

For written notes on this lecture, please read chapter 19 of *The Practical Bioinformatician* and *Hawkins & Kihara, JBCB 5(1):1-30, 2007*

# CS2220: Introduction to Computational Biology

## Unit 7: Sequence Homology Interpretation

**Wong Limsoon**



# Plan

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

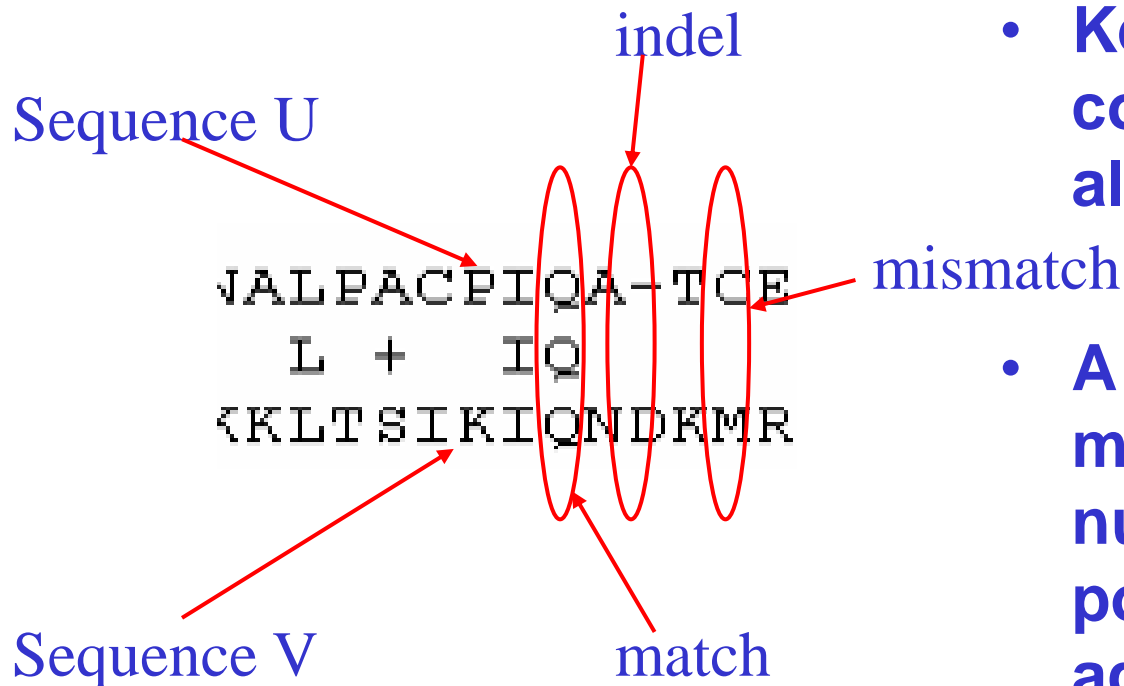
# Very Brief Recap of Sequence Comparison/Alignment



# 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      MPHNVH FVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVVE
                ...: . :.. ::
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  MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
          MK G L  ++          MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1  MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWNNDVVAHT 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 GMLKFLKVKACNP--QYAGAI VVHCSAGVGRTGTFVVIDAML D
gi|2499753     FHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPI VVHCSAGAGRTGCIYIVIDIML D
gi|462550|     YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVI VVHCSAGVGRTGTIYIVIDSM LQ
gi|2499751     FHFTSWPDHGVPD TTDLLINFRYLVRDYMKQSPPE SPILVHCSAGVGRTGTFIAIDRLIY
gi|1709906     FQFTA WPDHGVP EHP T PFLAFLRRVKTCNP--PDAGPM VVHCSAGVGRTGCFIVIDAM L E
gi|126471|     LHFTSWPDFGVPFTP I GMLKFLKVKTLNP--VHAGPI VVHCSAGVGRTGTFIVIDAM M A
gi|548626|     FHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPI VVHCSAGAGRTGCIYIVIDIML D
gi|131570|     FHFTGWPDHGVPYHATGLLGFVRQVKSKSP--PNAGPL VVHCSAGAGRTGCFIVIDIML D
gi|2144715     FHFTSWPDHGVPD TTDLLINFRYLVRDYMKQSPPE SPILVHCSAGVGRTGTFIAIDRLIY
                ..*  ***  ***          .  *                ..*****  *****  **  ..
  
```

Conserved sites

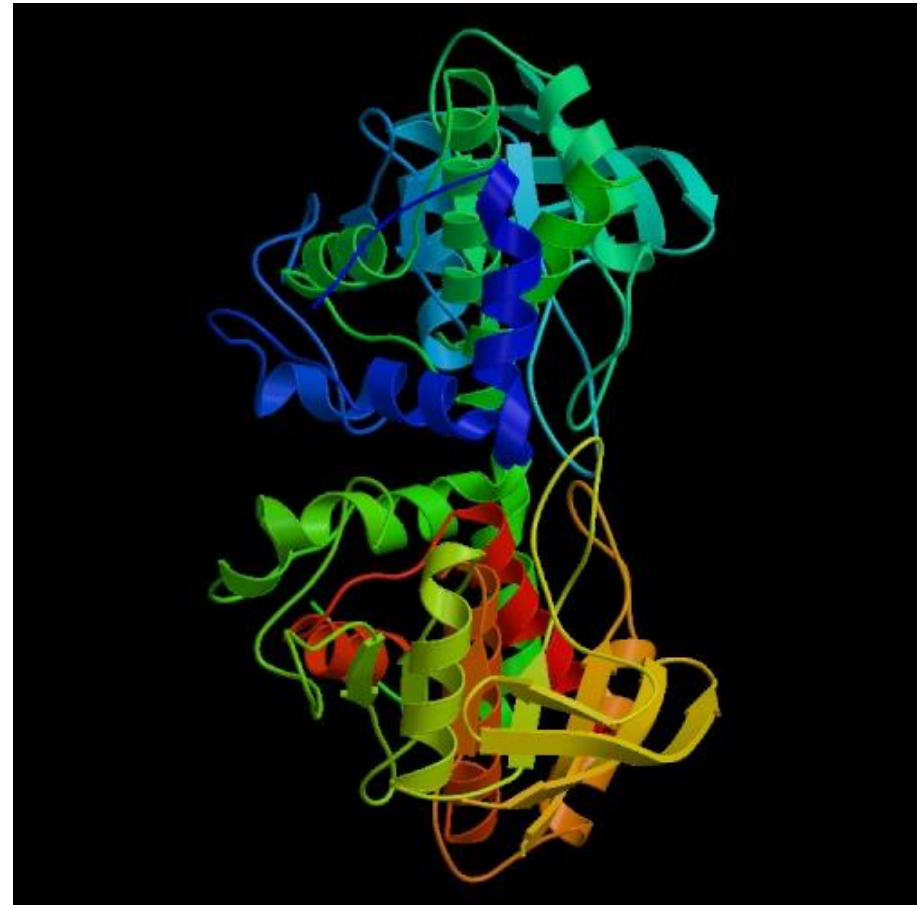


# Application of Sequence Comparison: Guilt-by-Association



## A protein is a ...

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



# Function assignment to protein seq

SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR  
YVNILPYDHSRVHLTPVEGVPDSYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE  
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD  
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG  
TFVVIDAMLDMMSERKVDVYGFVSRIRAQRCQMVQTDMQYVFYQALLEHYLYGDTELE  
VT

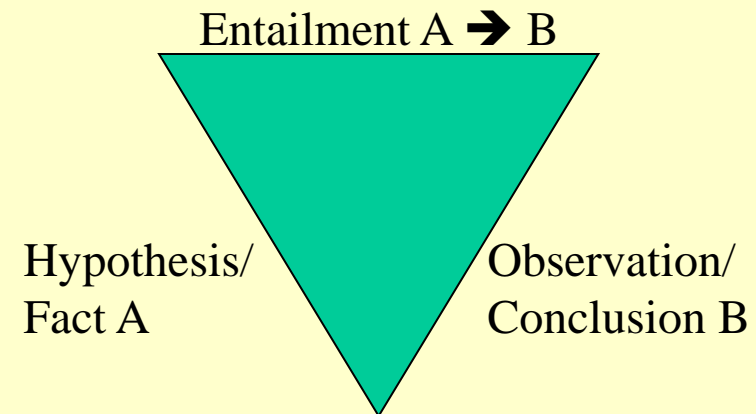
- How do we attempt to assign a function to a new protein sequence?

# Invariant and abductive reasoning

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

⇒ **Abductive reasoning**

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



⇒ **“Guilt by association”**

## 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  
 $\Rightarrow$  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
MPHNVHFVAGVLGSAALKGPMMKKEQAYSLSLTFTEAGTYDYHCTPHFFMRGKVVV
                . . . . .
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGE7FFYNFPVDNPGTFFYHGHLGMORSAGLYG
                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  
 $\Rightarrow$  The two proteins are likely to be homologous

```

>gi113476732|ref|NP_108301.1| unknown protein [Mesorhizobium loti]
gi114027493|db|BAE53762.1| 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 MKPORLASIALAIIFLPMVFAHAATIEITMENLVISPTIEVSAKVQDTIRVFNKDVFAHT 60
          MK G L ++ MA PA AATIE+T++ LV SP V AKVGDIT VVN DV AHT
Sbjct: 1 MKAGALIELSLAALALMAFAAAATIEVTIDKLVFSPATVEAKVGDITVFNKDVVAHT 60
  
```

good match between Amicyanin and unknown M. loti protein

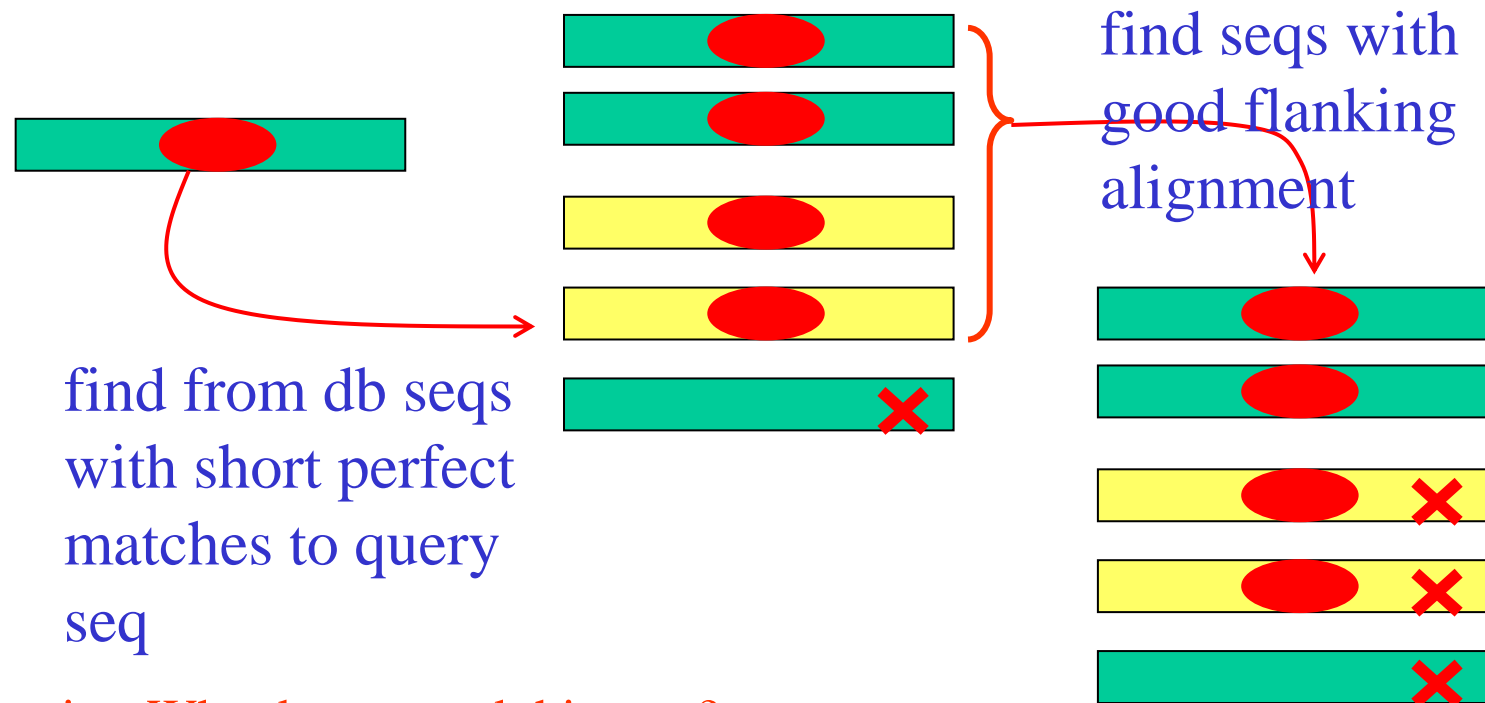
Assign to  $T$  same function as homologs

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**



Exercise: Why do we need this step?

# 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="#">621</a> <b>L</b>	e-177
<a href="#">gi 126467 sp P18433 PTRA_HUMAN</a> Protein-tyrosine phosphatase...	<a href="#">621</a> <b>L</b>	e-177
<a href="#">gi 4506303 ref NP_002827.1 </a> protein tyrosine phosphatase, r...	<a href="#">621</a> <b>L</b>	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="#">621</a> <b>L</b>	e-176
<a href="#">gi 32067 emb CAA37447.1 </a> tyrosine phosphatase precursor [Ho...	<a href="#">611</a> <b>L</b>	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="#">611</a> <b>L</b>	e-176
<a href="#">gi 2098414 pdb 1YFO A</a> Chain A, Receptor Protein Tyrosine Ph...	<a href="#">61</a> <b>S</b>	e-174
<a href="#">gi 32313 emb CAA38662.1 </a> protein-tyrosine phosphatase [Homo...	<a href="#">61</a> <b>L</b>	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> <b>L</b>	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 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFC IQQVGD 180
          QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFC IQQVGD
Sbjct: 322 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFC IQQVGD 381

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

Query: 241 TFVVIDAMLDMHSEKVDVYGFVSRIRAQRCQMVQTD MQYVF IYQALLEHYLYGDTELE 300
          TFVVIDAMLDMHSEKVDVYGFVSRIRAQRCQMVQTD MQYVF IYQALLEHYLYGDTELE
Sbjct: 442 TFVVIDAMLDMHSEKVDVYGFVSRIRAQRCQMVQTD MQYVF IYQALLEHYLYGDTELE 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!

Note:  $P = 1 - e^{-E}$

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**



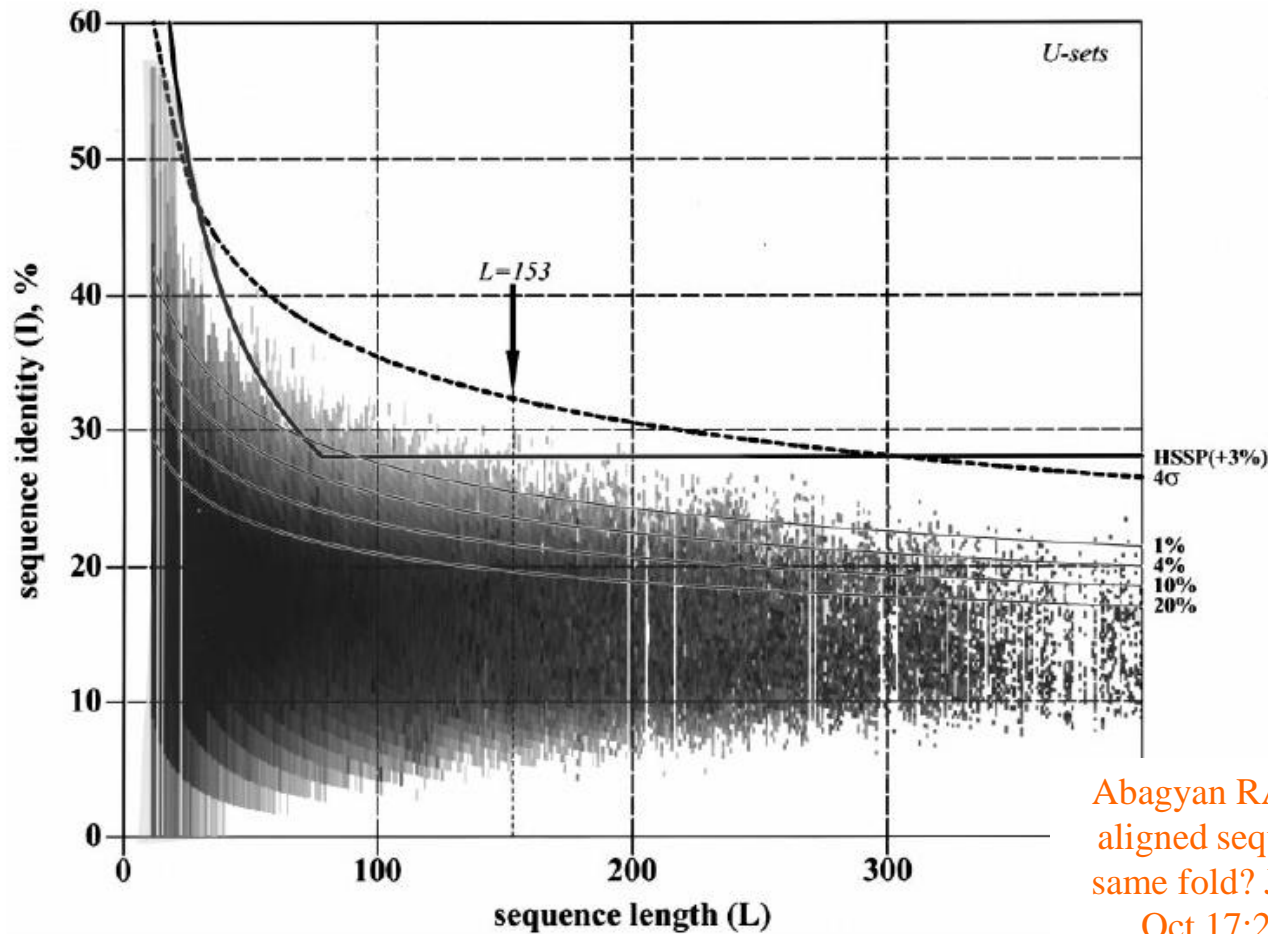
Cartoon: Ron Hipschman  
Data: David Hand

## Effect of seq compositional bias

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

Source: NCBI

## Effect of sequence length



Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol. 1997 Oct 17;273(1):355-68

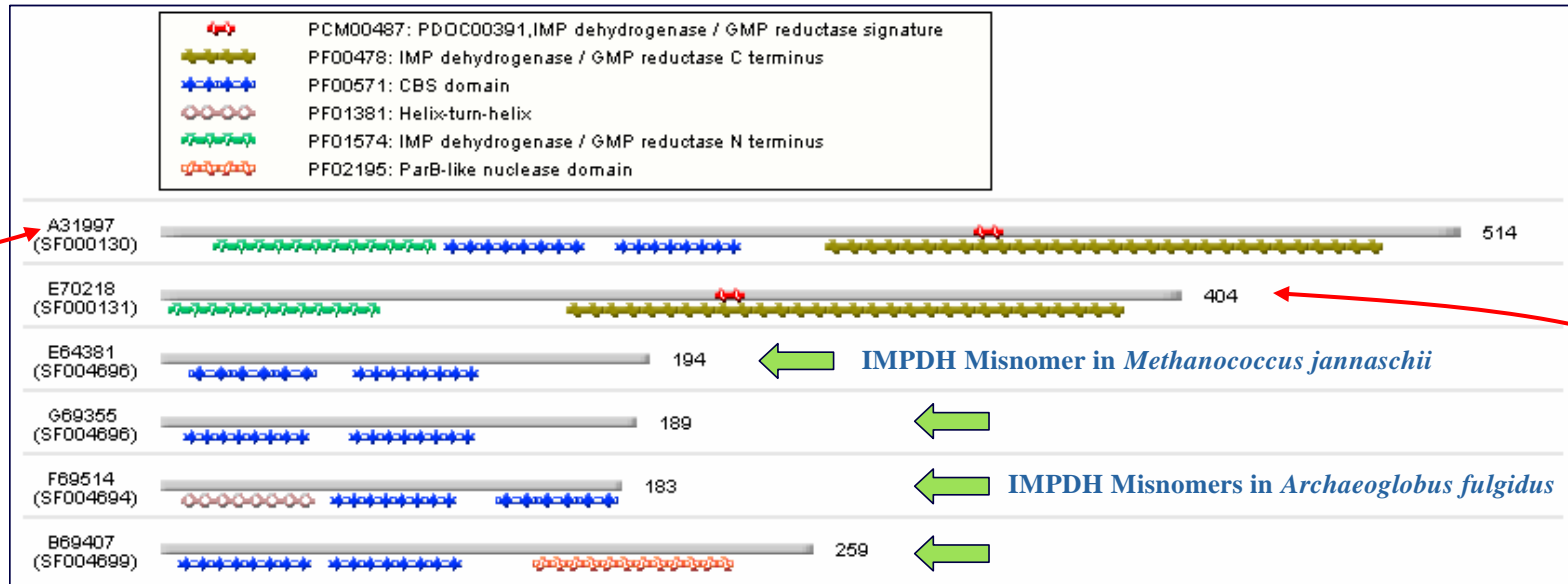
# Examples of invalid function assignment: IMP dehydrogenases (IMPDH)

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="#">G69636</a> MJ0653 homolog AF111220 <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	












# IMPDH domain structure

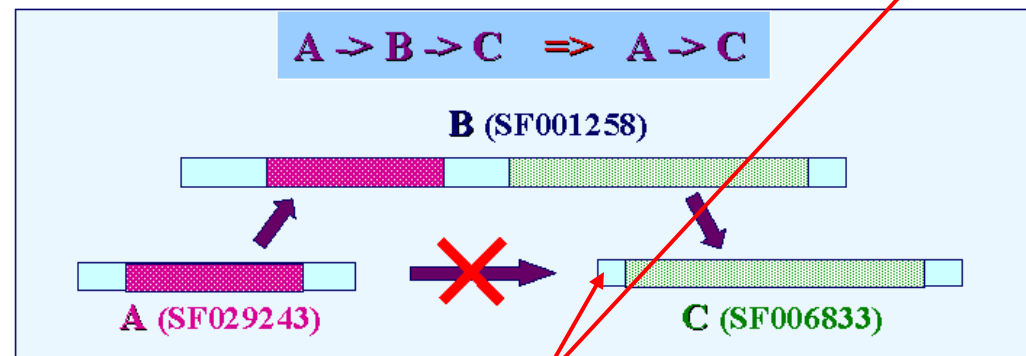


- Typical IMPDHs have 2 IMPDH domains that form the catalytic core and 2 CBS domains.
- A less common but functional IMPDH (E70218) lacks the CBS domains.
- Misnomers show similarity to the CBS domains

# Invalid transitive assignment

## Root of invalid transitive assignment

<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>	Prnk/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>								



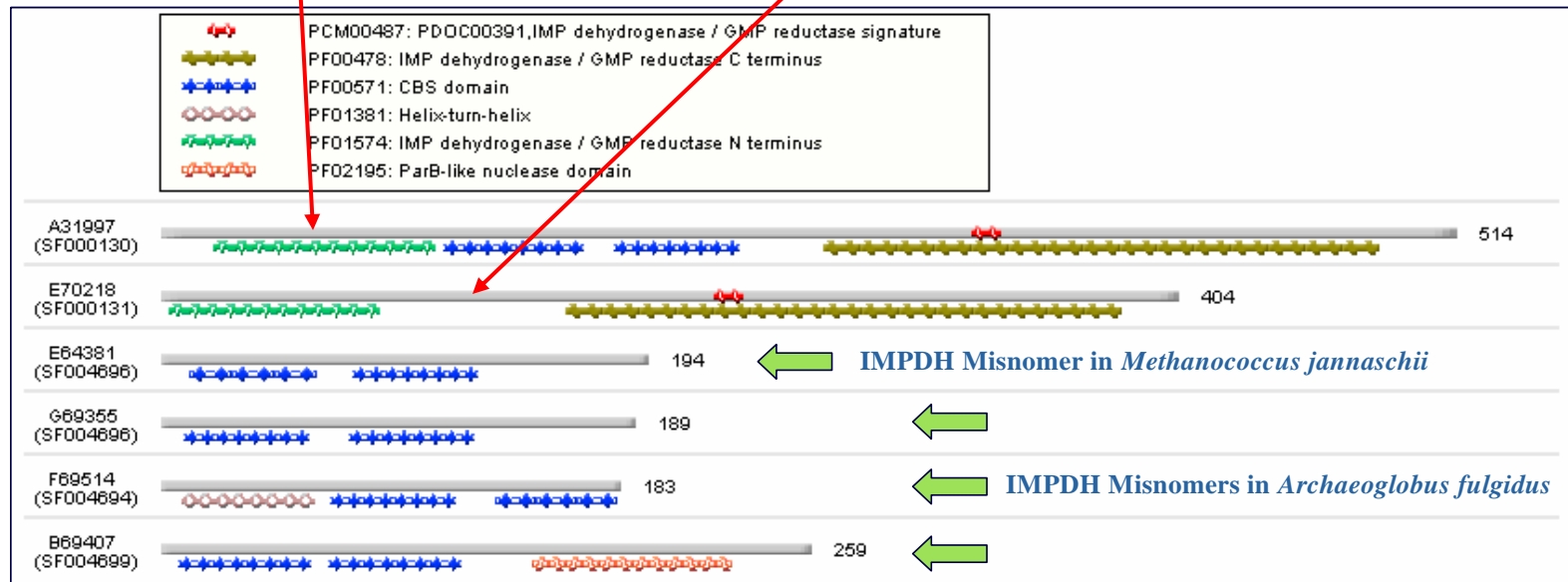
Mis-assignment  
of function

No IMPDH domain

# Emerging pattern

Typical IMPDH

Functional IMPDH w/o CBS



- 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

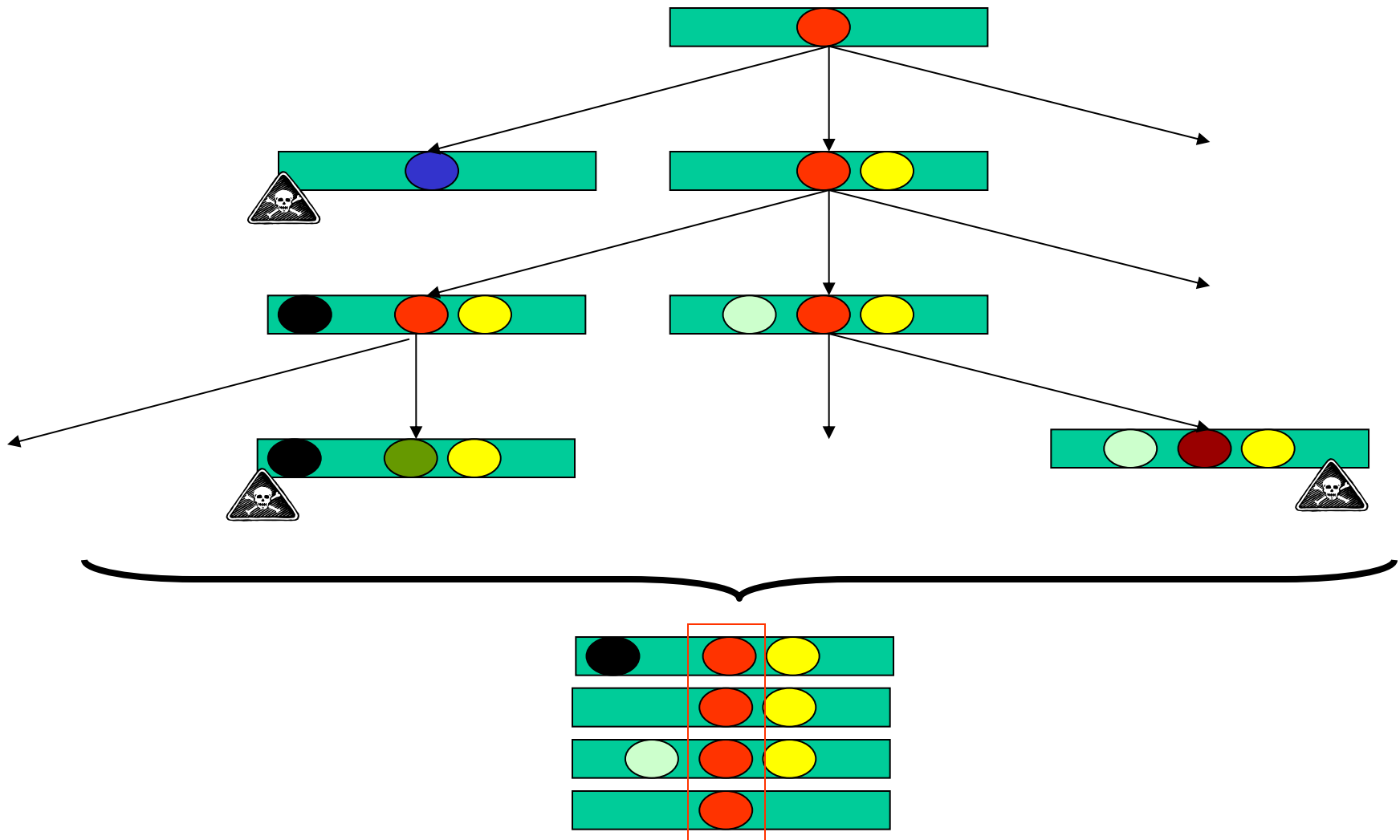


## Discover active site and/or domain

- **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?

# In the course of evolution...



## Multiple alignment of PTPs

```

gi|126467|      FHFTSWPDFGVPFTP I GMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAMLD
gi|2499753     FHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPIVVHCSAGAGRTGCIYIVIDIMLD
gi|462550|     YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTIYIVIDSMLQ
gi|2499751     FHFTSWPDHGVPD TTDLLINFRYLVRDYMKQSPPEPILVHCSAGVGRTGTFIAIDRLIY
gi|1709906     FQFTA WPDHGVP EHP T PFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAMLE
gi|126471|     LHFTSWPDFGVPFTP I GMLKFLKKVKT LNP--VHAGPIVVHCSAGVGRTGTFIVIDAMMA
gi|548626|     FHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPIVVHCSAGAGRTGCIYIVIDIMLD
gi|131570|     FHFTGWPDHGVPYHATGLLGFVRQVKS KSP--PNAGPLVVHCSAGAGRTGCFIVIDIMLD
gi|2144715     FHFTSWPDHGVPD TTDLLINFRYLVRDYMKQSPPEPILVHCSAGVGRTGTFIAIDRLIY
                ..*  ***  ***          .  *                               ..*****  ****...  **  ..
  
```

- 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

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





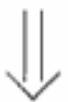
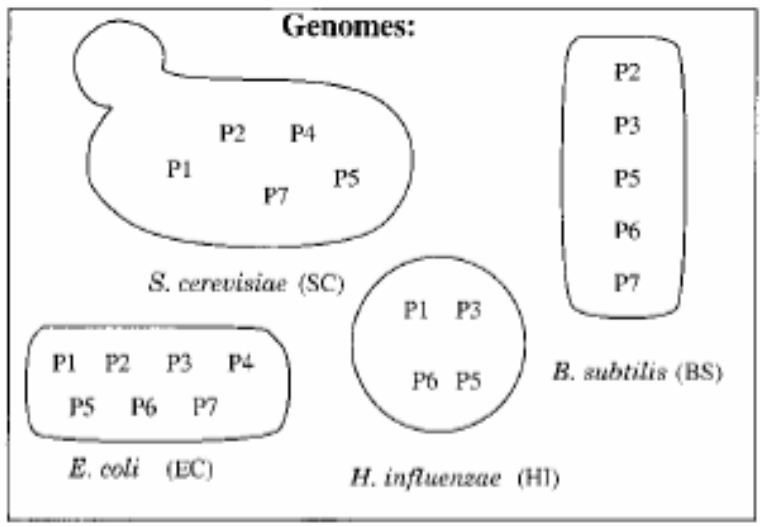
# What if there is no useful seq homolog?

- **Guilt by other types of association!**
  - Domain modeling (e.g., HMMPFAM)
  - ✓ Similarity of phylogenetic profiles
  - ✓ Similarity of dissimilarities (e.g., SVM-PAIRWISE)
  - Similarity of subcellular co-localization & other physico-chemico properties (e.g., PROTFUN)
  - Similarity of gene expression profiles
  - ✓ Similarity of protein-protein interaction partners
  - ...
  - Fusion of multiple types of info

# Phylogenetic profiling

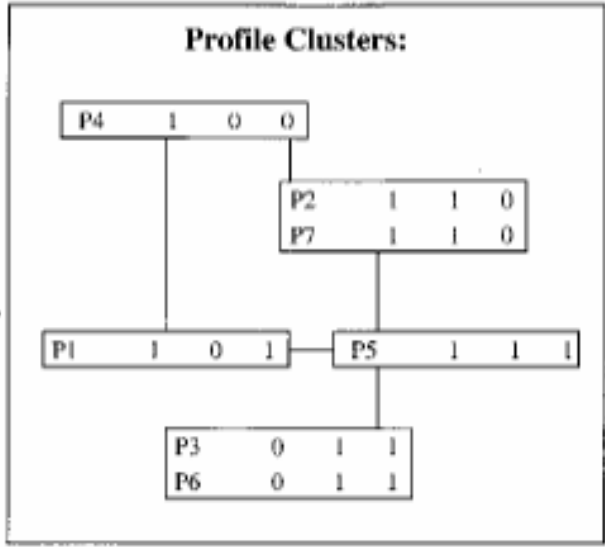
Pellegrini et al., *PNAS*, 96:4285--4288, 1999

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



**Phylogenetic Profile:**

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



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

# Phylogenetic profiling: How it works

# Phylogenetic profiling: P-value

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

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

where

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

**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

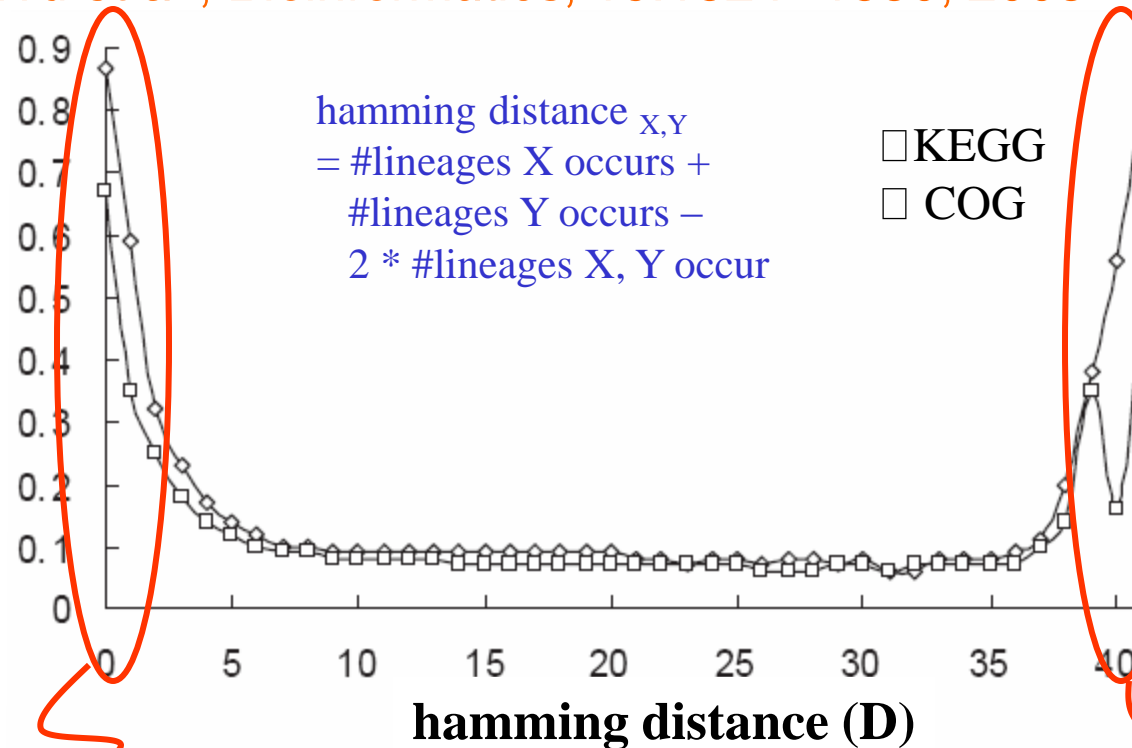
Keyword	No. of non-homologous proteins in group	No. neighbors in keyword group	No. neighbors in random group
Ribosome	60	197	27
Transcription	36	17	10
tRNA synthase and ligase	26	11	5
Membrane proteins*	25	89	5
Flagellar	21	89	3
Iron, ferric, and ferritin	19	31	2
Galactose metabolism	18	31	2
Molybdoterin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






# Guilt by association of dissimilarities



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



“unknown” is an “apple”!

	 Orange <sub>1</sub>	 Banana <sub>1</sub>	...
 Apple <sub>1</sub>	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 <sub>2</sub>	Color = orange vs orange Skin = rough vs rough Size = small vs small Shape = round vs round	Color = orange vs yellow Skin = rough vs smooth Size = small vs small Shape = round vs oblong	...
 Unknown <sub>1</sub>	Color = red vs orange Skin = smooth vs rough Size = small vs small Shape = round vs round	Color = red vs yellow Skin = smooth vs smooth Size = small vs small Shape = round vs oblong	...
...	...	...	...

# SVM-Pairwise framework

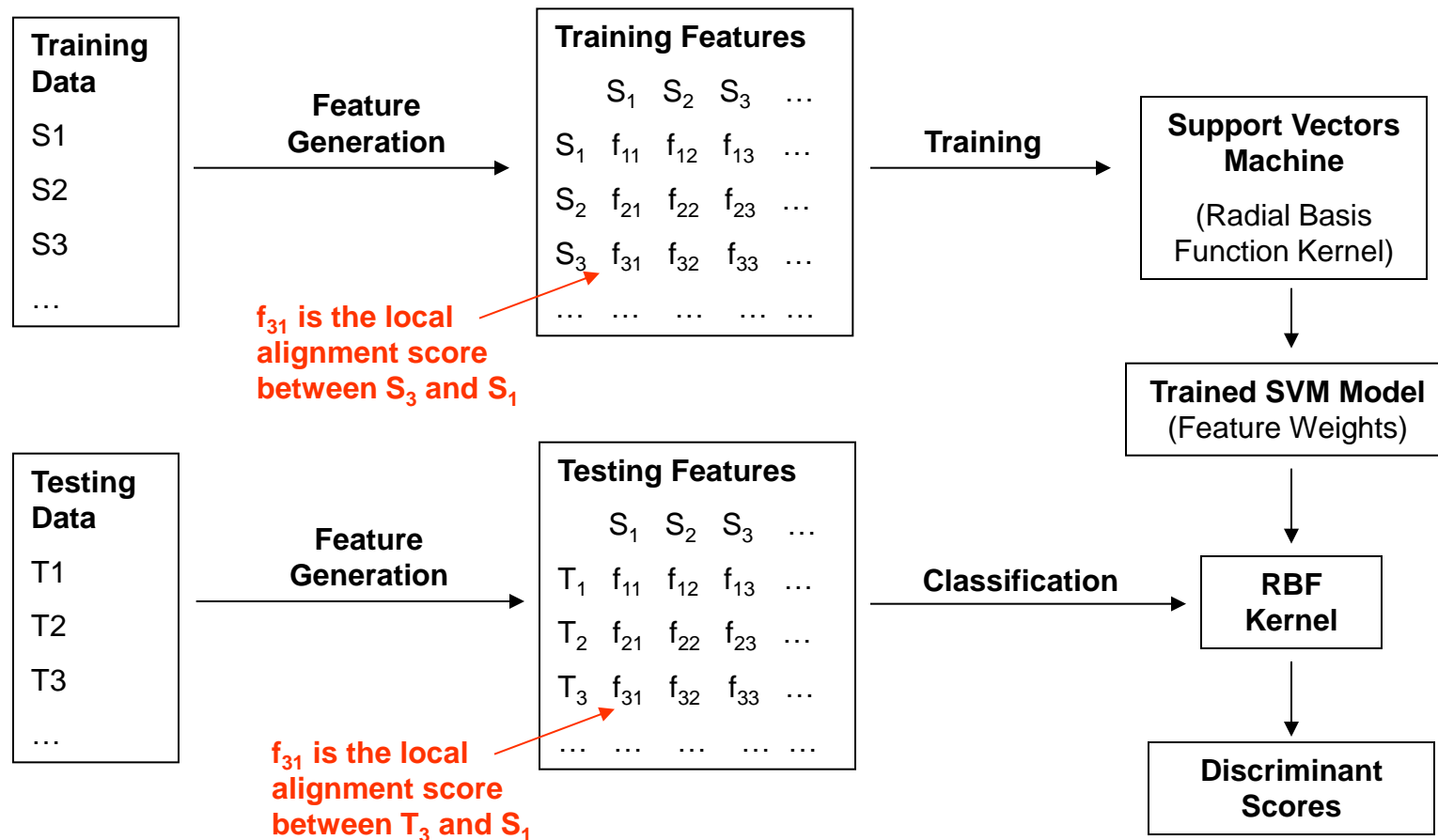
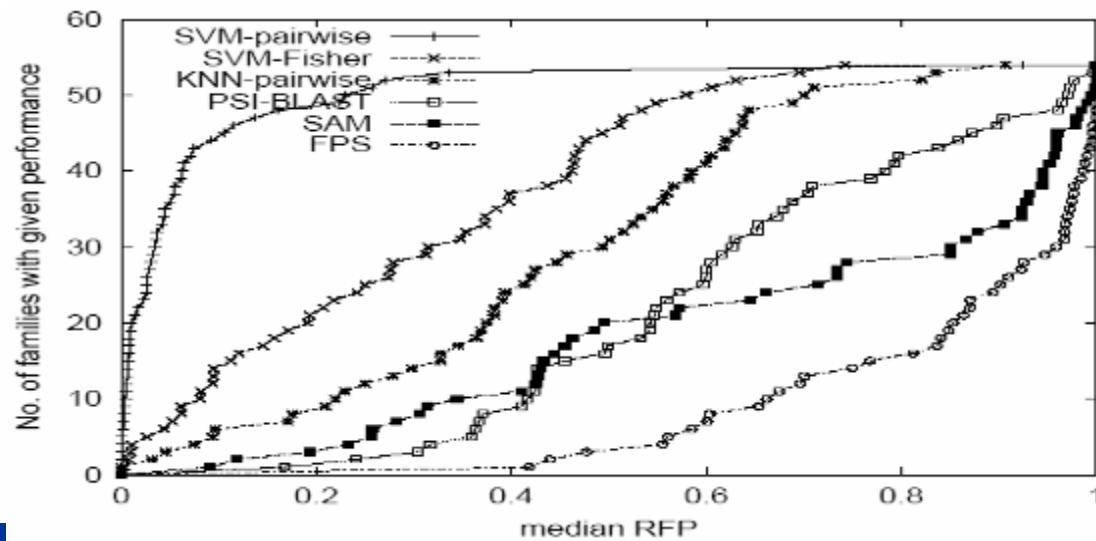
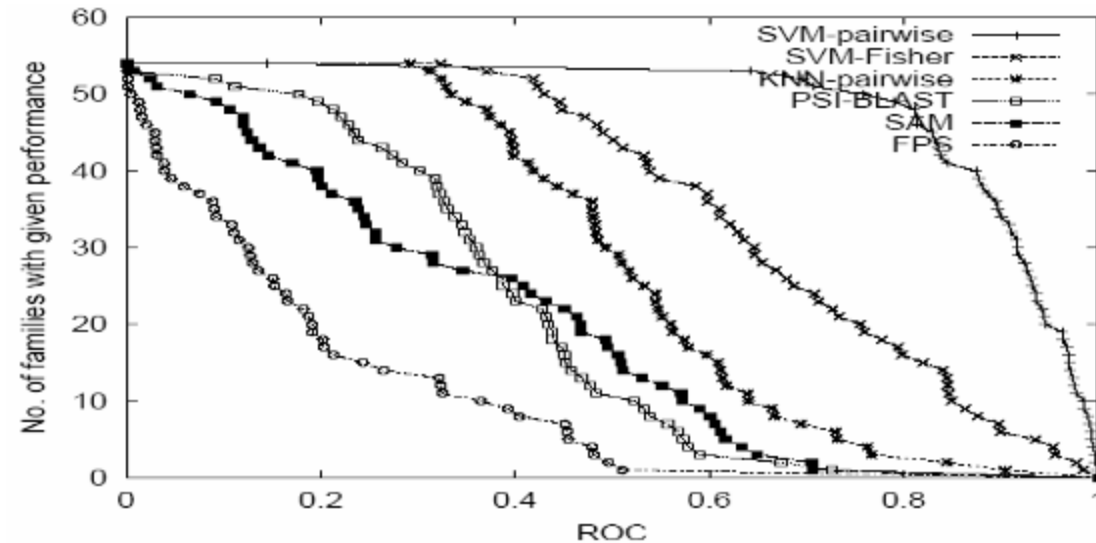


Image credit: Kenny Chua



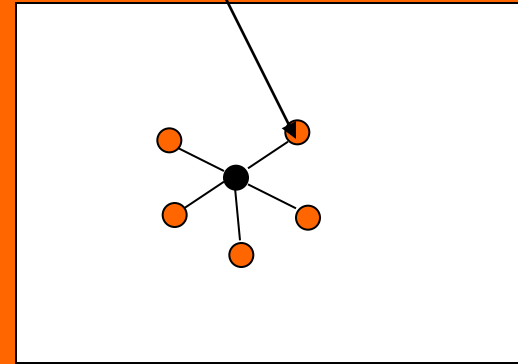
# Performance of SVM-Pairwise

- **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.

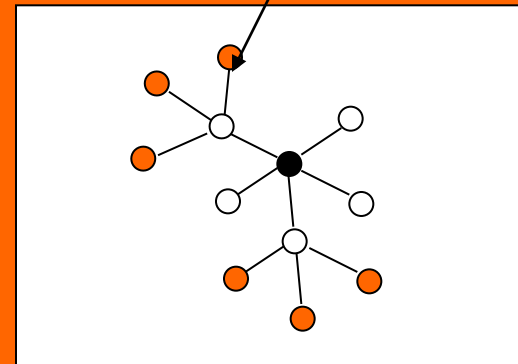


# Protein Function Prediction from Protein Interactions

Level-1 neighbour



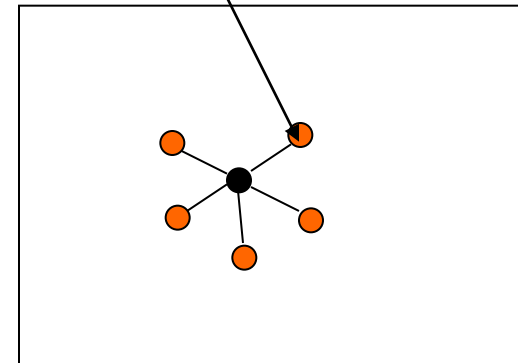
Level-2 neighbour



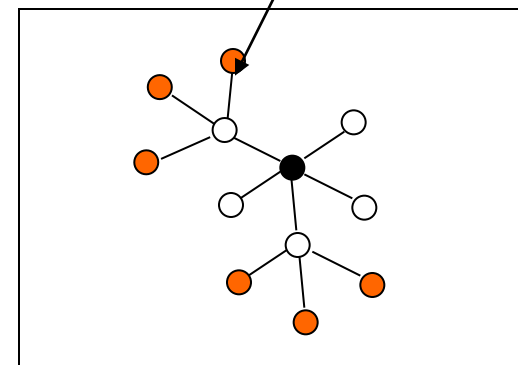
# Functional association thru interactions

- **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

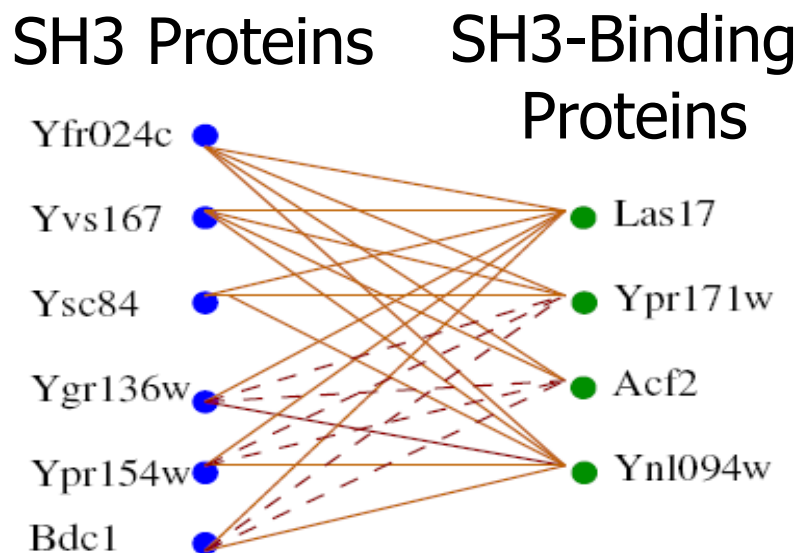
Level-1 neighbour



Level-2 neighbour

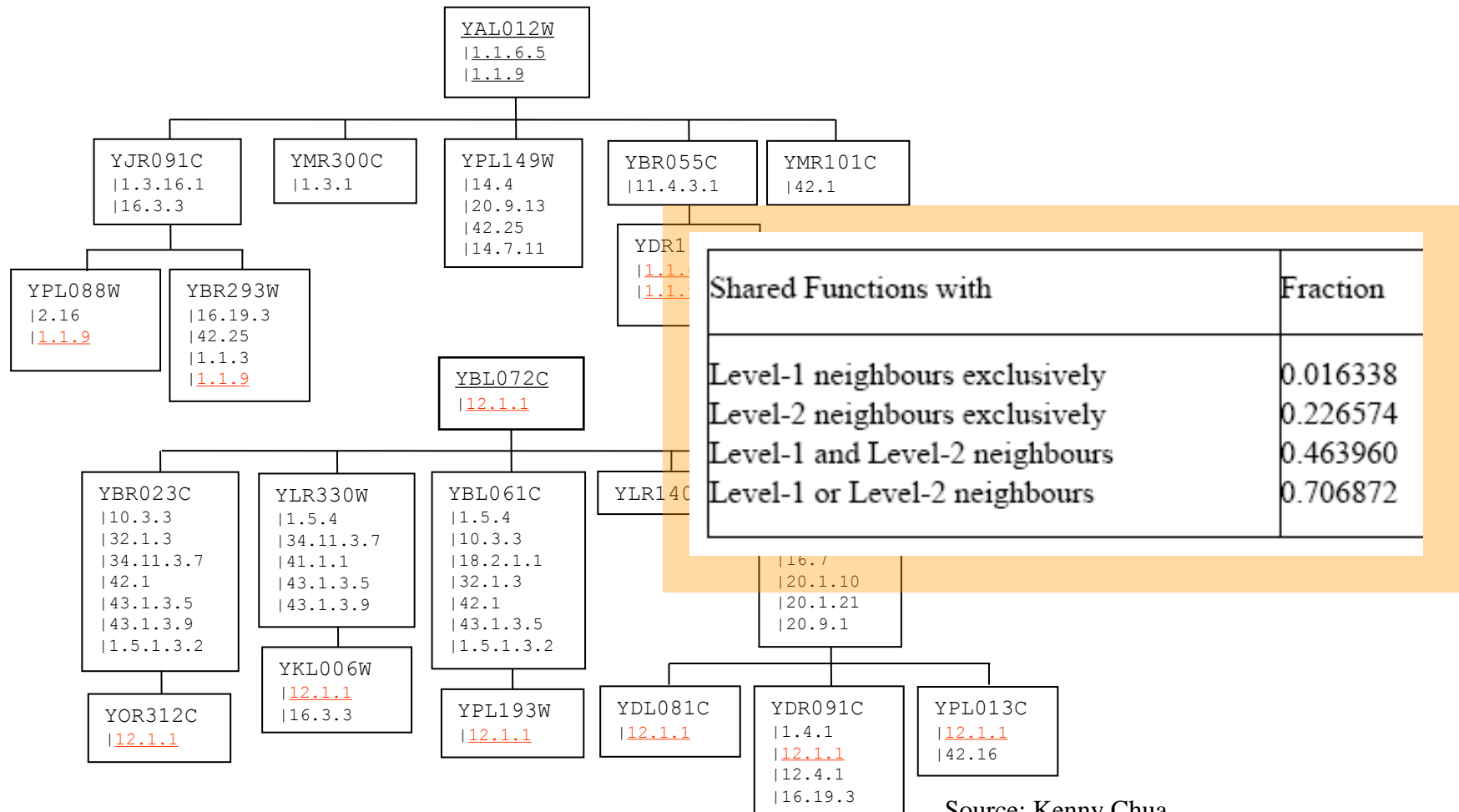


## An illustrative case of indirect functional association?



- 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



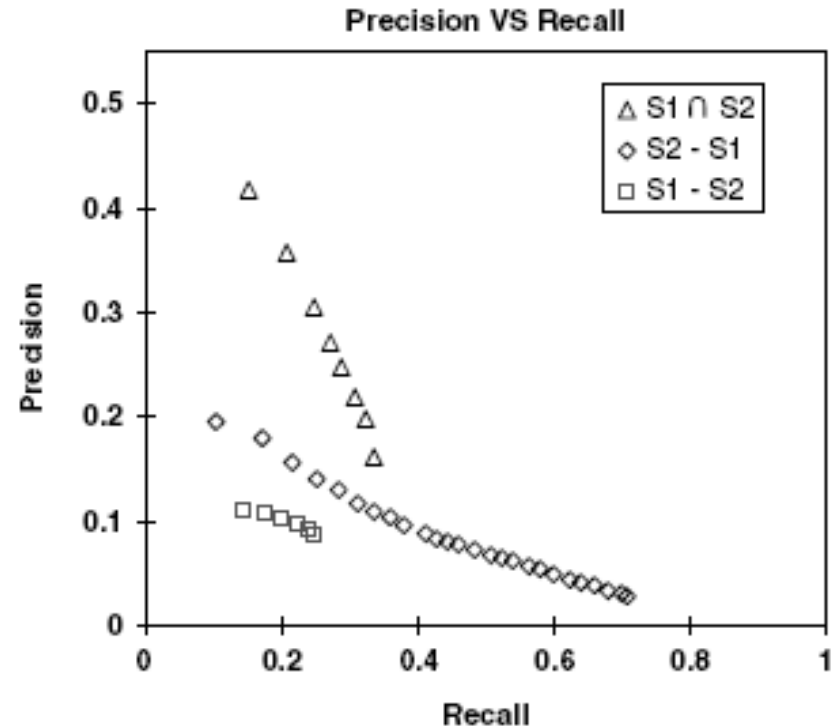
Source: Kenny Chua

# 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

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

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



⇒ “level-2 only” neighbours performs better

⇒ L1 ∩ L2 neighbours has greatest prediction power

# Functional similarity estimate: Czekanowski-Dice distance

- **Functional distance between two proteins** (Brun et al, 2003)

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

- $N_k$  is the set of interacting partners of  $k$
- $X \Delta Y$  is symmetric diff betw two sets  $X$  and  $Y$
- Greater weight given to similarity

⇒ **Similarity can be defined as**

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

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

# Functional similarity estimate: FS-weighted measure

- FS-weighted measure**

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

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

⇒ **Rewriting this as**

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



## Correlation w/ functional similarity

- Correlation betw functional similarity & estimates

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

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

# Reliability of expt sources

- **Diff Expt Sources have diff reliabilities**
  - Assign reliability to an interaction based on its expt sources (Nabieva et al, 2004)

- **Reliability betw u and v computed by:**

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

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

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

# Functional similarity estimate: FS-weighted measure with reliability

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

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

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

⇒ Rewriting

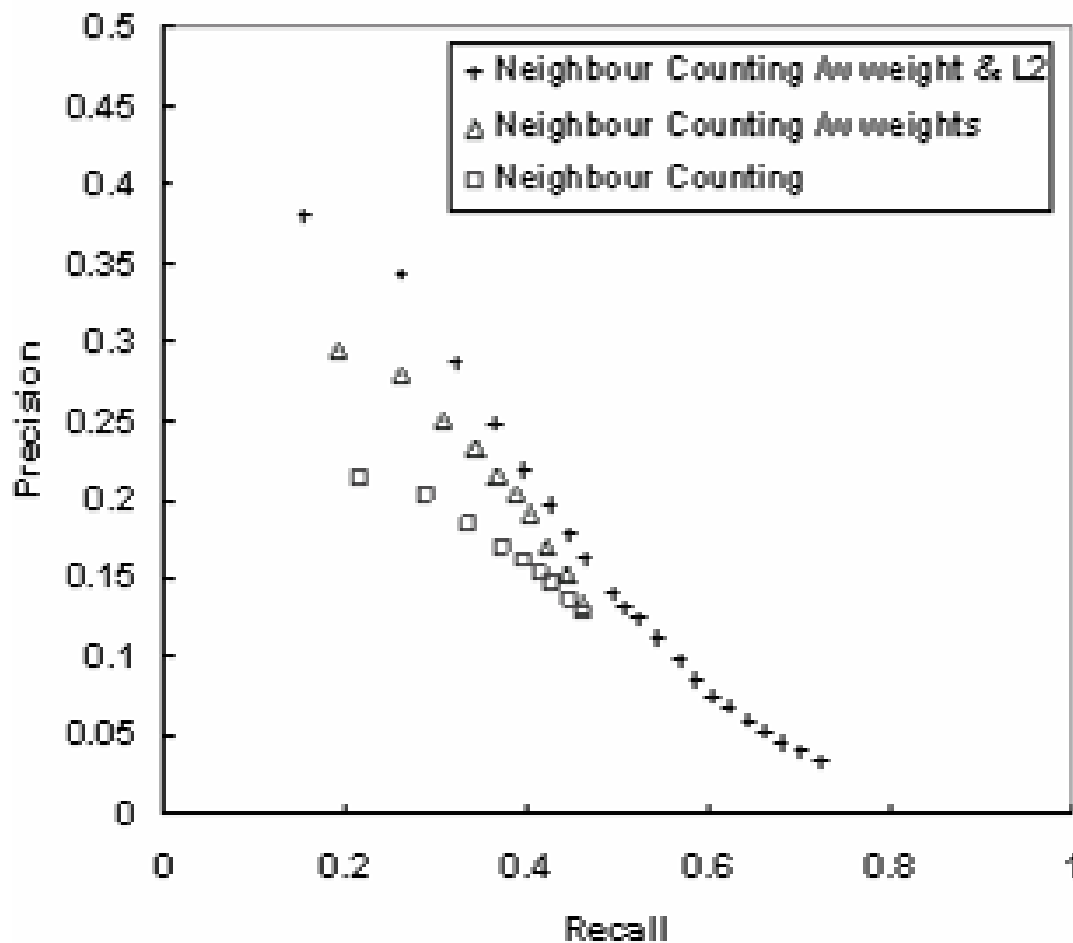
$$S(u, v) = \frac{2X}{2X + Y} \times \frac{2X}{2X + Z}$$

## Integrating reliabilities

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

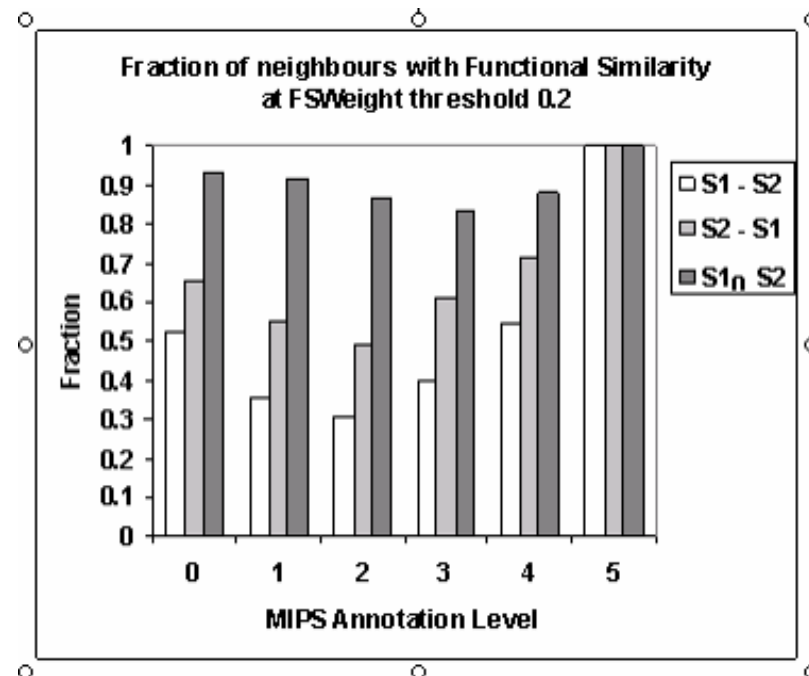
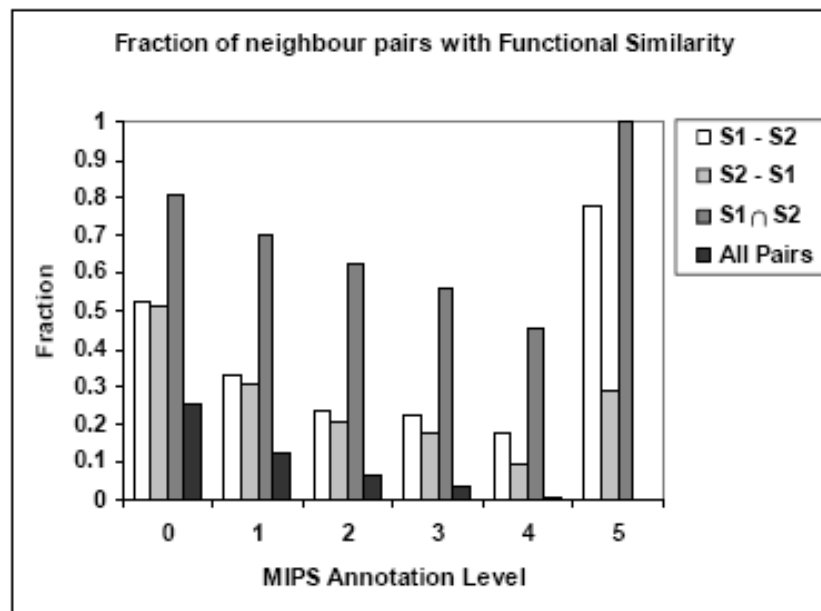
Neighbours	CD-Distance	FS-Weight	FS-Weight R
S <sub>1</sub>	0.471810	0.498745	0.532596
S <sub>2</sub>	0.224705	0.298843	0.375317
S <sub>1</sub> ∪ S <sub>2</sub>	0.224581	0.29629	0.363025

# 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**

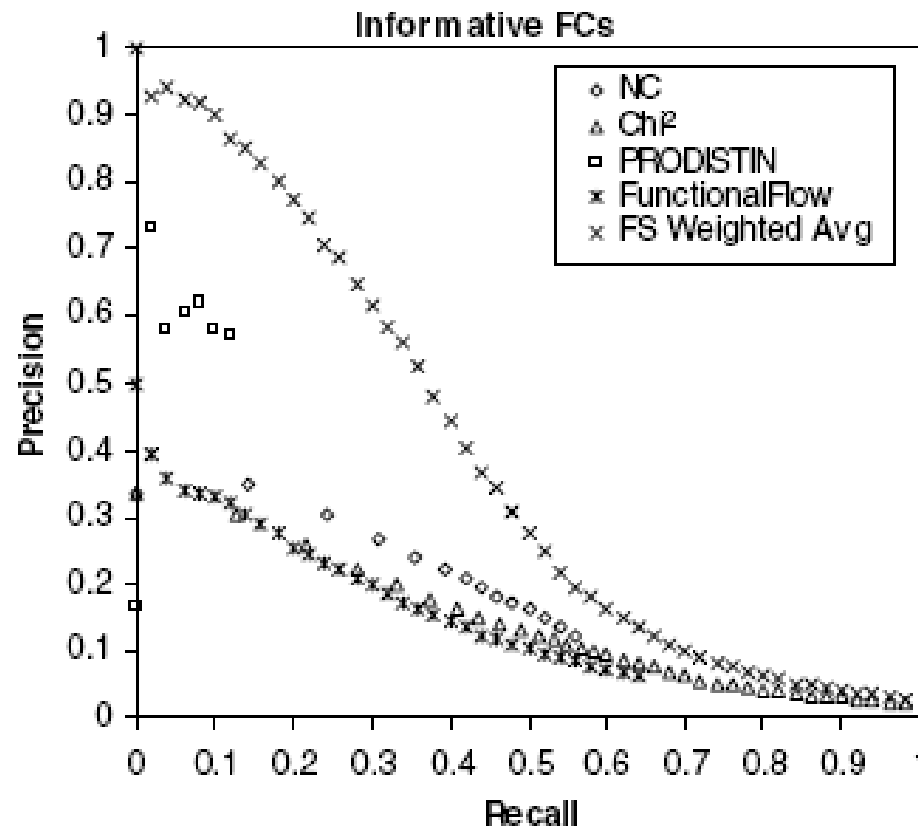
$$f_x(u) = \frac{1}{Z} \left[ \lambda r_{\text{int}} \pi_x + \sum_{v \in N_u} \left( S_{TR}(u, v) \delta(v, x) + \sum_{w \in N_v} S_{TR}(u, w) \delta(w, x) \right) \right]$$

- $r_{\text{int}}$  is fraction of all interaction pairs sharing function
- $\lambda$  is weight of contribution of background freq
- $\delta(k, x) = 1$  if  $k$  has function  $x$ , 0 otherwise
- $N_k$  is the set of interacting partners of  $k$
- $\pi_x$  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





# About the inventor: Chua Hon Nian

- **Chua Hon Nian**
  - PhD, NUS, 2008
  - Postdoc at Harvard & Univ of Toronto
  - 49<sup>th</sup> hottest paper in Computer Science published in 2006
  - Winner, DREAM2 challenge PPI subnetwork, 2007
  - Now Data Scientist at Data Robot



# 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|PTP&-D2
```

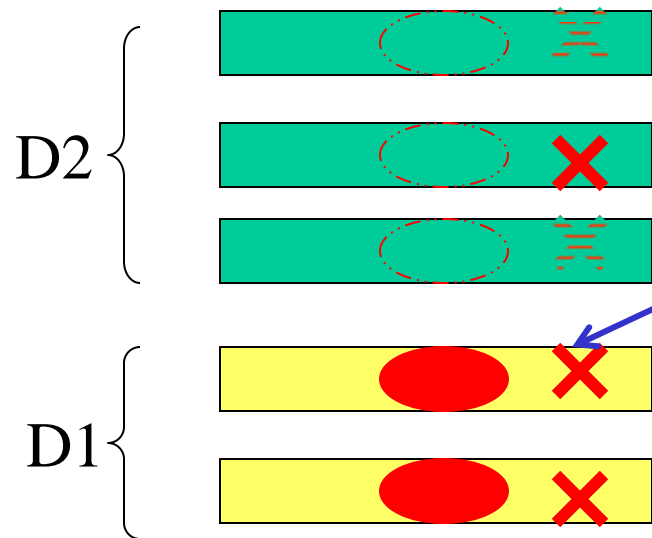
```
EEEFKKLTSIKIQNDKMRTGNLFPANMKKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASF
IDGYRQKDSYIASQGPLLHTIEDFWRMIWEWKSCSIVMLTELEERGQEKCAQYWPSDGLV
SYGDIITVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFHGWPEVGIIPSDGKGMISII
AAVQKQQQQSGNHPITVHCSAGAGRTGTFCALSTVLERVKAEGILDVVFQTVKSLRLQRPH
MVQTLQYEFQYKVVQYIDAFSDYANFK
```

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

## Emerging patterns of PTP D1 vs D2

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

# Emerging patterns of PTP D1 vs D2



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

```

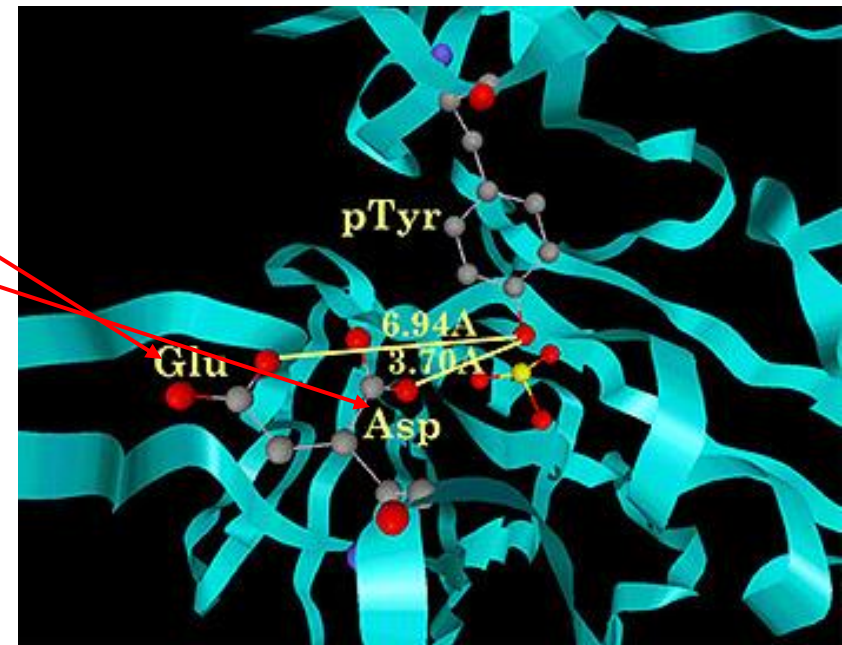
      ?  !  ?          ?          ?          ?          ?  ??
gi|00000|P  D2  QFHFHGWPEVGIPSDGKGMISIIAAVQKQQQQ-SGNHPITVHCSAGAGRTGTFCALSTVL
gi|126467|  QFHFTSWPDFGVPFTP I GMLKFLKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAML
gi|2499753  QFHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIML
gi|462550|  QYHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSML
gi|2499751  QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPEPILVHCSAGVGRTGTFIAIDRLI
gi|1709906  D1  QFQFTA WPDHGVP EHP T PFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAML
gi|126471|  QLHFTSWPDFGVPFTP I GMLKFLKVKTLNP--VHAGPIVVHCSAGVGRTGTFIVIDAMM
gi|548626|  QFHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIML
gi|131570|  QFHFTGWPDHGVPYHATGLLG FVRQVKS KSP--PNAGPLVVHCSAGAGRTGCFIVIDIML
gi|2144715  QFHFTSWPDHGVPDTTDLLINFRYLVRDYMKQSPPEPILVHCSAGVGRTGTFIAIDRLI
      * ..  ** . *. *          .          . ***** ****.. . ..
  
```

- 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  QFHFGWPEHGIPSDGK
gi|126467|    QFHFTSWPDPFGVFFTPIC
gi|2499753   QFHFTGWPDPHGVPYHATC
gi|462550|   QYHYTQWPDMGVPEYALI
gi|2499751   QFHFTSWPDPHGVPDTTDI
gi|1709906 D1  QFQFTAWPDPHGVPYHPTI
gi|126471|   QLHFTSWPDPFGVFFTPIC
gi|548626|   QFHFTGWPDPHGVPYHATC
gi|131570|   QFHFTGWPDPHGVPYHATC
gi|2144715   QFHFTSWPDPHGVPDTTDI
                * .. **.*.*
  
```



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

# Confirmation by mutagenesis

- **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?



# About the inventor: Prasanna Kolatkar

- **Prasanna Kolatkar**
  - Research Fellow, BIC, NUS, 1997-1999
  - Currently Senior Scientist at Qatar Biomedical Research Institute



# Concluding Remarks



## What have we learned?

- **General methodologies & applications**
  - Guilt by association for protein function inference
  - Invariants for active site discovery
  - Emerging patterns for mutation site discovery
- **Important tactics**
  - Genome phylogenetic profiling
  - SVM-Pairwise
  - Protein-protein interactions

Any Question?



# 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):95-114, 2000
- T. Hawkins and D. Kihara. Function prediction of uncharacterized proteins. *JBCB*, 5(1):1-30, 2007