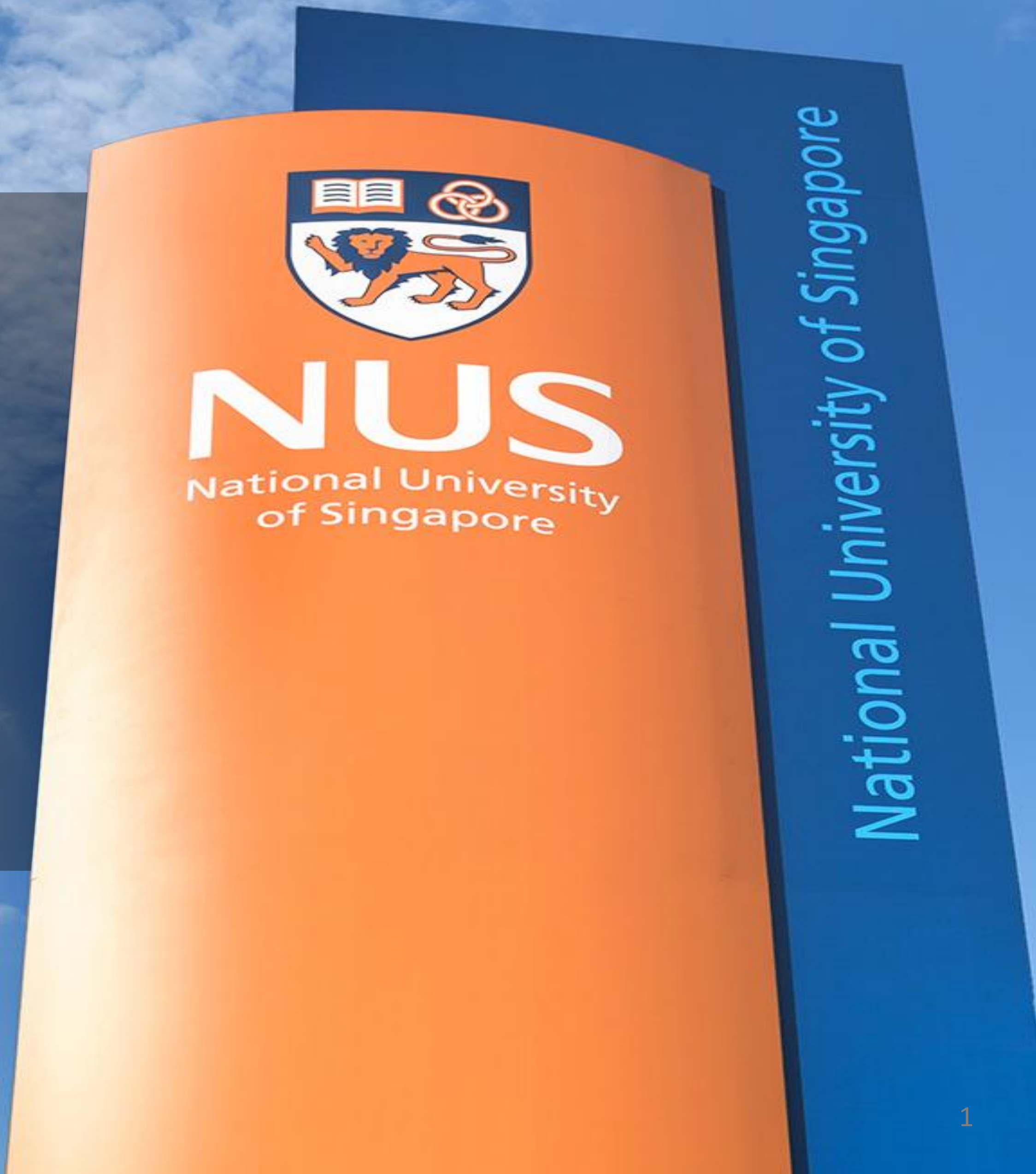
CS2220: Introduction to Computational Biology

Sequence Database Search

Wong Limsoon



Outline

Popular tools for fast database search

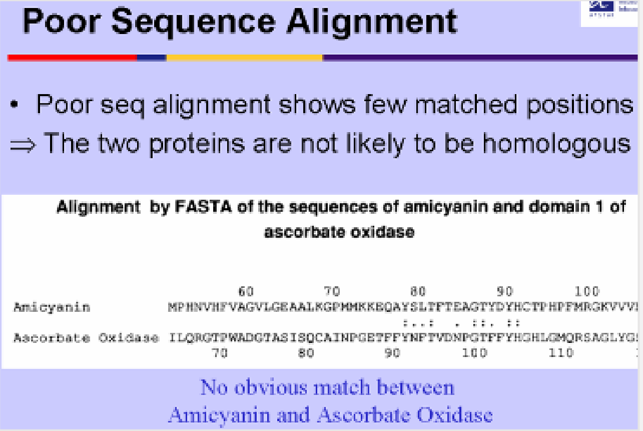
FASTA

BLAST

Pattern Hunter, …

Cautionary tales

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



Poor Sequence Alignment

- Poor seq alignment shows few matched positions

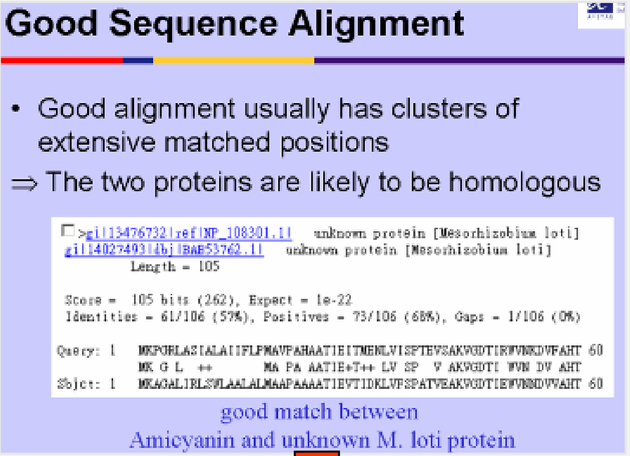
⇒ The two proteins are not likely to be homologous

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

```
                         60        70        80        90        100
Amicyanin         MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVV
                                                 :..:   . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYG
                     70        80        90        100       110
```

No obvious match between
Amicyanin and Ascorbate Oxidase

Discard this function as a candidate



Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

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

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

Query: 1   MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
            MK G L  ++        MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

good match between
Amicyanin and unknown M. loti protein

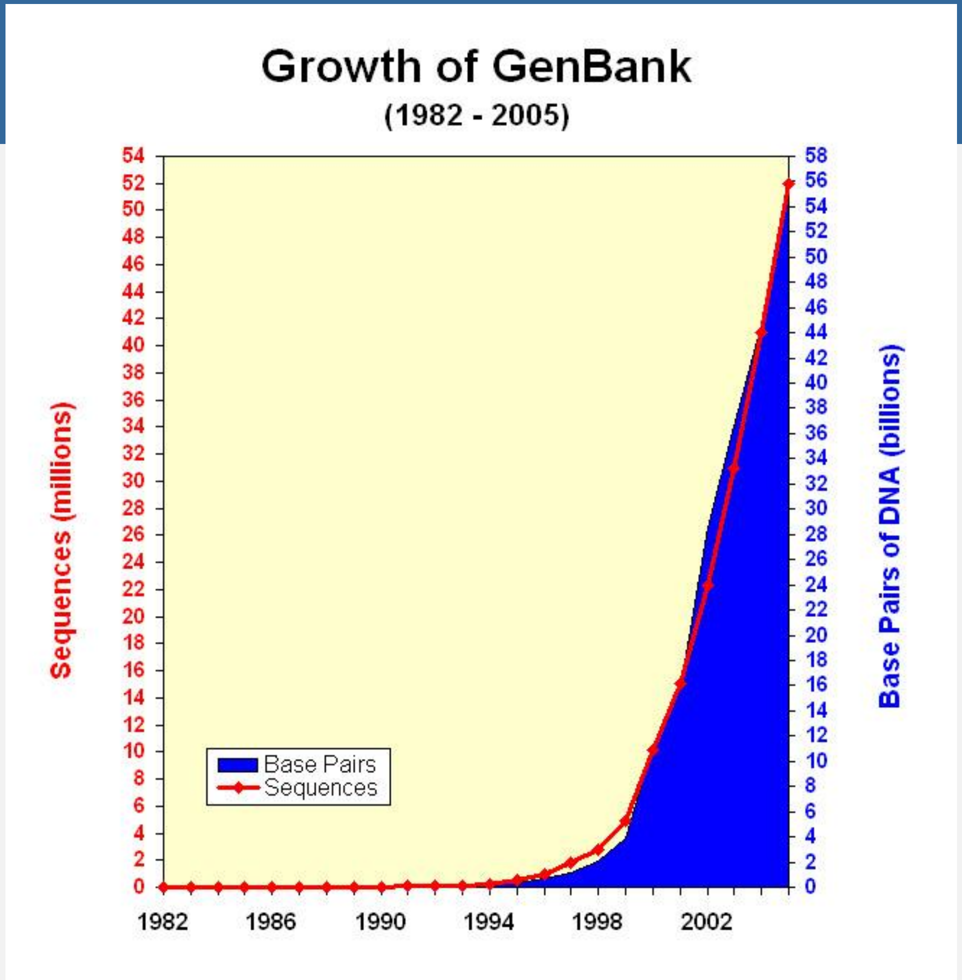
Assign to *T* same function as homologs

Confirm with suitable wet experiments

Scaling challenge

Increasing # of sequenced genomes: yeast, human, rice, mouse, fly, …

S/w must be “linearly” scalable to large datasets



Performance of a search algorithm

Sensitivity

Ability to detect “true positive”

Measured as the probability of finding the match given the query and the database sequence has only x% similarity

Specificity

Ability to reject “false positive”

A good search algorithm should be both sensitive and specific

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, so …

Break entire string into substrings, index the substrings

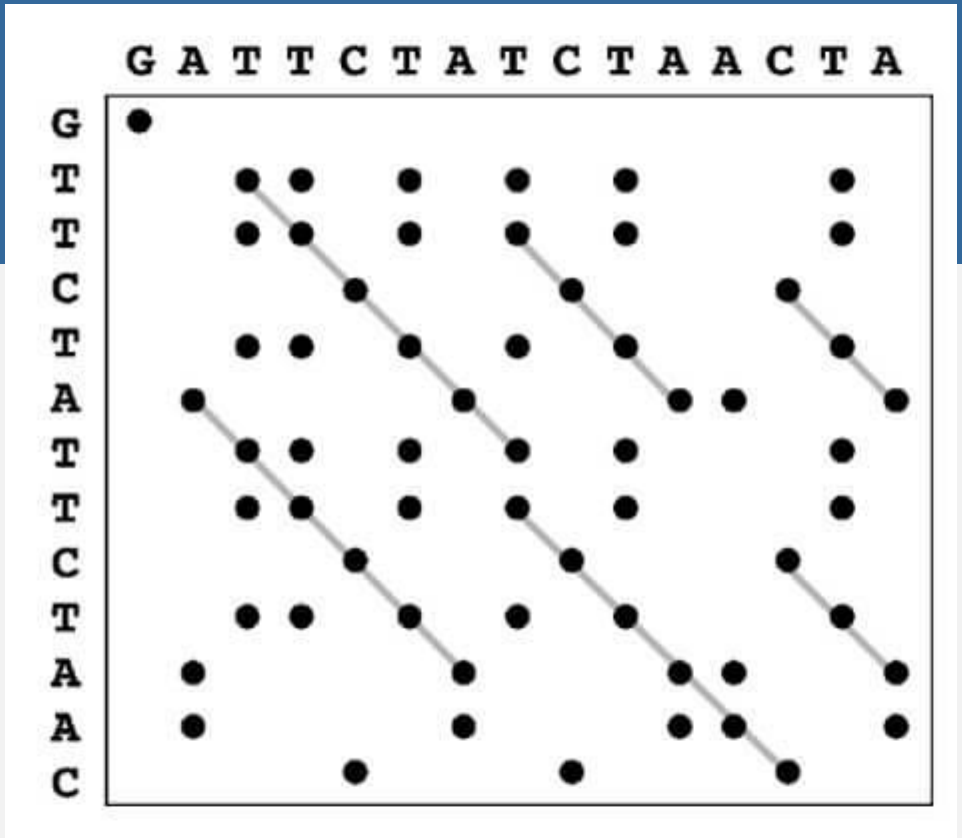
Search for matching short substrings and use as seed for further analysis

Extend to entire string find the most significant local alignment segment

FASTA

Overview of FASTA

- Fast sequence search
- Based on dotplot
- Identify identical words (k-tuples)
- Search significant diagonals
- Use PAM 250 for further refinement
- Dynamic programming for narrow region



Dotplot: For a simple visual representation of the similarity between two sequences, individual cells in the matrix can be shaded black if residues are identical, so that matching sequence segments appear as runs of diagonal lines across the matrix.

Image credit: https://microbenotes.com/local-global-multiple-sequence-alignment/

FASTA algorithm

Divide query sequence into its constituent overlapping words (ktup) of length k; default: 2 for proteins and 6 for nucleic acids

Each sequence in the database is also broken up in the same way

Two word lists are compared to find all identical words in both sequences

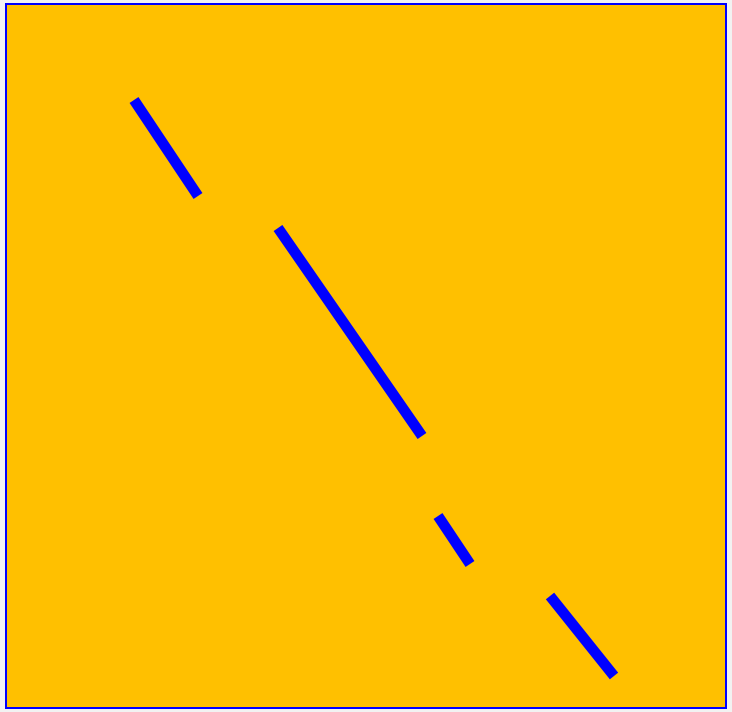
```
CTGCACTA        AGCTGACGCA
CTG               CTG
 TGC                     GCA
  GCA
     etc.
```

FASTA algorithm, cont’d

Ktup matches can be depicted in a matrix; diagonals indicate matches

For every library sequence, the 10 best diagonals are constructed from the ktup matches using a distance formula

The top 10 diagonals are rescored using substitution matrices; each of these rescored diagonals is called an initial region



FASTA, further cont’d

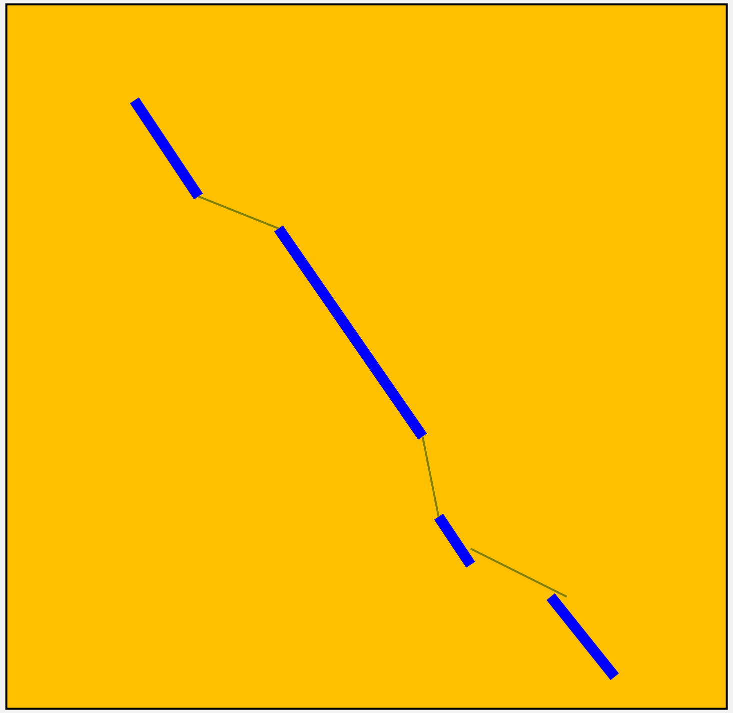
Initial regions are joined with a joining penalty (similar to a gap penalty)

The highest joined score defines the library sequence’s score

Library sequences are ranked by this score

If the score is high, a Smith–Waterman alignment is run in the same dot-plot region using the same window

The resulting score is reported as the optimal score



BLAST

Overview of BLAST

Similarity matching of words

3 aa's, 11 bases

No need identical words

If no words are similar

No alignment

Won't find matches for very short sequences

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

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

Step 1 of BLAST: For the query, find the list of high scoring words of length w

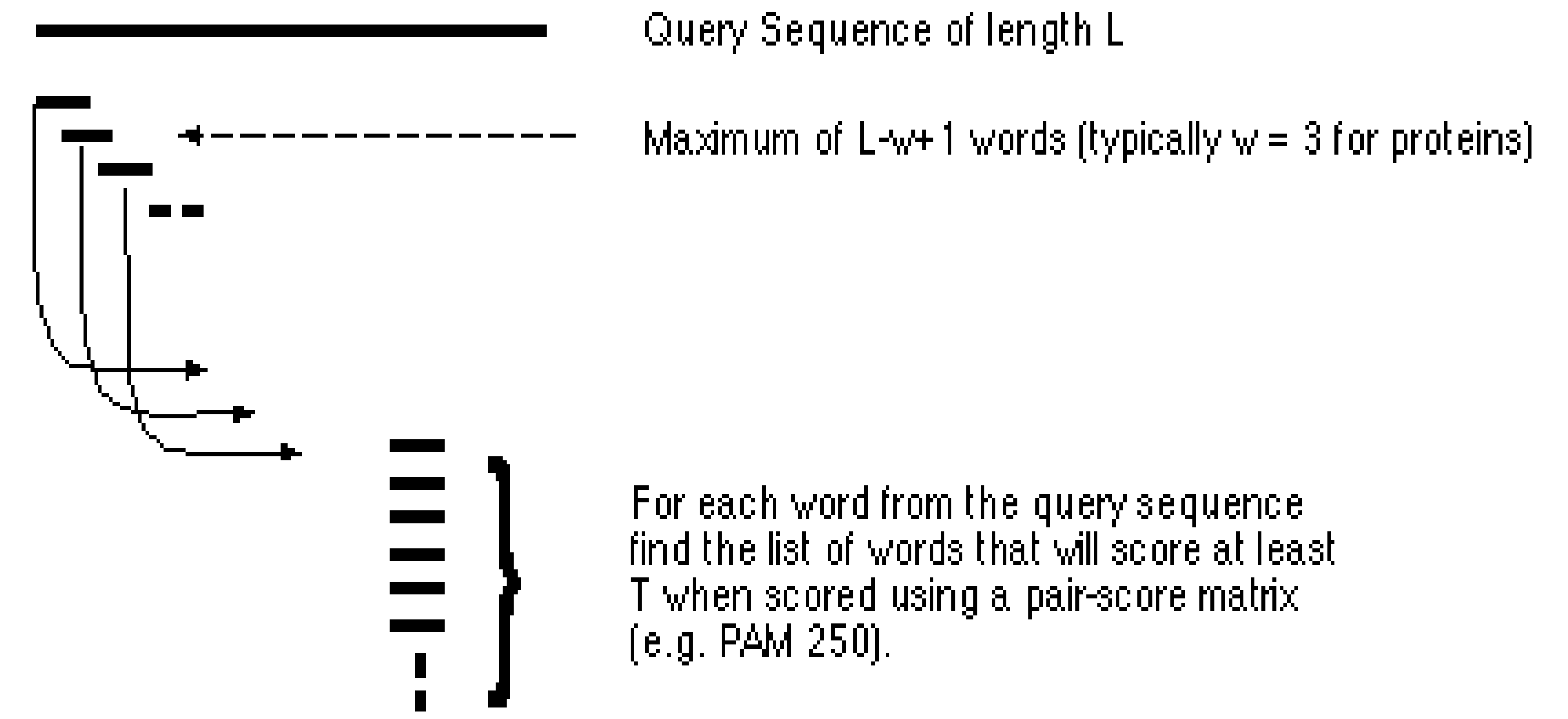

Query Sequence of length L
Maximum of L-w+1 words (typically w = 3 for proteins)
For each word from the query sequence find the list of words that will score at least T when scored using a pair-score matrix (e.g. PAM 250).

Image credit: Barton

Step 2 of BLAST: Compare word list to db & find exact matches

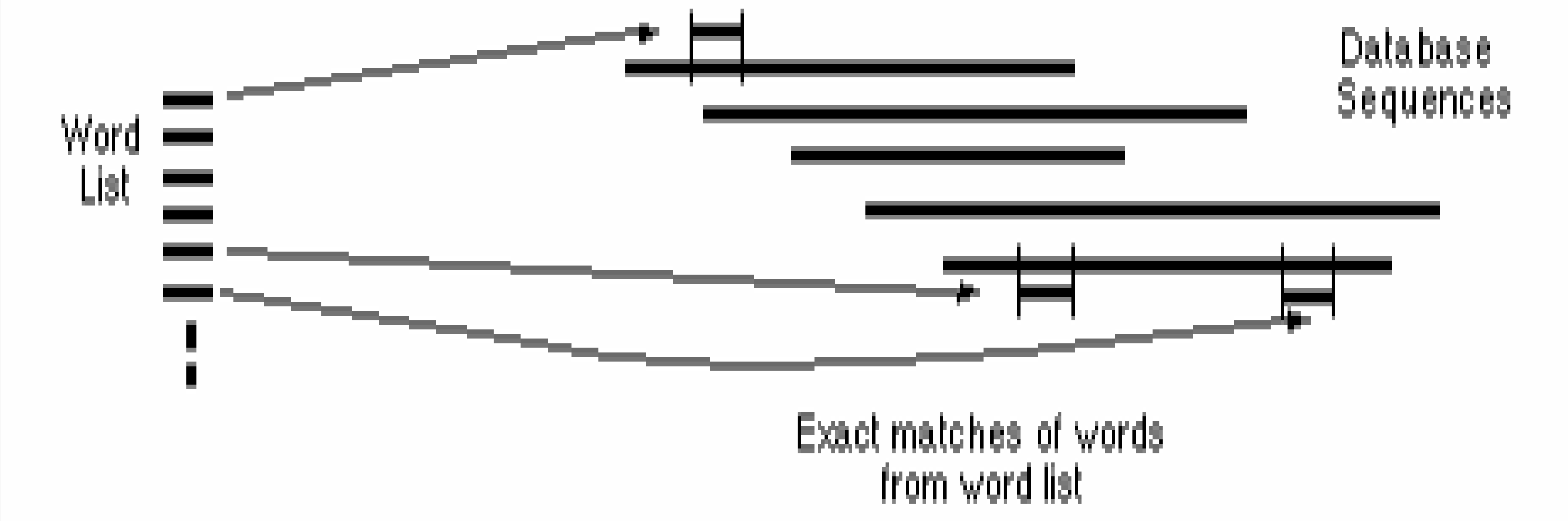


Image credit: Barton

Step 3 of BLAST: For each word match, extend alignment in both directions to find alignment that score greater than a threshold s

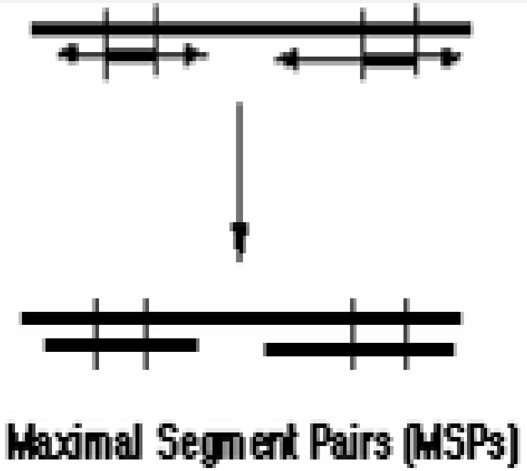


Image credit: Barton

Exercise

Is BLAST generally more sensitive and more efficient than FASTA? Why?

Spaced seeds

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
```

11111111111 is the BLAST seed model for comparing DNA seqs

Observations on spaced seeds

Seed models w/ different shapes can detect different homologies

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 homologies

Approaching 100% sensitive homology search

PatternHunter II

PatternHunter I

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

⇒ Wasteful

```
TTGACCTCACC?
|||||||||||?
TTGACCTCACC?
11111111111
 11111111111
```

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

Spaced seeds uses fewer hits to detect one homology

⇒ Efficient

```
CAA?A??A?C??TA?TGG?
|||?|??|?|??||?|||?
CAA?A??A?C??TA?TGG?
111010010100110111
 111010010100110111
```

$1/4^6$ chances to have 2nd hit next to the 1st hit

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 $(L – M + 1) * p^W$

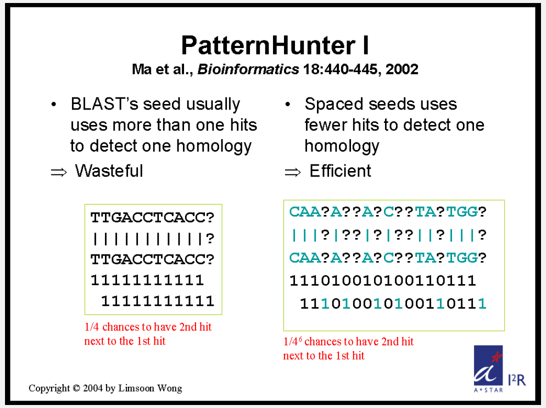
Proof:

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

There are $L – M + 1$ candidate positions

The proposition follows

Implication



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

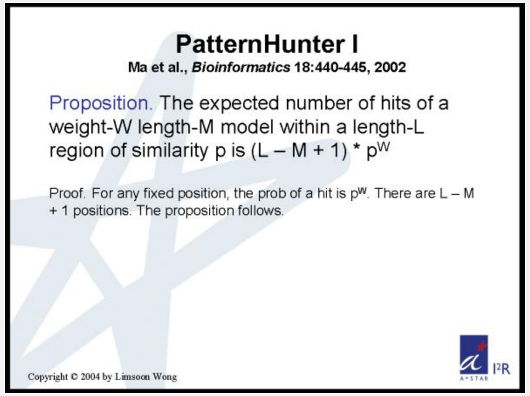
```
TTGACCTCACC?
|||||||||||?
TTGACCTCACC?
11111111111
 11111111111
```

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

```
CAA?A??A?C??TA?TGG?
|||?|??|?|??||?|||?
CAA?A??A?C??TA?TGG?
111010010100110111
 111010010100110111
```

$1/4^6$ 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

For L = 1017

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

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

Spaced seeds likely to be more sensitive & more efficient

Sensitivity of PatternHunter I

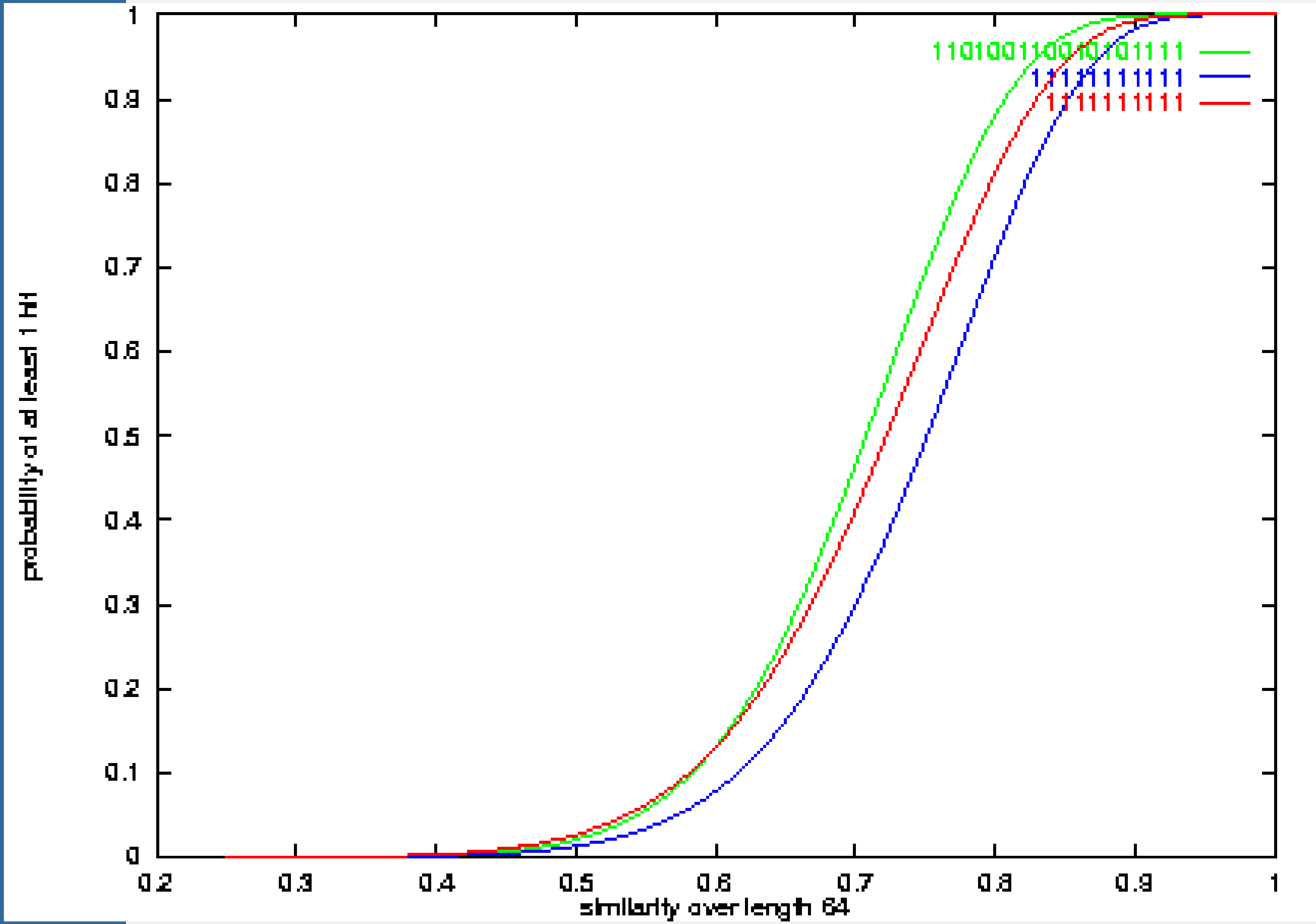
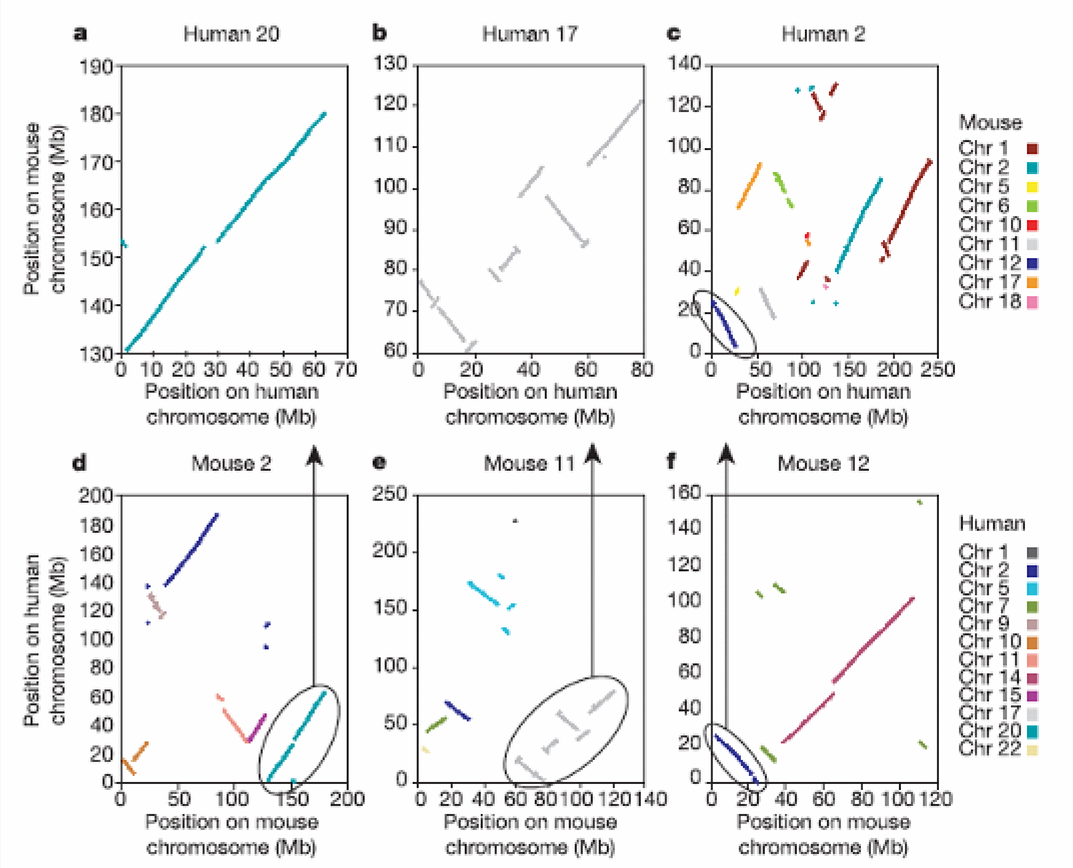


Image credit: Ming Li

Speed of PatternHunter I



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!

Nature, 420:520-522, 2002

How to increase sensitivity?

Ways to increase sensitivity:

“Optimal” seed

Reduce weight by 1

Increase number of spaced seeds by 1

Intuitively, for DNA sequences,

Reducing weight by 1 will increase number of matches 4 folds

Doubling number of seeds will increase number of matches 2 folds

Exercise

Is it better to use two spaced seeds or a reduced weight spaced seed?

PatternHunter II

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

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

See also Ilie & Ilie, “Multiple spaced seeds for homology search”, *Bioinformatics*, 23(22):2969-2977, 2007

Algo to select multiple spaced seeds

Let A be an empty set

Let s be the seed such that A U {s} has the highest hit probability

A = A U {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

⇒ "Double the seed number" gains better sensitivity than "decrease the weight by 1"

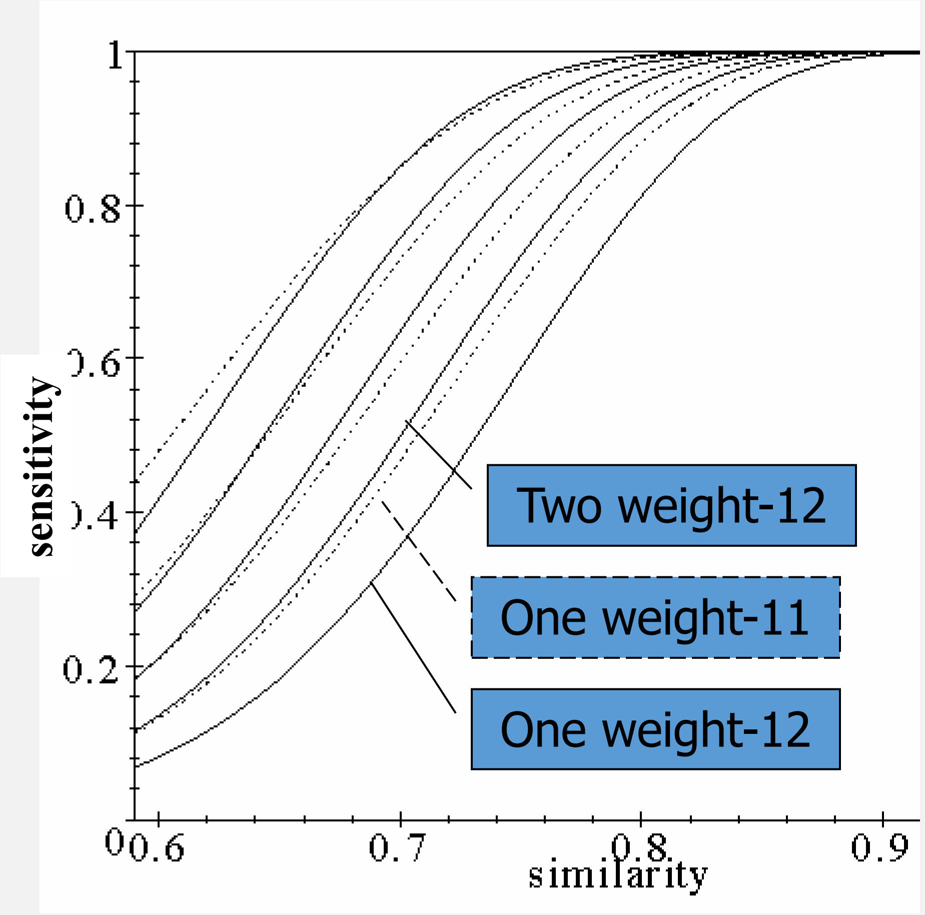


Image credit: Bin Ma

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

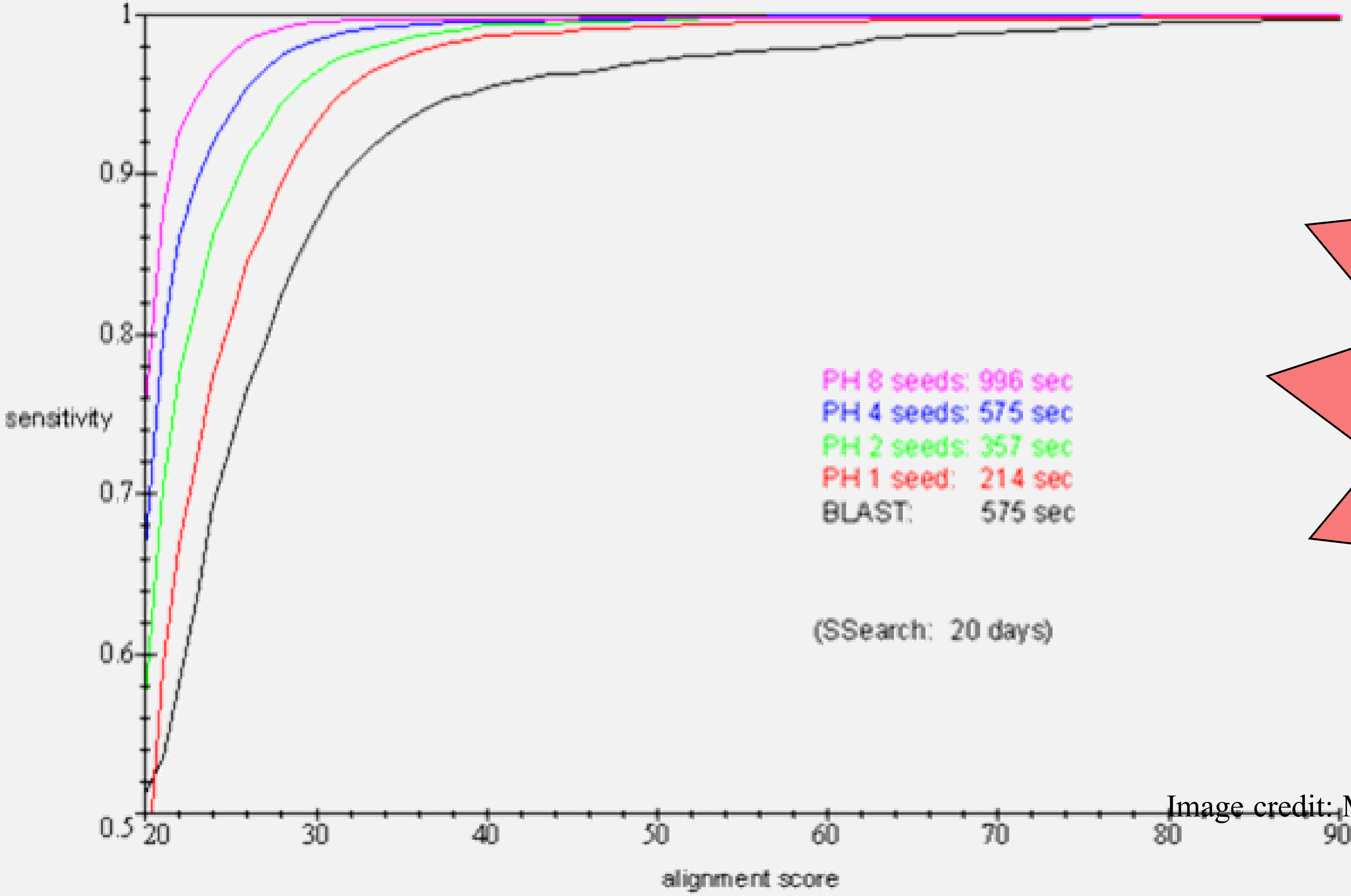
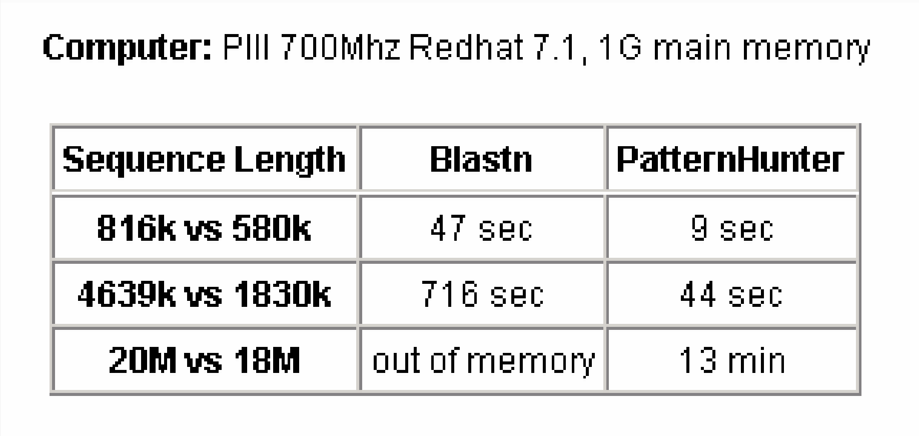


Image credit: Ma

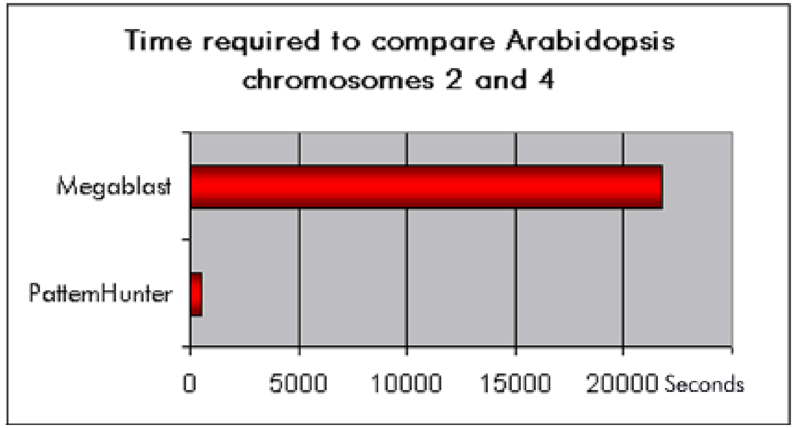
In fact, at 80% similarity, 100% sensitivity can be achieved using 40 weight-9 seeds

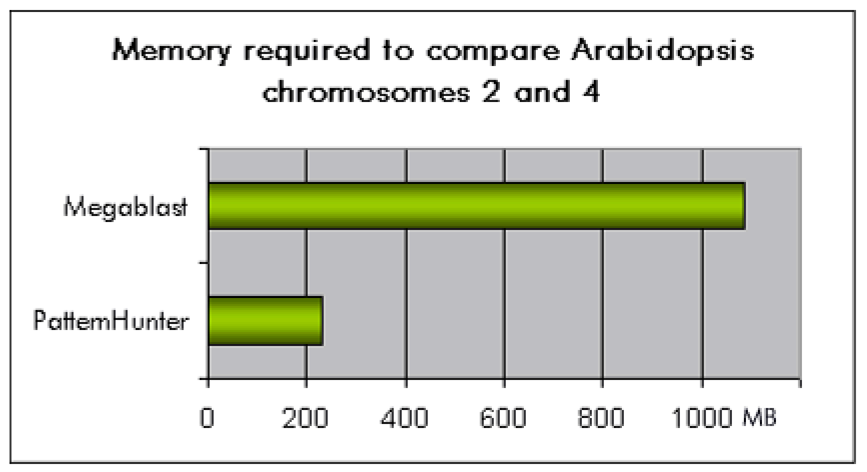
Farewell to Supercomputer Age of sequence comparison!



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





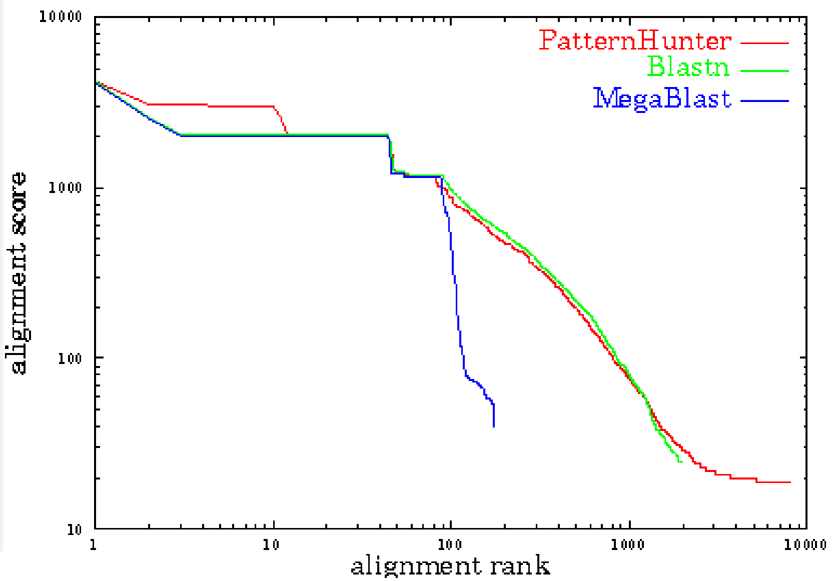


Image credit: Bioinformatics Solutions Inc

About the inventor: Ming Li

Ming Li

University Professor, Univ of Waterloo

Fellow, Royal Society of Canada

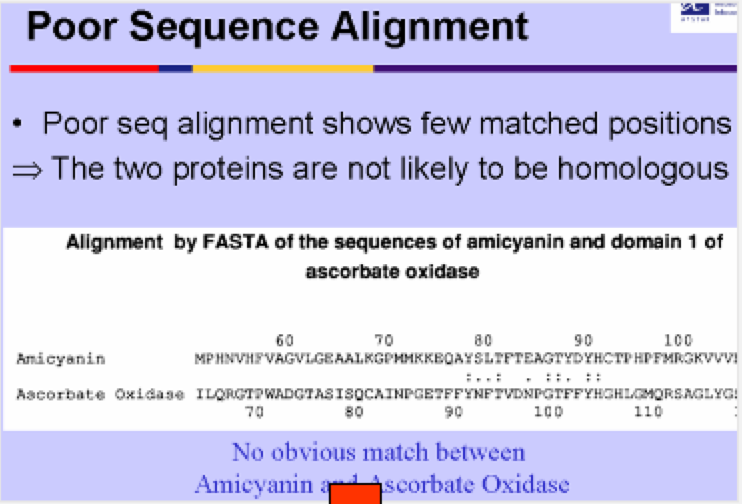
Fellow, ACM

Fellow, IEEE

Cautionary tales

Guilt by association

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



Poor Sequence Alignment

- Poor seq alignment shows few matched positions

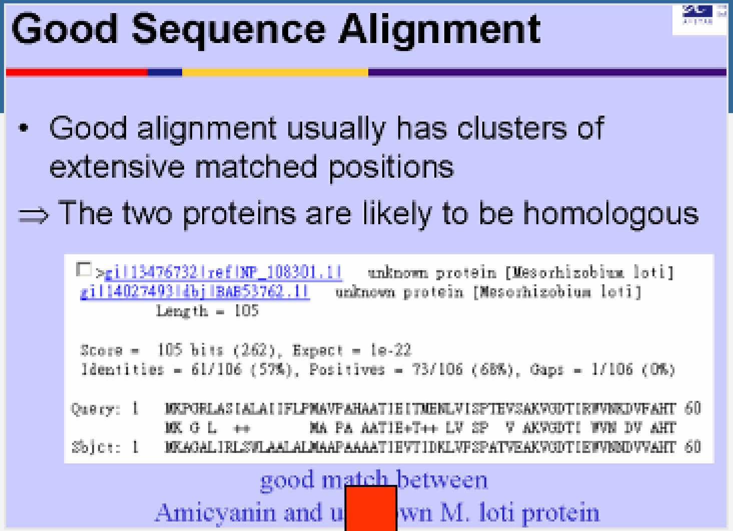
⇒ The two proteins are not likely to be homologous

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

```
                       60        70        80        90        100
Amicyanin         MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVV
                                             :..:   . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYG
                     70        80        90        100       110
```

No obvious match between
Amicyanin and Ascorbate Oxidase

Discard this function as a candidate



Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

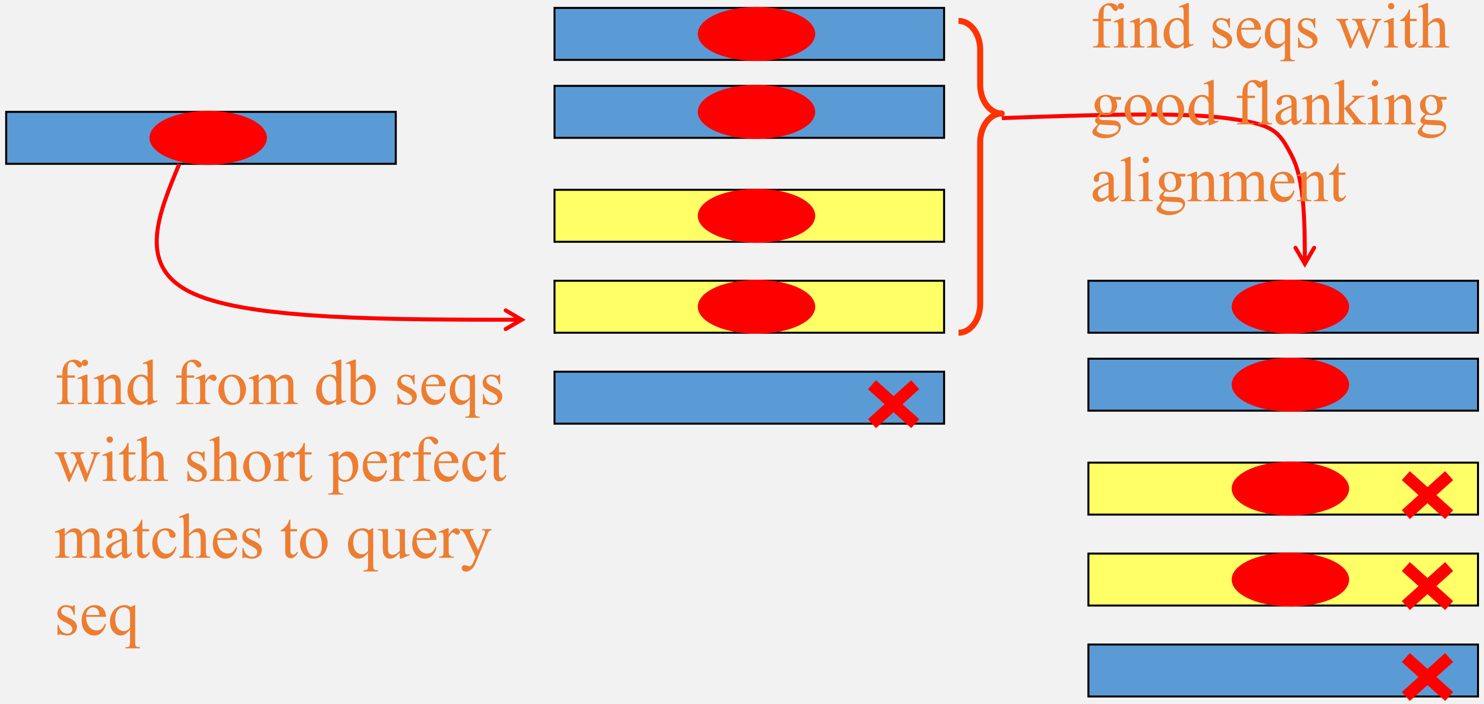
```
□ >gi|13476732|ref|NP_108301.1|   unknown protein [Mesorhizobium loti]
 gi|14027493|dbj|BAB53762.1|   unknown protein [Mesorhizobium loti]
          Length = 105

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

Query: 1   MKPGRLASIALAIIFLPMAVPAHAATIEITMENLVISPTEVSAKVGDTIRWVNKDVFAHT 60
            MK G L  ++        MA PA AATIE+T++ LV SP  V AKVGDTI WVN DV AHT
Sbjct: 1   MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDTIEWVNNDVVAHT 60
```

good match between
Amicyanin and unknown M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments

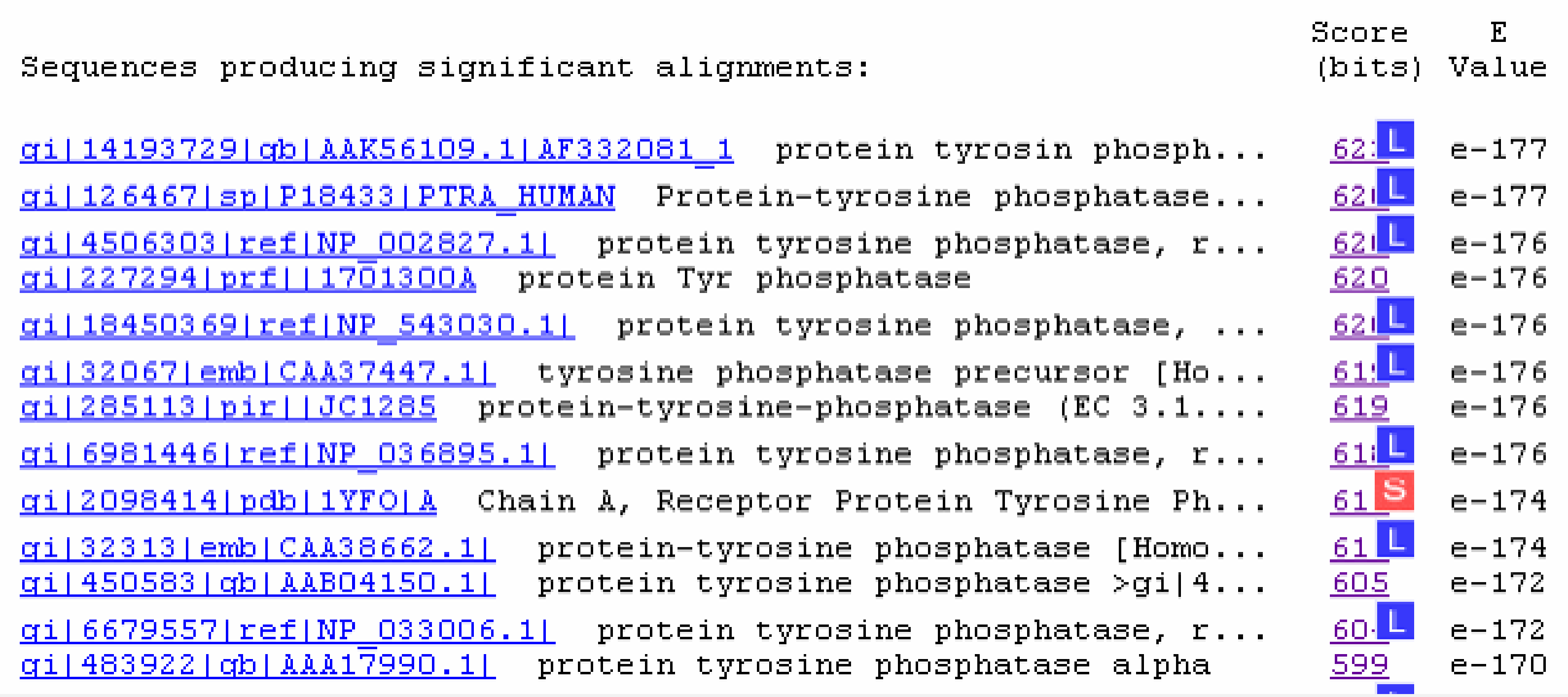
BLAST is one of the most popular tool for doing "guilt-by-association" sequence homology search



find from db seqs with short perfect matches to query seq

find seqs with good flanking alignment

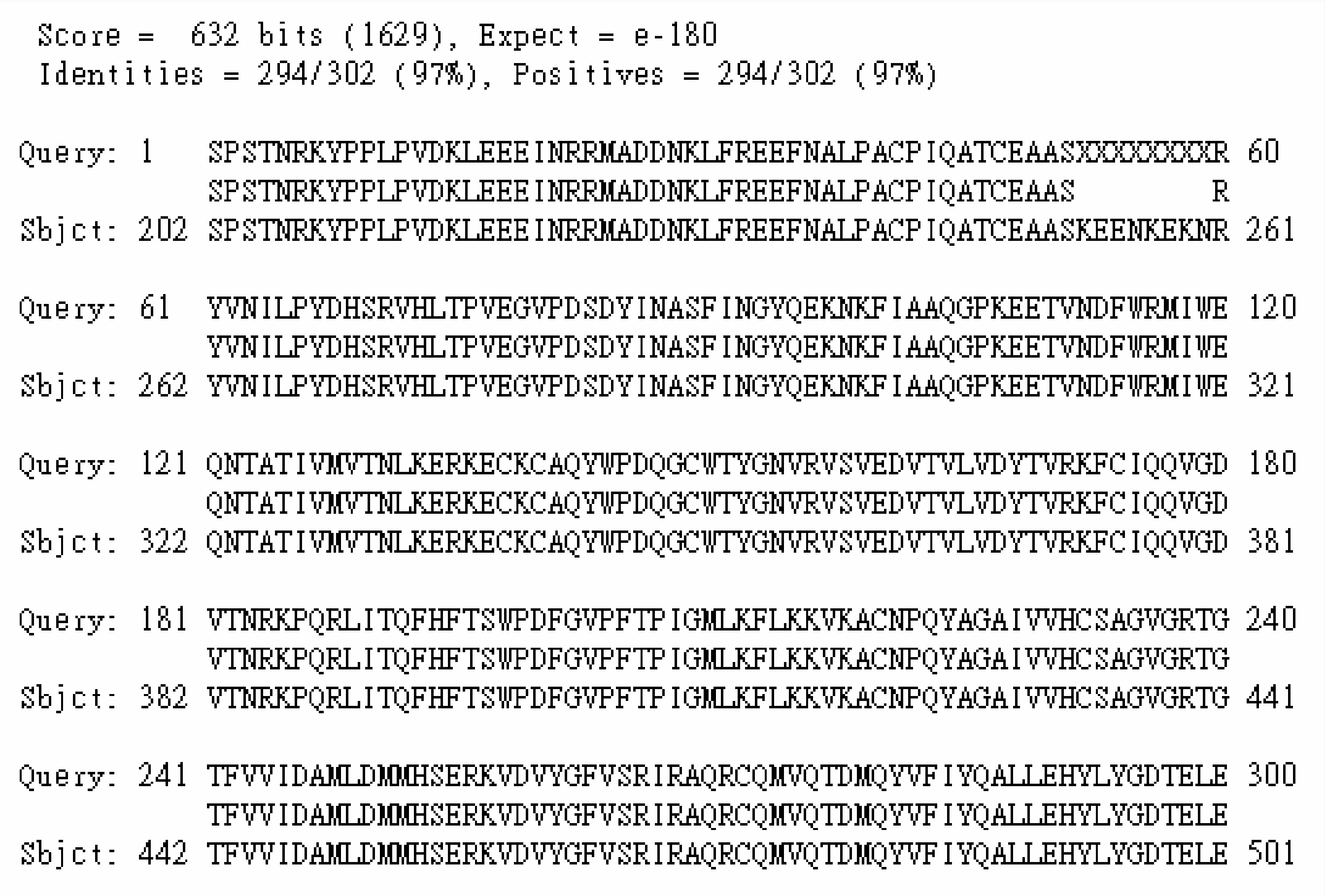
Homologs obtained by BLAST for a query sequence



```
                                                                     Score    E
Sequences producing significant alignments:                          (bits)  Value

gi|14193729|gb|AAK56109.1|AF332081_1  protein tyrosin phosph...       62 L   e-177
gi|126467|sp|P18433|PTRA_HUMAN  Protein-tyrosine phosphatase...       62 L   e-177
gi|4506303|ref|NP_002827.1|  protein tyrosine phosphatase, r...       62 L   e-176
gi|227294|prf||1701300A  protein Tyr phosphatase                      620    e-176
gi|18450369|ref|NP_543030.1|  protein tyrosine phosphatase, ...       62 L   e-176
gi|32067|emb|CAA37447.1|  tyrosine phosphatase precursor [Ho...       61 L   e-176
gi|285113|pir||JC1285  protein-tyrosine-phosphatase (EC 3.1....       619    e-176
gi|6981446|ref|NP_036895.1|  protein tyrosine phosphatase, r...       61 L   e-176
gi|2098414|pdb|1YFO|A  Chain A, Receptor Protein Tyrosine Ph...       61 S   e-174
gi|32313|emb|CAA38662.1|  protein-tyrosine phosphatase [Homo...       61 L   e-174
gi|450583|gb|AAB04150.1|  protein tyrosine phosphatase >gi|4...       605    e-172
gi|6679557|ref|NP_033006.1|  protein tyrosine phosphatase, r...       60 L   e-172
gi|483922|gb|AAA17990.1|  protein tyrosine phosphatase alpha          599    e-170
```

Thus, the query sequence could be a protein tyrosine phosphatase α (PTPα)

Example alignment with PTPα



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

Query: 1   SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASXXXXXXXXR 60
            SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAAS        R
Sbjct: 202 SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR 261

Query: 61  YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE 120
            YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE
Sbjct: 262 YVNILPYDHSRVHLTPVEGVPDSDYINASFINGYQEKNKFIAAQGPKEETVNDFWRMIWE 321

Query: 121 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD 180
            QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
Sbjct: 322 QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD 381

Query: 181 VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG 240
            VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG
Sbjct: 382 VTNRKPQRLITQFHFTSWPDFGVPFTPIGMLKFLKKVKACNPQYAGAIVVHCSAGVGRTG 441

Query: 241 TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE 300
            TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE
Sbjct: 442 TFVVIDAMLDMMHSERKVDVYGFVSRIRAQRCQMVQTDMQYVFIYQALLEHYLYGDTELE 501
```

Guilt by association: Caveats

Ensure that the effect of database size has been accounted for

Ensure that the function of the 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 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

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

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

$\Rightarrow$ Correct for database size if your seq comparison prog does not

Lightning does strike twice!

Roy Sullivan, a former park ranger, 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

Exercise

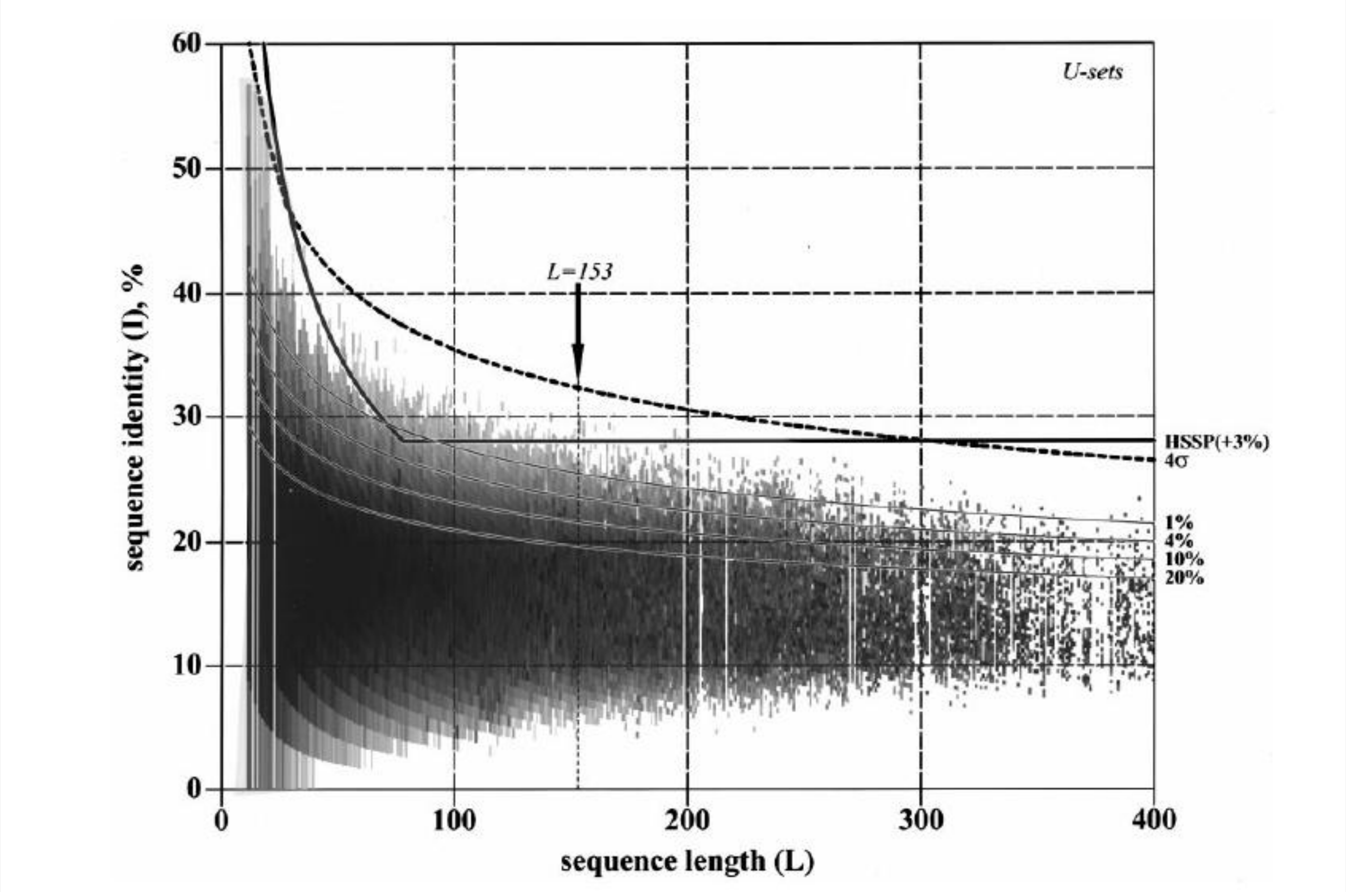
One fourth of all residues in protein sequences occur in regions with biased amino acid composition

What happens when you align protein sequences containing biased amino acid composition?

What should you do about this?

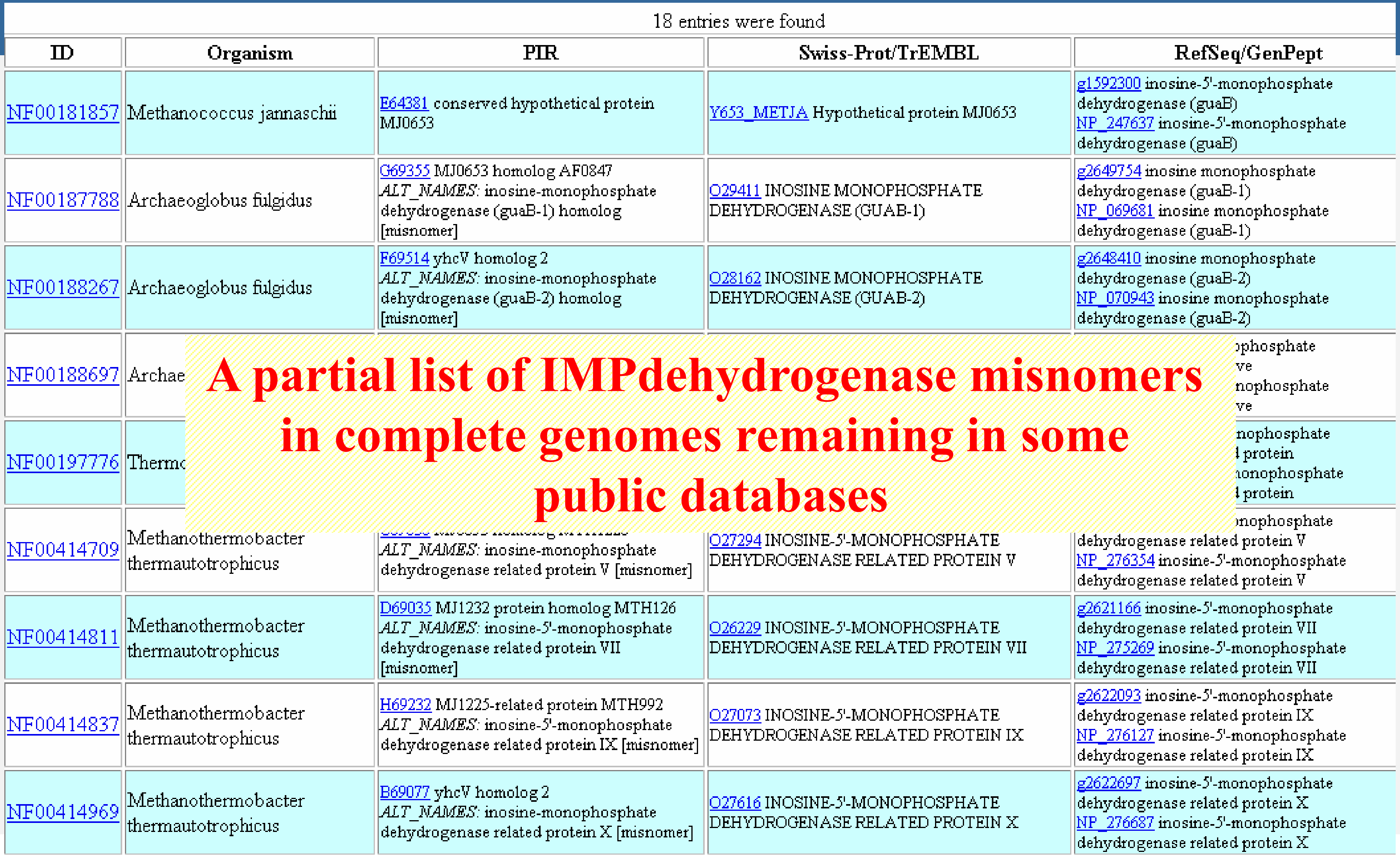
Source: NCBI

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

IMP dehydrogenases (IMPDH)

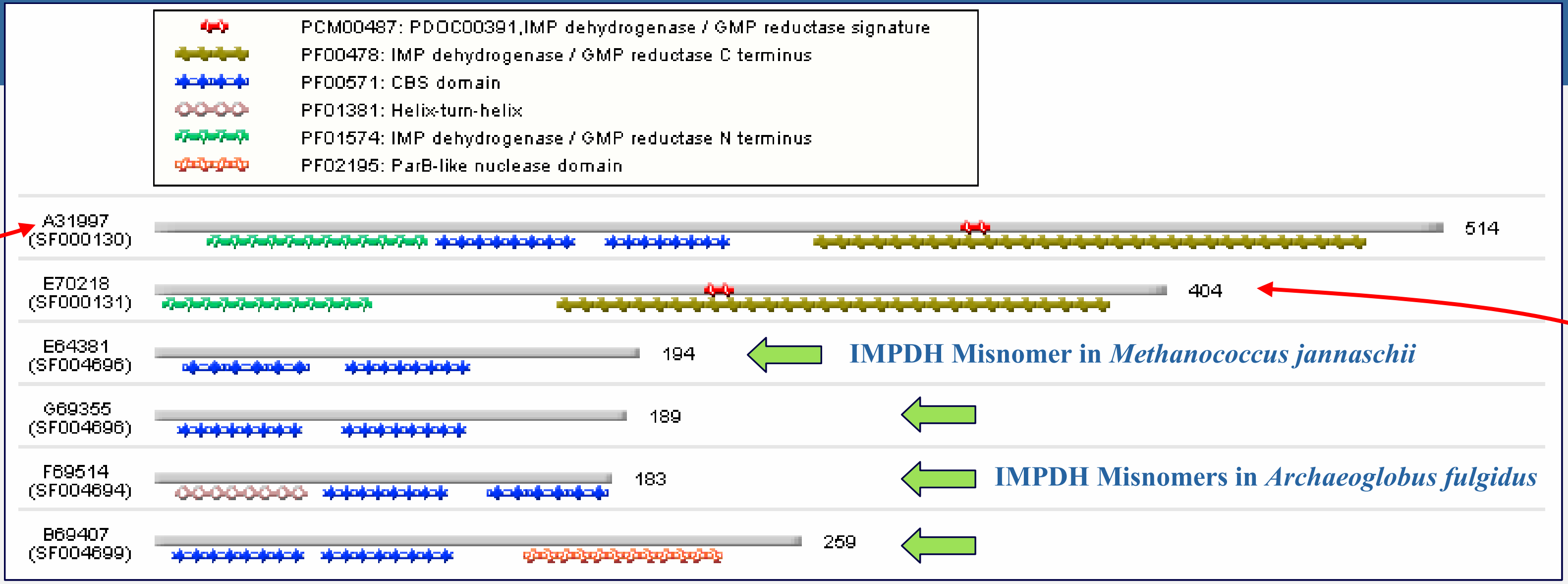


18 entries were found

ID	Organism	PIR	Swiss-Prot/TrEMBL	RefSeq/GenPept
NF00181857	Methanococcus jannaschii	E64381 conserved hypothetical protein MJ0653	Y653_METJA Hypothetical protein MJ0653	g1592300 inosine-5'-monophosphate dehydrogenase (guaB) NP_247637 inosine-5'-monophosphate dehydrogenase (guaB)
NF00187788	Archaeoglobus fulgidus	G69355 MJ0653 homolog AF0847 *ALT_NAMES*: inosine-monophosphate dehydrogenase (guaB-1) homolog [misnomer]	O29411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)	g2649754 inosine monophosphate dehydrogenase (guaB-1) NP_069681 inosine monophosphate dehydrogenase (guaB-1)
NF00188267	Archaeoglobus fulgidus	F69514 yhcV homolog 2 *ALT_NAMES*: inosine-monophosphate dehydrogenase (guaB-2) homolog [misnomer]	O28162 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-2)	g2648410 inosine monophosphate dehydrogenase (guaB-2) NP_070943 inosine monophosphate dehydrogenase (guaB-2)
NF00188697	Archae			ophosphate ve nophosphate ve
NF00197776	Thermo			nophosphate d protein onophosphate d protein
NF00414709	Methanothermobacter thermautotrophicus	[illegible] *ALT_NAMES*: inosine-monophosphate dehydrogenase related protein V [misnomer]	O27294 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN V	nophosphate dehydrogenase related protein V NP_276354 inosine-5'-monophosphate dehydrogenase related protein V
NF00414811	Methanothermobacter thermautotrophicus	D69035 MJ1232 protein homolog MTH126 *ALT_NAMES*: inosine-5'-monophosphate dehydrogenase related protein VII [misnomer]	O26229 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN VII	g2621166 inosine-5'-monophosphate dehydrogenase related protein VII NP_275269 inosine-5'-monophosphate dehydrogenase related protein VII
NF00414837	Methanothermobacter thermautotrophicus	H69232 MJ1225-related protein MTH992 *ALT_NAMES*: inosine-5'-monophosphate dehydrogenase related protein IX [misnomer]	O27073 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN IX	g2622093 inosine-5'-monophosphate dehydrogenase related protein IX NP_276127 inosine-5'-monophosphate dehydrogenase related protein IX
NF00414969	Methanothermobacter thermautotrophicus	B69077 yhcV homolog 2 *ALT_NAMES*: inosine-monophosphate dehydrogenase related protein X [misnomer]	O27616 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN X	g2622697 inosine-5'-monophosphate dehydrogenase related protein X NP_276687 inosine-5'-monophosphate dehydrogenase related protein X

A partial list of IMPdehydrogenase misnomers in complete genomes remaining in some public databases

IMPDH domain structure



PCM00487: PDOC00391,IMP dehydrogenase / GMP reductase signature

PF00478: IMP dehydrogenase / GMP reductase C terminus

PF00571: CBS domain

PF01381: Helix-turn-helix

PF01574: IMP dehydrogenase / GMP reductase N terminus

PF02195: ParB-like nuclease domain

IMPDH Misnomer in *Methanococcus jannaschii*

IMPDH Misnomers in *Archaeoglobus fulgidus*

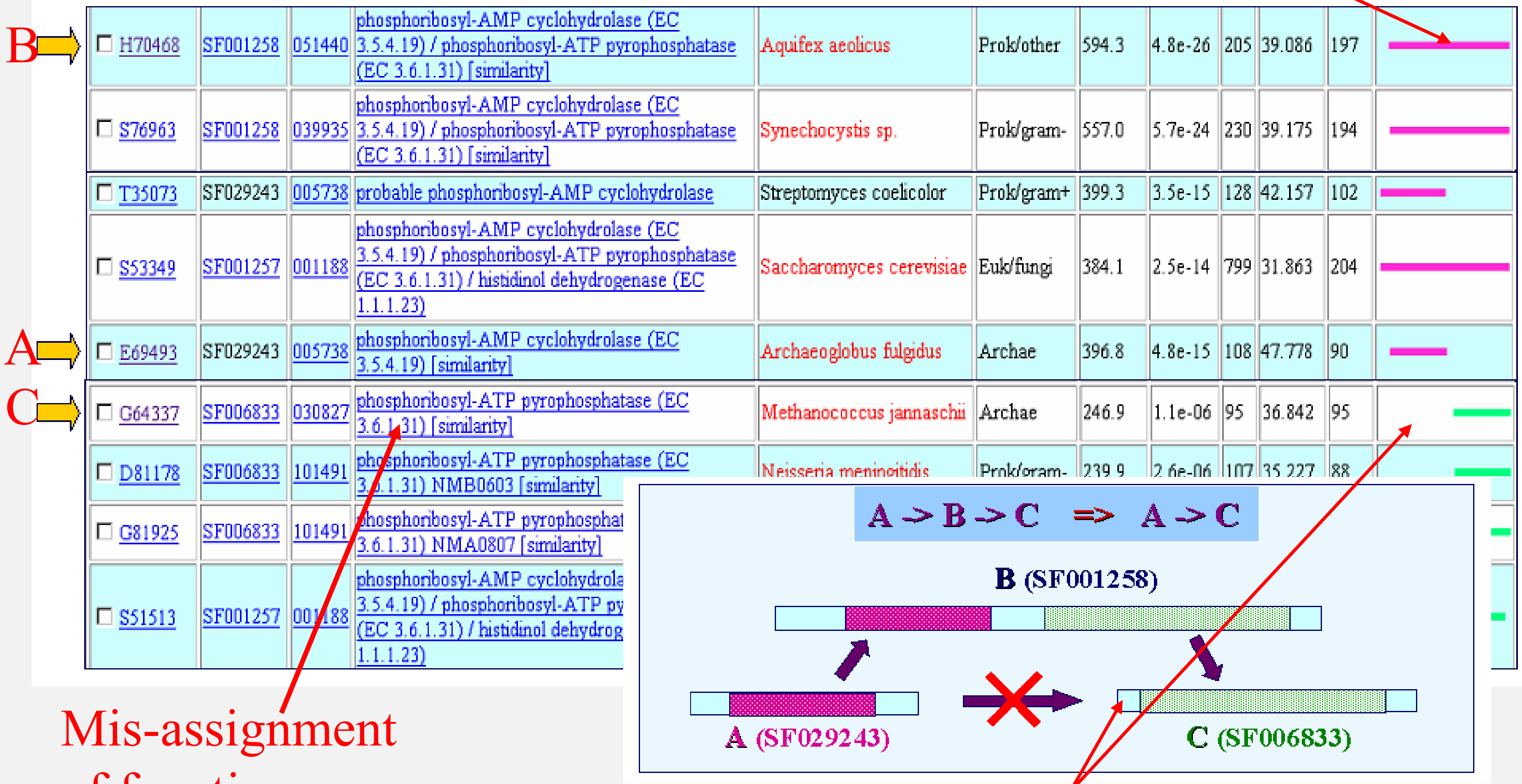
Typical IMPDHs have IMPDH domains as catalytic core and CBS domains

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

Misnomers show similarity only to the CBS domains

Invalid transitive assignment

Root of invalid transitive assignment

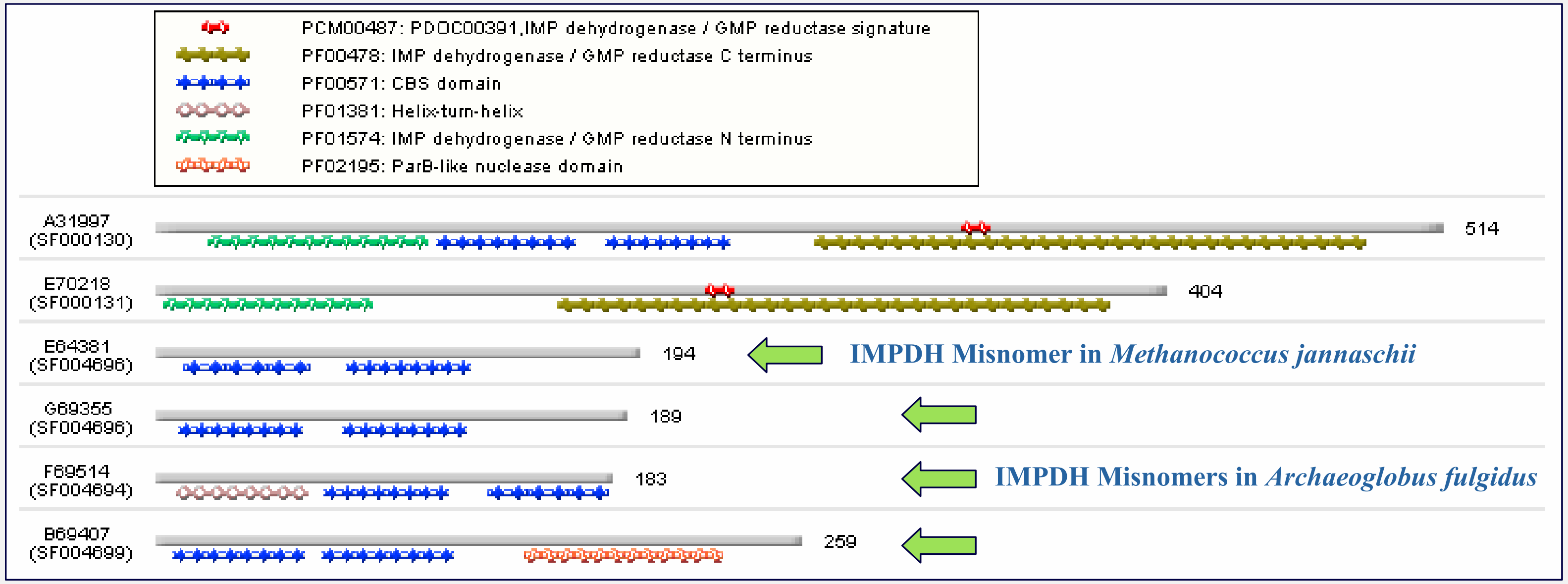


B	□ H70468	SF001258	051440	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Aquifex aeolicus	Prok/other	594.3	4.8e-26	205	39.086	197	▬
	□ S76963	SF001258	039935	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase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


PCM00487: PDOC00391,IMP dehydrogenase / GMP reductase signature
PF00478: IMP dehydrogenase / GMP reductase C terminus
PF00571: CBS domain
PF01381: Helix-turn-helix
PF01574: IMP dehydrogenase / GMP reductase N terminus
PF02195: ParB-like nuclease domain
A31997 (SF000130)
514
E70218 (SF000131)
404
E64381 (SF004696)
194
G69355 (SF004696)
189
F69514 (SF004694)
183
B69407 (SF004699)
259

IMPDH Misnomer in *Methanococcus jannaschii*

IMPDH Misnomers in *Archaeoglobus fulgidus*

Most IMPDHs have 2 IMPDH and 2 CBS domains

Some IMPDH (E70218) lacks CBS domains

IMPDH domain is the emerging pattern

Concluding remarks

Exercise

What have you learned?

Acknowledgements

Some slides on popular sequence alignment tools are based on those given to me by Bin Ma and Dong Xu

Good to read

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

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 2003*, 164-175