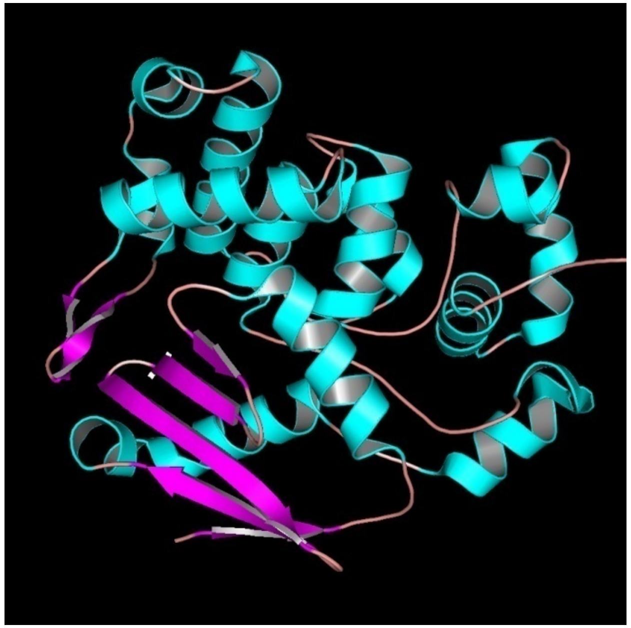
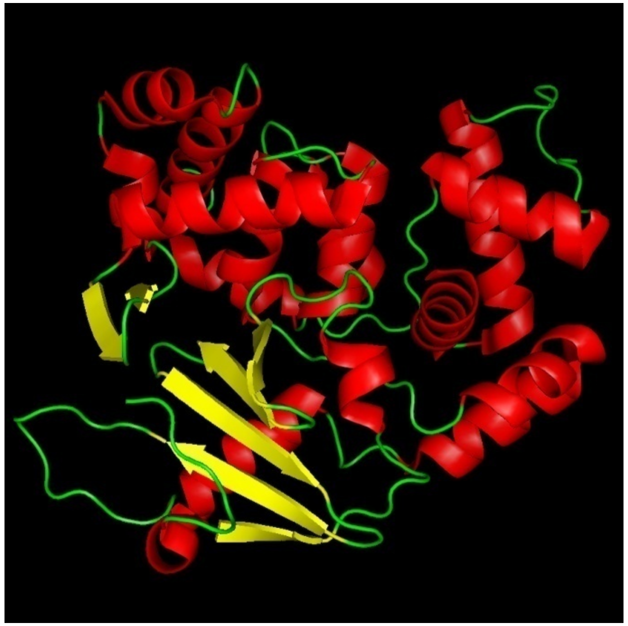
What is similarity and homology?
What is a good match?
How does BLAST work?

Structure and sequence alignment



E.coli AlkA

Hollis *et al.* (2000) *EMBO J.* **19**, 758-766 (PDB ID 1DIZ)



Human OGG1

Source: Bruner *et al.* (2000) *Nature* **403**, 859-866 (PDB ID 1EBM)

```
E.c. AlkA 127 SVAMAAKLTARVAQLYGERLDDFPE--YICFPTPQRLAAADPQA-LKALGMPLKRAEALI 183
              ++|    +  |+ | +| ||    +  |  ||+ | ||  + +| |+ ||+   ||  +
H.s. OGG1 151 NIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGY-RARYVS 209

E.c. AlkA 184 HLANAALE-----GTLPMTIPGDVEQAMKTLQTFPGIGRWTANYFAL              225
                | | ||       |        |+| | |   ||+|   |+   |
H.s. OGG1 210 ASARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICL              256
```

Similarity and homology

Two very important basic concepts:

- **Similarity**: Degree of likeness between two sequences, usually expressed as a percentage of similar (or identical) residues over a given length of the alignment. Can usually be easily calculated.
- **Homology**: Statement about common evolutionary ancestry of two sequences. Can only be true or false. We can rarely be certain about this, it is therefore usually a hypothesis that may be more or less probable.

A high degree if similarity implies a high probability of homology

- If two sequences are very similar, the sequences are usually homologous
- If two sequence are not similar, we don’t know if they are homologous
- If two sequences are not homologous, their sequences are usually not similar (but may be by chance)
- If two sequences are homologous, their sequences may or may not be similar; we don’t know

Sequence similarity and homology

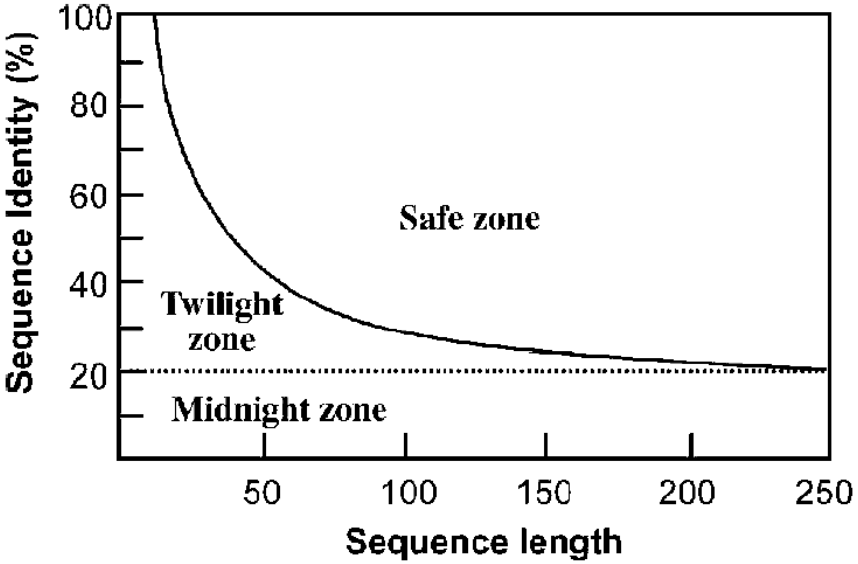


Figure 3.1: The three zones of protein sequence alignments. Two protein sequences can be regarded as homologous if the percentage sequence identity falls in the safe zone. Sequence identity values below the zone boundary, but above 20%, are considered to be in the twilight zone, where homologous relationships are less certain. The region below 20% is the midnight zone, where homologous relationships cannot be reliably determined. (*Source:* Modified from Rost 1999).

Common alignment scoring system

- Substitution score matrix
  - Score for aligning any two residues to each other
  - Identical residues have large positive scores
  - Similar residues have small positive scores
  - Very different residues have large negative scores

BLOSUM62 amino acid substituition score matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

- Gap penalties
  - Penalty for opening a gap in a sequence (Q)
  - Penalty for extending a gap (R)
  - Typical gap function: G = Q + R * L, where L is length of gap
  - Example: Q=11, R=1

```
E.c. AlkA 127 SVAMAAKLTARVAQLYGERLDDFPE--YICFPTPQRLAAADPQA-LKALGMPLKRAEALI 183
              ++|    +  |+ | +| ||    +  |  ||+ | ||  + +| |+ ||+   ||  +
H.s. OGG1 151 NIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGY-RARYVS 209

E.c. AlkA 184 HLANAALE-----GTLPMTIPGDVEQAMKTLQTFPGIGRWTANYFAL              225
                | | ||       |        |+| | |   ||+|   |+   |
H.s. OGG1 210 ASARAILEEQGGLAWLQQLRESSYEEAHKALCILPGVGTKVADCICL              256
```

Amino acid substitution score matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLOSUM62

Significance of alignments

- Even random sequences may reach a high score when aligned optimally, so when is a sequence alignment significant?

- How can we know that sequences are homologous? Homology means that a common ancestor is assumed

- Statistical methods compare the score of a match with the distribution of alignment scores found by aligning random sequences

- The most commonly used indicator of significance:

  E-value = Expect value = expected number of random matches at least as good as this one (with at least this alignment score)

- Some other simple indicators of significance (less accurate):
  - Percentage of identical residues
  - Percentage of similar residues
  - Bit score
  - Raw alignment score

Expect value (E-value)

Expected number of random matches with at least a given alignment score

$$E = K\,M\,N\,e^{-\lambda S}$$

Here,

- S is the raw alignment score
- K and λ are constants that depends on the score matrix and gap penalties used.
- M and N are the lengths of the query and database sequences

Normalized score (bitscore):

$$S' = (\lambda S - \ln K) / \ln 2$$

Interpreting E values

Low E-values indicate high statistical significance.

Rules of thumb:

- $E < 0.05$: probably related (homologous)
- $E < 1$: may be related
- $E >= 1$: no statistical significance, but may be biologically significant anyway

Repeats and low complexity regions

- Repeats and low complexity regions constitute more than one third of the human genome.
- Highly locally biased composition occurs in regions of many proteins and in DNA. E.g. structural proteins in hair.
- Low complexity regions may give rise to high alignment scores – but are usually biologically uninteresting
- They can (and should usually) be masked using programs like RepeatMasker, DUST or SEG before a database search is caried out. The sequence in each region is then replaced by Ns or Xs.

- Examples:
  - interspersed repeats:
    - Short interspersed elements (SINEs)
    - Long interspersed elements (LINEs)
  - simple repeats (microsatellites)
    - usually 1 to 7 nucleotides are repeated a large number of times
    - E.g. …AGAGAGAGAGAGAGAG…
    - E.g. …CCGCCGCCGCCGCCGCCG…
  - low complexity regions,
    - Protein example: PPCDPPPPPKDKKKDDGPP
    - DNA example: AAATAAAAAAAATAAAAAAT

Database search algorithms

- Based on local alignments of query sequence with every database sequence
- Exhaustive / Optimal / Brute-force: Smith-Waterman
- Heuristic: BLAST, FASTA, PARALIGN, ...
- Heuristic algorithms are faster but less accurate

Search performance

Three important performance indicators :

- Sensitivity (Recall)
  - Ability to detect the homologous sequences in the database
  - The fraction of truly homologous sequences found (with a score above a certain threshold) among all homologous sequences
  - True positives / (True positives + False negatives)

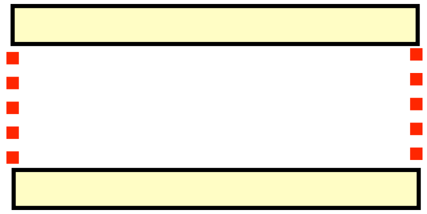
- Precision (PPV)
  - Ability to distinguish between homologous sequences and non-homologous sequences
  - The fraction of truly homologous sequences found (with a score above a certain threshold) among all sequences found
  - True positives / (True positives + False positives)

- Speed

Global and local alignments

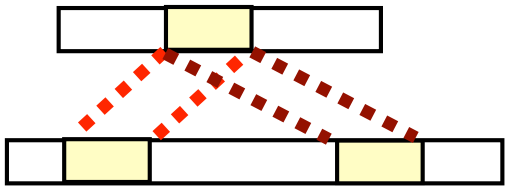
Global alignment:

- Alignment of entire sequences (all symbols)
- May be used when the sequences are of approximately equal length and are expected to be related over their entire length.



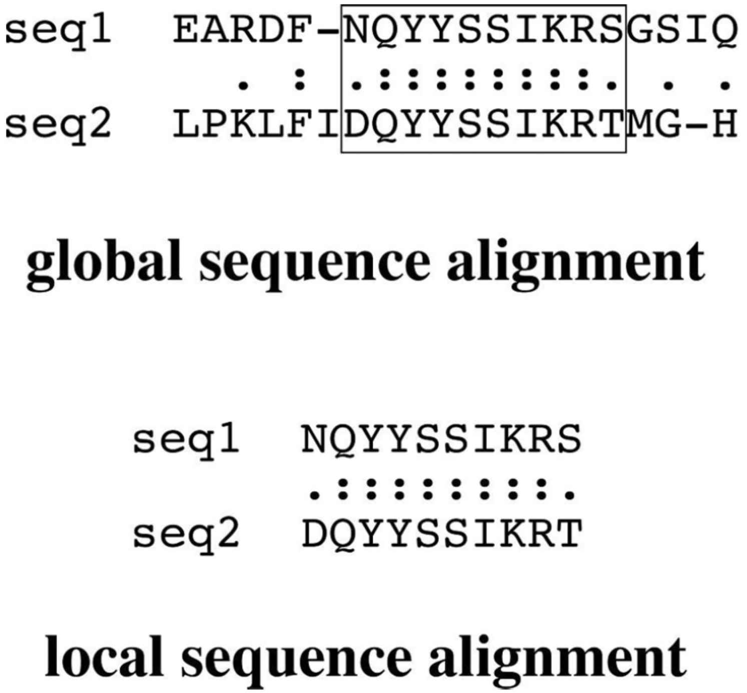
Local alignment:

- Alignment of subsequences from each sequence
- Part of the problem is to identify which parts of the sequences should be included
- Is used when the sequences are of inequal length; and/ or only certain regions in the sequences are assumed to be related (conserved domains).



Global and local alignments

Figure 3.2: An example of pairwise sequence comparison showing the distinction between global and local alignment. The global alignment (*top*) includes all residues of both sequences. The region with the highest similarity is highlighted in a box. The local alignment only includes portions of the two sequences that have the highest regional similarity. In the line between the two sequences, ":" indicates identical residue matches and "." indicates similar residue matches.



```
seq1  EARDF-NQYYSSIKRSGSIQ
        . : .::::::::. . .
seq2  LPKLFIDQYYSSIKRTMG-H
```

global sequence alignment

```
seq1  NQYYSSIKRS
      .::::::::.
seq2  DQYYSSIKRT
```

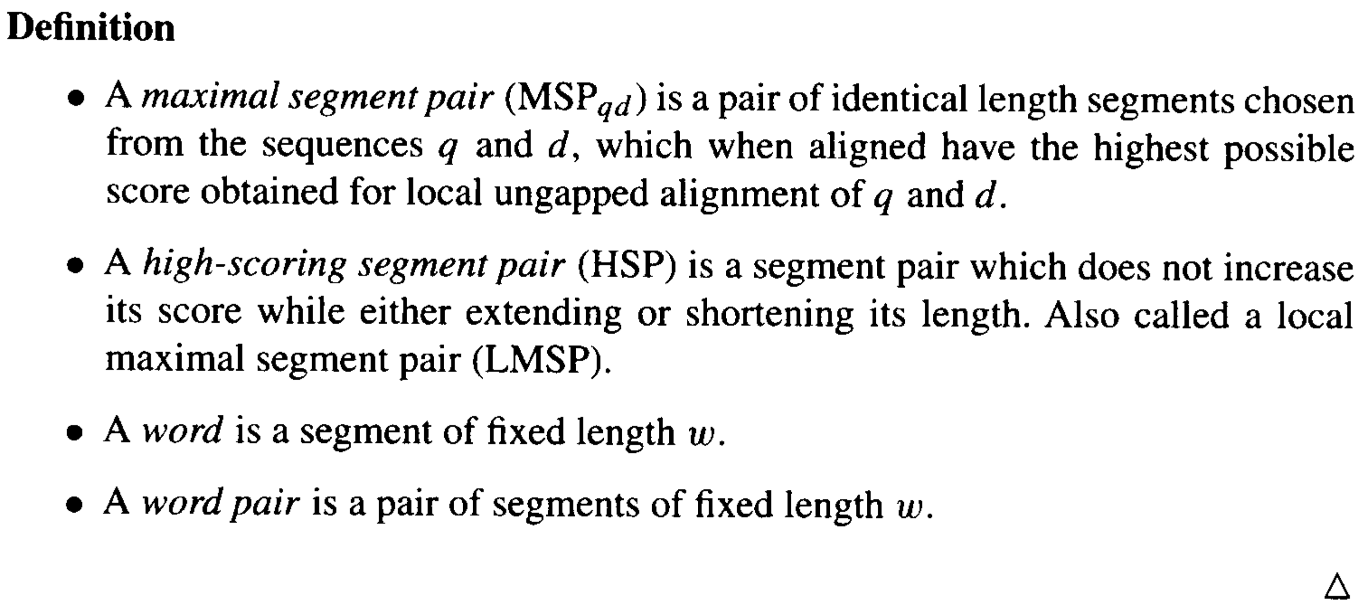
local sequence alignment

BLAST

- BLAST = Basic local alignment search tool
- Very popular, probably most commonly used tool in bioinformatics
- First version in 1990 (no gaps)
- Second version in 1997 (with gaps, + PSI-BLAST etc)
- References
  - Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. J Mol Biol., 215, 403-410.
  - Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389-3402.

BLAST: pre-processing

- BLAST looks for so-called maximal segment pairs (MSPs) with a high score. The goal is to find all MSPs with score at least V.
- Within a MSP with score at least V there is a high probability that there will be a word pair with score at least T. These are called hits.
- Initially BLAST will look for word pairs with score of at least T



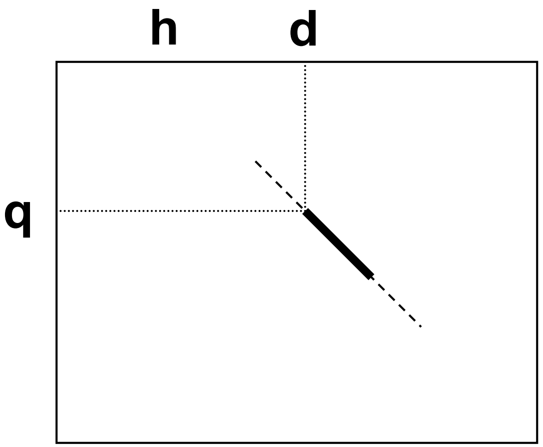
Definition

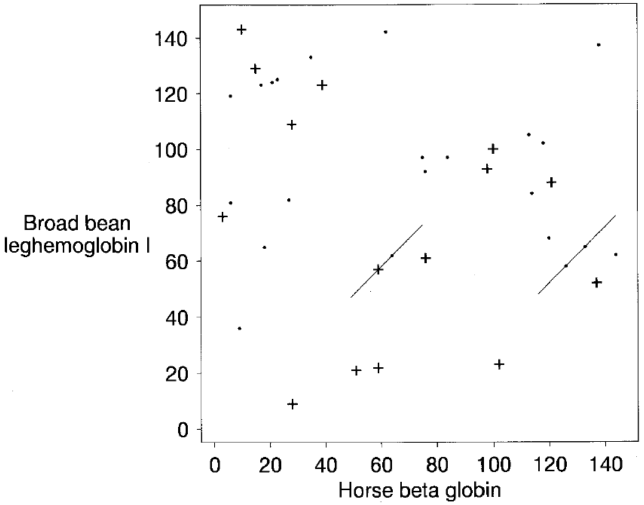
- A *maximal segment pair* (MSP_{qd}) is a pair of identical length segments chosen from the sequences q and d, which when aligned have the highest possible score obtained for local ungapped alignment of q and d.
- A *high-scoring segment pair* (HSP) is a segment pair which does not increase its score while either extending or shortening its length. Also called a local maximal segment pair (LMSP).
- A *word* is a segment of fixed length w.
- A *word pair* is a pair of segments of fixed length w.

△

BLAST for proteins, step 1

- Search through the database sequence and identify the position of all words matching the query sequence
- Keep track of the starting positions of the words, both in the query sequence (q) and in the database sequence (p)
- Compute the diagonal number h = d - q

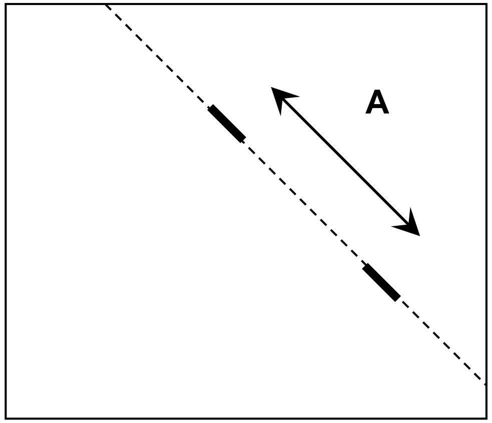




BLAST for proteins, step 2

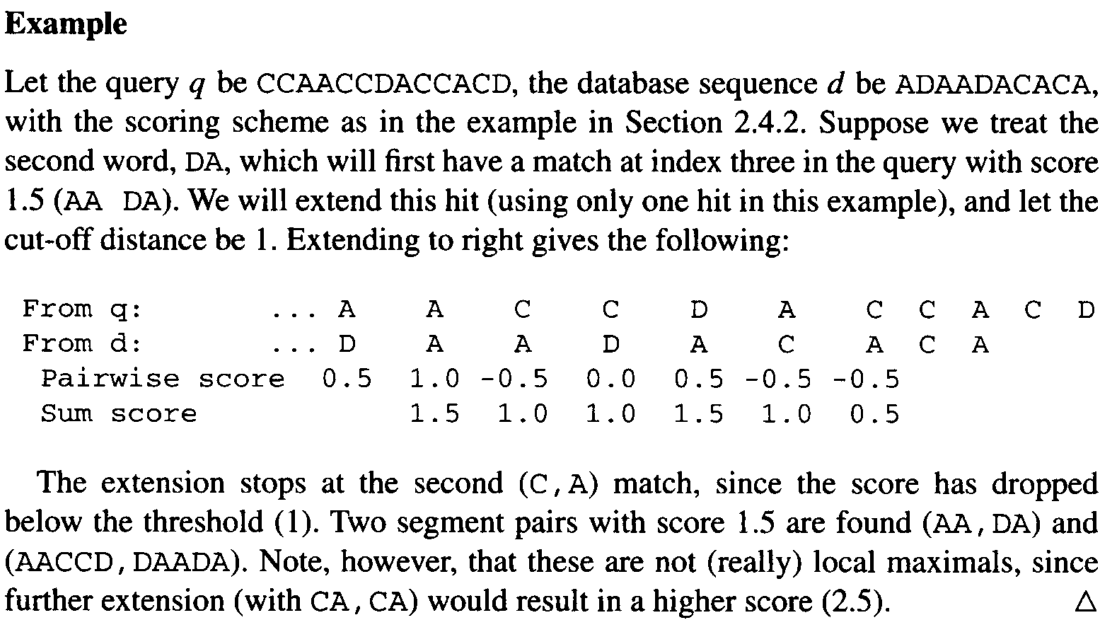
- Keep hits if there are two hits on the same diagonal within a maximal distance A (typical 40)

```
              d
       L U K A L W Y A R . . .
  i\j  1 2 3 4 5 6 7 8 9
  1  E
  2  A       *
  3  L         *
q 4  C
  5  K     *
  6  A       *h=-2   *
  7  R                 *h=2
  8  V
  9  A               *
 10  R                 *h=-1
     .
```



BLAST for proteins, step 3

- Expand the hits into HSPs in both directions (no gaps) by adding score values from the substitution score matrix.
- In each direction, stop when the score decreases more than a threshold X from the highest score seen so far.

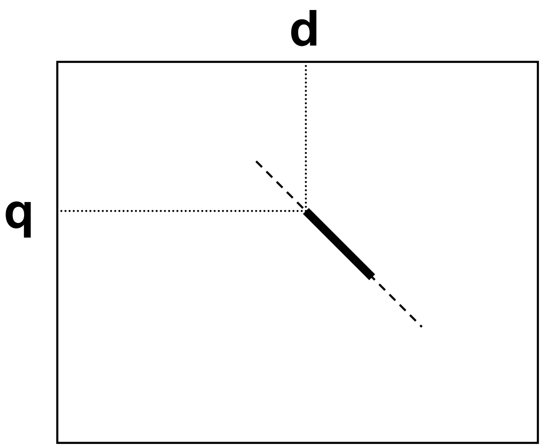


Example

Let the query *q* be CCAACCDACCACD, the database sequence *d* be ADAADACACA, with the scoring scheme as in the example in Section 2.4.2. Suppose we treat the second word, DA, which will first have a match at index three in the query with score 1.5 (AA DA). We will extend this hit (using only one hit in this example), and let the cut-off distance be 1. Extending to right gives the following:

```
From q:        ... A    A    C    C    D    A    C  C  A  C  D
From d:        ... D    A    A    D    A    C    A  C  A
 Pairwise score   0.5  1.0 -0.5  0.0  0.5 -0.5 -0.5
 Sum score             1.5  1.0  1.0  1.5  1.0  0.5
```

The extension stops at the second (C,A) match, since the score has dropped below the threshold (1). Two segment pairs with score 1.5 are found (AA,DA) and (AACCD,DAADA). Note, however, that these are not (really) local maximals, since further extension (with CA,CA) would result in a higher score (2.5). △

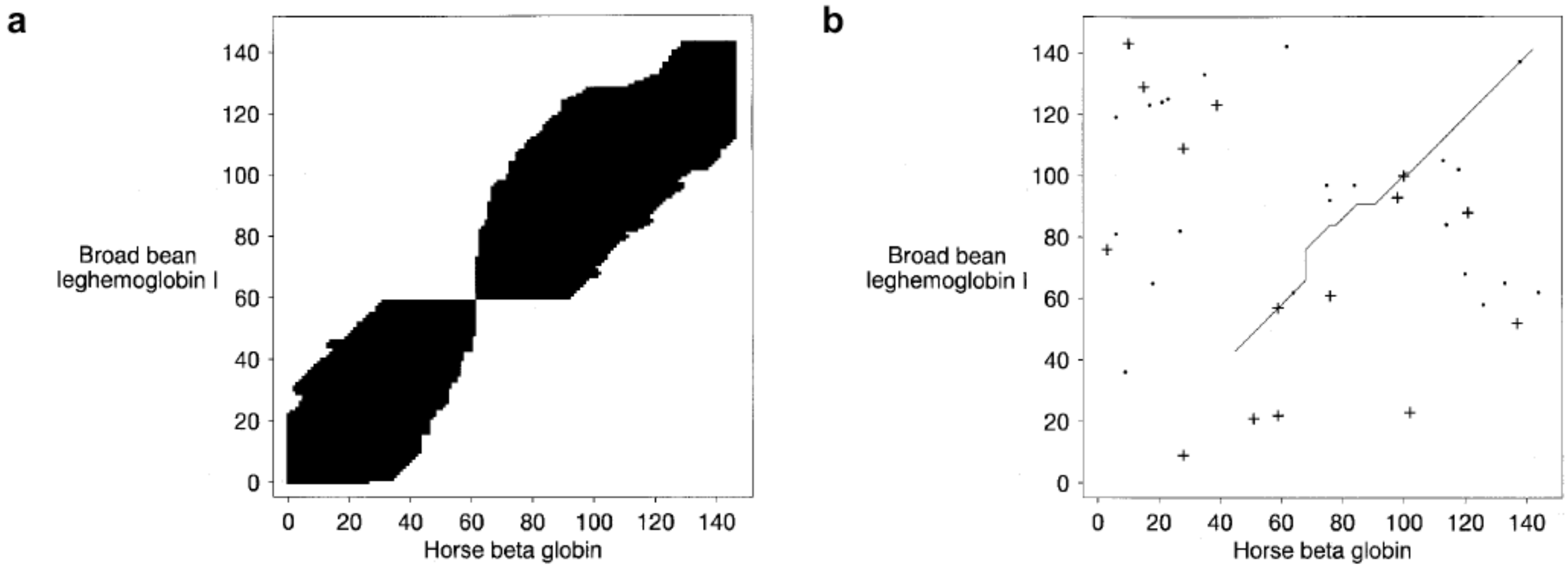


BLAST for proteins, step 4

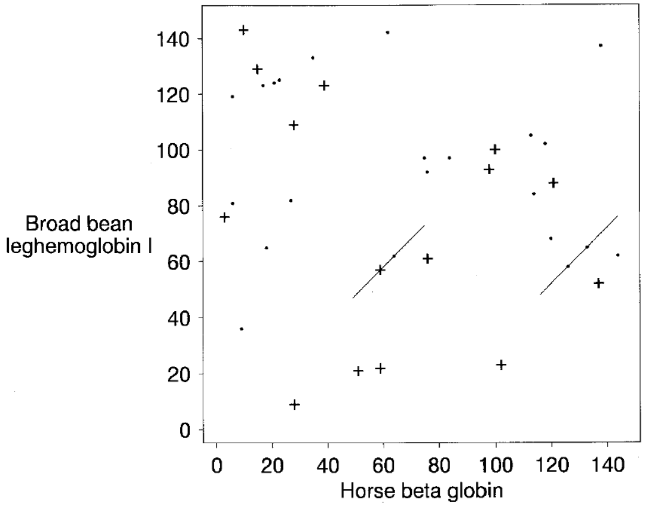
- Keep HSPs with score of at least S_g.
- The threshold is set to corresponds to approximately 2% of the database sequences on average

BLAST for proteins, step 5

- Recalculate the score again by computing an optimal local alignment score within an area around a ”seed” in the middle of the HSP.
- The area is limited by the H-value in the DP-matrix not dropping more than a certain value (X_g) below the current optimal alignment score



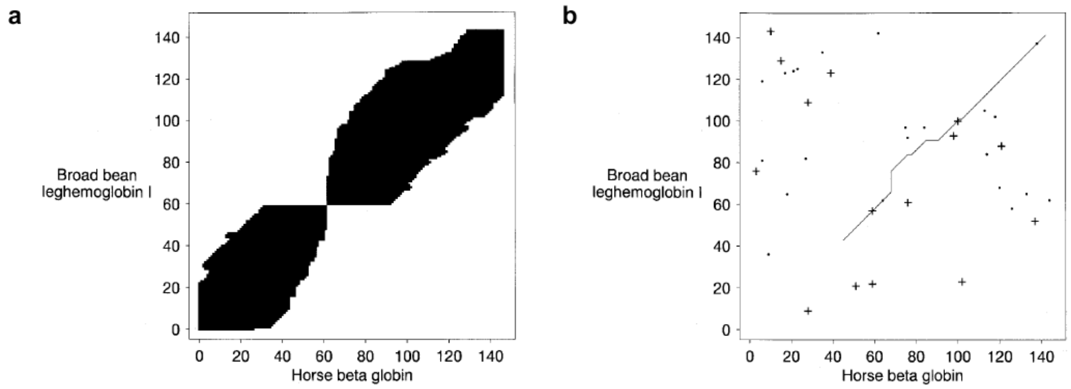
BLAST example



BLAST hits in the alignment

+ Hits with score >= 13

• Hits with score >= 11



a) Areas explored by BLAST during final alignment

b) Graph of the alignment

```
Leghemoglobin  43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------  90
                  F  L +   V+ +PK+ AH +KV          L + GE V  LD   G+
Beta globin    45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE  90


Leghemoglobin  91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                  +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
Beta globin    91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

Alignment created by BLAST

Differences between nucleotide and protein searches

- The databases are often larger (e.g. several complete eukaryote genomes)
- The required sensitivity is usually lower (except when looking for ncRNA)
- Often we would like to find almost identical matches, allowing only a few mismatches or small gaps due to sequencing errors or a few mutations (polymorphisms)
- We have only four symbols: a, c, g and t
- We usually do not use a scoring matrix, we just use:
  - one single score for matches (e.g. +5)
  - one single penalty for mismatches (e.g. -4)
  - a gap penalty (e.g. 12-4k)

Typical usage of nucleotide searches

- Identify the genomic location of an mRNA, a cDNA, an exon or an EST (from the same species), i.e. mapping part of a transcript to the genome sequence
- Identify similar (corresponding) genomic regions in relatively closely related species (e.g. mouse and human genomes) (synteny)

Other examples:

- Identify homologous non-protein coding regions (e.g ribosomal RNA) (often requires more sensitivity)

BLASTN and MegaBLAST

BLASTN

- Word length is W=11 by default
- Only identical words considered hits

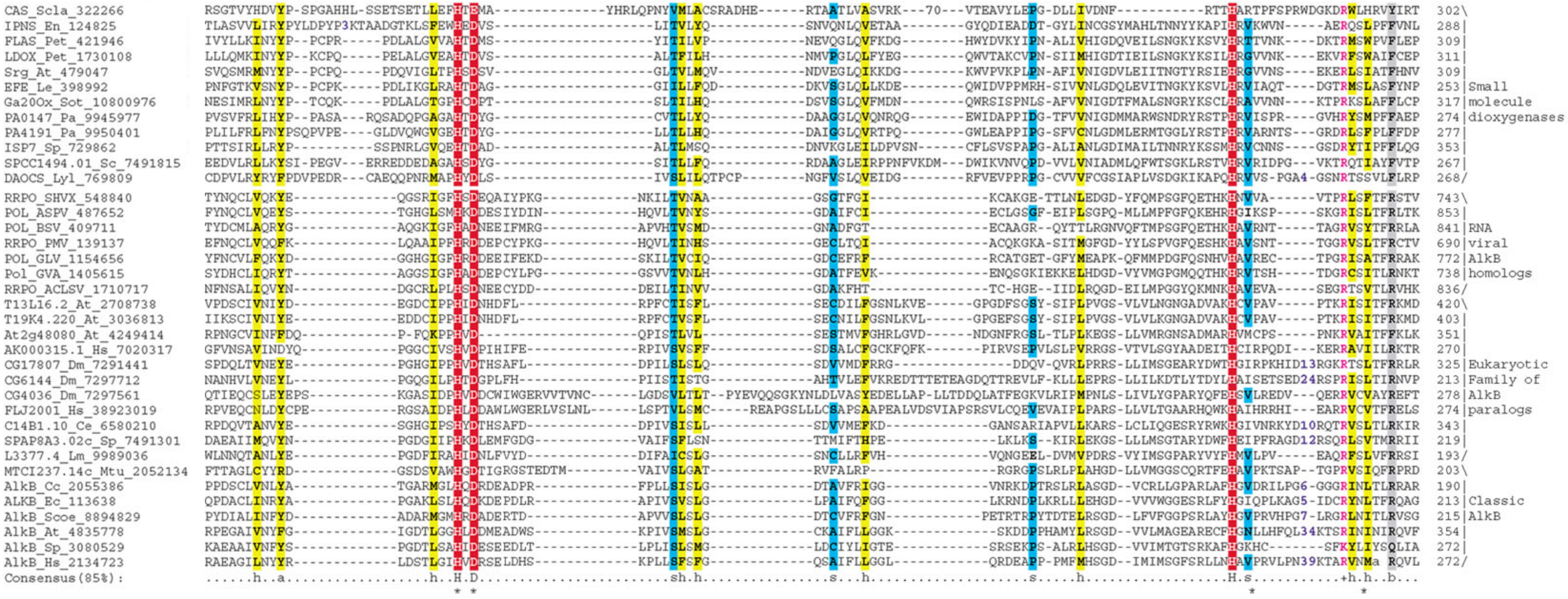
MegaBLAST

- Similar to BLASTN
- Optimized for longer sequences and almost perfect matches
- Uses default word length W=28
- Requires 28 consecutive matching nucleotides between the query and a database sequence
- Much faster than BLASTN, but reduced sensitivity
- Reference:
  Zhang Z, Schwartz S, Wagner L, Miller W (2000)
  A greedy algorithm for aligning DNA sequences.
  J Comput Biol., 7 (1-2), 203-14.

What is PSI-BLAST?

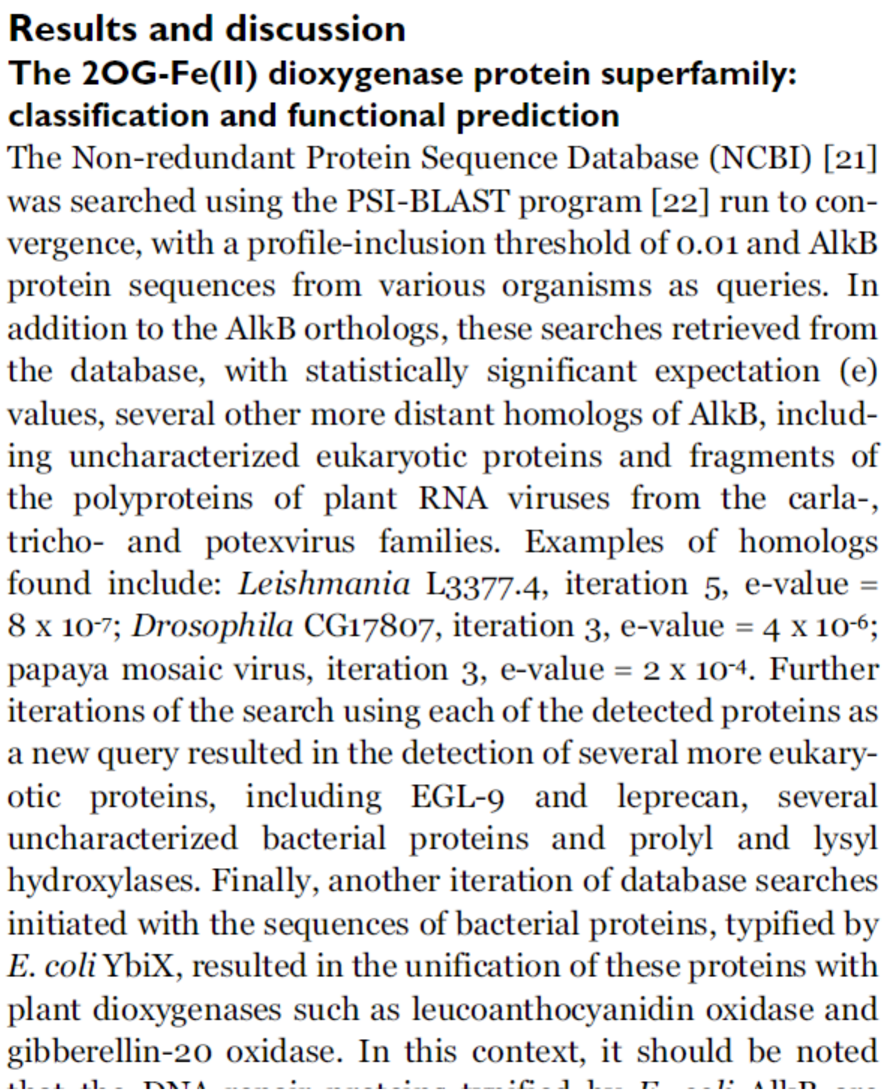
Back to the example...

How are all these sequences found? Ordinary BLAST is not enough...



```
CAS_Scla_322266         RSGTVYHDVYP-SPGAHHL-SSETSETLLEFHTEMA-------------YHRLQPNYVMLACSRADHE------RTAATLVASVRK---70---VTEAVYLEPG-DLLIVDNF----------RTTHARTPFSPRWDGKDRWLHRVYIRT  302\
IPNS_En_124825          TLASVVLIRYPYLDPYP3KTAADGTKLSFEWHEDVS------------------LITVLYQ-----------SNVQNLQVETAA--------GYQDIEADDT-GYLINCGSYMAHLTNNYYKAPIHRVKWVN-----AERQSLPFFVNL  288|
FLAS_Pet_421946         IVYLLKINYYP--PCPR-----PDLALGVVAHTDMS------------------YITILVP-----------NEVQGLQVFKDG--------HWYDVKYIPN-ALIVHIGDQVEILSNGKYKSVYHRTTVNK----DKTRMSWPVFLEP  309|
LDOX_Pet_1730108        LLLQMKINYYP--KCPQ-----PELALGVEAHTDVS------------------ALTFILH-----------NMVPGLQLFYEG--------QWVTAKCVPN-SIIMHIGDTIEILSNGKYKSILHRGVVNK----EKVRFSWAIFCEP  311|
Srg_At_479047           SVQSMRMNYYP--PCPQ-----PDQVIGLTPHSDSV------------------GLTVLMQV----------NDVEGLQIKKDG--------KWVPVKPLPN-AFIVNIGDVLEIITNGTYRSIEHRGVVNS----EKERLSIATFHNV  309|
EFE_Le_398992           PNFGTKVSNYP--PCPK-----PDLIKGLRAHTDAG------------------GIILLFQD----------DKVSGLQLLKDE--------QWIDVPPMRH-SIVVNLGDQLEVITNGKYKSVLHRVIAQT----DGTRMSLASFYNP  253|Small
Ga20Ox_Sot_10800976     NESIMRLNYYP--TCQK-----PDLALGTGPHCDPT------------------SLTILHQ-----------DSVSGLQVFMDN--------QWRSISPNLS-AFVVNIGDTFMALSNGRYKSCLHRAVVNN----KTPRKSLAFFLCP  317|molecule
PA0147_Pa_9945977       PVSVFRLIHYP--PASA---RQSADQPGAGAHTDYG------------------CVTLLYQ-----------DAAGGLQVQNRQG-------EWIDAPPIDG-TFVVNIGDMMARWSNDRYRSTPHRVISPR----GVHRYSMPFFAEP  274|dioxygenases
PA4191_Pa_9950401       PLILFRLFNYPSQPVPE----GLDVQWGVGEHTDYG------------------LLTLLHQ-----------DAIGGLQVRTPQ--------GWLEAPPIPG-SFVCNLGDMLERMTGGLYRSTPHRVARNTS---GRDRLSFPLFFDP  277|
ISP7_Sp_729862          PTTSIRLLRYP----------SSPNRLGVQEHTDAD------------------ALTLMSQ-----------DNVKGLEILDPVSN------CFLSVSPAPG-ALIANLGDIMAILTNNRYKSSMHRVCNNS----GSDRYTIPFFLQG  353|
SPCC1494.01_Sc_7491815  EEDVLRLLKYSI-PEGV---ERREDDEDAGAHSDYG------------------SITLLFQ-----------RDAAGLEIRPPNFVKDM---DWIKVNVQPD-VVLVNIADMLQFWTSGKLRSTVHRVRIDPG---VKTRQTIAYFVTP  267|
DAOCS_Lyl_769809        CDPVLRYRYFPDVPEDR---CAEQQPNRMAPHYDLS------------------IVSLILQTPCP--------NGFVSLQVEIDG--------RFVEVPPRPG-CVVVFCGSIAPLVSDGKIKAPQHRVVS-PGA4-GSNRTSSVLFLRP  268/
RRPO_SHVX_548840        TYNQCLVQKYE-------------QGSRIGFHSDEQAIYPKG-----------NKILTVNAA-------------GSGTFGI----------------KCAKGE-TTLNLEDGD-YFQMPSGFQETHKHNVVA------VTPRLSFTFRSTV  743\
POL_ASPV_487652         FYNQCLVQEYS-------------TGHGLSMHKDDESIYDIN-----------HQVLTVNYS-------------GDAIFCI----------------ECLGSGF-EIPLSGPQ-MLLMPFGFQKEHRHGIKSP-----SKGRISLTFRLTK  853|
POL_BSV_409711          TYDCMLAQRYG-------------AQGKIGFHADNEEIFMRG-----------APVHTVSMD-------------GNADFGT----------------ECAAGR--QYTTLRGNVQFTMPSGFQETHKHAVRNT-----TAGRVSYTFRRLA  841|RNA
RRPO_PMV_139137         EFNQCLVQQFK-------------LQAAIPFHRDDEPCYPKG-----------HQVLTINHS-------------GECLTQI----------------ACQKGKA-SITMGFGD-YYLSPVGFQESHKHAVSNT-----TGGRVSLTFRCTV  690|viral
POL_GLV_1154656         YFNCVLFQKYD-------------GGHGIGFHRDDEEIFEKD-----------SKILTVCIQ-------------GDCEFRF----------------RCATGET-GFYMEAPK-QFMMPDGFQSNHVHAVREC-----TPGRISATFRRAK  772|AlkB
Pol_GVA_1405615         SYDHCLIQRYT-------------AGGSIGFHADDEPCYLPG-----------GSVVTVNLH-------------GDATFEVK---------------ENQSGKIEKKELHDGD-VYVMGPGMQQTHKHRVTSH-----TDGRCSITLRNKT  738|homologs
RRPO_ACLSV_1710717      NFNSALIQVYN-------------DGCRLPLHSDNEECYDD------------DEILTINVV-------------GDAKFHT----------------TC-HGE--IIDLRQGD-EILMPGGYQKMNKHAVEVA-----SEGRTSVTLRVHK  836/
T13L16.2_At_2708738     VPDSCIVNIYD-------------EGDCIPPHIDNHDFL--------------RPFCTISFL-------------SECDILFGSNLKVE------GPGDFSGSY-SIPLPVGS-VLVLNGNGADVAKHCVPAV-----PTKRISITFRKMD  420\
T19K4.220_At_3036813    IIKSCIVNIYE-------------EDDCIPPHIDNHDFL--------------RPFCTVSFL-------------SECNILFGSNLKVL------GPGEFSGSY-SIPLPVGS-VLVLKGNGADVAKHCVPAV-----PTKRISITFRKMD  403|
At2g48080_At_4249414    RPNGCVINFFDQ------------P-FQKPPHVD-------------------QPISTLVL-------------SESTMVFGHRLGVD------NDGNFRGSL-TLPLKEGS-LLVMRGNSADMARHVMCPS-----PNKRVAITFFKLK  351|
AK000315.1_Hs_7020317   GFVNSAVINDYQ------------PGGCIVSHVDPIHIFE-------------RPIVSVSFF-------------SDSALCFGCKFQFK-------PIRVSEPVLSLPVRRGS-VTVLSGYAADEITHCIRPQDI---KERRAVIILRKTR  270|
CG17807_Dm_7291441      SPDQLTVNEYE-------------PGHGIPPHVDTHSAFL-------------DPILSLSLQ-------------SDVVMDFRRG---------------DDQV-QVRLPRRS-LLIMSGEARYDWTHGIRPKHID13RGKRTSLTFRRLR  325|Eukaryotic
CG6144_Dm_7297712       NANHVLVNEYL-------------PGQGILPHTDGPLFH--------------PIISTISTG-------------AHTVLEFVKREDTTTETEAGDQTTREVLF-KLLLEPRS-LLILKDTLYTDYLHAISETSED24RSPRISLTIRNVP  213|Family of
CG4036_Dm_7297561       QTIEQCSLEYEPS-----------KGASIDPHVDDCWIWGERVVTVNC------LGDSVLTLT--PYEVQQSGKYNLDLVASYEDELLAP-LLTDDQLATFEGKVLRIPMPNLS-LIVLYGPARYQFEHSVLREDV---QERRVCVAYREFT  278|AlkB
FLJ2001_Hs_38923019     RPVEQCNLDYCPE-----------RGSAIDPHLDDAWLWGERLVSLNL------LSPTVLSMC-----REAPGSLLLCSAPSAAPEALVDSVIAPSRSVLCQEVEVAIPLPARS-LLVLTGAARHQWKHAIHRRHI---EARRVCVTFRELS  274|paralogs
C14B1.10_Ce_6580210     RPDQVTANVYE-------------SGHGIPSHYDTHSAFD-------------DPIVSISLL-------------SDVVMEFKD-------------GANSARIAPVLLKARS-LCLIQGESRYRWKHGIVNRKYD10RQTRVSLTLRKIR  343|
SPAP8A3.02c_Sp_7491301  DAEAIIMQVYN-------------PGDGIIPHKDLEMFGDG------------VAIFSFLSN-------------TTMIFTHPE---------------LKLKS--KIRLEKGS-LLLMSGTARYDWFHEIPFRAGD12RSQRLSVTMRRII  219|
L3377.4_Lm_9989036      WLNNQTANLYE-------------PGDFIRAHIDNLFVYD-------------DIFAICSLG-------------SNCLLRFVH--------------VQNGEEL-DVMVPDRS-VYIMSGPARYVYFHMVLPV-----EAQRFSLVFRRSI  193/
MTCI237.14c_Mtu_2052134 FTTAGLCYYRD-------------GSDSVAWHGDTIGRGSTEDTM--------VAIVSLGAT-------------RVFALRP----------------RGRGPSLRLPLAHGD-LLVMGGSCQRTFEHAVPKTSAP--TGPRVSIQFRPRD  203\
AlkB_Cc_2055386         PPDSCLVNLYA-------------TGARMGLHQDRDEADPR------------FPLLSISLG-------------DTAVFRIGG--------------VNRKDPTRSLRLASGD--VCRLLGPARLAFHGVDRILPG6-GGGRINLTLRRAR  190|
ALKB_Ec_113638          QPDACLINRYA-------------PGAKLSLHQDKDEPDLR------------APIVSVSLG-------------LPAIFQFGG--------------LKRNDPLKRLLLEHGD--VVVWGGESRLFYHGIQPLKAG5-IDCRYNLTFRQAG  213|Classic
AlkB_Scoe_8894829       PYDIALINFYD-------------ADARMGMHRDADERTD-------------APVVSLSLG-------------DTCVFRFGN------------PETRTRPYTDTELRSGD--LFVFGGPSRLAYHGVPRVHPG7-LRGRLNITLRVSG  215|AlkB
AlkB_At_4835778         RPEGAIVNYFG-------------IGDTLGGHLDDMEADWS------------KPIVSMSLG-------------CKAIFLLGGK-------------SKDDPPHAMYLRSGD--VVLMAGEARECFHGNLLHFQL34KTSRININIRQVF  354|
AlkB_Sp_3080529         KAEAAIVNFYS-------------PGDTLSAHIDESEEDLT------------LPLISLSMG-------------LDCIYLIGTE-------------SRSEKPS-ALRLHSGD--VVIMTGTSRKAFHGKHC-------SFKYLIYSQLIA  272|
AlkB_Hs_2134723         RAEAGILNYYR-------------LDSTLGIHVDRSELDHS------------KPLLSFSFG-------------QSAIFLLGGL-------------QRDEAPP-PMFMHSGD--IMIMSGFSRLLNHAVPRVLPN39KTARVNMa RQVL  272/
Consensus(85%):         ......h..a.................h..H.D......................sh.h...............s...h..................s.....h...............H.s...........+h.h..b...
                                                     * *                                                                                   *           *
```

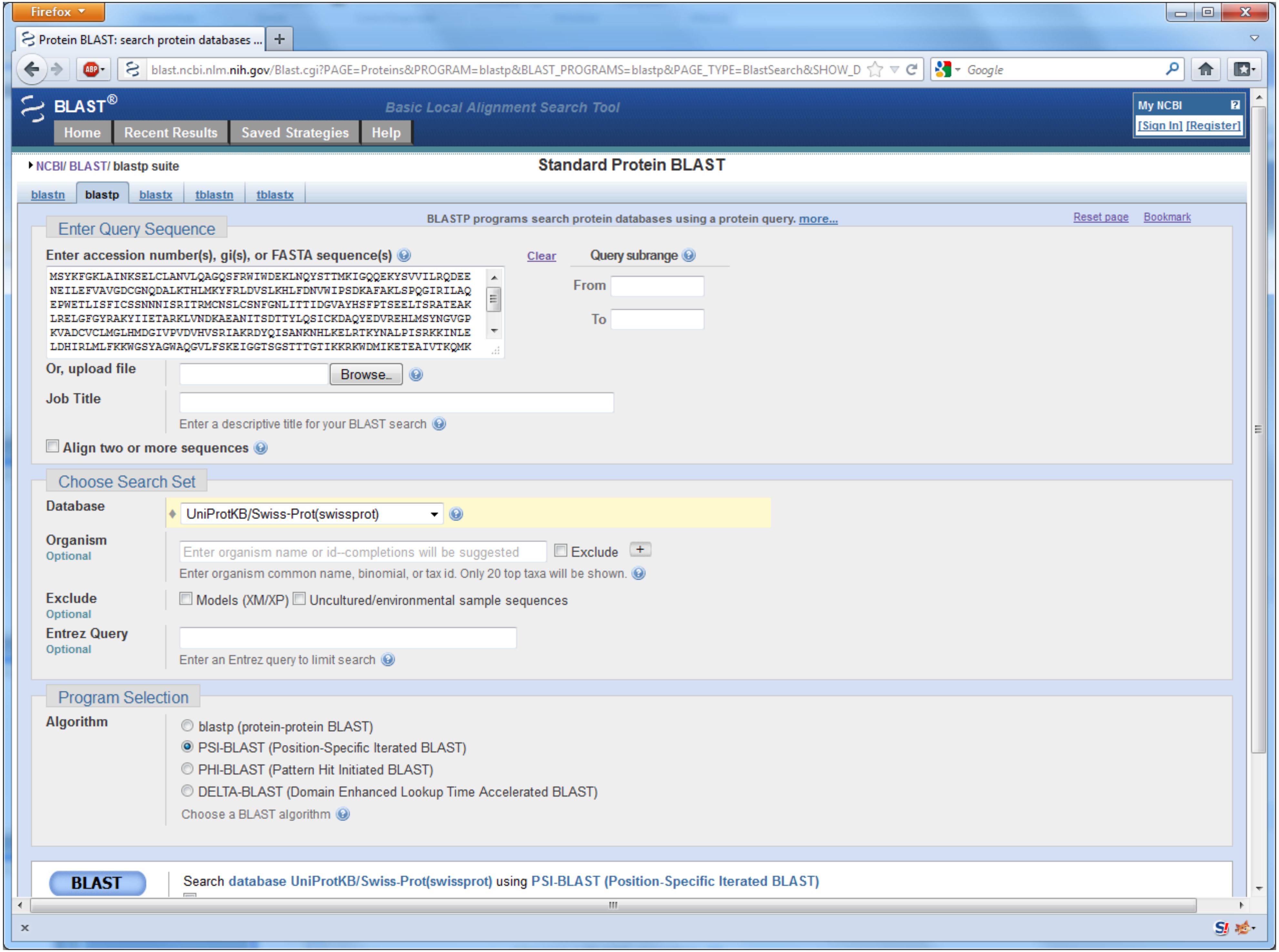
Excerpt from the AlkB paper



Results and discussion

The 2OG-Fe(II) dioxygenase protein superfamily: classification and functional prediction

The Non-redundant Protein Sequence Database (NCBI) [21] was searched using the PSI-BLAST program [22] run to convergence, with a profile-inclusion threshold of 0.01 and AlkB protein sequences from various organisms as queries. In addition to the AlkB orthologs, these searches retrieved from the database, with statistically significant expectation (e) values, several other more distant homologs of AlkB, including uncharacterized eukaryotic proteins and fragments of the polyproteins of plant RNA viruses from the carla-, tricho- and potexvirus families. Examples of homologs found include: *Leishmania* L3377.4, iteration 5, e-value = 8 x 10^{-7}; *Drosophila* CG17807, iteration 3, e-value = 4 x 10^{-6}; papaya mosaic virus, iteration 3, e-value = 2 x 10^{-4}. Further iterations of the search using each of the detected proteins as a new query resulted in the detection of several more eukaryotic proteins, including EGL-9 and leprecan, several uncharacterized bacterial proteins and prolyl and lysyl hydroxylases. Finally, another iteration of database searches initiated with the sequences of bacterial proteins, typified by *E. coli* YbiX, resulted in the unification of these proteins with plant dioxygenases such as leucoanthocyanidin oxidase and gibberellin-20 oxidase. In this context, it should be noted that the DNA repair proteins typified by *E. coli* AlkB are



Firefox
Protein BLAST: search protein databases ...
blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_D
Google
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite
blastn
blastp
blastx
tblastn
tblastx
Reset page
Bookmark
Clear
Browse...
BLAST

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
MSYKFGKLAINKSELCLANVLQAGQSFRWIWDEKLNQYSTTMKIGQQEKYSVVILRQDEE
NEILEFVAVGDCGNQDALKTHLMKYFRLDVSLKHLFDNVWIPSDKAFAKLSPQGIRILAQ
EPWETLISFICSSNNNISRITRMCNSLCSNFGNLITTIDGVAYHSFPTSEELTSRATEAK
LRELGFGYRAKYIIETARKLVNDKAEANITSDTTYLQSICKDAQYEDVREHLMSYNGVGP
KVADCVCLMGLHMDGIVPVDVHVSRIAKRDYQISANKNHLKELRTKYNALPISRKKINLE
LDHIRLMLFKKWGSYAGWAQGVLFSKEIGGTSGSTTTGTIKKRKWDMIKETEAIVTKQMK
```

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: UniProtKB/Swiss-Prot(swissprot)

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query
Optional

Enter an Entrez query to limit search

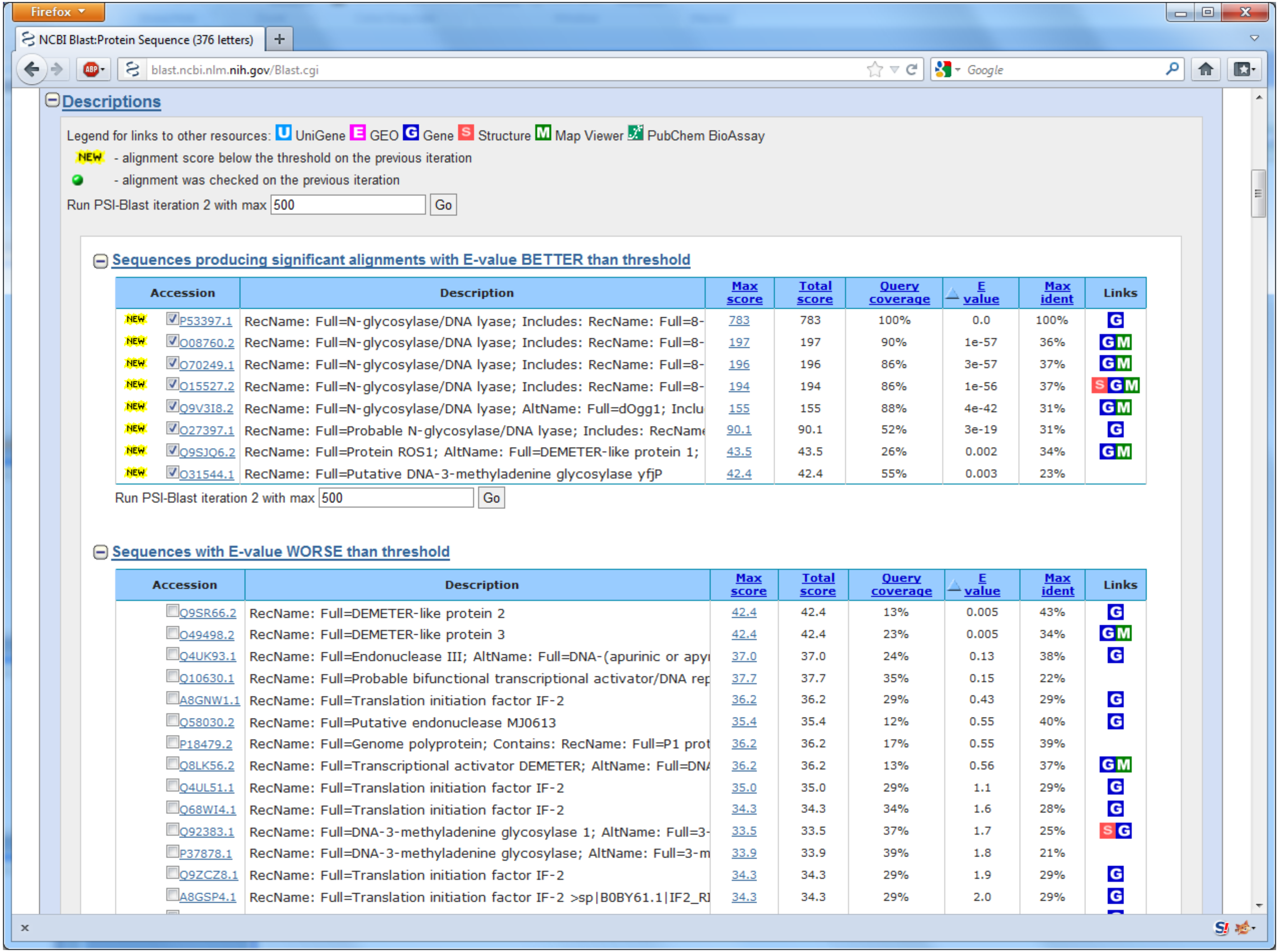
Program Selection

Algorithm

○ blastp (protein-protein BLAST)
● PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Search database UniProtKB/Swiss-Prot(swissprot) using PSI-BLAST (Position-Specific Iterated BLAST)


Firefox
NCBI Blast:Protein Sequence (376 letters)
blast.ncbi.nlm.nih.gov/Blast.cgi
Google

Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

NEW - alignment score below the threshold on the previous iteration

● - alignment was checked on the previous iteration

Run PSI-Blast iteration 2 with max 500 Go

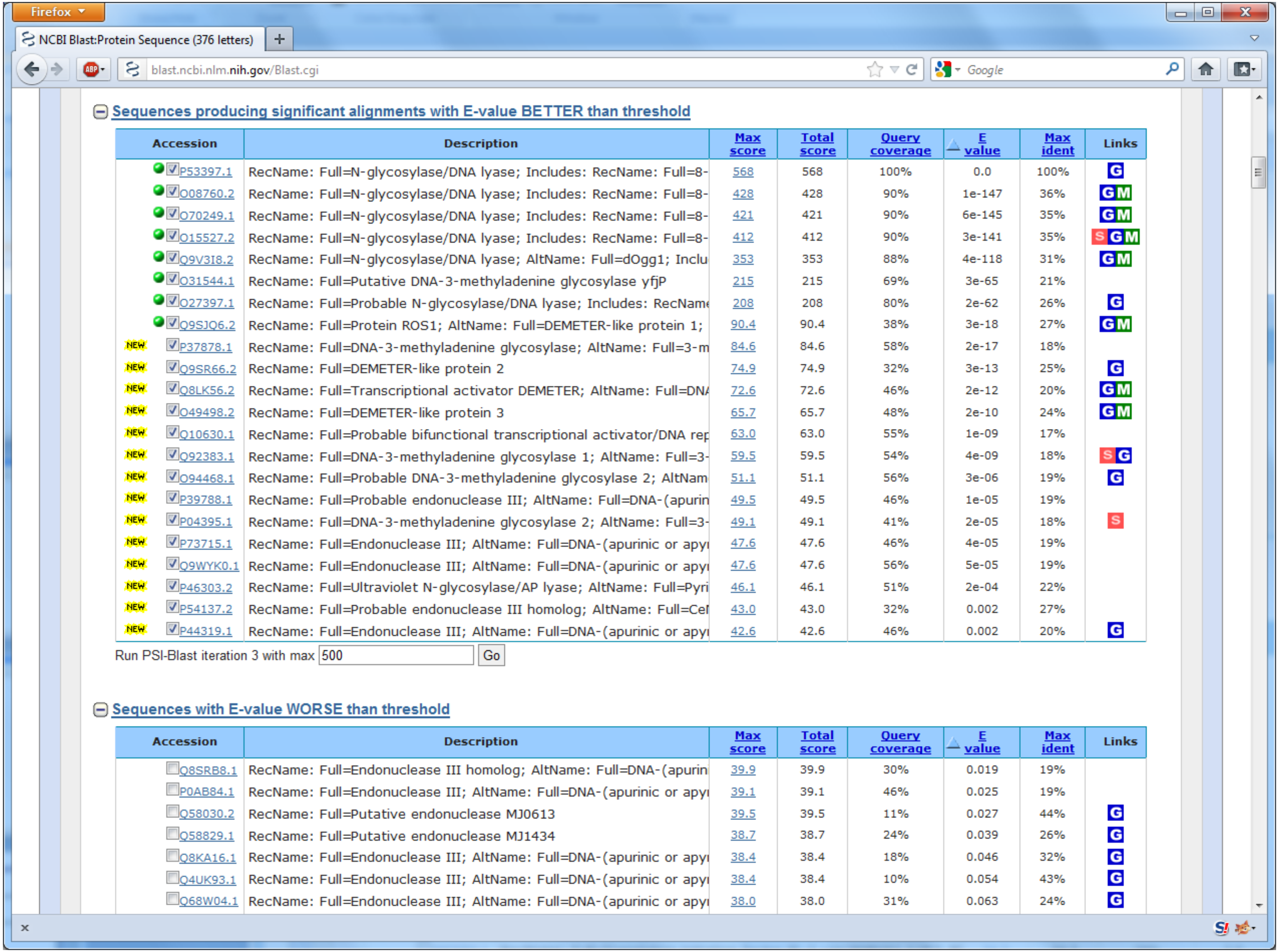
Sequences producing significant alignments with E-value BETTER than threshold

Accession		Description	Max score	Total score	Query coverage	E value	Max ident	Links
NEW	☑ P53397.1	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	783	783	100%	0.0	100%	G
NEW	☑ O08760.2	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	197	197	90%	1e-57	36%	G M
NEW	☑ O70249.1	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	196	196	86%	3e-57	37%	G M
NEW	☑ O15527.2	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	194	194	86%	1e-56	37%	S G M
NEW	☑ Q9V3I8.2	RecName: Full=N-glycosylase/DNA lyase; AltName: Full=dOgg1; Inclu	155	155	88%	4e-42	31%	G M
NEW	☑ O27397.1	RecName: Full=Probable N-glycosylase/DNA lyase; Includes: RecName	90.1	90.1	52%	3e-19	31%	G
NEW	☑ Q9SJQ6.2	RecName: Full=Protein ROS1; AltName: Full=DEMETER-like protein 1;	43.5	43.5	26%	0.002	34%	G M
NEW	☑ O31544.1	RecName: Full=Putative DNA-3-methyladenine glycosylase yfjP	42.4	42.4	55%	0.003	23%	

Run PSI-Blast iteration 2 with max 500 Go

Sequences with E-value WORSE than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
□ Q9SR66.2	RecName: Full=DEMETER-like protein 2	42.4	42.4	13%	0.005	43%	G
□ O49498.2	RecName: Full=DEMETER-like protein 3	42.4	42.4	23%	0.005	34%	G M
□ Q4UK93.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyı	37.0	37.0	24%	0.13	38%	G
□ Q10630.1	RecName: Full=Probable bifunctional transcriptional activator/DNA rep	37.7	37.7	35%	0.15	22%	
□ A8GNW1.1	RecName: Full=Translation initiation factor IF-2	36.2	36.2	29%	0.43	29%	G
□ O58030.2	RecName: Full=Putative endonuclease MJ0613	35.4	35.4	12%	0.55	40%	G
□ P18479.2	RecName: Full=Genome polyprotein; Contains: RecName: Full=P1 prot	36.2	36.2	17%	0.55	39%	
□ Q8LK56.2	RecName: Full=Transcriptional activator DEMETER; AltName: Full=DN/	36.2	36.2	13%	0.56	37%	G M
□ Q4UL51.1	RecName: Full=Translation initiation factor IF-2	35.0	35.0	29%	1.1	29%	G
□ Q68WI4.1	RecName: Full=Translation initiation factor IF-2	34.3	34.3	34%	1.6	28%	G
□ Q92383.1	RecName: Full=DNA-3-methyladenine glycosylase 1; AltName: Full=3-	33.5	33.5	37%	1.7	25%	S G
□ P37878.1	RecName: Full=DNA-3-methyladenine glycosylase; AltName: Full=3-m	33.9	33.9	39%	1.8	21%	
□ Q9ZCZ8.1	RecName: Full=Translation initiation factor IF-2	34.3	34.3	29%	1.9	29%	G
□ A8GSP4.1	RecName: Full=Translation initiation factor IF-2 >sp|B0BY61.1|IF2_RI	34.3	34.3	29%	2.0	29%	G


Firefox
NCBI Blast:Protein Sequence (376 letters)
blast.ncbi.nlm.nih.gov/Blast.cgi
Google

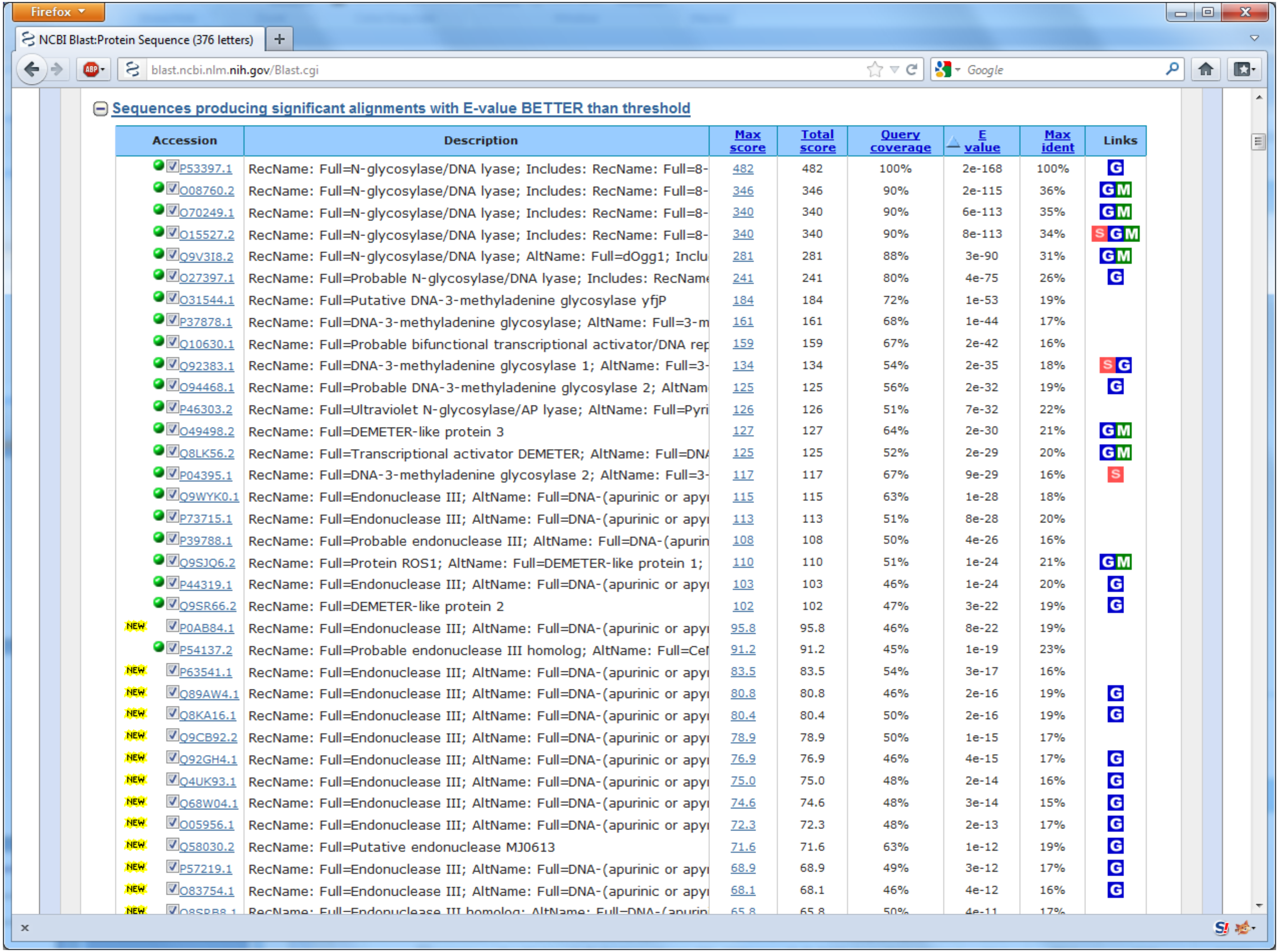
Sequences producing significant alignments with E-value BETTER than threshold

	Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
●	☑ P53397.1	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	568	568	100%	0.0	100%	G
●	☑ O08760.2	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	428	428	90%	1e-147	36%	GM
●	☑ O70249.1	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	421	421	90%	6e-145	35%	GM
●	☑ O15527.2	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	412	412	90%	3e-141	35%	SGM
●	☑ Q9V3I8.2	RecName: Full=N-glycosylase/DNA lyase; AltName: Full=dOgg1; Inclu	353	353	88%	4e-118	31%	GM
●	☑ O31544.1	RecName: Full=Putative DNA-3-methyladenine glycosylase yfjP	215	215	69%	3e-65	21%	
●	☑ O27397.1	RecName: Full=Probable N-glycosylase/DNA lyase; Includes: RecName	208	208	80%	2e-62	26%	G
●	☑ Q9SJQ6.2	RecName: Full=Protein ROS1; AltName: Full=DEMETER-like protein 1;	90.4	90.4	38%	3e-18	27%	GM
NEW	☑ P37878.1	RecName: Full=DNA-3-methyladenine glycosylase; AltName: Full=3-m	84.6	84.6	58%	2e-17	18%	
NEW	☑ Q9SR66.2	RecName: Full=DEMETER-like protein 2	74.9	74.9	32%	3e-13	25%	G
NEW	☑ Q8LK56.2	RecName: Full=Transcriptional activator DEMETER; AltName: Full=DNA	72.6	72.6	46%	2e-12	20%	GM
NEW	☑ O49498.2	RecName: Full=DEMETER-like protein 3	65.7	65.7	48%	2e-10	24%	GM
NEW	☑ Q10630.1	RecName: Full=Probable bifunctional transcriptional activator/DNA rep	63.0	63.0	55%	1e-09	17%	
NEW	☑ Q92383.1	RecName: Full=DNA-3-methyladenine glycosylase 1; AltName: Full=3-	59.5	59.5	54%	4e-09	18%	SG
NEW	☑ O94468.1	RecName: Full=Probable DNA-3-methyladenine glycosylase 2; AltNam	51.1	51.1	56%	3e-06	19%	G
NEW	☑ P39788.1	RecName: Full=Probable endonuclease III; AltName: Full=DNA-(apurin	49.5	49.5	46%	1e-05	19%	
NEW	☑ P04395.1	RecName: Full=DNA-3-methyladenine glycosylase 2; AltName: Full=3-	49.1	49.1	41%	2e-05	18%	S
NEW	☑ P73715.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	47.6	47.6	46%	4e-05	19%	
NEW	☑ Q9WYK0.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	47.6	47.6	56%	5e-05	19%	
NEW	☑ P46303.2	RecName: Full=Ultraviolet N-glycosylase/AP lyase; AltName: Full=Pyri	46.1	46.1	51%	2e-04	22%	
NEW	☑ P54137.2	RecName: Full=Probable endonuclease III homolog; AltName: Full=Cel	43.0	43.0	32%	0.002	27%	
NEW	☑ P44319.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	42.6	42.6	46%	0.002	20%	G

Run PSI-Blast iteration 3 with max 500 Go

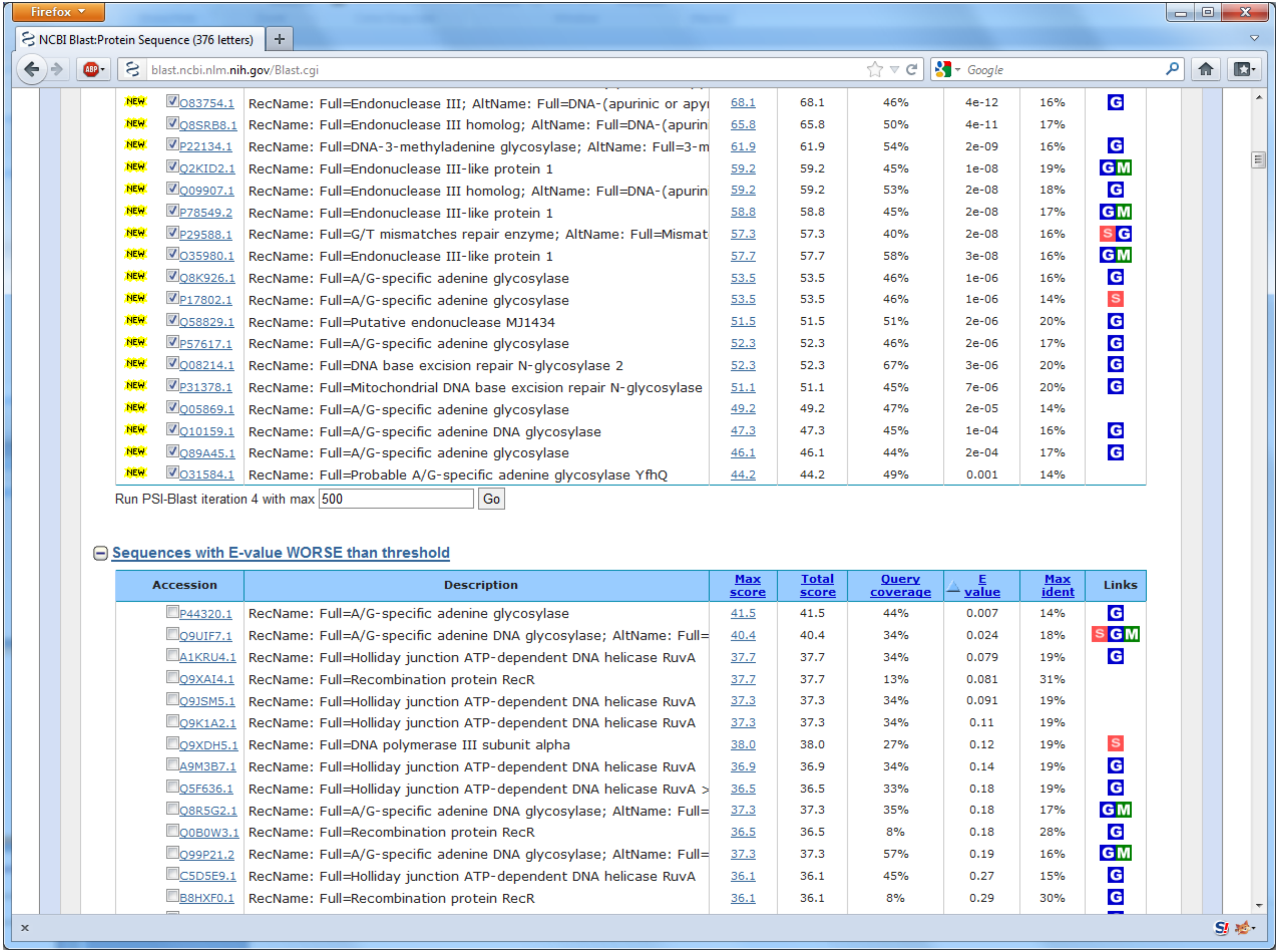
Sequences with E-value WORSE than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
□ Q8SRB8.1	RecName: Full=Endonuclease III homolog; AltName: Full=DNA-(apurin	39.9	39.9	30%	0.019	19%	
□ P0AB84.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	39.1	39.1	46%	0.025	19%	
□ Q58030.2	RecName: Full=Putative endonuclease MJ0613	39.5	39.5	11%	0.027	44%	G
□ Q58829.1	RecName: Full=Putative endonuclease MJ1434	38.7	38.7	24%	0.039	26%	G
□ Q8KA16.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	38.4	38.4	18%	0.046	32%	G
□ Q4UK93.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	38.4	38.4	10%	0.054	43%	G
□ Q68W04.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	38.0	38.0	31%	0.063	24%	G


Firefox
NCBI Blast:Protein Sequence (376 letters)
blast.ncbi.nlm.nih.gov/Blast.cgi
Google

Sequences producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
● ☑ P53397.1	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	482	482	100%	2e-168	100%	G
● ☑ O08760.2	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	346	346	90%	2e-115	36%	G M
● ☑ O70249.1	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	340	340	90%	6e-113	35%	G M
● ☑ O15527.2	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	340	340	90%	8e-113	34%	S G M
● ☑ Q9V3I8.2	RecName: Full=N-glycosylase/DNA lyase; AltName: Full=dOgg1; Inclu	281	281	88%	3e-90	31%	G M
● ☑ O27397.1	RecName: Full=Probable N-glycosylase/DNA lyase; Includes: RecName	241	241	80%	4e-75	26%	G
● ☑ O31544.1	RecName: Full=Putative DNA-3-methyladenine glycosylase yfjP	184	184	72%	1e-53	19%	
● ☑ P37878.1	RecName: Full=DNA-3-methyladenine glycosylase; AltName: Full=3-m	161	161	68%	1e-44	17%	
● ☑ Q10630.1	RecName: Full=Probable bifunctional transcriptional activator/DNA rep	159	159	67%	2e-42	16%	
● ☑ Q92383.1	RecName: Full=DNA-3-methyladenine glycosylase 1; AltName: Full=3-	134	134	54%	2e-35	18%	S G
● ☑ O94468.1	RecName: Full=Probable DNA-3-methyladenine glycosylase 2; AltNam	125	125	56%	2e-32	19%	G
● ☑ P46303.2	RecName: Full=Ultraviolet N-glycosylase/AP lyase; AltName: Full=Pyri	126	126	51%	7e-32	22%	
● ☑ O49498.2	RecName: Full=DEMETER-like protein 3	127	127	64%	2e-30	21%	G M
● ☑ Q8LK56.2	RecName: Full=Transcriptional activator DEMETER; AltName: Full=DN/	125	125	52%	2e-29	20%	G M
● ☑ P04395.1	RecName: Full=DNA-3-methyladenine glycosylase 2; AltName: Full=3-	117	117	67%	9e-29	16%	S
● ☑ Q9WYK0.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	115	115	63%	1e-28	18%	
● ☑ P73715.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	113	113	51%	8e-28	20%	
● ☑ P39788.1	RecName: Full=Probable endonuclease III; AltName: Full=DNA-(apurin	108	108	50%	4e-26	16%	
● ☑ Q9SJQ6.2	RecName: Full=Protein ROS1; AltName: Full=DEMETER-like protein 1;	110	110	51%	1e-24	21%	G M
● ☑ P44319.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	103	103	46%	1e-24	20%	G
● ☑ Q9SR66.2	RecName: Full=DEMETER-like protein 2	102	102	47%	3e-22	19%	G
NEW ☑ P0AB84.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	95.8	95.8	46%	8e-22	19%	
● ☑ P54137.2	RecName: Full=Probable endonuclease III homolog; AltName: Full=Cel	91.2	91.2	45%	1e-19	23%	
NEW ☑ P63541.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	83.5	83.5	54%	3e-17	16%	
NEW ☑ Q89AW4.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	80.8	80.8	46%	2e-16	19%	G
NEW ☑ Q8KA16.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	80.4	80.4	50%	2e-16	19%	G
NEW ☑ Q9CB92.2	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	78.9	78.9	50%	1e-15	17%	
NEW ☑ Q92GH4.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	76.9	76.9	46%	4e-15	17%	G
NEW ☑ Q4UK93.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	75.0	75.0	48%	2e-14	16%	G
NEW ☑ Q68W04.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	74.6	74.6	48%	3e-14	15%	G
NEW ☑ O05956.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	72.3	72.3	48%	2e-13	17%	G
NEW ☑ Q58030.2	RecName: Full=Putative endonuclease MJ0613	71.6	71.6	63%	1e-12	19%	G
NEW ☑ P57219.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	68.9	68.9	49%	3e-12	17%	G
NEW ☑ O83754.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	68.1	68.1	46%	4e-12	16%	G
NEW ☑ Q8SPB8.1	RecName: Full=Endonuclease III homolog; AltName: Full=DNA-(apurin	65.8	65.8	50%	4e-11	17%	


Firefox
NCBI Blast:Protein Sequence (376 letters)
blast.ncbi.nlm.nih.gov/Blast.cgi
Google

NEW	☑ O83754.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyı	68.1	68.1	46%	4e-12	16%	G
NEW	☑ Q8SRB8.1	RecName: Full=Endonuclease III homolog; AltName: Full=DNA-(apurin	65.8	65.8	50%	4e-11	17%	
NEW	☑ P22134.1	RecName: Full=DNA-3-methyladenine glycosylase; AltName: Full=3-m	61.9	61.9	54%	2e-09	16%	G
NEW	☑ Q2KID2.1	RecName: Full=Endonuclease III-like protein 1	59.2	59.2	45%	1e-08	19%	GM
NEW	☑ Q09907.1	RecName: Full=Endonuclease III homolog; AltName: Full=DNA-(apurin	59.2	59.2	53%	2e-08	18%	G
NEW	☑ P78549.2	RecName: Full=Endonuclease III-like protein 1	58.8	58.8	45%	2e-08	17%	GM
NEW	☑ P29588.1	RecName: Full=G/T mismatches repair enzyme; AltName: Full=Mismat	57.3	57.3	40%	2e-08	16%	SG
NEW	☑ O35980.1	RecName: Full=Endonuclease III-like protein 1	57.7	57.7	58%	3e-08	16%	GM
NEW	☑ Q8K926.1	RecName: Full=A/G-specific adenine glycosylase	53.5	53.5	46%	1e-06	16%	G
NEW	☑ P17802.1	RecName: Full=A/G-specific adenine glycosylase	53.5	53.5	46%	1e-06	14%	S
NEW	☑ Q58829.1	RecName: Full=Putative endonuclease MJ1434	51.5	51.5	51%	2e-06	20%	G
NEW	☑ P57617.1	RecName: Full=A/G-specific adenine glycosylase	52.3	52.3	46%	2e-06	17%	G
NEW	☑ Q08214.1	RecName: Full=DNA base excision repair N-glycosylase 2	52.3	52.3	67%	3e-06	20%	G
NEW	☑ P31378.1	RecName: Full=Mitochondrial DNA base excision repair N-glycosylase	51.1	51.1	45%	7e-06	20%	G
NEW	☑ Q05869.1	RecName: Full=A/G-specific adenine glycosylase	49.2	49.2	47%	2e-05	14%	
NEW	☑ Q10159.1	RecName: Full=A/G-specific adenine DNA glycosylase	47.3	47.3	45%	1e-04	16%	G
NEW	☑ Q89A45.1	RecName: Full=A/G-specific adenine glycosylase	46.1	46.1	44%	2e-04	17%	G
NEW	☑ O31584.1	RecName: Full=Probable A/G-specific adenine glycosylase YfhQ	44.2	44.2	49%	0.001	14%	

Run PSI-Blast iteration 4 with max 500 Go

⊟ Sequences with E-value WORSE than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
□ P44320.1	RecName: Full=A/G-specific adenine glycosylase	41.5	41.5	44%	0.007	14%	G
□ Q9UIF7.1	RecName: Full=A/G-specific adenine DNA glycosylase; AltName: Full=	40.4	40.4	34%	0.024	18%	SGM
□ A1KRU4.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA	37.7	37.7	34%	0.079	19%	G
□ Q9XAI4.1	RecName: Full=Recombination protein RecR	37.7	37.7	13%	0.081	31%	
□ Q9JSM5.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA	37.3	37.3	34%	0.091	19%	
□ Q9K1A2.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA	37.3	37.3	34%	0.11	19%	
□ Q9XDH5.1	RecName: Full=DNA polymerase III subunit alpha	38.0	38.0	27%	0.12	19%	S
□ A9M3B7.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA	36.9	36.9	34%	0.14	19%	G
□ Q5F636.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA >	36.5	36.5	33%	0.18	19%	G
□ Q8R5G2.1	RecName: Full=A/G-specific adenine DNA glycosylase; AltName: Full=	37.3	37.3	35%	0.18	17%	GM
□ Q0B0W3.1	RecName: Full=Recombination protein RecR	36.5	36.5	8%	0.18	28%	G
□ Q99P21.2	RecName: Full=A/G-specific adenine DNA glycosylase; AltName: Full=	37.3	37.3	57%	0.19	16%	GM
□ C5D5E9.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA	36.1	36.1	45%	0.27	15%	G
□ B8HXF0.1	RecName: Full=Recombination protein RecR	36.1	36.1	8%	0.29	30%	G

Using a family of proteins as query

Instead of searching with a simple sequence, we can search with a family of proteins, represented by a model.

Models for the representation of a family of protein sequences:

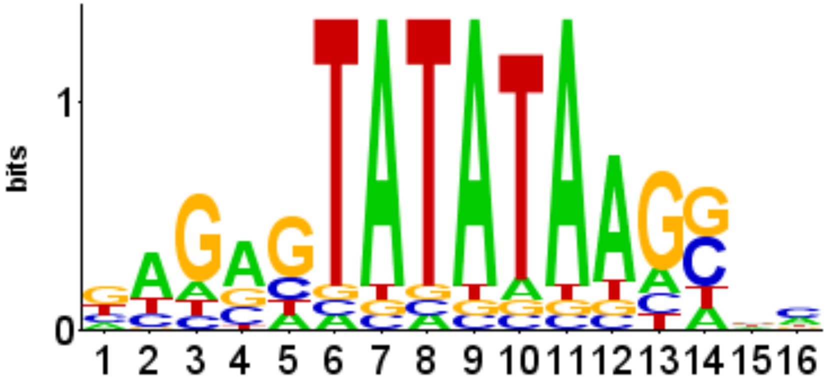
- Set of sequences
- Consensus sequence
- Patterns: Simplified ”regular expressions”
- Profiles: position-specific scoring matrices (PSSMs) based on probabilities of amino acid substitutions (Gribskov *et al.* 1987)
- Hidden Markov models (HMMs): probabilistic model for linear sequences (Haussler *et al.* 1993)

A good multiple alignment of the sequences in the family is essential for most of these models.

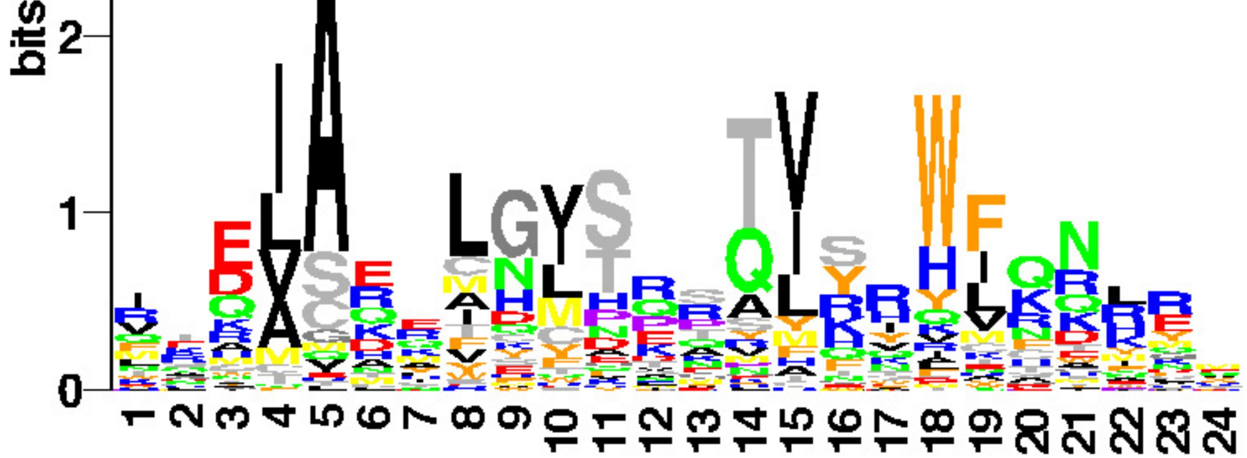
Sequence profiles (PSSMs)

- Position-specific scoring matrices
- Based on a multiple alignment of proteins in a family
- A matrix of 21 x L cells, where L is the length of the alignment (21 for the 20 amino acids + gap)
- Scores in each cell are calculated as a weighted average of the scores from a substitution score matrix (e.g. BLOSUM62) for matching a certain amino acid with each of the amino acids present in the proteins in a specific position in the multiple alignment.
- Sequences are weighted in order to reduce the effect of many similar sequences.

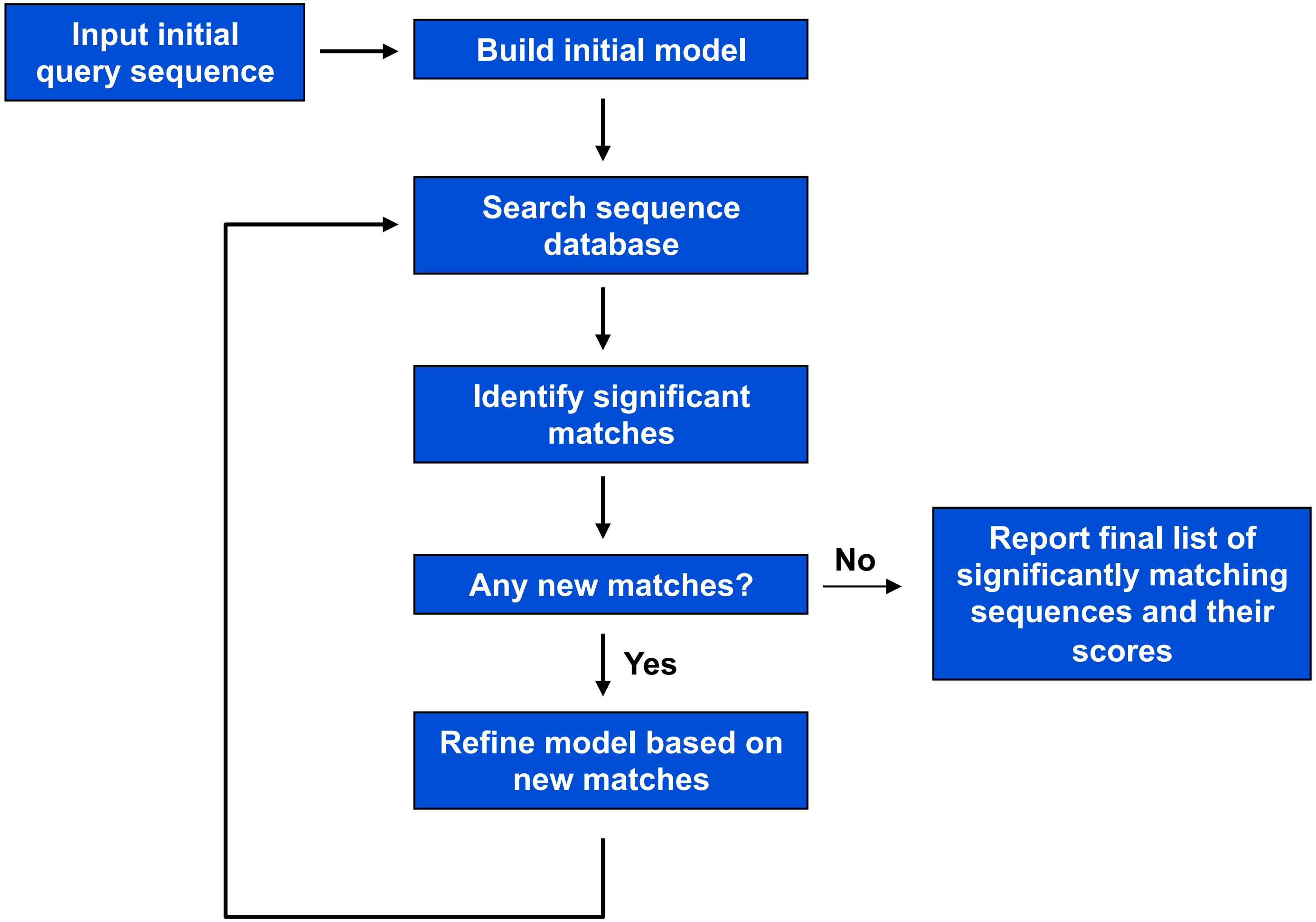
DNA and protein sequences logos



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	0.27	0.40	0.14	0.40	0.18	0.03	0.91	0.03	0.91	0.11	0.91	0.62	0.34	0.08	0.24	0.21
C	0.21	0.18	0.18	0.05	0.34	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.08	0.42	0.26	0.29
G	0.32	0.26	0.52	0.32	0.45	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.55	0.42	0.26	0.24
T	0.19	0.16	0.16	0.23	0.03	0.91	0.03	0.91	0.03	0.83	0.03	0.32	0.03	0.08	0.24	0.26



Iterated searches



Input initial query sequence

Build initial model

Search sequence database

Identify significant matches

Any new matches?

No

Report final list of significantly matching sequences and their scores

Yes

Refine model based on new matches

Literature

PSI-BLAST paper

- *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*
  Altschul SF *et al.* (1997)
  Nucleic Acids Research, 25, 3389-3402.
  http://nar.oupjournals.org/cgi/content/abstract/25/17/3389

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs

Stephen F. Altschul*, Thomas L. Madden, Alejandro A. Schäffer[1], Jinghui Zhang, Zheng Zhang[2], Webb Miller[2] and David J. Lipman

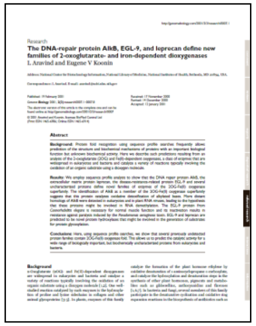
ABSTRACT

The BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities. For protein comparisons, a variety of definitional, algorithmic and statistical refinements described here permits the execution time of the BLAST programs to be decreased substantially while enhancing their sensitivity to weak similarities. A new criterion for triggering the extension of word hits, combined with a new heuristic for generating gapped alignments, yields a gapped BLAST program that runs at approximately three times the speed of the original. In addition, a method is introduced for automatically combining statistically significant alignments produced by BLAST into a position-specific score matrix, and searching the database using this matrix. The resulting Position-Specific Iterated BLAST (PSI-BLAST) program runs at approximately the same speed per iteration as gapped BLAST, but in many cases is much more sensitive to weak but biologically relevant sequence similarities. PSI-BLAST is used to uncover several new and interesting members of the BRCT superfamily.

INTRODUCTION

AlkB paper

- *The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases*
  Aravind L, Koonin EV (2001)
  Genome Biology, 2(3):RESEARCH0007.
  http://genomebiology.com/2001/2/3/RESEARCH/0007



Research

The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases

L Aravind and Eugene V Koonin

Abstract

Background: Protein fold recognition using sequence profile searches frequently allows prediction of the structure and biochemical mechanisms of proteins with an important biological function but unknown biochemical activity. Here we describe such predictions resulting from an analysis of the 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenases, a class of enzymes that are widespread in eukaryotes and bacteria and catalyze a variety of reactions typically involving the oxidation of an organic substrate using a dioxygen molecule.

Results: We employ sequence profile analysis to show that the DNA repair protein AlkB, the extracellular matrix protein leprecan, the disease-resistance-related protein EGL-9 and several uncharacterized proteins define novel families of enzymes of the 2OG-Fe(II) oxygenase superfamily. The identification of AlkB as a member of the 2OG-Fe(II) oxygenase superfamily suggests that this protein catalyzes oxidative detoxification of alkylated bases. More distant homologs of AlkB were detected in eukaryotes and in plant RNA viruses, leading to the hypothesis that these proteins might be involved in RNA demethylation. The EGL-9 protein from *Caenorhabditis elegans* is necessary for normal muscle function and its inactivation results in resistance against paralysis induced by the *Pseudomonas aeruginosa* toxin. EGL-9 and leprecan are predicted to be novel protein hydroxylases that might be involved in the generation of substrates for protein glycosylation.

Conclusions: Here, using sequence profile searches, we show that several previously undetected protein families contain 2OG-Fe(II) oxygenase fold. This allows us to predict the catalytic activity for a wide range of biologically important, but biochemically uncharacterized proteins from eukaryotes and bacteria.

Background

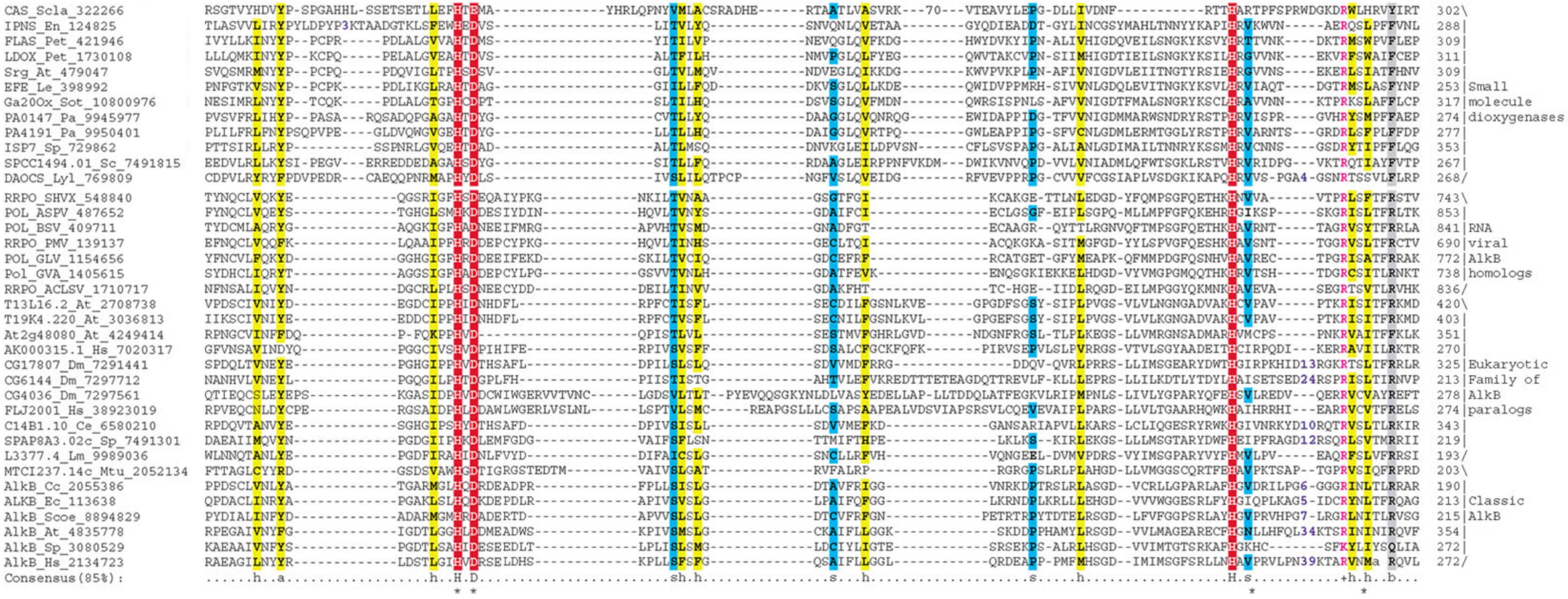
Multiple sequence alignment

What is a multiple alignment (MSA)?

- Extension of pairwise alignments to three or more sequences
- Usually global alignments – entire sequences included
- Indicates common conserved residues in all or most sequences – usually important for function / activity
- Indicates accepted residues in the different positions
- Indicates positions where gaps are more likely

- Basis for construction of phylogenetic trees
- Basis for sequence motifs and profiles
- Essential for evolutionary studies and phylogenetics

Example



```
CAS_Scla_322266            RSGTVYHDVYP-SPGAHHL-SSETSETLLEFHTEMA-------------YHRLQPNYVMLACSRADHE------RTAATLVASVRK---70---VTEAVYLEPG-DLLIVDNF-----------RTTHARTPFSPRWDGKDRWLHRVYIRT  302\
IPNS_En_124825             TLASVVLIRYPYLDPYP3KTAADGTKLSFEWHEDVS-------------------LITVLYQ------------SNVQNLQVETAA--------GYQDIEADDT-GYLINCGSYMAHLTNNYYKAPIHRVKWVN-----AERQSLPFFVNL  288|
FLAS_Pet_421946            IVYLLKINYYP--PCPR-----PDLALGVVAHTDMS-------------------YITILVP------------NEVQGLQVFKDG--------HWYDVKYIPN-ALIVHIGDQVEILSNGKYKSVYHRTTVNK----DKTRMSWPVFLEP  309|
LDOX_Pet_1730108           LLLQMKINYYP--KCPQ-----PELALGVEAHTDVS-------------------ALTFILH------------NMVPGLQLFYEG--------QWVTAKCVPN-SIIMHIGDTIEILSNGKYKSILHRGVVNK----EKVRFSWAIFCEP  311|
Srg_At_479047              SVQSMRMNYYP--PCPQ-----PDQVIGLTPHSDSV-------------------GLTVLMQV-----------NDVEGLQIKKDG--------KWVPVKPLPN-AFIVNIGDVLEIITNGTYRSIEHRGVVNS----EKERLSIATFHNV  309|
EFE_Le_398992              PNFGTKVSNYP--PCPK-----PDLIKGLRAHTDAG-------------------GIILLFQD-----------DKVSGLQLLKDE--------QWIDVPPMRH-SIVVNLGDQLEVITNGKYKSVLHRVIAQT----DGTRMSLASFYNP  253|Small
Ga20Ox_Sot_10800976        NESIMRLNYYP--TCQK-----PDLALGTGPHCDPT-------------------SLTILHQ------------DSVSGLQVFMDN--------QWRSISPNLS-AFVVNIGDTFMALSNGRYKSCLHRAVVNN----KTPRKSLAFFLCP  317|molecule
PA0147_Pa_9945977          PVSVFRLIHYP--PASA---RQSADQPGAGAHTDYG-------------------CVTLLYQ------------DAAGGLQVQNRQG-------EWIDAPPIDG-TFVVNIGDMMARWSNDRYRSTPHRVISPR----GVHRYSMPFFAEP  274|dioxygenases
PA4191_Pa_9950401          PLILFRLFNYPSQPVPE----GLDVQWGVGEHTDYG-------------------LLTLLHQ------------DAIGGLQVRTPQ--------GWLEAPPIPG-SFVCNLGDMLERMTGGLYRSTPHRVARNTS---GRDRLSFPLFFDP  277|
ISP7_Sp_729862             PTTSIRLLRYP----------SSPNRLGVQEHTDAD-------------------ALTLMSQ------------DNVKGLEILDPVSN------CFLSVSPAPG-ALIANLGDIMAILTNNRYKSSMHRVCNNS----GSDRYTIPFFLQG  353|
SPCC1494.01_Sc_7491815     EEDVLRLLKYSI-PEGV---ERREDDEDAGAHSDYG-------------------SITLLFQ------------RDAAGLEIRPPNFVKDM---DWIKVNVQPD-VVLVNIADMLQFWTSGKLRSTVHRVRIDPG---VKTRQTIAYFVTP  267|
DAOCS_Lyl_769809           CDPVLRYRYFPDVPEDR---CAEQQPNRMAPHYDLS-------------------IVSLILQTPCP--------NGFVSLQVEIDG--------RFVEVPPRPG-CVVVFCGSIAPLVSDGKIKAPQHRVVS-PGA4-GSNRTSSVLFLRP  268/

RRPO_SHVX_548840           TYNQCLVQKYE-------------QGSRIGFHSDEQAIYPKG-----------NKILTVNAA-------------GSGTFGI---------------KCAKGE-TTLNLEDGD-YFQMPSGFQETHKHNVVA------VTPRLSFTFRSTV  743\
POL_ASPV_487652            FYNQCLVQEYS-------------TGHGLSMHKDDESIYDIN-----------HQVLTVNYS-------------GDAIFCI---------------ECLGSGF-EIPLSGPQ-MLLMPFGFQKEHRHGIKSP-----SKGRISLTFRLTK  853|
POL_BSV_409711             TYDCMLAQRYG-------------AQGKIGFHADNEEIFMRG-----------APVHTVSMD-------------GNADFGT---------------ECAAGR--QYTTLRGNVQFTMPSGFQETHKHAVRNT-----TAGRVSYTFRRLA  841|RNA
RRPO_PMV_139137            EFNQCLVQQFK-------------LQAAIPFHRDDEPCYPKG-----------HQVLTINHS-------------GECLTQI---------------ACQKGKA-SITMGFGD-YYLSPVGFQESHKHAVSNT-----TGGRVSLTFRCTV  690|viral
POL_GLV_1154656            YFNCVLFQKYD-------------GGHGIGFHRDDEEIFEKD-----------SKILTVCIQ-------------GDCEFRF---------------RCATGET-GFYMEAPK-QFMMPDGFQSNHVHAVREC-----TPGRISATFRRAK  772|AlkB
Pol_GVA_1405615            SYDHCLIQRYT-------------AGGSIGFHADDEPCYLPG-----------GSVVTVNLH-------------GDATFEVK--------------ENQSGKIEKKELHDGD-VYVMGPGMQQTHKHRVTSH-----TDGRCSITLRNKT  738|homologs
RRPO_ACLSV_1710717         NFNSALIQVYN-------------DGCRLPLHSDNEECYDD------------DEILTINVV-------------GDAKFHT---------------TC-HGE--IIDLRQGD-EILMPGGYQKMNKHAVEVA-----SEGRTSVTLRVHK  836/
T13L16.2_At_2708738        VPDSCIVNIYD-------------EGDCIPPHIDNHDFL--------------RPFCTISFL-------------SECDILFGSNLKVE------GPGDFSGSY-SIPLPVGS-VLVLNGNGADVAKHCVPAV-----PTKRISITFRKMD  420\
T19K4.220_At_3036813       IIKSCIVNIYE-------------EDDCIPPHIDNHDFL--------------RPFCTVSFL-------------SECNILFGSNLKVL------GPGEFSGSY-SIPLPVGS-VLVLKGNGADVAKHCVPAV-----PTKRISITFRKMD  403|
At2g48080_At_4249414       RPNGCVINFFDQ------------P-FQKPPHVD-------------------QPISTLVL--------------SESTMVFGHRLGVD------NDGNFRGSL-TLPLKEGS-LLVMRGNSADMARHVMCPS-----PNKRVAITFFKLK  351|
AK000315.1_Hs_7020317      GFVNSAVINDYQ------------PGGCIVSHVDPIHIFE-------------RPIVSVSFF-------------SDSALCFGCKFQFK-------PIRVSEPVLSLPVRRGS-VTVLSGYAADEITHCIRPQDI---KERRAVIILRKTR  270|
CG17807_Dm_7291441         SPDQLTVNEYE-------------PGHGIPPHVDTHSAFL-------------DPILSLSLQ-------------SDVVMDFRRG---------------DDQV-QVRLPRRS-LLIMSGEARYDWTHGIRPKHID13RGKRTSLTFRRLR  325|Eukaryotic
CG6144_Dm_7297712          NANHVLVNEYL-------------PGQGILPHTDGPLFH--------------PIISTISTG-------------AHTVLEFVKREDTTTETEAGDQTTREVLF-KLLLEPRS-LLILKDTLYTDYLHAISETSED24RSPRISLTIRNVP  213|Family of
CG4036_Dm_7297561          QTIEQCSLEYEPS-----------KGASIDPHVDDCWIWGERVVTVNC-----LGDSVLTLT--PYEVQQSGKYNLDLVASYEDELLAP-LLTDDQLATFEGKVLRIPMPNLS-LIVLYGPARYQFEHSVLREDV---QERRVCVAYREFT  278|AlkB
FLJ2001_Hs_38923019        RPVEQCNLDYCPE-----------RGSAIDPHLDDAWLWGERLVSLNL-----LSPTVLSMC-----REAPGSLLLCSAPSAAPEALVDSVIAPSRSVLCQEVEVAIPLPARS-LLVLTGAARHQWKHAIHRRHI---EARRVCVTFRELS  274|paralogs
C14B1.10_Ce_6580210        RPDQVTANVYE-------------SGHGIPSHYDTHSAFD-------------DPIVSISLL-------------SDVVMEFKD-------------GANSARIAPVLLKARS-LCLIQGESRYRWKHGIVNRKYD10RQTRVSLTLRKIR  343|
SPAP8A3.02c_Sp_7491301     DAEAIIMQVYN-------------PGDGIIPHKDLEMFGDG------------VAIFSFLSN-------------TTMIFTHPE--------------LKLKS--KIRLEKGS-LLLMSGTARYDWFHEIPFRAGD12RSQRLSVTMRRII  219|
L3377.4_Lm_9989036         WLNNQTANLYE-------------PGDFIRAHIDNLFVYD-------------DIFAICSLG-------------SNCLLRFVH-------------VQNGEEL-DVMVPDRS-VYIMSGPARYVYFHMVLPV-----EAQRFSLVFRRSI  193/
MTCI237.14c_Mtu_2052134    FTTAGLCYYRD-------------GSDSVAWHGDTIGRGSTEDTM--------VAIVSLGAT-------------RVFALRP----------------RGRGPSLRLPLAHGD-LLVMGGSCQRTFEHAVPKTSAP--TGPRVSIQFRPRD  203\
AlkB_Cc_2055386            PPDSCLVNLYA-------------TGARMGLHQDRDEADPR------------FPLLSISLG-------------DTAVFRIGG-------------VNRKDPTRSLRLASGD--VCRLLGPARLAFHGVDRILPG6-GGGRINLTLRRAR  190|
ALKB_Ec_113638             QPDACLINRYA-------------PGAKLSLHQDKDEPDLR------------APIVSVSLG-------------LPAIFQFGG-------------LKRNDPLKRLLLEHGD--VVVWGGESRLFYHGIQPLKAG5-IDCRYNLTFRQAG  213|Classic
AlkB_Scoe_8894829          PYDIALINFYD-------------ADARMGMHRDADERTD-------------APVVSLSLG-------------DTCVFRFGN------------PETRTRPYTDTELRSGD--LFVFGGPSRLAYHGVPRVHPG7-LRGRLNITLRVSG  215|AlkB
AlkB_At_4835778            RPEGAIVNYFG-------------IGDTLGGHLDDMEADWS------------KPIVSMSLG-------------CKAIFLLGGK-------------SKDDPPHAMYLRSGD--VVLMAGEARECFHGNLLHFQL34KTSRININIRQVF  354|
AlkB_Sp_3080529            KAEAAIVNFYS-------------PGDTLSAHIDESEEDLT------------LPLISLSMG-------------LDCIYLIGTE------------SRSEKPS-ALRLHSGD--VVIMTGTSRKAFHGKHC-------SFKYLIYSQLIA  272|
AlkB_Hs_2134723            RAEAGILNYYR-------------LDSTLGIHVDRSELDHS------------KPLLSFSFG-------------QSAIFLLGGL------------QRDEAPP-PMFMHSGD--IMIMSGFSRLLNHAVPRVLPN39KTARVNMa RQVL  272/
Consensus(85%):            ......h..a..................h..H.D.......................sh.h................s...h....................s.....h..................H.s...........+h.h..b...
                                                          * *                                                                                               *              *
```

Approaches to multiple alignment

Some of the major approaches used to construct MSAs:

- Brute force optimal alignment (very hard)
- Centre-star alignment (simple, used in PSI-BLAST)
- Progressive alignment (e.g. Clustal W)
- Iterative alignment (e.g. Muscle)

A lot of software...

- Clustal W - progressive
- T-Coffee – progressive
- MUSCLE - iterative
- MAFFT – various technqiues
- ProbCons – probabilistic
- Dialign, Dialign2 – blocks-based
- MSA – full DP
- DCA – divide and conquer
- DbClustal - progressive
- Poa - progressive
- PRALINE - progressive
- PRRN - iterative
- Match-Box – blocks-based
- …

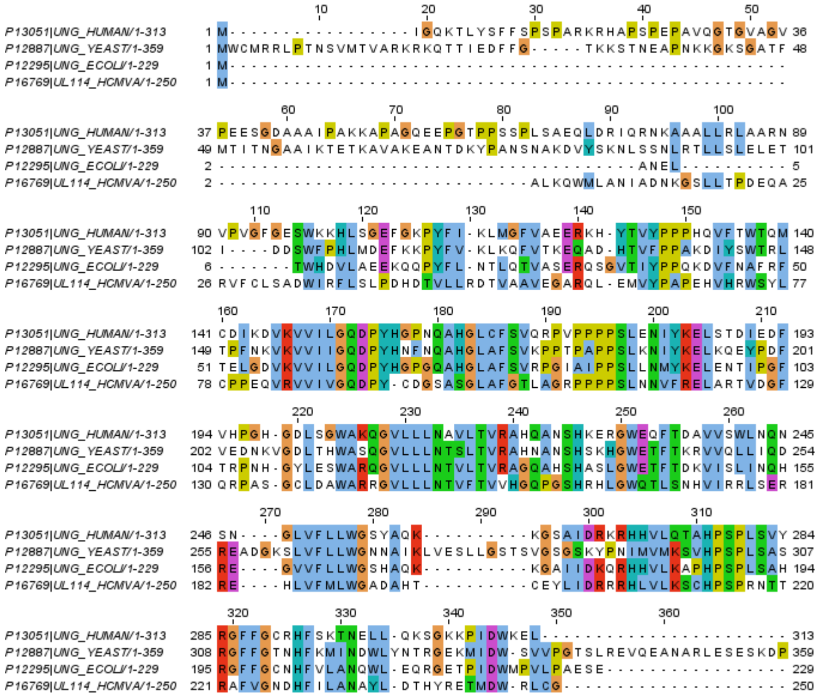
Jalview demo/example

- Jalview is a multiple sequence alignment editor

- www.jalview.org

- Can run the algorithms Clustal W, MUSCLE and MAFFT from within the program

- Very useful for making nice colorful figures



```
                                       10        20        30        40        50
P13051|UNG_HUMAN/1-313       1 M-----------------IGQKTLYSFFSPSPARKRHAPSPEPAVQGTGVAGV 36
P12887|UNG_YEAST/1-359       1 MWCMRRLPTNSVMTVARKRKQTTIEDFFG-----TKKSTNEAPNKKGKSGATF 48
P12295|UNG_ECOLI/1-229       1 M----------------------------------------------------
P16769|UL114_HCMVA/1-250     1 M----------------------------------------------------

                                    60        70        80        90        100
P13051|UNG_HUMAN/1-313      37 PEESGDAAAIPAKKAPAGQEEPGTPPSSPLSAEQLDRIQRNKAAALLRLAARN 89
P12887|UNG_YEAST/1-359      49 MTITNGAAIKTETKAVAKEANTDKYPANSNAKDVYSKNLSSNLRTLLSLELET 101
P12295|UNG_ECOLI/1-229       2 ---------------------------------------ANEL---------- 5
P16769|UL114_HCMVA/1-250     2 -----------------------------ALKQWMLANIADNKGSLLTPDEQA 25

                                 110       120       130       140       150
P13051|UNG_HUMAN/1-313      90 VPVGFGESWKKHLSGEFGKPYFI-KLMGFVAEERKH-YTVYPPPHQVFTWTQM 140
P12887|UNG_YEAST/1-359     102 I----DDSWFPHLMDEFKKPYFV-KLKQFVTKEQAD-HTVFPPAKDIYSWTRL 148
P12295|UNG_ECOLI/1-229       6 -------TWHDVLAEEKQQPYFL-NTLQTVASERQSGVTIYPPQKDVFNAFRF 50
P16769|UL114_HCMVA/1-250    26 RVFCLSADWIRFLSLPDHDTVLLRDTVAAVEGARQL-EMVYPAPEHVHRWSYL 77

                               160       170       180       190       200       210
P13051|UNG_HUMAN/1-313     141 CDIKDVKVVILGQDPYHGPNQAHGLCFSVQRPVPPPPSLENIYKELSTDIEDF 193
P12887|UNG_YEAST/1-359     149 TPFNKVKVVIIGQDPYHNFNQAHGLAFSVKPPTPAPPSLKNIYKELKQEYPDF 201
P12295|UNG_ECOLI/1-229      51 TELGDVKVVILGQDPYHGPGQAHGLAFSVRPGIAIPPSLLNMYKELENTIPGF 103
P16769|UL114_HCMVA/1-250    78 CPPEQVRVVIVGQDPY-CDGSASGLAFGTLAGRPPPPSLNNVFRELARTVDGF 129

                                      220       230       240       250       260
P13051|UNG_HUMAN/1-313     194 VHPGH-GDLSGWAKQGVLLLNAVLTVRAHQANSHKERGWEQFTDAVVSWLNQN 245
P12887|UNG_YEAST/1-359     202 VEDNKVGDLTHWASQGVLLLNTSLTVRAHNANSHSKHGWETFTKRVVQLLIQD 254
P12295|UNG_ECOLI/1-229     104 TRPNH-GYLESWARQGVLLLNTVLTVRAGQAHSHASLGWETFTDKVISLINQH 155
P16769|UL114_HCMVA/1-250   130 QRPAS-GCLDAWARRGVLLLNTVFTVVHGQPGSHRHLGWQTLSNHVIRRLSER 181

                                   270       280       290       300       310
P13051|UNG_HUMAN/1-313     246 SN----GLVFLLWGSYAQK----------KGSAIDRKRHHVLQTAHPSPLSVY 284
P12887|UNG_YEAST/1-359     255 READGKSLVFLLWGNNAIKLVESLLGSTSVGSGSKYPNIMVMKSVHPSPLSAS 307
P12295|UNG_ECOLI/1-229     156 RE----GVVFLLWGSHAQK----------KGAIIDKQRHHVLKAPHPSPLSAH 194
P16769|UL114_HCMVA/1-250   182 RE----HLVFMLWGADAHT----------CEYLIDRRRHLVLKSCHPSPRNTT 220

                                320       330       340       350       360
P13051|UNG_HUMAN/1-313     285 RGFFGCRHFSKTNELL-QKSGKKPIDWKEL----------------------- 313
P12887|UNG_YEAST/1-359     308 RGFFGTNHFKMINDWLYNTRGEKMIDW-SVVPGTSLREVQEANARLESESKDP 359
P12295|UNG_ECOLI/1-229     195 RGFFGCNHFVLANQWL-EQRGETPIDWMPVLPAESE----------------- 229
P16769|UL114_HCMVA/1-250   221 RAFVGNDHFILANAYL-DTHYRETMDW-RLCG--------------------- 250
```

Finding the best multiple alignment

- To find the best multiple sequence alignments the MSA programs will try to find the one with the highest score
- The score is usually the sum-of-pairs-score or similar
- Corresponds approximately to the sum of all pairwise alignment scores
- For the alignment A of m sequences s^1 til s^m we have the sum-of-pairs score S(A):

$$S(\mathcal{A}) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} S(\bar{s}^i, \bar{s}^j).$$

- *S*(*a*,*b*) is the pairwise score of a and b, and s^{-i} is the projection of s^i, that is, s^i with inserted gaps

The sum-of-pairs score

M	Q	P	I	L	L	L
M	L	R	-	L	L	-
M	K	-	I	L	L	L
M	P	P	V	L	I	L

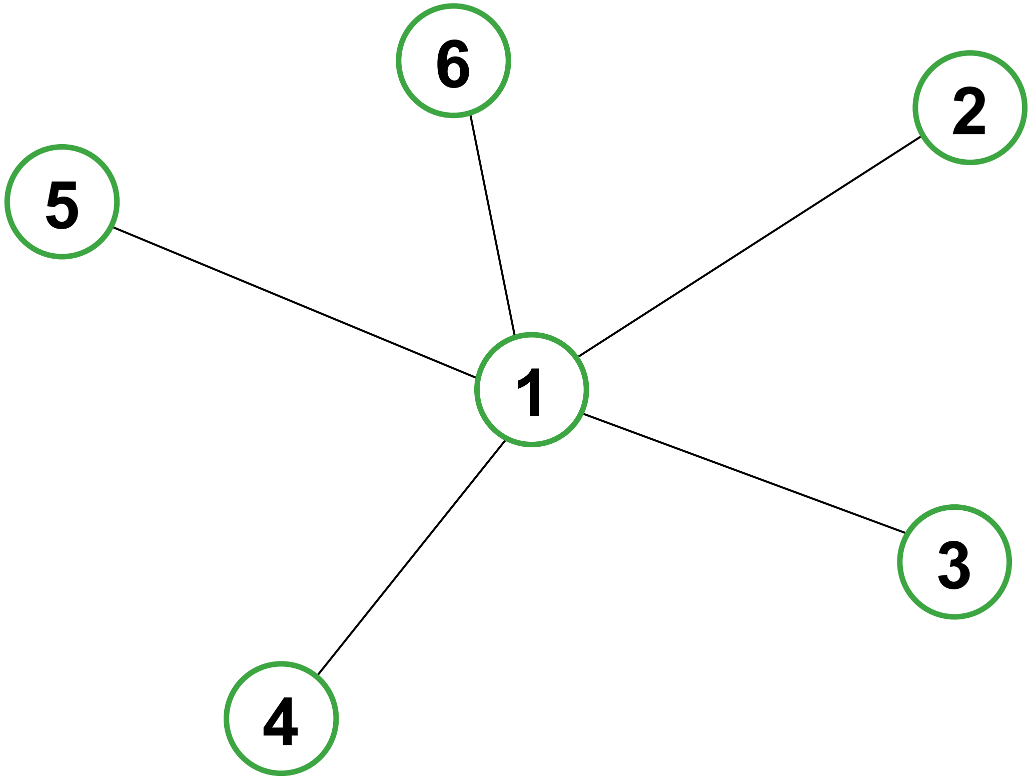
score(k) = S(P,R) + S(P,–) + S(P,P) + S(R,–) + S(R,P) + S(–,P)

score for column k = 3

We have S(–,–) = 0

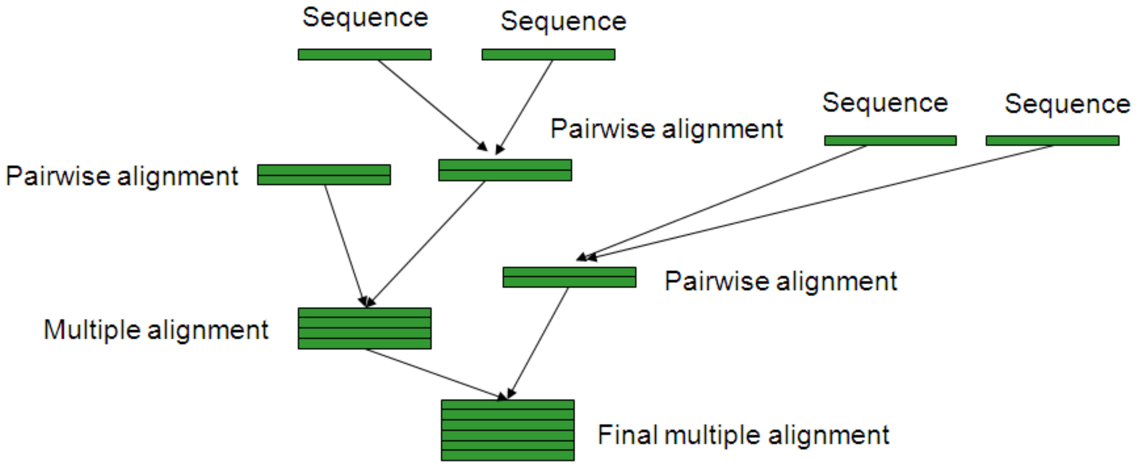
Total score = score(1) + score(2) + ... + score(N)

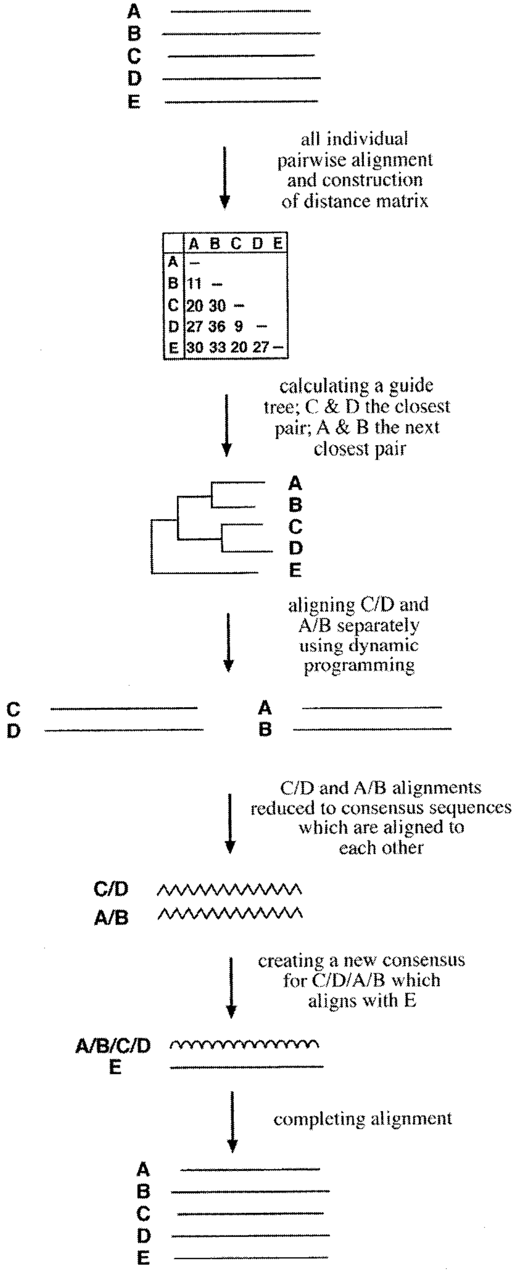
Centre star multiple alignment



Clustal W

- One of the most commonly used and well-known tools for multiple sequence alignment. Now somewhat outdated and surpassed by other tools.
- Uses a progressive algorithm: Always starts with the most similar sequences and then aligns less similar sequences with each other.



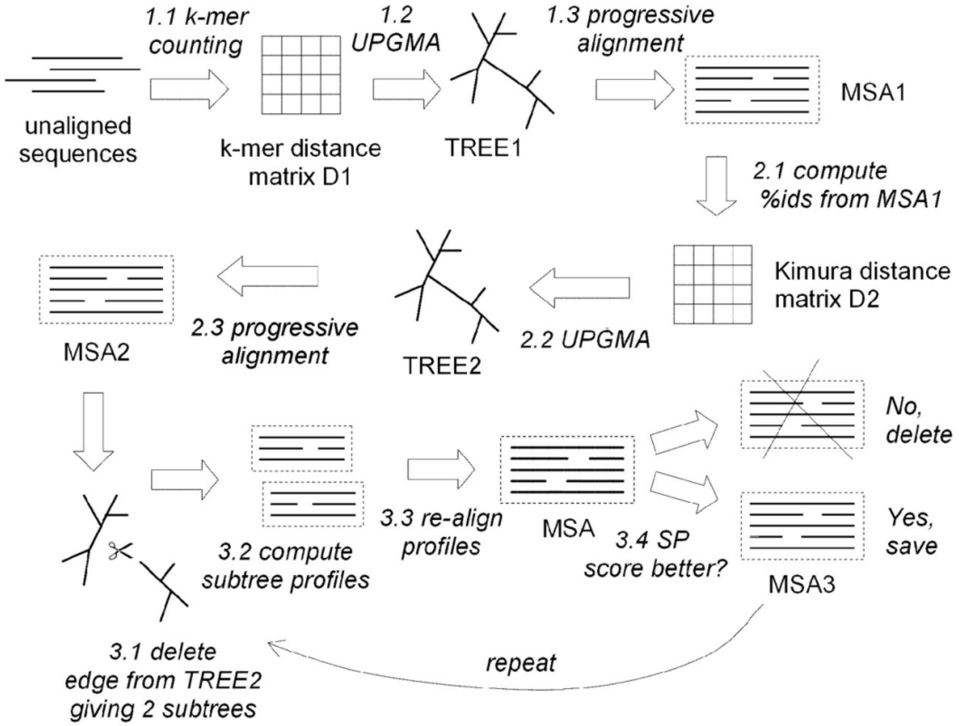


A
B
C
D
E
all individual pairwise alignment and construction of distance matrix
calculating a guide tree; C & D the closest pair; A & B the next closest pair
A
B
C
D
E
aligning C/D and A/B separately using dynamic programming
C
D
A
B
C/D and A/B alignments reduced to consensus sequences which are aligned to each other
C/D
A/B
creating a new consensus for C/D/A/B which aligns with E
A/B/C/D
E
completing alignment
A
B
C
D
E

	A	B	C	D	E
A	–				
B	11	–			
C	20	30	–		
D	27	36	9	–	
E	30	33	20	27	–

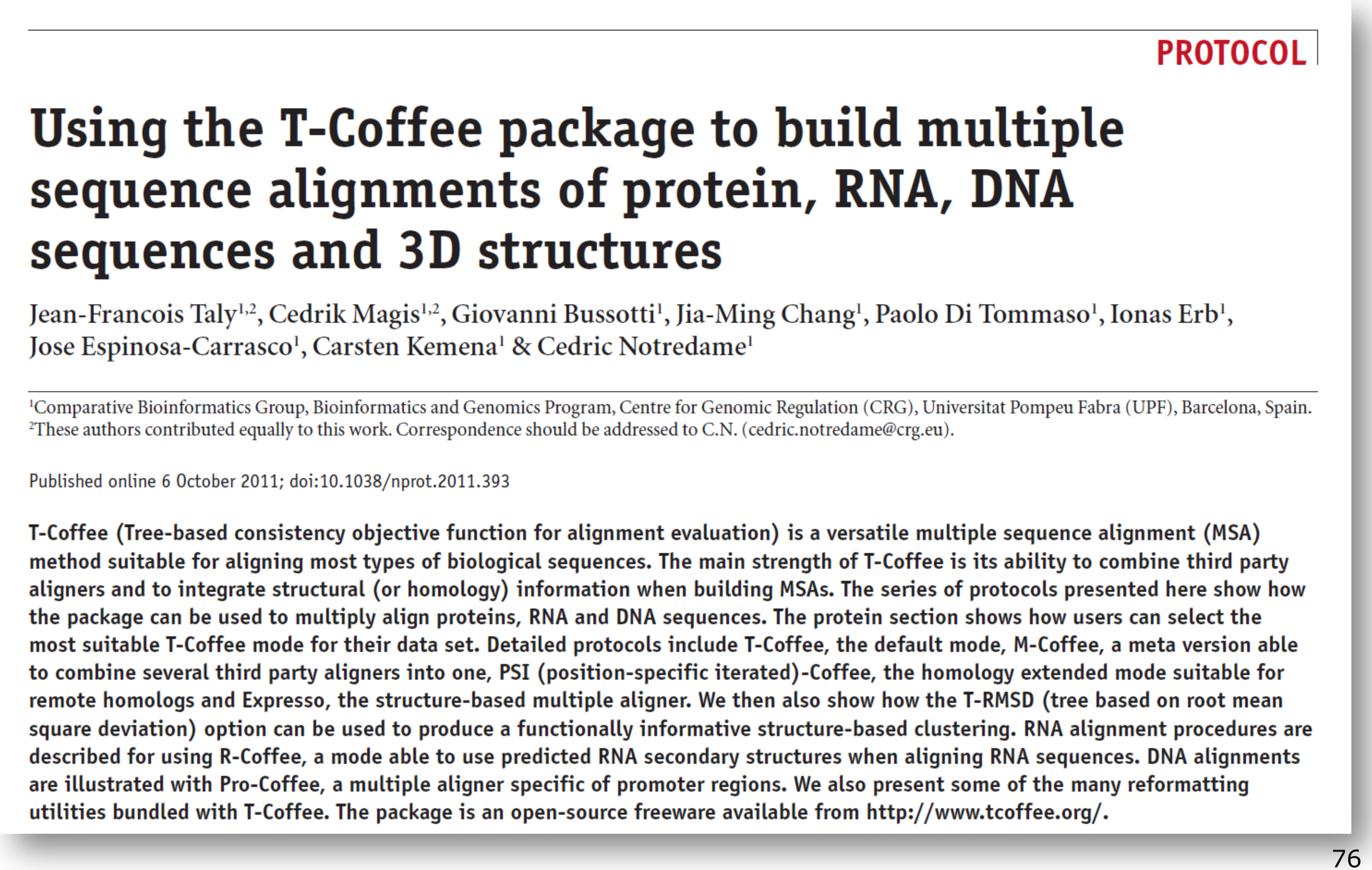
MUSCLE

- MUSCLE = Multiple Sequence Comparison by Log Expectation
- Iterative procedure: improves the alignment gradually until good enough by introducing random changes in the alignment
- Very high quality of alignments
- Much faster than Clustal W



1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
No, delete
Yes, save
MSA3
3.3 re-align profiles
MSA
3.4 SP score better?
3.2 compute subtree profiles
3.1 delete edge from TREE2 giving 2 subtrees
repeat

More here



PROTOCOL

Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures

Jean-Francois Taly[1,2], Cedrik Magis[1,2], Giovanni Bussotti[1], Jia-Ming Chang[1], Paolo Di Tommaso[1], Ionas Erb[1], Jose Espinosa-Carrasco[1], Carsten Kemena[1] & Cedric Notredame[1]

[1]Comparative Bioinformatics Group, Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Universitat Pompeu Fabra (UPF), Barcelona, Spain. [2]These authors contributed equally to this work. Correspondence should be addressed to C.N. (cedric.notredame@crg.eu).

Published online 6 October 2011; doi:10.1038/nprot.2011.393

T-Coffee (Tree-based consistency objective function for alignment evaluation) is a versatile multiple sequence alignment (MSA) method suitable for aligning most types of biological sequences. The main strength of T-Coffee is its ability to combine third party aligners and to integrate structural (or homology) information when building MSAs. The series of protocols presented here show how the package can be used to multiply align proteins, RNA and DNA sequences. The protein section shows how users can select the most suitable T-Coffee mode for their data set. Detailed protocols include T-Coffee, the default mode, M-Coffee, a meta version able to combine several third party aligners into one, PSI (position-specific iterated)-Coffee, the homology extended mode suitable for remote homologs and Expresso, the structure-based multiple aligner. We then also show how the T-RMSD (tree based on root mean square deviation) option can be used to produce a functionally informative structure-based clustering. RNA alignment procedures are described for using R-Coffee, a mode able to use predicted RNA secondary structures when aligning RNA sequences. DNA alignments are illustrated with Pro-Coffee, a multiple aligner specific of promoter regions. We also present some of the many reformatting utilities bundled with T-Coffee. The package is an open-source freeware available from http://www.tcoffee.org/.