Searching sequence databases

MBV-INF4410
Thursday 10 September 2009

Torbjørn Rognes
Centre for Molecular Biology and Neuroscience (CMBN), Rikshospitalet
& Department of Informatics, University of Oslo
torognes@ifi.uio.no

Overview of the presentation

- An example showing how useful database searches can be
- Searching sequence databases
- A walk-through of the BLAST search service
  - BLAST variants
  - About sequence databases
- What is a good match?
- What is homology?
- Significance of alignments
- Problematic sequences

One example of how useful database searches can be

- The protein AlkB was discovered in the bacterium *E.coli* more than 20 years ago.
- 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

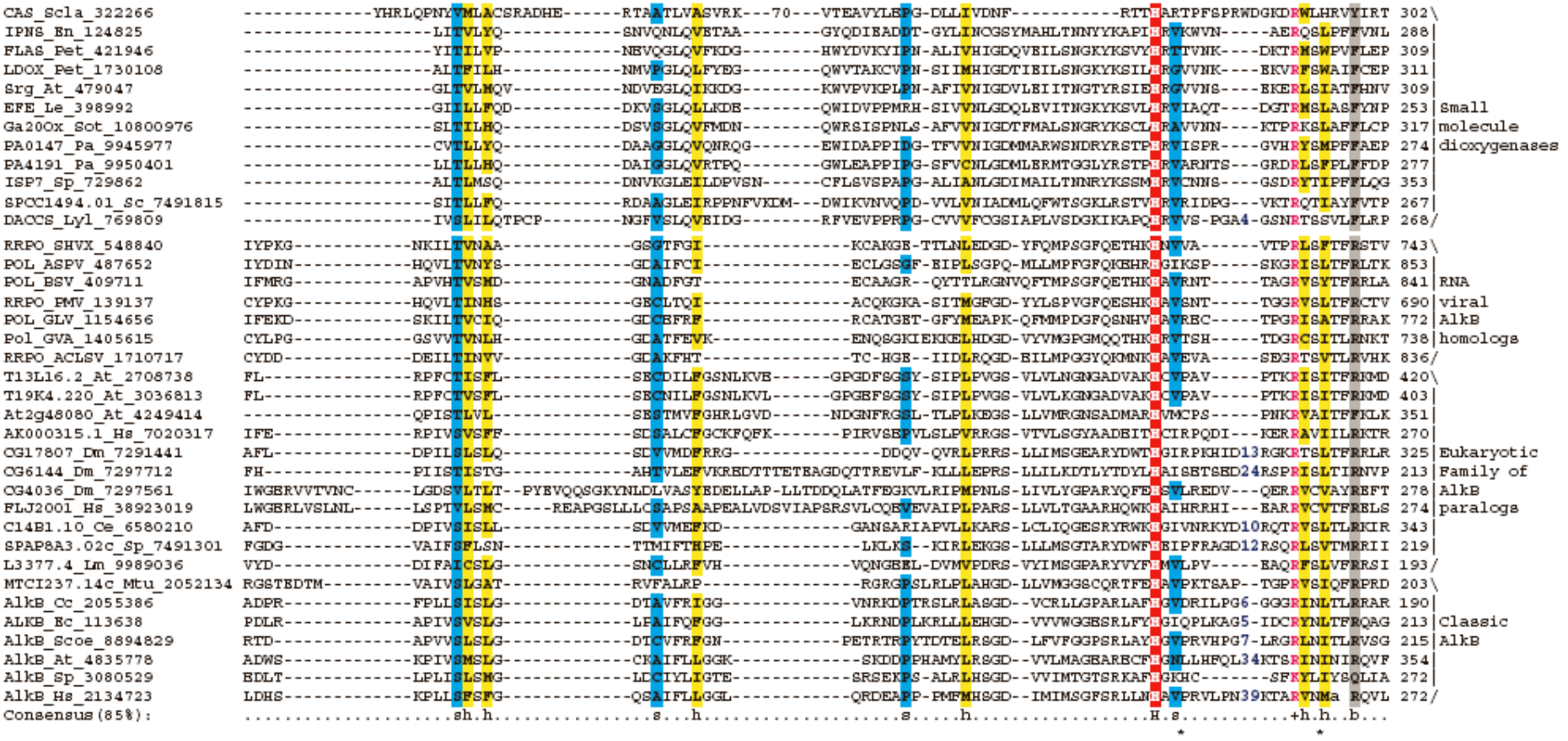
L Aravind and Eugene V Koonin

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…

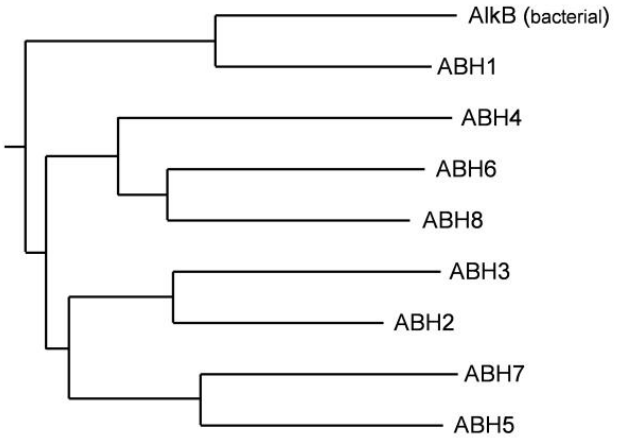
- By comparing *E.coli* AlkB to other sequences in the database it was found that AlkB had some features in common with more well-known enzymes

- 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…

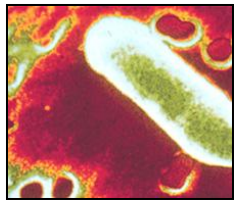
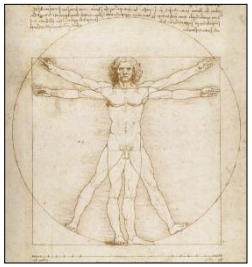
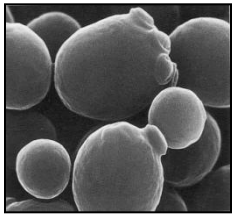
- By further sequence analysis 3 AlkB-like sequences were found in humans:
  - ALKBH1
  - ALKBH2
  - ALKBH3
- And by even more advanced analysis another 5 homologs were found in humans:
  - ALKBH4
  - ALKBH5
  - ALKBH6
  - ALKBH7
  - ALKBH8



- 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 900 completely sequenced genomes available – an enormous source of information. More than 3 000 other genomes underway...
- More than 250 000 000 000 basepairs sequenced to date *
- Database sizes are growing exponentially – doubling in about 18 months
- 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.




* Source: 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**

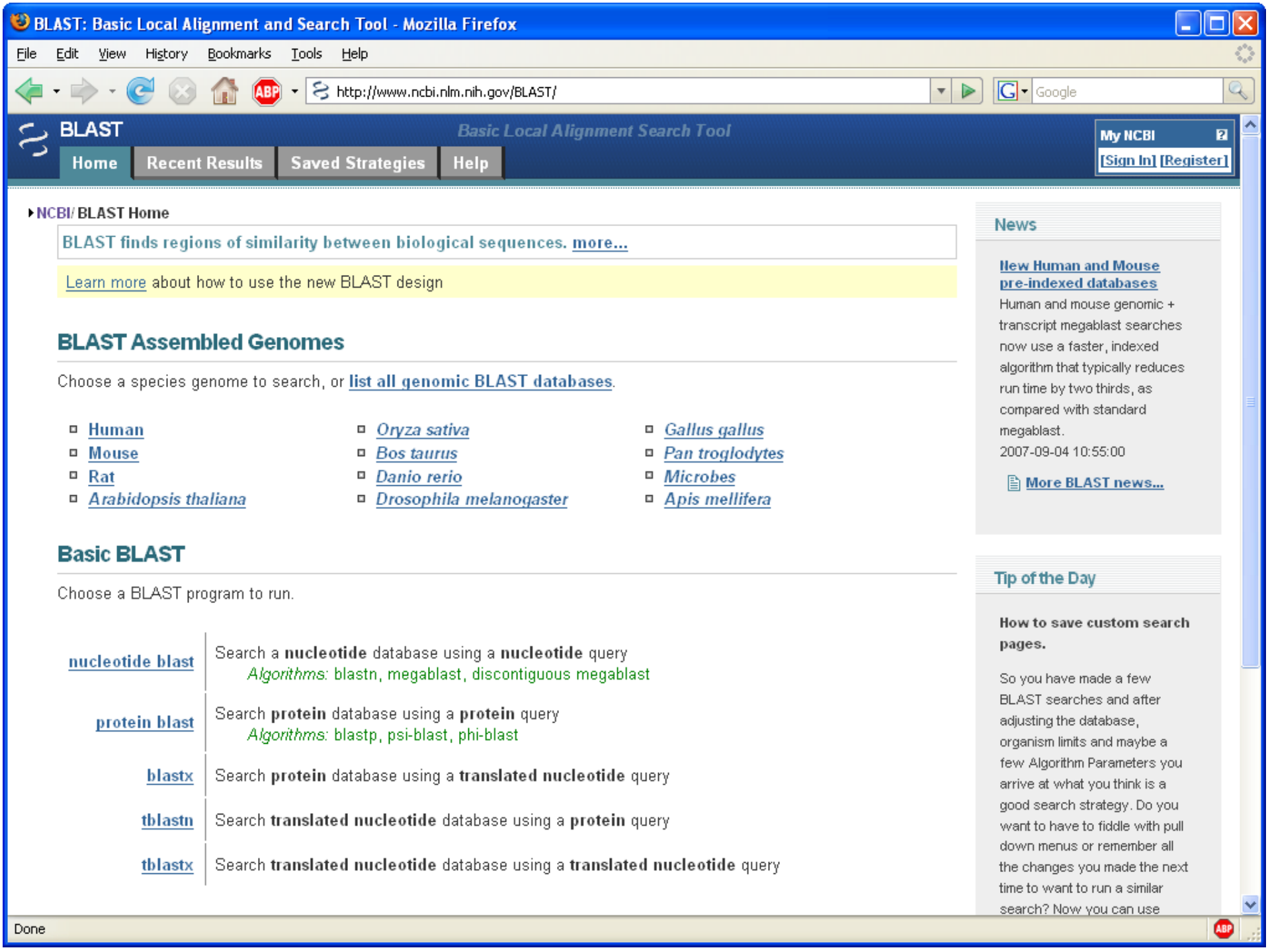
Database sequences:

Identical (trivial): **acgatcgattagcca**

Very similar (easy): **acgaccgatgagcca**

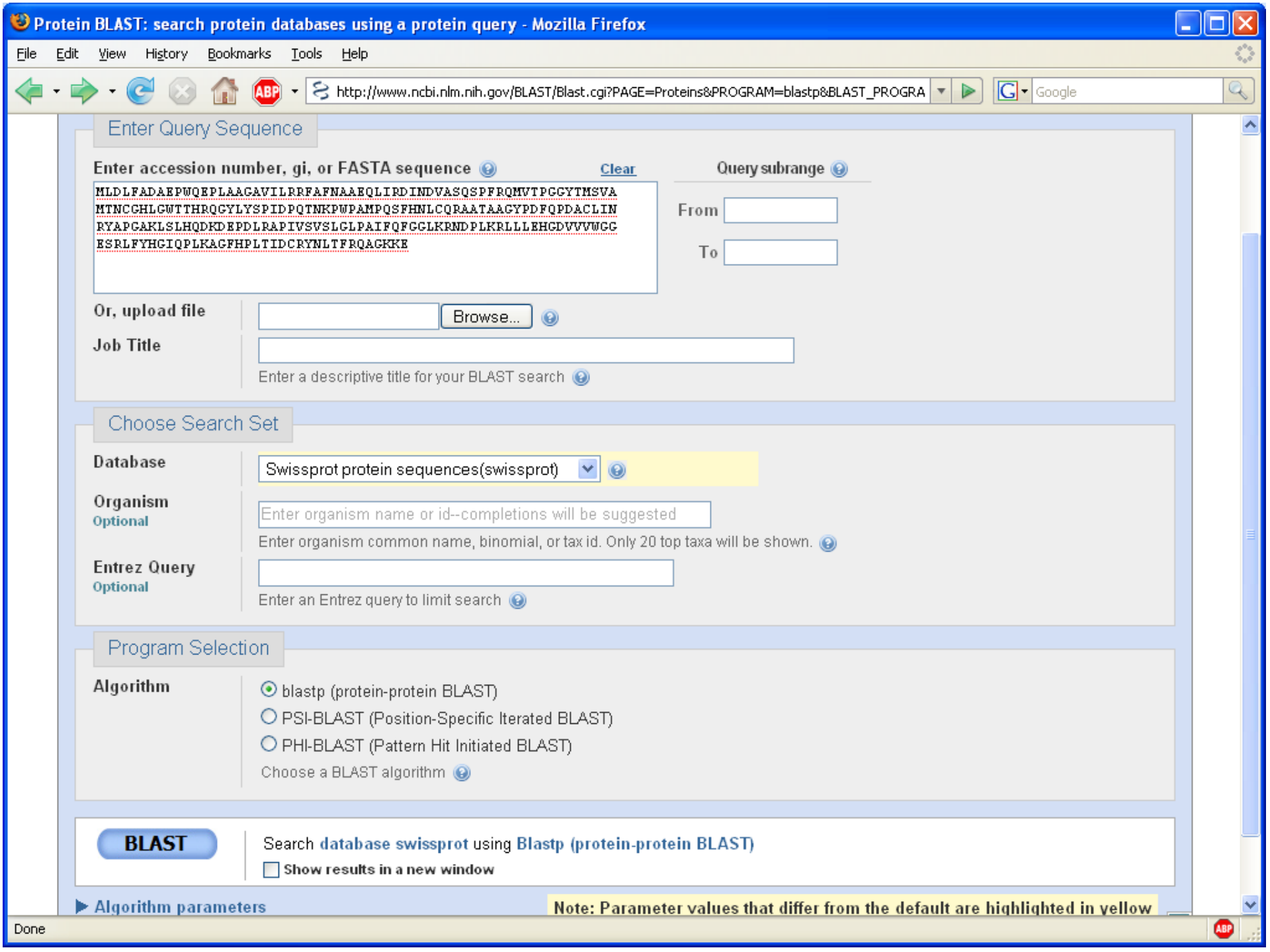
Similar (moderate): **atgacggatgagcga**

Very diverged (hard): **atgacgggatgagcga**


BLAST: Basic Local Alignment and Search Tool - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/BLAST/
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...
Learn more about how to use the new BLAST design
BLAST Assembled 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
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
News
New Human and Mouse pre-indexed databases
Human and mouse genomic + transcript megablast searches now use a faster, indexed algorithm that typically reduces run time by two thirds, as compared with standard megablast.
2007-09-04 10:55:00
More BLAST news...
Tip of the Day
How to save custom search pages.
So you have made a few BLAST searches and after adjusting the database, organism limits and maybe a few Algorithm Parameters you arrive at what you think is a good search strategy. Do you want to have to fiddle with pull down menus or remember all the changes you made the next time to want to run a similar search? Now you can use
Done

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.

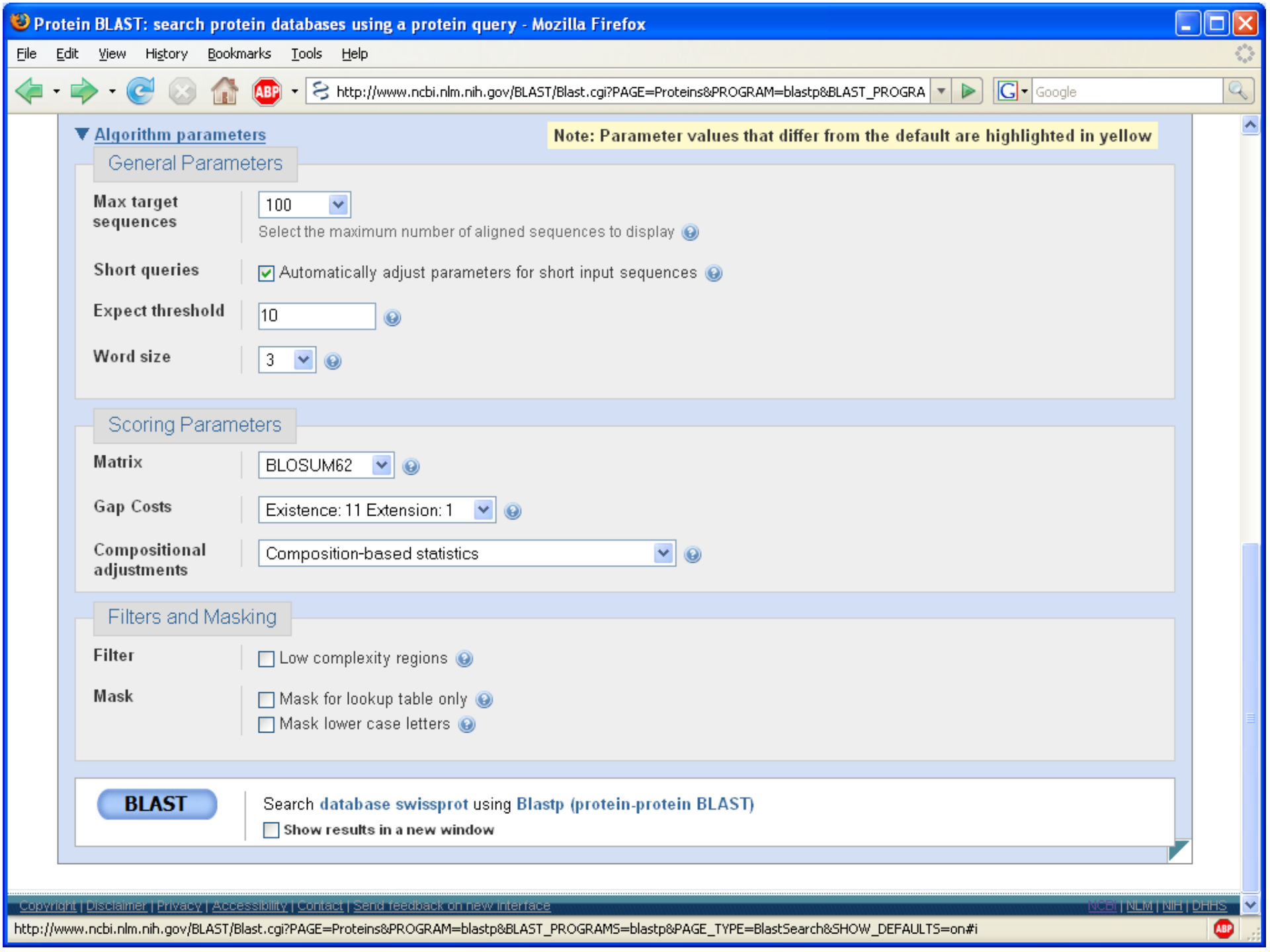

Protein BLAST: search protein databases using a protein query - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRA
Google
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
MLDLFADAEPWQEPLAAGAVILRRFAFNAAEQLIRDINDVASQSPFRQMVTPGGYTMSVA
MTNCGHLGWTTHRQGYLYSPIDPQTNKPWPAMPQSFHNLCQRAATAAGYPDFQPDACLIN
RYAPGAKLSLHQDKDEPDLRAPIVSVSLGLPAIFQFGGLKRNDPLKRLLLEHGDVVVWGG
ESRLFYHGIQPLKAGFHPLTIDCRYNLTFRQAGKKE
Query subrange
From
To
Or, upload file
Browse...
Job Title
Enter a descriptive title for your BLAST search
Choose Search Set
Database
Swissprot protein sequences(swissprot)
Organism
Optional
Enter organism name or id--completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
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)
Choose a BLAST algorithm
BLAST
Search database 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
Done

BLAST databases (protein)

nr:	All non-redundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF
refseq:	RefSeq protein sequences from NCBI's Reference Sequence Project.
swissprot:	Last major release of the SWISS-PROT protein sequence database (no updates).
pat:	Proteins from the Patent division of GenPept.
pdb:	Sequences derived from the 3-dimensional structure from Brookhaven Protein Data Bank.
env_nr:	Protein sequences from 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
est:	Database of GenBank + EMBL + DDBJ sequences from EST Divisions
est_human:	Human subset of est.
est_mouse:	Mouse subset.
est_others:	Non-Mouse, non-Human subset of est.
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 Brookhaven Protein Data Bank
month:	All new or revised GenBank + EMBL + DDBJ + PDB sequences released in the last 30 days.
dbsts:	Database of GenBank+EMBL+DDBJ sequences from STS Divisions .
chromosome:	A database with complete genomes and chromosomes from the NCBI Reference Sequence project..
wgs:	A database for whole genome shotgun sequence entries.
env_nt:	Nucleotide sequences from environmental samples, including those from Sargasso Sea and Mine Drainage projects.


Protein BLAST: search protein databases using a protein query - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRA
Google
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow
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
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Composition-based statistics
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database swissprot using Blastp (protein-protein BLAST)
Show results in a new window
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback on new interface
NCBI | NLM | NIH | DHHS
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on#i

BLAST *Basic Local Alignment Search Tool*

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp/ Formatting Results - JE920YVS012 [Formatting options]

Job Title: Protein sequence(216 letters)

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

Request ID	**JE920YVS012**
Status	Searching
Submitted at	Tue Oct 30 08:32:00 2007
Current time	Tue Oct 30 08:32:07 2007
Time since submission	

This page will be automatically updated in **12** seconds

Job Title: Protein sequence(216 letters)

Show Conserved Domains

BLASTP 2.2.17 (Aug-26-2007)

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Reference:
Schäffer, Alejandro A., L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", Nucleic Acids Res. 29:2994-3005.

RID: JE920YVS012

Database: Non-redundant SwissProt sequences
259,844 sequences; 98,643,457 total letters

If you have any problems or questions with the results of this search please refer to the **BLAST FAQs**
Taxonomy reports

Query=
Length=216

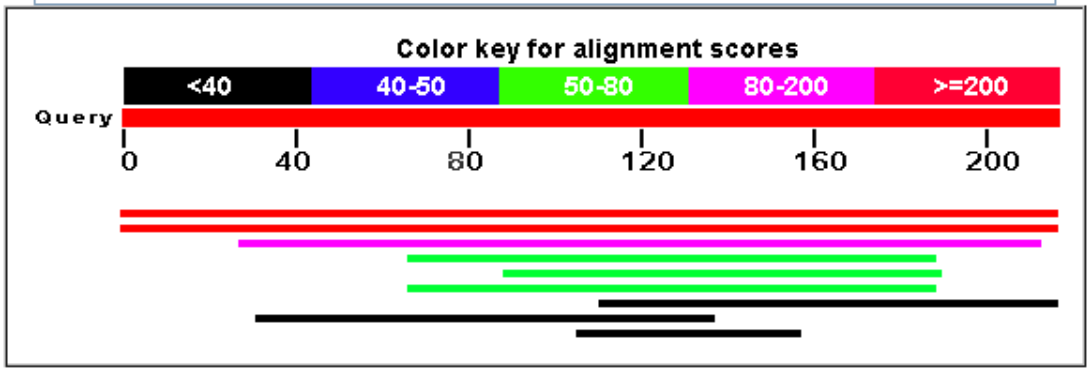
Distribution of 9 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Color key for alignment scores

Distribution of 9 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments



Distance tree of results NEW

Sequences producing significant alignments:		Score (Bits)	E Value	
sp|P05050.1|ALKB_ECOLI	Alpha-ketoglutarate-dependent dioxygen...	450	2e-126	
sp|P37462|ALKB_SALTY	Alpha-ketoglutarate-dependent dioxygenas...	363	3e-100	
sp|O05725|ALKB_CAUCR	Alpha-ketoglutarate-dependent dioxygenase a	139	6e-33	
sp|Q9SA98|ALKBH_ARATH	Alkylated DNA repair protein alkB homolog	75.5	2e-13	
sp|Q13686|ALKB1_HUMAN	Alkylated DNA repair protein alkB homolog	69.3	1e-11	G
sp|O60066|ALKBH_SCHPO	Alkylated DNA repair protein alkB homolog	68.2	2e-11	
sp|Q5UR03|YL905_MIMIV	Uncharacterized protein L905	36.2	0.085	
sp|Q5UR02|CHLE_MIMIV	Probable cholinesterase precursor (Acylchol	30.8	4.0	G
sp|Q8N661|TM86B_HUMAN	Transmembrane protein 86B	30.0	7.6	G

Alignments

Get selected sequences | Select all | Deselect all | Distance tree of results

>sp|P05050.1|ALKB_ECOLI Alpha-ketoglutarate-dependent dioxygenase alkB (Alkylated DNA repair protein alkB)
Length=216

>sp|P05050.1|ALKB_ECOLI Alpha-ketoglutarate-dependent dioxygenase alkB (Alkylated DNA repair protein alkB)
Length=216

Score = 450 bits (1157), Expect = 2e-126, Method: Composition-based stats.
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
```

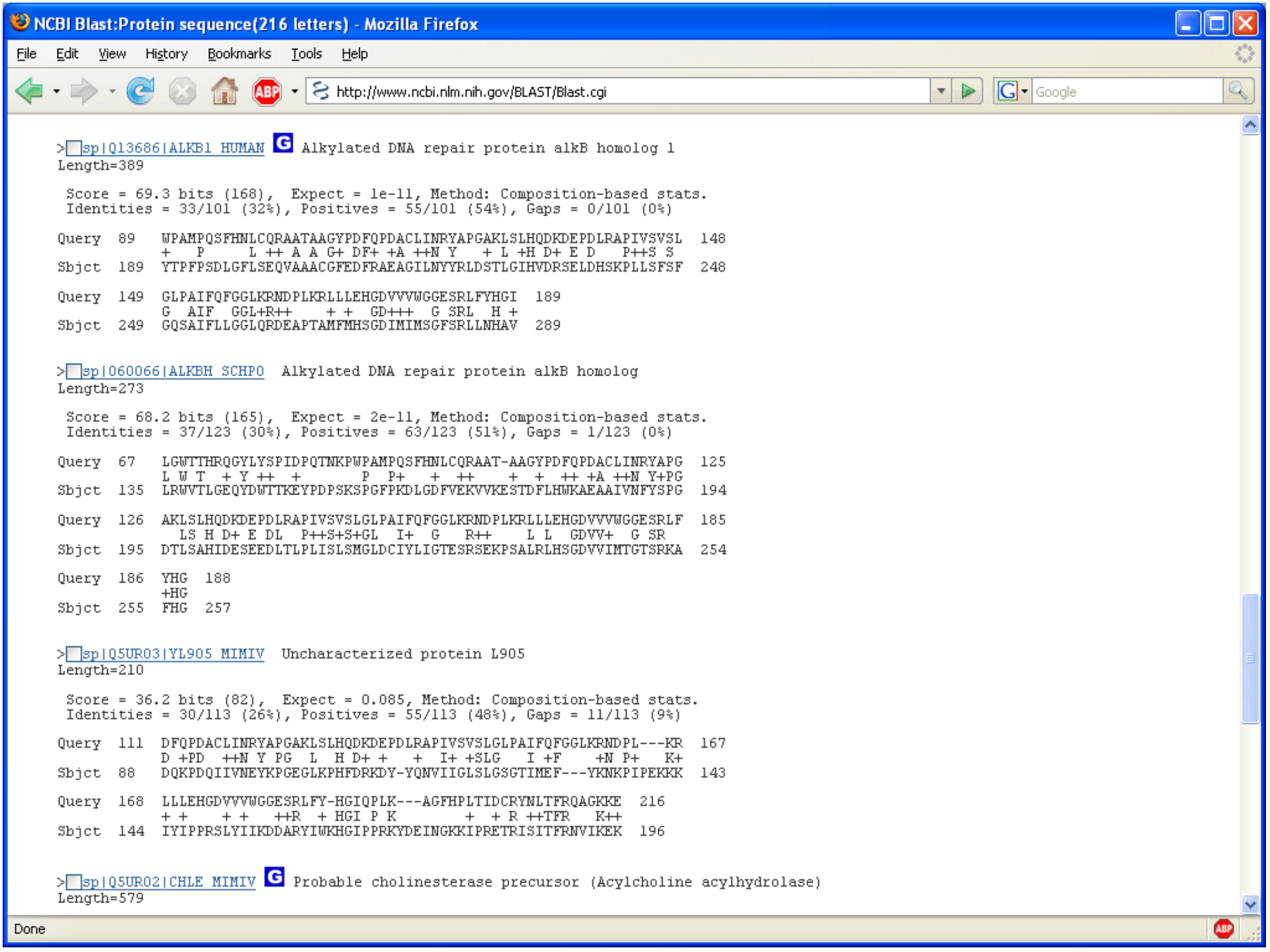
>sp|P37462|ALKB_SALTY Alpha-ketoglutarate-dependent dioxygenase alkB (Alkylated DNA repair protein alkB)
Length=216

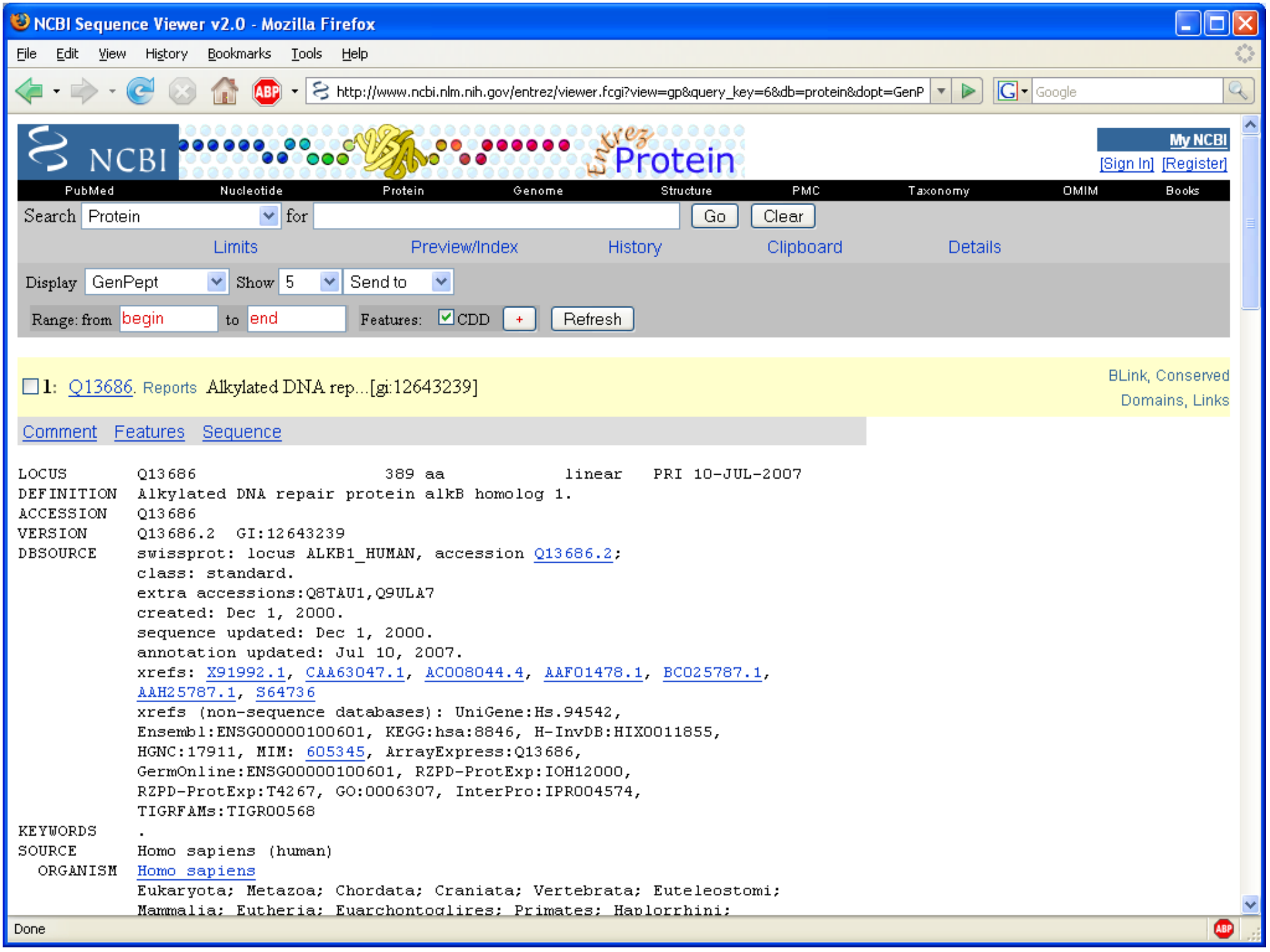
Score = 363 bits (932), Expect = 3e-100, Method: Composition-based stats.
Identities = 172/216 (79%), Positives = 193/216 (89%), Gaps = 0/216 (0%)

```
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|O05725|ALKB_CAUCR Alpha-ketoglutarate-dependent dioxygenase alkB homolog
Length=220

Score = 139 bits (351), Expect = 6e-33, Method: Composition-based stats.


NCBI Blast:Protein sequence(216 letters) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Google
>sp|Q13686|ALKB1_HUMAN Alkylated DNA repair protein alkB homolog 1
Length=389
Score = 69.3 bits (168), Expect = 1e-11, Method: Composition-based stats.
Identities = 33/101 (32%), Positives = 55/101 (54%), Gaps = 0/101 (0%)
Query 89 WPAMPQSFHNLCQRAATAAGYPDFQPDACLINRYAPGAKLSLHQDKDEPDLRAPIVSVSL 148
+ P L ++ A A G+ DF+ +A ++N Y + L +H D+ E D P++S S
Sbjct 189 YTPFPSDLGFLSEQVAAACGFEDFRAEAGILNYYRLDSTLGIHVDRSELDHSKPLLSFSF 248
Query 149 GLPAIFQFGGLKRNDPLKRLLLEHGDVVVWGGESRLFYHGI 189
G AIF GGL+R++ + + GD+++ G SRL H +
Sbjct 249 GQSAIFLLGGLQRDEAPTAMFMHSGDIMIMSGFSRLLNHAV 289
>sp|O60066|ALKBH_SCHPO Alkylated DNA repair protein alkB homolog
Length=273
Score = 68.2 bits (165), Expect = 2e-11, Method: Composition-based stats.
Identities = 37/123 (30%), Positives = 63/123 (51%), Gaps = 1/123 (0%)
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 YHG 188
+HG
Sbjct 255 FHG 257
>sp|Q5UR03|YL905_MIMIV Uncharacterized protein L905
Length=210
Score = 36.2 bits (82), Expect = 0.085, Method: Composition-based stats.
Identities = 30/113 (26%), Positives = 55/113 (48%), Gaps = 11/113 (9%)
Query 111 DFQPDACLINRYAPGAKLSLHQDKDEPDLRAPIVSVSLGLPAIFQFGGLKRNDPL---KR 167
D +PD ++N Y PG L H D+ + + I+ +SLG I +F +N P+ K+
Sbjct 88 DQKPDQIIVNEYKPGEGLKPHFDRKDY-YQNVIIGLSLGSGTIMEF---YKNKPIPEKKK 143
Query 168 LLLEHGDVVVWGGESRLFY-HGIQPLK---AGFHPLTIDCRYNLTFRQAGKKE 216
+ + + + ++R + HGI P K + + R ++TFR K++
Sbjct 144 IYIPPRSLYIIKDDARYIWKHGIPPRKYDEINGKKIPRETRISITFRNVIKEK 196
>sp|Q5UR02|CHLE_MIMIV Probable cholinesterase precursor (Acylcholine acylhydrolase)
Length=579
Done


NCBI Sequence Viewer v2.0 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?view=gp&query_key=6&db=protein&dopt=GenP
Google
NCBI
Entrez Protein
My NCBI
[Sign In] [Register]
PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books
Search Protein for Go Clear
Limits Preview/Index History Clipboard Details
Display GenPept Show 5 Send to
Range: from begin to end Features: CDD + Refresh
1: Q13686. Reports Alkylated DNA rep...[gi:12643239]
BLink, Conserved Domains, Links
Comment Features Sequence
LOCUS Q13686 389 aa linear PRI 10-JUL-2007
DEFINITION Alkylated DNA repair protein alkB homolog 1.
ACCESSION Q13686
VERSION Q13686.2 GI:12643239
DBSOURCE swissprot: locus ALKB1_HUMAN, accession Q13686.2;
class: standard.
extra accessions:Q8TAU1,Q9ULA7
created: Dec 1, 2000.
sequence updated: Dec 1, 2000.
annotation updated: Jul 10, 2007.
xrefs: X91992.1, CAA63047.1, AC008044.4, AAF01478.1, BC025787.1,
AAH25787.1, S64736
xrefs (non-sequence databases): UniGene:Hs.94542,
Ensembl:ENSG00000100601, KEGG:hsa:8846, H-InvDB:HIX0011855,
HGNC:17911, MIM: 605345, ArrayExpress:Q13686,
GermOnline:ENSG00000100601, RZPD-ProtExp:IOH12000,
RZPD-ProtExp:T4267, GO:0006307, InterPro:IPR004574,
TIGRFAMs:TIGR00568
KEYWORDS .
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Done

```
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 389)
  AUTHORS   Wei,Y.F., Carter,K.C., Wang,R.P. and Shell,B.K.
  TITLE     Molecular cloning and functional analysis of a human cDNA encoding
            an Escherichia coli AlkB homolog, a protein involved in DNA
            alkylation damage repair
  JOURNAL   Nucleic Acids Res. 24 (5), 931-937 (1996)
   PUBMED   8600462
  REMARK    NUCLEOTIDE SEQUENCE [MRNA].
            TISSUE=Synovial sarcoma
REFERENCE   2  (residues 1 to 389)
  AUTHORS   Heilig,R., Eckenberg,R., Petit,J.L., Fonknechten,N., Da Silva,C.,
            Cattolico,L., Levy,M., Barbe,V., de Berardinis,V., Ureta-Vidal,A.,
            Pelletier,E., Vico,V., Anthouard,V., Rowen,L., Madan,A., Qin,S.,
            Sun,H., Du,H., Pepin,K., Artiguenave,F., Robert,C., Cruaud,C.,
            Bruls,T., Jaillon,O., Friedlander,L., Samson,G., Brottier,P.,
            Cure,S., Segurens,B., Aniere,F., Samain,S., Crespeau,H., Abbasi,N.,
            Aiach,N., Boscus,D., Dickhoff,R., Dors,M., Dubois,I., Friedman,C.,
            Gouyvenoux,M., James,R., Madan,A., Mairey-Estrada,B., Mangenot,S.,
            Martins,N., Menard,M., Oztas,S., Ratcliffe,A., Shaffer,T.,
            Trask,B., Vacherie,B., Bellemere,C., Belser,C., Besnard-Gonnet,M.,
            Bartol-Mavel,D., Boutard,M., Briez-Silla,S., Combette,S.,
            Dufosse-Laurent,V., Ferron,C., Lechaplais,C., Louesse,C.,
            Muselet,D., Magdelenat,G., Pateau,E., Petit,E.,
            Sirvain-Trukniewicz,P., Trybou,A., Vega-Czarny,N., Bataille,E.,
            Bluet,E., Bordelais,I., Dubois,M., Dumont,C., Guerin,T.,
            Haffray,S., Hammadi,R., Muanga,J., Pellouin,V., Robert,D.,
            Wunderle,E., Gauguet,G., Roy,A., Sainte-Marthe,L., Verdier,J.,
            Verdier-Discala,C., Hillier,L., Fulton,L., McPherson,J.,
            Matsuda,F., Wilson,R., Scarpelli,C., Gyapay,G., Wincker,P.,
            Saurin,W., Quetier,F., Waterston,R., Hood,L. and Weissenbach,J.
  TITLE     The DNA sequence and analysis of human chromosome 14
  JOURNAL   Nature 421 (6923), 601-607 (2003)
   PUBMED   12508121
  REMARK    NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
```

```
  CONSRTM    MGC Project Team
  TITLE      The status, quality, and expansion of the NIH full-length cDNA
             project: the Mammalian Gene Collection (MGC)
  JOURNAL    Genome Res. 14 (10B), 2121-2127 (2004)
   PUBMED    15489334
  REMARK     NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
             TISSUE=Pancreas
             Erratum:[Genome Res. 2006 Jun;16(6):804. Morrin, Ryan [corrected to
             Morin, Ryan]]
COMMENT      On Jul 17, 2007 this sequence version replaced gi:2134723.
             [TISSUE SPECIFICITY] Ubiquitous.
             [SIMILARITY] Belongs to the alkB family.
             [SEQUENCE CAUTION] Sequence=CAA63047.1; Type=Frameshift;
             Positions=Several.
FEATURES             Location/Qualifiers
     source          1..389
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
     gene            1..389
                     /gene="ALKBH1"
                     /note="synonyms: ABH, ALKBH"
     Protein         1..389
                     /gene="ALKBH1"
                     /product="Alkylated DNA repair protein alkB homolog 1"
     Region          1..389
                     /gene="ALKBH1"
                     /region_name="Mature chain"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Alkylated DNA repair protein alkB homolog 1.
                     /FTId=PRO_0000066668."
     Region          133
                     /gene="ALKBH1"
                     /region_name="Conflict"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="K -> T (in Ref. 1)."
     Region          <160..296
                     /gene="ALKBH1"
                     /region_name="AlkB"
```

```
     Region          1..389
                     /gene="ALKBH1"
                     /region_name="Mature chain"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="Alkylated DNA repair protein alkB homolog 1.
                     /FTId=PRO_0000066668."
     Region          133
                     /gene="ALKBH1"
                     /region_name="Conflict"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="K -> T (in Ref. 1)."
     Region          <160..296
                     /gene="ALKBH1"
                     /region_name="AlkB"
                     /note="Alkylated DNA repair protein [DNA replication,
                     recombination, and repair]; COG3145"
                     /db_xref="CDD:32959"
     Region          266..267
                     /gene="ALKBH1"
                     /region_name="Conflict"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="TA -> PP (in Ref. 1)."
     Region          388
                     /gene="ALKBH1"
                     /region_name="Conflict"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="D -> H (in Ref. 3)."
ORIGIN
        1 mgkmaaavgs vatlatepge dafrklfrfy rqsrpgtadl egvidfsaah aargkgpgaq
       61 kviksqlnvs svseqnayra glqpvskwqa yglkgypgfi fipnpflpgy qwhwvkqclk
      121 lysqkpnvcn ldkhmskeet qdlweqskef lrykeatkrr prslleklrw vtvgyhynwd
      181 skkysadhyt pfpsdlgfls eqvaaacgfe dfraeagiln yyrldstlgi hvdrseldhs
      241 kpllsfsfgq saifllgglq rdeaptamfm hsgdimimsg fsrllnhavp rvlpnpegeg
      301 lphcleaplp avlprdsmve pcsmedwqvc asylktarvn mtvrqvlatd qnfplepied
      361 ekrdistegf chlddqnsev krarinpds
//
```

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

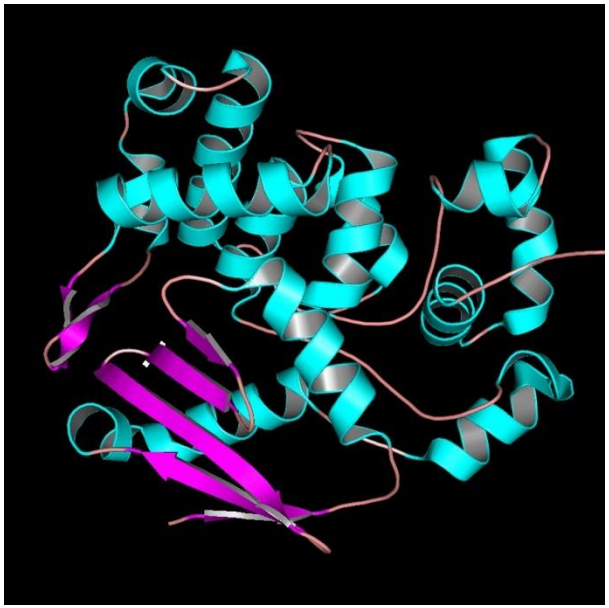
```
E.c. AlkA 127 SVAMAAKLTARVAQLYGERLDDFPE--YICFPTPQRLAAADPQA-LKALGMPLKRAEALI 183
               ++|    +  |+ | +| ||    +  |  ||+ | ||  + +| |+ ||+   ||  +
H.s. OGG1 151 NIARITGMVERLCQAFGPRLIQLDDVTYHGFPSLQALAGPEVEAHLRKLGLGY-RARYVS 209

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

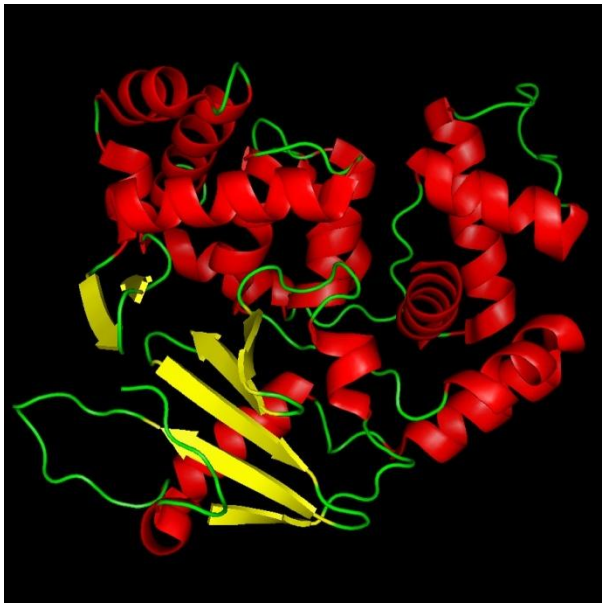
BLOSUM62

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
```

Significance of alignments

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

- Statistical methods compare the score of a match with the distribution of alignment scores found using 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 indicators of significance (less accurate):
  - Percentage of identical residues
  - Percentage of similar residues
  - Raw alignment score

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. …CCGCCGCCGCCGCCGCCG…
  - low complexity regions,
    - Protein example: PPCDPPPPPKDKKKKDDGPP
    - DNA example: AAATAAAAAAAATAAAAAAT

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

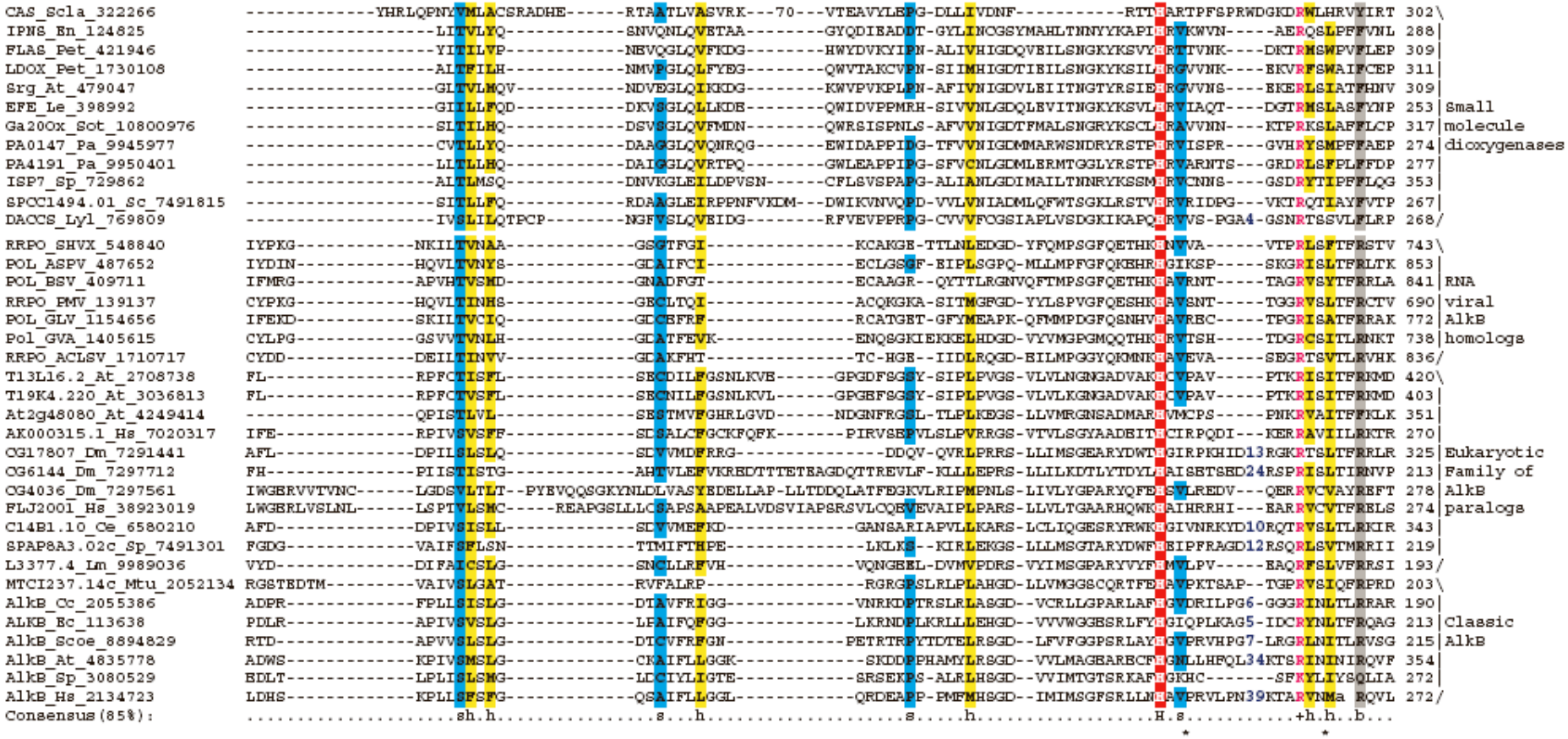
Finding families of proteins

Overview

- Searching with a family of proteins
  - Back to the example
  - PSI-BLAST example
  - Sequence patterns
  - Sequence profiles and logos
  - Iterated searches and PSI-BLAST
  - Advanced PSI-BLASTing

Back to the example...

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

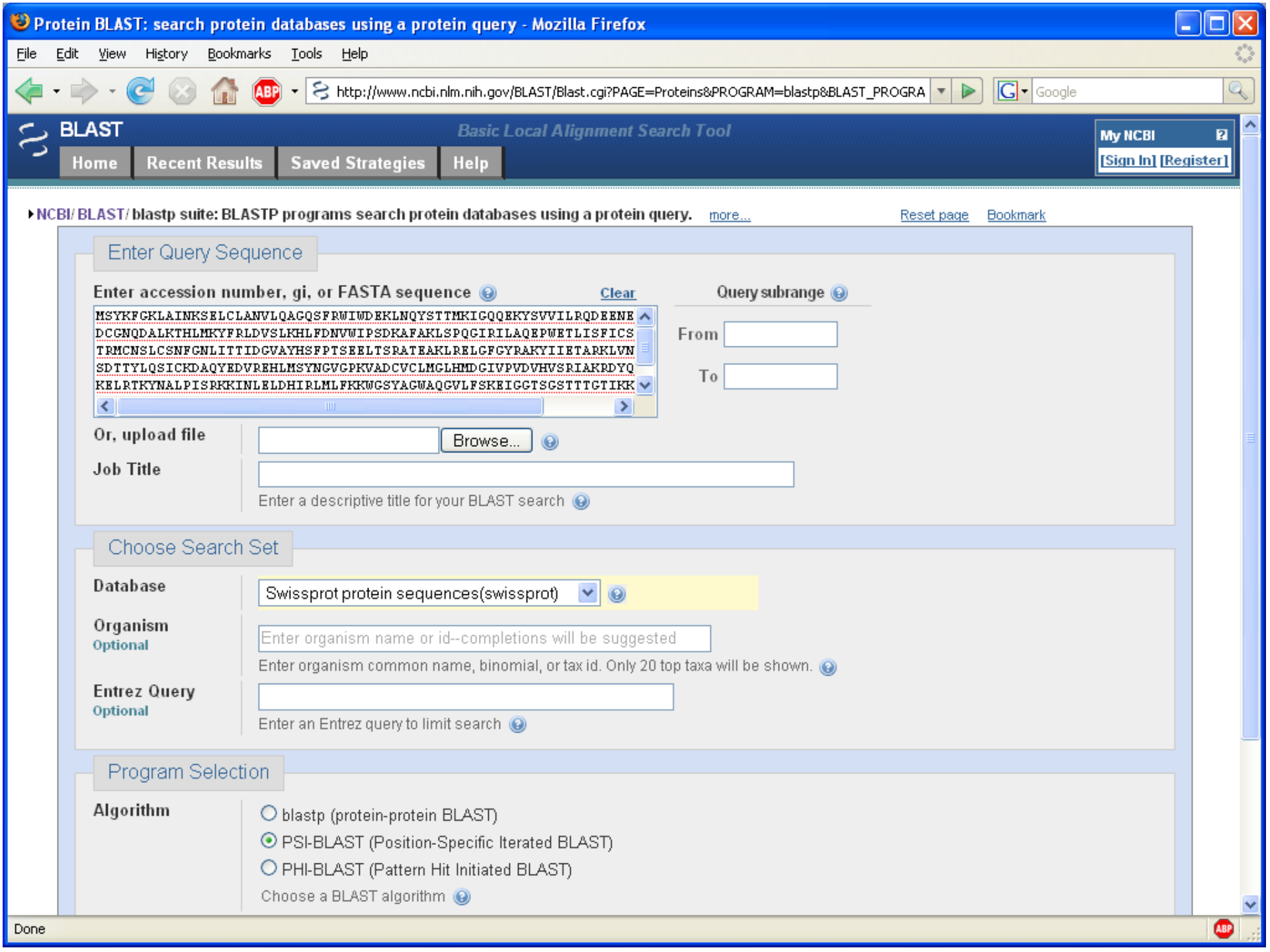


Excerpt from the AlkB paper

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


Protein BLAST: search protein databases using a protein query - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRA
Google
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query.
more...
Reset page
Bookmark
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Clear
MSYKFGKLAINKSELCLANVLQAGQSFRWIWDEKLNQYSTTMKIGQQEKYSVVILRQDEENE
DCGNQDALKTHLMKYFRLDVSLKHLFDNVWIPSDKAFAKLSPQGIRILAQEPWETLISFICS
TRMCNSLCSNFGNLITTIDGVAYHSFPTSEELTSPATEAKLRELGFGYRAKYIIETARKLVN
SDTTYLQSICKDAQYEDVREHLMSYNGVGPKVADCVCLMGLHMDGIVPVDVHVSRIAKRDYQ
KELRTKYNALPISRKKINLELDHIRLMLFKKWGSYAGWAQGVLFSKEIGGTSGSTTTGTIKK
Query subrange
From
To
Or, upload file
Browse...
Job Title
Enter a descriptive title for your BLAST search
Choose Search Set
Database
Swissprot protein sequences(swissprot)
Organism
Optional
Enter organism name or id--completions will be suggested
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
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)
Choose a BLAST algorithm
Done

Sequences with E-value BETTER than threshold

Sequences producing significant alignments:		Score (Bits)	E Value	
NEW ☑	sp|P53397|OGG1_YEAST N-glycosylase/DNA lyase [Includes: 8-oxo...	741	0.0	G
NEW ☑	sp|O70249|OGG1_RAT N-glycosylase/DNA lyase [Includes: 8-oxogu...	194	4e-49	G
NEW ☑	sp|O08760|OGG1_MOUSE N-glycosylase/DNA lyase [Includes: 8-oxo...	194	4e-49	G
NEW ☑	sp|O15527|OGG1_HUMAN N-glycosylase/DNA lyase [Includes: 8-oxo...	192	2e-48	G
NEW ☑	sp|Q9V3I8|OGG1_DROME N-glycosylase/DNA lyase (dOggl) [Include...	153	1e-36	G
NEW ☑	sp|O27397|OGG1_METTH Probable N-glycosylase/DNA lyase [Includ...	90.5	1e-17	G
NEW ☑	sp|Q9SJQ6|ROS1_ARATH Protein ROS1 (Repressor of silencing 1) ...	43.9	0.001	G
NEW ☑	sp|O49498|DML3_ARATH DEMETER-like protein 3	42.4	0.004	G

Run PSI-Blast iteration 2

Sequences with E-value WORSE than threshold

☐	sp|Q9SR66|DML2_ARATH DEMETER-like protein 2	40.8	0.009	
☐	sp|Q4UK93|END3_RICFE Endonuclease III (DNA-(apurinic or apyrimid	35.8	0.30	
☐	sp|Q8LK56|DME_ARATH Transcriptional activator DEMETER (DNA gl...	35.4	0.36	G
☐	sp|P18479|POLG_ZYMVC Genome polyprotein [Contains: P1 protein...	35.4	0.36	
☐	sp|P75506.1|Y269_MYCPN UPF0144 protein MG130 homolog	33.5	1.3	
☐	sp|Q10630|ADA_MYCTU Putative regulatory protein ada (Regulato...	33.5	1.6	
☐	sp|Q58030|Y613_METJA Putative endonuclease MJ0613	33.1	2.0	
☐	sp|Q63072|BST1_RAT ADP-ribosyl cyclase 2 precursor (Cyclic AD...	32.7	2.9	G
☐	sp|Q64277|BST1_MOUSE ADP-ribosyl cyclase 2 precursor (Cyclic ...	32.3	3.0	G
☐	sp|Q40002|VATA_HORVU Vacuolar ATP synthase catalytic subunit ...	32.0	4.0	
☐	sp|Q89330|POLG_ZYMVR Genome polyprotein [Contains: P1 protein...	32.0	4.8	
☐	sp|Q84934|POLG_PPVSK Genome polyprotein [Contains: P1 protein...	31.6	5.2	
☐	sp|P17767|POLG_PPVRA Genome polyprotein [Contains: P1 protein...	31.2	7.4	
☐	sp|P13529|POLG_PPVD Genome polyprotein [Contains: P1 proteina...	31.2	7.4	

Run PSI-Blast iteration 2

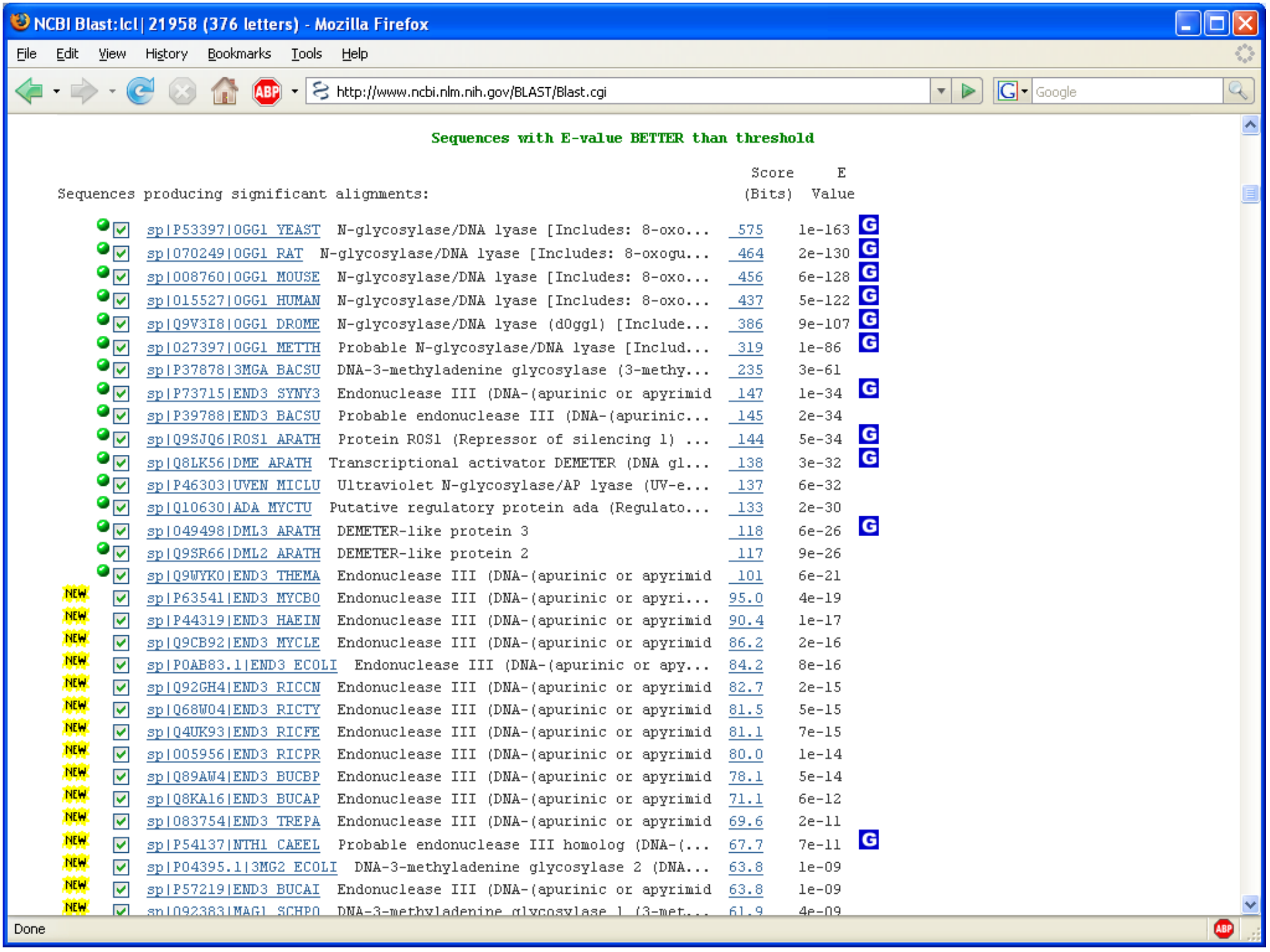
Sequences with E-value BETTER than threshold

Sequences producing significant alignments:	Score (Bits)	E Value	
sp|P53397|OGG1_YEAST N-glycosylase/DNA lyase [Includes: 8-oxo...	617	2e-176	G
sp|O08760|OGG1_MOUSE N-glycosylase/DNA lyase [Includes: 8-oxo...	462	1e-129	G
sp|O70249|OGG1_RAT N-glycosylase/DNA lyase [Includes: 8-oxogu...	459	7e-129	G
sp|O15527|OGG1_HUMAN N-glycosylase/DNA lyase [Includes: 8-oxo...	450	4e-126	G
sp|Q9V3I8|OGG1_DROME N-glycosylase/DNA lyase (dOggl) [Include...	417	5e-116	G
sp|O27397|OGG1_METTH Probable N-glycosylase/DNA lyase [Includ...	247	6e-65	G
sp|Q9SJQ6|ROS1_ARATH Protein ROS1 (Repressor of silencing 1) ...	114	7e-25	G
sp|O49498|DML3_ARATH DEMETER-like protein 3	104	6e-22	G
NEW sp|Q9SR66|DML2_ARATH DEMETER-like protein 2	88.5	4e-17	
NEW sp|Q8LK56|DME_ARATH Transcriptional activator DEMETER (DNA gl...	85.8	3e-16	G
NEW sp|P37878|3MGA_BACSU DNA-3-methyladenine glycosylase (3-methy...	52.7	2e-06	
NEW sp|P73715|END3_SYNY3 Endonuclease III (DNA-(apurinic or apyrimid	46.9	2e-04	G
NEW sp|Q9WYK0|END3_THEMA Endonuclease III (DNA-(apurinic or apyrimid	44.9	5e-04	
NEW sp|Q10630|ADA_MYCTU Putative regulatory protein ada (Regulato...	44.9	5e-04	
NEW sp|P39788|END3_BACSU Probable endonuclease III (DNA-(apurinic...	43.4	0.002	
NEW sp|P46303|UVEN_MICLU Ultraviolet N-glycosylase/AP lyase (UV-e...	42.3	0.003	

Run PSI-Blast iteration 3

Sequences with E-value WORSE than threshold

sp|Q92383|MAG1_SCHPO DNA-3-methyladenine glycosylase 1 (3-met...	41.1	0.007	
sp|Q58030|Y613_METJA Putative endonuclease MJ0613	41.1	0.008	
sp|P54137|NTH1_CAEEL Probable endonuclease III homolog (DNA-(...	39.6	0.020	G
sp|P44319|END3_HAEIN Endonuclease III (DNA-(apurinic or apyrimid	39.6	0.023	
sp|Q8KA16|END3_BUCAP Endonuclease III (DNA-(apurinic or apyrimid	39.6	0.024	
sp|Q68W04|END3_RICTY Endonuclease III (DNA-(apurinic or apyrimid	38.8	0.040	
sp|Q4UK93|END3_RICFE Endonuclease III (DNA-(apurinic or apyrimid	38.8	0.041	
sp|O83754|END3_TREPA Endonuclease III (DNA-(apurinic or apyrimid	38.0	0.070	
sp|Q92GH4|END3_RICCN Endonuclease III (DNA-(apurinic or apyrimid	38.0	0.073	
sp|O35980|NTHL1_MOUSE Endonuclease III-like protein 1	37.6	0.093	G
sp|P78549|NTHL1_HUMAN Endonuclease III-like protein 1	37.2	0.11	G


NCBI Blast:lcl|21958 (376 letters) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Google
Sequences with E-value BETTER than threshold
Sequences producing significant alignments:
Score (Bits)
E Value
sp|P53397|OGG1_YEAST N-glycosylase/DNA lyase [Includes: 8-oxo... 575 1e-163
sp|O70249|OGG1_RAT N-glycosylase/DNA lyase [Includes: 8-oxogu... 464 2e-130
sp|O08760|OGG1_MOUSE N-glycosylase/DNA lyase [Includes: 8-oxo... 456 6e-128
sp|O15527|OGG1_HUMAN N-glycosylase/DNA lyase [Includes: 8-oxo... 437 5e-122
sp|Q9V3I8|OGG1_DROME N-glycosylase/DNA lyase (dOggl) [Include... 386 9e-107
sp|O27397|OGG1_METTH Probable N-glycosylase/DNA lyase [Includ... 319 1e-86
sp|P37878|3MGA_BACSU DNA-3-methyladenine glycosylase (3-methy... 235 3e-61
sp|P73715|END3_SYNY3 Endonuclease III (DNA-(apurinic or apyrimid 147 1e-34
sp|P39788|END3_BACSU Probable endonuclease III (DNA-(apurinic... 145 2e-34
sp|Q9SJQ6|ROS1_ARATH Protein ROS1 (Repressor of silencing 1) ... 144 5e-34
sp|Q8LK56|DME_ARATH Transcriptional activator DEMETER (DNA gl... 138 3e-32
sp|P46303|UVEN_MICLU Ultraviolet N-glycosylase/AP lyase (UV-e... 137 6e-32
sp|Q10630|ADA_MYCTU Putative regulatory protein ada (Regulato... 133 2e-30
sp|O49498|DML3_ARATH DEMETER-like protein 3 118 6e-26
sp|Q9SR66|DML2_ARATH DEMETER-like protein 2 117 9e-26
sp|Q9WYK0|END3_THEMA Endonuclease III (DNA-(apurinic or apyrimid 101 6e-21
NEW sp|P63541|END3_MYCBO Endonuclease III (DNA-(apurinic or apyri... 95.0 4e-19
NEW sp|P44319|END3_HAEIN Endonuclease III (DNA-(apurinic or apyrimid 90.4 1e-17
NEW sp|Q9CB92|END3_MYCLE Endonuclease III (DNA-(apurinic or apyrimid 86.2 2e-16
NEW sp|P0AB83.1|END3_ECOLI Endonuclease III (DNA-(apurinic or apy... 84.2 8e-16
NEW sp|Q92GH4|END3_RICCN Endonuclease III (DNA-(apurinic or apyrimid 82.7 2e-15
NEW sp|Q68W04|END3_RICTY Endonuclease III (DNA-(apurinic or apyrimid 81.5 5e-15
NEW sp|Q4UK93|END3_RICFE Endonuclease III (DNA-(apurinic or apyrimid 81.1 7e-15
NEW sp|O05956|END3_RICPR Endonuclease III (DNA-(apurinic or apyrimid 80.0 1e-14
NEW sp|Q89AW4|END3_BUCBP Endonuclease III (DNA-(apurinic or apyrimid 78.1 5e-14
NEW sp|Q8KA16|END3_BUCAP Endonuclease III (DNA-(apurinic or apyrimid 71.1 6e-12
NEW sp|O83754|END3_TREPA Endonuclease III (DNA-(apurinic or apyrimid 69.6 2e-11
NEW sp|P54137|NTH1_CAEEL Probable endonuclease III homolog (DNA-(... 67.7 7e-11
NEW sp|P04395.1|3MG2_ECOLI DNA-3-methyladenine glycosylase 2 (DNA... 63.8 1e-09
NEW sp|P57219|END3_BUCAI Endonuclease III (DNA-(apurinic or apyrimid 63.8 1e-09
NEW sp|O92383|MAG1_SCHPO DNA-3-methyladenine glycosylase 1 (3-met... 61.9 4e-09
Done

NEW	☑	sp|P57219|END3_BUCAI	Endonuclease III (DNA-(apurinic or apyrimid	63.8	1e-09
NEW	☑	sp|Q92383|MAG1_SCHPO	DNA-3-methyladenine glycosylase 1 (3-met...	61.9	4e-09
NEW	☑	sp|Q58030|Y613_METJA	Putative endonuclease MJ0613	61.5	6e-09
NEW	☑	sp|P78549|NTHL1_HUMAN	Endonuclease III-like protein 1	59.2	3e-08 G
NEW	☑	sp|Q2KID2|NTHL1_BOVIN	Endonuclease III-like protein 1	58.4	5e-08 G
NEW	☑	sp|P17802.1|MUTY_ECOLI	A/G-specific adenine glycosylase	56.1	2e-07
NEW	☑	sp|O94468|MAG2_SCHPO	Probable DNA-3-methyladenine glycosylase...	53.8	1e-06
NEW	☑	sp|Q05869|MUTY_SALTY	A/G-specific adenine glycosylase	53.4	1e-06
NEW	☑	sp|Q58829|Y1434_METJA	Putative endonuclease MJ1434	52.3	3e-06
NEW	☑	sp|O35980|NTHL1_MOUSE	Endonuclease III-like protein 1	52.3	4e-06 G
NEW	☑	sp|P31378|NTG1_YEAST	DNA base excision repair N-glycosylase 1...	51.5	5e-06 G
NEW	☑	sp|Q8K926|MUTY_BUCAP	A/G-specific adenine glycosylase	51.5	6e-06
NEW	☑	sp|Q10159|MYH1_SCHPO	A/G-specific adenine DNA glycosylase	51.1	7e-06
NEW	☑	sp|Q09907|END3_SCHPO	Endonuclease III homolog (DNA-(apurinic ...	51.1	8e-06
NEW	☑	sp|P57617|MUTY_BUCAI	A/G-specific adenine glycosylase	50.7	1e-05
NEW	☑	sp|P29588|GTMR_METTF	G/T mismatches repair enzyme (Thymine-DN...	49.2	3e-05 G
NEW	☑	sp|Q9UIF7|MUTYH_HUMAN	A/G-specific adenine DNA glycosylase (MutY	46.5	2e-04 G
NEW	☑	sp|Q08214|NTG2_YEAST	DNA base excision repair N-glycosylase 2	45.3	5e-04 G
NEW	☑	sp|Q8R5G2|MUTYH_RAT	A/G-specific adenine DNA glycosylase (MutY h	44.5	7e-04 G
NEW	☑	sp|Q89A45|MUTY_BUCBP	A/G-specific adenine glycosylase	44.5	7e-04
NEW	☑	sp|P44320|MUTY_HAEIN	A/G-specific adenine glycosylase	43.4	0.002
NEW	☑	sp|Q9XAI4|RECR_STRCO	Recombination protein recR	41.9	0.005

Run PSI-Blast iteration 4

Sequences with E-value WORSE than threshold

	☐	sp|Q99P21|MUTYH_MOUSE	A/G-specific adenine DNA glycosylase (MutY	41.1	0.008 G
	☐	sp|P22134|MAG_YEAST	DNA-3-methyladenine glycosylase (3-methyl...	40.7	0.009 G
	☐	sp|Q65UP1|RUVA_MANSM	Holliday junction ATP-dependent DNA helicas	39.5	0.022
	☐	sp|P21269|CCA1_YEAST	tRNA nucleotidyltransferase, mitochondri...	39.5	0.023 G
	☐	sp|Q82EQ7|RECR_STRAW	Recombination protein recR	38.4	0.054
	☐	sp|O29876|OGG1_ARCFU	Probable N-glycosylase/DNA lyase [Includ...	36.8	0.13
	☐	sp|P74925|MUTL_THEMA	DNA mismatch repair protein mutL	35.3	0.42
	☐	sp|Q9Z512|UVRC_STRCO	UvrABC system protein C (Protein uvrC) (...	34.9	0.52
	☐	sp|Q49Y80|RUVA_STAS1	Holliday junction ATP-dependent DNA helicas	34.1	0.83 G

Searching with a family of proteins

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.

Pattern example

Entry for alkylbase DNA glycosylases (AlkA) in PROSITE:

G-I-G-x-W-[ST]-[AV]-x-[LIVMFY](2)-x-[LIVM]-x(8)-[MF]-x(2)- [ED]-D

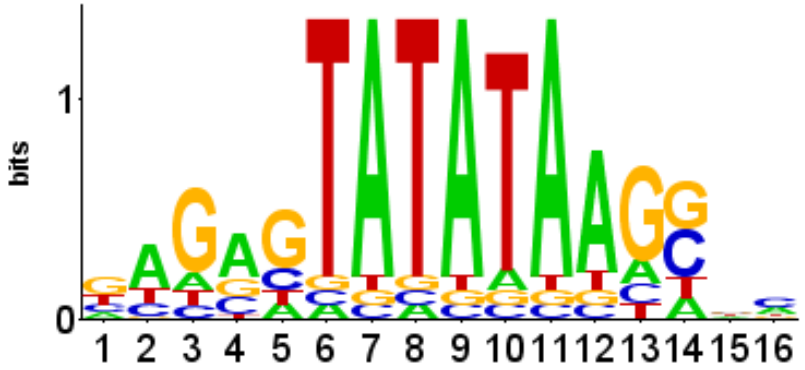
Syntax:
x denotes any amino acid
[ST] denotes alternative amino acids, e.g. serine (S) or threonine (T)
(2) denotes that the previous part should occur 2 times

- Derived from a multiple alignment of a family of proteins
- Generated automatically (e.g. by PRATT) or manually
- Collected in PROSITE database (www.expasy.org/prosite)

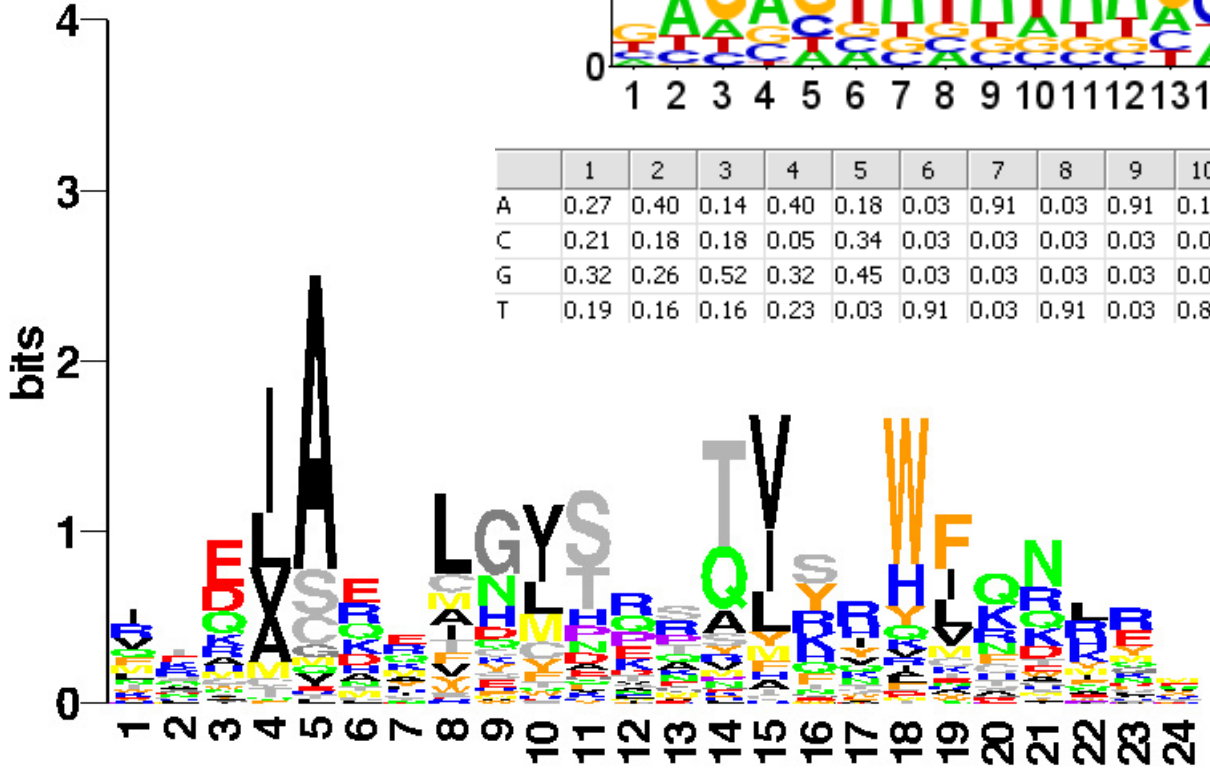
Sequence profiles (PSSMs)

- 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



Calculating PSSMs

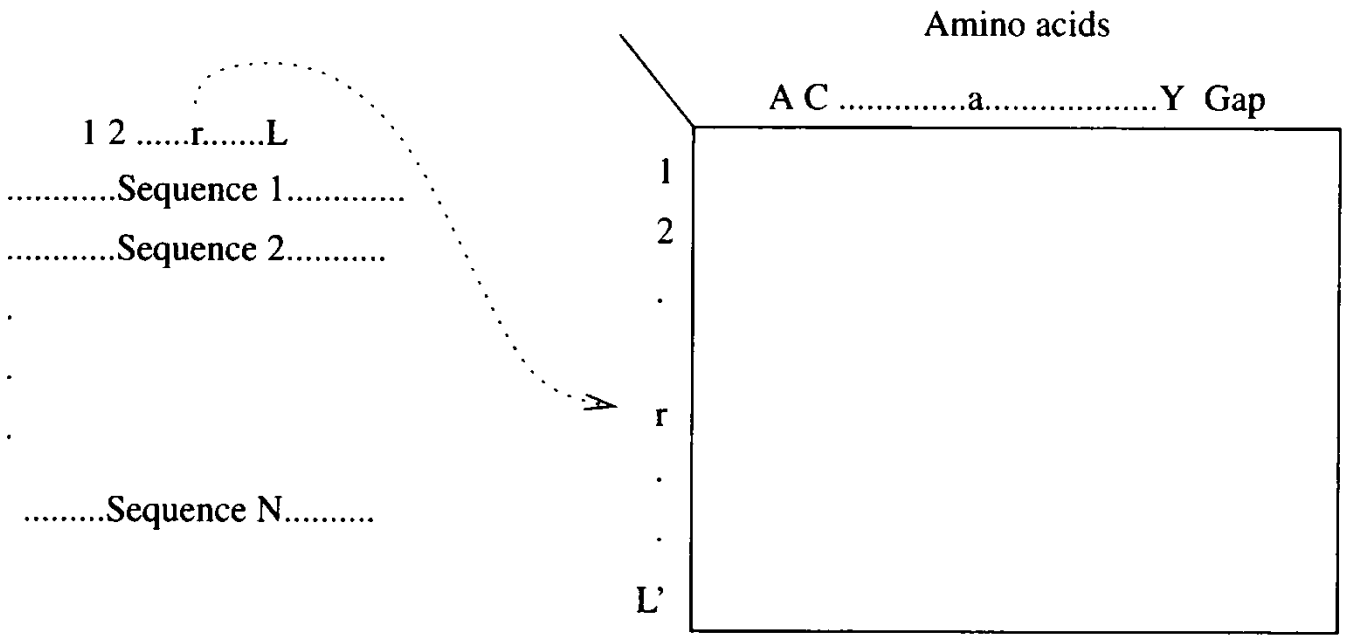


Figure 6.1 The figure shows the connection between a block of a multiple alignment and a derived profile. Positions $1 \ldots r \ldots L$ are positions in the profile. Gap specifies gap penalties for the gap in profile positions. (L' can be different from L if some of the columns are left out.)

Profile example

Position-specific scoring matrix (PSSM) derived from a family of ATP binding RNA helicases (”DEAD" box family)

Protein	Sequence
rhle_ecoli	EILVL**DEAD**RMLDMGFIHDI
dbp2_schpo	TYLVL**DEAD**RMLDMGFEPQI
dbp2_yeast	TYLVL**DEAD**RMLDMGFEPQI
dbpa_ecoli	NTLVM**DEAD**RMLDMGFSDAI
rm62_drome	TYLVL**DEAD**RMLDMGFEPQI
p68_human	TYLVL**DEAD**RMLDMGFEPQI
rhlb_ecoli	QVVVL**DEAD**RMYDLGFIKDI
yn21_caeel	KFLIM**DEAD**RILNMDFEVEL
yhm5_yeast	KFLVM**DEAD**RLLDMEFGPVL
me31_drome	RILVL**DEAD**KLLSLDFQGML
drs1_yeast	EILVM**DEAD**RMLEEGFQDEL
if4a_rabit	KMFVL**DEAD**EMLSRGFKDQI
if41_human	KMFVL**DEAD**EMLSRGFKDQI
yk04_yeast	RYIVL**DEGD**KLMELGFDETI
...	

Pos	Cons	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Gap
1	K	4	-11	16	17	-21	6	9	-7	27	-12	0	16	7	16	20	12	8	-7	0	-18	100
2	F	-3	3	-14	-10	29	-10	-3	20	-9	26	19	-7	-12	-13	-10	-6	5	14	10	20	100
3	L	-5	-19	-21	-14	42	-16	-8	34	-12	51	40	-15	-12	-10	-15	-12	-3	32	16	17	100
4	V	10	11	-11	-11	15	8	-17	66	-11	46	34	-17	4	-12	-17	-6	11	83	-43	-4	100
5	L	-4	-37	-24	-14	53	-23	-11	39	-10	73	67	-19	-14	-4	-14	-19	-4	39	17	11	100
6	D	30	-50	150	100	-100	70	40	-20	30	-50	-40	70	10	70	0	20	20	-20	-110	-50	100
7	E	30	-60	100	150	-70	50	40	-20	30	-30	-20	50	10	70	0	20	20	-20	-110	-50	100
8	A	121	24	25	25	-41	58	-8	0	0	-9	0	17	41	16	-24	33	33	16	-66	-25	100
9	D	30	-50	150	100	-100	70	40	-20	30	-50	-40	70	10	70	0	20	20	-20	-110	-50	100
10	R	-7	-16	7	10	-23	-7	17	-9	36	-14	7	8	10	18	49	5	0	-10	35	-25	100
11	M	-1	-28	-19	-10	33	-16	-12	31	1	60	62	-15	-10	-2	-1	-14	-1	31	-2	2	100
12	L	-6	-35	-26	-16	61	-26	-10	41	-15	74	64	-20	-17	-7	-20	-20	-5	39	24	18	100
13	D	12	-9	41	32	-30	23	11	-7	12	-17	-12	23	7	20	3	19	10	-7	-27	-19	100
14	M	5	-15	-3	1	8	-3	-3	13	6	25	30	-3	0	5	6	-1	2	13	-1	-6	100
15	G	26	4	31	23	-25	54	-5	-8	-1	-16	-9	17	11	10	-10	23	17	8	-41	-27	100
16	F	-31	-6	-63	-44	96	-36	-4	44	-45	77	31	-31	-45	-50	-32	-19	-19	13	82	90	100
17	E	7	-7	10	13	-7	5	4	7	10	3	5	6	2	10	3	6	5	5	-9	-7	100
18	D	12	-12	31	27	-27	19	12	-6	12	-12	-8	17	12	21	4	8	9	-3	-30	-19	100
19	D	8	-14	25	22	-18	11	13	0	9	-3	-1	13	5	24	5	3	8	0	-24	-12	100
20	I	0	-5	-8	-7	27	-10	-9	43	-7	36	29	-10	-8	-7	-11	-6	6	35	-9	5	100

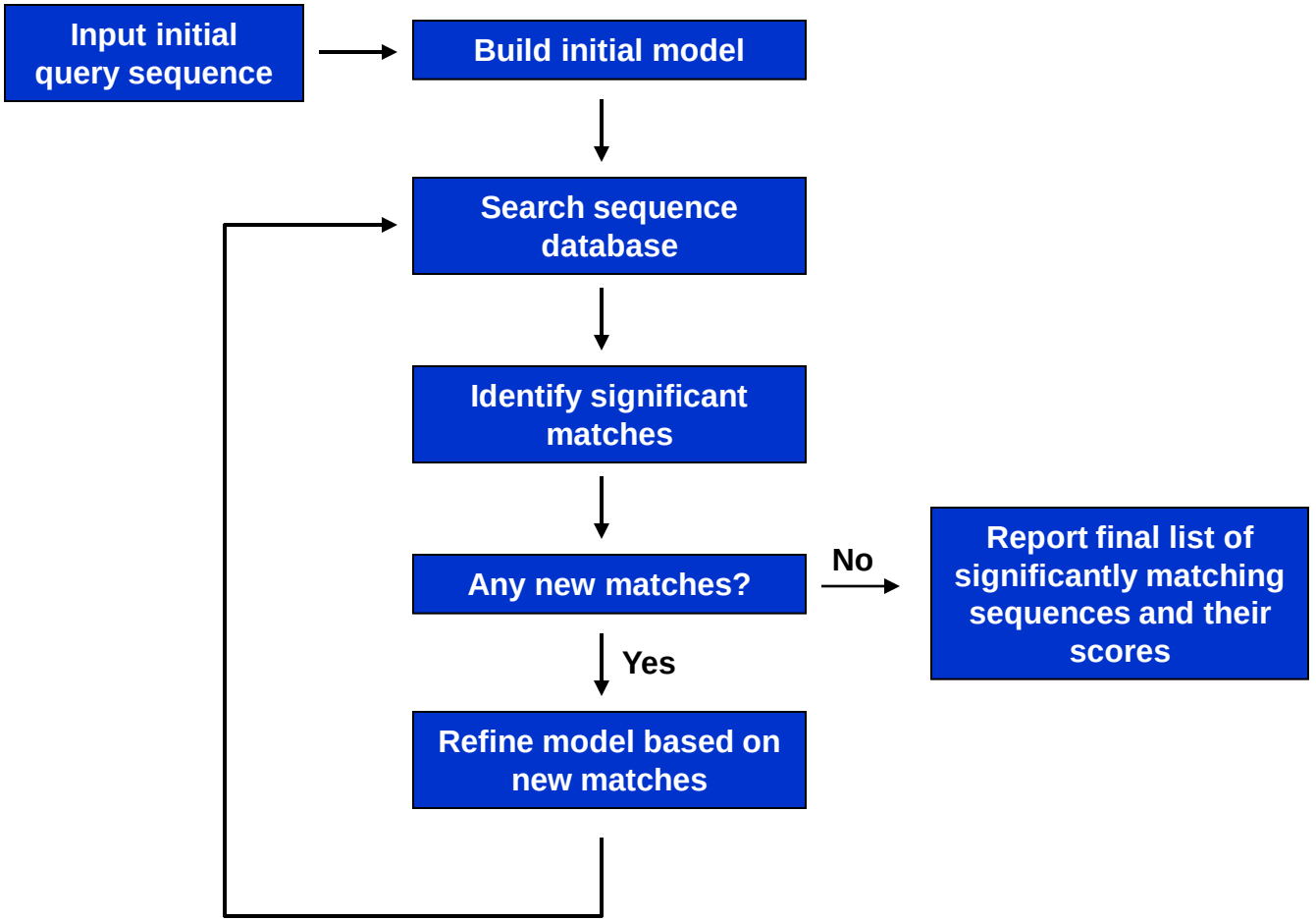
Source: Gribskov and Veretnik (1996) *Meth.Enz.* **266**, 198-212

Amino acid substitution score matrix

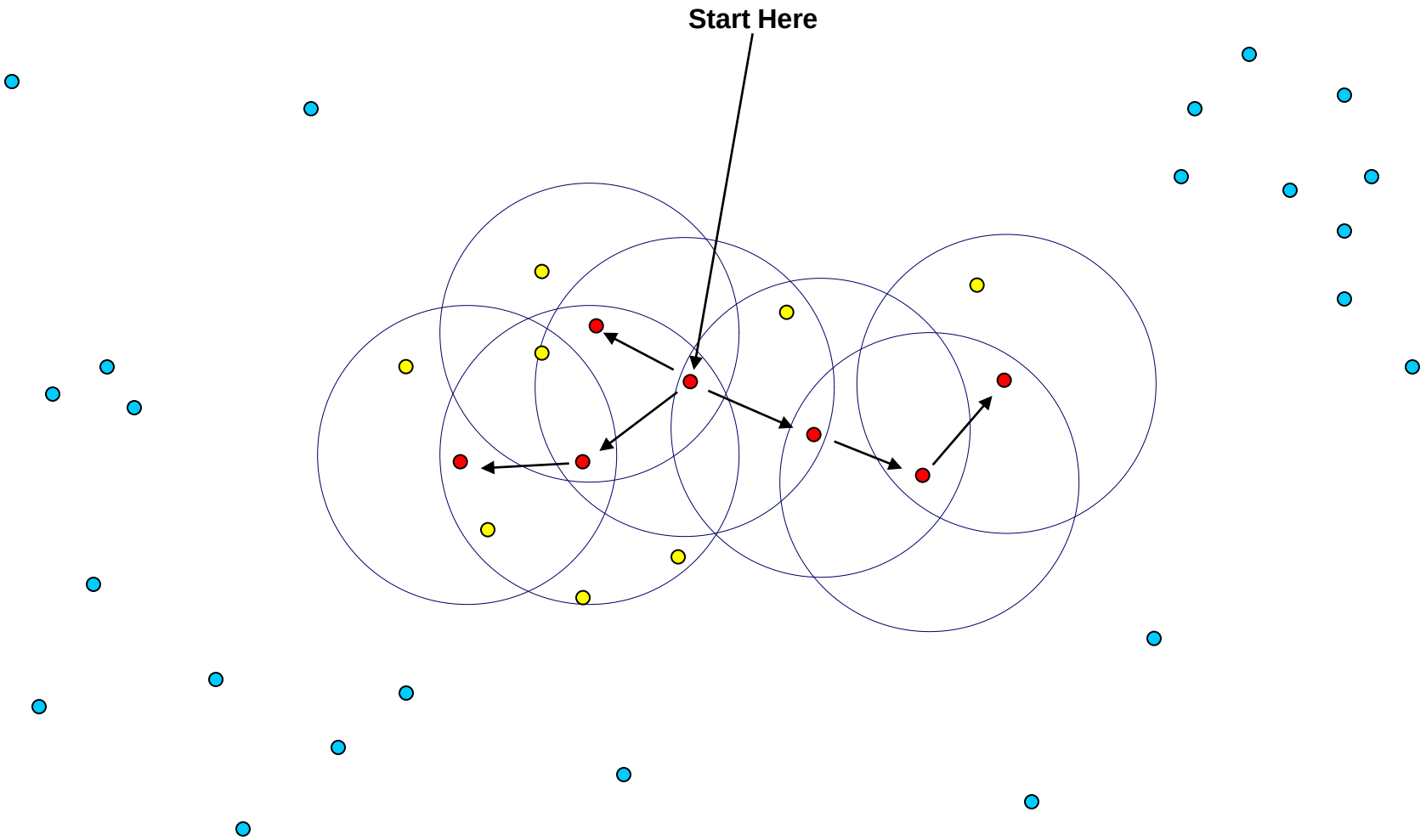
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2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

Iterated searches

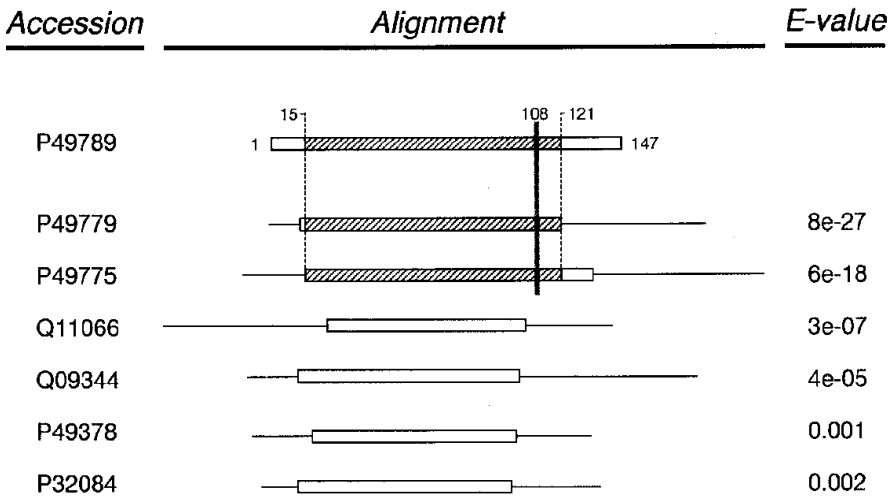


Iterated searches in protein space



PSI-BLAST: Calculating the profile

- Search database and collect clearly significant matches (e.g. E< 0.002)
- Create multiple sequence alignment (MSA) after each iteration
- Simplistic MSA based on pairwise alignment to the original query sequence, ignore gaps
- Only one copy of very similar sequences (98% identical) are used, other sequences are weighted



Source: Altschul *et al.* (1997) *Nucleic Acids Res.* **25,** 3389-3402

Advanced PSI-Blasting

PSI-BLAST search with 5 rounds and inclusion threshold 0.002:

```
blastpgp -d nr –i query.fsa –JT –sT –j 5 –h 0.002
```

Same PSI-BLAST search, but save profile:

```
blastpgp -d nr –i query.fsa –JT –sT –j 5 –h 0.002 –C pssm
```

Run search with a saved profile:

```
blastpgp -d nr –i query.fsa –JT –sT –R pssm
```

Run search in translated nucleotide database (tblastn-like) with saved protein profile:

```
blastall –p psitblastn -d nt –i query.fsa –JT –FF –R pssm
```

Recommended options for blastpgp: -JT –sT

Recommended options for blastall: -JT -FF

Using your own alignment

Run PSI-BLAST search based on own multiple alignment:

```
blastpgp -d nr –i query.fsa –JT –sT –j 5 –h 0.002 –B alignment
```

The alignment file:

```
mdap12itam  AETESPYQELQG-QRPEVYSDLNTQ
rdap12itam  AETESPYQELQG-QRPEVYSDLNTQ
hdap12itam  TETESPYQELQG-QRSDVYSDLNTQ
ssdap12itam AETESAYQELQG-QRSDVYSDLNTQ
btdap12itam PETESPYQELQG-QRTDVYSDLNTQ
ggnfam1itam PPPSPVYDCLDS-QQVEVYSVLKNN
ccnilt1itam VDPDQIYTELNASRQSDVYQSLRTD
btzetaitam2 NPNEVVYNELRKDKMAEAYSEIGMK
rzetaitam2  NPHEVVYNELRKDKMAEAYSEIGMK
mzetaitam2  NPQEGVYNALQKDKMAEAYSEIGTK
ggzetaitam2 NPHDTVYSSLQKDKMGEAYSEIGKK
mgraspitam  APGEELYAALEDYHPAELYRALAVS
```

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

Recommended reading

- *Getting the most from PSI-BLAST*
  David T Jones and Mark B Swindells (2002)
  Trends in Biochemical Sciences, 27, 161-164.
  http://dx.doi.org/10.1016/S0968-0004(01)02039-4


Getting the most from PSI-BLAST

Multiple sequence alignment

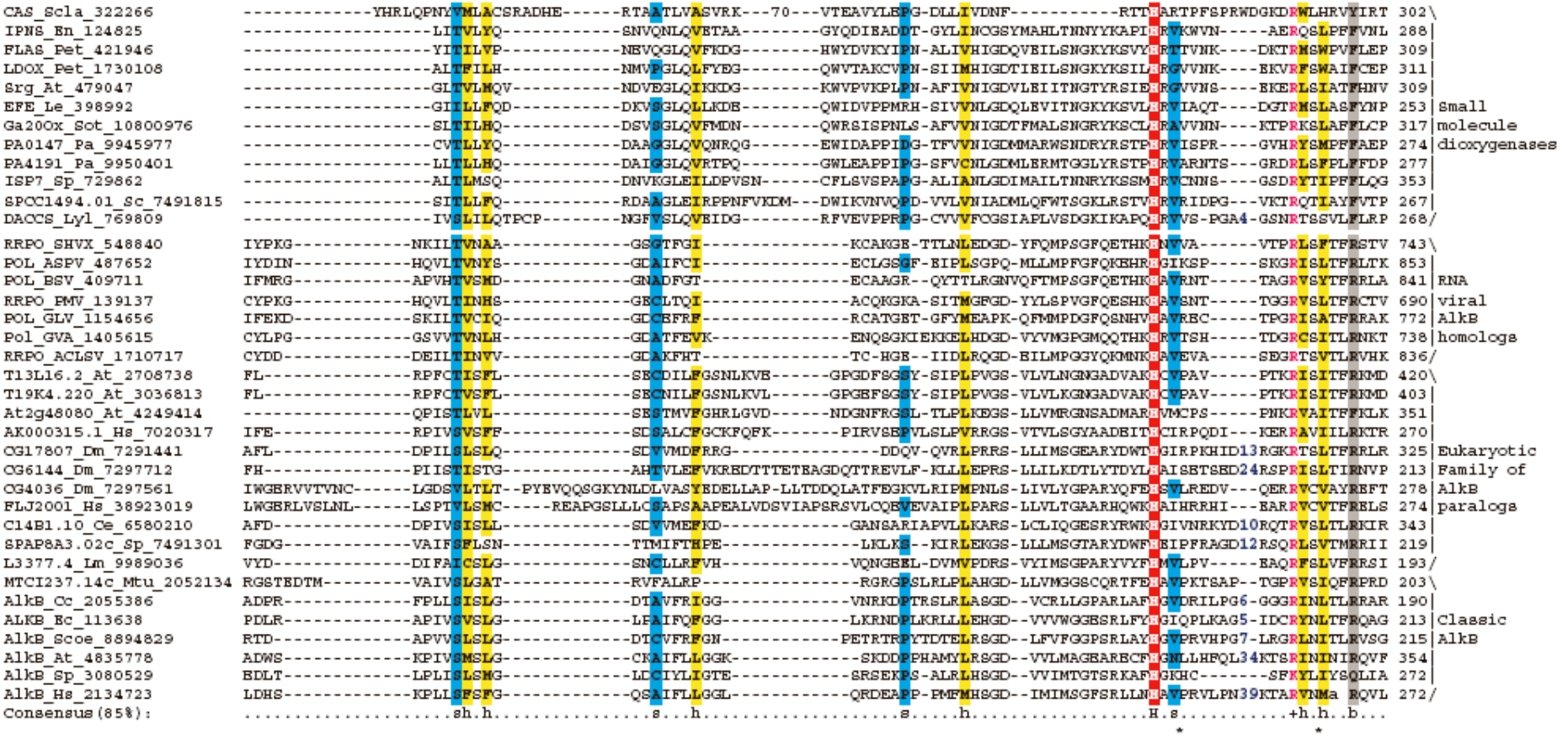
Overview

- What is multiple sequence alignment?
- Examples
- An objective function – sum of pairs score
- Exhaustive algorithms
- Heuristic algorithms
  - Progressive alignment
    - CLUSTAL W
  - Iterative alignment
    - MUSCLE
  - MAFFT
- Aligning protein-coding DNA sequences

What is a multiple alignment?

- 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

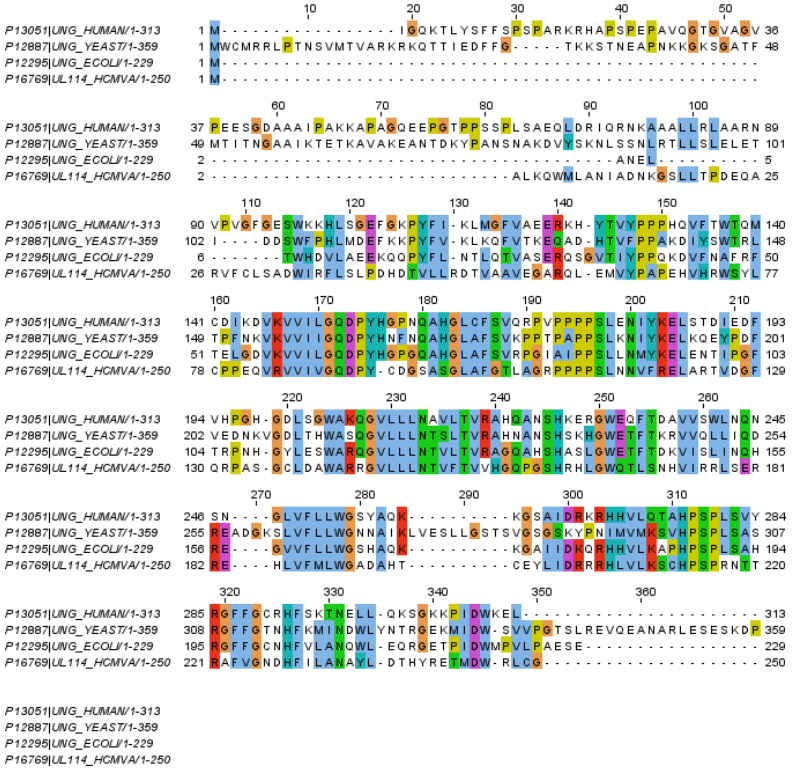

CAS_Scla_322266
IPNS_En_124825
FLAS_Pet_421946
LDOX_Pet_1730108
Srg_At_479047
EFE_Le_398992
Ga20Ox_Sot_10800976
PA0147_Pa_9945977
PA4191_Pa_9950401
ISP7_Sp_729862
SPCC1494.01_Sc_7491815
DACCS_Lyl_769809
RRPO_SHVX_548840
POL_ASPV_487652
POL_BSV_409711
RRPO_PMV_139137
POL_GLV_1154656
Pol_GVA_1405615
RRPO_ACLSV_1710717
T13L16.2_At_2708738
T19K4.220_At_3036813
At2g48080_At_4249414
AK000315.1_Hs_7020317
CG17807_Dm_7291441
CG6144_Dm_7297712
CG4036_Dm_7297561
FLJ20011_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
AlkB homologa
Eukaryotic Family of AlkB paralogs
Classic AlkB

A lot of software...

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

Jalview demo/example

- Jalview is a multiple sequence alignment editor
- www.jalview.org
- Can run the algorithms Clustal W, MUSCLE and MAFFT from within the program
- Very useful for making nice colorful figures



Finding the best MSA

- How should we design a score model for MSA that leads us to the best biological alignment?

- Given a scoring model, how do we find the alignment that achieves the highest score?

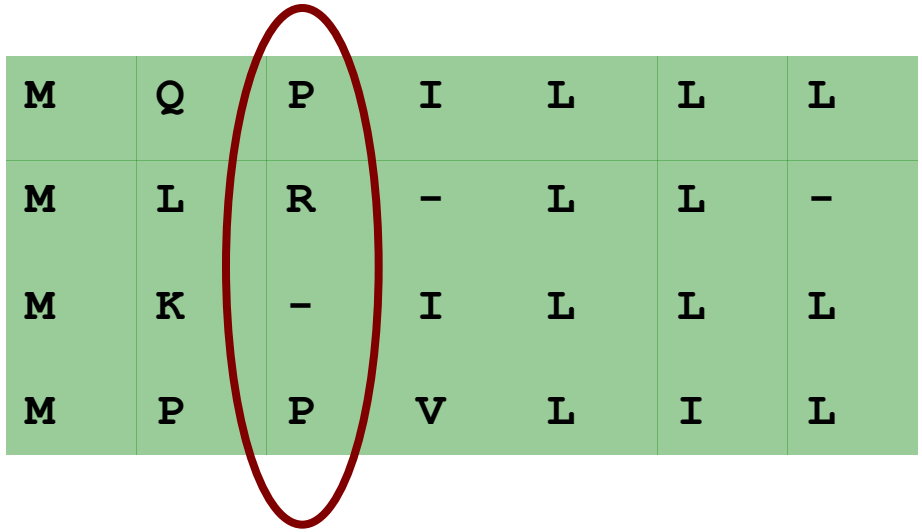
Sum-of-pairs (SP) score

- Objective function indicating multiple alignment quality
- The sum-of-pairs score (SP score) is a simple and commonly used objective function
- 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
- This leads to an alignment that optimizes all pairwise alignments with equal weight on all pairs.
- Columns with gaps in all sequences are ignored.
- The SP score does not necessarily give the biologically best alignment

Calculating 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)

Exhaustive algorithms

- Identifies the mathematically optimal alignment that is guaranteed to have the maximum sum-of-pairs score

- Dynamic programming
- Smith-Waterman generalized to N dimensions
- MSA
- (DCA)

- Computationally very demanding
- Only practical for a few short sequences
- Impossible for many and/or long sequences

Progressive algorithms

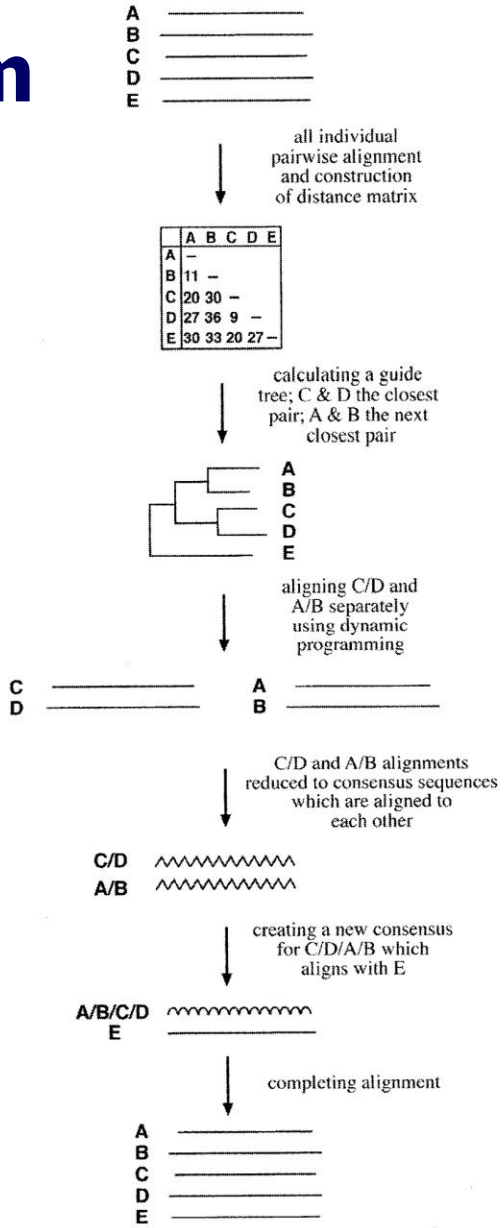
- Clustal W
- T-Coffee
- DbClustal
- Poa
- PRALINE

Progressive alignment

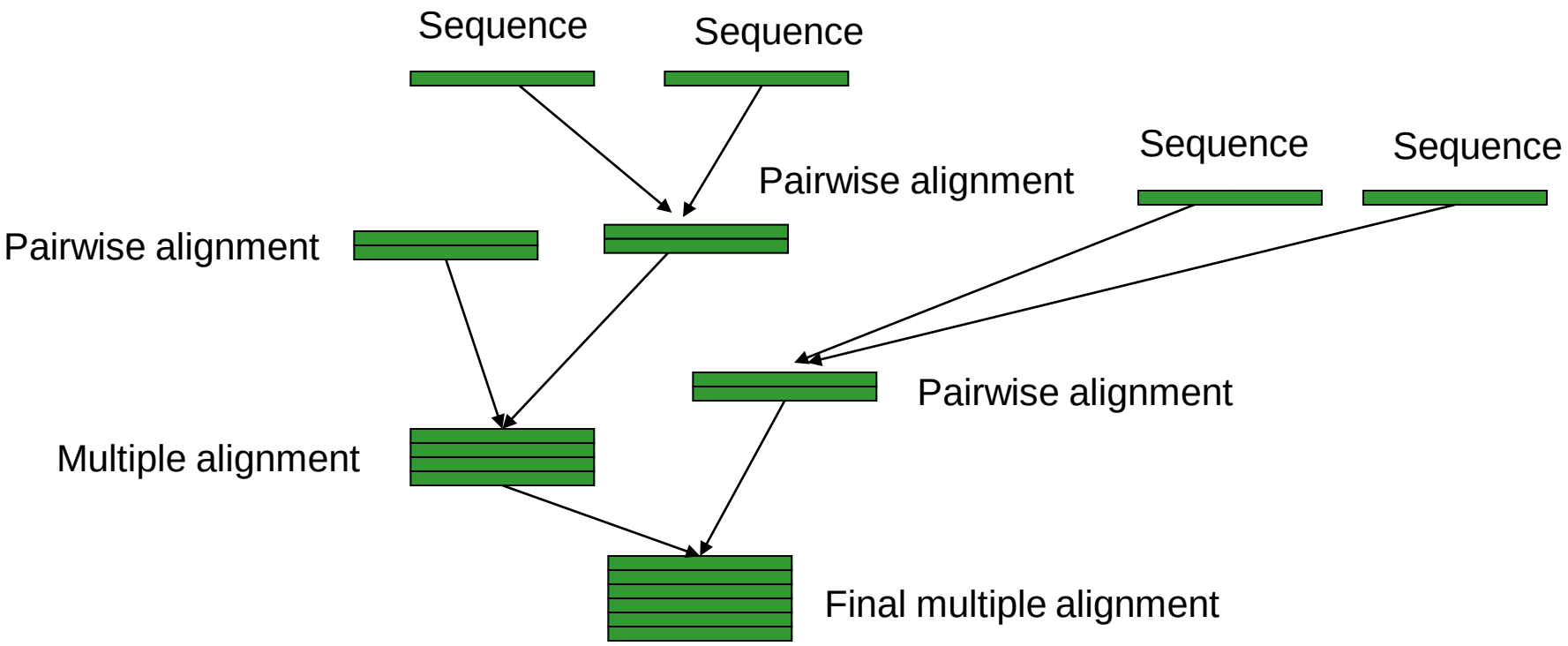
- Start with one sequence and add progressivily more sequences into the alignment
- Align one sequence with an alignment, or align to alignments with each other
- Disadvantage: Early errors in alignment may be fatal and cannot be fixed later
- The choice of which sequences to align and in which order is usually based on clustering or phylogenetic analysis and a guide tree.

Progressive alignment algorithm

1. Pairwise alignment of all n(n-1)/2 pairs of sequences. Construct distance matrix from scores.
2. Group sequences progressively based on pairwise distances and produce ”guide tree”
3. Based on guide tree, start by realigning the two closest sequences and treat them as one ”consensus” sequence in later steps
4. Gradually realign all pairs of sequences according to guide tree, always starting with the closest sequences, until all sequences are aligned
5. Repeat step 4 until all seqeunces are aligned.



Profile 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.
- Many versions, including one with a graphical user interface (Clustal X).

- Paper: Thompson JD, Higgins DG, Gibson TJ. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-4680.

- http://www.ebi.ac.uk/clustalw/

Iterative alignment

- Iteration is used to gradually increase the quality of the alignment by repeating some steps in the algorithm until a good alignment is reached
- Reduces/eliminates the effect of the order of aligning sequences used in progressive alignment

Programs:

- PRRN
- Muscle

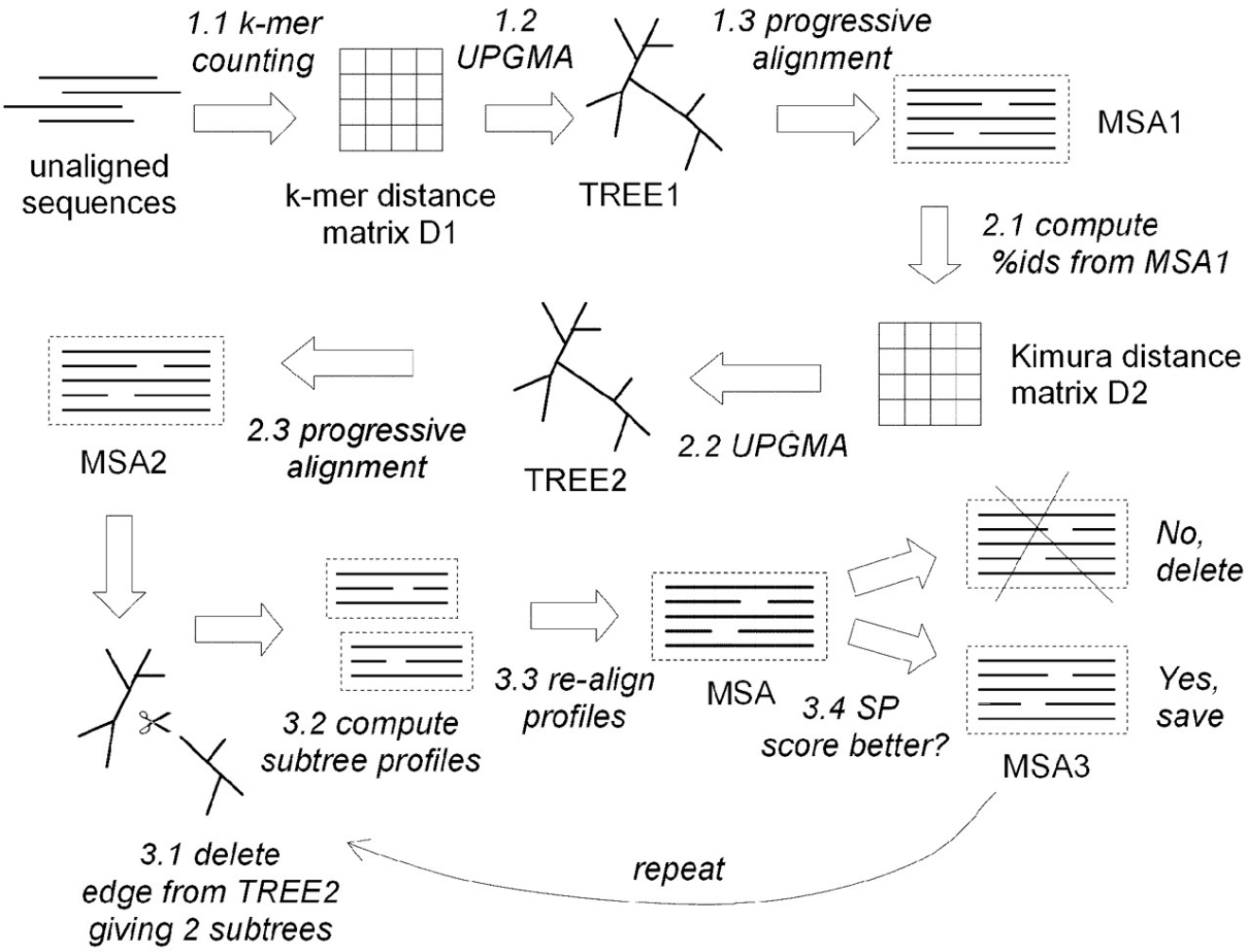
MUSCLE

- MUSCLE = Multiple Sequence Comparison by Log Expectation
- Very high quality of alignments
- Much faster than Clustal W

- Improvements:
  - Bases initial sequence distances on k-mer counting
  - Progressive alignment based on a new and better score function
  - Iterated partitioning of the tree and realignment

- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., 32, 1792-1797.

MUSCLE - Algorithm



MAFFT

- MAFFT is one of the most accurate MSA programs currently available
- Uses a combination of approaches
- Improved objective function
- http://align.bmr.kyushu-u.ac.jp/mafft/software/

Alignment of protein-coding DNA sequences

- Protein-coding DNA sequences should be aligned at the amino acid level
- Alignment at the DNA level may introduce unwanted gaps within codons
- Translate DNA sequences to protein before alignment, then translate back if necessary
- RevTrans and PROTA2DNA are multiple sequence alignment programs that align sequences by translating them to proteins

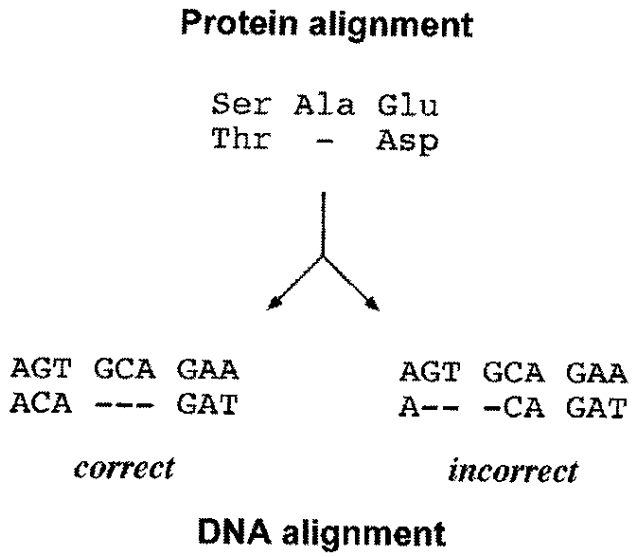


Figure 5.5: Comparison of alignment at the protein level and DNA level. The DNA alignment on the left is the correct one and consistent with amino acid sequence alignment, whereas the DNA alignment on the right, albeit more optimal in matching similar residues, is incorrect because it disregards the codon boundaries.