BLAST for proteins, step 3

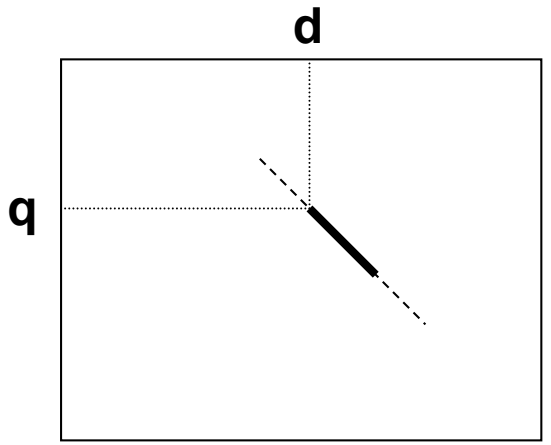
- Expand the hits into HSPs in both directions (no gaps) by adding score values from the substitution score matrix.
- In each direction, stop when the score decreases more than a threshold X from the highest score seen so far.

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). △

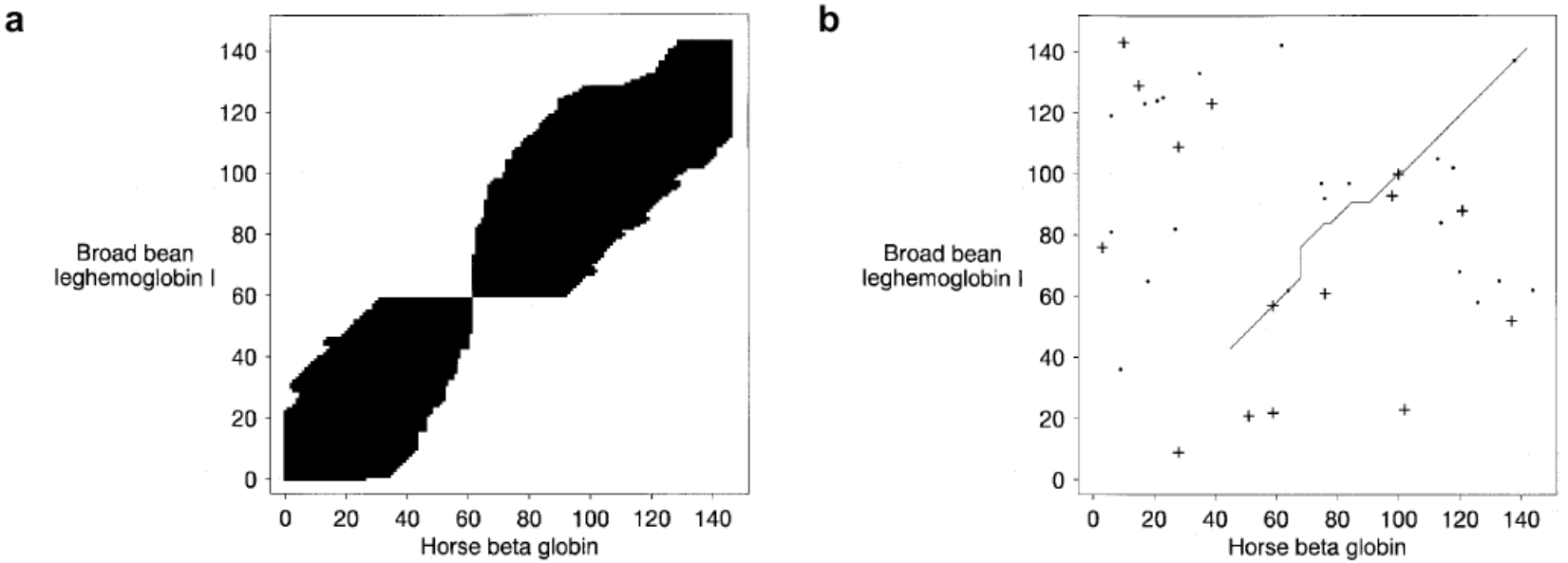


BLAST for proteins, step 4

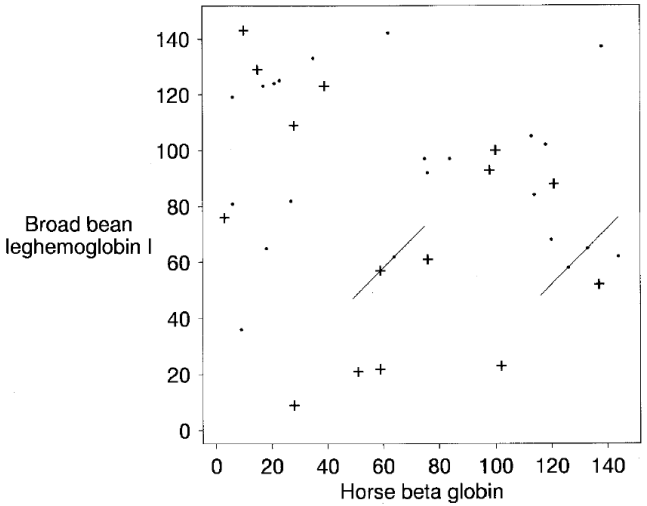
- Keep HSPs with score of at least S_g.
- The threshold is set to corresponds to approximately 2% of the database sequences on average

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



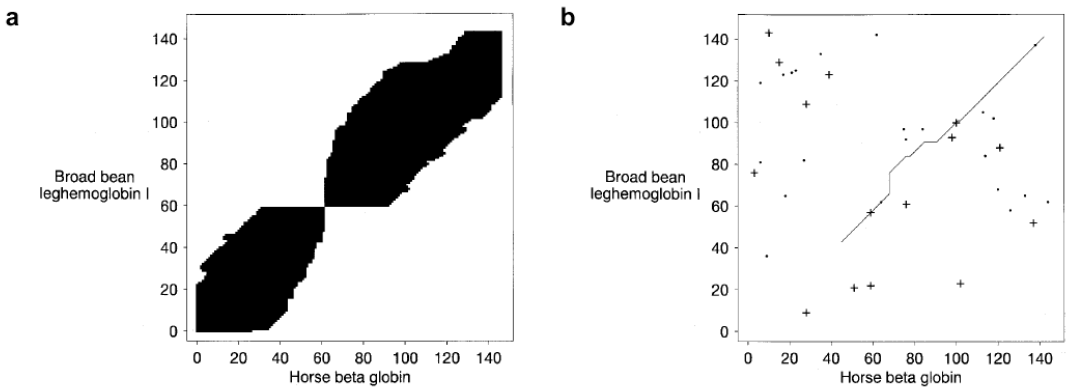
BLAST example



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

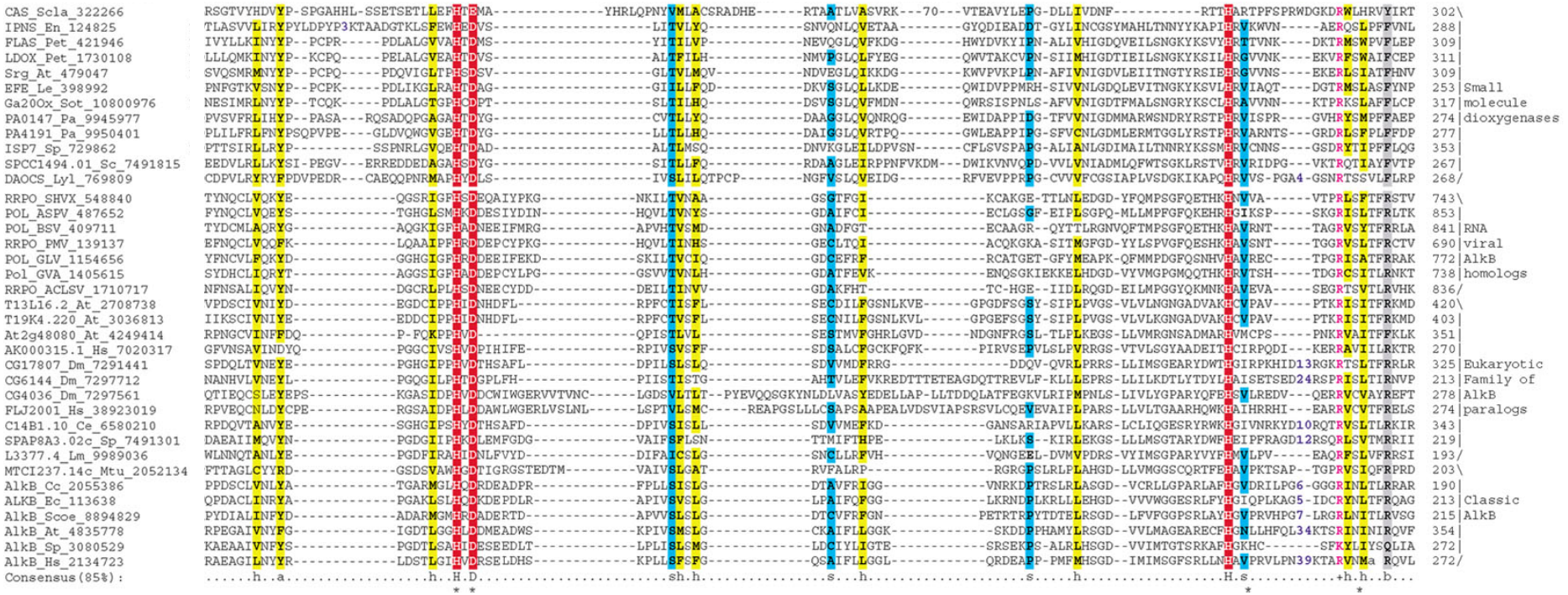
- Word length is W=11 by default
- Only identical words considered hits

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.

Back to the example...

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

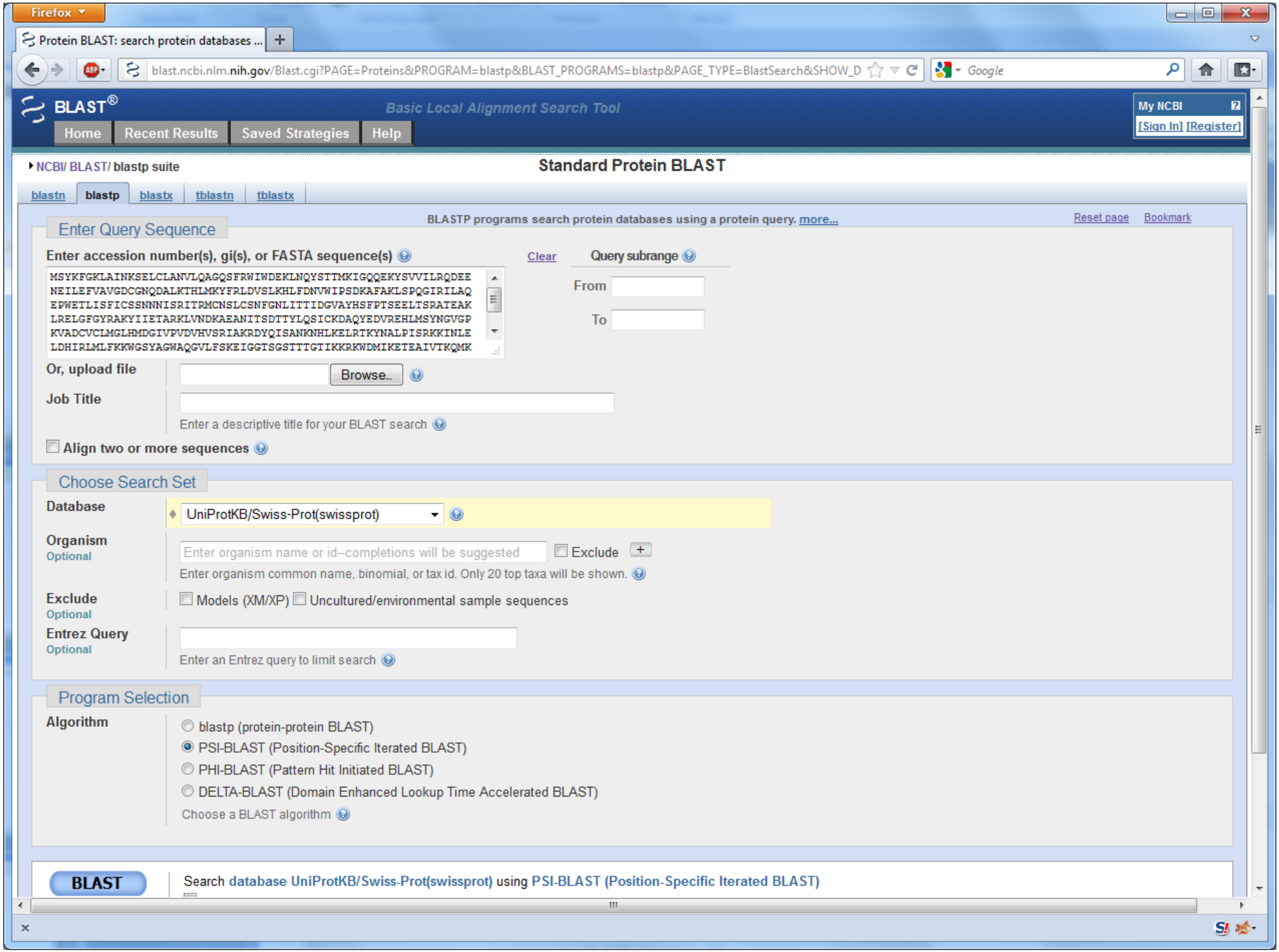


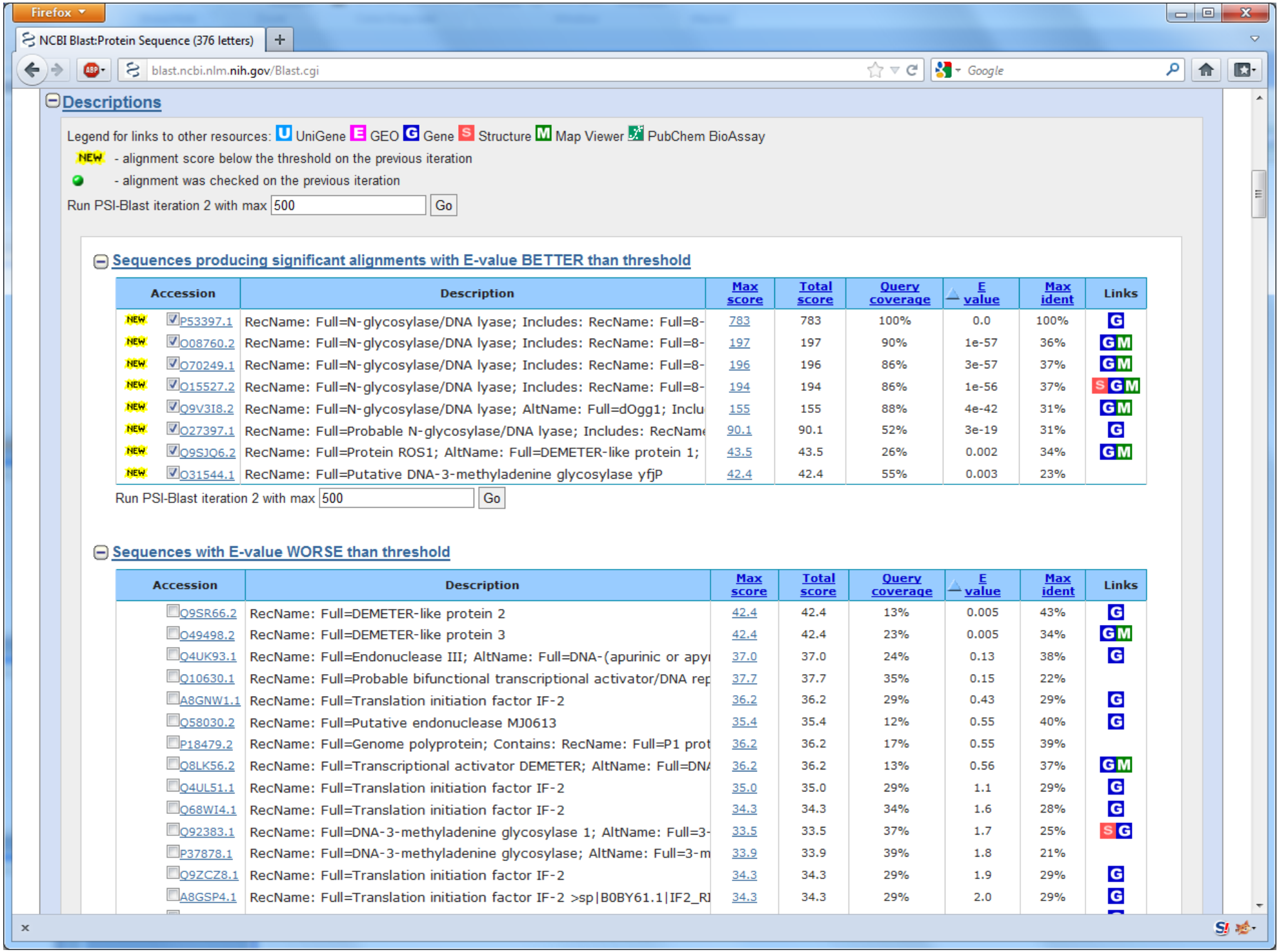
Excerpt from the AlkB paper

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 = 8×10^{-7}; *Drosophila* CG17807, iteration 3, e-value = 4×10^{-6}; papaya mosaic virus, iteration 3, e-value = 2×10^{-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


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
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein query. more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
MSYKFGKLAINKSELCLANVLQAGQSFRWIWDEKLNQYSTTMKIGQQEKYSVVILRQDEE
NEILEFVAVGDCGNQDALKTHLMKYFRLDVSLKHLFDNVWIPSDKAFAKLSPQGIRILAQ
EPWETLISFICSSNNNISRITRMCNSLCSNFGNLITTIDGVAYHSFPTSEELTSRATEAK
LRELGFGYRAKYIIETARKLVNDKAEANITSDTTYLQSICKDAQYEDVREHLMSYNGVGP
KVADCVCLMGLHMDGIVPVDVHVSRIAKRDYQISANKNHLKELRTKYNALPISRKKINLE
LDHIRLMLFKKWGSYAGWAQGVLFSKEIGGTSGSTTTGTIKKRKWDMIKETEAIVTKQMK
Or, upload file
Browse...
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
BLAST
Search database UniProtKB/Swiss-Prot(swissprot) using PSI-BLAST (Position-Specific Iterated BLAST)



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

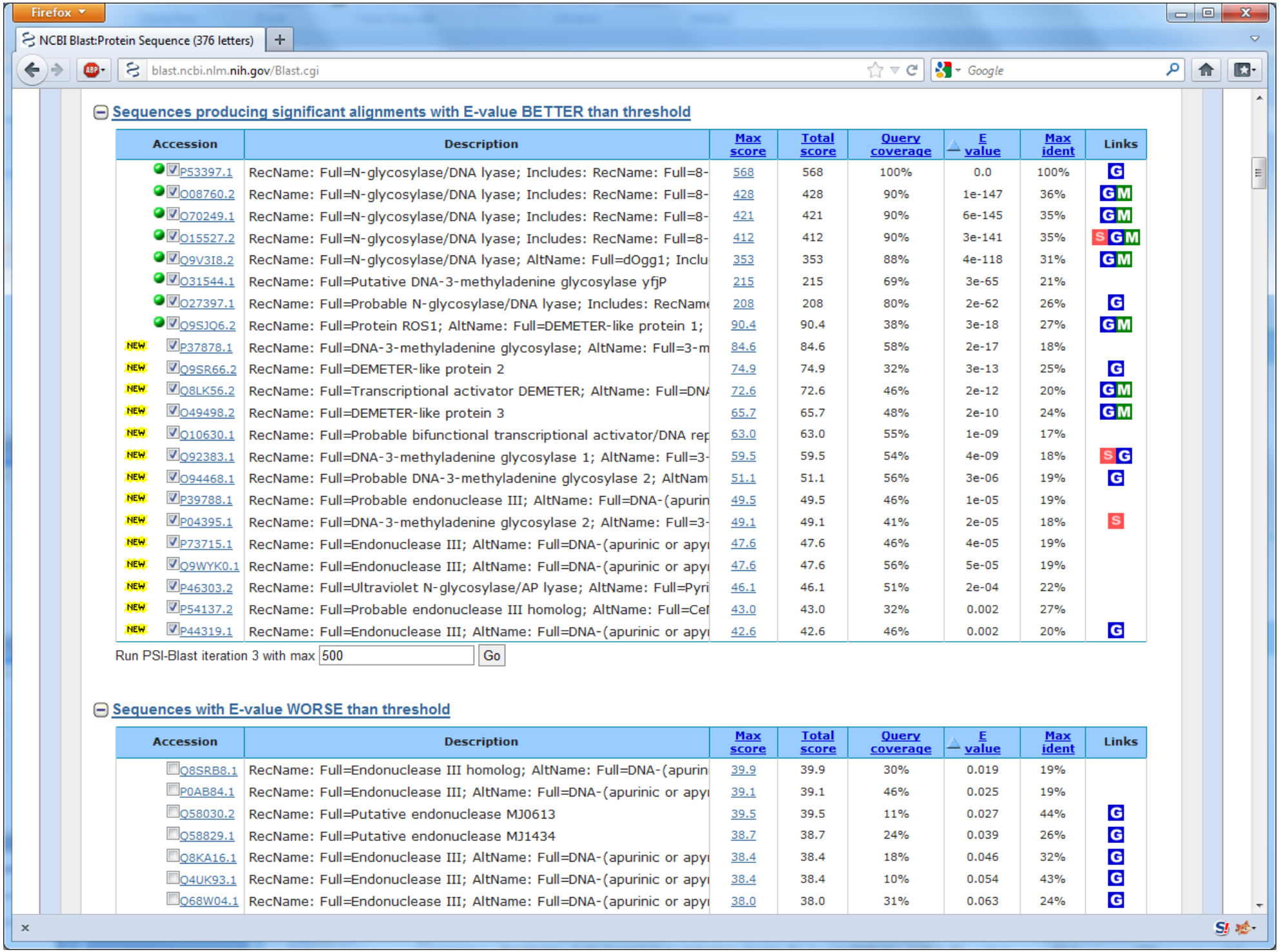
Sequences producing significant alignments with E-value BETTER than threshold

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
Q49498.2	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 apyr	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
Q58030.2	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=DNA	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



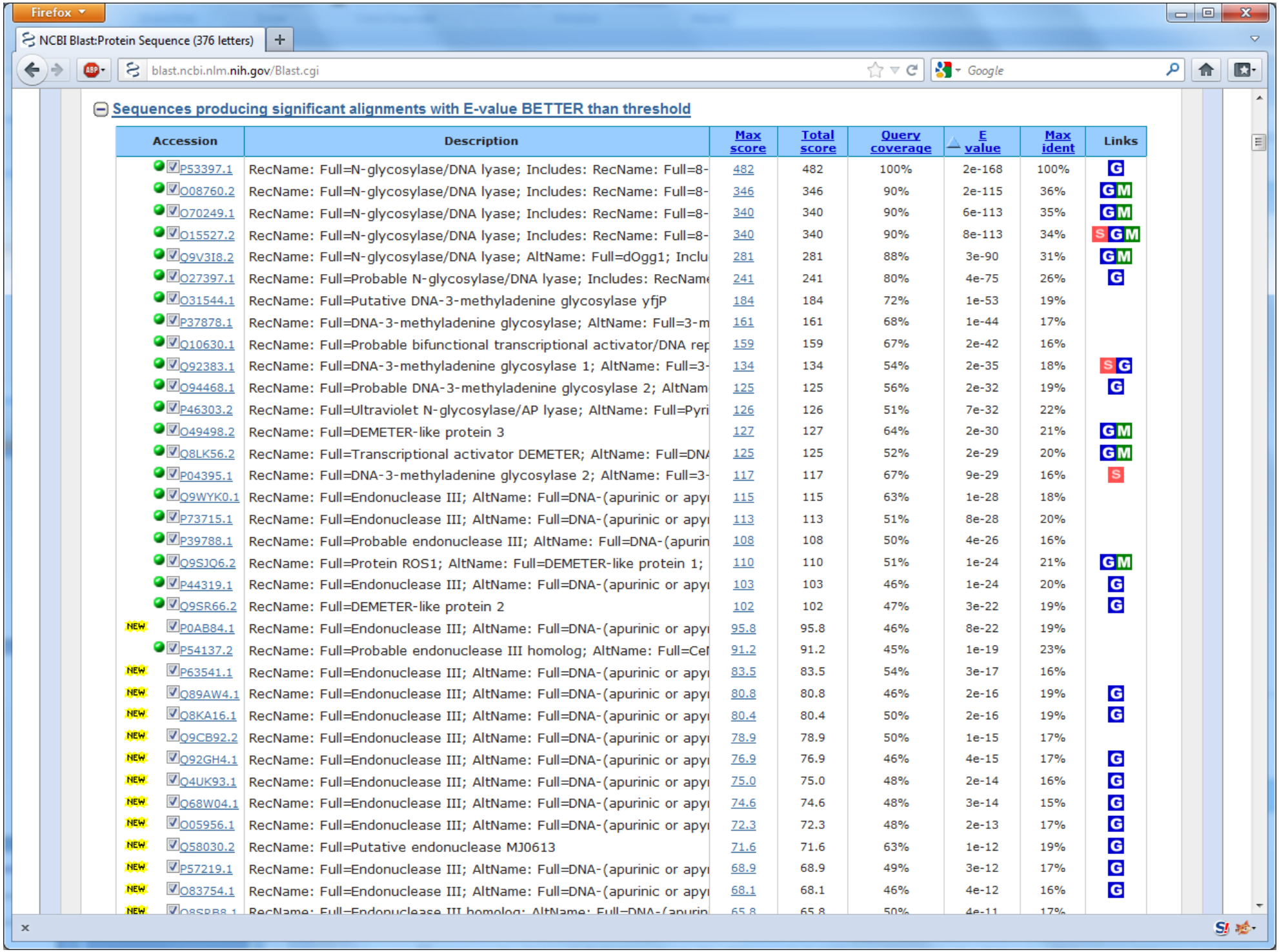
Sequences producing significant alignments with E-value BETTER than threshold

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%	G M
O70249.1	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	421	421	90%	6e-145	35%	G M
O15527.2	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	412	412	90%	3e-141	35%	S G M
Q9V3I8.2	RecName: Full=N-glycosylase/DNA lyase; AltName: Full=dOgg1; Inclu	353	353	88%	4e-118	31%	G M
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%	G M
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%	G M
NEW O49498.2	RecName: Full=DEMETER-like protein 3	65.7	65.7	48%	2e-10	24%	G M
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%	S G
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 apyr	47.6	47.6	46%	4e-05	19%	
NEW Q9WYK0.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	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 apyr	42.6	42.6	46%	0.002	20%	G

Run PSI-Blast iteration 3 with max 500 Go

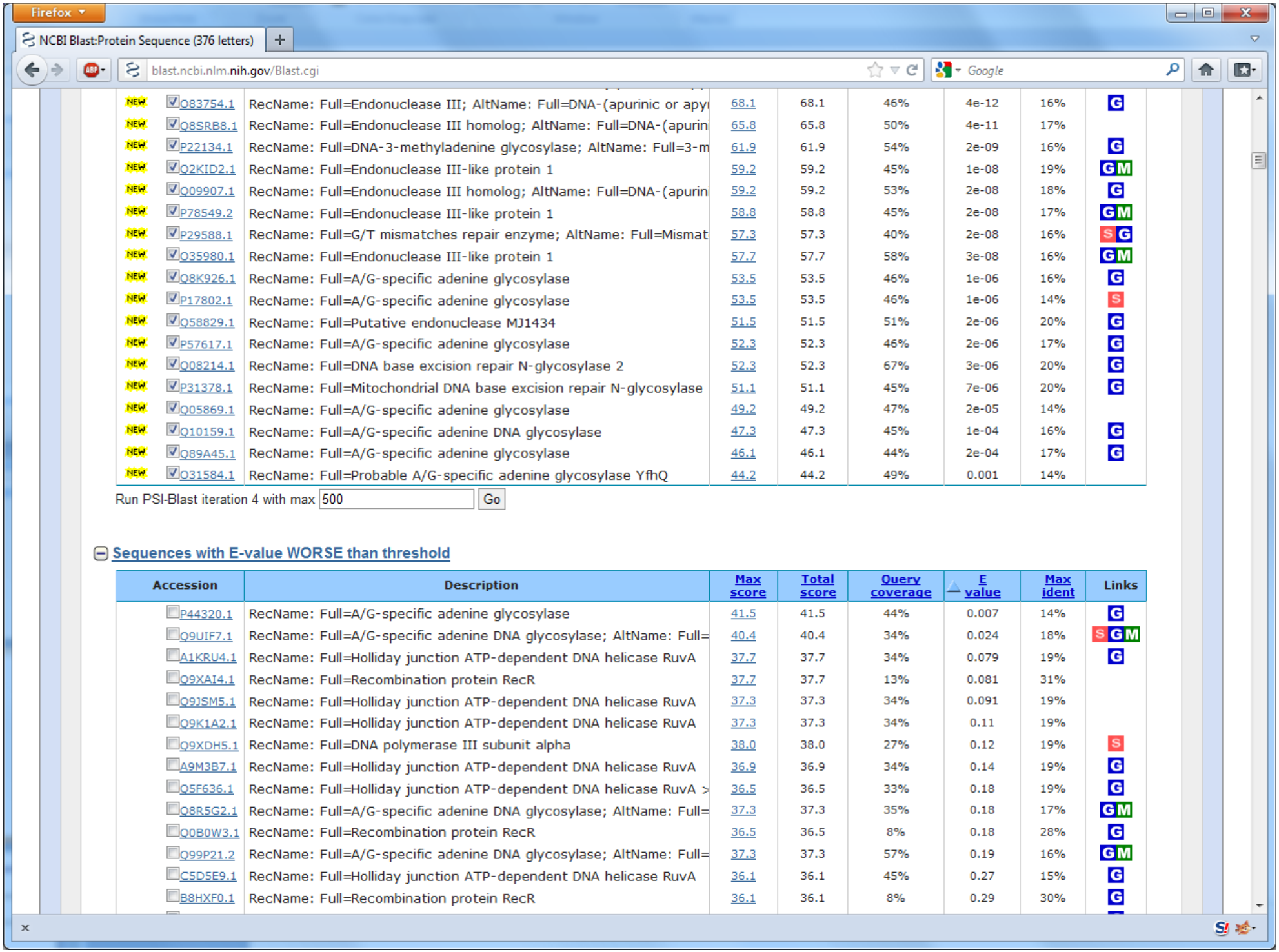
Sequences with E-value WORSE than threshold

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 apyr	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 apyr	38.4	38.4	18%	0.046	32%	G
Q4UK93.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	38.4	38.4	10%	0.054	43%	G
Q68W04.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	38.0	38.0	31%	0.063	24%	G



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=DNA	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 apyr	115	115	63%	1e-28	18%	
P73715.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	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 apyr	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 apyr	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 apyr	83.5	83.5	54%	3e-17	16%	
NEW Q89AW4.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	80.8	80.8	46%	2e-16	19%	G
NEW Q8KA16.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	80.4	80.4	50%	2e-16	19%	G
NEW Q9CB92.2	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	78.9	78.9	50%	1e-15	17%	
NEW Q92GH4.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	76.9	76.9	46%	4e-15	17%	G
NEW Q4UK93.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	75.0	75.0	48%	2e-14	16%	G
NEW Q68W04.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	74.6	74.6	48%	3e-14	15%	G
NEW O05956.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	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 apyr	68.9	68.9	49%	3e-12	17%	G
NEW O83754.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apyr	68.1	68.1	46%	4e-12	16%	G
NEW O8SRB8.1	RecName: Full=Endonuclease III homolog; AltName: Full=DNA-(apurin	65.8	65.8	50%	4e-11	17%	



NEW	Q83754.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	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:

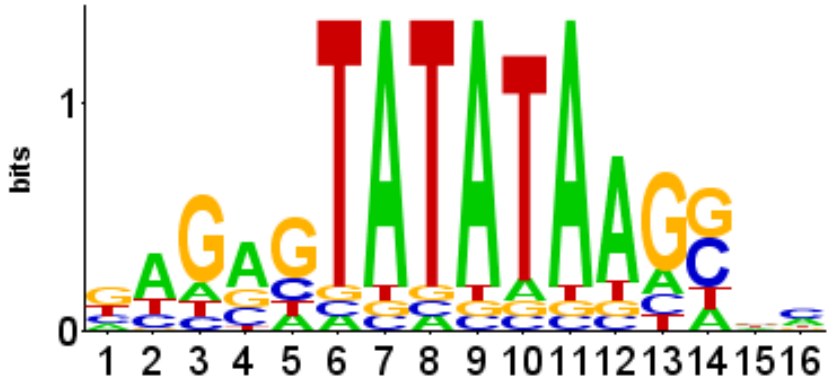
- Set of sequences
- Consensus sequence
- Patterns: Simplified ”regular expressions”
- Profiles: position-specific scoring matrices (PSSMs) based on probabilities of amino acid substitutions (Gribskov *et al.* 1987)
- Hidden Markov models (HMMs): probabilistic model for linear sequences (Haussler *et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

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

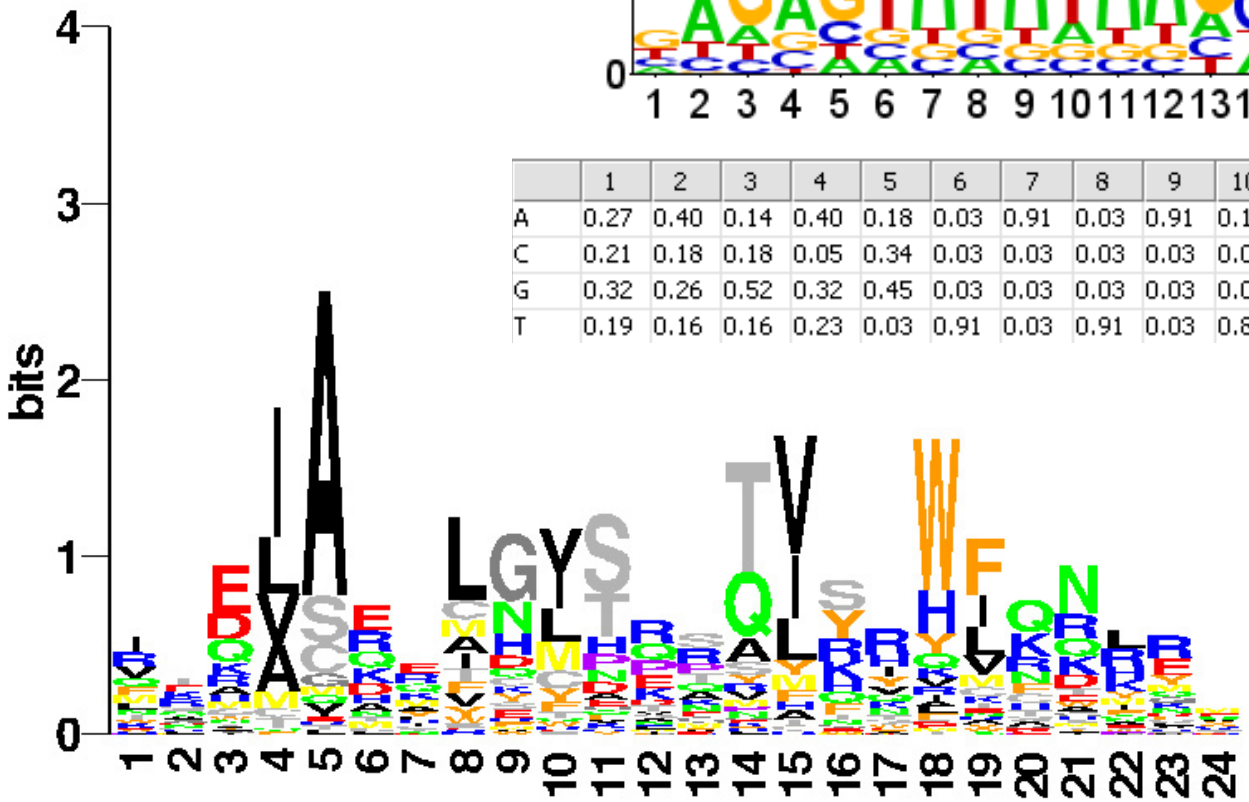
Sequence profiles (PSSMs)

- 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.

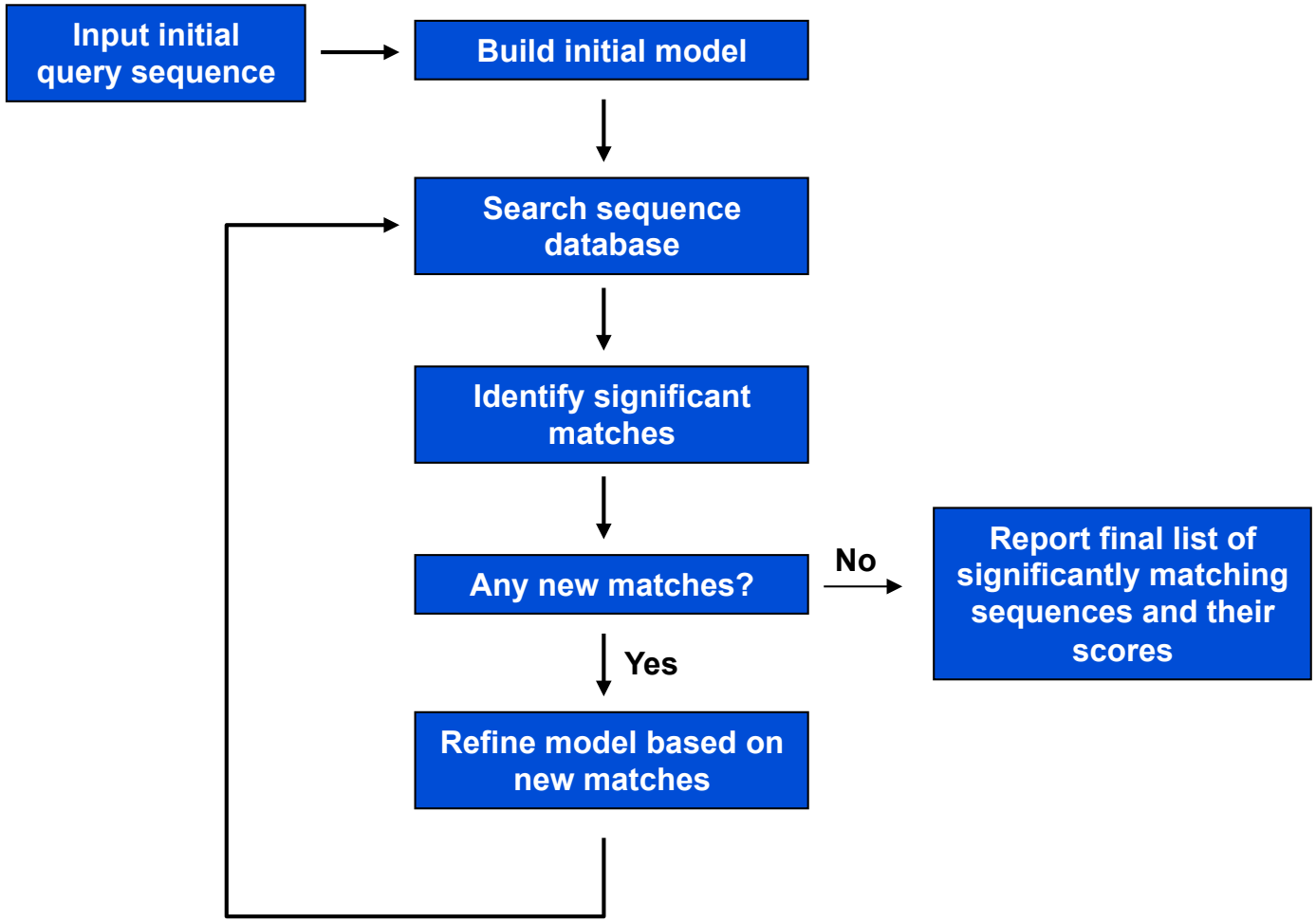
DNA and protein sequences logos



	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



BLAST online resources

- NCBI BLAST website
  http://www.ncbi.nlm.nih.gov/BLAST/
- NCBI tutorial on BLAST
  http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html
- NCBI Handbook, Chapter 16, BLAST
  http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch16
- BLAST FAQ
  http://www.ncbi.nlm.nih.gov/blast/blast_FAQs.shtml
- Wikipedia on BLAST
  http://en.wikipedia.org/wiki/BLAST

Literature

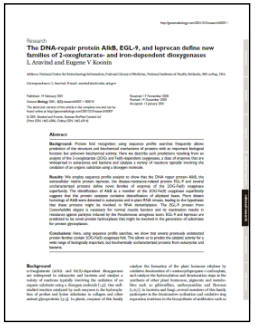
PSI-BLAST paper

- *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

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

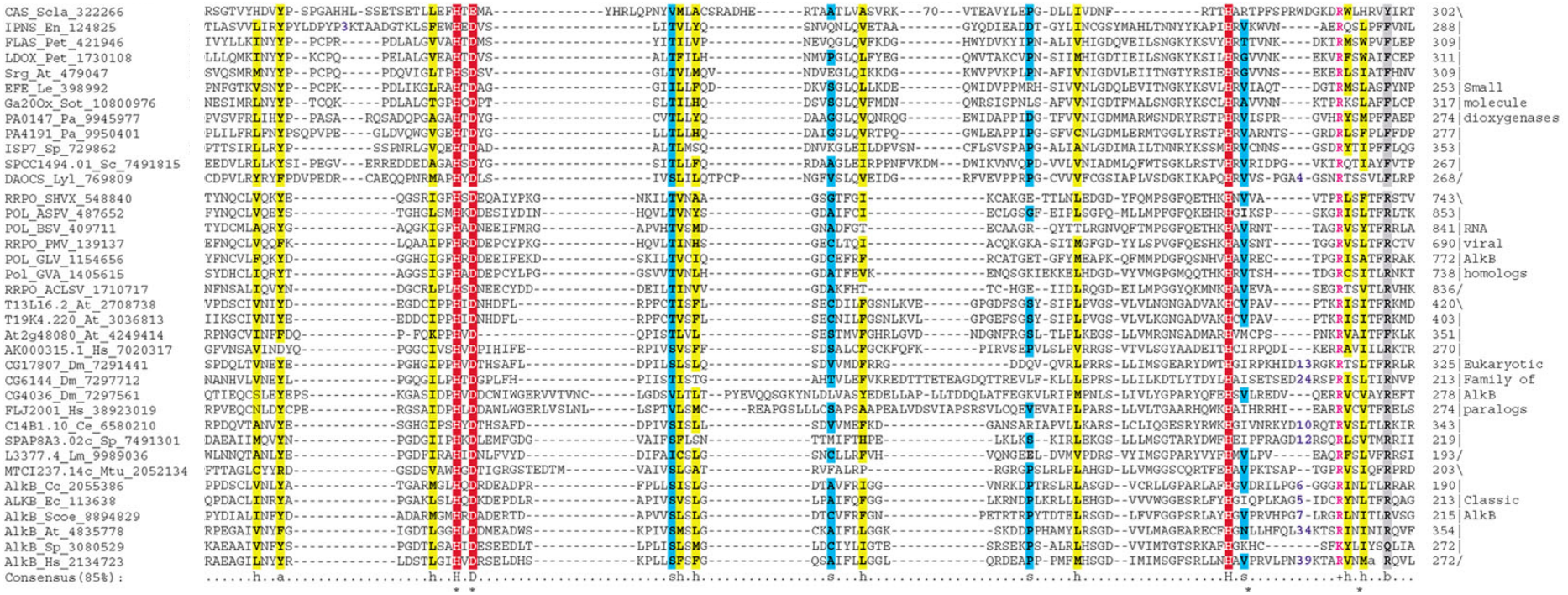

The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases

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
IPNS_En_124825
FLAS_Pet_421946
LDOX_Pet_1730108
Srg_At_479047
EFE_Le_398992
Ga20ox_Sot_10800976
PA0147_Pa_9945977
PA4191_Pa_9950401
ISP7_Sp_729862
SPCC1494.01_Sc_7491815
DAOCS_Lyl_769809
RRPO_SHVX_548840
POL_ASPV_487652
POL_BSV_409711
RRPO_PMV_139137
POL_GLV_1154656
Pol_GVA_1405615
RRPO_ACLSV_1710717
T13L16.2_At_2708738
T19K4.220_At_3036813
At2g48080_At_4249414
AK000315.1_Hs_7020317
CG17807_Dm_7291441
CG6144_Dm_7297712
CG4036_Dm_7297561
PLJ2001_Hs_38923019
C14B1.10_Ce_6580210
SPAP8A3.02c_Sp_7491301
L3377.4_Lm_9989036
MTCI237.14c_Mtu_2052134
AlkB_Cc_2055386
ALKB_Ec_113638
AlkB_Scoe_8894829
AlkB_At_4835778
AlkB_Sp_3080529
AlkB_Hs_2134723
Consensus(85%):
Small
molecule
dioxygenases
RNA
viral
homologs
Eukaryotic
Family of
paralogs
Classic
AlkB

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
- ...

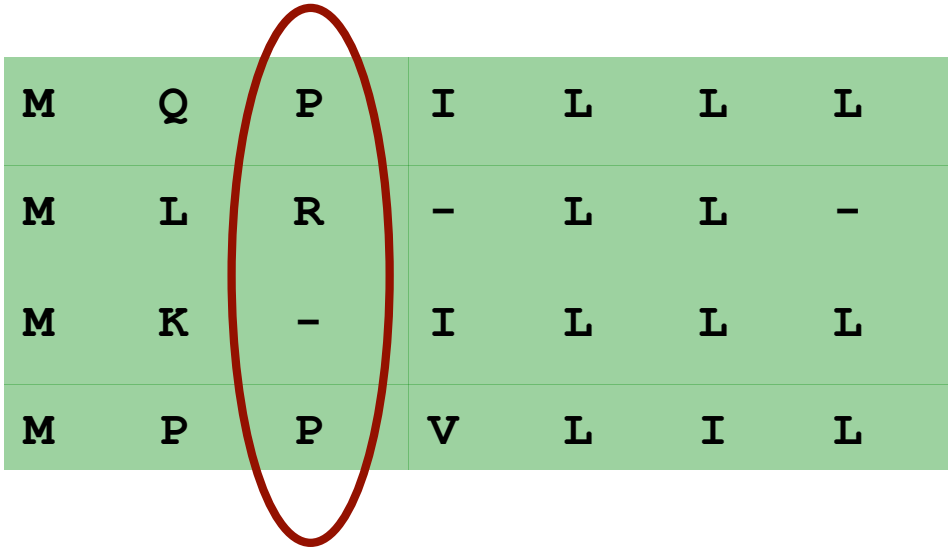
Finding the best multiple alignment

- 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

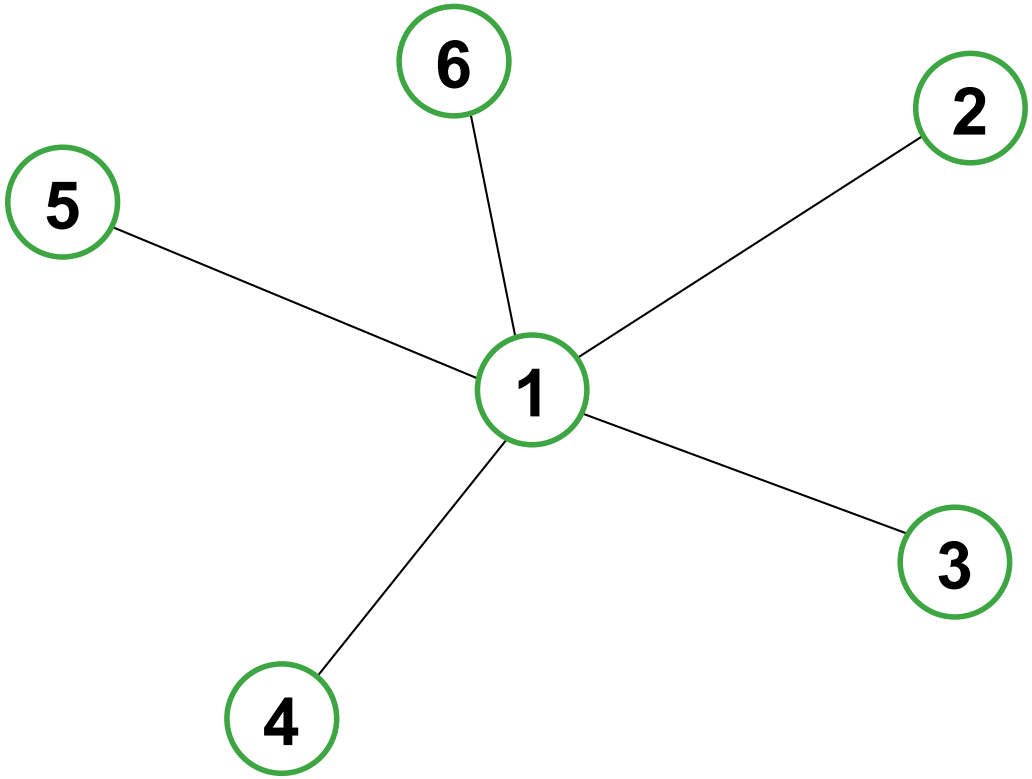
score(k) = S(P,R) + S(P,-) + S(P,P) + S(R,-) + S(R,P) + S(-,P)

score for column k = 3

We have S(-,-) = 0

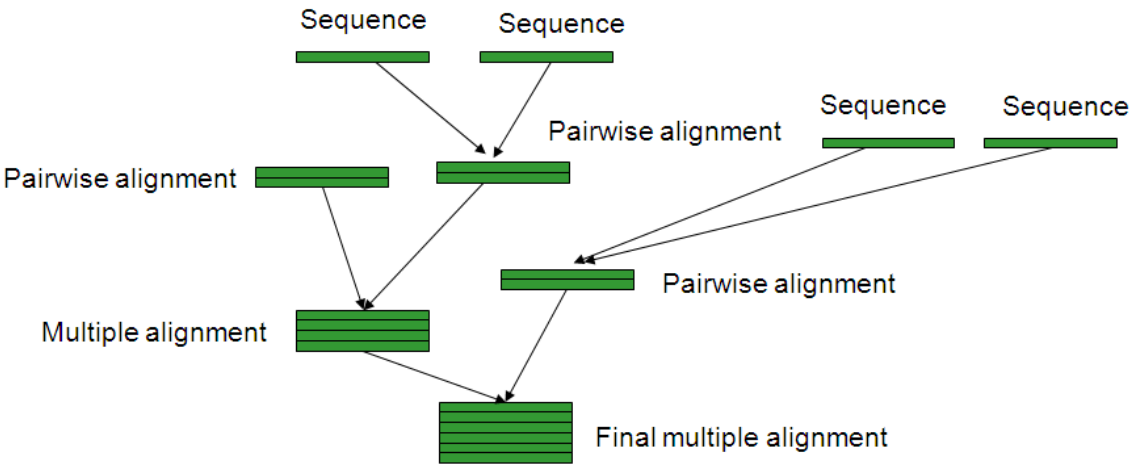
Total score = score(1) + score(2) + ... + score(N)

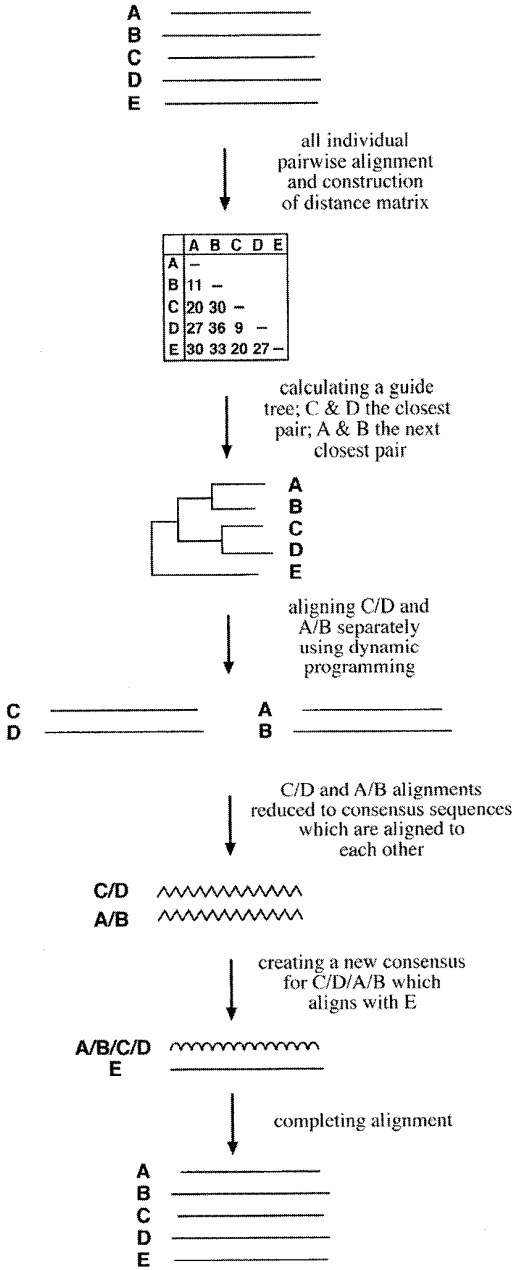
Centre star multiple alignment



Clustal W

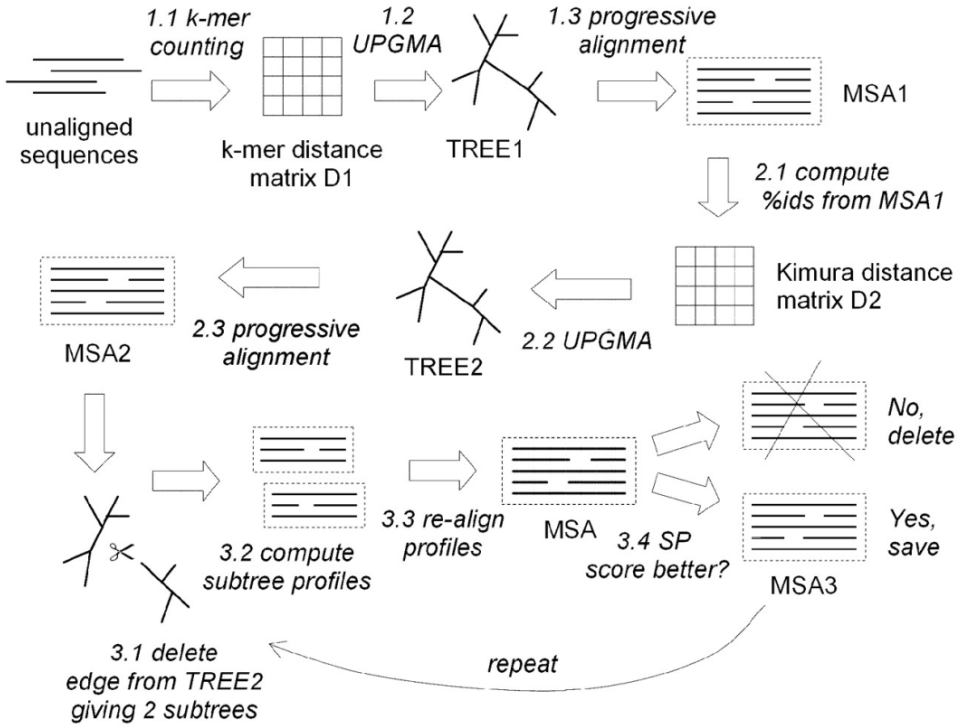
- One of the most commonly used and well-known tools for multiple sequence alignment. Now somewhat outdated and surpassed by other tools.
- Uses a progressive algorithm: Always starts with the most similar sequences and then aligns less similar sequences with each other.





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



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/.