPQYAGAI VVHCSAGVGRGTGTFIVIDAMLDMHAERKVDVYGFVSRIRAQRCQM
VQTD MQYVFIYQALLEHYLYGDTELEVTSLEIHLQKIYNKVPGTSSNGLEEEFKKLTSIKIQNDKMRTGN
LPANMKKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASFIDGYRRRTPTCQPRPVQHTIEDFWRMIWEWK
SCSIVMLTELEERGQEKCAQYWPSDGSVSYGDINVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFH
GWPEVGIPSDGKGMINI IAAVQKQQQQSGNHMPMHCHCSAGAGRTGTFCALSTVLERVKAEGILDVVFQTVK
SLRLQRPHMVQTTLEQYEFQYKVVQEYIDAFSDYANFK
```

- **How do we find the domain and associated active sites in the protein above?**

# In the course of evolution...



# Domain/Active Sites as Emerging Patterns

- **How to discover active site and/or domain?**
- **If you are lucky, domain has already been modelled**
  - BLAST,
  - HMMPFAM, ...
- **If you are unlucky, domain not yet modelled**
  - Find homologous seqs
  - Do multiple alignment of homologous seqs
  - Determine conserved positions
  - ⇒ Emerging patterns relative to background
  - ⇒ Candidate active sites and/or domains

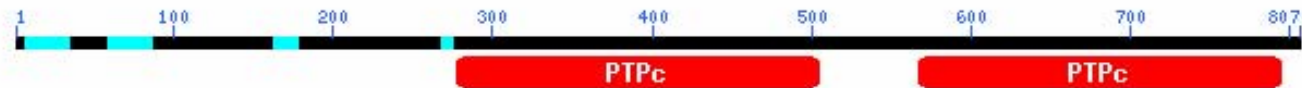
# Lucky Case: Try BLAST



Your request has been successfully submitted and put into the Blast Queue.

Query = (807 letters)

Putative conserved domains have been detected, click on the image below for detailed results.



The request ID is

or

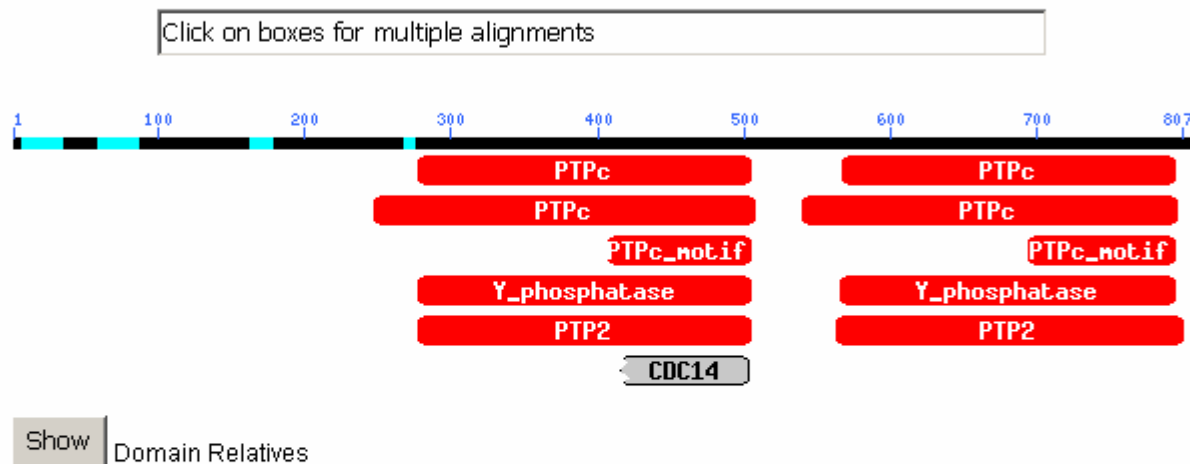
- Just run BLAST on your protein sequence
- If has known domain, BLAST will highlight it ...

# .. And you can navigate the output for more details about the known domains

RPS-BLAST 2.2.13 [Nov-27-2005]

Query= local sequence:  
(807 letters)

Database: cdd.v2.06



PSSMs producing significant alignments:

	Score	E value
<a href="#">gnl CDD 28929</a> cd00047, PTPc, Protein tyrosine phosphatases (PTP) catalyze th...	<a href="#">302</a>	9e-83
<a href="#">gnl CDD 28929</a> cd00047, PTPc, Protein tyrosine phosphatases (PTP) catalyze th...	<a href="#">297</a>	2e-81
<a href="#">gnl CDD 24216</a> smart00194, PTPc, Protein tyrosine phosphatase, catalytic doma...	<a href="#">303</a>	7e-83
<a href="#">gnl CDD 24216</a> smart00194, PTPc, Protein tyrosine phosphatase, catalytic doma...	<a href="#">301</a>	2e-82

# Unlucky Case: Domain/Active Sites Not Already Modelled

- **Find homologous seqs**
  - Literature search
  - BLAST, ...
  - It is better to use distance homologs (why?)
  - “Adjust” the seqs if necessary
- **Do multiple alignment of homologous seqs**
  - ClustalW
  - T-Coffee, ...
- **Determine conserved positions**

## Some Homologs of Our Example Protein

- **P18433**: Receptor-type tyrosine-protein phosphatase alpha precursor (R-PTP- $\alpha$ ) [gi|126467|sp|P18433|PTPRA\\_HUMAN\[126467\]](#)
- **Q15262**: Receptor-type tyrosine-protein phosphatase kappa precursor (R-PTP- $\kappa$ ) [gi|2499753|sp|Q15262|PTPRK\\_HUMAN\[2499753\]](#)
- **P23470**: Receptor-type tyrosine-protein phosphatase gamma precursor (R-PTP- $\gamma$ ) [gi|462550|sp|P23470|PTPRG\\_HUMAN\[462550\]](#)
- **P28828**: Receptor-type tyrosine-protein phosphatase mu precursor (R-PTP- $\mu$ ) [gi|131570|sp|P28828|PTPRM\\_MOUSE\[131570\]](#)
- **P35822**: Receptor-type tyrosine-protein phosphatase kappa precursor (R-PTP-  $\kappa$ ) [gi|548626|sp|P35822|PTPRK\\_MOUSE\[548626\]](#)



Address http://www.ebi.ac.uk/clustalw/

Google quence multiple alignment Search PageRank 6 blocked Check AutoLink AutoFill

## SEQUENCE ANALYSIS

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- ClustalW Help
- ClustalW FAQ
- Jalview Help
- Scores Table
- Alignment
- Guide Tree
- Colours

## ClustalW Submission Form

Clustal W is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. **[New users, please read the FAQ.](#)**

>> **Download Software**

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
<input type="text"/>	<input type="text" value="Sequence"/>	<input type="text" value="interactive"/>	<input type="text" value="full"/>	<input type="text" value="single"/>
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="percent"/>	<input type="text" value="def"/>	<input type="text" value="def"/>
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
<input type="text" value="aln w/numbers"/>	<input type="text" value="aligned"/>	<input type="text" value="none"/>	<input type="text" value="off"/>	<input type="text" value="off"/>

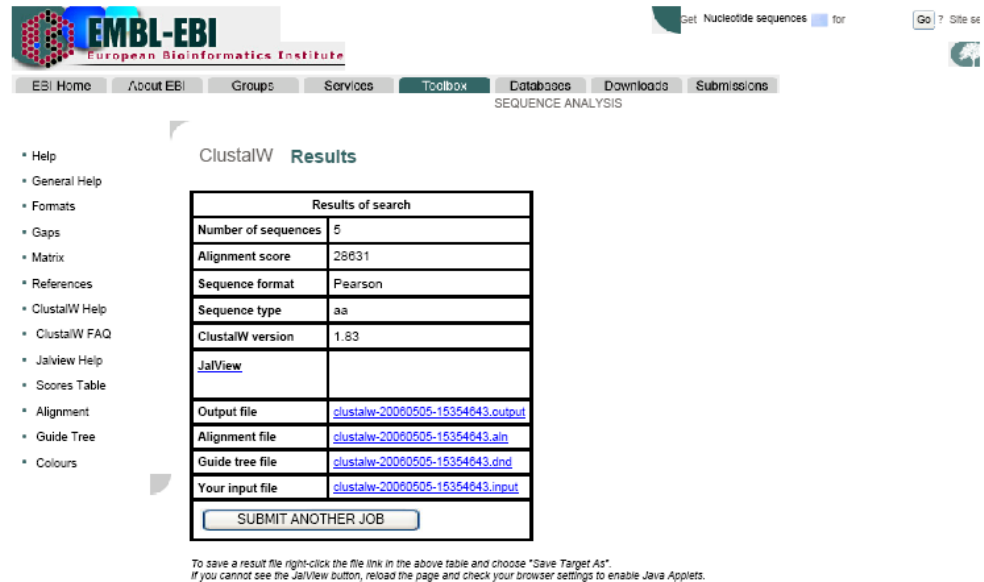
Enter or Paste a set of Sequences in any supported format:

# Example Output from ClustalW

- [clustalw-output.html](#)

ClustalW

Page 1 of 8



EMBL-EBI  
European Bioinformatics Institute

Get Nucleotide sequences for   Site set

EBI Home About EBI Groups Services **Toolbox** Databases Downloads Submissions

SEQUENCE ANALYSIS

ClustalW **Results**

Results of search	
Number of sequences	5
Alignment score	28031
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.83
<a href="#">JalView</a>	
Output file	<a href="#">clustalw-20060505-15354643.output</a>
Alignment file	<a href="#">clustalw-20060505-15354643.aln</a>
Guide tree file	<a href="#">clustalw-20060505-15354643.dnd</a>
Your input file	<a href="#">clustalw-20060505-15354643.input</a>
<input type="button" value="SUBMIT ANOTHER JOB"/>	

To save a result file right-click the file link in the above table and choose "Save Target As".  
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

<http://www.ebi.ac.uk/cgi-bin/clustalw/result?tool=clustalw&jobid=clustalw-20060505-15354643&treendispl=hide&treetype=new&sortby...> 05/05/2006

## Let's put in a few more distance homologs

- **>gi|18859295|ref|NP\_571963.1| protein tyrosine phosphatase, receptor type, A [Danio rerio]**
- **>gi|7248657|gb|AAF43605.1|AF197944\_1 receptor protein tyrosine phosphatase delta [Xenopus laevis]**
- **>gi|15027042|emb|CAC44759.1| receptor protein-tyrosine phosphatase sigma [Danio rerio]**
- **>gi|6093855|sp|Q98936|PTPRG\_CHICK Receptor-type tyrosine-protein phosphatase gamma precursor (Protein-tyrosine phosphatase gamma) (R-PTP-gamma)**

# Example Output from ClustalW

- [more-clustalw-output.html](#)

Page 1 of 8

## ClustalW Results

Results of search	
Number of sequences	9
Alignment score	73502
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.83
JalView	
Output file	<a href="#">clustalw-20080505-16575265.output</a>
Alignment file	<a href="#">clustalw-20080505-16575265.aln</a>
Guide tree file	<a href="#">clustalw-20080505-16575265.dnd</a>
Your input file	<a href="#">clustalw-20080505-16575265.input</a>
<input type="button" value="SUBMIT ANOTHER JOB"/>	

To save a result file right-click the file link in the above table and choose "Save Target As".  
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

## Scores Table

Sort by

SeqA Name	Len(aa)	SeqB Name	Len(aa)	Score		
1	g 1126467 sp F18433 PTPRA_HUMA	802	2	g 12499753 sp Q15262 PTPRK_HUM	1439	32
1	g 1126467 sp F18433 PTPRA_HUMA	802	3	g 1462550 sp P23470 PTPRG_HUMA	1445	30
1	g 1126467 sp F18433 PTPRA_HUMA	802	4	g 1131570 sp P28828 PTPRM_MOUSE	1452	33
1	g 1126467 sp F18433 PTPRA_HUMA	802	5	g 1548626 sp P35822 PTPRK_MOUSE	1457	32
1	g 1126467 sp F18433 PTPRA_HUMA	802	6	g 138859295 ref NP_571963.1	833	74
1	g 1126467 sp F18433 PTPRA_HUMA	802	7	g 17248657 gb AAP43605.1 AF197	1896	41
1	g 1126467 sp F18433 PTPRA_HUMA	802	8	g 115027042 emb CAC44759.1	857	38
1	g 1126467 sp F18433 PTPRA_HUMA	802	9	g 16093855 sp Q98936 PTPRG_CHI	1422	31
2	g 12499753 sp Q15262 PTPRK_HUM	1439	3	g 1462550 sp P23470 PTPRG_HUMA	1445	18
2	g 12499753 sp Q15262 PTPRK_HUM	1439	4	g 1131570 sp P28828 PTPRM_MOUSE	1452	61
2	g 12499753 sp Q15262 PTPRK_HUM	1439	5	g 1548626 sp P35822 PTPRK_MOUSE	1457	98
2	g 12499753 sp Q15262 PTPRK_HUM	1439	6	g 138859295 ref NP_571963.1	833	33
2	g 12499753 sp Q15262 PTPRK_HUM	1439	7	g 17248657 gb AAP43605.1 AF197	1896	25
2	g 12499753 sp Q15262 PTPRK_HUM	1439	8	g 115027042 emb CAC44759.1	857	35
2	g 12499753 sp Q15262 PTPRK_HUM	1439	9	g 16093855 sp Q98936 PTPRG_CHI	1422	20
3	g 1462550 sp P23470 PTPRG_HUMA	1445	4	g 1131570 sp P28828 PTPRM_MOUSE	1452	18
3	g 1462550 sp P23470 PTPRG_HUMA	1445	5	g 1548626 sp P35822 PTPRK_MOUSE	1457	18
3	g 1462550 sp P23470 PTPRG_HUMA	1445	6	g 138859295 ref NP_571963.1	833	30
3	g 1462550 sp P23470 PTPRG_HUMA	1445	7	g 17248657 gb AAP43605.1 AF197	1896	22
3	g 1462550 sp P23470 PTPRG_HUMA	1445	8	g 115027042 emb CAC44759.1	857	32
3	g 1462550 sp P23470 PTPRG_HUMA	1445	9	g 16093855 sp Q98936 PTPRG_CHI	1422	87
4	g 1131570 sp P28828 PTPRM_MOUSE	1452	5	g 1548626 sp P35822 PTPRK_MOUSE	1457	61
4	g 1131570 sp P28828 PTPRM_MOUSE	1452	6	g 138859295 ref NP_571963.1	833	31
4	g 1131570 sp P28828 PTPRM_MOUSE	1452	7	g 17248657 gb AAP43605.1 AF197	1896	26
4	g 1131570 sp P28828 PTPRM_MOUSE	1452	8	g 115027042 emb CAC44759.1	857	35
4	g 1131570 sp P28828 PTPRM_MOUSE	1452	9	g 16093855 sp Q98936 PTPRG_CHI	1422	18
5	g 1548626 sp P35822 PTPRK_MOUSE	1457	6	g 138859295 ref NP_571963.1	833	33
5	g 1548626 sp P35822 PTPRK_MOUSE	1457	7	g 17248657 gb AAP43605.1 AF197	1896	25
5	g 1548626 sp P35822 PTPRK_MOUSE	1457	8	g 115027042 emb CAC44759.1	857	35
5	g 1548626 sp P35822 PTPRK_MOUSE	1457	9	g 16093855 sp Q98936 PTPRG_CHI	1422	20

<http://www.ebi.ac.uk/cgi-bin/printable?F=http://www.ebi.ac.uk/cgi-bin/clustalw/result...> 06/05/2006

## Multiple Alignment of PTPs

```

gi|126467|      FHFTSWPDFGVPFTP I GMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAMLD
gi|2499753     FHFTGWPDHGVPHYHATGLLSF IRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|462550|     YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSMLQ
gi|2499751     FHFTSWPDHGVPD TTDLLINFRYLVRDYMKQSPPE SPILVHCSAGVGRTGTFIAIDRLIY
gi|1709906     FQFTA WPDHGVPEHPTPFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAMLE
gi|126471|     LHFTSWPDFGVPFTP I GMLKFLKKVKT LNP--VHAGPIVVHCSAGVGRTGTFIVIDAMMA
gi|548626|     FHFTGWPDHGVPHYHATGLLSF IRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|131570|     FHFTGWPDHGVPHYHATGLLGFVRQVKS KSP--PNAGPLVVHCSAGAGRTGCFIVIDIMLD
gi|2144715     FHFTSWPDHGVPD TTDLLINFRYLVRDYMKQSPPE SPILVHCSAGVGRTGTFIAIDRLIY
                ..*  ***  ***          .  *                               ..*****  ****...  **  ..

```

- Notice the PTPs agree with each other on some positions more than other positions
  - These positions are more imp't wrt PTPs
  - Else they wouldn't be conserved by evolution
- ⇒ They are candidate active sites

# Application of Sequence Comparison: Key Mutation Site Discovery




# 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
EEEFKKLTSIKIQNDKMRTGNLPA NMKKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASF
IDGYRQKDSYIASQGPLLHTIEDFWRMIWEWKSCSIVMLTELEERGQEKCAQYWPSDGLV
SYGDITVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFHGWPEVGIPSDGKGMISII
AAVQKQQQSGNHPITVHCSAGAGRTGTFCALSTVLERVKAEGILDVVFQTVKSLRLQRP
MVQTLQYEFQYKVVQYIDAFSDYANFK
```



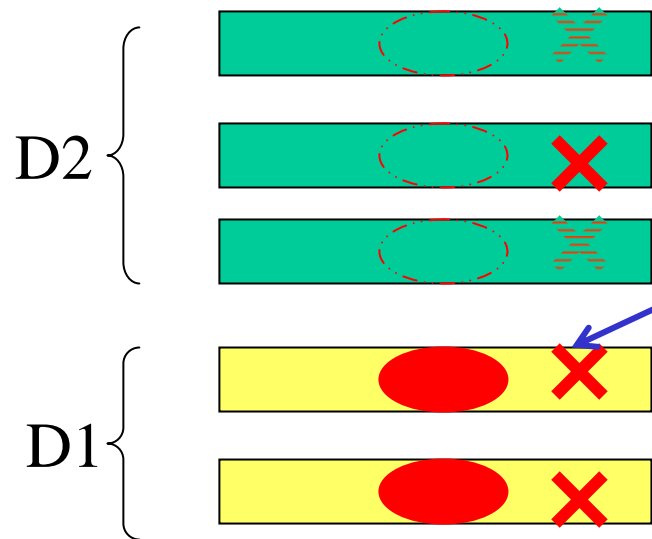
- Some PTPs have 2 PTP domains
- PTP domain D1 is has much more activity than PTP domain D2
- Why? And how do you figure that out?

# 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

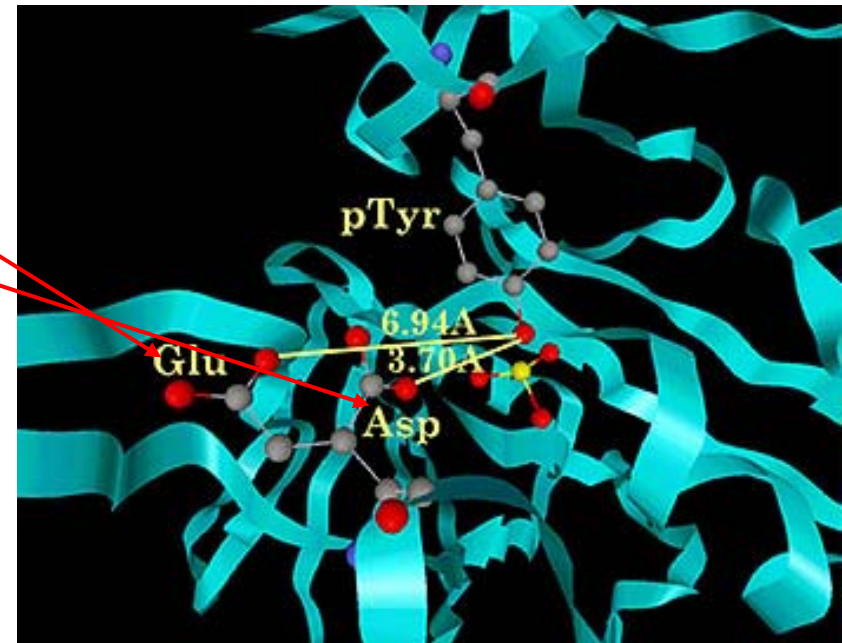
 absent  
 present



## Key Mutation Site: PTP D1 vs D2

```

                ?  !  ?
gi|00000|P D2  QFHFGWPEVNGIPSDGK
gi|126467|      QFHFTSWPDFGVFFTPIC
gi|2499753     QFHFTGWPDHGVPYHATC
gi|462550|     QYHYTQWPDMGVPEYALI
gi|2499751     QFHFTSWPDHGVPDTTDI
gi|1709906 D1  QFQFTAAMPDHGVPEHPTI
gi|126471|     QLHFTSWPDFGVFFTPIC
gi|548626|     QFHFTGWPDHGVPYHATC
gi|131570|     QFHFTGWPDHGVPYHATC
gi|2144715     QFHFTSWPDHGVPDTTDI
                * .. **.*.*
  
```



- 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  $\rightarrow$  D in D2 and see if there is gain in PTP activity
  - Mutate D  $\rightarrow$  E in D1 and see if there is loss in PTP activity

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

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

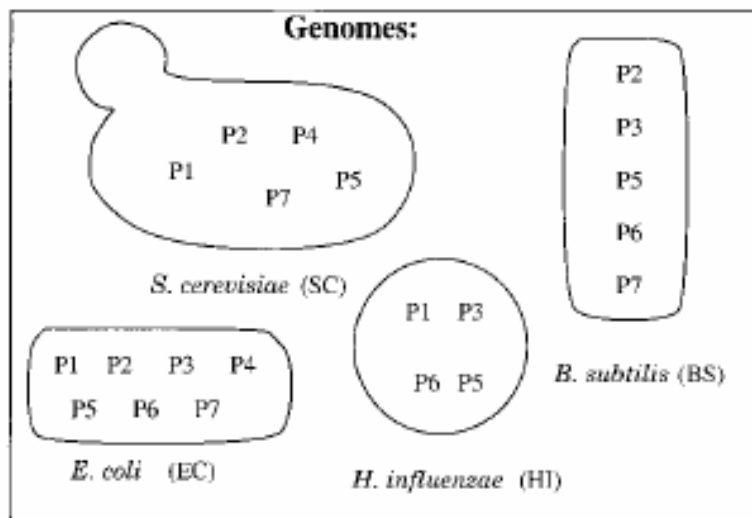
**genome phylogenetic profiles  
protfun's feature profiles**



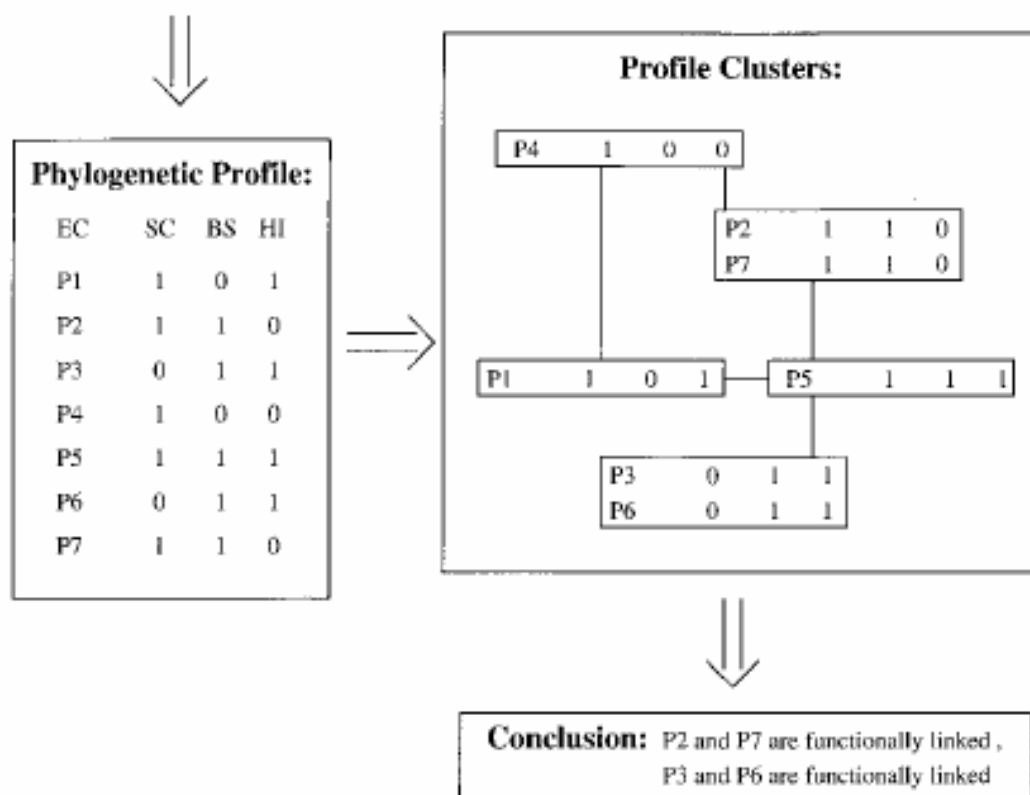
# Phylogenetic Profiling

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

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

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

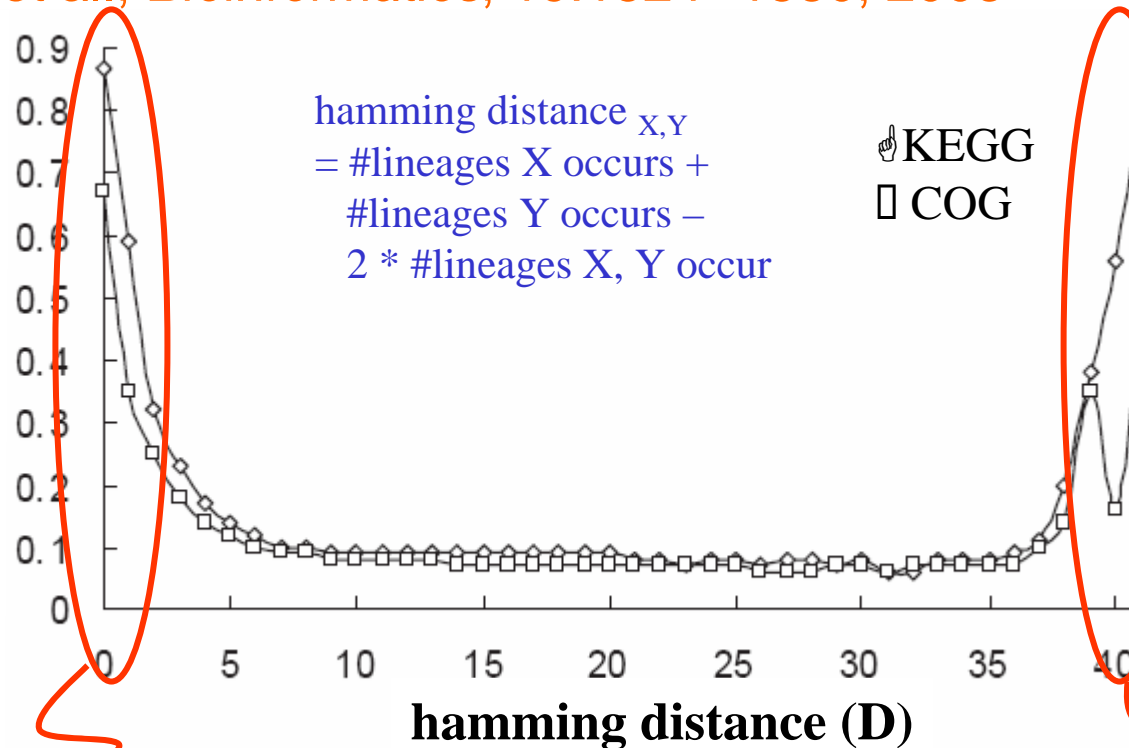
Keyword	No. of non-homologous proteins in group	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins*	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdoterin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

fraction of gene pairs  
having hamming distance D  
and share a common pathway  
in KEGG/COG



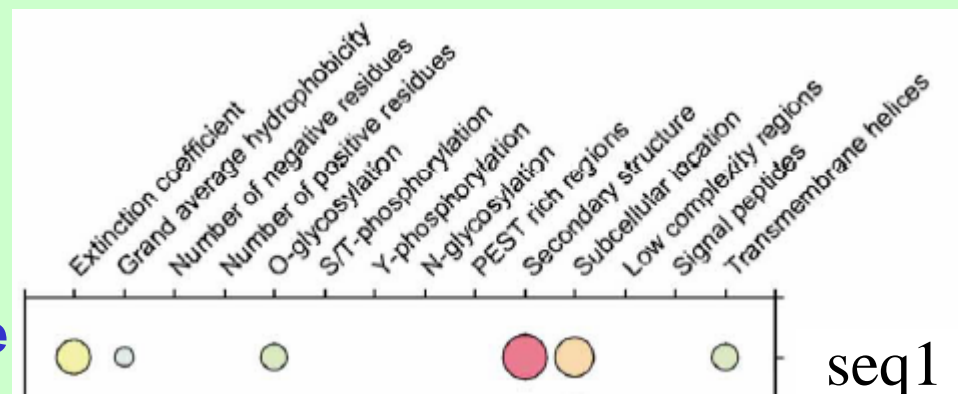
- Proteins having low hamming distance (thus highly similar phylogenetic profiles) tend to share common pathways

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

# The ProtFun Approach

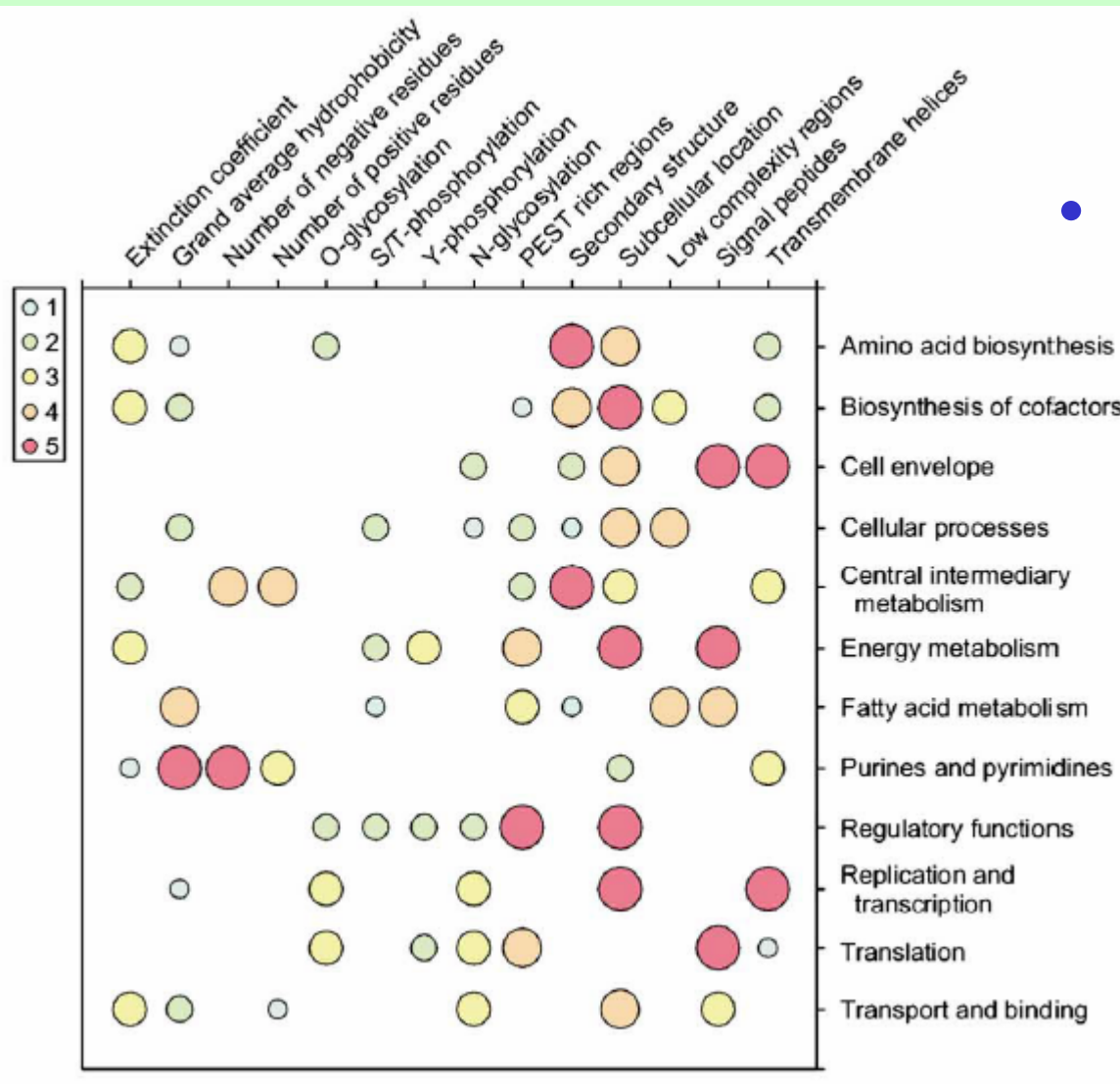
Jensen, *JMB*, 319:1257--1265, 2002

- 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, 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?

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

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

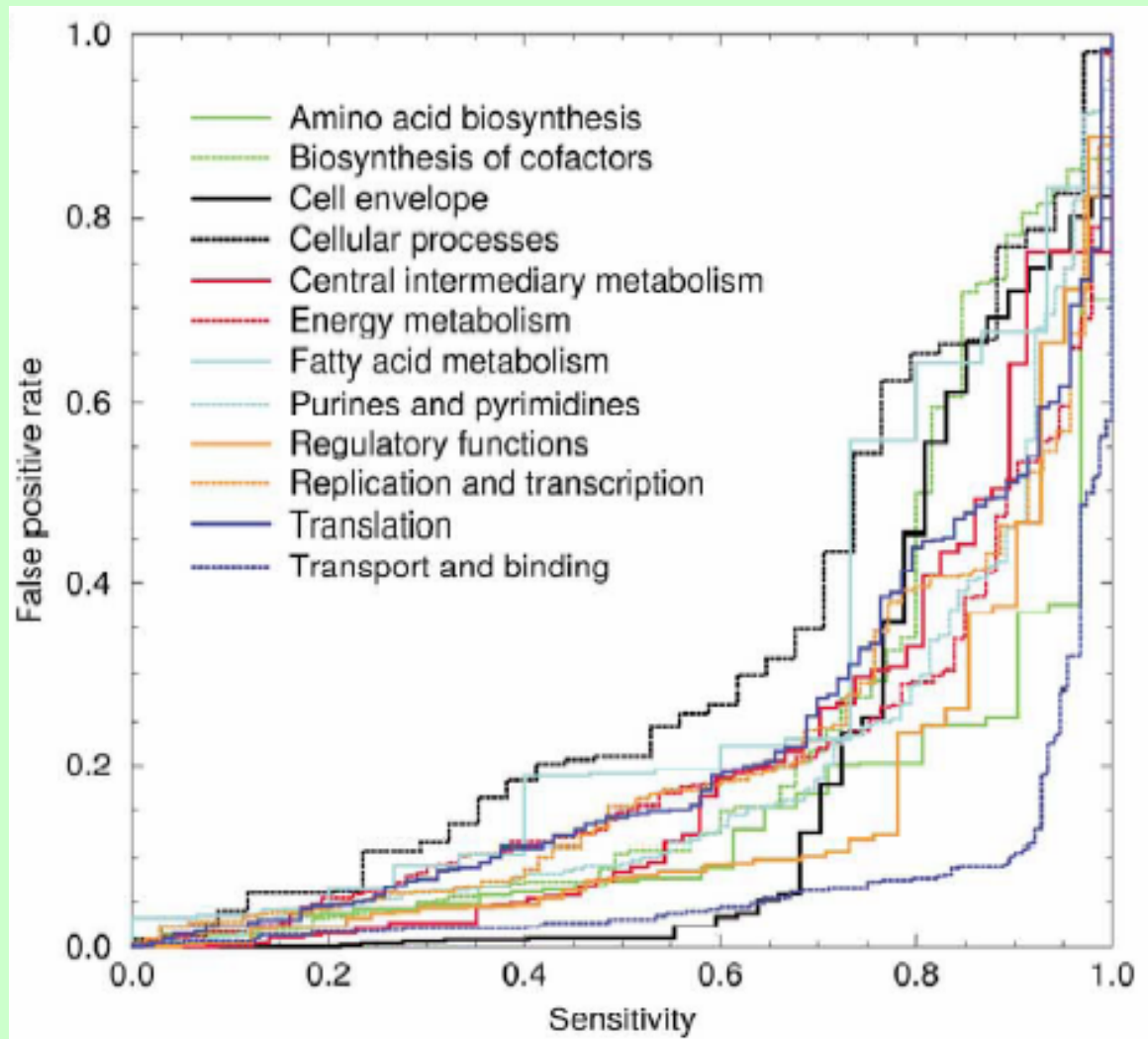
Average the output of the 5 component ANNs

# ProtFun: 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



# SVM-Pairwise Framework

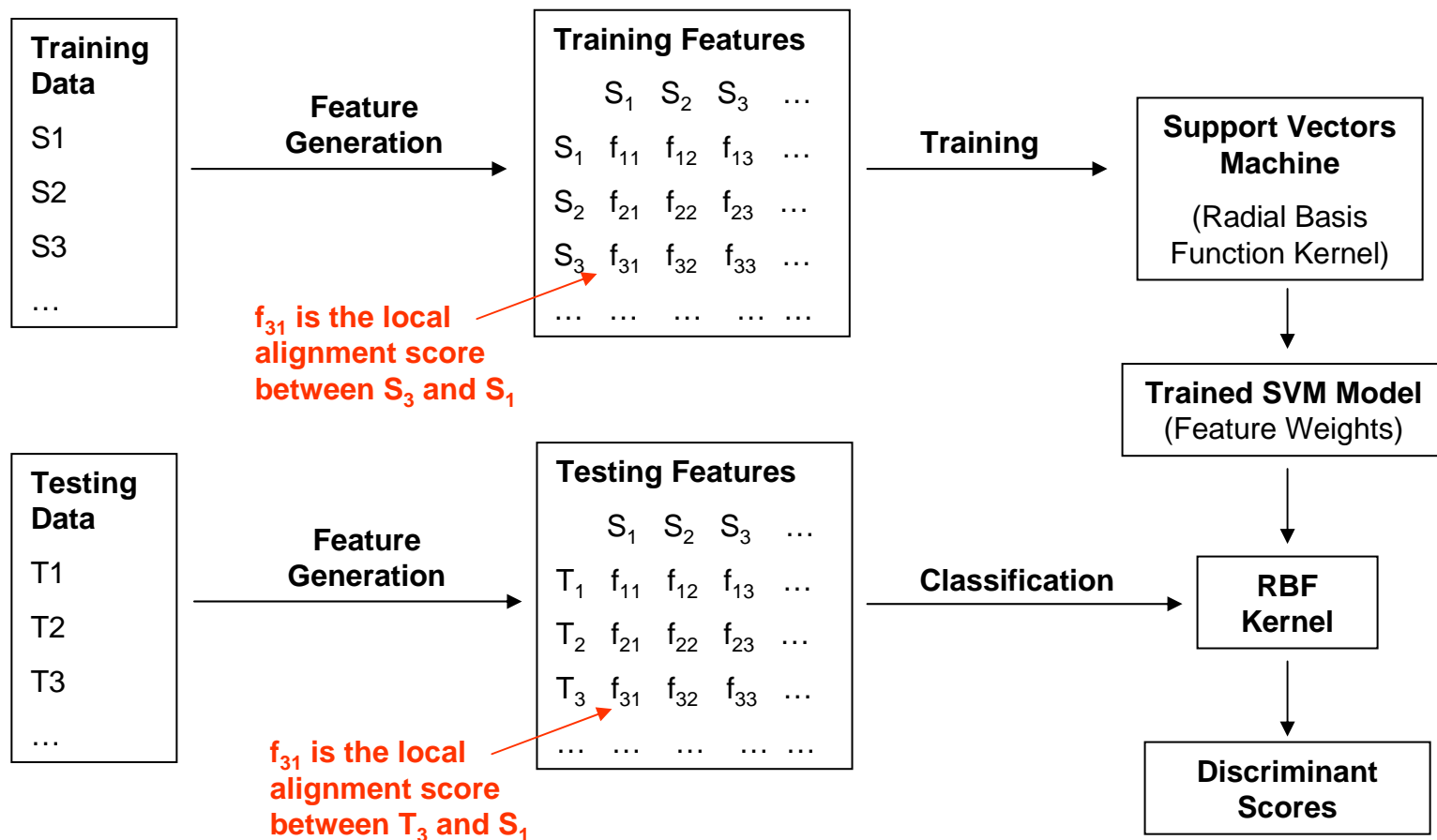
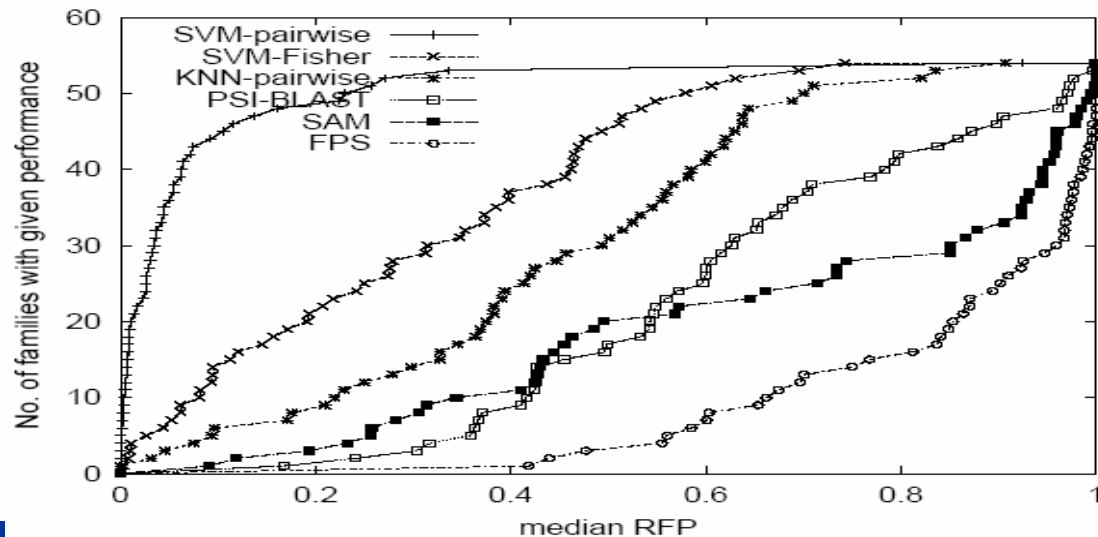
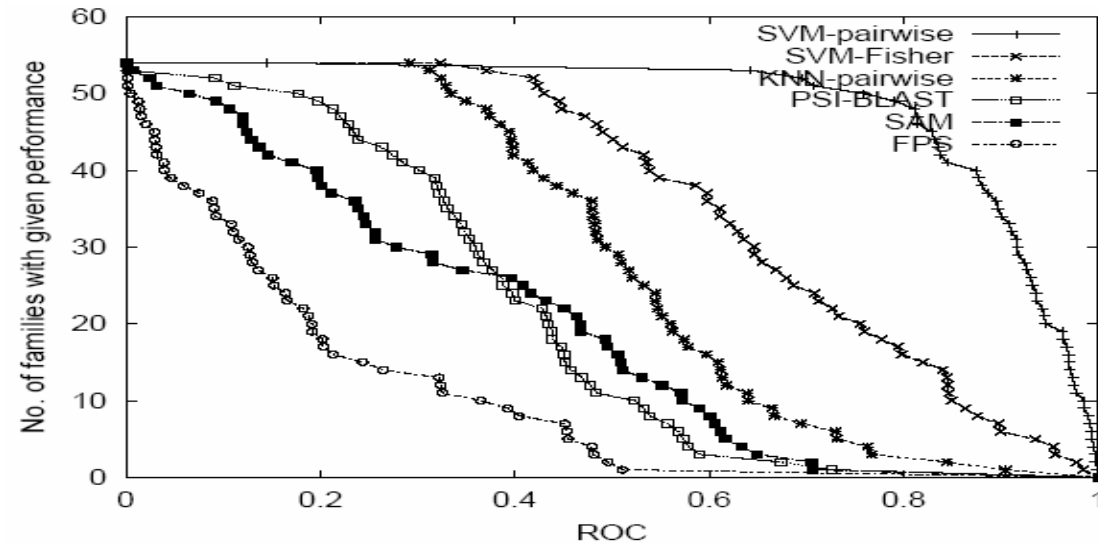


Image credit: Kenny Chua



# 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.



Any Questions?



# 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
- RA Abagyan, S Batalov. Do aligned sequences share the same fold? *JMB*, 273(1):355-68, 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
- C. Wu, W. Barker. “A Family Classification Approach to Functional Annotation of Proteins”, *The Practical Bioinformatician*, Chapter 19, pages 401—416, WSPC, 2004
- H.N. Chua, W.-K. Sung. [A better gap penalty for pairwise SVM](#). Proc. APBC05, pages 11-20
- T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote homologies. *JCB*, 7(1-2):95—11, 2000