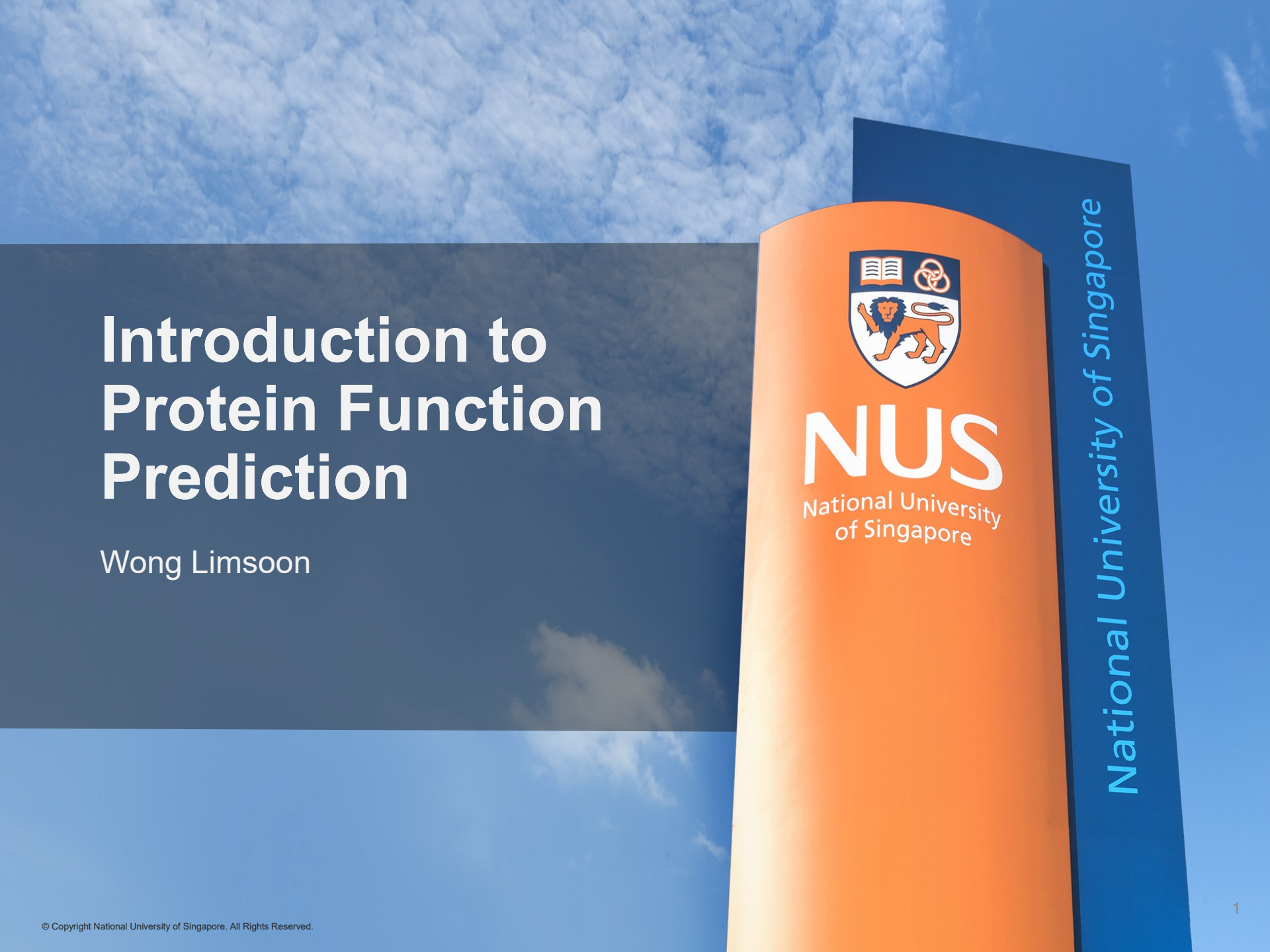
Introduction to Protein Function Prediction

Wong Limsoon



Plan

Sequence alignment

Guilt by association

Key mutation site discovery

What if no homology of known function is found?

Sequence comparison / alignment

Motivations for seq comparison

Evolution is related to changes in DNA

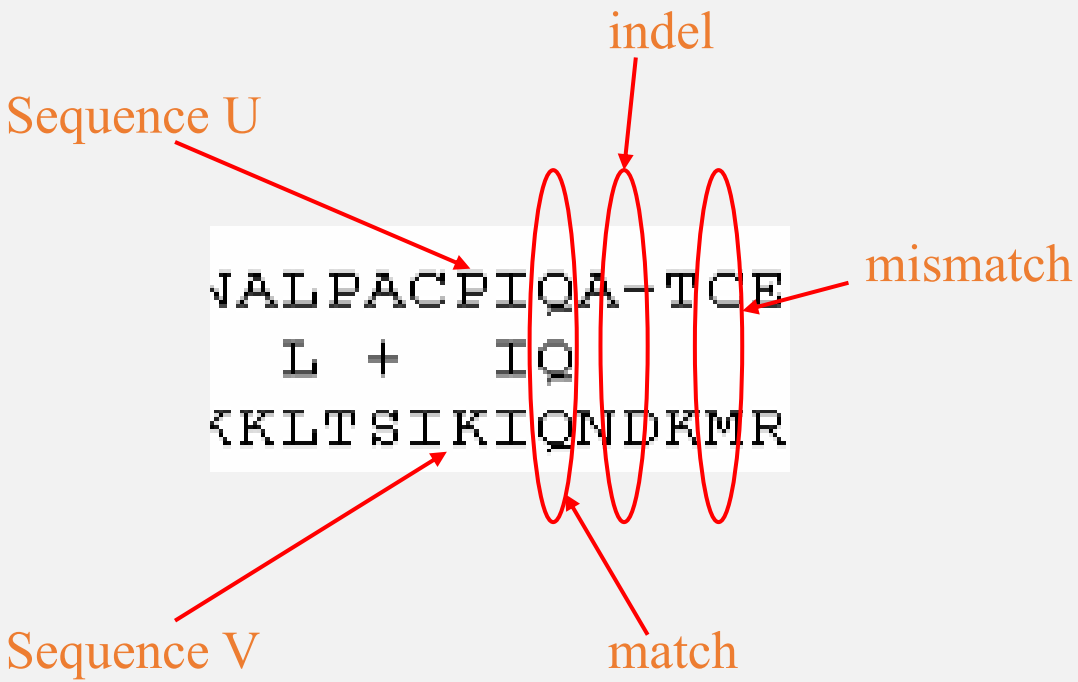
By comparing DNA sequences, we can infer evolutionary relationships between the sequences w/o knowledge of the evolutionary events themselves

Thus, sequence comparison is a 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 agree in two sequences



Sequence alignment: Poor example

Poor seq alignment shows few matched positions

⇒ The 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 has clusters of 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 betw Amicyanin & an 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

Guilt by association

Proteins

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

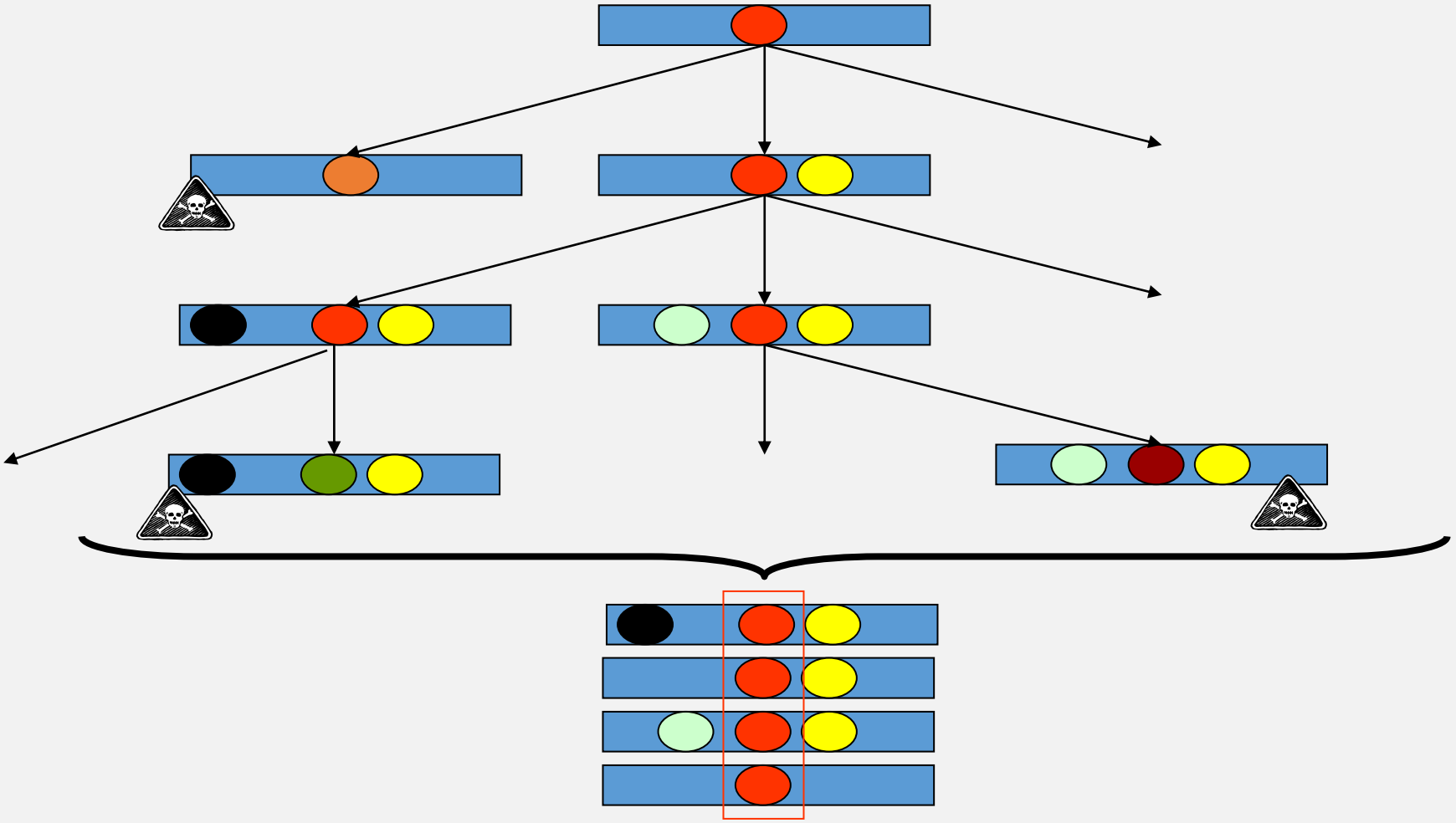
Proteins perform a wide variety of activities in the cell

Function assignment to protein seq

SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
VT

How to assign a function to a new protein sequence?

In the course of evolution…



Evolution takes time …

Let a = AFPHQHRVP

Let b = PQVYNIMKE

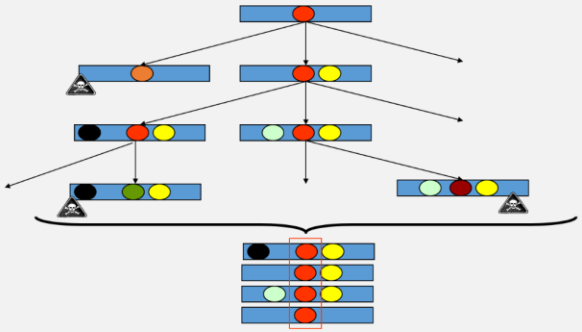
Suppose each generation differs from the previous by 1 residue

What is the max difference between the 2nd generation of a

What is the min difference between the 2nd generation of a and b?

The triumph of logic

In the course of evolution…



Evolution takes time …

Let a = AFPHQHRVP
Let b = PQVYNIMKE

Suppose each generation differs from the previous by 1 residue

What is the max difference between the 2nd generation of a

What is the min difference between the 2nd generation of a and b?

Two proteins inheriting their function from a common ancestor have very similar amino acid sequences

Discussion

How to guess the function of a protein?

Guilt by association: Caveats

Ensure that the effect of database size has been accounted for

Ensure that the function of the homolog 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 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

$\Rightarrow$ Need to correct for database size if your seq comparison prog does not do that!

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

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

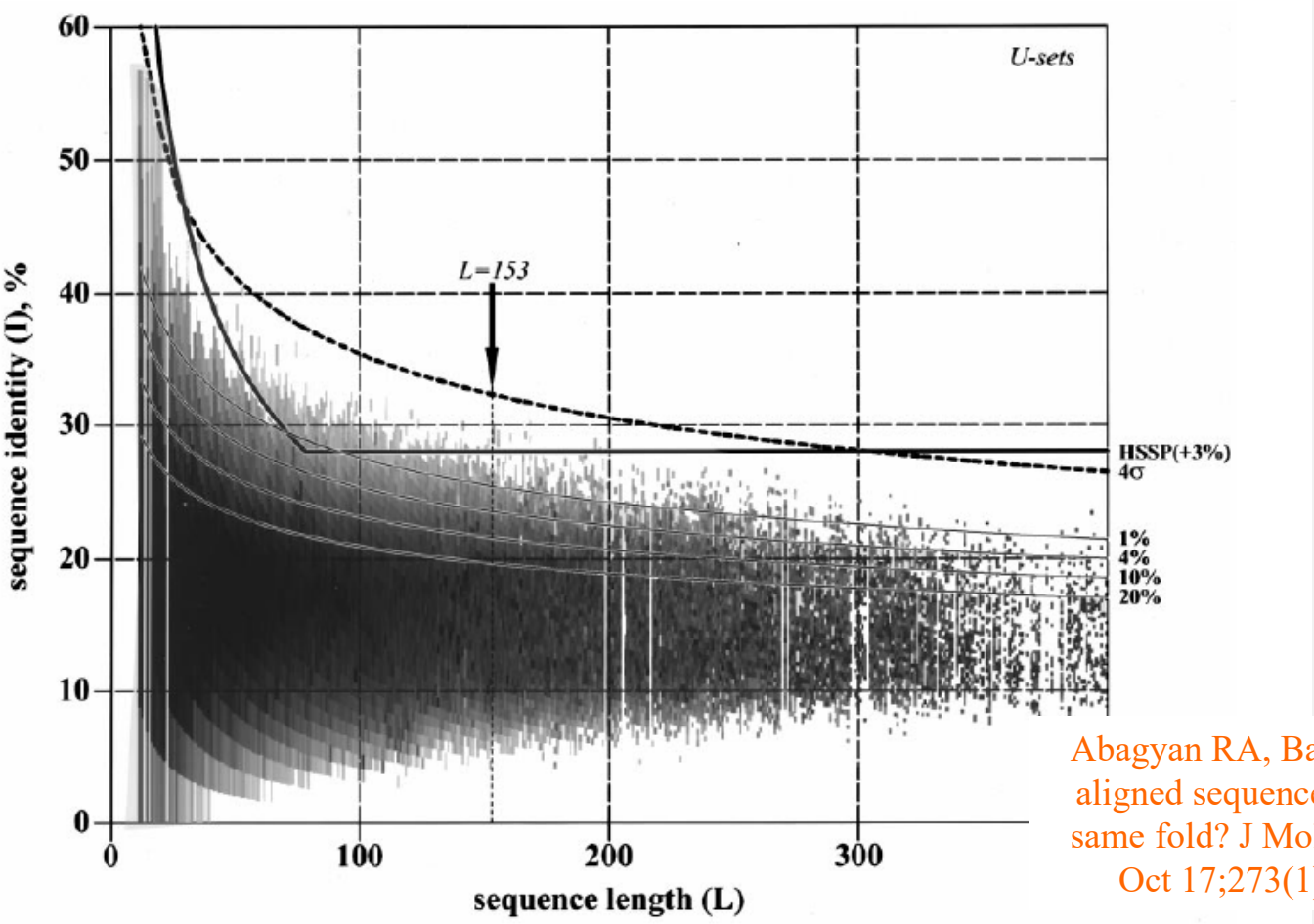
Alignment of two such regions achieves high score purely due to segment composition

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

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

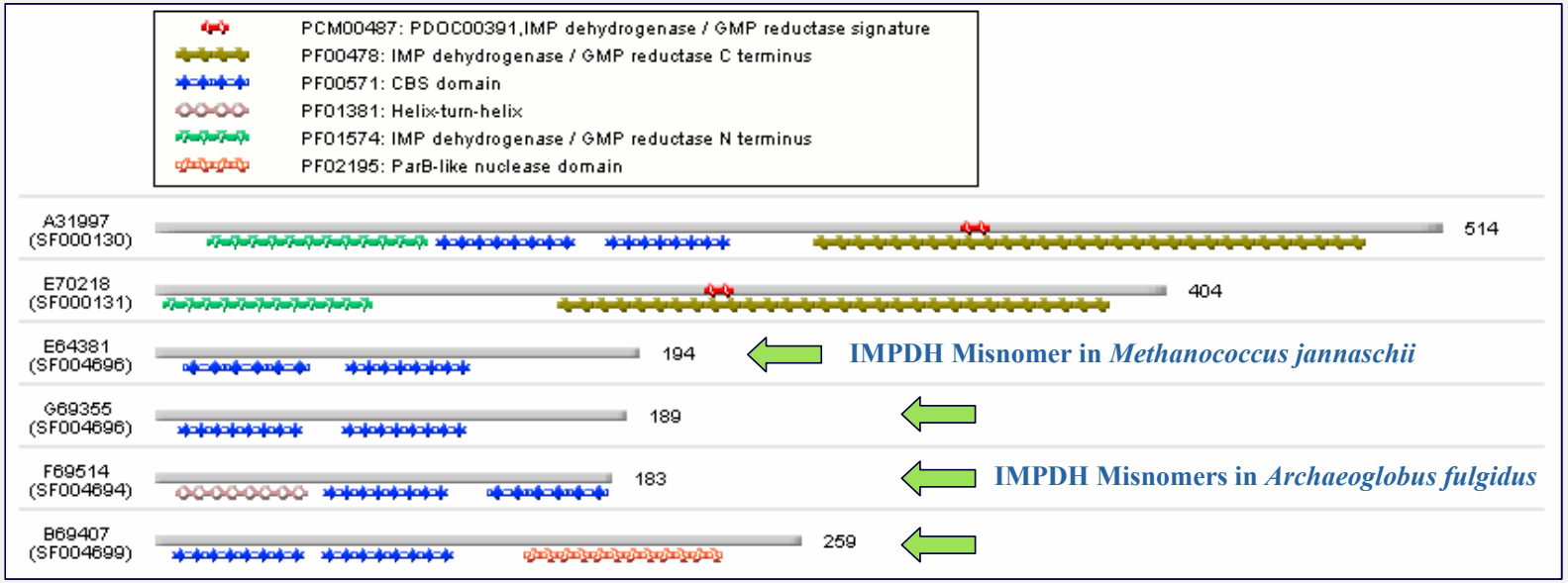
Examples of invalid function assignment: IMP dehydrogenases (IMPDH)

18 entries were found

ID	Organism	PIR	Swiss-Prot/TrEMBL	RefSeq/GenPept
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	Archaeoglobus fulgidus	B69407 MJ0188 homolog *ALT_NAMES*: inosine monophosphate dehydrogenase homolog [misnomer]	O29009 Hypothetical protein AF1259	g2649320 inosine monophosphate dehydrogenase, putative NP_070087 inosine monophosphate putative
NF00197776	Thermo			-5-monophosphate related protein monophosphate d protein
NF00414709	Methanothermobacter thermautotrophicus	G69030 MJ0653 homolog MTH1226 *ALT_NAMES*: inosine-monophosphate dehydrogenase related protein V [misnomer]	O27294 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN V	g2622337 inosine-5'-monophosphate 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



Typical IMPDHs have 2 IMPDH domains that form the catalytic core and 2 CBS domains.

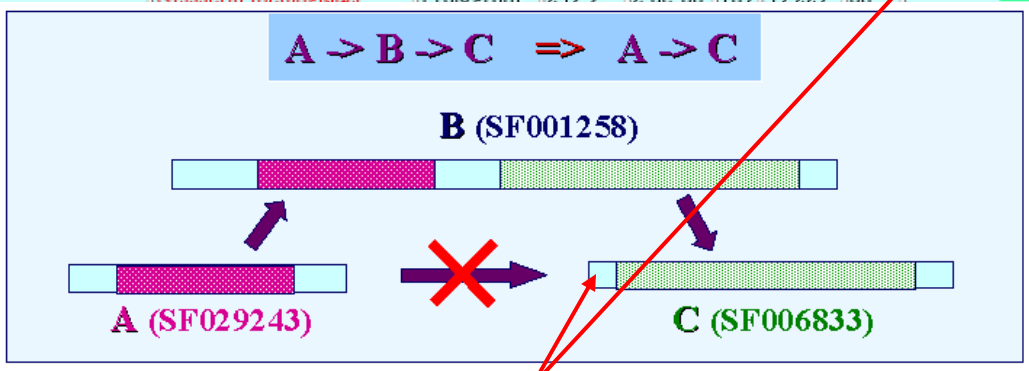
A less common but functional IMPDH (E70218) lacks the CBS domains.

Misnomers show similarity to the CBS domains

Invalid transitive assignment

Root of invalid transitive assignment

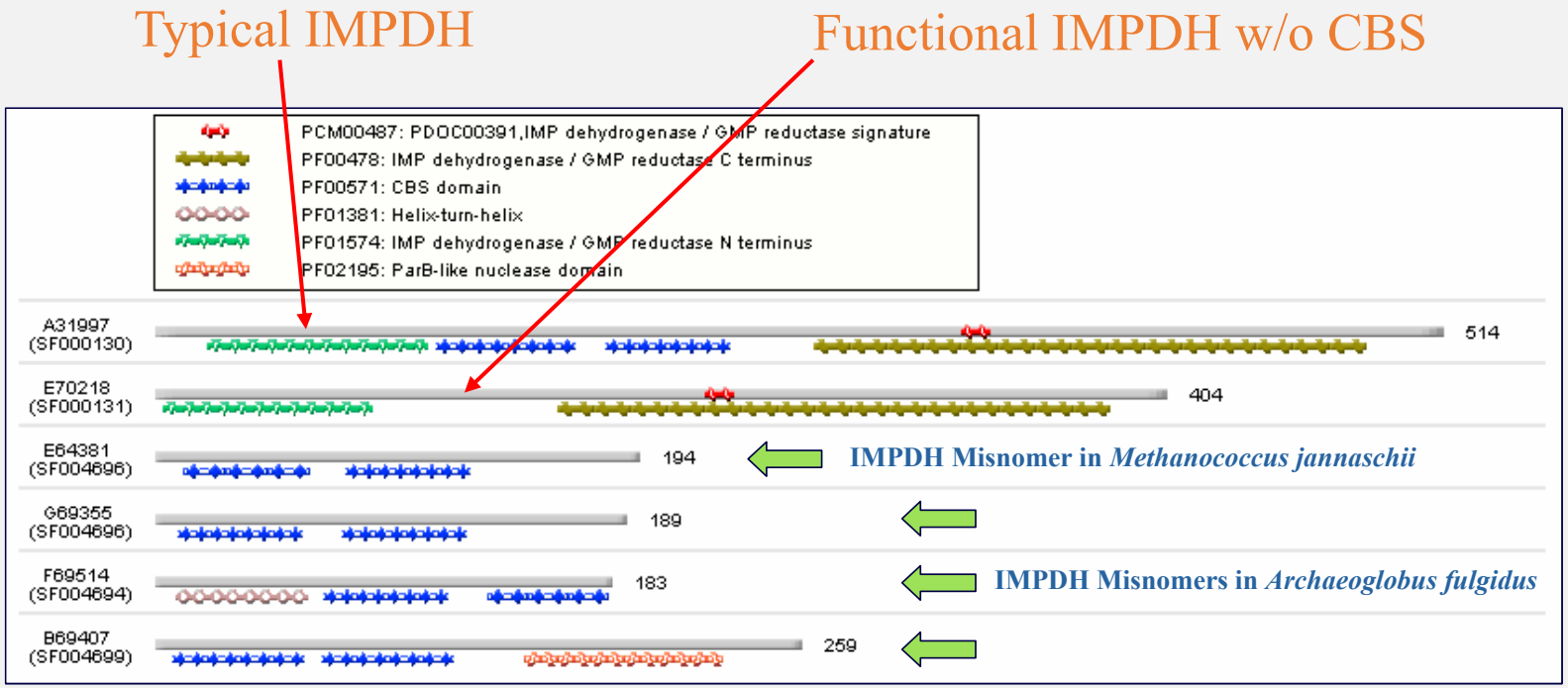
B	H70468	SF001258	051440	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Aquifex aeolicus	Prok/other	594.3	4.8e-26	205	39.086	197
	S76963	SF001258	039935	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (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



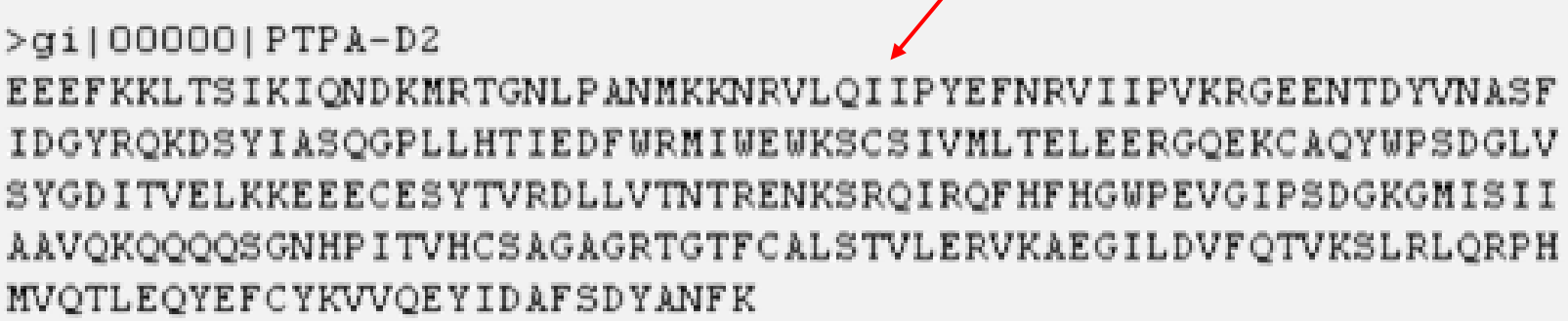
Most IMPDHs have 2 IMPDH and 2 CBS domains

Some IMPDH (E70218) lacks CBS domains

$\Rightarrow$ IMPDH domain is the characterising pattern of IMPDH

Key mutation site discovery

Sequence from a typical PTP domain D2



Discussion

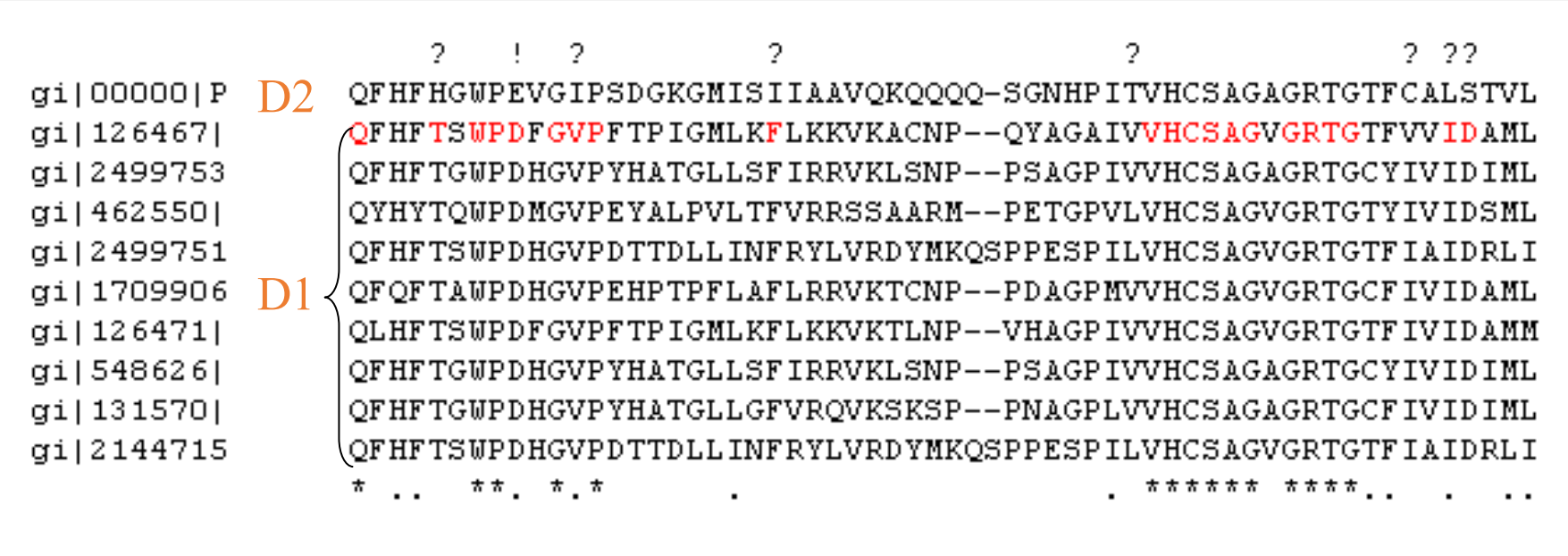
Some PTPs have 2 PTP domains

PTP domain D1 has more activity than PTP domain D2

Why? How do you figure that out?

Key mutation site: PTP D1 vs D2

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



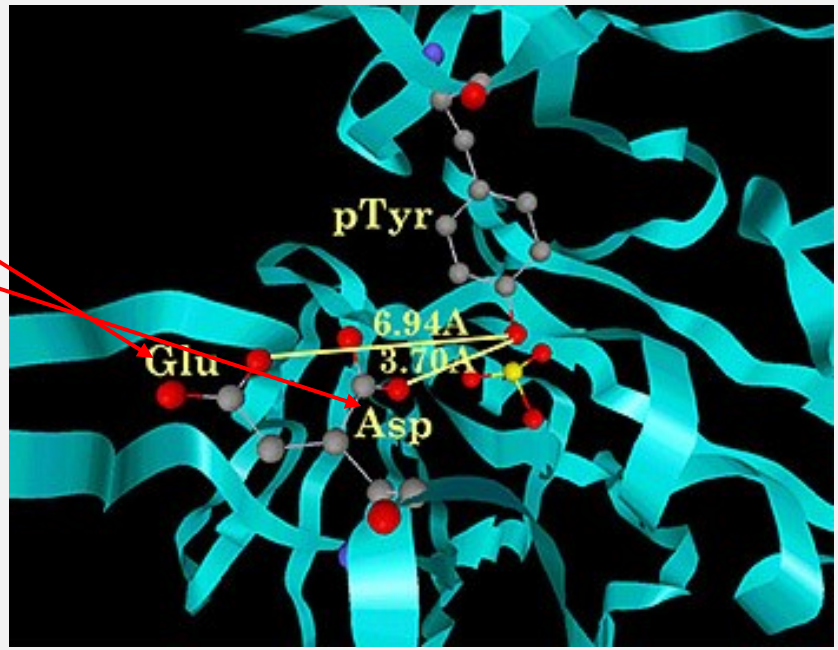
Positions marked by “!” and “?” are likely places responsible for reduced PTP activity

PTP D1 agree on them; PTP D2 disagree on them

Key mutation site: PTP D1 vs D2

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

```
                          ?    !    ?
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 modelling predicts they induce large distortion to structure

Confirmation by mutagenesis

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

Why do you need this 2-way expt?

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

…

Hidden Markov Model for biological seqs

https://webpages.math.luc.edu/~tobrien/courses/bioinf/krogh.pdf

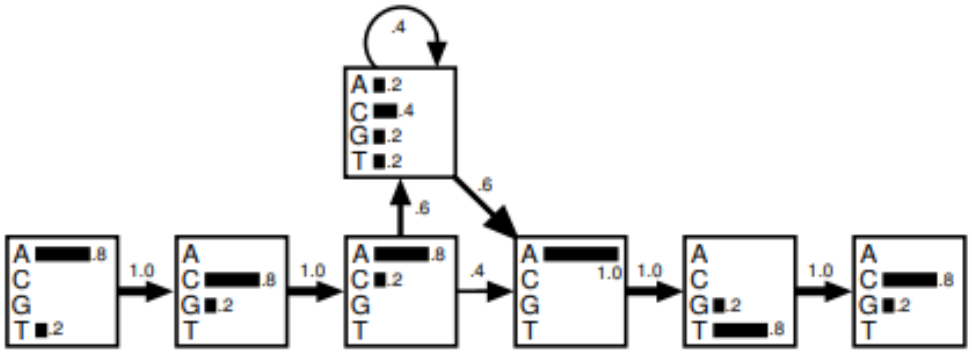
A multiple alignment of sequences of the same family/function

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

(I use DNA only because of the smaller number of letters than for amino acids).

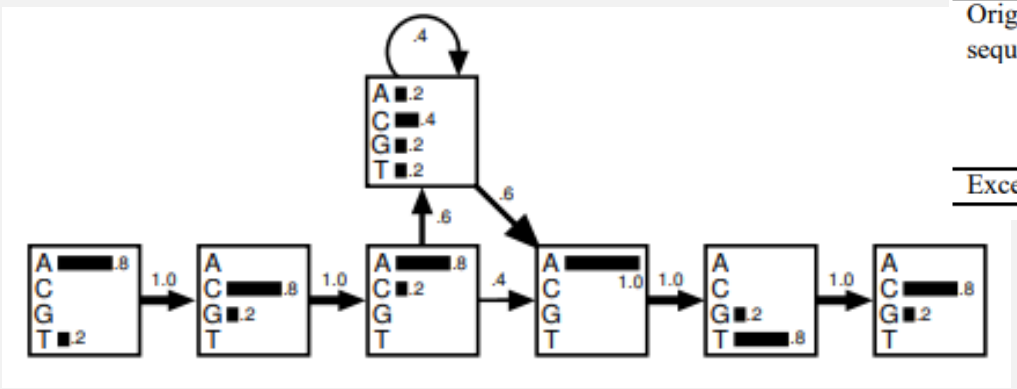
For each protein family, create multiple alignment of its members

HMM derived from the multiple alignment



Scoring by HMM

https://webpages.math.luc.edu/~tobrien/courses/bioinf/krogh.pdf



	Sequence	Probability ×100	Log odds
Consensus	A C A C - - A T C	4.7	6.7
Original sequences	A C A - - - A T G	3.3	4.9
	T C A A C T A T C	0.0075	3.0
	A C A C - - A G C	1.2	5.3
	A G A - - - A T C	3.3	4.9
	A C C G - - A T C	0.59	4.6
Exceptional	T G C T - - A G G	0.0023	-0.97

$$P(\texttt{ACACATC}) = 0.8 \times 1 \times 0.8 \times 1 \times 0.8 \times 0.6 \times 0.4 \times 0.6 \times 1 \times 1 \times 0.8 \times 1 \times 0.8 \simeq 4.7 \times 10^{-2}.$$

The prob depends on seq length; not very convenient score for interpretation

Log odds obtained by comparing to a null model of the same length

$$\text{log-odds for sequence } S = \log \frac{P(S)}{0.25^L} = \log P(S) - L \log 0.25.$$

For a 4-letter alphabet, each letter has 0.25 prob to occur

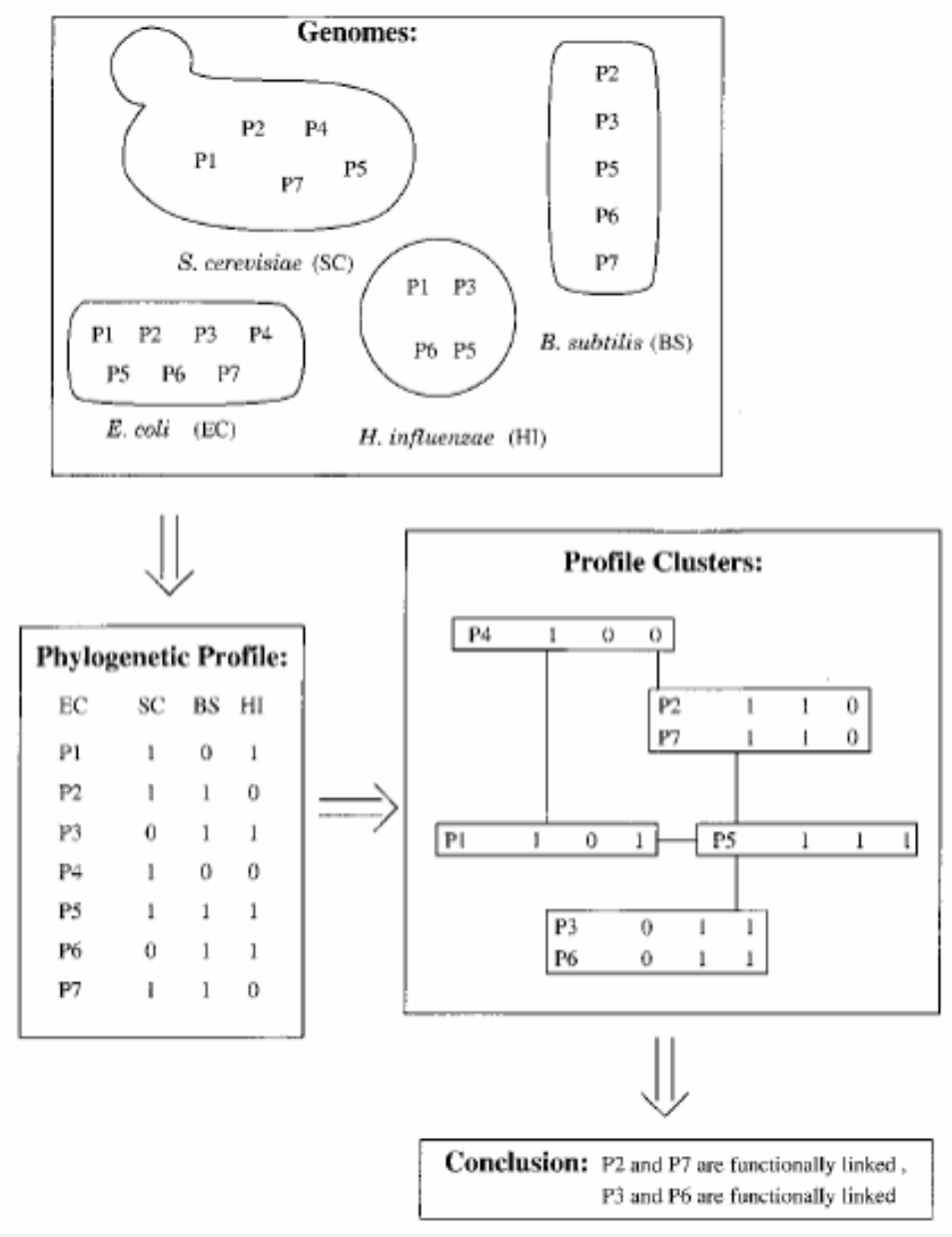
Phylogenetic profiling

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

Genes (and hence proteins) with identical patterns of occurrence across phyla tend to function together

$\Rightarrow$ 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

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

Phylogenetic profiles: Evidence

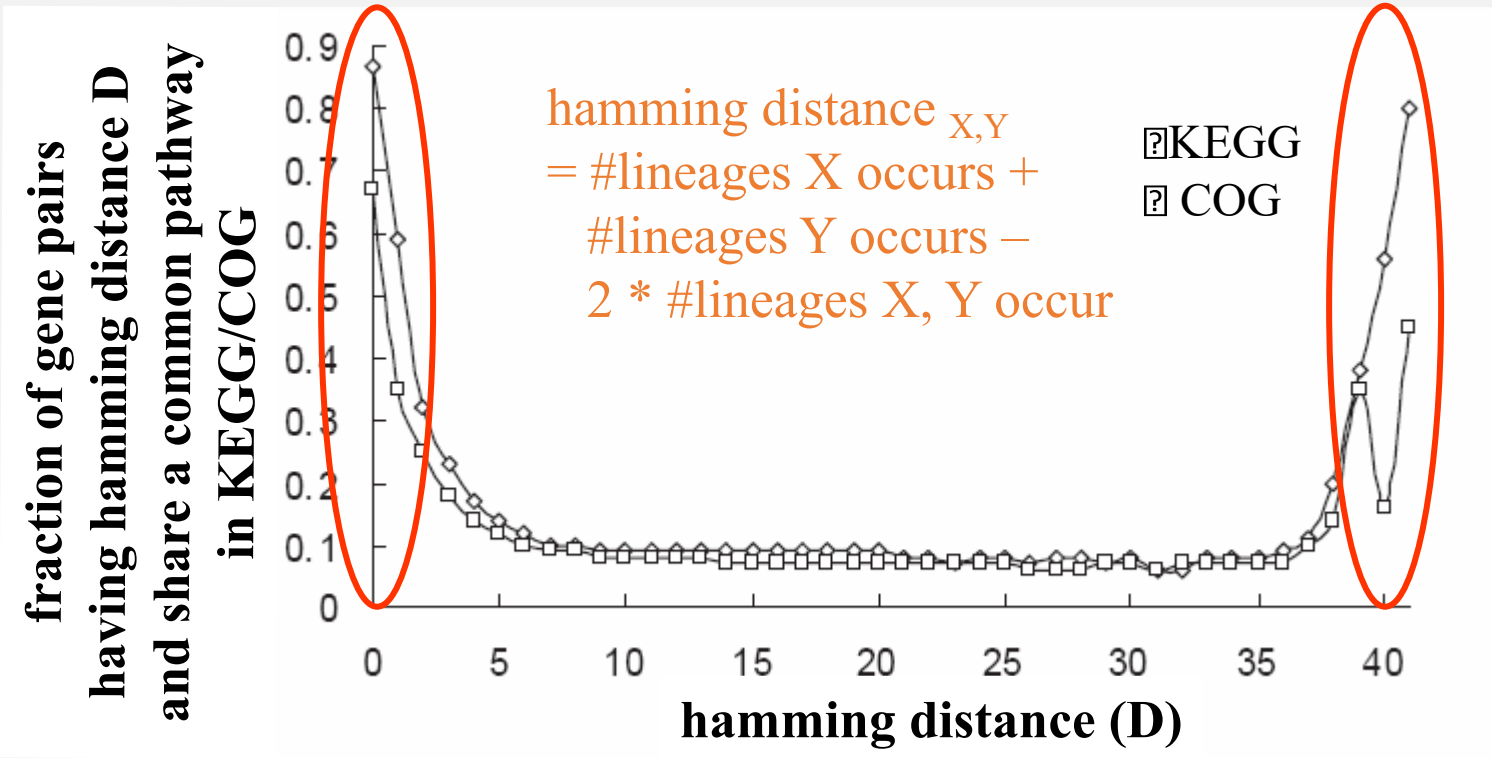
Pellegrini et al., PNAS, 96:4285--4288, 1999

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



Explain the two red ovals. Any surprise there?

Concluding remarks

Discussion

What have we learned?

Suggested readings

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

Suggested readings

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

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