

# Techniques & Applications of Sequence Comparison

**Limsoon Wong**  
**30 August 2006**



# Lecture Plan

- **Recap on sequence alignment**
- **Popular tools**
  - BLAST, Pattern Hunter
- **Applications**
  - Homologs, Active sites, Key mutation sites,  
Looking for SNPs, Determining origin of species,  
...

# Recap on Sequence Alignment



# Sequence Comparison: Motivations

- **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**

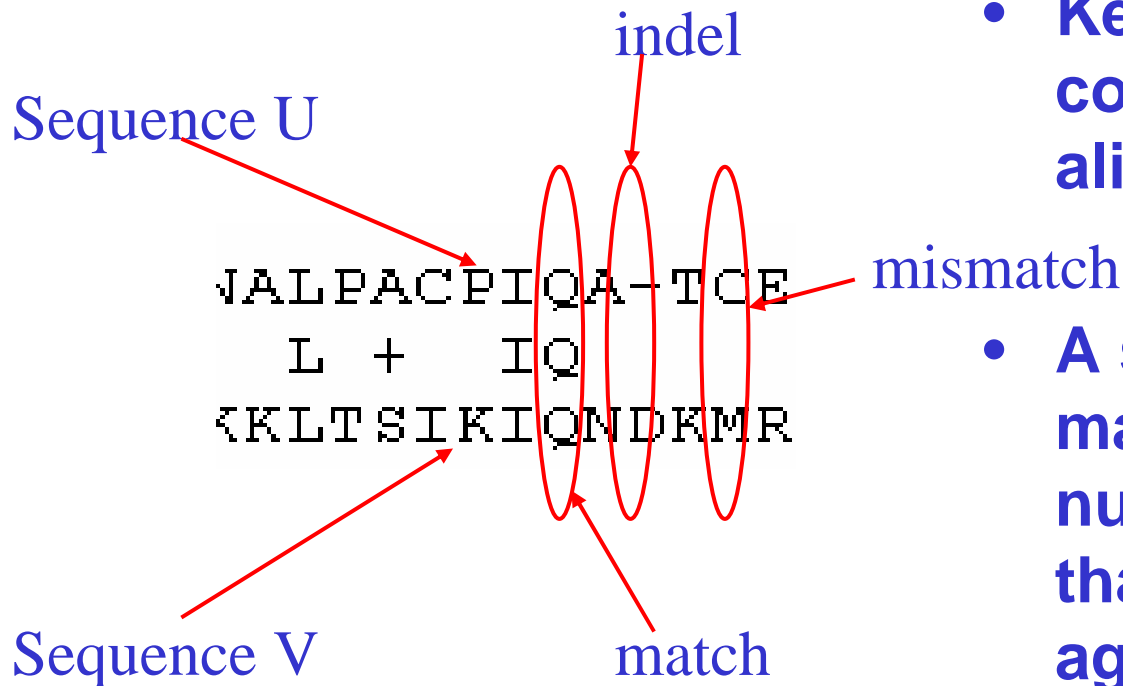
# Earliest Research in Seq Comparison

Source: Ken Sung

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

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

# Alignment



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

# 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      MPHNVHVFVAGVLGEAALKGPMMKKEQAYS�TFTEAGTYDYHCTPHPFMRGKVVVE
                ...:  .  :::  ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLMQRSAGLYGSLI
                70      80      90      100      110      120
```

No obvious match between  
Amicyanin and Ascorbate Oxidase

## 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



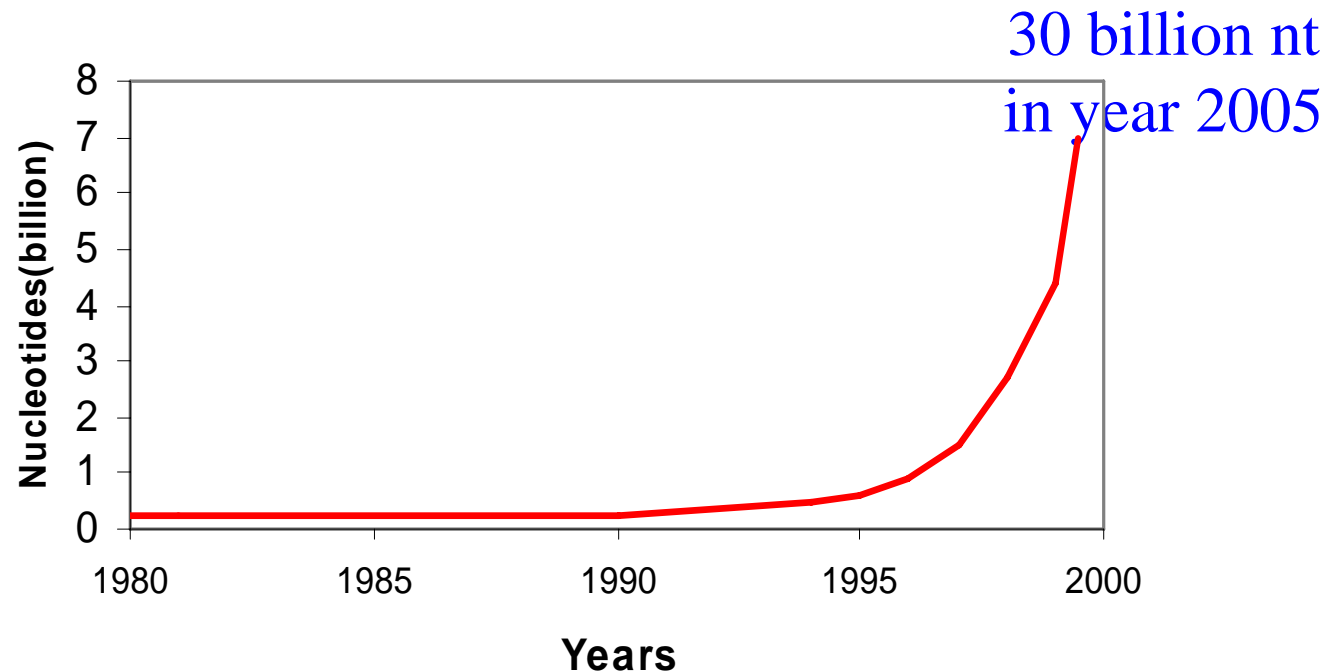
# Popular Tools for Sequence Comparison: FASTA, BLAST, Pattern Hunter

## Acknowledgements:

Some slides here are “borrowed” from Bin Ma & Dong Xu



# Scalability of Software



- Increasing number of sequenced genomes: yeast, human, rice, mouse, fly, ...
- ⇒ S/w must be “linearly” scalable to large datasets

# Need Heuristics for Sequence Comparison

- Time complexity for optimal alignment is  $O(n^2)$ , where  $n$  is sequence length
- ⇒ Given current size of sequence databases, use of optimal algorithms is not practical for database search
- Heuristic techniques:
  - BLAST
  - FASTA
  - Pattern Hunter
  - MUMmer, ...
- Speed up:
  - 20 min (optimal alignment)
  - 2 min (FASTA)
  - 20 sec (BLAST)

# Basic Idea: Indexing & Filtering

- **Good alignment includes short identical, or similar fragments**
  - ⇒ **Break entire string into substrings, index the substrings**
  - ⇒ **Search for matching short substrings and use as seed for further analysis**
  - ⇒ **Extend to entire string find the most significant local alignment segment**

# BLAST in 3 Steps

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

- **Word matching**
- **Similarity matching of words (3 aa's, 11 bases)**
  - no need identical words
- **If no words are similar, then no alignment**
  - won't find matches for very short sequences
- **MSP: Highest scoring pair of segments of identical length. A segment pair is locally maximal if it cannot be improved by extending or shortening the segments**
- **Find alignments w/ optimal max segment pair (MSP) score**
- **Gaps not allowed**
- **Homologous seqs will contain a MSP w/ a high score; others will be filtered out**

# BLAST in 3 Steps

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

## Step 1

- For the query, find the list of high scoring words of length  $w$

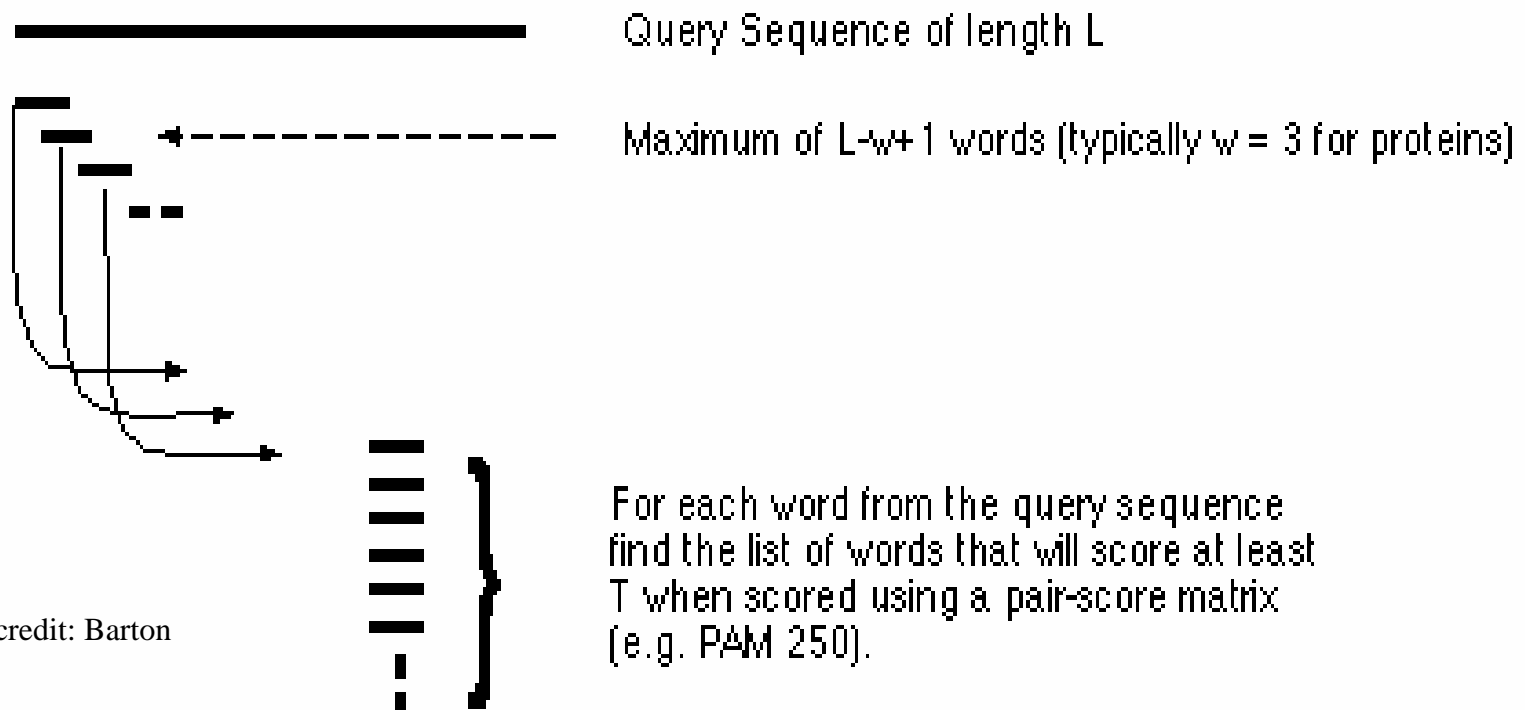


Image credit: Barton

# BLAST in 3 Steps

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

## Step 2

- Compare word list to db & find exact matches

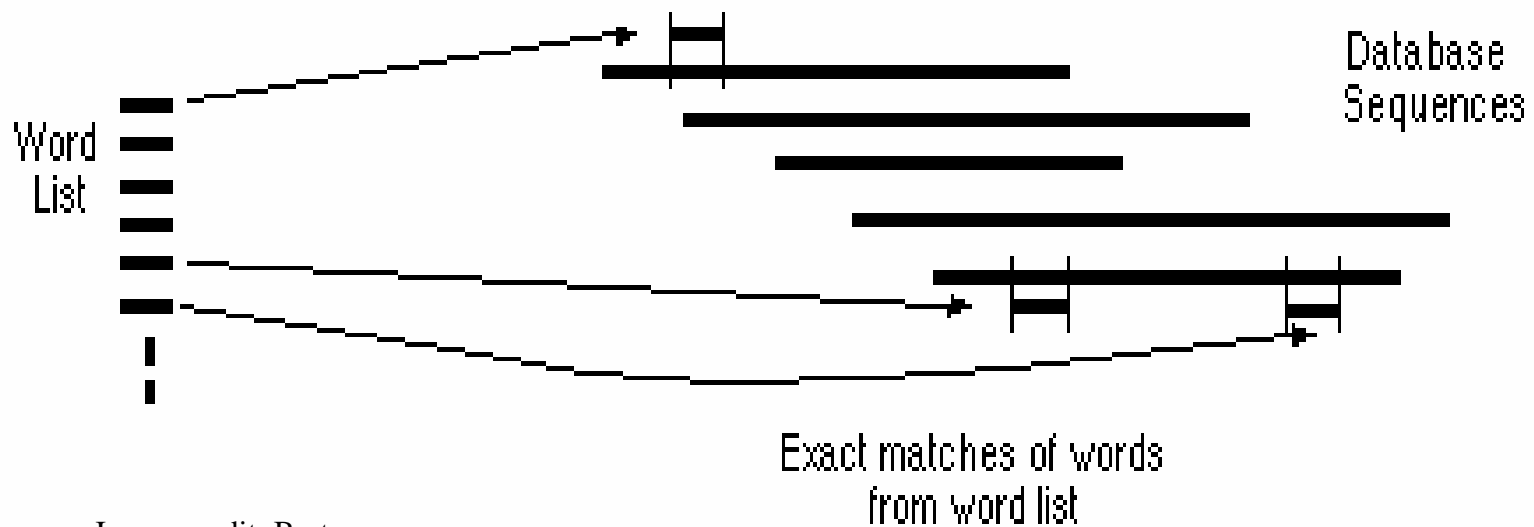


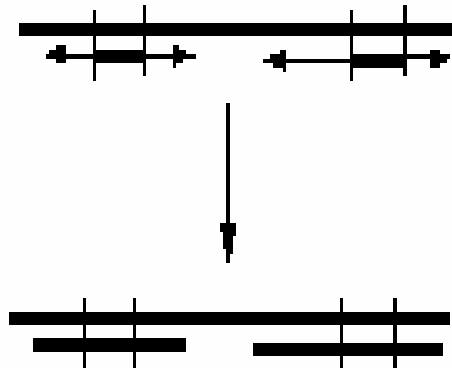
Image credit: Barton

# BLAST in 3 Steps

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

## Step 3

- For each word match, extend alignment in both directions to find alignment that score greater than a threshold  $s$



**Maximal Segment Pairs (MSPs)**

Image credit: Barton



# Spaced Seeds

- **111010010100110111** is an example of a spaced seed model with
  - 11 required matches (weight=11)
  - 7 “don’t care” positions

```

GAGTACTCAACACCAACATTAGTGGCAATGGAAAAT...
|| ||||| ||||| || ||||| |||||
GAATACTCAACAGCAACACTAATGGCAGCAGAAAAT...
      111010010100110111
  
```

- **1111111111** is the BLAST seed model for comparing DNA seqs

# Observations on Spaced Seeds

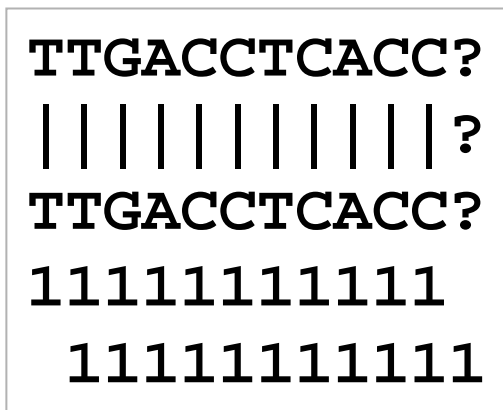
- **Seed models w/ different shapes can detect different homologies**
  - the 3rd base in a codon “wobbles” so a seed like 110110110... should be more sensitive when matching coding regions
- ⇒ **Some models detect more homologies**
  - More sensitive homology search
  - PatternHunter I
- ⇒ **Use >1 seed models to hit more homologs**
  - Approaching 100% sensitive homology search
  - PatternHunter II

# PatternHunter I

Ma et al., *Bioinformatics* 18:440-445, 2002

- BLAST's seed usually uses more than one hits to detect one homology

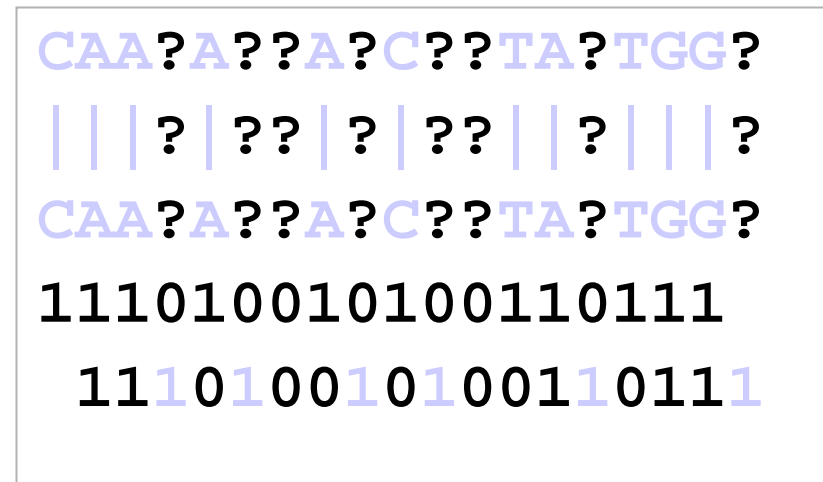
⇒ Wasteful



1/4 chances to have 2nd hit next to the 1st hit

- Spaced seeds uses fewer hits to detect one homology

⇒ Efficient



1/4<sup>6</sup> chances to have 2nd hit next to the 1st hit

# PatternHunter I

Ma et al., *Bioinformatics* 18:440-445, 2002

**Proposition.** The expected number of hits of a weight- $W$  length- $M$  model within a length- $L$  region of similarity  $p$  is  $(L - M + 1) * p^W$

**Proof.**

For any fixed position, the prob of a hit is  $p^W$ .

There are  $L - M + 1$  candidate positions.

The proposition follows.

# Implication

- For  $L = 1017$

- BLAST seed expects  $(1017 - 11 + 1) * p^{11} = 1007 * p^{11}$  hits

- But  $\sim 1/4$  of these overlap each other. So likely to have only  $\sim 750 * p^{11}$  distinct hits

- Our example spaced seed expects  $(1017 - 18 + 1) * p^{11} = 1000 * p^{11}$  hits

- But only  $1/4^6$  of these overlap each other. So likely to have  $\sim 1000 * p^{11}$  distinct hits



**PatternHunter I**  
Ma et al., *Bioinformatics* 18:440-445, 2002

- BLAST's seed usually uses more than one hits to detect one homology  
⇒ Wasteful
- Spaced seeds uses fewer hits to detect one homology  
⇒ Efficient

|              |                     |
|--------------|---------------------|
| TTGACCTCACC? | CAA?A??A?C??TA?TGG? |
|              | ? ?? ? ?? ? ? ?     |
| TTGACCTCACC? | CAA?A??A?C??TA?TGG? |
| 11111111111  | 111010010100110111  |
| 11111111111  | 111010010100110111  |

1/4 chances to have 2nd hit next to the 1st hit

1/4 chances to have 2nd hit next to the 1st hit

Copyright © 2004 by Limsoon Wong

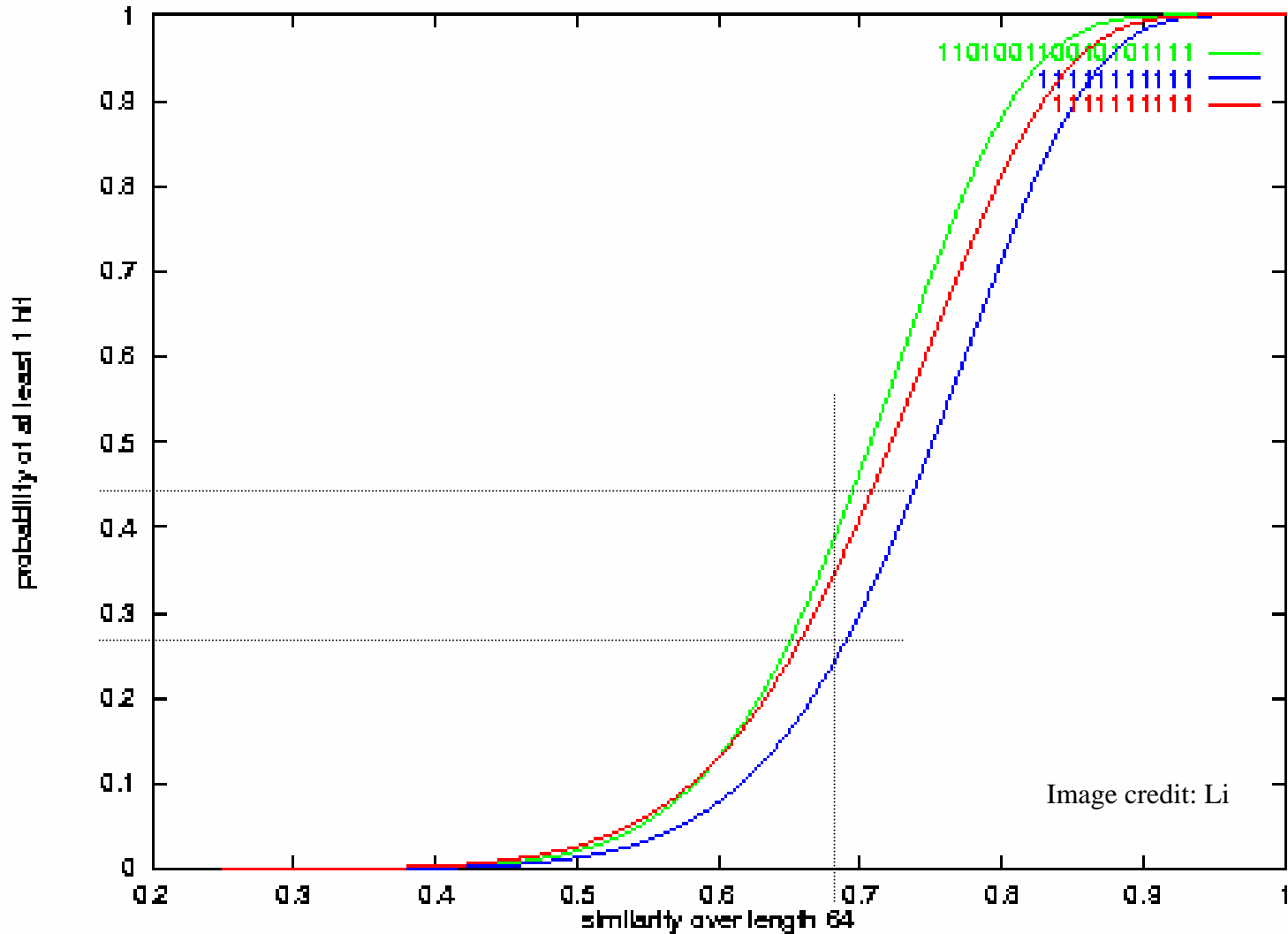
**PatternHunter I**  
Ma et al., *Bioinformatics* 18:440-445, 2002

Proposition. The expected number of hits of a weight- $W$  length- $M$  model within a length- $L$  region of similarity  $p$  is  $(L - M + 1) * p^W$

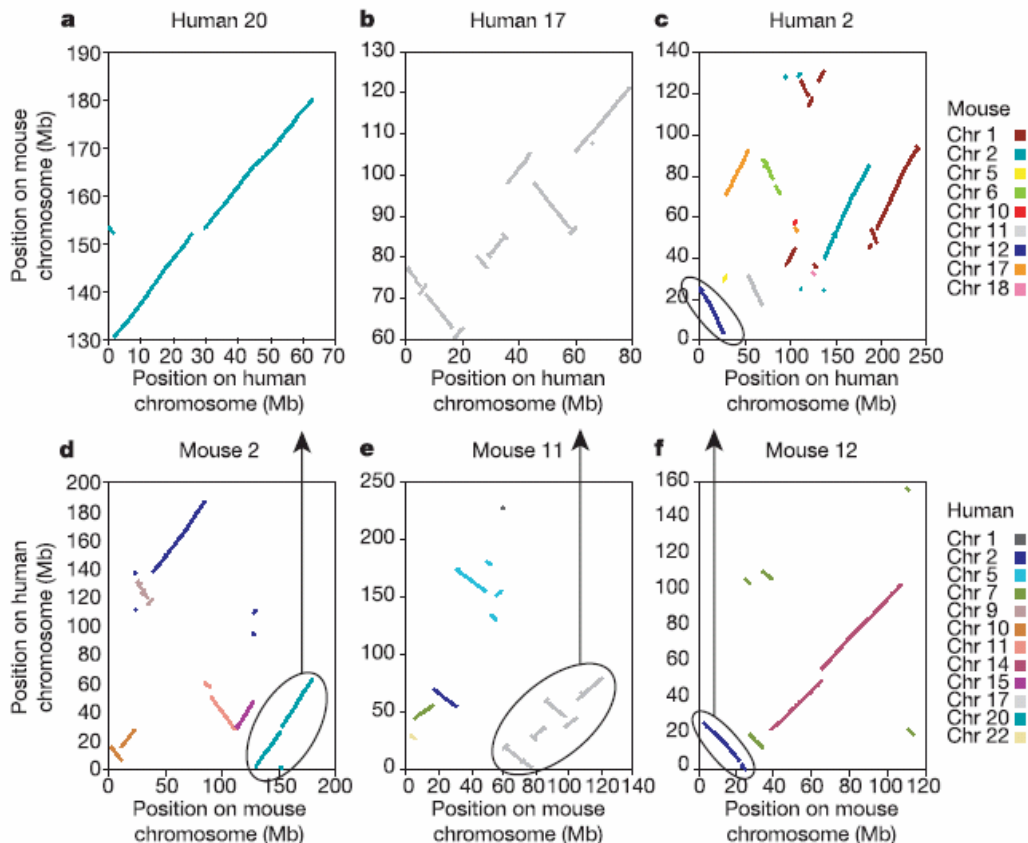
Proof. For any fixed position, the prob of a hit is  $p^W$ . There are  $L - M + 1$  positions. The proposition follows.

Copyright © 2004 by Limsoon Wong

# Sensitivity of PatternHunter I



# Speed of PatternHunter I



*Nature*, 420:520-522, 2002

- **Mouse Genome Consortium used PatternHunter to compare mouse genome & human genome**
- **PatternHunter did the job in a 20 CPU-days --- it would have taken BLAST 20 CPU-years!**

# How to Increase Sensitivity?

- **Ways to increase sensitivity:**

- “Optimal” seed
- Reduce weight by 1
- Increase number of spaced seeds by 1

Proposition. The expected number of hits of a weight- $W$  length- $M$  model within a length- $L$  region of similarity  $p$  is  $(L - M + 1) * p^W$

Proof. For any fixed position, the prob of a hit is  $p^W$ . There are  $L - M + 1$  positions. The proposition follows.

- **For  $L = 1017$  &  $p = 50\%$**

- 1 weight-11 length-18 model expects  $1000/2^{11}$  hits

- 2 weight-12 length-18 models expect  $2 * 1000/2^{12} = 1000/2^{11}$  hits

⇒ When comparing regions w/  $>50\%$  similarity, using 2 weight-12 spaced seeds together is more sensitive than using 1 weight-11 spaced seed!



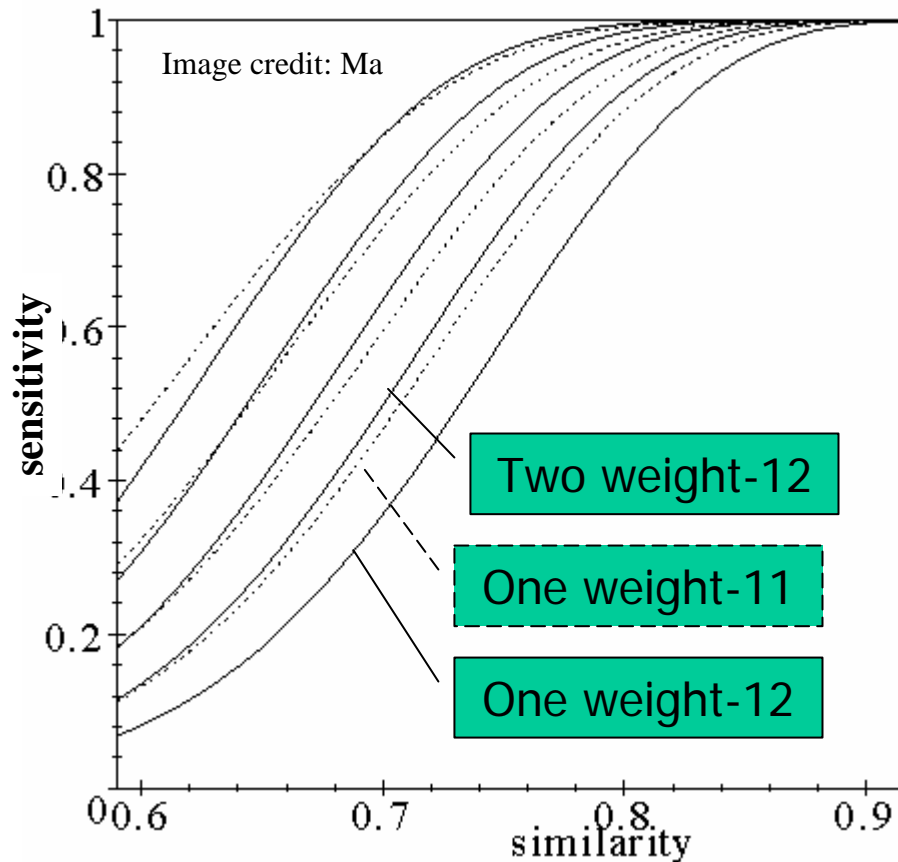


# PatternHunter II

Li et al, *GIW*, 164-175, 2003

- **Idea**
  - Select a group of spaced seed models
  - For each hit of each model, conduct extension to find a homology
- **Selecting optimal multiple seeds is NP-hard**
- **Algorithm to select multiple spaced seeds**
  - Let  $A$  be an empty set
  - Let  $s$  be the seed such that  $A \cup \{s\}$  has the highest hit probability
  - $A = A \cup \{s\}$
  - Repeat until  $|A| = K$
- **Computing hit probability of multiple seeds is NP-hard**

# Sensitivity of PatternHunter II

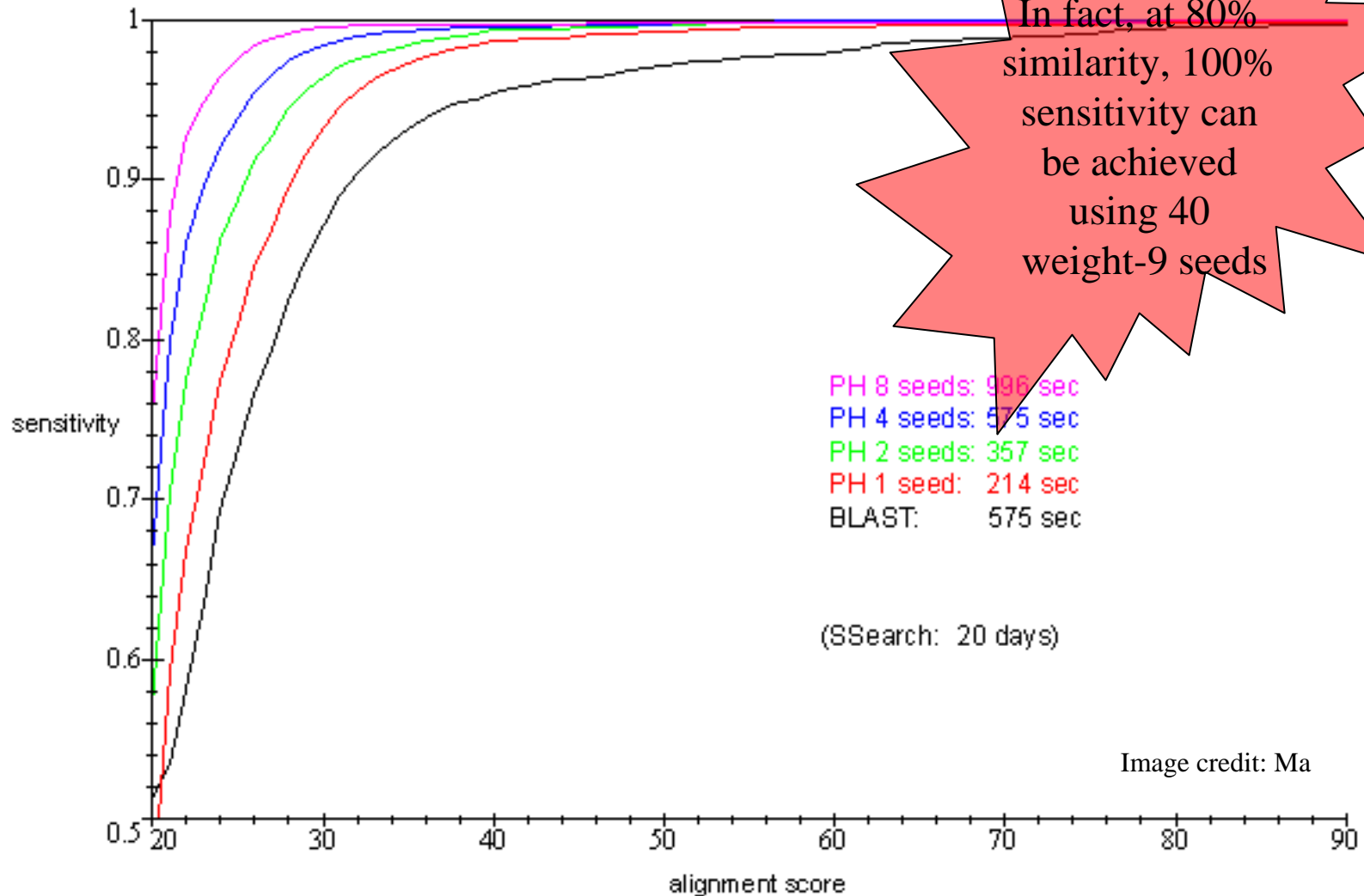


- **Solid curves: Multiple (1, 2, 4, 8, 16) weight-12 spaced seeds**
  - **Dashed curves: Optimal spaced seeds with weight = 11, 10, 9, 8**
- ⇒ “Doubling the seed number” gains better sensitivity than “decreasing the weight by 1”

## Expts on Real Data

- **30k mouse ESTs (25Mb) vs 4k human ESTs (3Mb)**
  - downloaded from NCBI genbank
  - “low complexity” regions filtered out
- **SSearch (Smith-Waterman method) finds “all” pairs of ESTs with significant local alignments**
- **Check how many percent of these pairs can be “found” by BLAST and different configurations of PatternHunter II**

# Results



# Farewell to the Supercomputer Age of Sequence Comparison!

Computer: PIII 700Mhz Redhat 7.1, 1G main memory

| Sequence Length | Blastn        | PatternHunter |
|-----------------|---------------|---------------|
| 816k vs 580k    | 47 sec        | 9 sec         |
| 4639k vs 1830k  | 716 sec       | 44 sec        |
| 20M vs 18M      | out of memory | 13 min        |

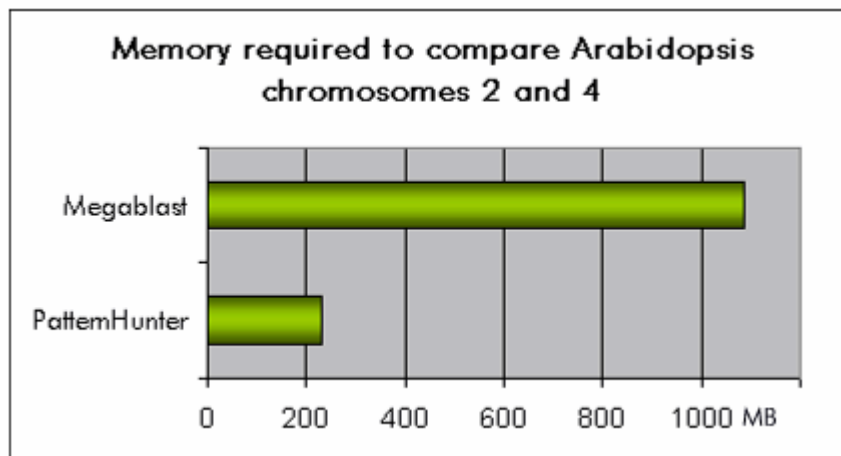
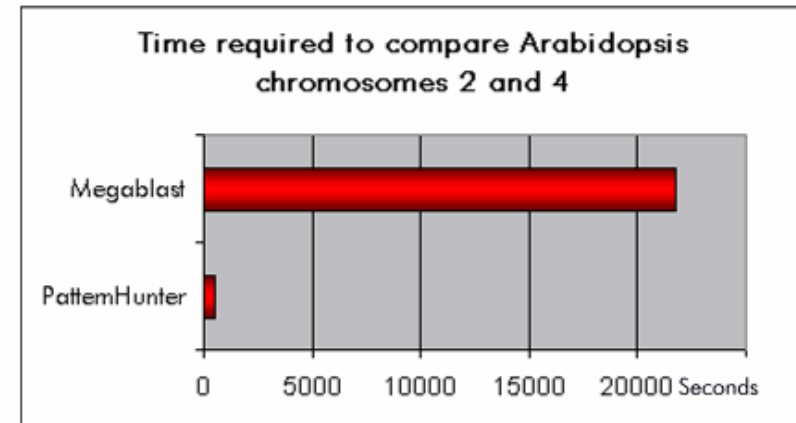
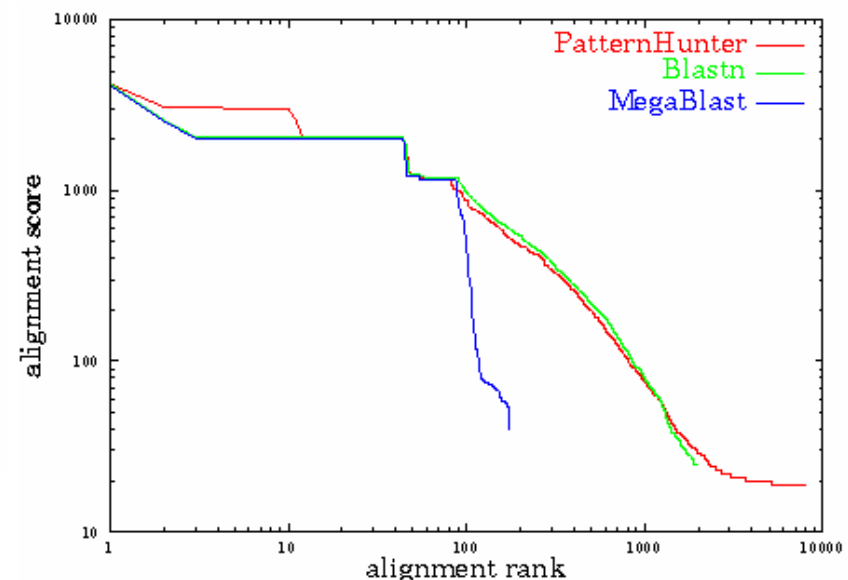


Image credit: Bioinformatics Solutions Inc

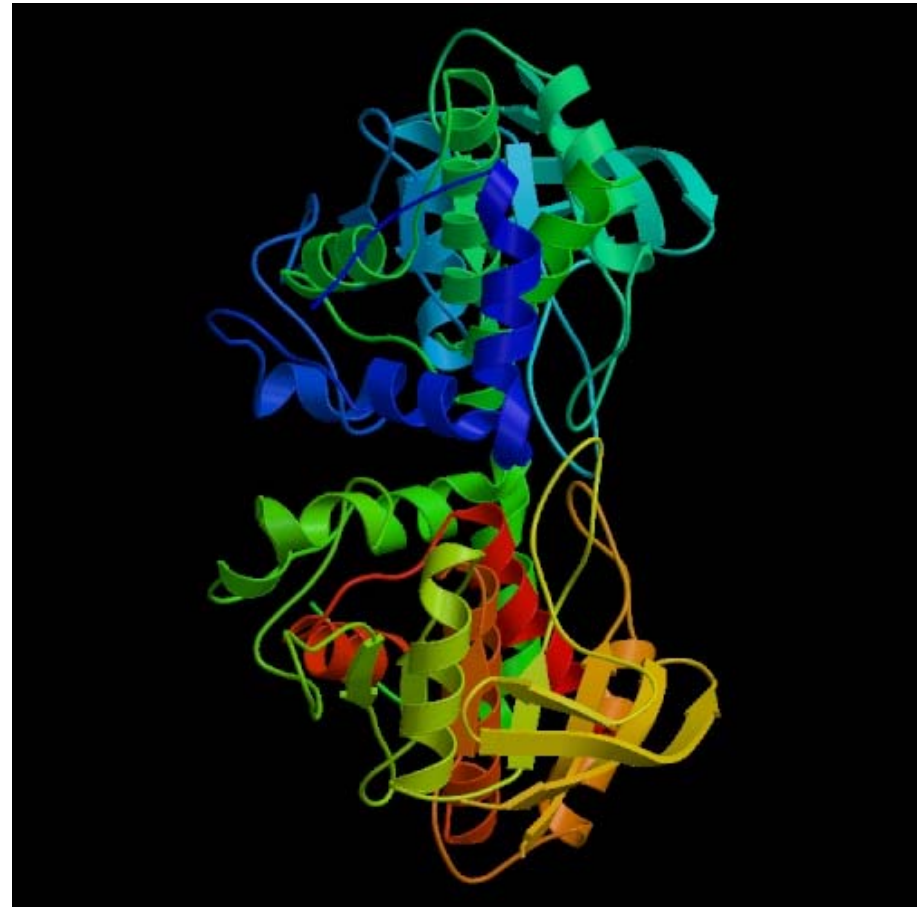


# Application of Sequence Comparison: Guilt-by-Association



# A protein is a ...

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



# Function Assignment to Protein Sequence

SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR  
YVNILPYDHSRVHLTPVEGVPDSYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE  
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD  
VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG  
TFVVIDAMLDMMSERKVDVYGFVSRIRAQRCQMVQTDMQYVFYQALLEHYLYGDTELE  
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

```

      60      70      80      90     100
Amicyanin  MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVV
           . . . . .
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNEFTVDNPGTFFYHGHLMQRSAGLYG
           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|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  MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTVEVSAKVGDTIRVWVKDVFVAHT 60
           MK G L ++      MA PA AATIE+T++ LV SP V AKVGDIT WVN DV AHT
Sbjct: 1  MKAGALIRLSVLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNDVVAHT 60
    
```

good match between Amicyanin and unknown *M. loti* protein

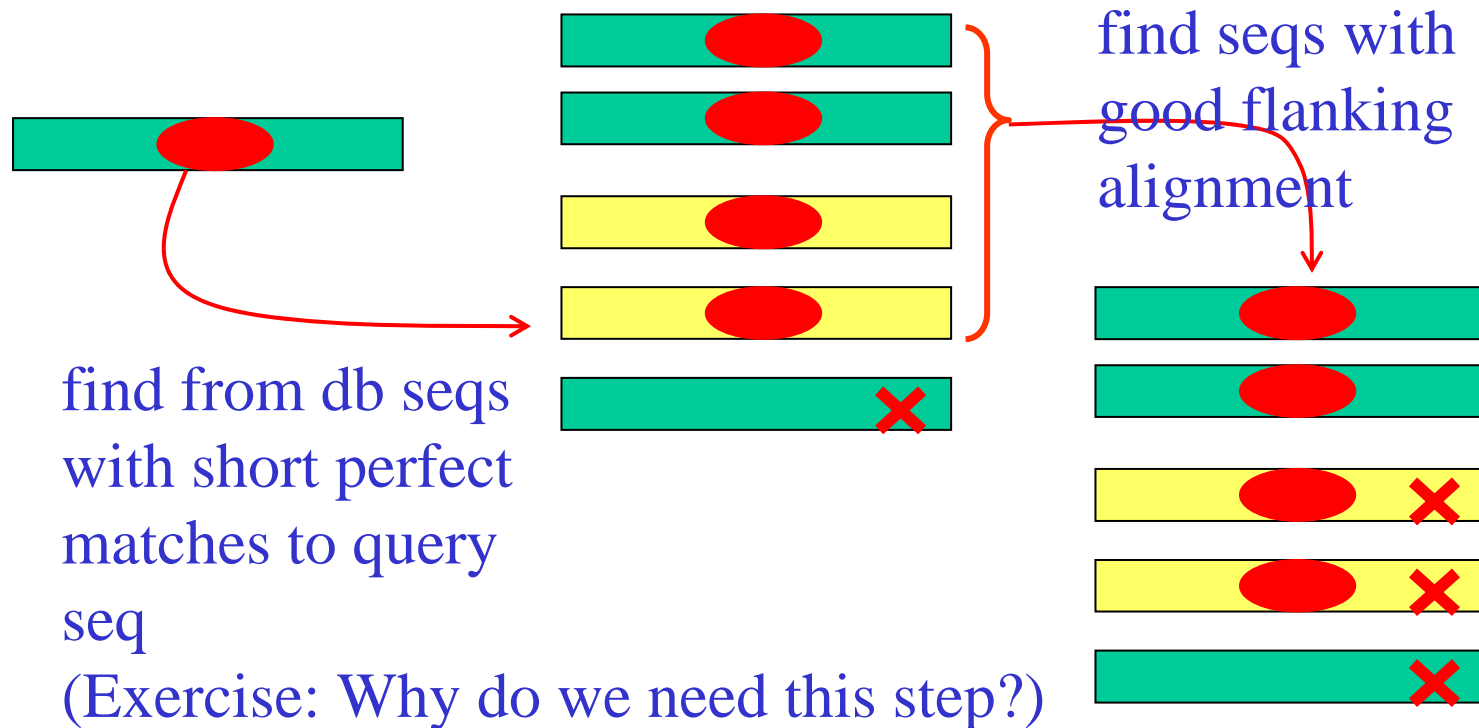
Assign to  $T$  same function as homologs

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**



# Homologs Obtained by BLAST

| Sequences producing significant alignments:                                    | Score<br>(bits) | E<br>Value |
|--|-----------------|------------|
| <a href="#">gi 14193729 gb AAK56109.1 AF332081_1</a> protein tyrosin phosph... | 62 L            | e-177      |
| <a href="#">gi 126467 sp P18433 PTRA_HUMAN</a> Protein-tyrosine phosphatase... | 62 L            | e-177      |
| <a href="#">gi 4506303 ref NP_002827.1</a> protein tyrosine phosphatase, r...  | 62 L            | e-176      |
| <a href="#">gi 227294 prf  1701300A</a> protein Tyr phosphatase                | 620             | e-176      |
| <a href="#">gi 18450369 ref NP_543030.1</a> protein tyrosine phosphatase, ...  | 62 L            | e-176      |
| <a href="#">gi 32067 emb CAA37447.1</a> tyrosine phosphatase precursor [Ho...  | 61 L            | e-176      |
| <a href="#">gi 285113 pir  JC1285</a> protein-tyrosine-phosphatase (EC 3.1.... | 619             | e-176      |
| <a href="#">gi 6981446 ref NP_036895.1</a> protein tyrosine phosphatase, r...  | 61 L            | e-176      |
| <a href="#">gi 2098414 pdb 1YFO A</a> Chain A, Receptor Protein Tyrosine Ph... | 61 S            | e-174      |
| <a href="#">gi 32313 emb CAA38662.1</a> protein-tyrosine phosphatase [Homo...  | 61 L            | e-174      |
| <a href="#">gi 450583 gb AAB04150.1</a> protein tyrosine phosphatase >gi 4...  | 605             | e-172      |
| <a href="#">gi 6679557 ref NP_033006.1</a> protein tyrosine phosphatase, r...  | 60 L            | e-172      |
| <a href="#">gi 483922 gb AAA17990.1</a> protein tyrosine phosphatase alpha     | 599             | 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 TFVVIDAMLDMMHSEKVDVYGFVSRIRAQRCQMVQTD MQYVF IYQALLEHYLYGDTELE 300
          TFVVIDAMLDMMHSEKVDVYGFVSRIRAQRCQMVQTD MQYVF IYQALLEHYLYGDTELE
Sbjct: 442 TFVVIDAMLDMMHSEKVDVYGFVSRIRAQRCQMVQTD 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!



# 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**



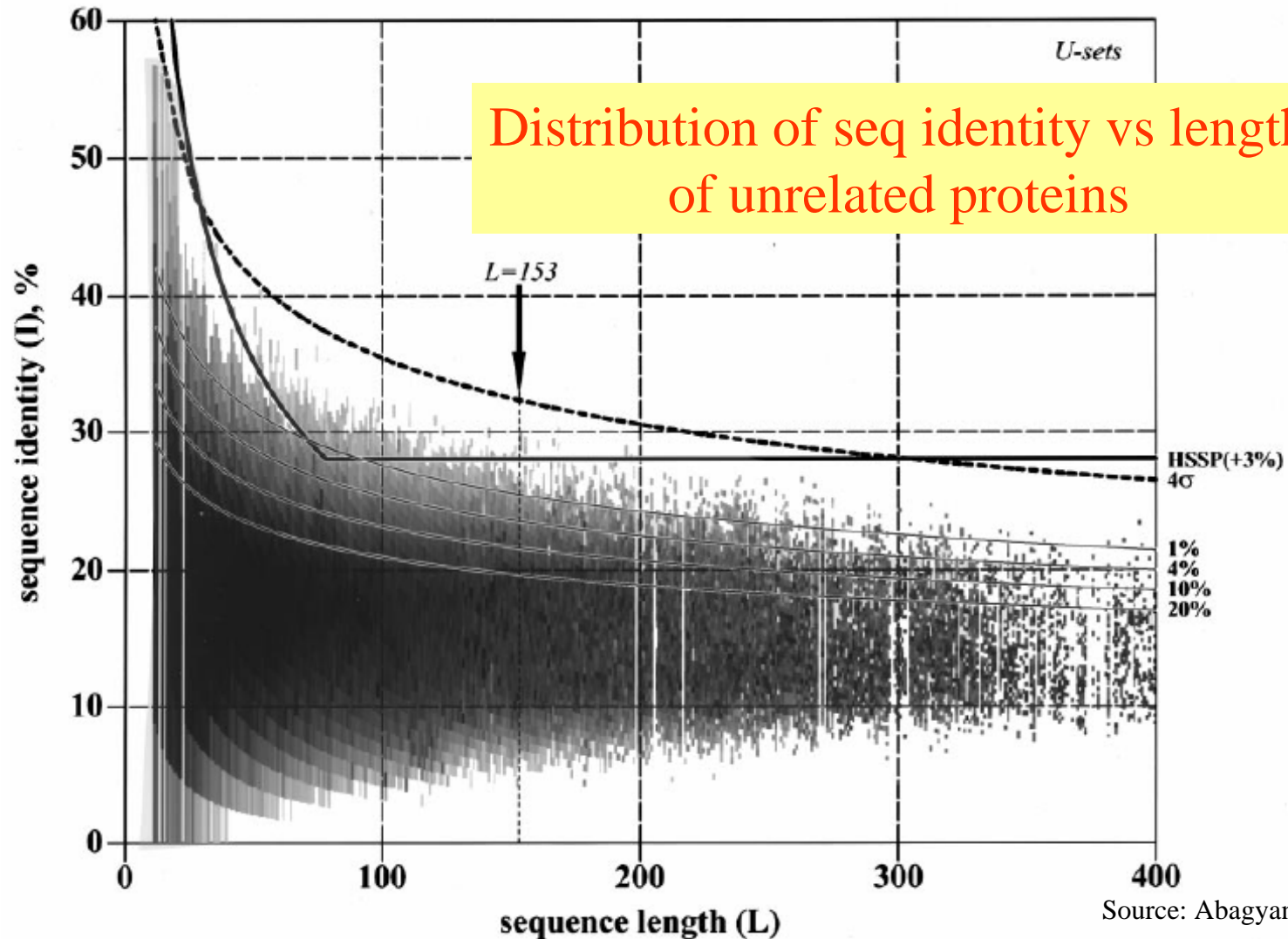
Cartoon: Ron Hipschman  
Data: David Hand

# 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



Source: Abagyan & Batalov

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

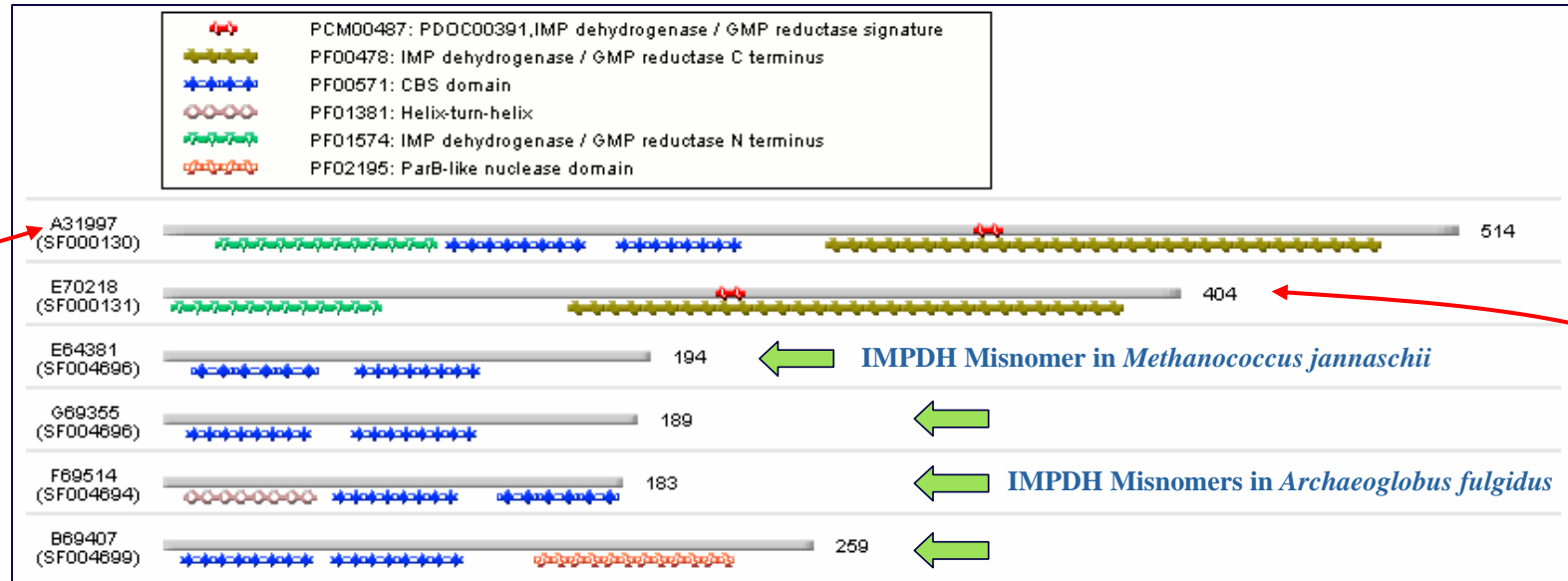


18 entries were found

| 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)<br><a href="#">NP_247637</a> inosine-5'-monophosphate dehydrogenase (guaB)  |
| <a href="#">NF00187788</a> | Archaeoglobus fulgidus                    | <a href="#">G69355</a> MJ0653 homolog AF0847<br><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)<br><a href="#">NP_069681</a> inosine monophosphate dehydrogenase (guaB-1)  |
| <a href="#">NF00188267</a> | Archaeoglobus fulgidus                    | <a href="#">F69514</a> yhcV homolog 2<br><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)<br><a href="#">NP_070943</a> inosine monophosphate dehydrogenase (guaB-2)  |
| <a href="#">NF00188697</a> | Archaeoglobus fulgidus                    | <a href="#">B69407</a> MJ0188 homolog<br><i>ALT_NAMES</i> : inosine monophosphate  | <a href="#">O29009</a> Hypothetical protein AF1259                                | <a href="#">g2649320</a> inosine monophosphate dehydrogenase, putative<br><a href="#">NP_070997</a> inosine monophosphate  |
| <a href="#">NF00197776</a> | Thermo                                    |  |   | ive<br>monophosphate<br>d protein<br>monophosphate<br>d protein  |
| <a href="#">NF00414709</a> | Methan<br>thermau                         |  |   | monophosphate<br>d protein V<br>monophosphate  |
| <a href="#">NF00414811</a> | Methanothermobacter<br>thermautotrophicus | <a href="#">D69035</a> MJ1232 protein homolog MTH126<br><i>ALT_NAMES</i> : inosine-5'-monophosphate dehydrogenase related protein VII [misnomer] | <a href="#">O26229</a> INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN VII | dehydrogenase related protein V<br><a href="#">g2621166</a> inosine-5'-monophosphate dehydrogenase related protein VII<br><a href="#">NP_275269</a> inosine-5'-monophosphate dehydrogenase related protein VII |
| <a href="#">NF00414837</a> | Methanothermobacter<br>thermautotrophicus | <a href="#">H69232</a> MJ1225-related protein MTH992<br><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<br><a href="#">NP_276127</a> inosine-5'-monophosphate dehydrogenase related protein IX                                      |
| <a href="#">NF00414969</a> | Methanothermobacter<br>thermautotrophicus | <a href="#">B69077</a> yhcV homolog 2<br><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<br><a href="#">NP_276687</a> inosine-5'-monophosphate dehydrogenase related protein X  |

**A partial list of IMP dehydrogenase misnomers  
in complete genomes remaining in some  
public databases**










# IMPDH Domain Structure

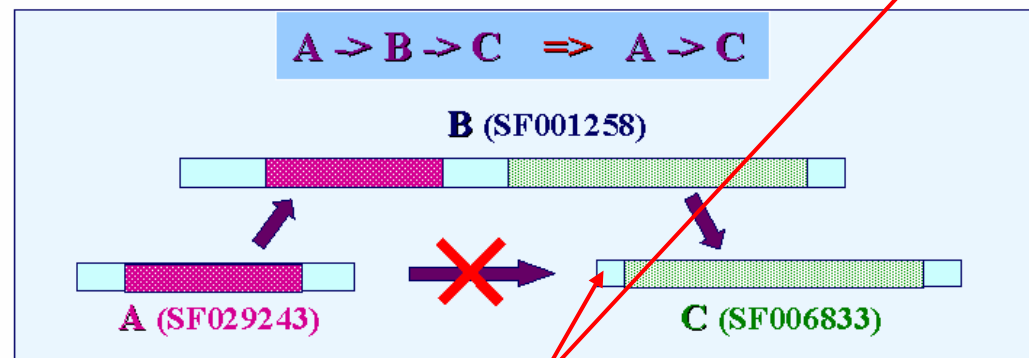


- Typical IMPDHs have 2 IMPDH domains that form the catalytic core and 2 CBS domains.
- A less common but functional IMPDH (E70218) lacks the CBS domains.
- Misnomers show similarity to the CBS domains

# Invalid Transitive Assignment

## Root of invalid transitive assignment

|            |                        |                          |                        |   |                                 |            |       |         |     |        |     |   |
|------------|------------------------|--------------------------|------------------------|---|---------------------------------|------------|-------|---------|-----|--------|-----|---|
| <b>B</b> → | <a href="#">H70468</a> | <a href="#">SF001258</a> | <a href="#">051440</a> | <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 |    |
|            | <a href="#">S76963</a> | <a href="#">SF001258</a> | <a href="#">039935</a> | <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 |    |
|            | <a href="#">T35073</a> | <a href="#">SF029243</a> | <a href="#">005738</a> | <a href="#">probable phosphoribosyl-AMP cyclohydrolase</a>  | <i>Streptomyces coelicolor</i>  | Prok/gram+ | 399.3 | 3.5e-15 | 128 | 42.157 | 102 |    |
|            | <a href="#">S53349</a> | <a href="#">SF001257</a> | <a href="#">001188</a> | <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> → | <a href="#">E69493</a> | <a href="#">SF029243</a> | <a href="#">005738</a> | <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> → | <a href="#">G64337</a> | <a href="#">SF006833</a> | <a href="#">030827</a> | <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  |    |
|            | <a href="#">D81178</a> | <a href="#">SF006833</a> | <a href="#">101491</a> | <a href="#">phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) NMB0603 [similarity]</a>   | <i>Neisseria meningitidis</i>   | Prok/gram- | 239.9 | 2.6e-06 | 107 | 35.227 | 88  |    |
|            | <a href="#">G81925</a> | <a href="#">SF006833</a> | <a href="#">101491</a> | <a href="#">phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) NMA0807 [similarity]</a>   |                                 |            |       |         |     |        |     |    |
|            | <a href="#">S51513</a> | <a href="#">SF001257</a> | <a href="#">001188</a> | <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

# 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**)

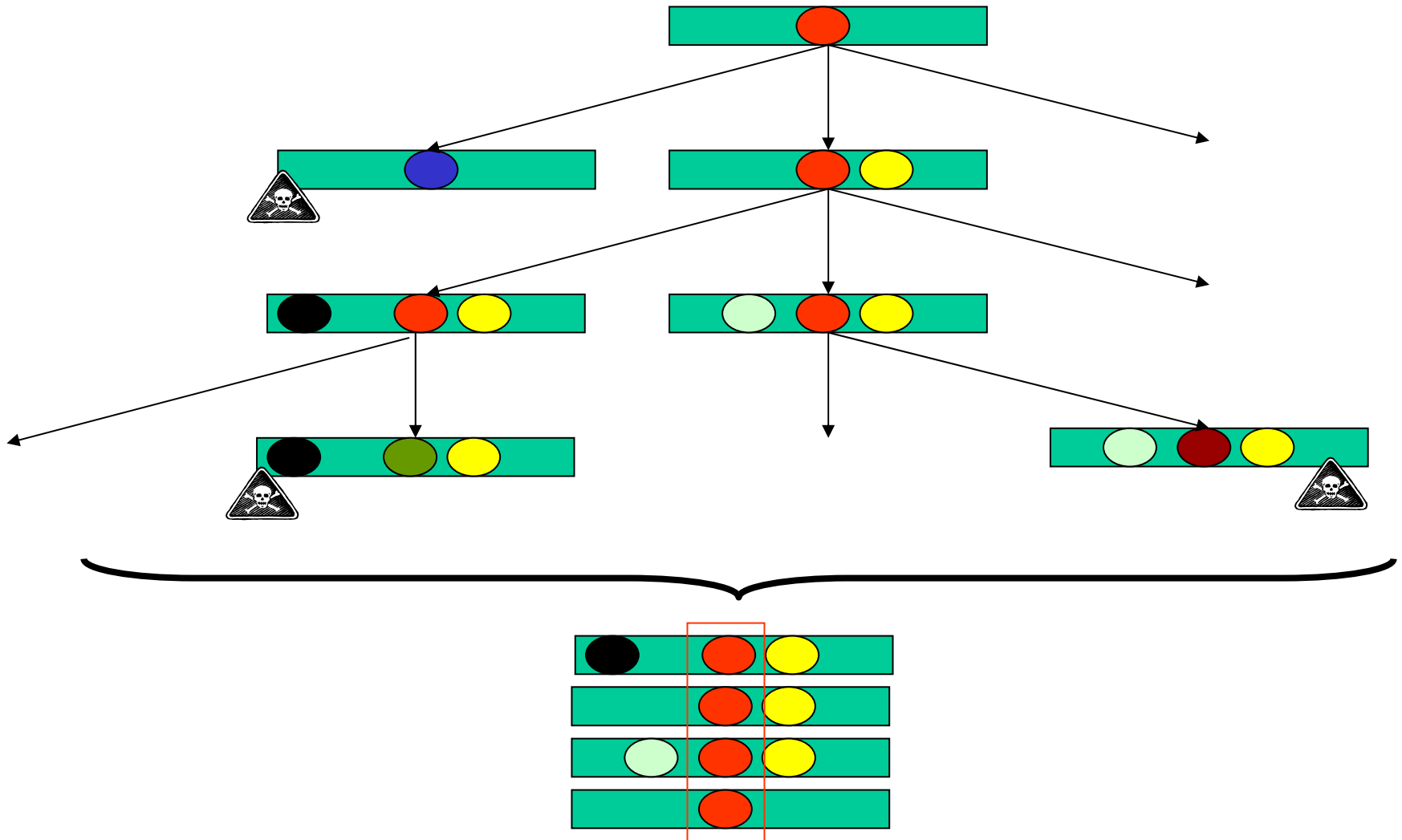


# Discovering Domain and Active Sites

```
>gi|475902|emb|CAA83657.1| protein-tyrosine-phosphatase alpha
MDLWFFVLLLGSGLISVGATNVTTEPPTTVPTSTRIPTKAPTAAPDGGTTPRVSSLNVSSPMTTSAPASE
PPTTTATSISPNATTASLNASTPGTSVPTSAPVAISLPPSATPSALLTALPSTEAEMTERNVSATVTTQE
TSSASHNGNSDRRDETPIIAVMVALSLLVIVFIIIVLYMLRFKKYKQAGSHSNSFRLPNGRTDDAEPQS
MPLLARSPSTNRKYPPPLPVDKLEEEINRRIGDDNKLFFREEFNALPACPIQATCEAASKEENKEKNRYVNI
LPYDHSRVHLTPVEGV PDSHYINTSFINSYQEKNKFI AAQGPKEETVND FWRMIWEQNTATIVMVTNLKE
RKECKCAQYWPDQGCWTYGNIRVSVEDVTVLVDYTVRKFCIQQVGDVTNKKPQRLVTQFHFTSWPDFGVP
FTP I GMLKFLKKVKTCNPQYAGAI VVHCSAGVGRTGTFIVIDAMLDMHAERKVDVYGFVSRIRAQRCQM
VQTD MQYVFIYQALLEHYLYGDTELEVTSLEIHLQKIYNKVPGTSSNGLEEEFKKLTSIKI QNDKMRTGN
LPANMKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASFIDGYRRRTPTCQPRPVQHTIEDFWRMIWEWK
SCSIVMLTELEERGQEKCAQYWPSDGSVSYGDINVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFH
GWPEVGIPSDGKGMINI IAAVQKQQQQSGNHPMHCHCSAGAGRTGTF CALSTVLERVKAEGILDVVFQTVK
SLRLQRPHMVQTLEQYEF CYKVVQEYIDAFSDYANFK
```

- **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

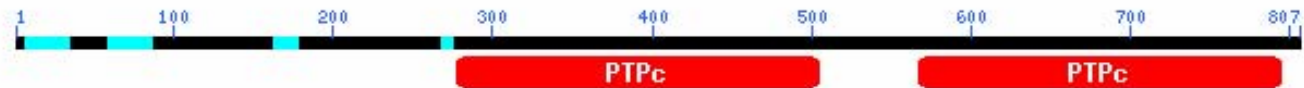
# Lucky Case: Try BLAST



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 ...

# 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**

# Multiple Alignment of PTPs

```

gi|126467|      FHFTSWPDFGVPFTP I GMLKFLKKVKACNP--QYAGAIVVHCSAGVGRTGTFVVIDAMLD
gi|2499753     FHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|462550|     YHYTQWPDMGVPEYALPVLTFVRRSSAARM--PETGPVLVHCSAGVGRTGTYIVIDSMLQ
gi|2499751     FHFTSWPDHGVPD TTDLLINFRYLVRDYMKQSPPE SPILVHCSAGVGRTGTFIAIDRLIY
gi|1709906     FQFTA WPDHGVP EHP T PFLAFLRRVKTCNP--PDAGPMVVHCSAGVGRTGCFIVIDAMLE
gi|126471|     LHFTSWPDFGVPFTP I GMLKFLKKVKT LNP--VHAGPIVVHCSAGVGRTGTFIVIDAMMA
gi|548626|     FHFTGWPDHGVPYHATGLLSF IRRVKLSNP--PSAGPIVVHCSAGAGRTGCYIVIDIMLD
gi|131570|     FHFTGWPDHGVPYHATGLLGFVRQVKS 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

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

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



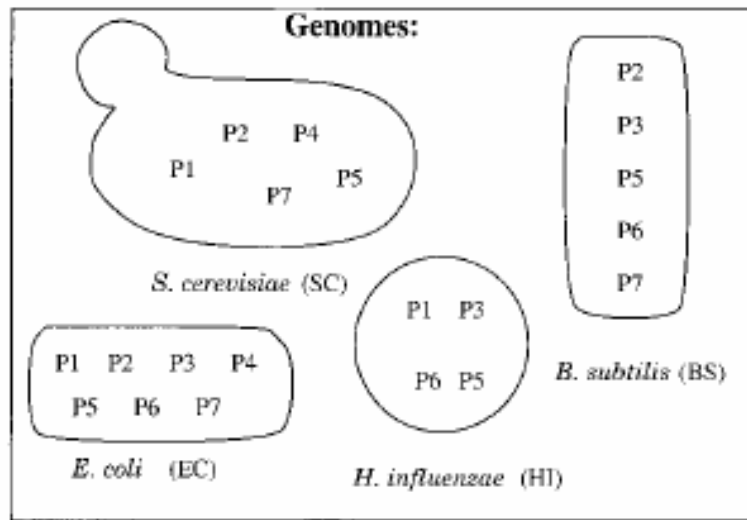
# Phylogenetic Profiling

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**

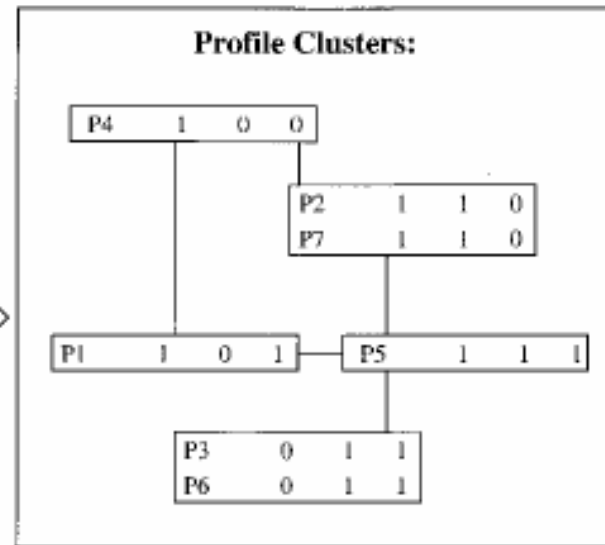


# Phylogenetic Profiling: How it Works



**Phylogenetic Profile:**

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



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

# Phylogenetic Profiling: P-value

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

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

where

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

**No. of ways to distribute  $z$  co-occurrences over  $N$  lineage's** (points to  $w_z$ )  
**No. of ways to distribute the remaining  $x - z$  and  $y - z$  occurrences over the remaining  $N - z$  lineage's** (points to  $\overline{w}_z$ )  
**No. of ways of distributing  $X$  and  $Y$  over  $N$  lineage's without restriction** (points to  $W$ )

# Phylogenetic Profiles: Evidence

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

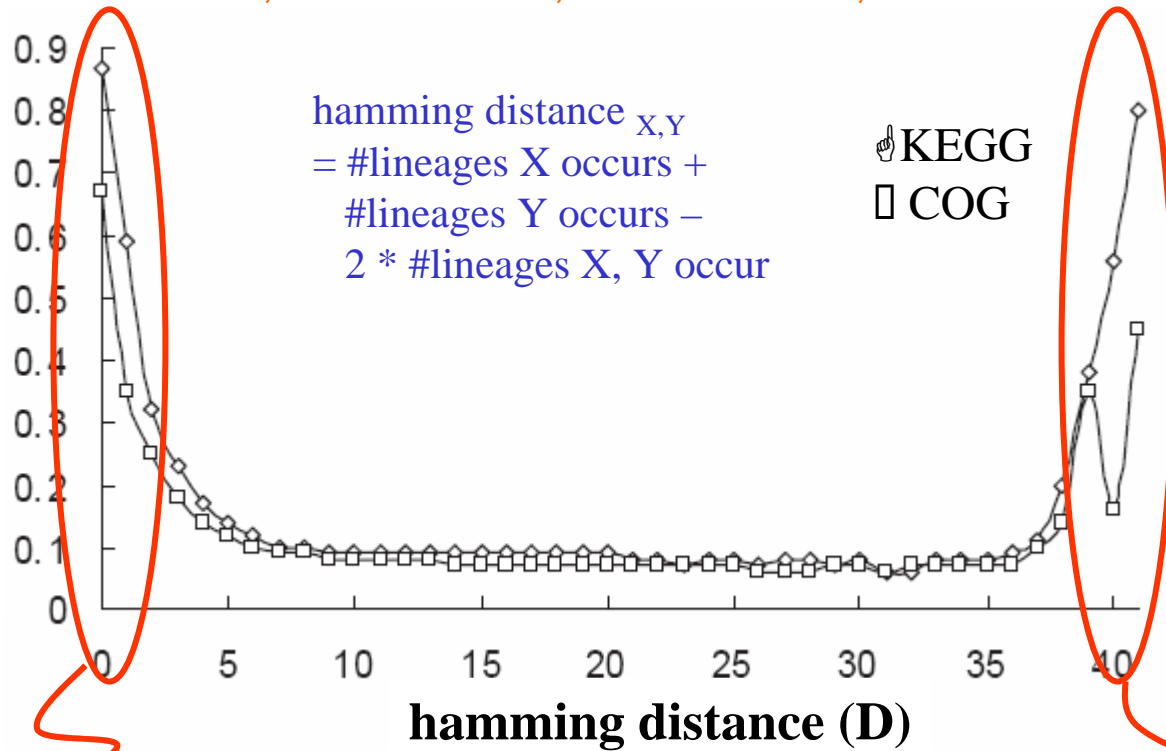
| Keyword                                       | No. of proteins in this group | No. of protein pairs in this group that differ by < 3 “bit” | No. of protein pairs in random group that differ by < 3 “bit” |
|---|-------------------------------|---|---|
| Ribosome                                      | 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   |

- **Proteins grouped based on similar keywords in SWISS-PROT have more 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

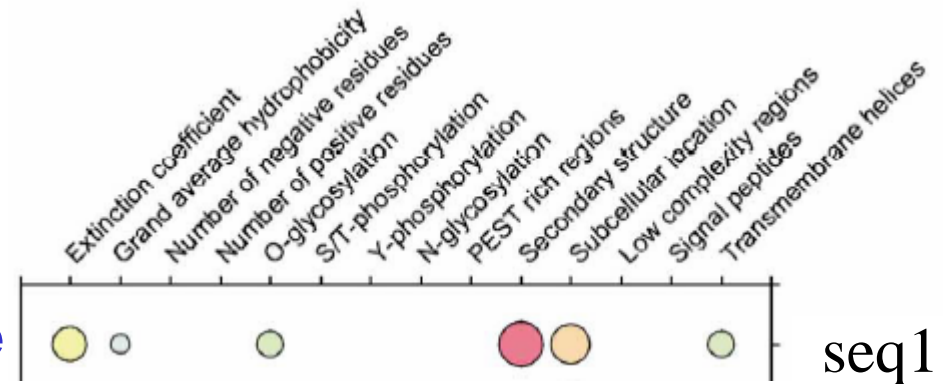


- Proteins having low hamming distance (thus highly similar phylogenetic profiles) tend to share common pathways
- Exercise: Why do proteins having high hamming distance also have this behaviour?

# 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: How it Works

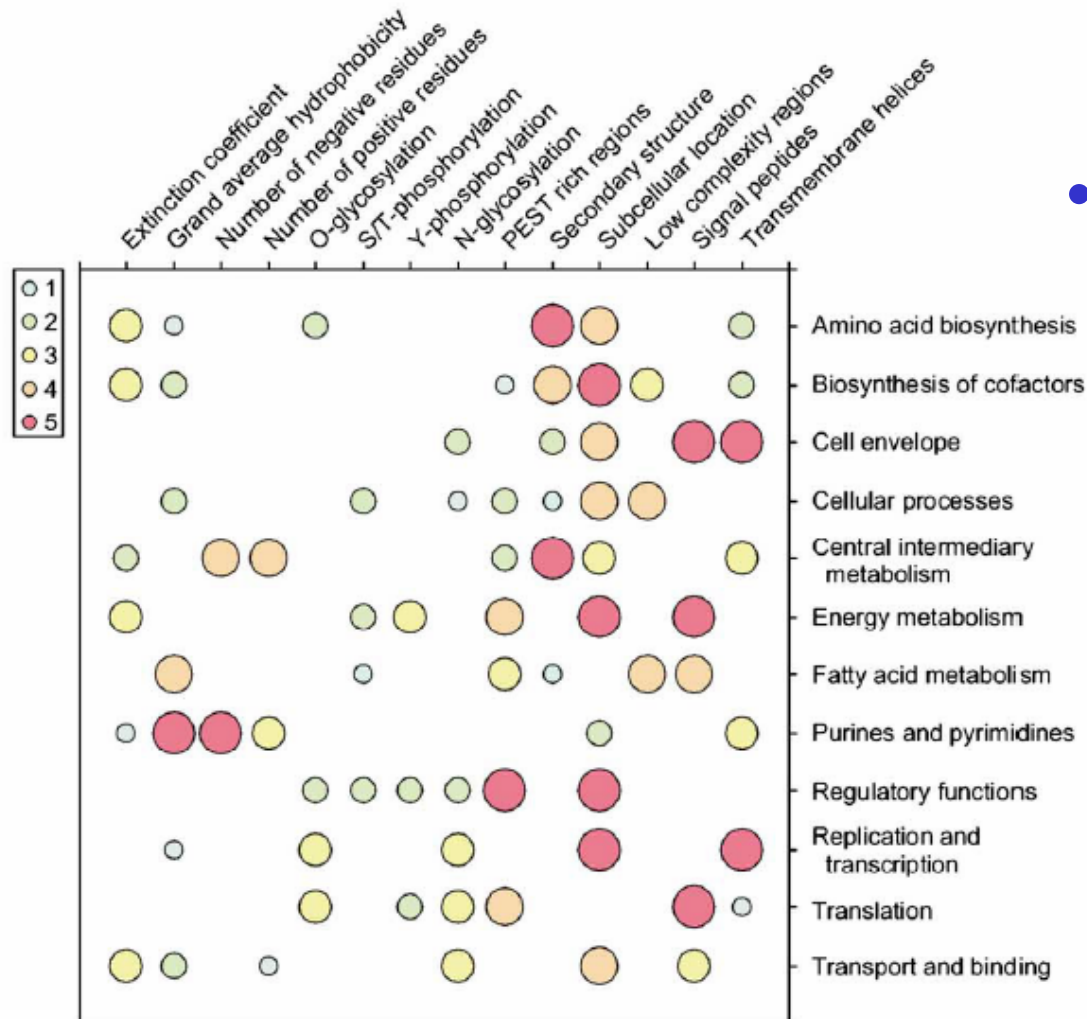
| Abbreviation | Encoding                            | Description   |
|--------------|-------------------------------------|---|
| ec           | single value                        | Extinction coefficient predicted by <a href="#">ExPASy ProtParam</a>              |
| gravy        | single value                        | 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: Evidence



- Some combinations of “features” seem to characterize some functional categories

# ProtFun: Example Output

|                                 | Prion | A4    | TTHY  |
|---------------------------------|-------|-------|-------|
| Amino acid biosynthesis         | 0.011 | 0.011 | 0.011 |
| Biosynthesis of cofactors       | 0.041 | 0.161 | 0.034 |
| Cell envelope                   | 0.146 | 0.804 | 0.698 |
| Cellular processes              | 0.027 | 0.027 | 0.051 |
| Central intermediary metabolism | 0.047 | 0.139 | 0.059 |
| Energy metabolism               | 0.029 | 0.023 | 0.046 |
| Fatty acid metabolism           | 0.017 | 0.017 | 0.023 |
| Purines and pyrimidines         | 0.528 | 0.417 | 0.153 |
| Regulatory functions            | 0.013 | 0.014 | 0.014 |
| Replication and transcription   | 0.020 | 0.029 | 0.040 |
| Translation                     | 0.035 | 0.027 | 0.032 |
| Transport and binding           | 0.831 | 0.827 | 0.812 |
| Enzyme                          | 0.233 | 0.367 | 0.227 |
| Non-enzyme                      | 0.767 | 0.633 | 0.773 |
| Oxidoreductase (EC 1.-.-.-)     | 0.070 | 0.024 | 0.055 |
| Transferase (EC 2.-.-.-)        | 0.031 | 0.208 | 0.037 |
| Hydrolase (EC 3.-.-.-)          | 0.101 | 0.090 | 0.208 |
| Isomerase (EC 4.-.-.-)          | 0.020 | 0.020 | 0.020 |
| Ligase (EC 5.-.-.-)             | 0.010 | 0.010 | 0.010 |
| Lyase (EC 6.-.-.-)              | 0.017 | 0.078 | 0.017 |

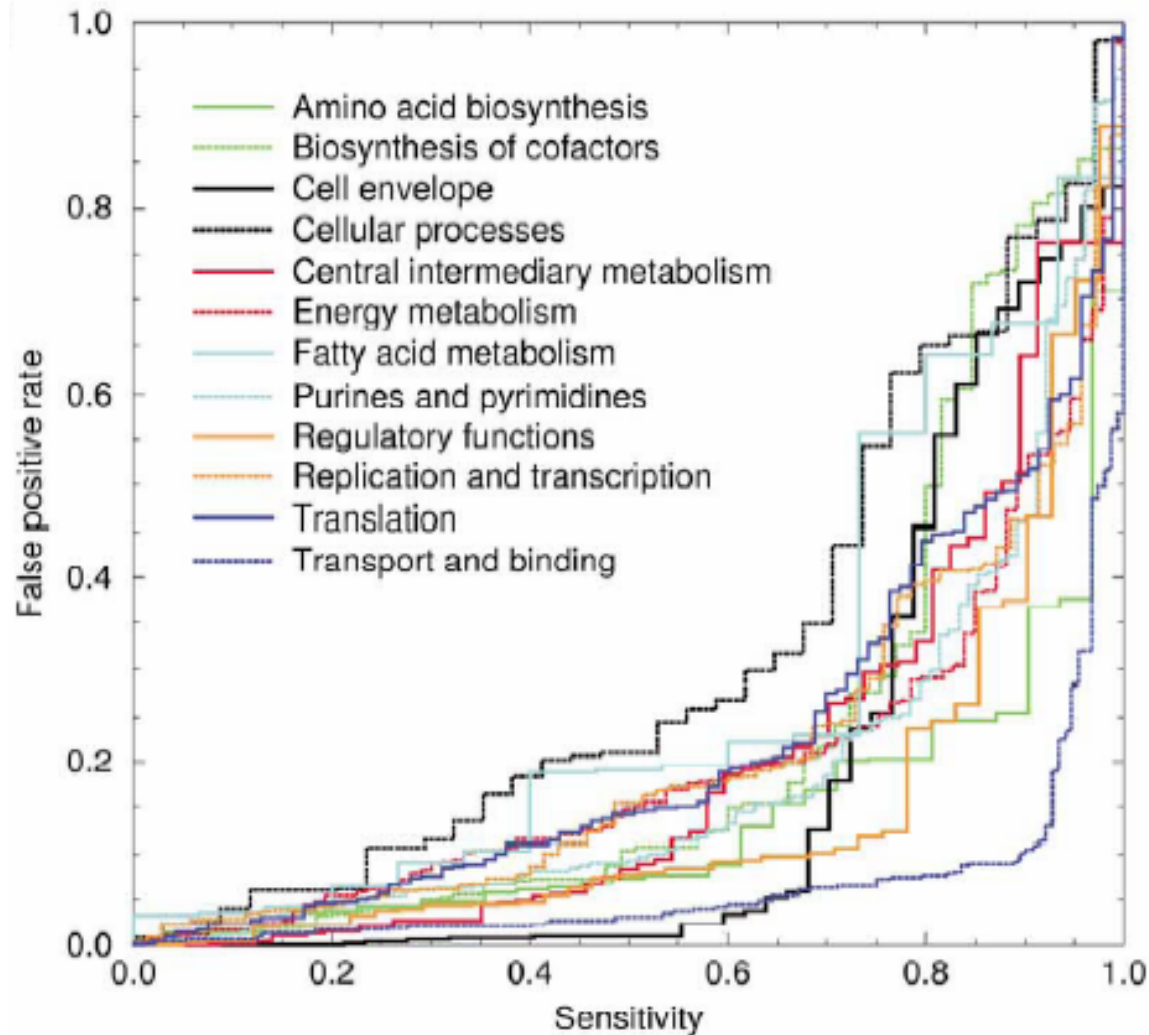
• At the seq level,  
**Prion, A4, & TTHY are  
 dissimilar**

• **ProtFun predicts  
 them to be cell  
 envelope-related,  
 transport & binding**

• **This is in agreement  
 with known  
 functionality of these  
 proteins**



# ProtFun: Performance



# Similarity of Dissimilarities

|                           | <b>orange<sub>1</sub></b>  | <b>banana<sub>1</sub></b>  | ... |
|---------------------------|--|--|-----|
| <b>apple<sub>1</sub></b>  | Color = red vs orange<br>Skin = smooth vs rough<br>Size = small vs small<br>Shape = round vs round   | Color = red vs yellow<br>Skin = smooth vs smooth<br>Size = small vs small<br>Shape = round vs oblong   | ... |
| <b>apple<sub>2</sub></b>  | Color = red vs orange<br>Skin = smooth vs rough<br>Size = small vs small<br>Shape = round vs round   | Color = red vs yellow<br>Skin = smooth vs smooth<br>Size = small vs small<br>Shape = round vs oblong   | ... |
| <b>orange<sub>2</sub></b> | Color = orange vs orange<br>Skin = rough vs rough<br>Size = small vs small<br>Shape = round vs round | Color = orange vs yellow<br>Skin = rough vs smooth<br>Size = small vs small<br>Shape = round vs oblong | ..  |
| ...                       | ...  | ...  | ... |

# SVM-Pairwise Framework

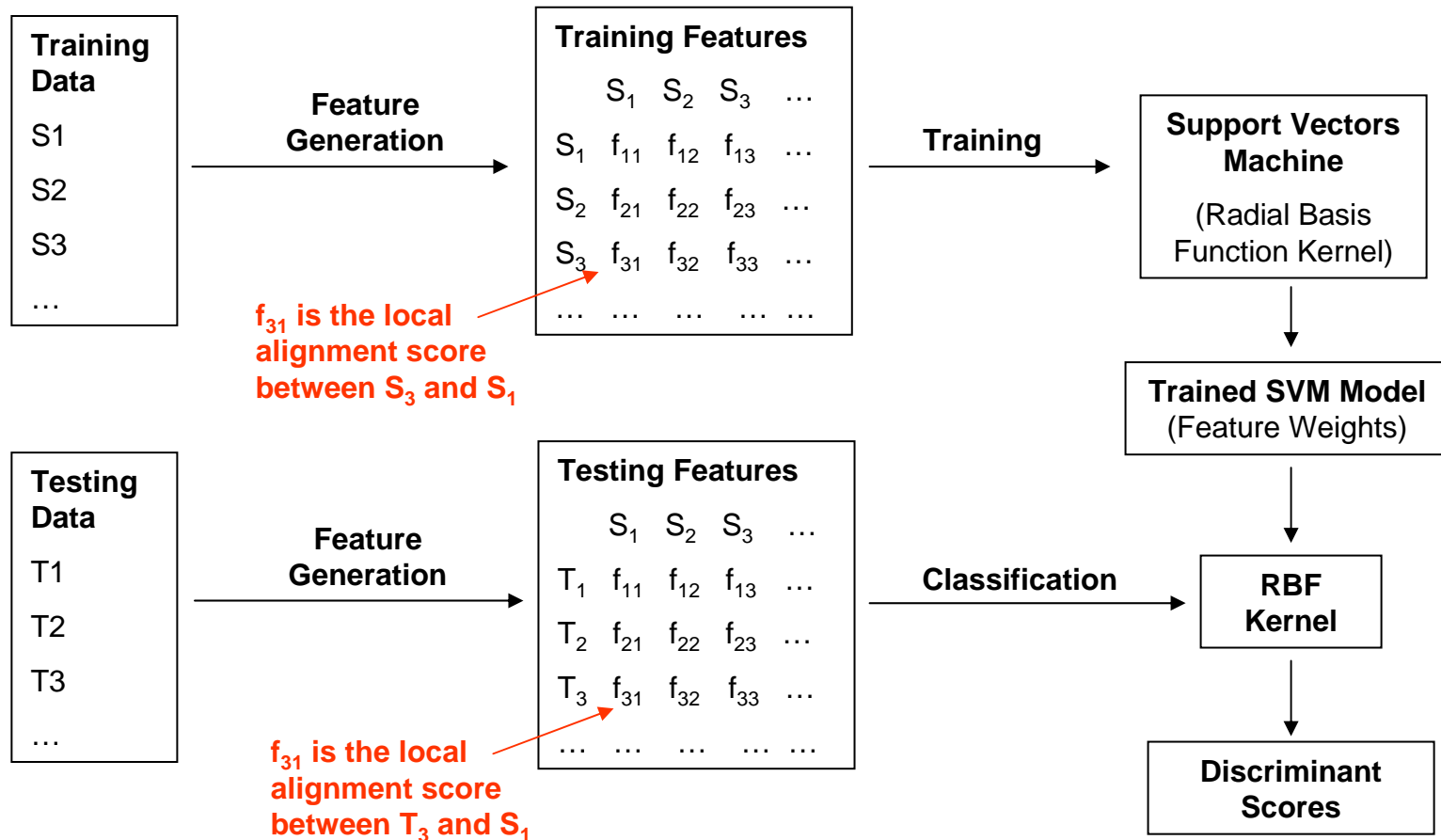
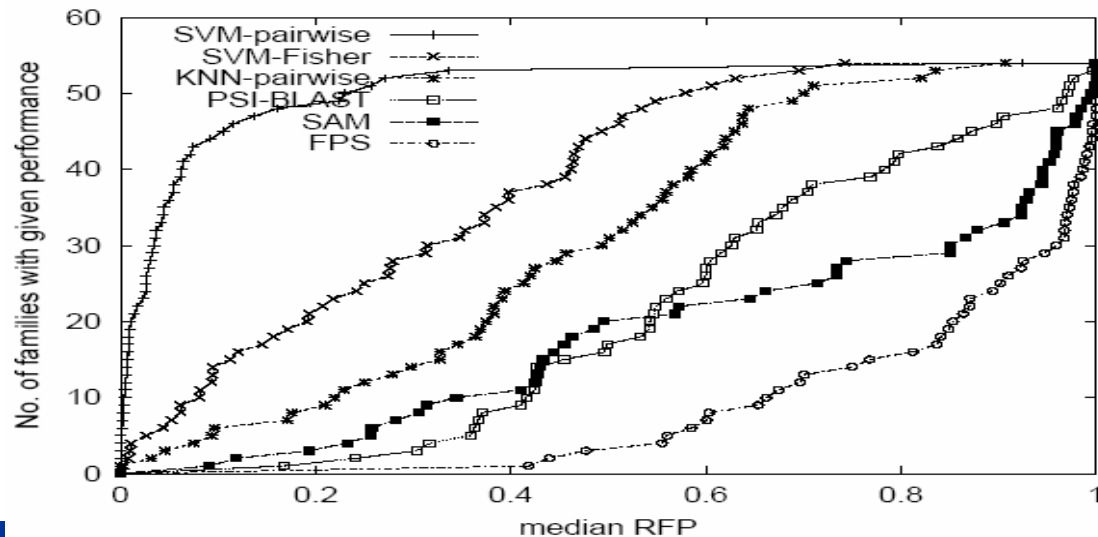
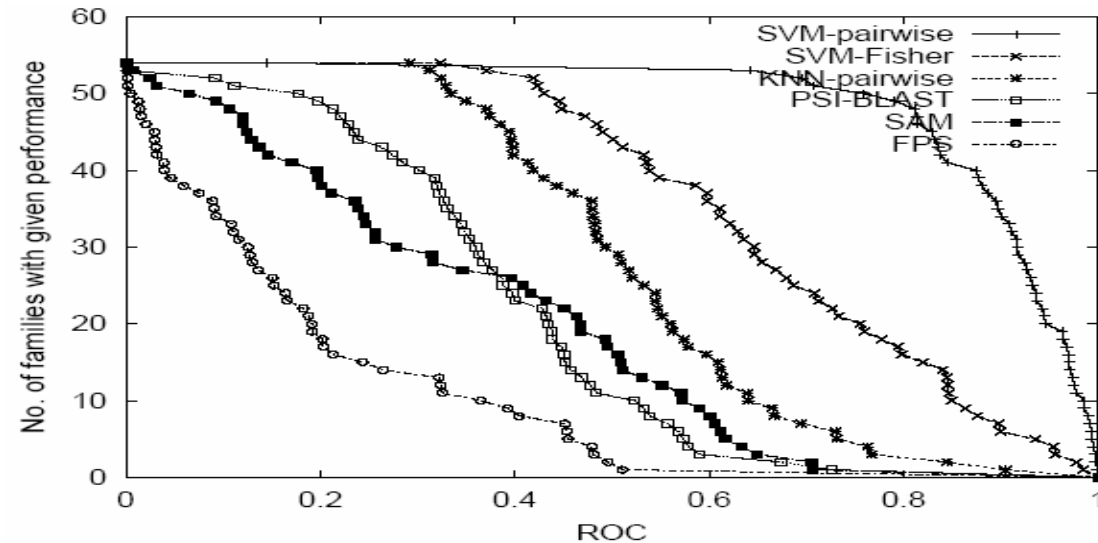


Image credit: Kenny Chua

# 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.



# 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
```

```
EEEFKKLTSIKIQNDKMRTGNLFPANMKKNRVLQIIPYEFNRVIIPVKRGEENTDYVNASF  
IDGYRQKDSYIASQGPLLHTIEDFWRMIWEWKSCSIVMLTELEERGQEKCAQYWPSDGLV  
SYGDITVELKKEEECESYTVRDLLVTNTRENKSRQIRQFHFHGWPEVVGIPSDGKGMISII  
AAVQKQQQQSGNHPITVHCSAGAGRTGTFCALSTVLERVKAEGILDVVFQTVKSLRLQRP  
MVQTLQYEFQYKVVQYIDAFSDYANFK
```

A red arrow points from the word 'typical' in the section header above to the sequence 'IPYEFNRVI' in the protein sequence.

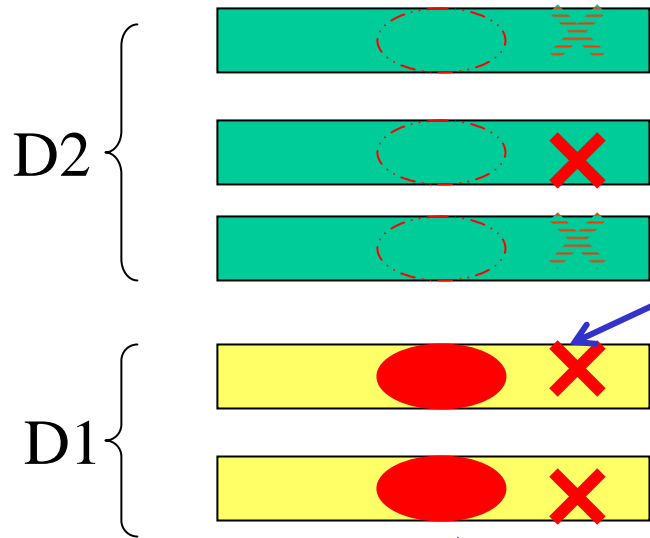
- **Some PTPs have 2 PTP domains**
- **PTP domain D1 is has much more activity than PTP domain D2**
- **Why? And how do you figure that out?**

# 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

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
* .. **.*.*
  
```

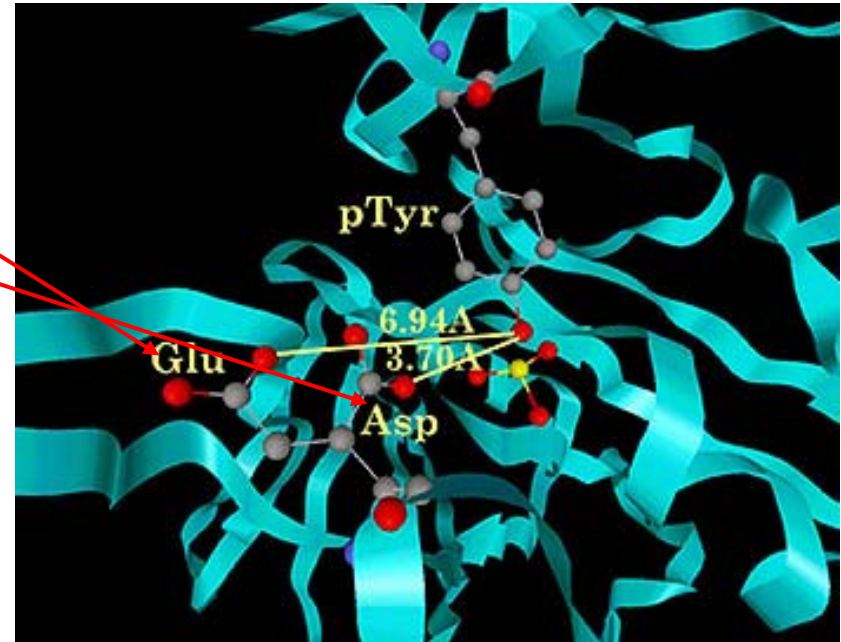


Image credit: Kolatkar

- 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?**

# Application of Sequence Comparison: From Looking for Similarities To Looking for Differences

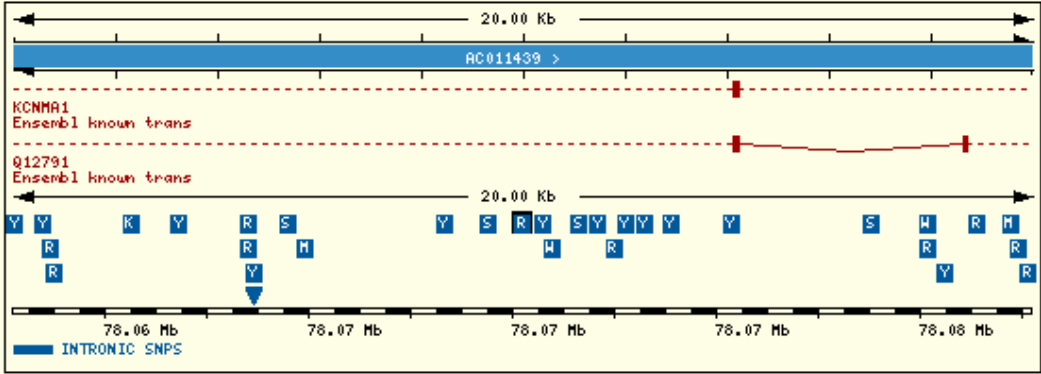


# Single Nucleotide Polymorphism

- SNP occurs when a single nucleotide replaces one of the other three nucleotide letters
- E.g., the alteration of the DNA segment **AAGGTTA** to **ATGGTTA**
- SNPs occur in human population > 1% of the time
- Most SNPs are found outside of "coding seqs" (Exercise: Why?)
  - ⇒ SNPs found in a coding seq are of great interest as they are more likely to alter function of a protein

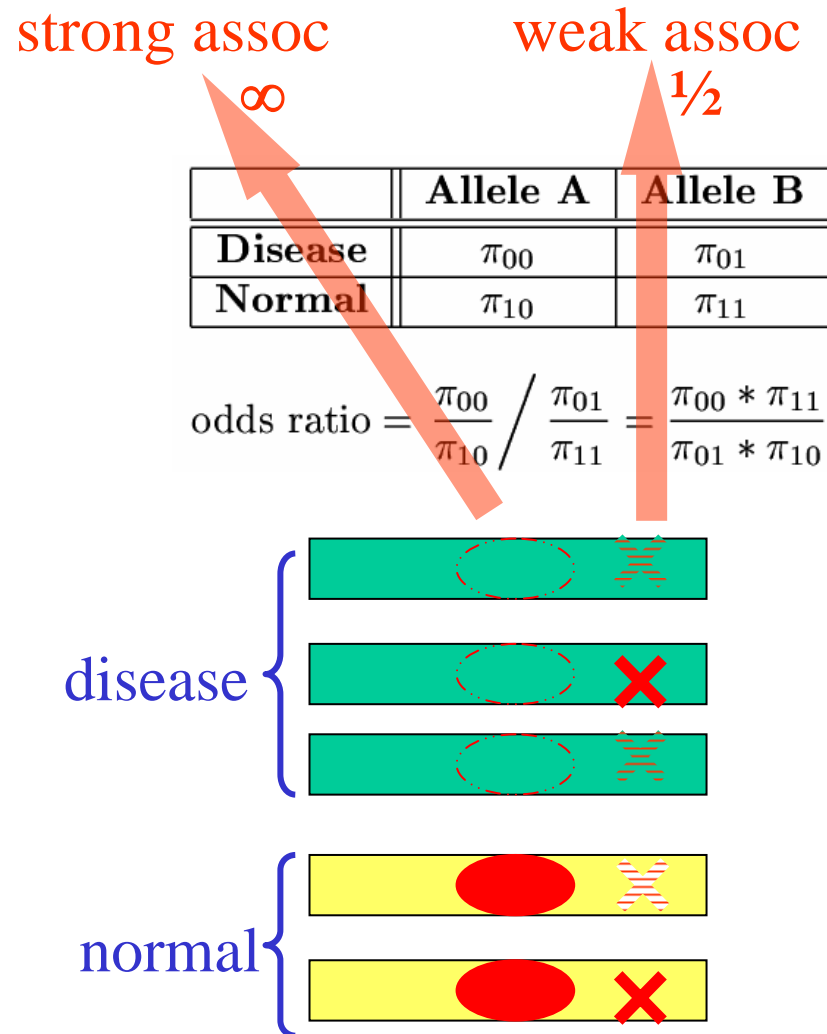
# Example SNP Report

## Ensembl SNP Report

|                          |  |
|--------------------------|--|
| <b>SNP</b>               | 1907745  |
| <b>Source</b>            | dbSNP  |
| <b>Synonyms</b>          | <b>dbSNP:</b> <a href="#">1907745</a><br><b>TSC:</b> <a href="#">TSC0953388</a><br><b>HGbase:</b> <a href="#">SNP001275703</a> |
| <b>Score</b>             | 1  |
| <b>Validation Status</b> | proven by cluster (SNP tested and validated by a non-computational method)   |
| <b>Alleles</b>           | A G (ambiguity code: <b>R</b> )  |
| <b>Sequence Region</b>   | AGGCATCCAGTCTCGGTAAACCTAG <b>R</b> CAAGTAATATTATTAGTTGAGCATT (SNP highlighted)   |
| <b>SNP neighbourhood</b> |   |

# SNP Uses

- **Association studies**
  - Analyze DNA of group affected by disease for their SNP patterns
  - Compare to patterns obtained from group unaffected by disease
  - Detect diff betw SNP patterns of the two
  - Find pattern most likely associated with disease-causing gene

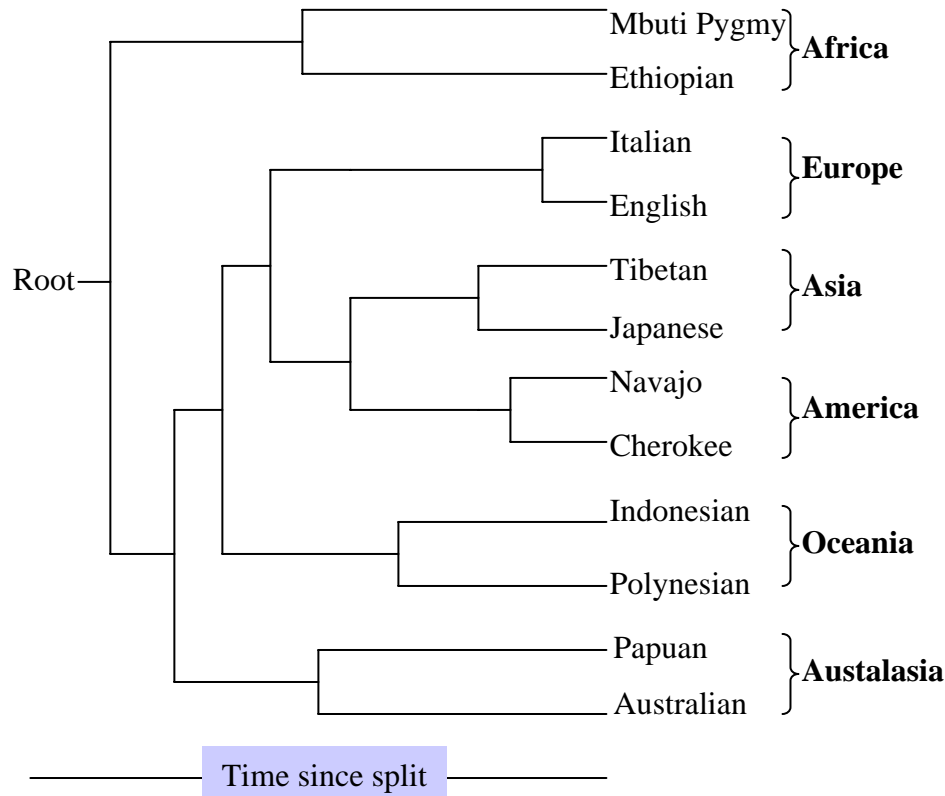


# Application of Sequence Comparison: The 7 Daughters of Eve



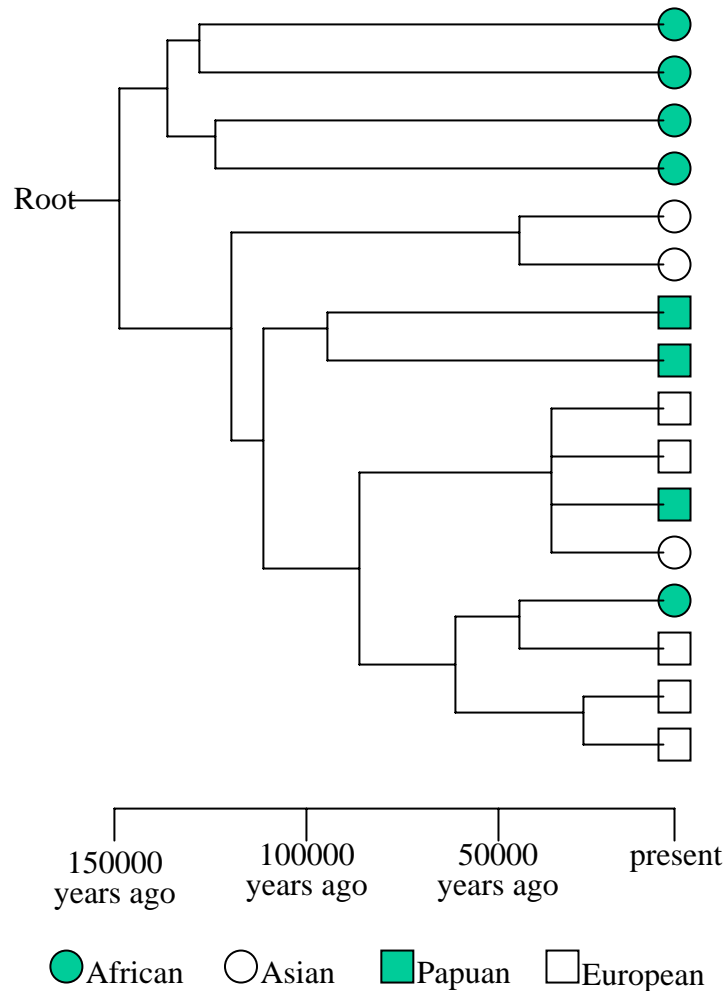


# Population Tree



- Estimate order in which “populations” evolved
- Based on assimilated freq of many different genes
- But ...
  - is human evolution a succession of population fissions?
  - Is there such thing as a proto-Anglo-Italian population which split, never to meet again, and became inhabitants of England and Italy?

# Evolution Tree



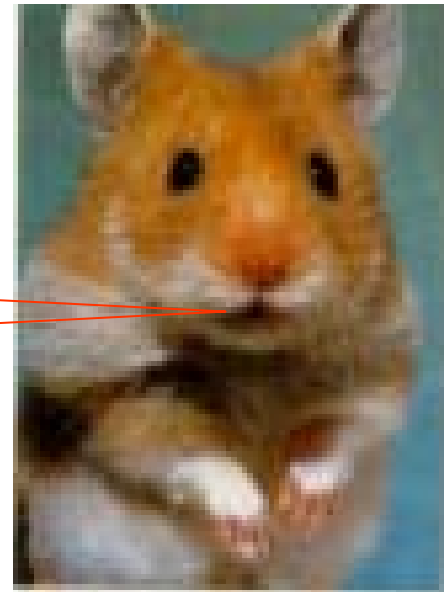
- Leaves and nodes are individual persons---real people, not hypothetical concept like “proto-population”
- Lines drawn to reflect genetic differences between them in one special gene called mitochondrial DNA

# Why Mitochondrial DNA

- **Present in abundance in bone fossils**
- **Inherited only from mother**
- **Sufficient to look at the 500bp control region**
- **Accumulate more neutral mutations than nuclear DNA**
- **Accumulate mutations at the “right” rate, about 1 every 10,000 years**
- **No recombination, not shuffled at each generation**

# Mutation Rates

- All pet golden hamsters in the world descend from a single female caught in 1930 in Syria
- Golden hamsters “manage” ~4 generations a year :-)
- So >250 hamster generations since 1930
- Mitochondrial control regions of 35 (independent) golden hamsters were sequenced and compared
- No mutation was found



⇒ Mitochondrial control region mutates at the “right” rate

# Contamination

- **Need to know if DNA extracted from old bones really from those bones, and not contaminated with modern human DNA**
  - **Apply same procedure to old bones from animals, check if you see modern human DNA**
- ⇒ **If none, then procedure is OK**

# Origin of Polynesians

- Do they come from Asia or America?

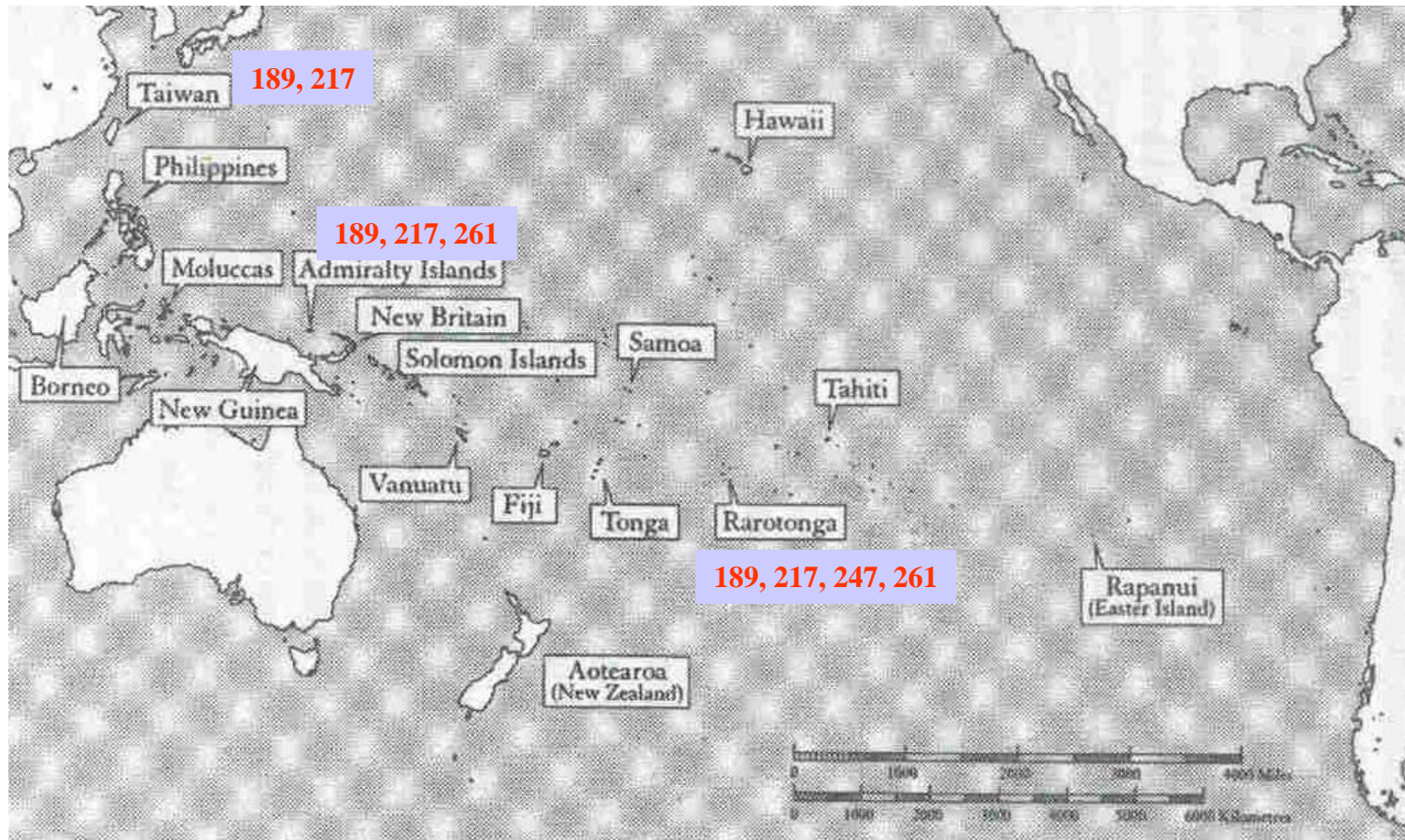
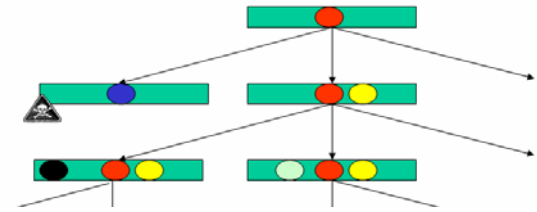


Image credit: Sykes



# Origin of Polynesians

- Common mitochondrial control seq from Rarotonga have variants at positions 189, 217, 247, 261. Less common ones have 189, 217, 261
- Seq from Taiwan natives have variants 189, 217
- Seq from regions in betw have variants 189, 217, 261.
- More 189, 217 closer to Taiwan. More 189, 217, 261 closer to Rarotonga
- 247 not found in America  
⇒ Polynesians came from Taiwan!
- Taiwan seq sometimes have extra mutations not found in other parts  
⇒ These are mutations that happened since Polynesians left Taiwan!



# Neanderthal vs Cro Magnon

- Are Europeans descended purely from Cro Magnons? Pure Neanderthals? Or mixed?



Neanderthal



Cro Magnon





# Neanderthal vs Cro Magnon

- Based on palaeontology, Neanderthal & Cro Magnon last shared an ancestor 250k years ago
- Mitochondrial control regions accumulate 1 mutation per 10k years
  - ⇒ If Europeans have mixed ancestry, mito-chondrial control regions betw 2 Europeans should have ~25 diff w/ high probability
- The number of diff betw Welsh is ~3, & at most 8.
- When compared w/ other Europeans, 14 diff at most
  - ⇒ Ancestor either 100% Neanderthal or 100% Cro Magnon
- Mitochondrial control seq from Neanderthal have 26 diff from Europeans
  - ⇒ Ancestor must be 100% Cro Magnon

# Suggested Readings



# References

- S.E.Brenner. “Errors in genome annotation”, *TIG*, 15:132--133, 1999
- 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.

# References

- J. Park et al. “Sequence comparisons using multiple sequences detect three times as many remote homologs as pairwise methods”, *JMB*, 284(4):1201-1210, 1998
- J. Park et al. “Intermediate sequences increase the detection of homology between sequences”, *JMB*, 273:349--354, 1997
- Z. Zhang et al. “Protein sequence similarity searches using patterns as seeds”, *NAR*, 26(17):3986--3990, 1996
- M.S.Gelfand et al. “Gene recognition via spliced sequence alignment”, *PNAS*, 93:9061--9066, 1996
- B. Ma et al. “PatternHunter: Faster and more sensitive homology search”, *Bioinformatics*, 18:440--445, 2002
- M. Li et al. “PatternHunter II: Highly sensitive and fast homology search”, *GIW*, 164--175, 2003

# References

- S.F.Altshcul et al. “Basic local alignment search tool”, *JMB*, 215:403--410, 1990
- S.F.Altschul et al. “Gapped BLAST and PSI-BLAST: A new generation of protein database search programs”, *NAR*, 25(17):3389--3402, 1997.
- 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
- B. Sykes. “The seven daughters of Eve”, *Gorgi Books*, 2002

# References

- P. Erdős & A. Rényi. “On a new law of large numbers”, *J. Anal. Math.*, 22:103--111, 1970
- R. Arratia & M. S. Waterman. “Critical phenomena in sequence matching”, *Ann. Prob.*, 13:1236--1249, 1985
- R. Arratia, P. Morris, & M. S. Waterman. “Stochastic scrabble: Large deviations for sequences with scores”, *J. Appl. Prob.*, 25:106--119, 1988
- R. Arratia, L. Gordon. “Tutorial on large deviations for the binomial distribution”, *Bull. Math. Biol.*, 51:125--131, 1989