Introduction to sequence similarity searches and sequence alignment

MBV-INF4410/9410
Monday 17 November 2014

Torbjørn Rognes
Department of Informatics, University of Oslo
& Department of Microbiology, Oslo University Hospital
torognes@ifi.uio.no

Overview of the presentation

PART 1

- An example showing how useful bioinformatics can be
- Searching sequence databases
- A walk-through of the BLAST search service

PART 2

- Alignments, sequence similarity and homology
- Significance of matches: What is a good match?
- How does BLAST work?

PART 3

- Iterative searching with a family of proteins (PSI-BLAST)

PART 4:

- Multiple sequence alignments

One example of how useful bioinformatics can be

- The protein AlkB was discovered in *E.coli* in 1984.
- It was known that it protected the bacterium when subjected to DNA-alkylation agents.
- No enzymatic activity was found.
- Perhaps some co-factors where missing?
- In 2001, a bioinformatics paper was published that shed light on the problem. Many similar sequences where found using advanced sequence similarity searches …

http://genomebiology.com/2001/2/3/research/0007.1

Research

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

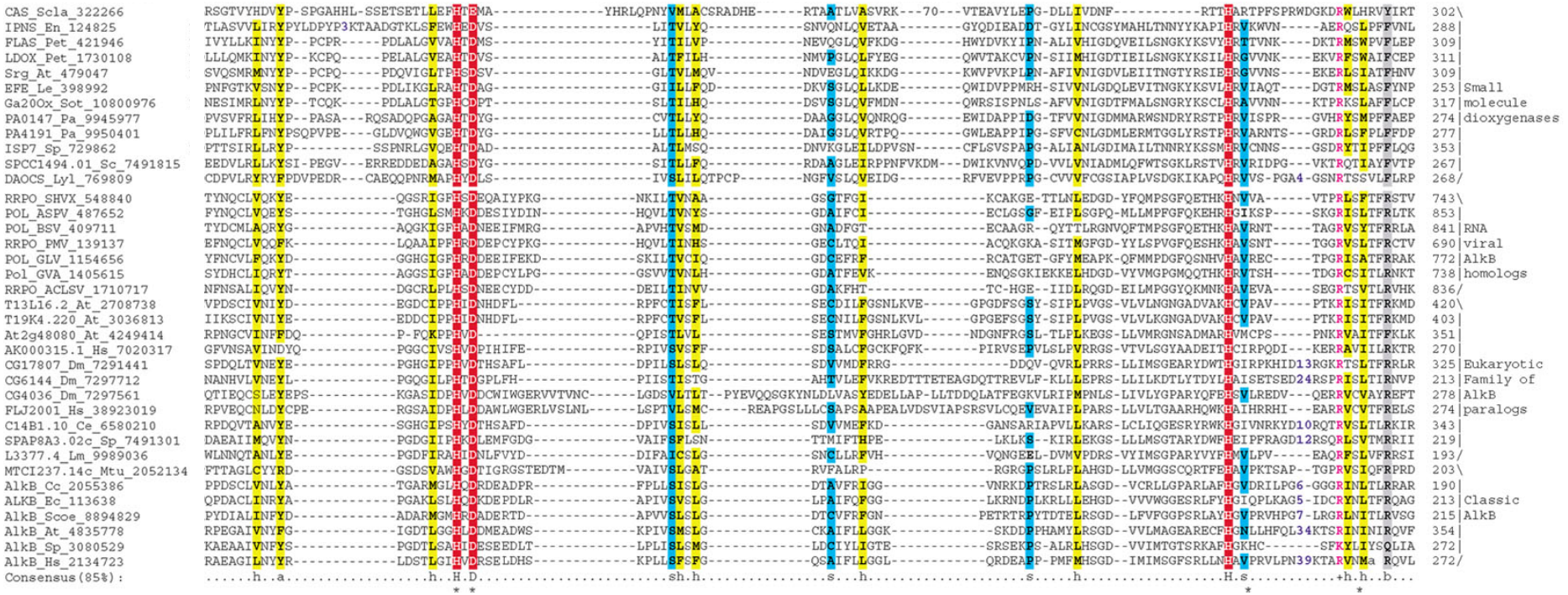
L Aravind and Eugene V Koonin

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

Correspondence: L Aravind. E-mail: aravind@ncbi.nlm.nih.gov

Example...

Alignment showing conserved amino acids among many sequences



Example…

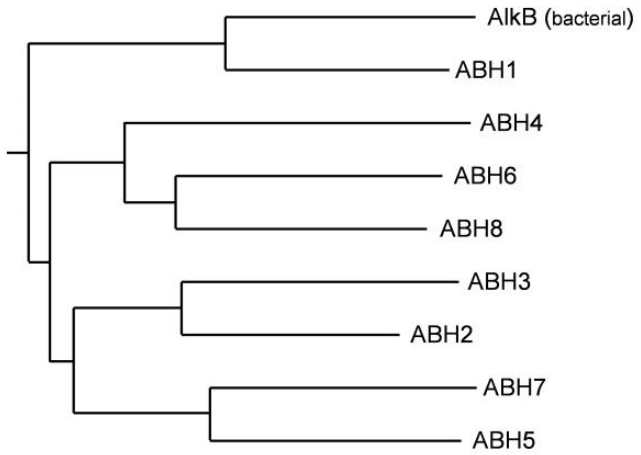
- By comparing *E.coli* AlkB to other sequences in the database it was found that AlkB had some features in common with more well-known enzymes

- Based on these similarities the following was suggested regarding AlkB:
  - That AlkB is a dioxygenase
  - That the enzyme is Iron(II) dependent
  - That the enzyme is 2-oxo-glutarate dependent
  - That AlkB repairs alkylated bases through a form of oxidation
  - That the enzyme could demethylate RNA as well (not just DNA)
  - That there were eukaryotic counterparts of the protein

- All of this was later verified in the lab and resulted in three publications in *Nature*.

Example...

- By further sequence analysis 3 AlkB-like sequences were found in humans:
  - ALKBH1
  - ALKBH2
  - ALKBH3
- And by even more advanced analysis another 5 homologs were found in humans:
  - ALKBH4
  - ALKBH5
  - ALKBH6
  - ALKBH7
  - ALKBH8

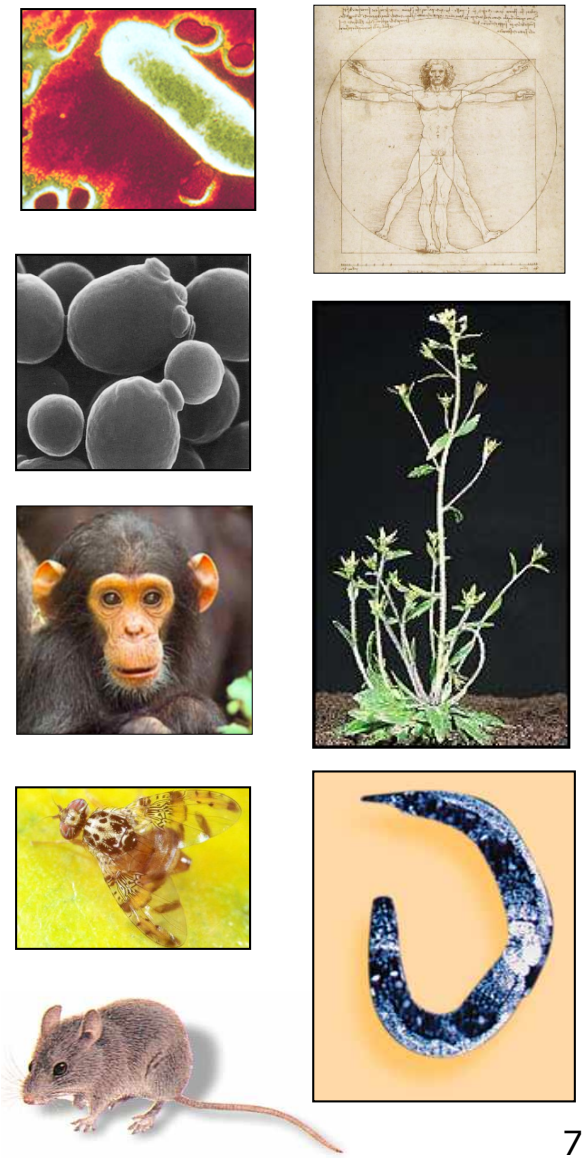


- The function of these 8 enzymes are now being studied in detail. Some of them may be related to human diseases.

Genomes are a huge source of information

- More than 6000 ”completely” sequenced genomes available – an enormous source of information. Many thousands of other genomes in progress*
- Almost 1 000 000 000 000 basepairs in GenBank (2014)*
- Database sizes are growing exponentially – doubling in about 18 months since 1982
- Searching sequence databases for a similar sequence is fundamental in many types of analyses in bioinformatics
- Searching a sequence database with a new amino acid or nucleotide sequence allow us to find out more about:
  - Gene function
  - Conserved and probably important residues
  - 3D structure of a protein
  - Distribution of the gene among species
  - Gene structure
  - Chromosomal localisation
- Save time in the lab!
- Database searching is highly compute intensive and is probably the task consuming the largest amount of computing time within bioinformatics.

* Sources: genomesonline.org & NCBI (ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt)

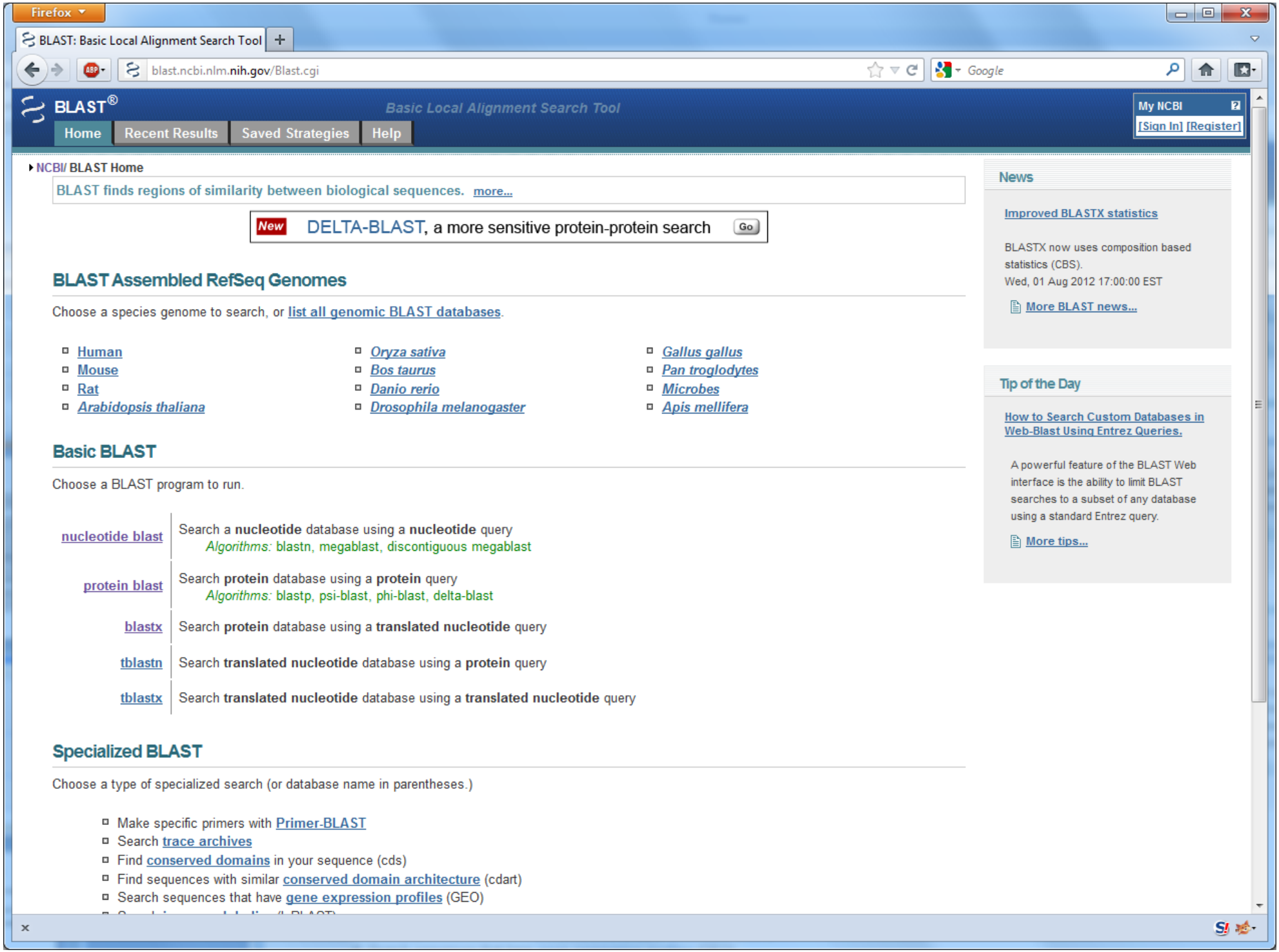


Searching sequence databases

- Goal: Identify which sequences in a database are significantly similar to a given DNA, RNA or protein sequence.
- How: The query sequence is compared (aligned) with each of the database sequences, and the amount of similarity is determined for each database sequence.

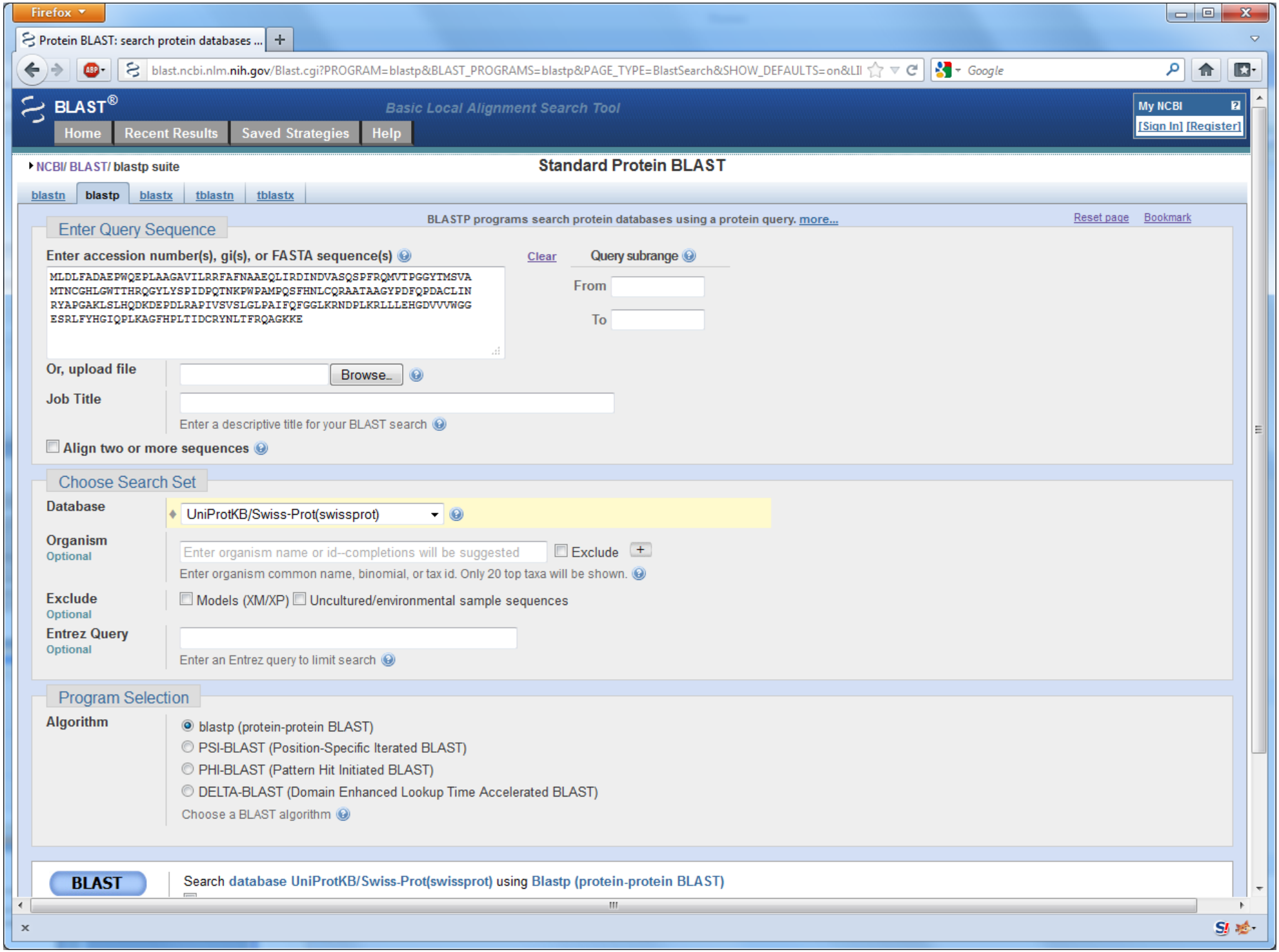
Example:

Query sequence:	**acgatcgattagcca**
Database sequences:	
Identical (trivial):	**acgatcgattagcca**
Very similar (easy):	**acgaccgatgagcca**
Similar (moderate):	**atgacggatgagcga**
Very diverged (hard):	**atgacgggatgagcga**


Firefox
BLAST: Basic Local Alignment Search Tool
blast.ncbi.nlm.nih.gov/Blast.cgi
Google
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST Home
BLAST finds regions of similarity between biological sequences. more...
New
DELTA-BLAST, a more sensitive protein-protein search
Go
BLAST Assembled RefSeq Genomes
Choose a species genome to search, or list all genomic BLAST databases.
Human
Mouse
Rat
Arabidopsis thaliana
Oryza sativa
Bos taurus
Danio rerio
Drosophila melanogaster
Gallus gallus
Pan troglodytes
Microbes
Apis mellifera
Basic BLAST
Choose a BLAST program to run.
nucleotide blast
Search a nucleotide database using a nucleotide query
Algorithms: blastn, megablast, discontiguous megablast
protein blast
Search protein database using a protein query
Algorithms: blastp, psi-blast, phi-blast, delta-blast
blastx
Search protein database using a translated nucleotide query
tblastn
Search translated nucleotide database using a protein query
tblastx
Search translated nucleotide database using a translated nucleotide query
Specialized BLAST
Choose a type of specialized search (or database name in parentheses.)
Make specific primers with Primer-BLAST
Search trace archives
Find conserved domains in your sequence (cds)
Find sequences with similar conserved domain architecture (cdart)
Search sequences that have gene expression profiles (GEO)
News
Improved BLASTX statistics
BLASTX now uses composition based statistics (CBS).
Wed, 01 Aug 2012 17:00:00 EST
More BLAST news...
Tip of the Day
How to Search Custom Databases in Web-Blast Using Entrez Queries.
A powerful feature of the BLAST Web interface is the ability to limit BLAST searches to a subset of any database using a standard Entrez query.
More tips...

Search program variants

Query	Database	Comparisons	FASTA	BLAST	Description
Nucleotide	Nucleotide	Nucleotide (2)	fasta (fastn)	blastn	Compares directly both strands (forward and reverse complement) of the nucleotide query sequence to the nucleotide sequences in the database.
Amino acid	Amino acid	Amino acid (1)	fasta (fastp)	blastp	Compares the amino acid query sequence with the amino acid sequences in the database.
Amino acid	Nucleotide	Amino acid (6)	tfasta, tfastx, tfasty	tblastn	Translates the database nucleotide sequences into all six frames and compares the resulting amino acid sequences with the amino acid query sequences. tfasty allows intra-codon substitutions and frameshifts.
Nucleotide	Amino acid	Amino acid (6)	fastx, fasty	blastx	Translates the nucleotide query sequence into all six frames and compares the resulting amino acid sequences with the amino acid sequences in the database. fasty allows intra-codon substitutions and frameshifts.
Nucleotide	Nucleotide	Amino acid (36)	-	tblastx	Translates both the query nucleotide sequence and the database nucleotide sequences into all six frames and compares the resulting amino acid sequences with each other.

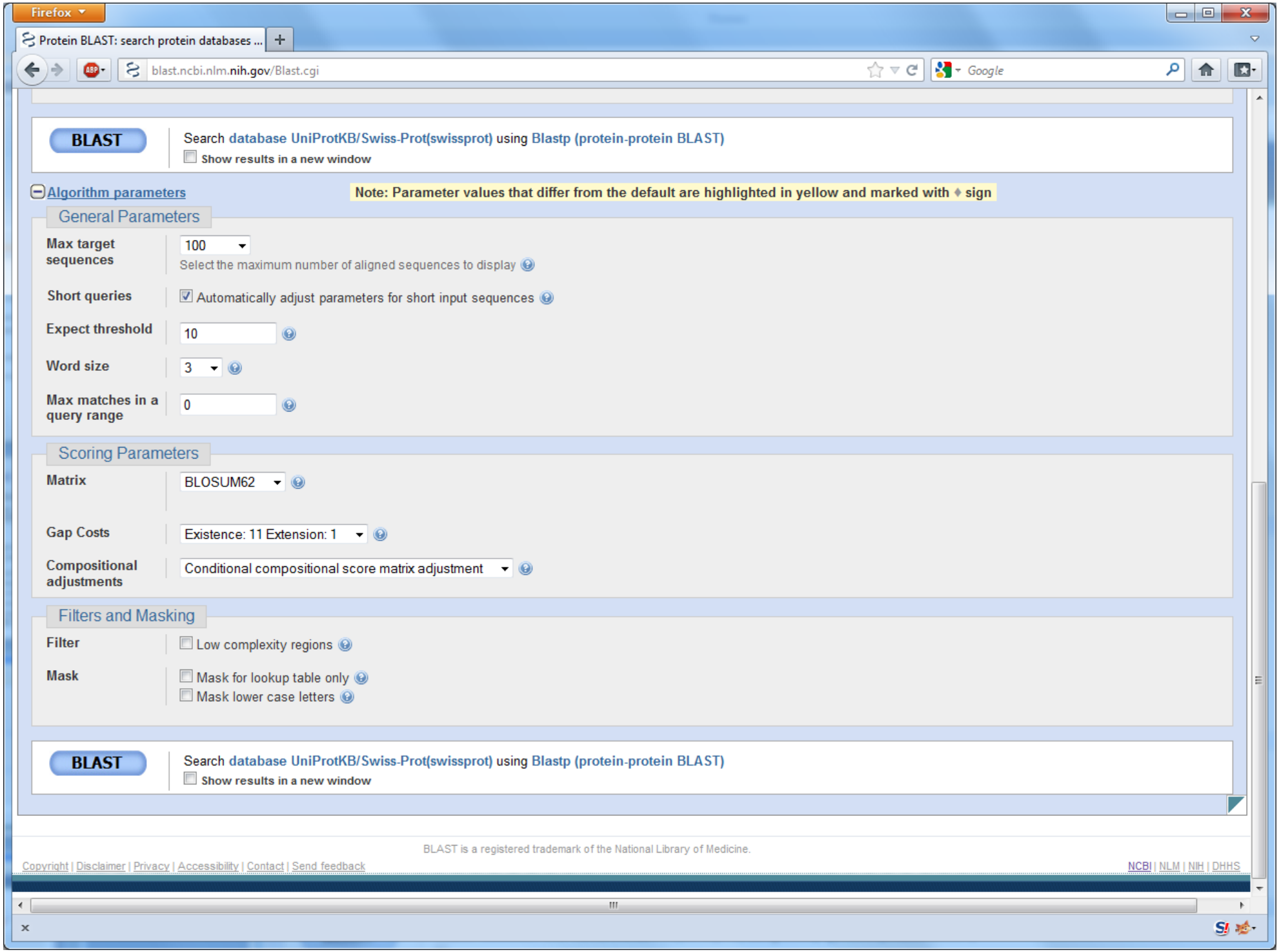

Firefox
Protein BLAST: search protein databases ...
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LI
Google
BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp 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
MLDLFADAEPWQEPLAAGAVILRRFAFNAAEQLIRDINDVASQSPFRQMVTPGGYTMSVA
MTNCGHLGWTTHRQGYLYSPIDPQTNKPWPAMPQSFHNLCQRAATAAGYPDFQPDACLIN
RYAPGAKLSLHQDKDEPDLRAPIVSVSLGLPAIFQFGGLKRNDPLKRLLLEHGDVVVWGG
ESRLFYHGIQPLKAGFHPLTIDCRYNLTFRQAGKKE
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 Blastp (protein-protein BLAST)

BLAST databases (protein)

nr:	All non-redundant GenBank CDS translations + RefSeq Proteins + PDB + UniProtKB/SwissProt + PIR + PRF
refseq:	RefSeq protein sequences from NCBI's Reference Sequence Project.
swissprot:	The SWISSPROT part of UniProt Knowledge Base (UniProtKB)
pat:	Patented protein sequences
pdb:	Sequences of proteins in the Protein Data Bank (PDB) containing the 3-dimensional structure of proteins
env_nr:	Protein sequences from metagenomic projects and environmental samples.

BLAST databases (nucleotides)

nr:	All GenBank + RefSeq Nucleotides + EMBL + DDBJ + PDB sequences (excluding HTGS0,1,2, EST, GSS, STS, PAT, WGS). No longer "non-redundant”.
refseq_rna:	RNA entries from NCBI's Reference Sequence project
refseq_genomic:	Genomic entries from NCBI's Reference Sequence project
chromosome:	A database with complete genomes and chromosomes from the NCBI Reference Sequence project..
est:	Database of GenBank + EMBL + DDBJ sequences from EST Divisions
gss:	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
htgs:	Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2 (finished, phase 3 HTG sequences are in nr)
pat:	Nucleotides from the Patent division of GenBank.
pdb:	Sequences derived from the 3-dimensional structure from Protein Data Bank (PDB)
alu_repeats:	Human ALU repeat elements
dbsts:	Database of GenBank+EMBL+DDBJ sequences from STS Divisions .
wgs:	A database for whole genome shotgun sequence entries
tsa:	Transcriptome shotgun assembly
16S:	16S ribosomal RNA from Bacteria and Archaea


Firefox
Protein BLAST: search protein databases ...
blast.ncbi.nlm.nih.gov/Blast.cgi
Google
BLAST
Search database UniProtKB/Swiss-Prot(swissprot) using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
10
Word size
3
Max matches in a query range
0
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database UniProtKB/Swiss-Prot(swissprot) using Blastp (protein-protein BLAST)
Show results in a new window
BLAST is a registered trademark of the National Library of Medicine.
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

Firefox

NCBI Blast:Protein Sequence (216 letters)

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

Google

BLAST®

Basic Local Alignment Search Tool

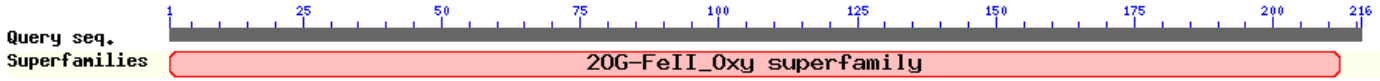
Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ **Formatting Results - 1ZKJE2S701N** [Formatting options]

Job Title: Protein Sequence (216 letters)

Putative conserved domains have been detected, click on the image below for detailed results.



Request ID	**1ZKJE2S701N**
Status	Searching
Submitted at	Mon Aug 6 10:01:50 2012
Current time	Mon Aug 6 10:01:53 2012
Time since submission	00:00:02

This page will be automatically updated in **7** seconds

Firefox

NCBI Blast:Protein Sequence (216 letters)

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

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - 1ZKJE2S701N

Edit and Resubmit | Save Search Strategies | Formatting options | Download

Protein Sequence (216 letters)

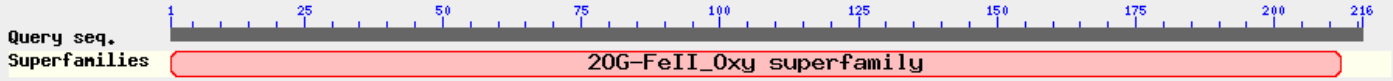
Query ID	lcl|94622	**Database Name**	swissprot
Description	unnamed protein product	**Description**	Non-redundant SwissProt sequences
Molecule type	amino acid	**Program**	BLASTP 2.2.26+ Citation
Query Length	216		

Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

Graphic Summary

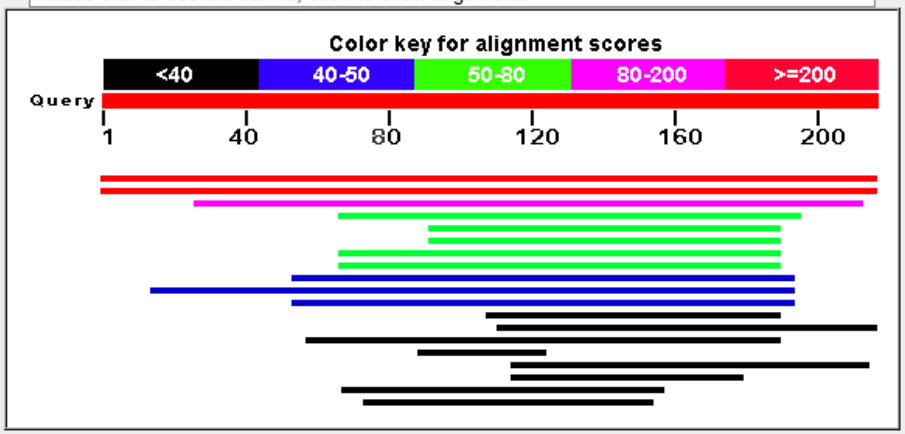
Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

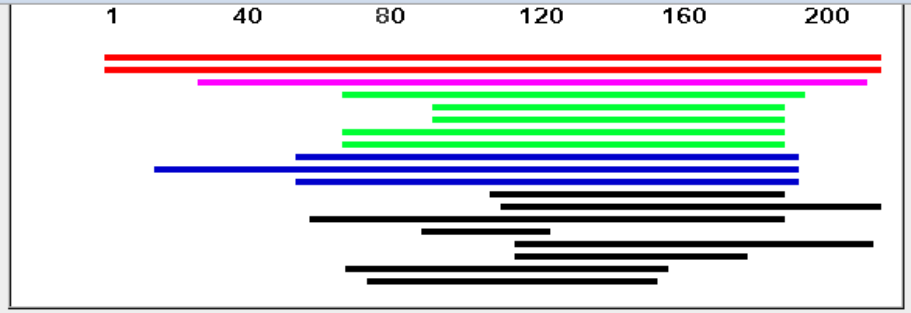


Distribution of 19 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments



Descriptions



Descriptions

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

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
P05050.1	RecName: Full=Alpha-ketoglutarate-dependent dioxygenase AlkB; Alt	451	451	100%	2e-161	100%	S
P37462.2	RecName: Full=Alpha-ketoglutarate-dependent dioxygenase AlkB; Alt	366	366	100%	8e-128	80%	
P0CAT7.1	RecName: Full=Alpha-ketoglutarate-dependent dioxygenase AlkB hom	144	144	86%	6e-41	42%	
Q9SA98.2	RecName: Full=Alpha-ketoglutarate-dependent dioxygenase alkB; Alt	76.3	76.3	59%	3e-15	37%	G M
P0CB42.1	RecName: Full=Alkylated DNA repair protein alkB homolog 1; AltName:	72.8	72.8	45%	7e-14	37%	G M
Q13686.2	RecName: Full=Alkylated DNA repair protein alkB homolog 1; AltName:	70.9	70.9	45%	4e-13	34%	G M
O60066.2	RecName: Full=Alpha-ketoglutarate-dependent dioxygenase abh1; Al	68.6	68.6	56%	1e-12	29%	G
Q54N08.1	RecName: Full=Alpha-ketoglutarate-dependent dioxygenase alkB; Alt	68.6	68.6	56%	2e-12	29%	
Q80Y20.1	RecName: Full=Alkylated DNA repair protein alkB homolog 8; AltName:	44.3	44.3	64%	4e-04	28%	G M
Q07G10.2	RecName: Full=Alkylated DNA repair protein alkB homolog 8; AltName:	42.7	42.7	82%	0.001	24%	G
A1A4L5.1	RecName: Full=Alkylated DNA repair protein alkB homolog 8; AltName:	40.8	40.8	64%	0.005	27%	G M
Q96BT7.2	RecName: Full=Alkylated DNA repair protein alkB homolog 8; AltName:	38.1	38.1	37%	0.036	36%	S G M
Q5UR03.1	RecName: Full=Uncharacterized protein L905	37.0	37.0	49%	0.058	27%	G
Q95K79.1	RecName: Full=Alkylated DNA repair protein alkB homolog 8; AltName:	37.7	37.7	61%	0.059	29%	
Q6DDW2.1	RecName: Full=UPF0577 protein KIAA1324-like homolog; AltName: Ful	33.5	33.5	16%	1.1	33%	G
B1PS76.1	RecName: Full=RNA-directed RNA polymerase; Includes: RecName: Fu	33.5	33.5	46%	1.5	30%	
Q67704.1	RecName: Full=RNA replication protein; Includes: RecName: Full=RNA	33.1	33.1	30%	1.9	36%	
Q8N661.2	RecName: Full=Lysoplasmalogenase; AltName: Full=Transmembrane p	32.0	32.0	41%	2.6	31%	G M
Q64962.1	RecName: Full=RNA replication protein; Includes: RecName: Full=RNA	31.2	31.2	37%	8.5	26%	

Alignments

Alignments

Select All Get selected sequences Distance tree of results Multiple alignment

>sp|P05050.1|ALKB_ECOLI RecName: Full=Alpha-ketoglutarate-dependent dioxygenase AlkB;
AltName: Full=Alkylated DNA repair protein AlkB
Length=216

 Score = 451 bits (1161), Expect = 2e-161, Method: Compositional matrix adjust.
 Identities = 216/216 (100%), Positives = 216/216 (100%), Gaps = 0/216 (0%)

```
Query  1    MLDLFADAEPWQEPLAAGAVILRRFAFNAAEQLIRDINDVASQSPFRQMVTPGGYTMSVA  60
            MLDLFADAEPWQEPLAAGAVILRRFAFNAAEQLIRDINDVASQSPFRQMVTPGGYTMSVA
Sbjct  1    MLDLFADAEPWQEPLAAGAVILRRFAFNAAEQLIRDINDVASQSPFRQMVTPGGYTMSVA  60

Query  61   MTNCGHLGWTTHRQGYLYSPIDPQTNKPWPAMPQSFHNLCQRAATAAGYPDFQPDACLIN  120
            MTNCGHLGWTTHRQGYLYSPIDPQTNKPWPAMPQSFHNLCQRAATAAGYPDFQPDACLIN
Sbjct  61   MTNCGHLGWTTHRQGYLYSPIDPQTNKPWPAMPQSFHNLCQRAATAAGYPDFQPDACLIN  120

Query  121  RYAPGAKLSLHQDKDEPDLRAPIVSVSLGLPAIFQFGGLKRNDPLKRLLLEHGDVVVWGG  180
            RYAPGAKLSLHQDKDEPDLRAPIVSVSLGLPAIFQFGGLKRNDPLKRLLLEHGDVVVWGG
Sbjct  121  RYAPGAKLSLHQDKDEPDLRAPIVSVSLGLPAIFQFGGLKRNDPLKRLLLEHGDVVVWGG  180

Query  181  ESRLFYHGIQPLKAGFHPLTIDCRYNLTFRQAGKKE  216
            ESRLFYHGIQPLKAGFHPLTIDCRYNLTFRQAGKKE
Sbjct  181  ESRLFYHGIQPLKAGFHPLTIDCRYNLTFRQAGKKE  216
```

>sp|P37462.2|ALKB_SALTY RecName: Full=Alpha-ketoglutarate-dependent dioxygenase AlkB;
AltName: Full=Alkylated DNA repair protein AlkB
Length=216

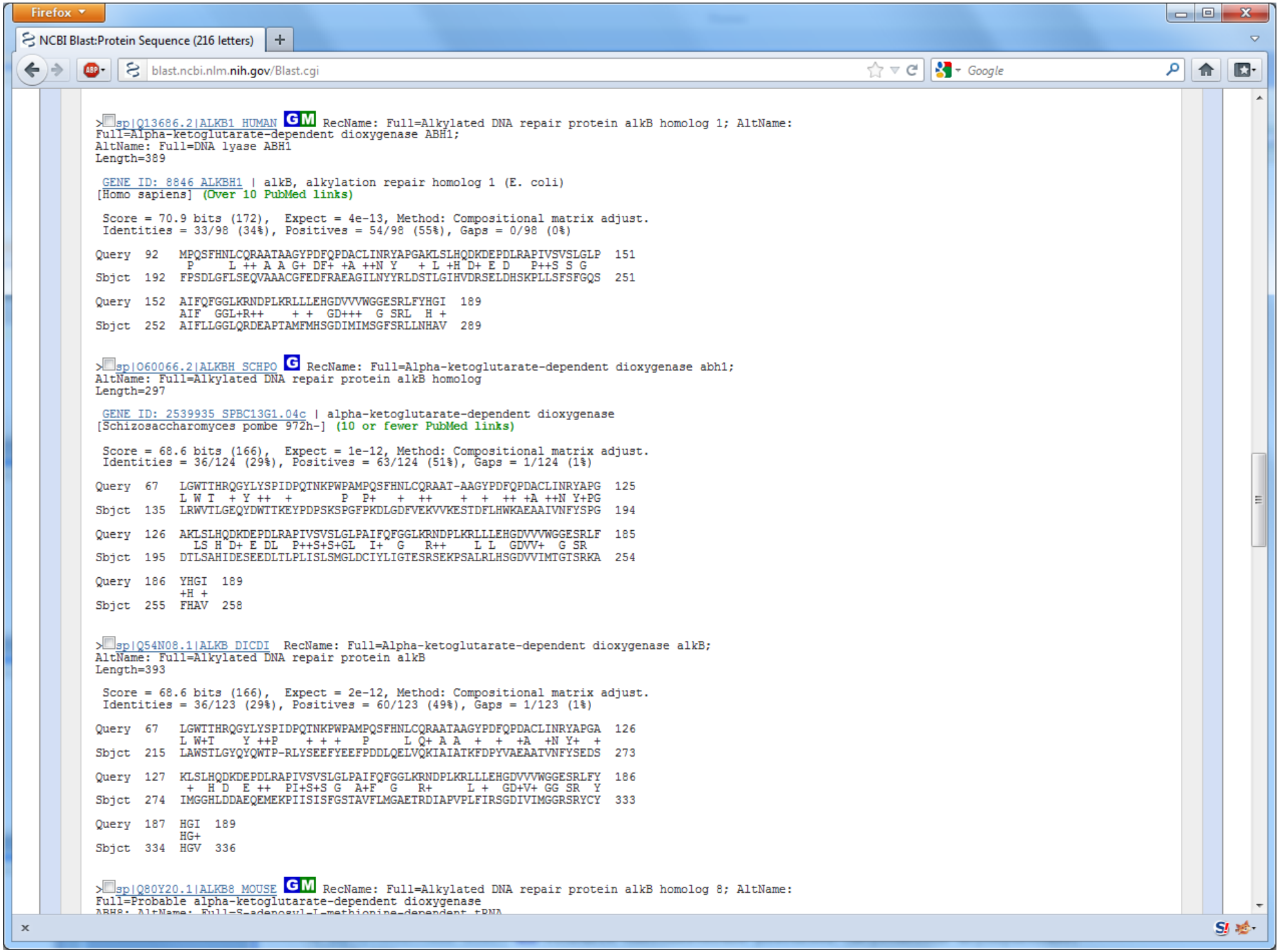
 Score = 366 bits (940), Expect = 8e-128, Method: Compositional matrix adjust.
 Identities = 172/216 (80%), Positives = 193/216 (89%), Gaps = 0/216 (0%)

```
Query  1    MLDLFADAEPWQEPLAAGAVILRRFAFNAAEQLIRDINDVASQSPFRQMVTPGGYTMSVA  60
            MLDLFAD  PWQEPLA GAV+LRRFAF AA+ L+ DI  VASQSPFRQMVTPGGYTMSVA
Sbjct  1    MLDLFADEAPWQEPLAPGAVVLRRFAFRAAQSLLDDIGFVASQSPFRQMVTPGGYTMSVA  60

Query  61   MTNCGHLGWTTHRQGYLYSPIDPQTNKPWPAMPQSFHNLCQRAATAAGYPDFQPDACLIN  120
            MTNCG LGWTT R GY Y+  DP T+KPWPA+P SF ++C++AA AAGY  FQPDACLIN
Sbjct  61   MTNCGALGWTTDRHGYCYAVRDPLTDKPWPALPLSFASVCRQAAIAAGYASFQPDACLIN  120

Query  121  RYAPGAKLSLHQDKDEPDLRAPIVSVSLGLPAIFQFGGLKRNDPLKRLLLEHGDVVVWGG  180
            RYAPGAKLSLHQDKDEPDLRAPIVSVSLG+PA+FQFGGL+R+DP++R+LLEHGD+VVWGG
Sbjct  121  RYAPGAKLSLHQDKDEPDLRAPIVSVSLGVPAVFQFGGLRRSDPIQRILLEHGDIVVWGG  180

Query  181  ESRLFYHGIQPLKAGFHPLTIDCRYNLTFRQAGKKE  216
            ESRLFYHGIQPLKAGFHP+T + RYNLTFRQA +KE
Sbjct  181  ESRLFYHGIQPLKAGFHPMTGEFRYNLTFRQAAEKE  216
```

>sp|P0CAT7.1|ALKB_CAUCR RecName: Full=Alpha-ketoglutarate-dependent dioxygenase AlkB
homolog
 sp|B8GWW6.2|ALKB_CAUCN RecName: Full=Alpha-ketoglutarate-dependent dioxygenase AlkB
homolog
Length=220

 Score = 144 bits (364), Expect = 6e-41, Method: Compositional matrix adjust.
 Identities = 79/187 (42%), Positives = 107/187 (57%), Gaps = 5/187 (3%)

```
Query  27   FNAAEQLIRDINDVASQSPFRQMVTPGGYTMSVAMTNCGHLGWTTHRQGYLYSPIDPQTN  86
```



>sp|Q13686.2|ALKB1_HUMAN GM RecName: Full=Alkylated DNA repair protein alkB homolog 1; AltName: Full=Alpha-ketoglutarate-dependent dioxygenase ABH1;
AltName: Full=DNA lyase ABH1
Length=389

GENE ID: 8846 ALKBH1 | alkB, alkylation repair homolog 1 (E. coli)
[Homo sapiens] (Over 10 PubMed links)

Score = 70.9 bits (172), Expect = 4e-13, Method: Compositional matrix adjust.
Identities = 33/98 (34%), Positives = 54/98 (55%), Gaps = 0/98 (0%)

```
Query  92   MPQSFHNLCQRAATAAGYPDFQPDACLINRYAPGAKLSLHQDKDEPDLRAPIVSVSLGLP  151
            P     L ++ A A G+ DF+ +A ++N Y   + L +H D+ E D   P++S S G
Sbjct  192  FPSDLGFLSEQVAAACGFEDFRAEAGILNYYRLDSTLGIHVDRSELDHSKPLLSFSFGQS  251

Query  152  AIFQFGGLKRNDPLKRLLLEHGDVVVWGGESRLFYHGI  189
            AIF  GGL+R++    + +  GD+++  G SRL  H +
Sbjct  252  AIFLLGGLQRDEAPTAMFMHSGDIMIMSGFSRLLNHAV  289
```

>sp|O60066.2|ALKBH_SCHPO G RecName: Full=Alpha-ketoglutarate-dependent dioxygenase abh1;
AltName: Full=Alkylated DNA repair protein alkB homolog
Length=297

GENE ID: 2539935 SPBC13G1.04c | alpha-ketoglutarate-dependent dioxygenase
[Schizosaccharomyces pombe 972h-] (10 or fewer PubMed links)

Score = 68.6 bits (166), Expect = 1e-12, Method: Compositional matrix adjust.
Identities = 36/124 (29%), Positives = 63/124 (51%), Gaps = 1/124 (1%)

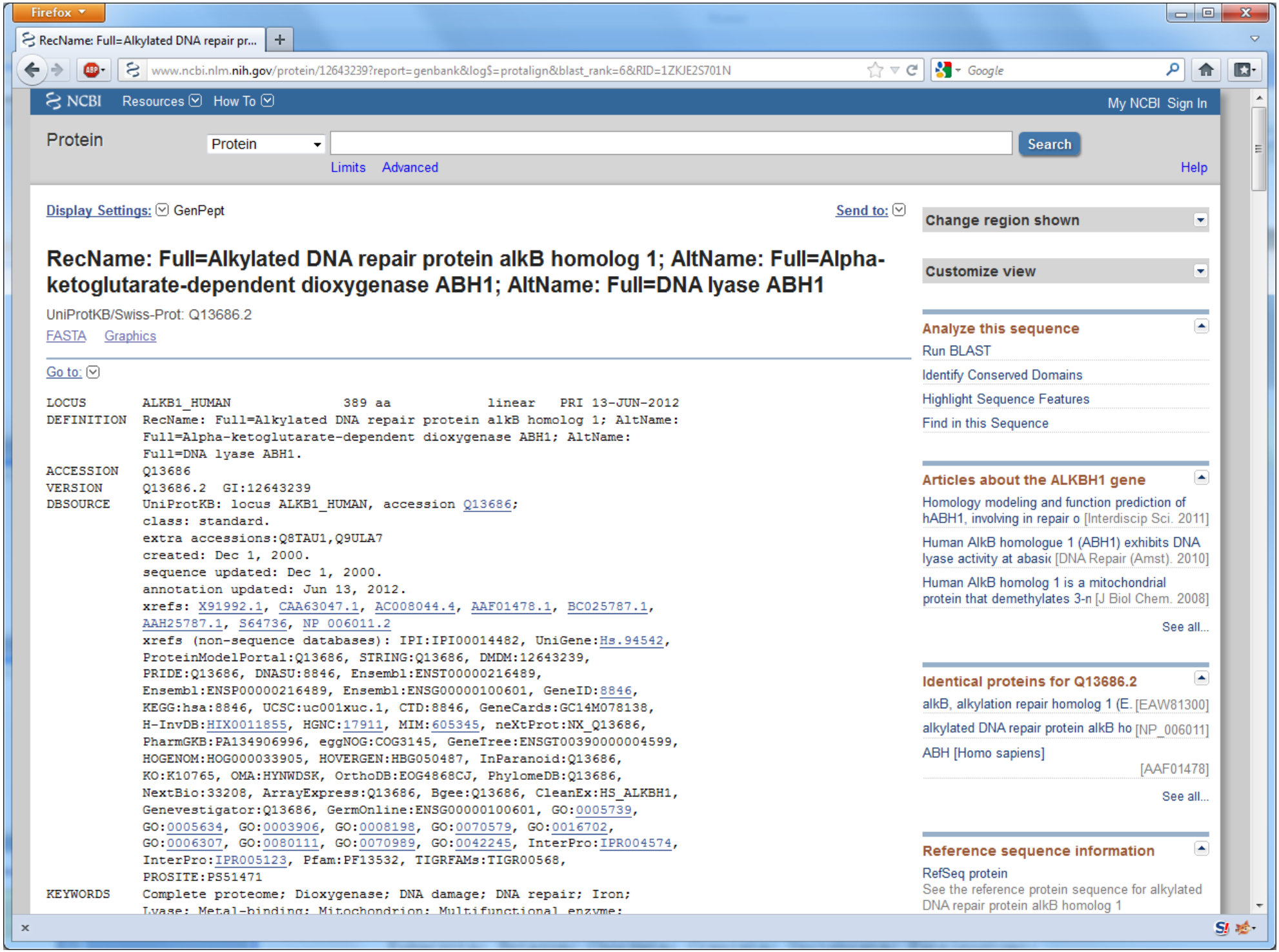
```
Query  67   LGWTTHRQGYLYSPIDPQTNKPWPAMPQSFHNLCQRAAT-AAGYPDFQPDACLINRYAPG  125
            L W T  + Y ++  +       P  P+   +  ++    +  +  ++ +A ++N Y+PG
Sbjct  135  LRWVTLGEQYDWTTKEYPDPSKSPGFPKDLGDFVEKVVKESTDFLHWKAEAAIVNFYSPG  194

Query  126  AKLSLHQDKDEPDLRAPIVSVSLGLPAIFQFGGLKRNDPLKRLLLEHGDVVVWGGESRLF  185
              LS H D+ E DL  P++S+S+GL  I+  G   R++    L L  GDVV+  G SR
Sbjct  195  DTLSAHIDESEEDLTLPLISLSMGLDCIYLIGTESRSEKPSALRLHSGDVVIMTGTSRKA  254

Query  186  YHGI  189
            +H +
Sbjct  255  FHAV  258
```

>sp|Q54N08.1|ALKB_DICDI RecName: Full=Alpha-ketoglutarate-dependent dioxygenase alkB;
AltName: Full=Alkylated DNA repair protein alkB
Length=393

Score = 68.6 bits (166), Expect = 2e-12, Method: Compositional matrix adjust.
Identities = 36/123 (29%), Positives = 60/123 (49%), Gaps = 1/123 (1%)

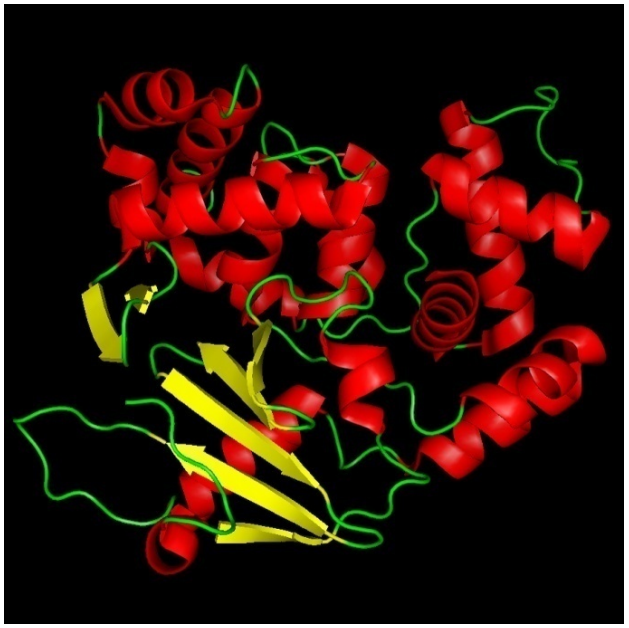
```
Query  67   LGWTTHRQGYLYSPIDPQTNKPWPAMPQSFHNLCQRAATAAGYPDFQPDACLINRYAPGA  126
            L W+I    Y ++P    + + +   P     L Q+ A A  +  +  +A  +N Y+  +
Sbjct  215  LAWSTLGYQYQWTP-RLYSEEFYEEFPDDLQELVQKIAIATKFDPYVAEAATVNFYSEDS  273

Query  127  KLSLHQDKDEPDLRAPIVSVSLGLPAIFQFGGLKRNDPLKRLLLEHGDVVVWGGESRLFY  186
             +  H D  E ++  PI+S+S G  A+F  G   R+     L +  GD+V+  G SR  Y
Sbjct  274  IMGGHLDDAEQEMEKPIISISFGSTAVFLMGAETRDIAPVPLFIRSGDIVIMGGRSRYCY  333

Query  187  HGI  189
            HG+
Sbjct  334  HGV  336
```

>sp|Q80Y20.1|ALKB8_MOUSE GM RecName: Full=Alkylated DNA repair protein alkB homolog 8; AltName:
Full=Probable alpha-ketoglutarate-dependent dioxygenase
ABH8; AltName: Full=S-adenosyl-L-methionine-dependent tRNA


Firefox
RecName: Full=Alkylated DNA repair pr...
www.ncbi.nlm.nih.gov/protein/12643239?report=genbank&log$=protalign&blast_rank=6&RID=1ZKJE2S701N
Google
NCBI Resources How To
My NCBI Sign In
Protein
Protein
Search
Limits Advanced
Help
Display Settings: GenPept
Send to:
Change region shown
Customize view
RecName: Full=Alkylated DNA repair protein alkB homolog 1; AltName: Full=Alpha-ketoglutarate-dependent dioxygenase ABH1; AltName: Full=DNA lyase ABH1
UniProtKB/Swiss-Prot: Q13686.2
FASTA Graphics
Go to:
LOCUS ALKB1_HUMAN 389 aa linear PRI 13-JUN-2012
DEFINITION RecName: Full=Alkylated DNA repair protein alkB homolog 1; AltName: Full=Alpha-ketoglutarate-dependent dioxygenase ABH1; AltName: Full=DNA lyase ABH1.
ACCESSION Q13686
VERSION Q13686.2 GI:12643239
DBSOURCE UniProtKB: locus ALKB1_HUMAN, accession Q13686;
class: standard.
extra accessions:Q8TAU1,Q9ULA7
created: Dec 1, 2000.
sequence updated: Dec 1, 2000.
annotation updated: Jun 13, 2012.
xrefs: X91992.1, CAA63047.1, AC008044.4, AAF01478.1, BC025787.1, AAH25787.1, S64736, NP_006011.2
xrefs (non-sequence databases): IPI:IPI00014482, UniGene:Hs.94542, ProteinModelPortal:Q13686, STRING:Q13686, DMDM:12643239, PRIDE:Q13686, DNASU:8846, Ensembl:ENST00000216489, Ensembl:ENSP00000216489, Ensembl:ENSG00000100601, GeneID:8846, KEGG:hsa:8846, UCSC:uc001xuc.1, CTD:8846, GeneCards:GC14M078138, H-InvDB:HIX0011855, HGNC:17911, MIM:605345, neXtProt:NX_Q13686, PharmGKB:PA134906996, eggNOG:COG3145, GeneTree:ENSGT00390000004599, HOGENOM:HOG000033905, HOVERGEN:HBG050487, InParanoid:Q13686, KO:K10765, OMA:HYNWDSK, OrthoDB:EOG4868CJ, PhylomeDB:Q13686, NextBio:33208, ArrayExpress:Q13686, Bgee:Q13686, CleanEx:HS_ALKBH1, Genevestigator:Q13686, GermOnline:ENSG00000100601, GO:0005739, GO:0005634, GO:0003906, GO:0008198, GO:0070579, GO:0016702, GO:0006307, GO:0080111, GO:0070989, GO:0042245, InterPro:IPR004574, InterPro:IPR005123, Pfam:PF13532, TIGRFAMs:TIGR00568, PROSITE:PS51471
KEYWORDS Complete proteome; Dioxygenase; DNA damage; DNA repair; Iron; Lyase; Metal-binding; Mitochondrion; Multifunctional enzyme;
Analyze this sequence
Run BLAST
Identify Conserved Domains
Highlight Sequence Features
Find in this Sequence
Articles about the ALKBH1 gene
Homology modeling and function prediction of hABH1, involving in repair o [Interdiscip Sci. 2011]
Human AlkB homologue 1 (ABH1) exhibits DNA lyase activity at abasi [DNA Repair (Amst). 2010]
Human AlkB homolog 1 is a mitochondrial protein that demethylates 3-n [J Biol Chem. 2008]
See all...
Identical proteins for Q13686.2
alkB, alkylation repair homolog 1 (E. [EAW81300]
alkylated DNA repair protein alkB ho [NP_006011]
ABH [Homo sapiens]
[AAF01478]
See all...
Reference sequence information
RefSeq protein
See the reference protein sequence for alkylated DNA repair protein alkB homolog 1

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

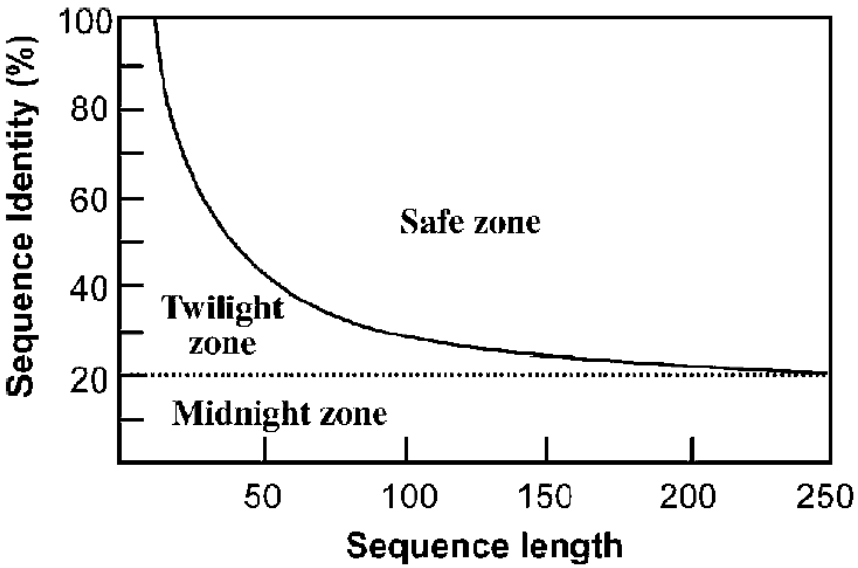


Figure 3.1: The three zones of protein sequence alignments. Two protein sequences can be regarded as homologous if the percentage sequence identity falls in the safe zone. Sequence identity values below the zone boundary, but above 20%, are considered to be in the twilight zone, where homologous relationships 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. …AGAGAGAGAGAGAGAGAG…
    - E.g. …CCGCCGCCGCCGCCGCCGCCG…
  - low complexity regions,
    - Protein example: PPCDPPPPPKDKKKKDDGPP
    - 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)

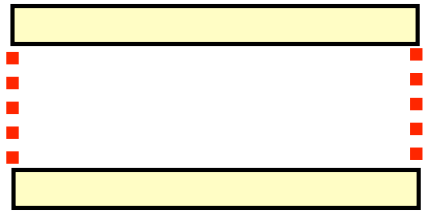
- Precision (PPV)
  - Ability to distinguish between homologous sequences and non-homologous sequences
  - The fraction of truly homologous sequences found (with a score above a certain threshold) among all sequences found
  - True positives / (True positives + False positives)

- Speed

Global and local alignments

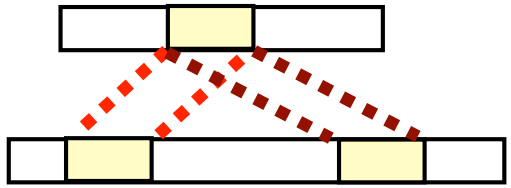
Global alignment:

- 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.



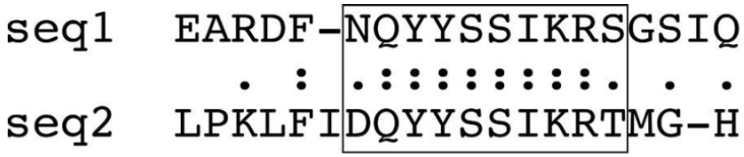
Local alignment:

- 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).

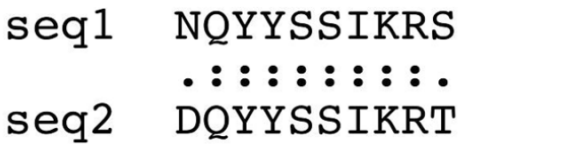


Global and local alignments

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.



global sequence alignment



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

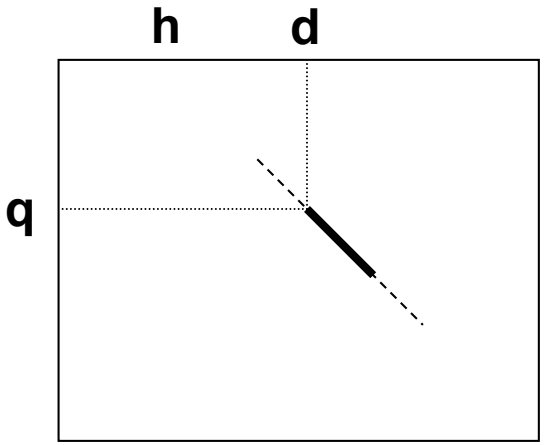
Definition

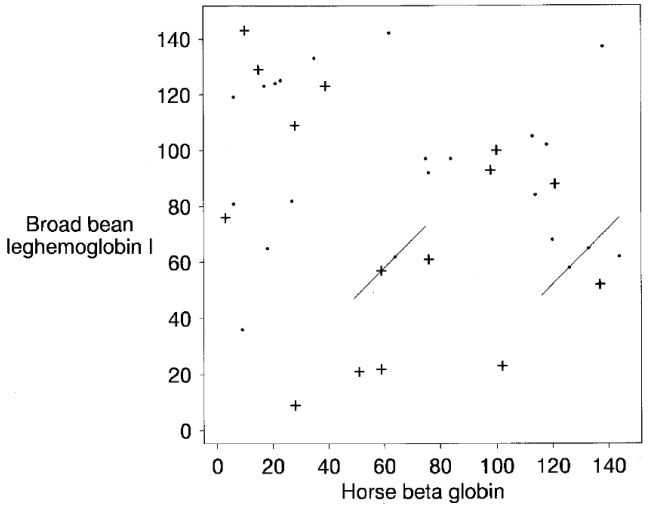
- A *maximal segment pair* (MSP_{qd}) is a pair of identical length segments chosen from the sequences q and d, which when aligned have the highest possible score obtained for local ungapped alignment of q and d.
- A *high-scoring segment pair* (HSP) is a segment pair which does not increase its score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

Δ

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
      .
```