

# BLAST for proteins, step 3

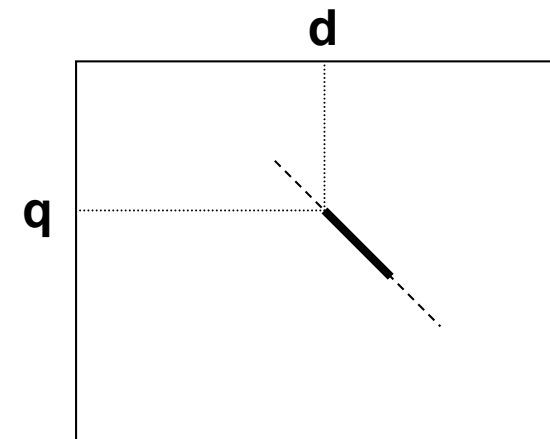
- Expand the hits into HSPs in both directions (no gaps) by adding score values from the substitution score matrix.
- In each direction, stop when the score decreases more than a threshold  $X$  from the highest score seen so far.

## 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 maximal, since further extension (with CA, CA) would result in a higher score (2.5).  $\triangle$



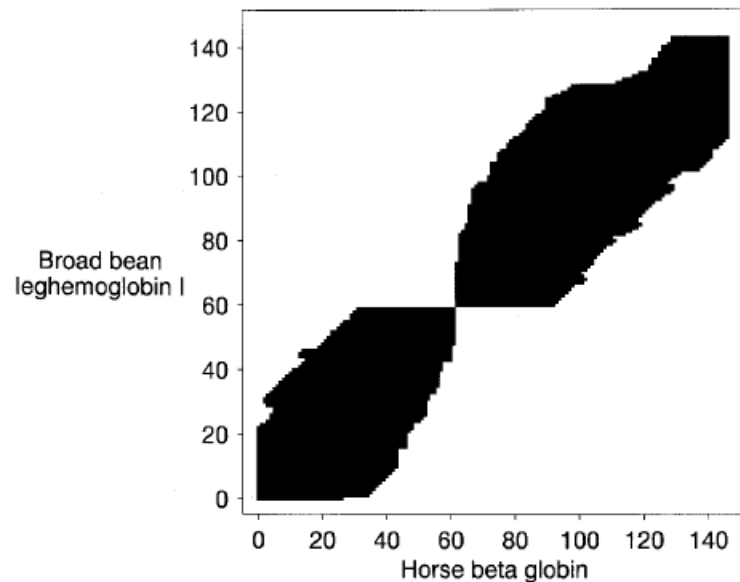
## **BLAST for proteins, step 4**

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

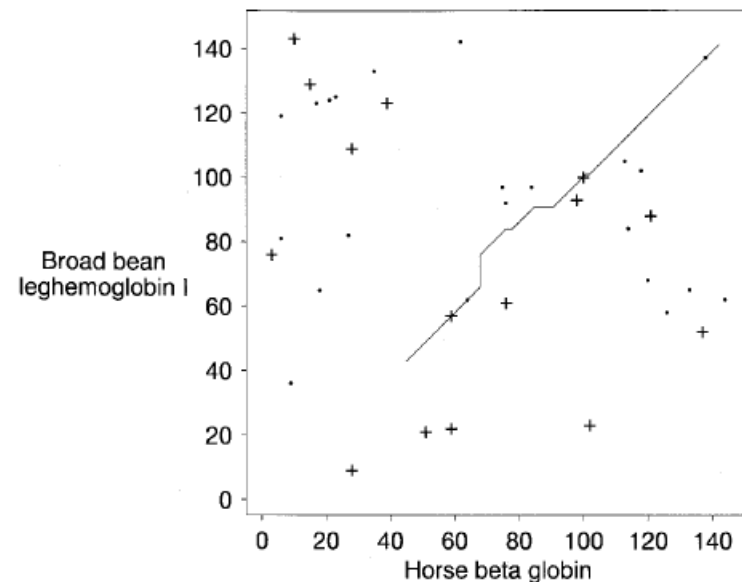
# BLAST for proteins, step 5

- 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

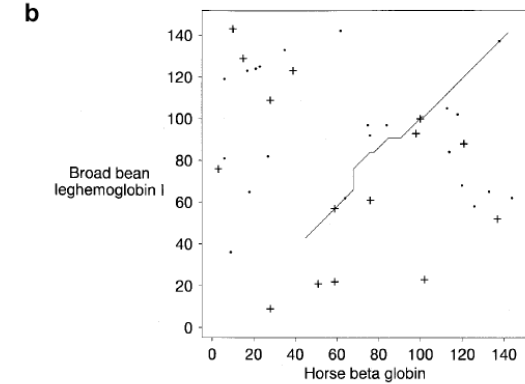
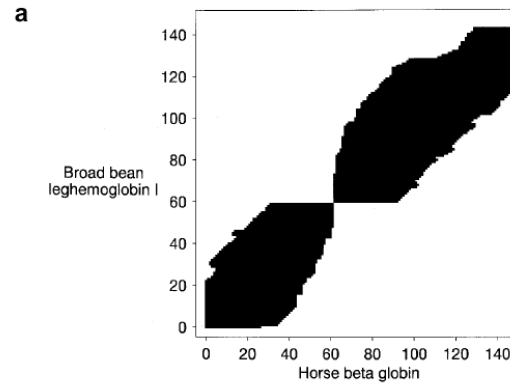
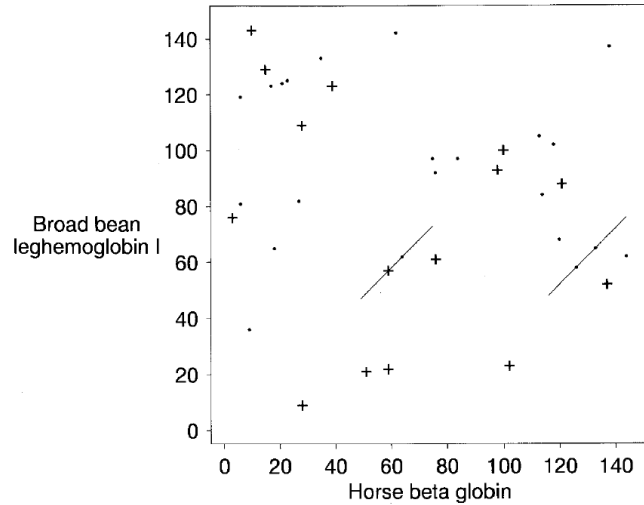
a



b



# BLAST example



## BLAST hits in the alignment

- + Hits with score  $\geq 13$
- Hits with score  $\geq 11$

## 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-----LHSPGEGVHHLNLDNLKGTFAALSE 90

Leghemoglobin 91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWAVAYDGLATAI 140
                  +H K +DP +F ++ L+ + G ++ EL A+++ G+A A+
Beta globin    91 LHCDKLHVDPENFRLLGNVLVVVVLARHFGKDFTPELQASYQKVVAGVANAL 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...

CAS_Sela_322266	RSQTVVHDVYYP-SPGAHHL-SSETSETLEFFHPEMA-----YHRLQPNYVMLACSRADHE----RTAATLVASVRK--70--VTEAVYLEEG-DLLIVDNF-----RTTHARTPFSPRWGDKDMLHRVYIIRT	302\
IPNS_En_124825	TLASVVLIRYPYLDYP3KTAADGTKLSFEWHDVDS-----LITVLYQ-----SNVQNLQVETAA-----GYQDI EADDT-GYLINCOSYMAHLTNNYKAPIHRKWNV----ABRQSLPFFVNL	288
FLAS_Pet_421946	IVYLLKINYP-PCPR----PDLALGVVAHDMS-----YITLIVP-----NEVQGLQVFKDG-----HWYDVKIEN-ALIVHIGDQVEILSNGKYKSVYHRTVTK----DKTRMSWPVLEP	309
LDOX_Pet_1730108	LLLQMKINYP-KCPQ----PELALGVVAHDMS-----ALTFILH-----NMVGLQLFYEG-----QWVTAKCVEN-SIIMHIGDTIEILSNGKYKSIHRGVNKK---EKVRFSAWVCEP	311
Srg_At_479047	SVQSMRMNYP-PCPQ----PDQVIGLTFHSDSV-----GLTVLMQV-----NDVEGLQIKKDG-----KWVPKPLEN-AFIVNIGDVLEIITNGYRSIHRGVNKK---EKERLSIATFHNV	309
EFE_Le_398992	PNFGTKVSNYP-PCPK----PDLIKGLRAHDADG-----GIILLFQD-----DKVGLQLKDE-----QWIDVPPMRH-SIVVNLGDQLEVIITNGYKSVIHRVIAQT---DGTFRMSLASFYNP	253
Ga200x_Sot_10800976	NESIMRLNYP-TQCK----PDLALGTGFHDPT-----SLTILHQ-----DSVGLQVFMFN-----QWRSISPNLS-AFVVNIGDTFMAISNGRYKSCIHRVAVNN---KTPRKSIAFFLCP	317
PA0147_Pa_9945977	PVSVFRLIHY-PASA---RQSADQPGAGAHDDYG-----CVTLLYQ-----DAAGGLQVQNRQG-----EWIDAPPIEG-FVFNVIGDMMARWSNDRYRSTPHRVISPR---GVHRYSMPPFAEP	274
PA4191_Pa_9950401	PLILFRLPNYPSQVPE-----GLDVQWVGGEHDYG-----LLTLLHQ-----DAIGLQVTRTPQ-----GWLEAPPIDG-SFVCNLGDMLEMTGGLYRSTPHRVARNTS---GRDRLSFLPFD	277
ISP7_Sp_729862	PTTSIRLLRY-P-----SSPNRLGVQHDAD-----ALTLMSQ-----DNVGLLEILDPVSN-----CFLSVSPARG-ALITANLGDIMAILTNNRYKSMHRVCNNS---GSDRYTIPFLQG	353
SPCC1494.01_Sc_7491815	EEDVLRLLKYSI-PEGV---ERREDDEAGAHSDYG-----SITLLFQ-----RDAAGLEIRPPNFVKDM---DWIKVNVQD-VVLVNIADMLQFPTSGKLRSTVHRVIDPG---VKTQRTIAYVTP	267
DAOCS_Ly1_769809	CDPVLRYRYPDVPEDR---CAEQQPNRMAFHDLS-----IVSLILQTPCP-----NGFVSLQVEIDG-----RFVEVPPFG-CVVVFCGSIAPLVSDGKIKAPQHRVVS-PGA4-GSNRTSIVLFLRP	268/
RRPO_SHVX_548840	TYNQCLVQKYE-----QGSRIGFHSDQAIYPKG-----NKILTVAIA-----GSGTFGI-----KCAKGE-TTLNLEDGD-YFQMPSGFQETHRKAIVA----VTPRLSFTFRSTV	743\
POL_ASPV_487652	FYNQCLVQEYS-----TGHGLSMHRDDES IYDIN-----HQVLTVNSY-----GDAIFCI-----ECLGSEF-EIPLSGPQ-MLMPPGFQKEHRHGKSP----SKGRISLTFRLLTK	853
POL_BSV_409711	TYDCMLAQRYG-----AQQKIGFHADNEE IFMRG-----APVHTVSM-----GNADFGT-----ECAAGR-QYTLRGNVQFTMPSGFQETHRKAIVRNT---TAGRVSYTFRRLA	841
RRPO_PMV_139137	EFNQCLVQCFK-----LQAAIPFHDDDEPCYPKG-----HQVLTINHS-----GECILCI-----ACQKGA-SITMGFGD-YYLSPVGFQESHKHAIVSNT---TGGRVSLTFRCTV	690
POL_GLV_1154656	YFNCVLFQKYD-----GGHGIGFHEDDEE IPEKD-----SKILTVCIQ-----GDEEFF-----RCATGET-GPYMEAPK-QFMMPDGFQSNHVAVREC---TPGRISATFRRAK	772
PoL_GVA_1405615	SYDHCLIQRYT-----AGSGIGFHADDEPCYLPF-----GSVVTVNLH-----GDATFEVK-----ENQSGKIEKELHDGD-YVVMGPGMQQTHRHRVTSH---TDGRCSITLNRKT	738
RRPO_ACLS_V_1710717	NFNSALIQVYN-----DGCRLLPESDNEE CYDD-----DEILTINVV-----DKAKFHT-----TC-HGE-IDLRLQGD-EI LMPGGYQKMKRKAIVEVA---SEGRTSVTLRVHK	836/
T13L16.2_At_2708738	VPSDCIVNIYD-----EGDCIIPFHDDNHDFL-----RPFCTISFL-----SECDIILFGSNLKV-----GPGDFSGY-SIFLPVGS-VLVLNGGADVAKHCVPAV---PTKRISITFRKMD	420\
T19K4.220_At_3036813	IIKSCIVNIYE-----EDDCIIPFHDDNHDFL-----RPFCTVSL-----SECNILFGSNLKV-----GPGDFSGY-SIFLPVGS-VLVLNGGADVAKHCVPAV---PTKRISITFRKMD	403
At2g48080_At_4249414	RPNGCVINFDQ-----P-FQKPPHYD-----QPISLVL-----SEBTMVFGHRLGDV-----NDGNFRGSL-TLFLKEGS-LLVMRGNADMARHVMCPS---PNKRVAITFFPKL	351
AK000315.1_Hs_7020317	GFVNSAVINDYQ-----PGGCIVSHVDPIHIFE-----RPIVSVSFP-----SDBALCFGCKFPQK-----PIRVSEVLSLFPVRRGS-VTVLSGYADEIITHCIRPDI---KERAVIILRTR	270
CG17807_Dm_7291441	SPDQLTVNEYE-----PDGHGIPFHDDTHSAFL-----SDVVMDFRRG-----DDQV-QVRLPRRS-LLVMSGEARYDWHGIRPKHID13RGRKSLTFRRLR	325
CG6144_Dm_7297712	NANHVLVNEYL-----PGQGLPFDGDLPH-----PIIISTISTG-----AHVVEVFKREDTTTETEAGDQTTREVLV-KLLLEPRS-LLILKDTLYTDYDPAISETSED24RSPRISLITRNVP	213
CG4036_Dm_7297561	QTIEQCSLEYEPS-----KGASIDPHVDWCWIGERVVTVNC-----LGDVSLTET-----PYEQQSGKYNLDLVASYEDLLAP-LLTDDQLATFEGKVLRIEMPNSL-LIVLYGPARYQFBSVLRREDV---QRRKVCVAREFT	278
PLJ2001_Hs_38923019	RPVQCNLDYCPE-----RGSADPHDDAMLWGERLVSLNL-----LSPTVLSMC-----REAPGSLLLCASPAAPEALVDSVIAPSRSVLCQEVVAIPLPARS-LLVLTGAARHQWKAHHRHHI---EARRVCVTFRELS	274
C14B1.10_Ce_6580210	RPDQVANVYE-----SGHGIPSHDDTHSAFD-----DPIVSLSL-----SDVVMDFKD-----GANSARIAPVLLKARS-LCLIQGESRYRWHGIVNRYD10RQTVSLTLRKR	343
SPAP8A3.02c_Sp_7491301	DAEAIIMQVYN-----PGDGIIPKRDLEMFGDG-----VAIFSPLSN-----LKLKQ-----KIRLEKGS-LLLMSGTARYDWPBEIPFRAGD12RSQRLSVTMRRII	219
L3377.4_Lm_9989036	WLNQTLANLYE-----PGDFIRAHDDNLFVYD-----DIFATCSL-----SNCLLRFVH-----VQNGEEL-DVMVPDRS-VYIMSGPARYVYRHWLVP---EAQRFSLVFRRSI	193/
MTC1237.14c_Mtu_2052134	FTTAGLCYRD-----GSDSVAMHDDTIGRSTEDTM-----VAIVSLGAT-----RVFALRP-----RGRGFSRLRFLAHGD-LLVMGSGCQRTFBEHAVPKTSAP---TGPRVSIQFRPRD	203\
AlkB_Cc_2055386	PPDSCLVNLXA-----TGARMLGHODRDEADPR-----FPLLSTSLG-----DTAVFRIGG-----VNRKDETRSLASGD-VCRLLGPARLARHGVDRILPG6-GGGRINLTLRRAR	190
AlkB_Ec_113638	QPDACLINRYA-----LPAKLSLGHODKDEPDLR-----APIVSVSLG-----LKRNDLKRLLLEHGD-VVVWGGESRLFYHGIQPLKAG5-LDCRYNLTFRQAG	213
AlkB_Scoe_8894829	PYDIALINFDG-----ADARMGSRDDADERTD-----APVSVSLG-----DTCVFRFNG-----PETRTFYDTELRSGD-LFVFGGSPRLAYHGVPRVHPG7-LRGLRNLTLRVSG	215
AlkB_At_4835778	RKEGAIVNYFG-----IGDTLSGHDDMEADWS-----KPIVSMSLG-----CKALFLLHGGK-----SKDDP PHAMYLRSGD-VVLMAGEARECPHNLHFLQ134TKSRIMINIRQVF	354
AlkB_Sp_3080529	KBAEAINVNFYS-----PGDTLSAHDDSEEDLT-----LPLISLMSG-----LDCIYLIGTE-----SRSEKFS-ALRHLHSGD-VVIMTGTSRKAPHGKHC---SKFYLIYSQLIA	272
AlkB_Hs_2134723	REBAGILNRYR-----LDSTLGIHVDSELDHS-----KPLISPSFG-----QSAIFLLGGL-----QRDEAR-PMFMHSGD-IMIMSGFSRLNLHAPRVLPN39KTAIVNMA ROVL	272/
Consensus (85%) :	.....h.h.a.....*.....h.H.D.....sh.h.....s.....s.....h.....h.s.....*.....h.h.....*.....h.h.....*	

# 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 =  $8 \times 10^{-7}$ ; *Drosophila* CG17807, iteration 3, e-value =  $4 \times 10^{-6}$ ; papaya mosaic virus, iteration 3, e-value =  $2 \times 10^{-4}$ . 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 sequence encoding the protein YbiX, which is

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

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](#)

```
MSYKFGKLAINKSELCLANVLQAGQSFRWIWDEKLNQYSTTMKIGQQEKYSVVILRQDEE
NEILEFVAVGDCGNQDALKTHLMKYFRLDVSLKHLFDNVWIPSDKAFKLSPOGIRILAQ
EPWETLISFICSSNNNISRITRMCNSLCSNFGNLITIDGVAYHSFPTSEELTSRATEAK
LRELGFYRAKYIIEETARKLVNDKAEANITSDTTYLQSI CKDAQYEDVREHLMSYNGVGP
KVADCVCLMGLHMDGIVPVDVHVHSRIAKRDYQISANKNHLKELRTKYNALPISRKKINLE
LDHIRLMLFKKWSYAGWAQGVLFKSKEIGGTSGSTTTGT IKKRKWDMIKETEAIIVTKQMK
```

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   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)

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 [P](#) 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

#### Sequences producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">P53397.1</a>	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">783</a>	783	100%	0.0	100%	<a href="#">G</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q08760.2</a>	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">197</a>	197	90%	1e-57	36%	<a href="#">GM</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q70249.1</a>	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">196</a>	196	86%	3e-57	37%	<a href="#">GM</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q15527.2</a>	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">194</a>	194	86%	1e-56	37%	<a href="#">S</a> <a href="#">GM</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q9V318.2</a>	RecName: Full=N-glycosylase/DNA lyase; AltName: Full=dOgg1; Inclu	<a href="#">155</a>	155	88%	4e-42	31%	<a href="#">GM</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q27397.1</a>	RecName: Full=Probable N-glycosylase/DNA lyase; Includes: RecName	<a href="#">90.1</a>	90.1	52%	3e-19	31%	<a href="#">G</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q9SJO6.2</a>	RecName: Full=Protein ROS1; AltName: Full=DEMETER-like protein 1;	<a href="#">43.5</a>	43.5	26%	0.002	34%	<a href="#">GM</a>
<b>NEW</b> <input checked="" type="checkbox"/> <a href="#">Q31544.1</a>	RecName: Full=Putative DNA-3-methyladenine glycosylase yfjP	<a href="#">42.4</a>	42.4	55%	0.003	23%	

Run PSI-Blast iteration 2 with max

#### Sequences with E-value WORSE than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
<input type="checkbox"/> <a href="#">Q9SR66.2</a>	RecName: Full=DEMETER-like protein 2	<a href="#">42.4</a>	42.4	13%	0.005	43%	<a href="#">G</a>
<input type="checkbox"/> <a href="#">Q49498.2</a>	RecName: Full=DEMETER-like protein 3	<a href="#">42.4</a>	42.4	23%	0.005	34%	<a href="#">GM</a>
<input type="checkbox"/> <a href="#">Q4UK93.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">37.0</a>	37.0	24%	0.13	38%	<a href="#">G</a>
<input type="checkbox"/> <a href="#">Q10630.1</a>	RecName: Full=Probable bifunctional transcriptional activator/DNA rep	<a href="#">37.7</a>	37.7	35%	0.15	22%	
<input type="checkbox"/> <a href="#">A8GNW1.1</a>	RecName: Full=Translation initiation factor IF-2	<a href="#">36.2</a>	36.2	29%	0.43	29%	<a href="#">G</a>
<input type="checkbox"/> <a href="#">Q58030.2</a>	RecName: Full=Putative endonuclease MJ0613	<a href="#">35.4</a>	35.4	12%	0.55	40%	<a href="#">G</a>
<input type="checkbox"/> <a href="#">P18479.2</a>	RecName: Full=Genome polyprotein; Contains: RecName: Full=P1 prot	<a href="#">36.2</a>	36.2	17%	0.55	39%	
<input type="checkbox"/> <a href="#">Q8LK56.2</a>	RecName: Full=Transcriptional activator DEMETER; AltName: Full=DNA	<a href="#">36.2</a>	36.2	13%	0.56	37%	<a href="#">GM</a>
<input type="checkbox"/> <a href="#">Q4UL51.1</a>	RecName: Full=Translation initiation factor IF-2	<a href="#">35.0</a>	35.0	29%	1.1	29%	<a href="#">G</a>
<input type="checkbox"/> <a href="#">Q68WI4.1</a>	RecName: Full=Translation initiation factor IF-2	<a href="#">34.3</a>	34.3	34%	1.6	28%	<a href="#">G</a>
<input type="checkbox"/> <a href="#">Q92383.1</a>	RecName: Full=DNA-3-methyladenine glycosylase 1; AltName: Full=3-	<a href="#">33.5</a>	33.5	37%	1.7	25%	<a href="#">S</a> <a href="#">G</a>
<input type="checkbox"/> <a href="#">P37878.1</a>	RecName: Full=DNA-3-methyladenine glycosylase; AltName: Full=3-m	<a href="#">33.9</a>	33.9	39%	1.8	21%	
<input type="checkbox"/> <a href="#">Q9ZCZ8.1</a>	RecName: Full=Translation initiation factor IF-2	<a href="#">34.3</a>	34.3	29%	1.9	29%	<a href="#">G</a>
<input type="checkbox"/> <a href="#">A8GSP4.1</a>	RecName: Full=Translation initiation factor IF-2 >sp B0BY61.1 IF2_R1	<a href="#">34.3</a>	34.3	29%	2.0	29%	<a href="#">G</a>

**Sequences producing significant alignments with E-value BETTER than threshold**

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
<input checked="" type="checkbox"/> P53397.1	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">568</a>	568	100%	0.0	100%	<a href="#">G</a>
<input checked="" type="checkbox"/> O08760.2	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">428</a>	428	90%	1e-147	36%	<a href="#">GM</a>
<input checked="" type="checkbox"/> Q70249.1	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">421</a>	421	90%	6e-145	35%	<a href="#">GM</a>
<input checked="" type="checkbox"/> O15527.2	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">412</a>	412	90%	3e-141	35%	<a href="#">SGM</a>
<input checked="" type="checkbox"/> Q9V3I8.2	RecName: Full=N-glycosylase/DNA lyase; AltName: Full=dOgg1; Inclu	<a href="#">353</a>	353	88%	4e-118	31%	<a href="#">GM</a>
<input checked="" type="checkbox"/> O31544.1	RecName: Full=Putative DNA-3-methyladenine glycosylase yfjP	<a href="#">215</a>	215	69%	3e-65	21%	
<input checked="" type="checkbox"/> Q27397.1	RecName: Full=Probable N-glycosylase/DNA lyase; Includes: RecName	<a href="#">208</a>	208	80%	2e-62	26%	<a href="#">G</a>
<input checked="" type="checkbox"/> Q9S1Q6.2	RecName: Full=Protein ROS1; AltName: Full=DEMETER-like protein 1;	<a href="#">90.4</a>	90.4	38%	3e-18	27%	<a href="#">GM</a>
NEW <input checked="" type="checkbox"/> P37878.1	RecName: Full=DNA-3-methyladenine glycosylase; AltName: Full=3-m	<a href="#">84.6</a>	84.6	58%	2e-17	18%	
NEW <input checked="" type="checkbox"/> Q9SR66.2	RecName: Full=DEMETER-like protein 2	<a href="#">74.9</a>	74.9	32%	3e-13	25%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> Q8LK56.2	RecName: Full=Transcriptional activator DEMETER; AltName: Full=DNA	<a href="#">72.6</a>	72.6	46%	2e-12	20%	<a href="#">GM</a>
NEW <input checked="" type="checkbox"/> Q49498.2	RecName: Full=DEMETER-like protein 3	<a href="#">65.7</a>	65.7	48%	2e-10	24%	<a href="#">GM</a>
NEW <input checked="" type="checkbox"/> Q10630.1	RecName: Full=Probable bifunctional transcriptional activator/DNA rep	<a href="#">63.0</a>	63.0	55%	1e-09	17%	
NEW <input checked="" type="checkbox"/> Q92383.1	RecName: Full=DNA-3-methyladenine glycosylase 1; AltName: Full=3-	<a href="#">59.5</a>	59.5	54%	4e-09	18%	<a href="#">SG</a>
NEW <input checked="" type="checkbox"/> Q94468.1	RecName: Full=Probable DNA-3-methyladenine glycosylase 2; AltNam	<a href="#">51.1</a>	51.1	56%	3e-06	19%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> P39788.1	RecName: Full=Probable endonuclease III; AltName: Full=DNA-(apurin	<a href="#">49.5</a>	49.5	46%	1e-05	19%	
NEW <input checked="" type="checkbox"/> P04395.1	RecName: Full=DNA-3-methyladenine glycosylase 2; AltName: Full=3-	<a href="#">49.1</a>	49.1	41%	2e-05	18%	<a href="#">S</a>
NEW <input checked="" type="checkbox"/> P73715.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">47.6</a>	47.6	46%	4e-05	19%	
NEW <input checked="" type="checkbox"/> Q9WYK0.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">47.6</a>	47.6	56%	5e-05	19%	
NEW <input checked="" type="checkbox"/> P46303.2	RecName: Full=Ultraviolet N-glycosylase/AP lyase; AltName: Full=Pyri	<a href="#">46.1</a>	46.1	51%	2e-04	22%	
NEW <input checked="" type="checkbox"/> P54137.2	RecName: Full=Probable endonuclease III homolog; AltName: Full=Cef	<a href="#">43.0</a>	43.0	32%	0.002	27%	
NEW <input checked="" type="checkbox"/> P44319.1	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">42.6</a>	42.6	46%	0.002	20%	<a href="#">G</a>

Run PSI-Blast iteration 3 with max

**Sequences with E-value WORSE than threshold**

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

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

Sequences producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
<input checked="" type="checkbox"/> <a href="#">P53397.1</a>	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">482</a>	482	100%	2e-168	100%	<a href="#">G</a>
<input checked="" type="checkbox"/> <a href="#">Q08760.2</a>	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">346</a>	346	90%	2e-115	36%	<a href="#">GM</a>
<input checked="" type="checkbox"/> <a href="#">Q70249.1</a>	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">340</a>	340	90%	6e-113	35%	<a href="#">GM</a>
<input checked="" type="checkbox"/> <a href="#">O15527.2</a>	RecName: Full=N-glycosylase/DNA lyase; Includes: RecName: Full=8-	<a href="#">340</a>	340	90%	8e-113	34%	<a href="#">SGM</a>
<input checked="" type="checkbox"/> <a href="#">Q9V318.2</a>	RecName: Full=N-glycosylase/DNA lyase; AltName: Full=dOgg1; Inclu	<a href="#">281</a>	281	88%	3e-90	31%	<a href="#">GM</a>
<input checked="" type="checkbox"/> <a href="#">Q27397.1</a>	RecName: Full=Probable N-glycosylase/DNA lyase; Includes: RecName	<a href="#">241</a>	241	80%	4e-75	26%	<a href="#">G</a>
<input checked="" type="checkbox"/> <a href="#">Q31544.1</a>	RecName: Full=Putative DNA-3-methyladenine glycosylase yfjP	<a href="#">184</a>	184	72%	1e-53	19%	
<input checked="" type="checkbox"/> <a href="#">P37878.1</a>	RecName: Full=DNA-3-methyladenine glycosylase; AltName: Full=3-m	<a href="#">161</a>	161	68%	1e-44	17%	
<input checked="" type="checkbox"/> <a href="#">Q10630.1</a>	RecName: Full=Probable bifunctional transcriptional activator/DNA rep	<a href="#">159</a>	159	67%	2e-42	16%	
<input checked="" type="checkbox"/> <a href="#">Q92383.1</a>	RecName: Full=DNA-3-methyladenine glycosylase 1; AltName: Full=3-	<a href="#">134</a>	134	54%	2e-35	18%	<a href="#">SG</a>
<input checked="" type="checkbox"/> <a href="#">Q94468.1</a>	RecName: Full=Probable DNA-3-methyladenine glycosylase 2; AltNam	<a href="#">125</a>	125	56%	2e-32	19%	<a href="#">G</a>
<input checked="" type="checkbox"/> <a href="#">P46303.2</a>	RecName: Full=Ultraviolet N-glycosylase/AP lyase; AltName: Full=Pyri	<a href="#">126</a>	126	51%	7e-32	22%	
<input checked="" type="checkbox"/> <a href="#">Q49498.2</a>	RecName: Full=DEMETER-like protein 3	<a href="#">127</a>	127	64%	2e-30	21%	<a href="#">GM</a>
<input checked="" type="checkbox"/> <a href="#">Q8LK56.2</a>	RecName: Full=Transcriptional activator DEMETER; AltName: Full=DNA	<a href="#">125</a>	125	52%	2e-29	20%	<a href="#">GM</a>
<input checked="" type="checkbox"/> <a href="#">P04395.1</a>	RecName: Full=DNA-3-methyladenine glycosylase 2; AltName: Full=3-	<a href="#">117</a>	117	67%	9e-29	16%	<a href="#">S</a>
<input checked="" type="checkbox"/> <a href="#">Q9WYK0.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">115</a>	115	63%	1e-28	18%	
<input checked="" type="checkbox"/> <a href="#">P73715.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">113</a>	113	51%	8e-28	20%	
<input checked="" type="checkbox"/> <a href="#">P39788.1</a>	RecName: Full=Probable endonuclease III; AltName: Full=DNA-(apurin	<a href="#">108</a>	108	50%	4e-26	16%	
<input checked="" type="checkbox"/> <a href="#">Q9SJO6.2</a>	RecName: Full=Protein ROS1; AltName: Full=DEMETER-like protein 1;	<a href="#">110</a>	110	51%	1e-24	21%	<a href="#">GM</a>
<input checked="" type="checkbox"/> <a href="#">P44319.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">103</a>	103	46%	1e-24	20%	<a href="#">G</a>
<input checked="" type="checkbox"/> <a href="#">Q9SR66.2</a>	RecName: Full=DEMETER-like protein 2	<a href="#">102</a>	102	47%	3e-22	19%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> <a href="#">P0AB84.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">95.8</a>	95.8	46%	8e-22	19%	
<input checked="" type="checkbox"/> <a href="#">P54137.2</a>	RecName: Full=Probable endonuclease III homolog; AltName: Full=Cef	<a href="#">91.2</a>	91.2	45%	1e-19	23%	
NEW <input checked="" type="checkbox"/> <a href="#">P63541.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">83.5</a>	83.5	54%	3e-17	16%	
NEW <input checked="" type="checkbox"/> <a href="#">Q89AW4.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">80.8</a>	80.8	46%	2e-16	19%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> <a href="#">Q8KA16.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">80.4</a>	80.4	50%	2e-16	19%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> <a href="#">Q9CB92.2</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">78.9</a>	78.9	50%	1e-15	17%	
NEW <input checked="" type="checkbox"/> <a href="#">Q92GH4.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">76.9</a>	76.9	46%	4e-15	17%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> <a href="#">Q4UK93.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">75.0</a>	75.0	48%	2e-14	16%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> <a href="#">Q68W04.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">74.6</a>	74.6	48%	3e-14	15%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> <a href="#">Q05956.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">72.3</a>	72.3	48%	2e-13	17%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> <a href="#">Q58030.2</a>	RecName: Full=Putative endonuclease MJ0613	<a href="#">71.6</a>	71.6	63%	1e-12	19%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> <a href="#">P57219.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">68.9</a>	68.9	49%	3e-12	17%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> <a href="#">Q83754.1</a>	RecName: Full=Endonuclease III; AltName: Full=DNA-(apurinic or apy	<a href="#">68.1</a>	68.1	46%	4e-12	16%	<a href="#">G</a>
NEW <input checked="" type="checkbox"/> <a href="#">Q8SR88.1</a>	RecName: Full=Endonuclease III homolog; AltName: Full=DNA-(apurin	<a href="#">65.8</a>	65.8	50%	4e-11	17%	

Firefox

NCBI Blast: Protein Sequence (376 letters)

blast.ncbi.nlm.nih.gov/Blast.cgi

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
<input checked="" type="checkbox"/> Q83754.1	RecName: Full=Endonuclease III; AltName: Full=DNA- (apurinic or apy...	68.1	68.1	46%	4e-12	16%	G
<input checked="" type="checkbox"/> Q8SRB8.1	RecName: Full=Endonuclease III homolog; AltName: Full=DNA- (apurin...	65.8	65.8	50%	4e-11	17%	
<input checked="" type="checkbox"/> P22134.1	RecName: Full=DNA- 3- methyladenine glycosylase; AltName: Full=3- m...	61.9	61.9	54%	2e-09	16%	G
<input checked="" type="checkbox"/> Q2KID2.1	RecName: Full=Endonuclease III-like protein 1	59.2	59.2	45%	1e-08	19%	GM
<input checked="" type="checkbox"/> Q09907.1	RecName: Full=Endonuclease III homolog; AltName: Full=DNA- (apurin...	59.2	59.2	53%	2e-08	18%	G
<input checked="" type="checkbox"/> P78549.2	RecName: Full=Endonuclease III-like protein 1	58.8	58.8	45%	2e-08	17%	GM
<input checked="" type="checkbox"/> P29588.1	RecName: Full=G/T mismatches repair enzyme; AltName: Full=Mismat...	57.3	57.3	40%	2e-08	16%	SG
<input checked="" type="checkbox"/> Q35980.1	RecName: Full=Endonuclease III-like protein 1	57.7	57.7	58%	3e-08	16%	GM
<input checked="" type="checkbox"/> Q8K926.1	RecName: Full=A/G-specific adenine glycosylase	53.5	53.5	46%	1e-06	16%	G
<input checked="" type="checkbox"/> P17802.1	RecName: Full=A/G-specific adenine glycosylase	53.5	53.5	46%	1e-06	14%	S
<input checked="" type="checkbox"/> Q58829.1	RecName: Full=Putative endonuclease MJ1434	51.5	51.5	51%	2e-06	20%	G
<input checked="" type="checkbox"/> P57617.1	RecName: Full=A/G-specific adenine glycosylase	52.3	52.3	46%	2e-06	17%	G
<input checked="" type="checkbox"/> Q08214.1	RecName: Full=DNA base excision repair N-glycosylase 2	52.3	52.3	67%	3e-06	20%	G
<input checked="" type="checkbox"/> P31378.1	RecName: Full=Mitochondrial DNA base excision repair N-glycosylase	51.1	51.1	45%	7e-06	20%	G
<input checked="" type="checkbox"/> Q05869.1	RecName: Full=A/G-specific adenine glycosylase	49.2	49.2	47%	2e-05	14%	
<input checked="" type="checkbox"/> Q10159.1	RecName: Full=A/G-specific adenine DNA glycosylase	47.3	47.3	45%	1e-04	16%	G
<input checked="" type="checkbox"/> Q89A45.1	RecName: Full=A/G-specific adenine glycosylase	46.1	46.1	44%	2e-04	17%	G
<input checked="" type="checkbox"/> Q31584.1	RecName: Full=Probable A/G-specific adenine glycosylase YfhQ	44.2	44.2	49%	0.001	14%	

Run PSI-Blast iteration 4 with max

**Sequences with E-value WORSE than threshold**

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
<input type="checkbox"/> P44320.1	RecName: Full=A/G-specific adenine glycosylase	41.5	41.5	44%	0.007	14%	G
<input type="checkbox"/> Q9UIF7.1	RecName: Full=A/G-specific adenine DNA glycosylase; AltName: Full=...	40.4	40.4	34%	0.024	18%	SGM
<input type="checkbox"/> A1KRU4.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA	37.7	37.7	34%	0.079	19%	G
<input type="checkbox"/> Q9XA14.1	RecName: Full=Recombination protein RecR	37.7	37.7	13%	0.081	31%	
<input type="checkbox"/> Q9JSM5.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA	37.3	37.3	34%	0.091	19%	
<input type="checkbox"/> Q9K1A2.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA	37.3	37.3	34%	0.11	19%	
<input type="checkbox"/> Q9XDH5.1	RecName: Full=DNA polymerase III subunit alpha	38.0	38.0	27%	0.12	19%	S
<input type="checkbox"/> A9M3B7.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA	36.9	36.9	34%	0.14	19%	G
<input type="checkbox"/> Q5F636.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA >	36.5	36.5	33%	0.18	19%	G
<input type="checkbox"/> Q8R5G2.1	RecName: Full=A/G-specific adenine DNA glycosylase; AltName: Full=...	37.3	37.3	35%	0.18	17%	GM
<input type="checkbox"/> Q0B0W3.1	RecName: Full=Recombination protein RecR	36.5	36.5	8%	0.18	28%	G
<input type="checkbox"/> Q99P21.2	RecName: Full=A/G-specific adenine DNA glycosylase; AltName: Full=...	37.3	37.3	57%	0.19	16%	GM
<input type="checkbox"/> C5D5E9.1	RecName: Full=Holliday junction ATP-dependent DNA helicase RuvA	36.1	36.1	45%	0.27	15%	G
<input type="checkbox"/> 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)

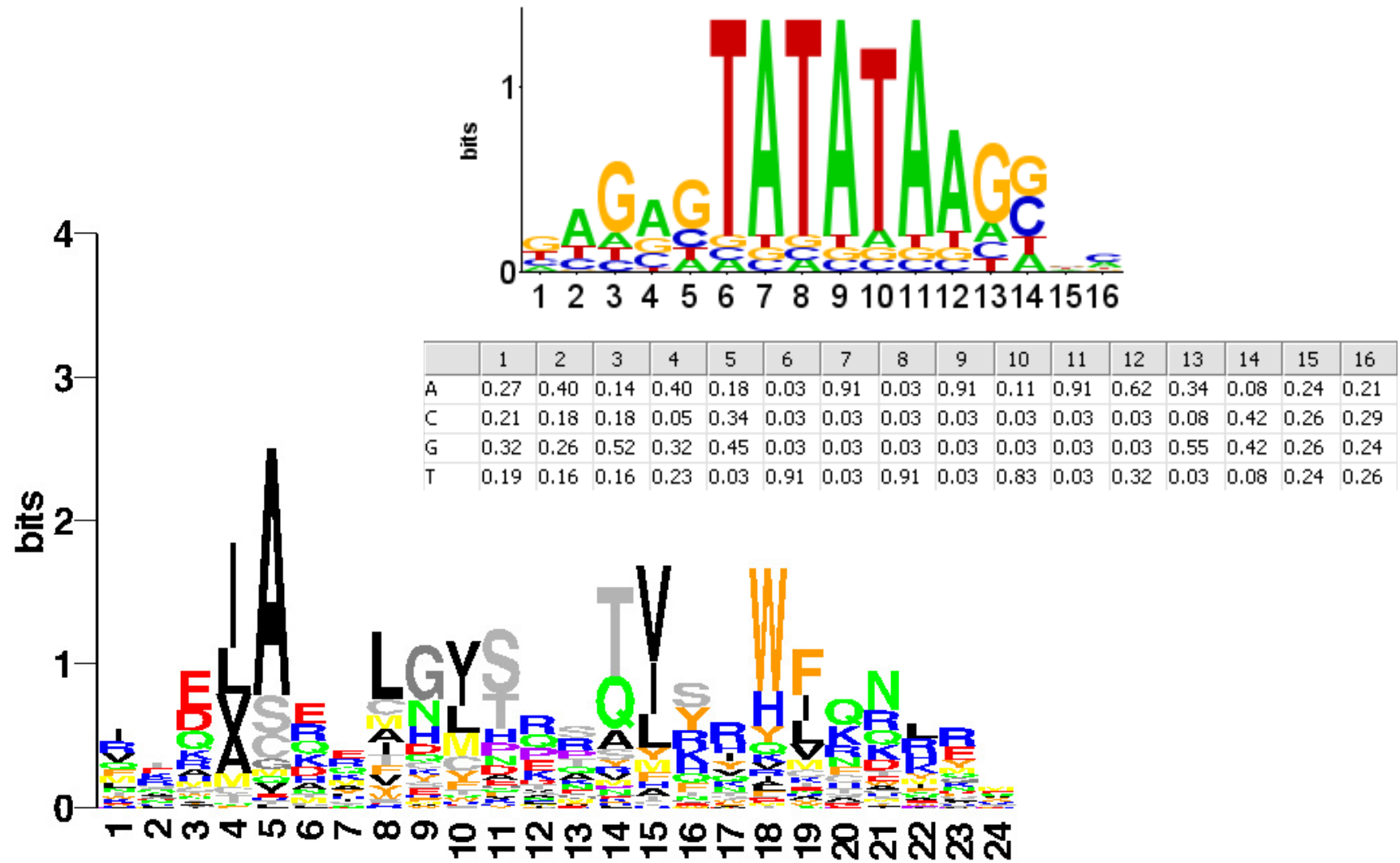
A good multiple alignment of the sequences in the family is essential for most of these models.



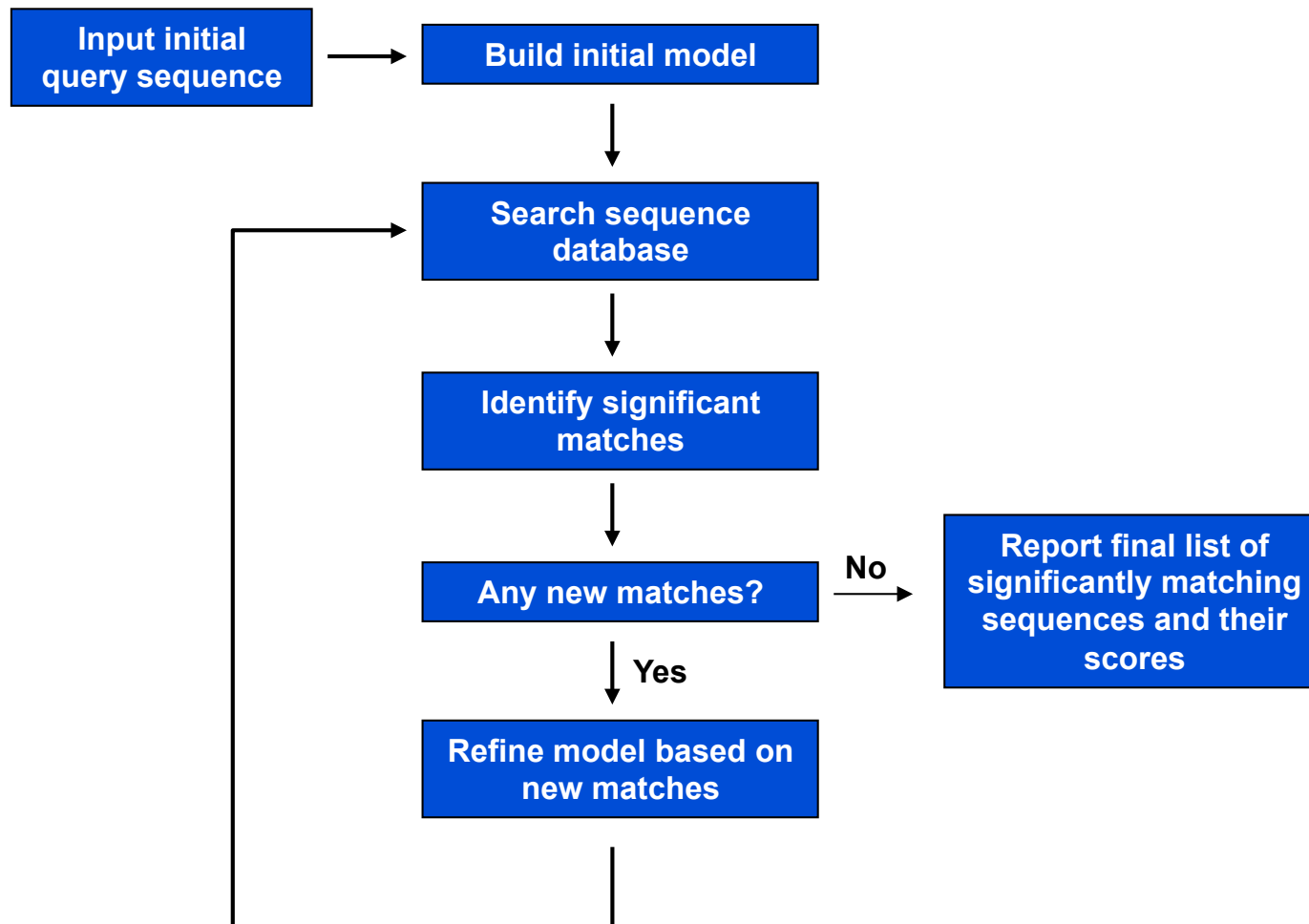
# Sequence profiles (PSSMs)

- Position-specific scoring matrices
- Based on a multiple alignment of proteins in a family
- A matrix of  $21 \times L$  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



# 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](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>



## AlkB paper

- *The DNA-repair protein AlkB, EGL-9, and Iprepcan 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>





# 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_Sola_322266	RSQTVVHVDYYP-SPGAHLL-SSETSETLLEFPHBMA-----YHRLQPNYVMLACSSRADHE----RTAATLVASVRK--70--VTEAVYLEEG-DLLIVDNF-----RTTHARTPFSPRWGDKDMLHRVYIRT	302\
IPNS_En_124825	TLASVVLIRYPYLDPPY3KTAADGTKLSFEWHDVDS-----LITVLYQ-----SNVQNLQVETAA-----GYQDIADDT-GYLINCOSYMAHLTNNYKAPIHRVWVN----ABRQSLPFFVNL	288
PLAS_Pet_421946	IVYLLKINYP-PCPR----PDLALGVVAHDMS-----YITILVP-----NEVQGLQVFKDG-----HWYDVKIEN-ALIVHIGDQVEILSNGKYKSVYHRTVTK----DKTRMSWPVPLEP	309
LDOX_Pet_1730108	LLLQMKINYP-KCPQ----PELALGVVAHDVDS-----ALTFILH-----NMVGLQLFYEG-----QWVTAKCVEN-SIIMHIGDTIEILSNGYKSIIDRRGVNK---EKVRFSAWAFCEP	311
Srg_At_479047	SVQSMRMNYP-PCPQ----PDQVIGLTFHSDSV-----GLTVLMQV-----NDVEGLQIKKDG-----KWVPKPLEN-AFIVNIGDVLEIITNGYRSIBRRGVVNS---EKERLSIATPHNV	309
EFE_Le_398992	PNFGTKVSNYP-PCPK----PDLIKGLRAHDADG-----GIILLLQD-----DKVGLQLKDE-----QWIDVPPMRH-SIVVNLGDQLEVITNGYKSVIHRVIAQT---DGTFRMSLASFYNP	253
Ga200x_Sot_10800976	NESIMRLNYP-TQCK----PDLALGTGFHDPT-----SLLTLHQ-----DSVGLQVFMND-----QWRSISPNLS-AFVVNIGDTFMALSNRYKSCIDRRVAVNN---KTFRKSLAFFLCP	317
PA0147_Pa_9945977	PVSVFRLIHY-PASA---RQSADQPGAGAHDDYG-----CVTLLYQ-----DAAGGLQVQNRQG-----EWIDAPPIEG-FVFNVIGDMMARWSNDRYRSTPHRVISPR---GVHRYSMPPFAEP	274
PA4191_Pa_9950401	PLILFRLPNYPSQPVPE-----GLDVQWVGQEHDDYG-----LLTLLHQ-----DAIGGLQVTRTPQ-----GWLEAPPIDG-SFVCNLGDMLERMTGGLYRSTPHRVARNTS---GRDRLSFPFLFDP	277
ISP7_Sp_729862	PTTSIRLLRY-P-----SSPNRLGVQEHDDAD-----ALTLMLQ-----DNVGLQVLDVPSN-----CFLSVSPARG-ALIANLGDIMAILTNNRYKSSMHRVCNNS---GSDRYTIFPFLQG	353
SPCC1494.01_Sc_7491815	EEDVLRLLKYSI-PEGV---ERREDDEAGAHSDYG-----SITLLFQ-----RDAAGLEIRPPNFVKDM---DWIKVNVQD-VVLVNIADMLQFPTSGLKLRSTVHRVIDPG---VKTQTIAYVTP	267
DAOCS_Ly1_769809	CDPVLRYRYPDVPEDR---CAEQQPNRMAFHDLS-----IVSLLQTPCP-----NGFVSLQVEIDG-----RFVEVPPRG-CVVVFCGSIAPLVS DGKIKAPQHRVVS-PGA4-GSNRTS VLVLELRP	268/
RRPO_SHVX_548840	TYNQCLVQKYE-----QGSRIGFHSDQAIYPKG-----NKILTVAIA-----GSGTFGI-----KCAKGE-TTLNLEDGD-YFQMPSGFQETHRKAIVA----VTPFLSFTFRSTV	743\
POL_ASPV_487652	FYNQCLVQEYS-----TGHGLSMHRDDES IYDIN-----HQVLTVNSY-----GDAIFCI-----ECLGSEF-EIPLSGPQ-MLLMPGPFQKEHRHGKSP----SKGRISLTFRRLTK	853
POL_BSV_409711	TYDCMLAQRYG-----AQQKIGFHADNEE IFMRG-----APVHTVSM-----GADFGT-----ECAAGR-QYTLRGNVQFTMPSGFQETHRKAIVRNT---TAGRVSYTFRRLA	841
RRPO_PMV_139137	EFNQCLVQCFK-----LQAAIPFHDDDEPCYPKG-----HQVLTINHS-----GECCTCI-----ACQKGA-SITMGFGD-YLSPVGFQESHKHAIVSNT---TGGRVSLTFRCTV	690
POL_GLV_1154656	YFNCVLFQKYD-----GGHGIGFHEDDEE IPEKD-----SKILTVCIQ-----GDEEFF-----RCATGET-GPYMEAPK-QFMMPDGFQSNHVAVREC---TPGRISATFRRAK	772
Pol_GVA_1405615	SYDHCLIQRYT-----AGSGIGFHADDEPCYLPF-----GSVVTVNLH-----GDATFEVK-----ENQSGIKKELHDGD-YVVMGPGMQQTHRHRVTSH---TDGRCSITLNRKT	738
RRPO_ACLSv_1710717	NFNSALIQVYN-----DGCRLLPESDNEE CYDD-----DEILTINVV-----DKAKFHT-----TC-HGE-IDLRLQGD-EI LMPGGYQKMKRHAIVEVA---SEGRTSVTLRVHK	836/
T13L16.2_Ac_2708738	VPDSCI VNIYD-----EGDCIIPFHDDNHDFL-----RPFCTISFL-----SECDIIFGSNLKVE-----GPGDFSGY-SIFLPVGS-VLVLNGGADVAKHCVPAV---PTKRISITFRKMD	420\
T19K4.220_Ac_3036813	IIKSCIVNIYE-----EDDCIIPFHDDNHDFL-----RPFCTVSL-----SECNILFGSNLKV-----GPGDFSGY-SIFLPVGS-VLVLKNGADVAKHCVPAV---PTKRISITFRKMD	403
At2g48080_At_4249414	RPNGCVINFDQ-----P-FQKPPHYD-----QPISLTVL-----SEBTMVFGHRLGVD-----NDGNFRGSL-TLFLKEGS-LLVMRGNADMARHVMCPS---PNKRVAITFFPKL	351
AK000315.1_Hs_7020317	GFVNSAVINDYQ-----PGGCIVSHVDPIHIFE-----RPIVSVSFP-----SDBALCFGCKFPQK-----PIRVSEVLSLFPVRRGS-VTVLSGYADEI THCIRPQDI---KERAVIILRTR	270
CG17807_Dm_7291441	SPDQLTVNEYE-----PDGHGIPFHDDTHSAFL-----SDVVMDFRRG-----DDQV-QVRLPRRS-LLVMSGEARYDWHGIRPKHID13RGRKSLTFRRLR	325
CG6144_Dm_7297142	NANHVLVNEYL-----PGQGLIPEHDDGFLPH-----PIIISTISTG-----AHVLEVFVREDTTTETETAGDQTTREVLV-KLLLEPRS-LLILKDTLYTDYDPAISETSED24RSPRISLITRNVP	213
CG4036_Dm_7297561	QTIEQCSLEYEPS-----KGASIDPHVDDCMIWGERVVTVNC-----LGDSVLTLT-----PYEVQSGKYNLDLVASYEDEL LAP-LLTDDQLATPEGKVLRI PMPNLS-LIVLYGPARYQFBHVLREDV---ERRRVCAVREFT	278
PLJ2001_Hs_38923019	RPVQCNLDYCPE-----RGSADIPHDDAMLWGERLVSLNL-----LSPTVLSMC-----REAPGSLLCAPSAPEALVDSVIAPSRSVLCQEVVAIFL PARS-LLVLTGAARHQWKAHHRHHI---EARRVCVTFRELS	274
C14B1.10_Ce_6580210	RPDQVANVYE-----SGHGIPSHDDTHSAFD-----DPIVSLSL-----SDVVMDFKD-----GANSARIAPVLLKARS-LCLIQGESRYRWHGIVNRYD10RQTFVSLTLRKIR	343
SPAP8A3.02c_Sp_7491301	DAEAIIMQVYN-----PGDGIIPKRDLEMFGDG-----VAIFSPLSN-----LKLKQ-----KIRLEKGS-LLLMSGTARYDWPBEIPFRAGD12RSQFLSVTMRRII	219
L3377.4_Lm_9989036	WLNQTLANLYE-----PGDFIRAHDDNLFVYD-----DIFATCSLG-----SNCLLRFVH-----VQNGEEL-DVMVPDRS-VYIMSGPARYVYRHWLVP---EAQFSLVFRRSI	193/
MTC1237.14c_Mtu_2052134	FTTAGLCYRD-----GSDSVAMHDDTIGRSTEDTM-----VAIVSLGAT-----RVFALRP-----RGRGFSRLRFLAHGD-LLVMGSGCQRTFEBHVPKTSAP---TGPRVSIQFRPRD	203\
AlkB_Cc_2055386	PPDSCLVNLYA-----TGARMLGHDDRDEADPR-----FPLLSTSLG-----DTAVFRIGG-----VNRKIDTRSLASGD-VCRLLGPARLAPHGVDRI LFG6-GGGRINLTLRRAR	190
AlkB_Ec_113638	QPDACLINRYA-----LPAKLSLGHDDKDE PDLR-----LPAKLSLGH-----LKRNDLKRLLLEHGD-VVVWGGESRLFYHGIQPLKAG5-LDCRYNLTFRQAG	213
AlkB_Scoe_8894829	PYDIALINFDG-----ADARMGSHDDADERTD-----APVVSLSLG-----DTCVFRFNG-----PETRTFYDTELRSGD-LFVFGGSPRLAYHGVPRVHPG7-LRGLRNLTLRVSG	215
AlkB_At_4835778	RPEGAI VNYFG-----IGDTLSGHDDMEADWS-----KPIVMSLG-----CKAIFLLSGK-----SKDDP PHMYLRSGD-VVLMAGEARECPHNLHFQL34TKSRIMINIRQVF	354
AlkB_Sp_3080529	KAEAAIVNFYS-----PGDTLSAHDDSEEDLT-----LPLISLSMG-----LDCIYLIGTE-----SRSEKFS-ALRRLHSGD-VVIMTGTSRKAPHGKHC---SKFYLIYSQLIA	272
AlkB_Hs_2134723	REBAGILNRYR-----LDSTLGIHVDSELDHS-----KPLLSPSFG-----QSAIFLLGGL-----QRDEARP-PMFMSHGD-IMIMSGFSRLNLHAPRVLPN39KTAIVNMA RQVL	272/
Consensus (85%):	.....h.h.a.....*.....h.H.D.....sh.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 techniques
- 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  $\bar{s}^i$  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

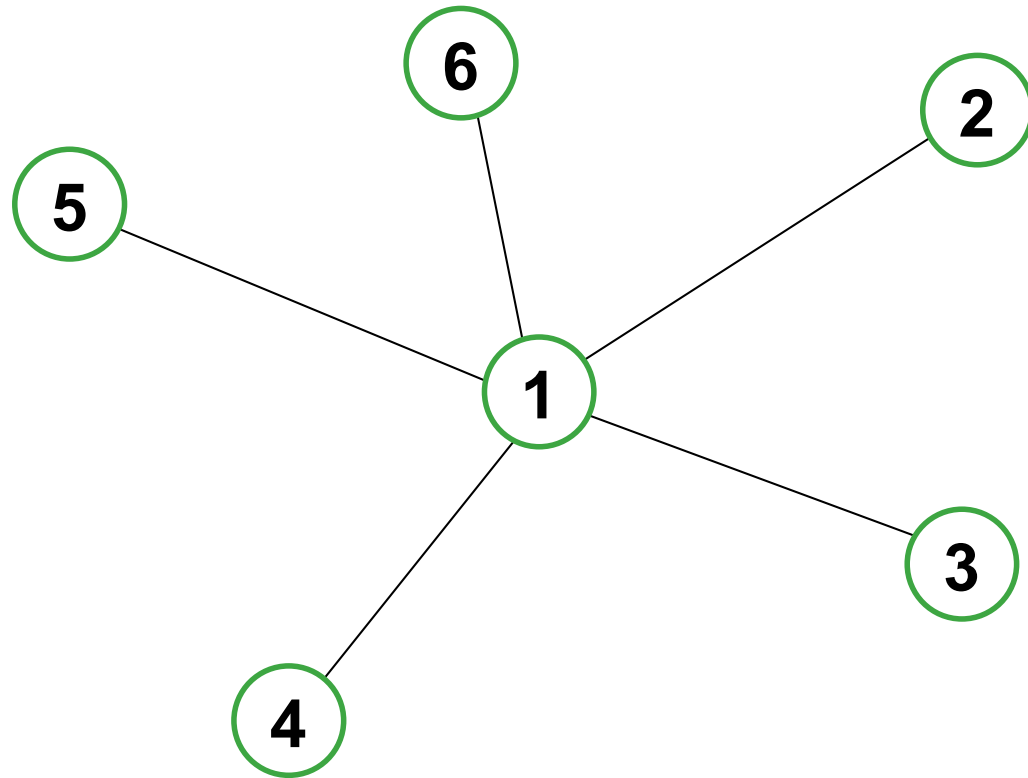
$$\text{score}(k) = S(P,R) + S(P,-) + S(P,P) + S(R,-) + S(R,P) + S(-,P)$$

↑  
score for  
column  $k = 3$

We have  $S(-,-) = 0$

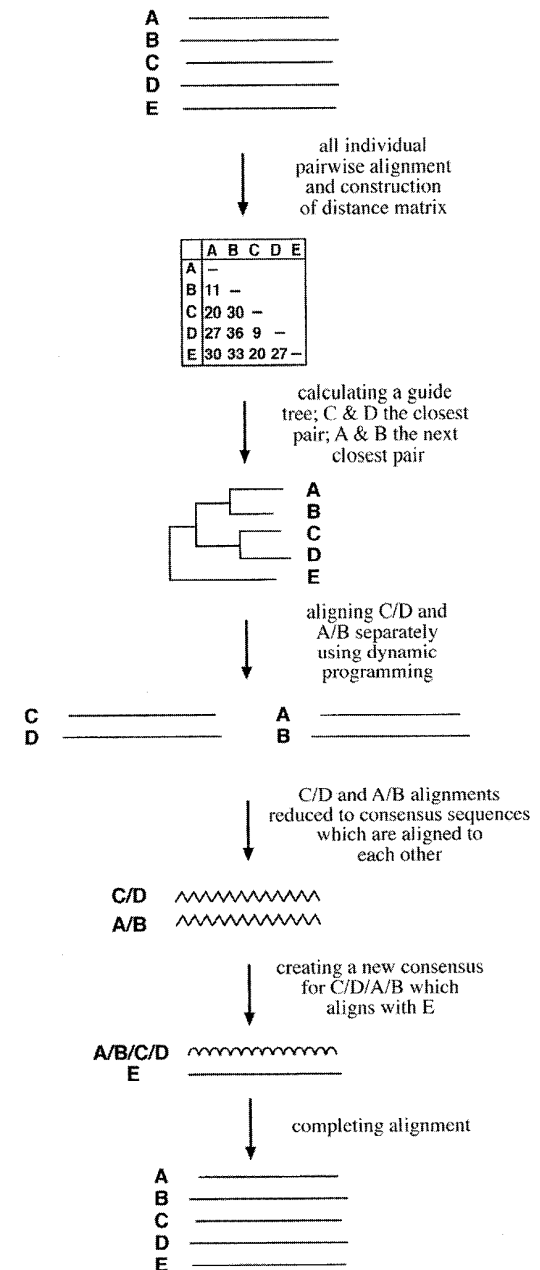
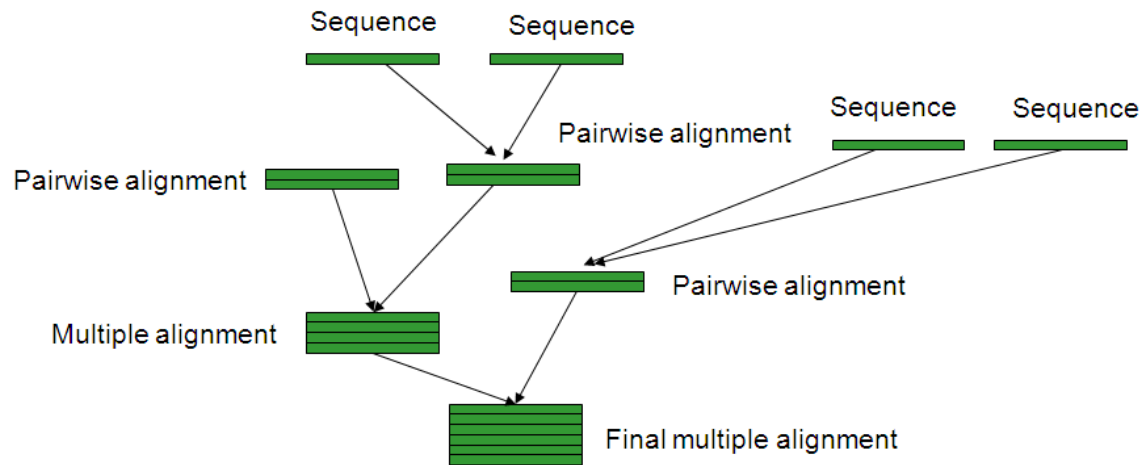
$$\text{Total score} = \text{score}(1) + \text{score}(2) + \dots + \text{score}(N)$$

# Centre star multiple alignment



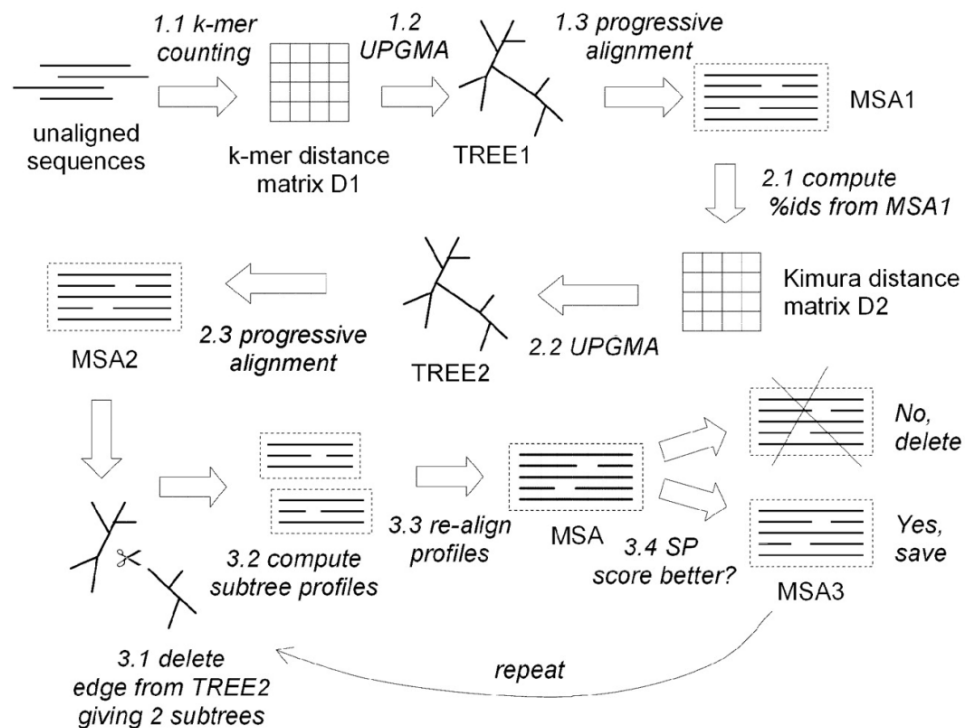
# Clustal W

- 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<sup>1,2</sup>, Cedrik Magis<sup>1,2</sup>, Giovanni Bussotti<sup>1</sup>, Jia-Ming Chang<sup>1</sup>, Paolo Di Tommaso<sup>1</sup>, Jonas Erb<sup>1</sup>, Jose Espinosa-Carrasco<sup>1</sup>, Carsten Kemena<sup>1</sup> & Cedric Notredame<sup>1</sup>

<sup>1</sup>Comparative Bioinformatics Group, Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Universitat Pompeu Fabra (UPF), Barcelona, Spain.

<sup>2</sup>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 Espresso, 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/>.**