

This week. Week 3

- Objective: To obtain the mapping of epitopes to complete reference proteomes

Next week. Week 4.

- Objectives:
 - To learn about protein disorderness → Tue @2pm
Lilia Iakuscheva will be given a presentation on this topic
 - Start calculating sequence conservation for epitopes
 - Start calculating disorderness for epitopes

Week 3

- Objective: To obtain the mapping of epitopes to complete reference proteomes
- Algorithm:
 - Input (SA – source antigen):
 - {epitope ID; epitope sequence; SA ID}
 - {SA ID; SA sequence; source organism ID; parent taxonomic ID}
 - {parent taxonomic ID; reference proteome sequences}
 - TODO....
 - Output:
 - {epitope ID; ref. sequence ID; ref. sequence; epitope positions in ref sequence}

What you will learn today

- Sequence databases for genes and proteins: NCBI, UniProt
- FASTA format for sequences
- Pairwise sequence alignment: a bit of theory & major algorithms: BLAST

The NAR online Molecular Biology Database Collection

www.oxfordjournals.org/nar/database/c/

1512 online databases

2013 NAR Database Summary Paper Category List

- [Nucleotide Sequence Databases](#) ←
- [RNA sequence databases](#)
- [Protein sequence databases](#) ←
- [Structure Databases](#)
- [Genomics Databases \(non-vertebrate\)](#)
- [Metabolic and Signaling Pathways](#)
- [Human and other Vertebrate Genomes](#)
- [Human Genes and Diseases](#)
- [Microarray Data and other Gene Expression Databases](#)
- [Proteomics Resources](#)
- [Other Molecular Biology Databases](#)
- [Organelle databases](#)
- [Plant databases](#)
- [Immunological databases](#)
- [Cell biology](#)

- ▶ Compilation Paper
- ▶ Category List
- ▶ Alphabetical List
- ▶ Category/Paper List
- ▶ Search Summary Papers

Primary nucleotide sequence databases

Nucleotide Sequence Databases

International Nucleotide Sequence Database Collaboration

DDBJ - DNA Data Bank of Japan

EBI BioSample Database

EBI patent sequences

European Genome-phenome Archive (EGA)

European Nucleotide Archive



GenBank®

NCBI BioSample/BioProject

neXtProt

The Sequence Read Archive (SRA)

Coding and non-coding DNA

Gene structure, introns and exons, splice sites

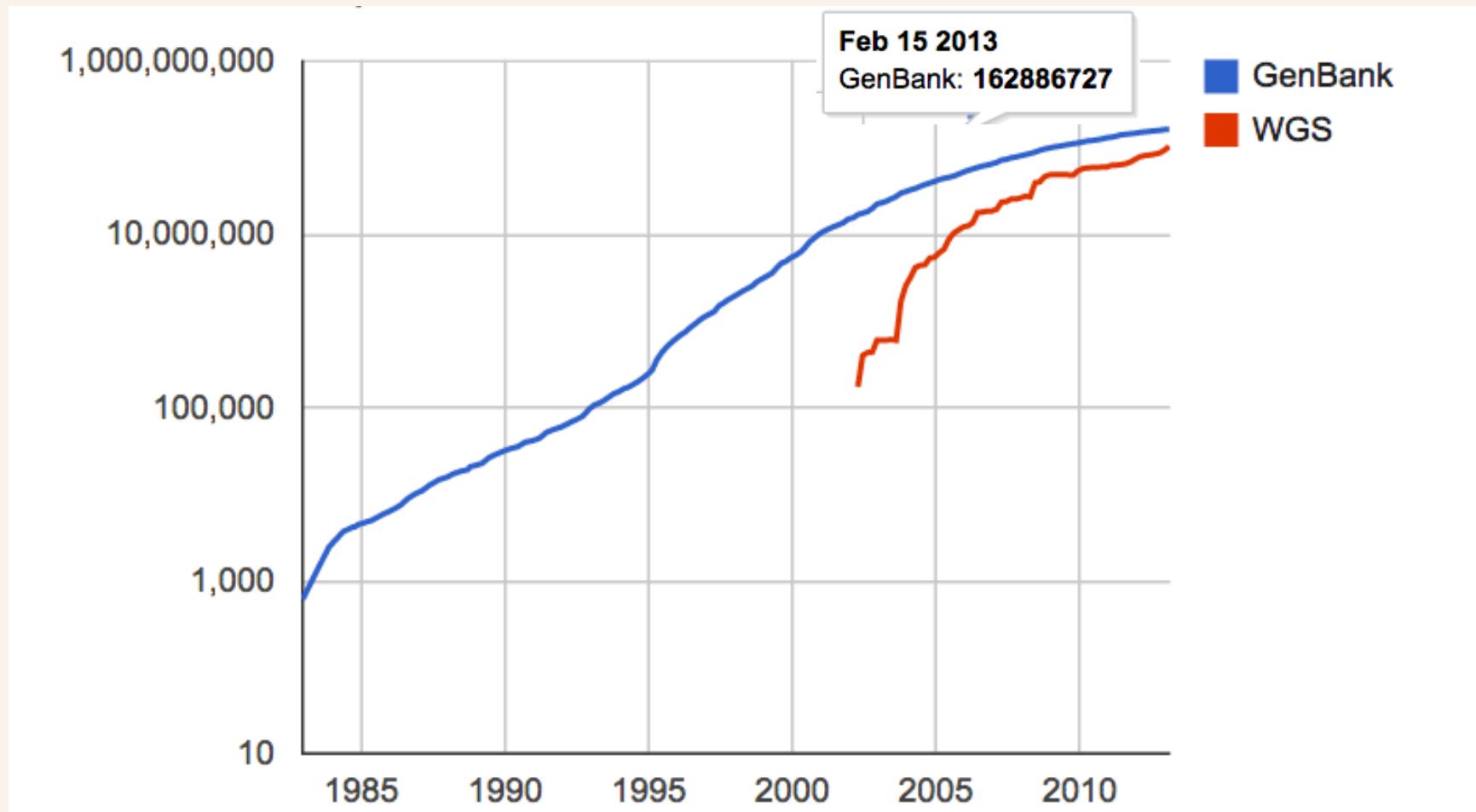
Transcriptional regulator sites and transcription factors

RNA sequence databases

Protein sequence databases

GenBank is part of the International Nucleotide Sequence Database Collaboration , which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

Growth of the number of sequences in NCBI GenBank



Whole Genome Shotgun (WGS) sequencing projects are incomplete genomes or incomplete chromosomes that are being sequenced by a whole genome shotgun strategy. WGS projects may be annotated, but annotation is not required.

Search for il1 in Entrez: must select “All Databases”

The screenshot shows the NCBI homepage. A red arrow points to the search bar, which has 'il1' typed into it. Above the search bar is a dropdown menu labeled 'All Databases' with a downward arrow, also highlighted by a red box.

NCBI Home

- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

NCBI Facebook page

Find out the latest news about NCBI resources and participate in community discussions.

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

Now Available: NCBI Insights Blog! 28 Jan 2013

NCBI has just released a new blog called *NCBI Insights*. Blog posts will provide an inside perspective to help

Come to the NCBI Discovery Workshops on February 4&5! 16 Jan 2013

Spaces are still available for the free, 2-day *Discovery Workshops* to be held on

New version of Genome Workbench available 06 Sep 2012

An integrated, downloadable application for viewing and analyzing sequences

[More...](#)

www.ncbi.nlm.nih.gov/gquery/?term=il1

NCBI

Entrez, The Life Sciences Search Engine.

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases il1 GO Clear Help

- Result counts displayed in gray indicate one or more terms not found

1411 PubMed: biomedical literature citations and abstracts	177 Books: online books
4575 PubMed Central: free, full text journal articles	135 OMIM: online Mendelian Inheritance in Man
107 Site Search: NCBI web and FTP sites	

1216 **Nucleotide:** Core subset of nucleotide sequence records

415 **EST:** Expressed Sequence Tag records

none **GSS:** Genome Survey Sequence records

2235 **Protein:** sequence database

2 **Genome:** whole genome sequences

50 **Structure:** three-dimensional macromolecular structures

none **Taxonomy:** organisms in GenBank

353 **SNP:** short genetic variations

99 **dbVar:** Genomic structural variation

335 **Gene:** gene-centered information

4 **SRA:** Sequence Read Archive

398 **BioSystems:** Pathways and systems of interacting molecules

10 **HomoloGene:** eukaryotic homology groups

4 **Probe:** sequence-specific reagents

21 **BioProject:** aggregated biological research project data

5 **dbGaP:** genotype and phenotype

15 **UniGene:** gene-oriented clusters of transcript sequences

4 **CDD:** conserved protein domain database

1128 **Clone:** integrated data for clone resources

1 **UniSTS:** markers and mapping data

37 **PopSet:** population study data sets

68772 **GEO Profiles:** expression and molecular abundance profiles

190 **GEO DataSets:** experimental sets of GEO data

none **Epigenomics:** Epigenetic maps and data sets

970 **PubChem BioAssay:** bioactivity screens of chemical substances

none **PubChem Compound:** unique small molecule chemical structures

7 **PubChem Substance:** deposited chemical substance records

3 **Protein Clusters:** a collection of related protein sequences

none **OMIA:** online Mendelian Inheritance in Animals

3 **BioSample:** biological material descriptions

3 **NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections

6 **MeSH:** detailed information about NLM's controlled vocabulary

A red arrow points to the "Nucleotide" entry in the left column.

Select human

Screenshot of the NCBI Nucleotide search results for "il1".

The search term "il1" was entered in the search bar. The results show 1631 nucleotide sequences.

Display Settings: Summary, 20 per page, Sorted by Default order.

Results: 1 to 20 of 1216

PREDICTED: Trichechus manatus latirostris interleukin 1 family, member 10 (theta) (IL1F10), mRNA

- 549 bp linear mRNA
Accession: XM_004388869.1 GI: 471414787
[GenBank](#) [FASTA](#) [Graphics](#)
- PREDICTED: Trichechus manatus latirostris interleukin 36, gamma (IL36G), mRNA
- 600 bp linear mRNA
Accession: XM_004391234.1 GI: 471414783
[GenBank](#) [FASTA](#) [Graphics](#)
- PREDICTED: Trichechus manatus latirostris interleukin 1, beta (IL1B), mRNA
- 834 bp linear mRNA
Accession: XM_004388864.1 GI: 471414775
[GenBank](#) [FASTA](#) [Graphics](#)
- PREDICTED: Trichechus manatus latirostris interleukin 1 receptor antagonist (IL1RN), mRNA
- 483 bp linear mRNA
Accession: XM_004388870.1 GI: 471414789
[GenBank](#) [FASTA](#) [Graphics](#)
- PREDICTED: Trichechus manatus latirostris interleukin 36, alpha (IL36A), mRNA
- 489 bp linear mRNA
Accession: XM_004388868.1 GI: 471414785
[GenBank](#) [FASTA](#) [Graphics](#)
- PREDICTED: Trichechus manatus latirostris interleukin 36 receptor antagonist (IL36RN), mRNA
- 468 bp linear mRNA
Accession: XM_004388865.1 GI: 471414777
[GenBank](#) [FASTA](#) [Graphics](#)
- PREDICTED: Trichechus manatus latirostris interleukin 1, alpha (IL1A), mRNA
- 807 bp linear mRNA
Accession: XM_004388863.1 GI: 471414773
[GenBank](#) [FASTA](#) [Graphics](#)
- PREDICTED: Trichechus manatus latirostris interleukin 18 (interferon-gamma-inducing factor) (IL18), mRNA
- 528 bp linear mRNA
Accession: XM_004391316.1 GI: 471398837
[GenBank](#) [FASTA](#) [Graphics](#)

Send to: [Email](#) [Print](#) [Copy](#) [Save](#)

Filter your results:

- All (1216)
- Bacteria (77)
- INSDC (GenBank) (777)
- mRNA (916)
- RefSeq (438)

[Manage Filters](#)

Top Organisms [Tree]

- Homo sapiens (129)
- Mus musculus (83)
- Helicobacter pylori (75)
- Rattus norvegicus (57)
- Variola virus (57)
- Bos taurus (41)
- synthetic construct (36)
- Equus caballus (29)
- Gallus gallus (29)
- Canis lupus familiaris (25)
- Gorilla gorilla (25)
- Sus scrofa (24)
- Gorilla gorilla gorilla (24)
- Ovis aries (22)
- Pan troglodytes (19)
- Oryctolagus cuniculus (18)
- Cavia porcellus (16)
- Macaca mulatta (16)
- Nomascus leucogenys (15)
- Felis catus (14)
- All other taxa (451)

[Less...](#)

Find related data

Database: [Select](#)

[Find Items](#)

Search details

il1[All Fields]

seqdump.txt

Select mRNA for IL1 alpha

Nucleotide Nucleotide (il1) AND "Homo sapiens"[porgn:__bxid9606] Search Help

Display Settings: Summary, 20 per page, Sorted by Default order Send to: Filter your results:

Found 316 nucleotide sequences. Nucleotide (129) EST (187)

Results: 1 to 20 of 129 << First < Prev Page 1 of 7 Next > Last >>

1. [Homo sapiens tripartite motif containing 38 \(TRIM38\), mRNA](#)
5,629 bp linear mRNA
Accession: NM_006355.3 GI: 393715115
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

2. [Homo sapiens trefoil factor 3 \(intestinal\) \(TFF3\), mRNA](#)
1,054 bp linear mRNA
Accession: NM_003226.3 GI: 281485607
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

3. [Homo sapiens ring finger protein 216 \(RNF216\), transcript variant 1, mRNA](#)
5,867 bp linear mRNA
Accession: NM_207111.3 GI: 319803088
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

4. [Homo sapiens toll interacting protein \(TOLLIP\), mRNA](#)
3,665 bp linear mRNA
Accession: NM_019009.3 GI: 296040496
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

5. [Homo sapiens ring finger protein 216 \(RNF216\), transcript variant 2, mRNA](#)
5,696 bp linear mRNA
Accession: NM_207116.2 GI: 319803089
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

6. [Homo sapiens interleukin 1, alpha \(IL1A\), mRNA](#)
2,943 bp linear mRNA
Accession: NM_000575.3 GI: 27894329
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

7. [Homo sapiens interleukin 1, beta \(IL1B\), mRNA](#)
1,498 bp linear mRNA
Accession: NM_000576.2 GI: 27894305
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

8. [Homo sapiens cytoskeleton associated protein 2-like \(CKAP2L\), mRNA](#)
3,273 bp linear mRNA

Send to: Filter your results:
All (129)
Bacteria (0)
[INSDC \(GenBank\) \(88\)](#)
[mRNA \(112\)](#)
[RefSeq \(41\)](#)
[Manage Filters](#)

Find related data
Database: Select Find items

Search details
il1[All Fields] AND "Homo sapiens"
[porgn]

Search See more...

Recent activity
Turn Off Clear
il1 AND "Homo sapiens"[porgn] (129) Nucleotide
il1 (1216) Nucleotide
gi|57618952|ref|NM_001009835.1| Felis catus... BLAST
gi|57618952|ref|NM_001009835.1| Felis catus... BLAST
gi|57618952|ref|NM_001009835.1| Felis catus... BLAST

See more...

Get FASTA-formatted sequence

Nucleotide Nucleotide Help

Display Settings: GenBank

Send:

Homo sapiens interleukin 1, alpha (IL1A), mRNA

NCBI Reference Sequence: NM_000575.3

[FASTA](#) [Graphics](#)

Go to:

Locus: NM_000575 2943 bp mRNA linear PRI 24-MAR-2013

Definition: Homo sapiens interleukin 1, alpha (IL1A), mRNA.

Accession: NM_000575

Version: NM_000575.3 GI:27894329

Keywords: .

Source: Homo sapiens (human)

Organism: Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominoidea; Homo.

Reference: 1 (bases 1 to 2943)

Authors: Kaarvatn,M.H., Jotanovic,Z., Mihelic,R., Etokebe,G.E., Mulac-Jericevic,B., Tijanic,T., Balen,S., Sestan,B. and Dembic,Z.

Title: Associations of the interleukin-1 gene locus polymorphisms with risk to hip and knee osteoarthritis: gender and subpopulation differences

Journal: Scand. J. Immunol. 77 (2), 151-161 (2013)

PubMed: 23216199

Remark: GeneRIF: Women carriers of the 1-2-1 haplotype [IL-1A(rs1800587) - IL-1B(rs1143634) - IL-1B(rs16944) - IL-1RN(VNTR)] had sixfold lower risk to develop knee osteoarthritis.

Reference: 2 (bases 1 to 2943)

Authors: Laine,M.L., Moustakis,V., Koumakis,L., Potamias,G. and Loos,B.G.

Title: Modeling susceptibility to periodontitis

Journal: J. Dent. Res. 92 (1), 45-50 (2013)

PubMed: 23100272

Remark: GeneRIF: Using decision tree analysis, we identified presence of bacterial species *Tannerella forsythia*, *Porphyromonas gingivalis*, *Aggregatibacter actinomycetemcomitans*; SNPs TNF -857 and IL-1A -889 as discriminators between periodontitis and non-periodontitis.

Reference: 3 (bases 1 to 2943)

Authors: Bajayo,A., Bar,A., Denes,A., Bachar,M., Kram,V., Attar-Namdar,M., Zallone,A., Kovacs,K.J., Yirmiya,R. and Bab,I.

Title: Skeletal parasympathetic innervation communicates central IL-1 signals regulating bone mass accrual

Journal: Proc. Natl. Acad. Sci. U.S.A. 109 (38), 15455-15460 (2012)

Send:

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Articles about the IL1A gene

Associations of the interleukin-1 gene locus polymorphisms with risk | [Scand J Immunol. 2013]

Modeling susceptibility to periodontitis. | [J Dent Res. 2013]

Skeletal parasympathetic innervation communicates c_x | [Proc Natl Acad Sci U S A. 2012]

See all...

Pathways for the IL1A gene

Spinal Cord Injury

Salmonella infection

Cytokine Signaling in Immune system

See all...

Reference sequence information

RefSeq protein product

See the reference protein sequence for interleukin-1 alpha proprotein (NP_000566.3).

More about the IL1A gene

FASTA format: the legacy of the a DNA and protein sequence alignment software package first described (as FASTP) by Lipman & Pearson in 1985

Sequences in this format are accepted by almost any bioinformatics program

Homo sapiens interleukin 1, alpha (IL1A), mRNA

NCBI Reference Sequence: NM_000575.3

[GenBank](#) [Graphics](#)

```
>gi|27894329|ref|NM_000575.3| Homo sapiens interleukin 1, alpha (IL1A), mRNA
ACCAGGCAACACCATTGAAGGCTCATATGTAAAAATCCATGCCTCCTTCTCCAATCTCCATTCCAA
ACTTAGGCCACTGGCTCTGGCTAGGGCTTACGCATACCTCCCAGGGCTTGACACACCTTCTACAG
AAGACACACCTTGGGCATATCTACAGAACGACCGAGCTCTCTGGTCCTGGTAGAGGGCTACTTAC
TGTAACAGGGCCAGGGTGGAGAGTTCTCTCTGAAGCTCCATCCCTCTATAGGAAATGTGTTGACAATA
TTCAGAAGAGTAAGAGGATCAAGACTCTTGTCTCAAATACCACTGTTCTCTACCCCTGCCCTA
ACCAGGAGCTTGTCACCCCAAACTCTGAGGTGATTATGCCCTAAAGCAAACCTCCCTTCAGAAA
AGATGGCTCATTTCCTCAAAAGTTGCCAGGAGCTGCCAAGTATTGCAATTCACCTGGAGCACAA
TCAACAAATTCAGCCAGAACACAACACTACAGCTACTATTAGAACTATTATTAAATAAATTCCCTCTCAA
ATCTAGCCCCCTGACTTCGGATTTCACGATTCTCCCTCCTCTAGAAACTGATAAGTTCCCGCGCT
TCCCTTTCTAAAGACTACATGTTGTCTTATAAAGCAAAGGGGTGAATAAATGAACCAAATCAATA
ACTTCTGGAATATCTGCAAACACAATAATATCAGCTATGCCATTTCACTATTTAGCCAGTATCGAG
TTGAATGAACATAGAAAAATACAAAAGTGAATTCTCCCTGTAATTCCCCGTTTGACGACGCAGTGT
AGCCACGTAGCCACGCCACTTAAGACAATTACAAAGGCGAAGAACAGTGAECTCAGGCTTAAGCTGCCA
GCCAGAGAGGGAGTCATTCATGGCTTGAGTCAGCAAAGAAGTCAAGATGCCAAAGTTCCAGACAT
GTTTGAAGACCTGAAGAACTGTACAGTGAAGAAGACAGTCCCTCATTGATCATCTGTCTG
AATCAGAAATCCTCTATCATGTAAGCTATGGCCCACCCATGAAGGCTGCATGGATCAATCTGTCTC
TGAGTATCTCTGAAACCTCTAAACATCCAAGCTTACCTCAAGGAGAGCATGGGGTAGTAGCAACCAA
CGGGAGGTTCTGAAGAAGAGACGGTTGAGTTAACCAATCCACTGATGATGACCTGGAGGCCATC
GCCAATGACTCAGAGGAAGAAATCATCAAGCCTAGGTAGCAGCACCTTGTAGCTCCTGAGCAATGTGAAAT
ACAACCTTATGAGGATCATCAAATACGAATTCTGCATGGCTCAATCAAAGTATAATTGAGC
CAATGATCAGTACCTCACGGCTGCTGCATTACATAACTGGATGAAGCAGTGAAGATGGCT
TATAAGTCATCAAAGGATGATGCTAAAATTACCGTATTCTAAGAATCTCAAAACTCAATTGTATGTGA
CTGCCCAAGATGAAGACCAACCAGTGCTGAGGAGATGCCAGGAGATACCCAAAACCATCACAGGTAG
TGAGACCAACCTCCTTCTGGAAACTCACGGCACTAAGAAGTATTGACATCAGTTGCCCATCCA
```

You can run BLAST to search for similar sequences from any sequence record in NCBI databases

Nucleotide Nucleotide Search Help

Display Settings: GenBank

Homo sapiens interleukin 1, alpha (IL1A), mRNA

NCBI Reference Sequence: NM_000575.3

FASTA Graphics

Go to:

LOCUS NM_000575 2943 bp mRNA linear PRI 24-MAR-2013

DEFINITION Homo sapiens interleukin 1, alpha (IL1A), mRNA.

ACCESSION NM_000575

VERSION NM_000575.3 GI:27894329

KEYWORDS

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 2943)

AUTHORS Kaarvatn,M.H., Jotanovic,Z., Mihelic,R., Etokebe,G.E., Mulac-Jericevic,B., Tijanic,T., Balen,S., Sestan,B. and Dembic,Z.

TITLE Associations of the interleukin-1 gene locus polymorphisms with risk to hip and knee osteoarthritis: gender and subpopulation differences

JOURNAL Scand. J. Immunol. 77 (2), 151-161 (2013)

PUBMED 23216199

REMARK GenoME: Women carriers of the 1-2-1-1 haplotype [IL-1A(rs1800587) - IL-1B(rs1143634) - IL-1B(rs16944) - IL-1RN(VNTR)] had sixfold lower risk to develop knee osteoarthritis.

REFERENCE 2 (bases 1 to 2943)

AUTHORS Laine,M.L., Moustakas,V., Koumakis,L., Potamias,G. and Loos,B.G.

TITLE Modeling susceptibility to periodontitis

JOURNAL J. Dent. Res. 92 (1), 45-50 (2013)

PUBMED 23100272

REMARK GenoME: Using decision tree analysis, we identified presence of bacterial species *Tannerella forsythia*, *Porphyromonas gingivalis*, *Aggregatibacter actinomycetemcomitans*; SNPs TNF -857 and IL-1 α -889 as discriminators between periodontitis and non-periodontitis.

REFERENCE 3 (bases 1 to 2943)

AUTHORS Bajayo,A., Bar,A., Denes,A., Bachar,M., Kram,V., Attar-Namdar,M., Zallone,A., Kovacs,K.J., Yirmiya,R. and Bab,I.

TITLE Skeletal parasympathetic innervation communicates central IL-1 signals regulating bone mass accrual

JOURNAL Proc. Natl. Acad. Sci. U.S.A. 109 (38), 15455-15460 (2012)

Send: Change region shown

Customize view

Analyze this sequence

- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

Articles about the IL1A gene

- Associations of the interleukin-1 gene locus polymorphisms with risk [Scand J Immunol. 2013]
- Modeling susceptibility to periodontitis. [J Dent Res. 2013]
- Skeletal parasympathetic innervation communicates with [Proc Natl Acad Sci U S A. 2012]

Pathways for the IL1A gene

- Spinal Cord Injury
- Salmonella infection
- Cytokine Signaling in Immune system

Reference sequence information

- RefSeq protein product
- See the reference protein sequence for interleukin-1 alpha proprotein (NP_000566.3).

More about the IL1A gene

NCBI BLAST to search NCBI databases

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

NM_000575.3

Query subrange [?](#)

From

To

Or, upload file [Choose File](#) No file chosen [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):
Nucleotide collection (nr/nt) [?](#)

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

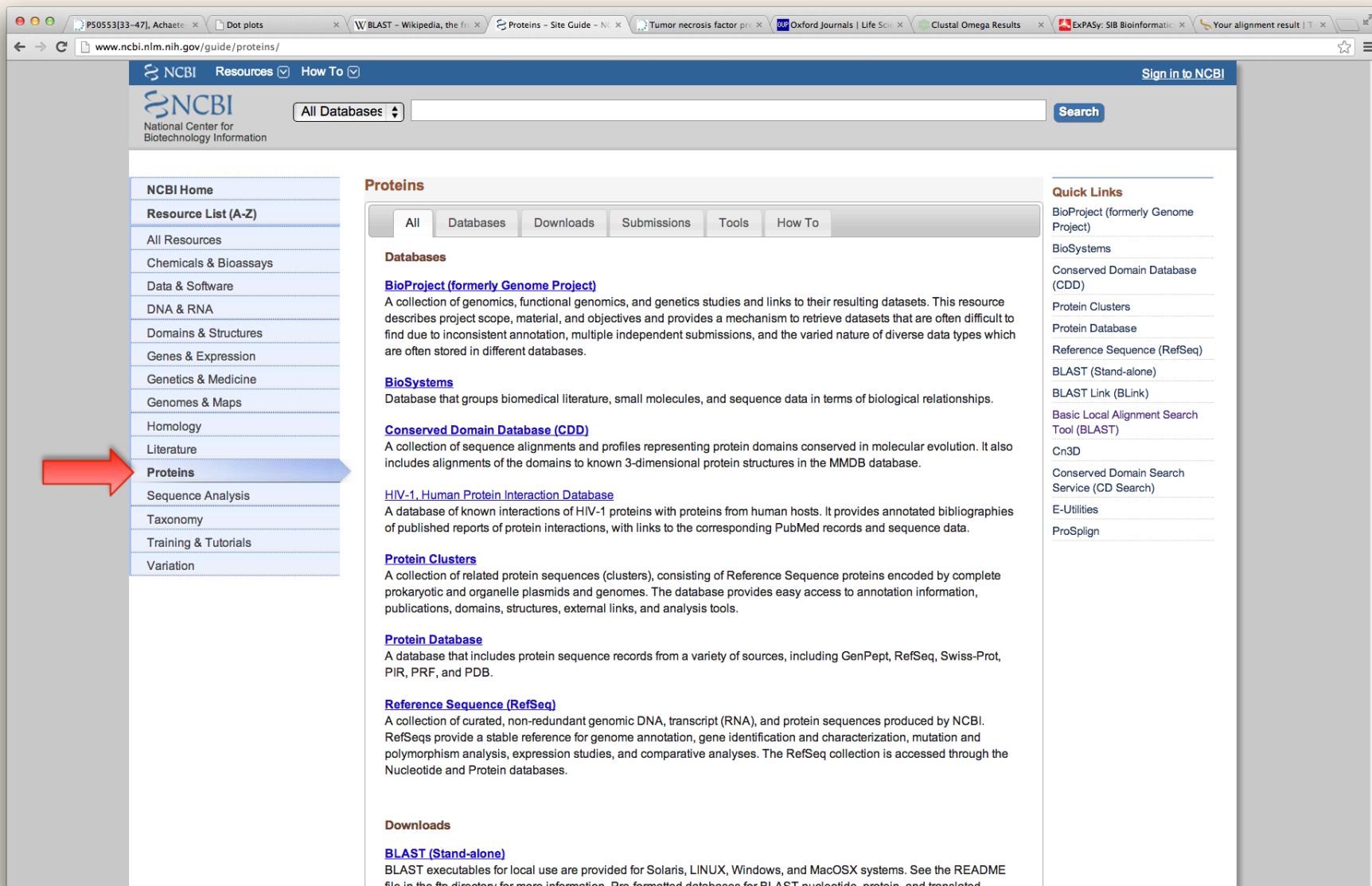
Entrez Query **Optional**
Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
Choose a BLAST algorithm [?](#)

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
 Show results in a new window

Here is the list of NCBI protein databases



The screenshot shows a web browser window with the NCBI Proteins Site Guide open. The URL is www.ncbi.nlm.nih.gov/guide/proteins/. The page has a blue header with the NCBI logo and a search bar. On the left, there's a sidebar with links like 'NCBI Home', 'Resource List (A-Z)', and 'Proteins' (which is highlighted with a red arrow). The main content area is titled 'Proteins' and contains sections for 'Databases', 'BioSystems', 'Conserved Domain Database (CDD)', 'HIV-1 Human Protein Interaction Database', 'Protein Clusters', 'Protein Database', and 'Reference Sequence (RefSeq)'. On the right, there's a 'Quick Links' sidebar with links to various NCBI services.

NCBI Resources How To

All Databases

Search

Sign in to NCBI

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Proteins

All Databases Downloads Submissions Tools How To

Databases

BioProject (formerly Genome Project)
A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.

BioSystems
Database that groups biomedical literature, small molecules, and sequence data in terms of biological relationships.

Conserved Domain Database (CDD)
A collection of sequence alignments and profiles representing protein domains conserved in molecular evolution. It also includes alignments of the domains to known 3-dimensional protein structures in the MMDB database.

HIV-1 Human Protein Interaction Database
A database of known interactions of HIV-1 proteins with proteins from human hosts. It provides annotated bibliographies of published reports of protein interactions, with links to the corresponding PubMed records and sequence data.

Protein Clusters
A collection of related protein sequences (clusters), consisting of Reference Sequence proteins encoded by complete prokaryotic and organelle plasmids and genomes. The database provides easy access to annotation information, publications, domains, structures, external links, and analysis tools.

Protein Database
A database that includes protein sequence records from a variety of sources, including GenPept, RefSeq, Swiss-Prot, PIR, PRF, and PDB.

Reference Sequence (RefSeq)
A collection of curated, non-redundant genomic DNA, transcript (RNA), and protein sequences produced by NCBI. RefSeqs provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies, and comparative analyses. The RefSeq collection is accessed through the Nucleotide and Protein databases.

Downloads

BLAST (Stand-alone)
BLAST executables for local use are provided for Solaris, LINUX, Windows, and MacOSX systems. See the README file in the [fto](#) directory for more information. Pre-formatted databases for BLAST nucleotide, protein, and translated

Quick Links

BioProject (formerly Genome Project)

BioSystems

Conserved Domain Database (CDD)

Protein Clusters

Protein Database

Reference Sequence (RefSeq)

BLAST (Stand-alone)

BLAST Link (BLink)

Basic Local Alignment Search Tool (BLAST)

Cn3D

Conserved Domain Search Service (CD Search)

E-Utilities

ProSalign

The NAR online Molecular Biology Database Collection

www.oxfordjournals.org/nar/database/c/

Nucleotide Sequence Databases

RNA sequence databases

Protein sequence databases

General sequence databases

CharProtDB

COMBREX

EXProt

MIPS resources

NCBI Protein database

PA-GOSUB

Patome

PIR - Protein Information Resource

PRF

RefSeq

TCDB

UniParc

UniProt

UniProt

UniRef

UniSave

Protein properties

Protein localization and targeting

Protein sequence motifs and active sites

Protein domain databases; protein classification

Databases of individual protein families

Structure Databases



- The Swiss-Prot, TrEMBL, and PIR protein database activities have united in 2004 to form **the Universal Protein Resource (UniProt)**, a central resource on protein sequences and functional annotation.
- **UniProtKB** is a “hub” for all available protein products produced by each gene. It has two parts: manually annotated **UniProtKB /Swiss-Prot** and automatically annotated **UniProtKB/TrEMBL**.
- **UniRef** speeds similarity searches via sequence space compression by merging sequences that are 100% (UniRef100), 90% (UniRef90), or 50% (UniRef50) identical.
- **UniProt Archive (UniParc)** stores all publicly available protein sequences, containing the history of sequence data with links to the source databases.

UniProt databases can be searched at uniprot.org via keyword or doing sequence search via BLAST

Screenshot of the UniProt homepage showing the search interface and various database sections.

The search bar at the top has "query" selected and contains the text "ii1". A red box highlights the search bar and input field.

WELCOME
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none">★ Swiss-Prot, which is manually annotated and reviewed.★ TrEMBL, which is automatically annotated and is not reviewed. Includes complete and reference proteome sets .
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more.

Getting started

- Text search
- Sequence similarity searches (BLAST)
- Sequence alignments
- Batch retrieval
- Database identifier mapping (ID Mapping)

NEWS

UniProt release 2013_03 - Mar 6, 2013
Latest from the prokaryotic world: bacterial Cas9, a new tool for genome engineering | Cross-references to ChiTaRS and SABIO-RK | Removal of cross-references to 8 2D gel databases and AGD

- Statistics for UniProtKB:
[Swiss-Prot](#) · [TrEMBL](#)
- Forthcoming changes
- News archives

[Follow @uniprot](#) 552 followers

SITE TOUR



Learn how to make best use of the tools and data on this site.

PROTEIN SPOTLIGHT

Output: sequences can be selected for download, further blasting or aligning using Clustal Omega

UniProtKB > UniProtKB

Search Blast Align Retrieve ID Mapping *

Search in Query

Protein Knowledgebase (UniProtKB) il1 Search Advanced Search » Clear

1 - 25 of 4,256 results for il1 in UniProtKB sorted by score descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% Download

Page 1 of 171 | Next »

Results Customize

Show only reviewed (497) ★ (UniProtKB/Swiss-Prot) or unreviewed (3,759) ★ (UniProtKB/TrEMBL) entries

Restrict term "il1" to protein family (542), gene name (25), protein name (276), strain (23), taxonomy (23), tissue (1)

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
Q90874	Q90874_CHICK	★	IL-1 receptor I	IL-1 receptor I	Gallus gallus (Chicken)	555
P21621	IL1B_SHEEP	★	Interleukin-1 beta	IL1B	Ovis aries (Sheep)	266
Q28386	IL1B_HORSE	★	Interleukin-1 beta	IL1B	Equus caballus (Horse)	268
P10749	IL1B_MOUSE	★	Interleukin-1 beta	Il1b	Mus musculus (Mouse)	269
P01582	IL1A_MOUSE	★	Interleukin-1 alpha	Il1a	Mus musculus (Mouse)	270
P01583	IL1A_HUMAN	★	Interleukin-1 alpha	IL1A IL1F1	Homo sapiens (Human)	271
P01584	IL1B_HUMAN	★	Interleukin-1 beta	IL1B IL1F2	Homo sapiens (Human)	269
Q28579	IL1A_SHEEP	★	Interleukin-1 alpha	IL1A	Ovis aries (Sheep)	268
P18430	IL1A_PIG	★	Interleukin-1 alpha	IL1A	Sus scrofa (Pig)	270
P08831	IL1A_BOVIN	★	Interleukin-1 alpha	IL1A	Bos taurus (Bovine)	268
Q28385	IL1A_HORSE	★	Interleukin-1 alpha	IL1A	Equus caballus (Horse)	270
P09428	IL1B_BOVIN	★	Interleukin-1 beta	IL1B	Bos taurus (Bovine)	266
P16598	IL1A_RAT	★	Interleukin-1 alpha	Il1a	Rattus norvegicus (Rat)	270
Q61730	IL1AP_MOUSE	★	Interleukin-1 receptor accessory protein	Il1rap	Mus musculus (Mouse)	570
P14628	IL1B_RABIT	★	Interleukin-1 beta	IL1B	Oryctolagus cuniculus (Rabbit)	268
Q9XS77	IL1B_TRIVU	★	Interleukin-1 beta	IL1B	Trichosurus vulpecula (Brush-tailed possum)	269
Q2MH07	IL1B_BUBCA	★	Interleukin-1 beta	IL1B	Bubalus carabanensis (Swamp type water buffalo) (Bubalus bubalis carabanensis)	266

Whole proteomes:
UniProt: complete vs. reference

www.uniprot.org/faq/15

UniProt FAQ Downloads · Contact · Documentation/Help

Search Blast Align Retrieve ID Mapping

Search in Query

FAQ Search Advanced Search » Clear

What are complete proteomes?

Last modified March 21, 2012

UniProt provides '[complete proteome](#)' sets of proteins thought to be expressed by organisms whose genomes have been completely sequenced.

What is a complete proteome?

A complete proteome is the entire set of proteins expressed by a specific organism. The majority of the UniProt complete proteomes are based on the translation of a completely sequenced genome, and will normally include sequences that derive from extra-chromosomal elements such as plasmids or organellar genomes in organisms where these occur. Some complete proteomes may also include protein sequences based on high quality cDNAs that cannot be mapped to the current genome assembly due to sequencing errors or gaps. These are only included in the complete proteome following manual review of the supporting evidence, including careful analysis of homologous sequences from closely related organisms.

What is the curation status of UniProt complete proteomes?

UniProt complete proteomes may include both manually reviewed (UniProtKB/Swiss-Prot) and unreviewed (UniProtKB/TrEMBL) entries. The proportion of reviewed entries varies between proteomes, and is obviously greater for the proteomes of intensively curated model organisms: some complete proteomes, such as those of *Saccharomyces cerevisiae* 288C and *Escherichia coli* strain K12 consist entirely of reviewed entries. Curation is a continuing process, and complete proteomes are updated in a regular manner as new information becomes available: pseudogenes and other dubious uncharacterized ORFs may be removed, other newly identified and characterized sequences may be added.

CROProgram....docx CRO-CID 2mts....pptx letter_NSF_UC....docx iedbProgRpt20....pdf pppt.1003390.s....pdf Show all downloads...

EN 4:55 PM 7/8/2013

www.uniprot.org/faq/15

1. Ensembl sequences are first mapped to their UniProtKB counterparts under stringent conditions, requiring 100% identity over 100% of the length of the two sequences. These entries are tagged with the keyword 'Complete proteome' and updated with an Ensembl cross-reference.
2. Ensembl sequences that are absent from UniProtKB are imported into UniProtKB/TrEMBL. These entries are tagged with the keyword 'Complete proteome' and have an Ensembl cross-reference.
3. All other UniProtKB/Swiss-Prot entries within the proteome that do not map to Ensembl are tagged with the keyword 'Complete proteome'.

Therefore, a complete proteome is formed from all UniProtKB/Swiss-Prot entries (irrespective of whether they map to Ensembl) plus those UniProtKB/TrEMBL entries mapping to Ensembl for that proteome.

To date this pipeline has been used to populate UniProtKB with additional sequences for the human and mouse proteomes (see headline [Complete proteomes for *Homo sapiens* and *Mus musculus*](#)) and many other vertebrates.

See also: [Where do the UniProtKB protein sequences come from?](#)

How to retrieve complete proteomes?

Complete proteomes for specific taxa can be retrieved by searching for the [taxonomic identifier](#) in the `organism` field together with the keyword 'Complete proteome'. For example, to retrieve the complete proteome for *Escherichia coli* (strain K12), which has the taxonomic identifier 83333, the required query would be:

• Query: [organism:83333 AND keyword:"Complete proteome"](#)

The taxonomic identifier can also be used to query the `taxonomy` field rather than the `organism` field. This will result in the retrieval of all complete proteome sequences at or below the taxonomic rank specified by the identifier. For example, to retrieve the complete proteome for *Escherichia coli* (strain K12) and all complete proteomes at lower taxonomic nodes (substrains such as *Escherichia coli* (strain K12 / DH10B)), then the required query would be:

• Query: [taxonomy:83333 AND keyword:"Complete proteome"](#)

CROProgram....docx CRO-CID 2mts....pptx letter_NSF_UC....docx iedbProgRpt20....pdf pppt.1003390.s....pdf Show all downloads... EN 4:56 PM 7/8/2013



Search

Blast

Align

Retrieve

ID Mapping *

Search in

Protein Knowledgebase (UniProtKB)

Query

organism:83333 AND keyword:"Complete proteome [KW-0181]"

Search

Advanced Search »

1 - 25 of 4,305 results for organism:"Escherichia coli (strain K12) [83333]" AND keyword:"Complete proteome [KW-0181]" in UniProtKB

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% [Download](#)

Page 1 of 173 | [Next](#)

Results [Customize](#)

- › Show only reviewed (4,303) ★ (UniProtKB/Swiss-Prot) or unreviewed (2) ★ (UniProtKB/TrEMBL) entries
- › Expand search to "Escherichia coli (strain K12) [83333]" to include lower taxonomic ranks
- › Show only entries from a reference proteome set (4,303)
- › Add columns: [Keywords](#)

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
P05100	3MG1_ECOLI	★	DNA-3-methyladenine glycosylase 1	tag b3549 JW3518	Escherichia coli (strain K12)	187
P04395	3MG2_ECOLI	★	DNA-3-methyladenine glycosylase 2	alkA aidA b2068 JW2053	Escherichia coli (strain K12)	282
P00350	6PGD_ECOLI	★	6-phosphogluconate dehydrogenase, decarboxyla...	gnd b2029 JW2011	Escherichia coli (strain K12)	468

www.iedb.org/sourceOrgId/467144

IMMUNE EPITOPE DATABASE AND ANALYSIS RESOURCE

Keyword Search

Home Browse Advanced Search Tools Support More

Modified Vaccinia Ankara virus epitopes

Source Organism

Source Organism: Modified Vaccinia Ankara virus
Source NCBI Taxonomy ID: 467144
Parent NCBI Taxonomy ID: 10245

Structure (93) Reference (10) Source Antigen (57) MHC Binding (38) T Cell Assay (127) MHC Ligand Elution (6)

93 item(s) found, displaying 1 to 25 (Click the column headers to adjust the sorting)

« previous 1 2 3 4 next » Go To 1

Export all results: (full)

Epitope ID ↑	Structure	Source Antigen	Source Organism
1780	AHINALEY	Protein A47	Modified Vaccinia Ankara virus
4868	ATAAVCLLFIQGYSIYENYGN	putative 16.3k protein	Modified Vaccinia Ankara virus
5364	AVHLIIYQLAGYILTVLGLG	putative 31.5k protein	Modified Vaccinia Ankara virus
6630	CLTEYILWV	putative 21.7k protein (2 more)	Vaccinia virus (2 more)
6637	CLVFEPPVNQSGIEILLYFK	nucleoside triphosphate phosphohydrolase I, DNA helicase	Modified Vaccinia Ankara virus
7566	DAKNNAAKLAVDKL	dsRNA dependent PK inhibitor	Modified Vaccinia Ankara virus
7892	DDYMFVVIKPNLGR	protein kinase	Modified Vaccinia Ankara virus
8263	DFFKFSFMYIESIKVDRIGDN	putative 49.1k protein	Modified Vaccinia Ankara virus
9432	DNCILANRCFVKI	putative 18.6k protein	Modified Vaccinia Ankara virus
11202	EATKLCDWV	Thymidine kinase (1 more)	Modified Vaccinia Ankara virus (1 more)

CROProgram....docx CRO-CID 2mts....pptx letter_NSF_UC....docx iedbProgRpt20....pdf ppal.1003390.s....pdf Show all downloads... 5:15 PM 7/8/2013

www.uniprot.org/uniprot/?query=taxonomy%3A467144+AND+keyword%3A"complete+proteome"&sort=score

UniProtKB

Downloads · Contact · Documentation/Help

Search Blast Align Retrieve ID Mapping

Search in Query

Protein Knowledgebase (UniProtKB)

taxonomy: 467144 AND keyword:"complete proteome"

Search Advanced Search » Clear

0 result for taxonomy:"Modified Vaccinia Ankara virus [467144]" AND keyword:"complete proteome" in UniProtKB

Results Customize

Can't find what you are looking for? Please contact us.

© 2002–2013 UniProt Consortium | License & Disclaimer | Contact



CROProgram....docx CRO-CID 2mts....pptx letter_NSF_UC....docx iedbProgRpt20....pdf ppal.1003390.s....pdf Show all downloads...

EN 5:18 PM
7/8/2013



Search

Blast

Align

Retrieve

ID Mapping *

Search in

Query

Protein Knowledgebase (UniProtKB)

taxonomy:10245 AND keyword:"complete proteome"

Search

Advanced Search »

1 - 25 of 474 results for taxonomy:"Vaccinia virus [10245]" AND keyword:"complete proteome" in UniProtKB sorted by score descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% [Download](#)



Page 1

of 19 | [Next](#)

Results [Customize](#)

- › Show only [reviewed \(473\)](#) (UniProtKB/Swiss-Prot) or [unreviewed \(1\)](#) (UniProtKB/TrEMBL) entries
- › Show only entries from a [reference proteome set \(217\)](#)
- › Add columns: [Keywords](#)

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
P20994	A19_VACCC		Protein A19	A19L	Vaccinia virus (strain Copenhagen) (VACV)	77
P21114	A26_VACCC		Putative A-type inclusion protein	A26L	Vaccinia virus (strain Copenhagen) (VACV)	322
P21080	E2_VACCC		Protein E2	E2L	Vaccinia virus (strain Copenhagen) (VACV)	737
P21020	F15_VACCC		Protein F15	F15L	Vaccinia virus (strain Copenhagen) (VACV)	158

www.uniprot.org/uniprot/?query=taxonomy%3A10245+AND+keyword%3A"complete+proteome"&sort=score

					(VACV)	
<input type="checkbox"/>	P21020	F15_VACCC		Protein F15	F15L	Vaccinia virus (strain Copenhagen) (VACV) 158
<input type="checkbox"/>	P24764	SEMA_VACCW		Semaphorin-like protein A39	VACWR163 A39R	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR)) 295
<input type="checkbox"/>	P20841	SPI2A_VACCC		Putative serine proteinase inhibitor 2 homolo...	B13R	Vaccinia virus (strain Copenhagen) (VACV) 116
<input type="checkbox"/>	P20842	SPI2B_VACCC		Putative serine proteinase inhibitor 2 homolo...	B14R	Vaccinia virus (strain Copenhagen) (VACV) 222
<input type="checkbox"/>	P21100	VC16_VACCC		Protein C16/B22	B22R C16L	Vaccinia virus (strain Copenhagen) (VACV) 181
<input type="checkbox"/>	P21103	VC19_VACCC		Protein C19/B25	B25R C19L	Vaccinia virus (strain Copenhagen) (VACV) 259
<input type="checkbox"/>	P21096	A31_VACCC		Protein A31	A31R	Vaccinia virus (strain Copenhagen) (VACV) 124
<input type="checkbox"/>	P24760	A31_VACCW		Protein A31	VACWR154 A31R	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR)) 124
<input type="checkbox"/>	P21060	A37_VACCC		Protein A37	A37R	Vaccinia virus (strain Copenhagen) (VACV) 263
<input type="checkbox"/>	P24762	A37_VACCW		Protein A37	VACWR160 A37R	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR)) 263
<input type="checkbox"/>	P21067	A47_VACCC		Protein A47	A47L	Vaccinia virus (strain Copenhagen) (VACV) 244
<input type="checkbox"/>	P26673	A47_VACCW		Protein A47	VACWR173 A47L	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR)) 252
<input type="checkbox"/>	P21068	A49_VACCC		Protein A49	A49R	Vaccinia virus (strain Copenhagen) (VACV) 162
<input type="checkbox"/>	P21069	A51_VACCC		Protein A51	A51R	Vaccinia virus (strain Copenhagen) (VACV) 334

Search

Blast

Align

Retrieve

ID Mapping *

Search in

Protein Knowledgebase (UniProtKB)

Query

taxonomy:10245 AND keyword:"complete proteome"

Search

Advanced Search »

1 - 25 of 474 results for taxonomy:"Vaccinia virus [10245]" AND keyword:"complete proteome" in UniProtKB sorted by score descending

[Browse by taxonomy, keyword, gene ontology, enzyme class or pathway](#) | [Reduce sequence redundancy to 100%, 90% or 50%](#)

Download



Page 1

of 19 | Next »

Results Customize

- › Show only reviewed (473) ★ (UniProtKB/Swiss-Prot) or unreviewed (1) ★ (UniProtKB/TrEMBL) entries
- › Show only entries in reference proteome set (217)
- › Add columns: [Keywords](#)

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
P20994	A19_VACCC	★	Protein A19	A19L	Vaccinia virus (strain Copenhagen) (VACV)	77
P21114	A26_VACCC	★	Putative A-type inclusion protein	A26L	Vaccinia virus (strain Copenhagen) (VACV)	322
P21080	E2_VACCC	★	Protein E2	E2L	Vaccinia virus (strain Copenhagen) (VACV)	737
P21020	F15_VACCC	★	Protein F15	F15L	Vaccinia virus (strain Copenhagen) (VACV)	158

CROProgram....docx

CRO-CID 2mts....pptx

letter_NSF_UC....docx

Show all downloads...

Search

Blast

Align

Retrieve

ID Mapping *

Search in

Query

Protein Knowledgebase (UniProtKB)

taxonomy:10245 AND keyword:"complete proteome" AND keyword:1185

Search

Advanced Search »

1 - 25 of 217 results for taxonomy:"Vaccinia virus [10245]" AND keyword:"complete proteome" AND keyword:"Reference proteome [1185]" in UniProtKB sorted by score descending

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% [Download](#)

Page 1

of 9 | [Next](#)

Results [Customize](#)

› Show only [reviewed](#) (216) (UniProtKB/Swiss-Prot) or [unreviewed](#) (1) (UniProtKB/TrEMBL) entries

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
P24764	SEMA_VACCW		Semaphorin-like protein A39	VACWR163 A39R	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))	295
P24760	A31_VACCW		Protein A31	VACWR154 A31R	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))	124
P24762	A37_VACCW		Protein A37	VACWR160 A37R	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))	263
P26673	A47_VACCW		Protein A47	VACWR173 A47I	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))	252

www.uniprot.org/taxonomy/complete-proteomes

UniProt Taxonomy Downloads · Contact · Documentation/Help

Search Blast Align Retrieve ID Mapping

Search in Taxonomy Query * AND complete:yes Search Advanced Search » Clear

COMPLETE PROTEOMES AND REFERENCE PROTEOMES

A **complete proteome** consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

A **reference proteome** is the complete proteome of a representative, well-studied model organism or an organism of interest for biomedical research.

These organisms can be searched via the taxonomy pages, which provide links to download complete and reference proteome sets when available, as well as links to the HAMAP web site.

Browse or list organisms with:

Complete proteomes	Reference proteomes
Browse by hierarchy	Browse by hierarchy
List all Bacteria	List all Bacteria
List all Archaea	List all Archaea
List all Eukaryota	List all Eukaryota
List all Viruses	List all Viruses

Search organisms with complete proteomes:

CROProgram....docx CRO-CID 2mts....pptx letter_NSF_UC....docx iedbProgRpt20....pdf pppt.1003390.s....pdf Show all downloads... 5:00 PM 7/8/2013

F A Q

- › [What are complete proteome sets?](#)
- › [What are reference proteome sets?](#)
- › [How to retrieve sets of protein sequences?](#)
- › [What is HAMAP?](#)
HAMAP is a system, based on manual protein annotation, that identifies and semi-automatically annotates proteins... [More](#)

www.uniprot.org/taxonomy/complete-proteomes

UniProt Taxonomy

Search Blast Align Retrieve ID Mapping

Search in Taxonomy Query * AND complete:yes Search Advanced Search » Clear

COMPLETE PROTEOMES AND REFERENCE PROTEOMES

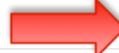
A **complete proteome** consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.

A **reference proteome** is the complete proteome of a representative, well-studied model organism or an organism of interest for biomedical research.

These organisms can be searched via the taxonomy pages, which provide links to download complete and reference proteome sets when available, as well as links to the HAMAP web site.

Browse or list organisms with:

Complete proteomes	Reference proteomes
Browse by hierarchy	Browse by hierarchy
List all Bacteria	List all Bacteria
List all Archaea	List all Archaea
List all Eukaryota	List all Eukaryota
List all Viruses	List all Viruses



Search organisms with complete proteomes:

F A Q

› [What are complete proteome sets?](#)

› [What are reference proteome sets?](#)

› [How to retrieve sets of protein sequences?](#)

› [What is HAMAP?](#)

HAMAP is a system, based on manual protein annotation, that identifies and semi-automatically annotates proteins... [More](#)

www.uniprot.org/taxonomy/?query=reference:yes%20ancestor:10239

UniProt Taxonomy Downloads · Contact · Documentation/Help

Search Blast Align Retrieve ID Mapping

Search in Taxonomy Query reference:yes ancestor:10239 Search Advanced Search » Clear

1 - 25 of 385 results for reference:yes AND ancestor:10239 in Taxonomy

Browse by hierarchy Download

Page 1 of 16 | Next »

Results Customize

› Show only taxa with annotated and reviewed (317) proteins

› Show only taxa with annotated (384) proteins

Acanthamoeba polyphaga mimivirus (APMV) ★

Viruses > dsDNA viruses, no RNA stage > Mimiviridae > Mimivirus

Reference proteome set (909)

Acholeplasma phage L2 (Bacteriophage L2) ★

Viruses > dsDNA viruses, no RNA stage > Plasmaviridae > Plasmavirus

Reference proteome set (14)

Acidianus bottle-shaped virus (isolate Italy/Pozzuoli) (ABV) ★

Viruses > dsDNA viruses, no RNA stage > Ampullaviridae > Ampullavirus

Reference proteome set (54)

Acidianus filamentous virus 1 (isolate United States/Yellowstone) (AFV-1) ★

Viruses > dsDNA viruses, no RNA stage > Lipothrixviridae > Gammalipothrixvirus

www.uniprot.org/taxonomy/654909



www.uniprot.org/taxonomy/?query=reference%3ayes+ancestor%3a10239&format=*

UniProt Taxonomy Downloads · Contact · Documentation/Help

Search Blast Align Retrieve ID Mapping

Search in Query

Taxonomy reference:yes ancestor:10239 Search Advanced Search » Clear

385 results for reference:yes AND ancestor:10239 in Taxonomy

› Download data compressed or uncompressed

Tab-Delimited
Summary information from the result view.
[Download (50 KB*) | Open | Open first 10]

Excel 
Summary information from the result view for MS Excel™.
[Download (50 KB*) | Open | Open first 10]

RDF/XML
Complete data in RDF format.
[Download (200 KB*) | Open | Open first 10]

List
List of NCBI taxonomy identifiers.
[Download (3 KB*) | Open | Open first 10]

* Estimate on the basis of the average entry size.

© 2002–2013 UniProt Consortium | License & Disclaimer | Contact



We need to map each epitope to a reference proteome

- Algorithm:
 - Input (SA – source antigen):
 - {epitope ID; epitope sequence; SA ID}
 - {SA ID; SA sequence; source organism ID; parent taxonomic ID}
 - **{parent taxonomic ID; reference proteome sequences}**
 - TODO....
 - Output:
 - {epitope ID; ref. sequence ID; ref. sequence; epitope positions in ref sequence}

To get {parent taxonomic ID; reference proteome sequences}

- For each Epitope → Source Organism → Source NCBI Taxonomy ID → “Parent NCBI Taxonomy ID”
- Scenario #1:
 - Get a list of Taxonomy IDs for all complete reference proteomes (UniProt or NCBI)
 - Look up if for “Parent NCBI Taxonomy ID” a reference proteome is available. If not, delete the epitope from the list.
 - Get the list of “Parent NCBI Taxonomy ID” from the filtered epitope list.
 - Get complete reference proteomes for this list.
 - The mapping program will work with filtered epitope list and a set of proteomes.
- Scenario #2 (more general):
 - Get the sequences of all complete reference proteomes.
 - The mapping program will work with initial list of all epitopes and all proteomes.

- You need to figure out how to get all complete proteome sequences for bacteria and viruses
- Or, alternatively, but worse for reproduction and slow approach, to implement Scenario #1.

Sequence alignment

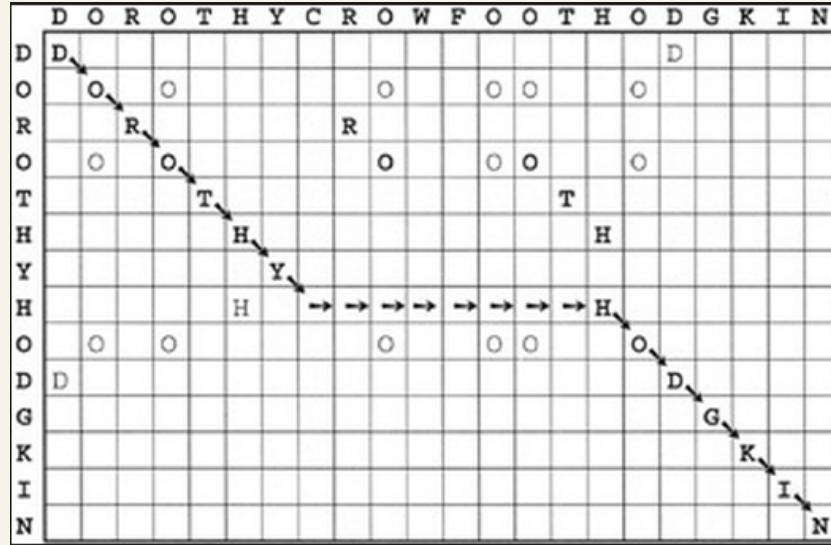
Why to align sequences?

- To search sequences in the databases.
- To annotate (assign function to) genes, proteins, genomes by means of transferring curated annotation of **homologous** -- orthologous (*conserved across species and direct descendants of a sequence in a common ancestor*) and paralogous (*conserved within species and separated by the event of gene duplication*) -- genes/proteins.
- The underlining hypothesis is that all genes and proteins are a product of evolution.
 - Therefore, conserved along evolution gene and protein sequences (*forming a functional family*) might have a similar function.
 - Likewise, a conserved region or a sequence motif might suggest that this region or motif has structural or functional importance (*e.g., active sites of enzymes, binding sites of protein receptors, cis-regulatory elements in DNA*).

Pairwise sequence alignment methods

- **Dot matrix methods:** applets Dotlet, DNADot, Dottup (*can be used for comparing entire genomes*); EBI LALIGN program generates dotplots
- **Global:**
 - **The dynamic programming algorithms:**
 - The Needleman-Wunsch algorithm (1970; NCBI BLAST web-site, also Needle at EBI www.ebi.ac.uk/Tools/psa/)
- **Local:**
 - **The dynamic programming algorithms:**
 - LALIGN (uses modified Smith-Waterman algorithm of 1981); EBI
 - **Word or k-tulip methods:**
 - NCBI BLAST server ([Altschuls et al., 1991](#)); used to search NCBI databases and UniProt
 - FASTA ([Lipman & Pearson, 1985](#)) - available at EBI to search for UniProt protein databases
- Wikipedia lists 15 methods for database search and 41 for pairwise alignment (some for protein or DNA only, some for both, 25 are available via web sites)

Pairwise sequence comparison: Dotplots



If two sequences are closely related, the alignment can be read directly from the dotplot:

DOROTHYCROWFOOTHODGKIN

DOROTHY-----HODGKIN

Let's generate a dotplot for the
human mucin using LALIGN

Open NCBI main page www.ncbi.nlm.nih.gov
Type “human mucin” and select “protein” database

The screenshot shows the NCBI homepage. A red box highlights the search bar where 'Protein' is selected from a dropdown menu and 'human mucin' is typed into the search field. The search button is visible to the right. The left sidebar contains a navigation menu with categories like NCBI Home, Resource List (A-Z), All Resources, Chemicals & Bioassays, Data & Software, DNA & RNA, Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, and Variation. The right sidebar features sections for Popular Resources (PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem) and NCBI Announcements (with a link to the NCBI Insights Blog). A central banner for the Genetic Testing Registry is also present.

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

NCBI

Resources

How To

Sign in to NCBI

Protein human mucin

Search

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genetic Testing Registry

A portal to clinical genetics resources with detailed information about genetic tests and laboratories.

GO

II 1 2 3 4 5 6 7 8

Popular Resources

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

NCBI Announcements

Now Available: NCBI Insights Blog! 28 Jan 2013

NCBI has just released a new blog called *NCBI Insights*. Blog posts will provide an insider's perspective to help

Come to the NCBI Discovery

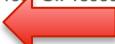
Select “FASTA” for the second protein

Protein Protein Save search Limits Advanced Help

Display Settings: Summary, 20 per page, Sorted by Default order **Send to:** Filter your results:

Results: 1 to 20 of 2661 << First < Prev Page of 134 Next > Last >>

[mucin \[Homo sapiens\]](#)
1. 1045 aa protein
Accession: CAA03985.1 GI: 3649741
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[mucin \[Homo sapiens\]](#)
2. 1255 aa protein
Accession: AAA60019.1 GI: 189599
[GenPept](#) [FASTA](#)  [Related Sequences](#) [Identical Proteins](#)

[mucin \[Homo sapiens\]](#)
3. 294 aa protein
Accession: AAA63229.2 GI: 8572538
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

[mucin, MG2 {N-terminal, fraction C-MG2a-T1} \[human, submandibular-sublingual saliva, Peptide Partial, 26 aa\]](#)
4. 26 aa protein
Accession: AAB28174.1 GI: 410695
[GenPept](#) [FASTA](#) [Graphics](#)

[mucin, MG2 {N-terminal, fraction C-MG2a-T2} \[human, submandibular-sublingual saliva, Peptide Partial, 20 aa\]](#)
5. 20 aa protein
Accession: AAB28175.1 GI: 410696
[GenPept](#) [FASTA](#) [Graphics](#)

[mucin \[Homo sapiens\]](#)

All (2661)
[Bacteria \(515\)](#)
[Related Structures \(1759\)](#)
[RefSeq \(599\)](#)
[Manage Filters](#)

Top Organisms [Tree]
Homo sapiens (913)
Mus musculus (219)
Schistosoma mansoni (211)
Trypanosoma cruzi (188)
Trypanosoma cruzi marinkellei (94)
All other taxa (1114)
[More...](#)

Find related data
Database:

Search details
("Homo sapiens"[Organism] OR human[All Fields]) AND mucin[All Fields]

This is a sequence of the human musin:
keep this window open and copy this sequence to clipboard

Open <http://www.ebi.ac.uk/Tools/psa/lalign/>

The screenshot shows the LALIGN tool interface on a web browser. The top navigation bar includes links for Services, Research, Training, Industry, and About us. Below the header, there are tabs for Protein alignment, Nucleotide alignment, Web services, and Help & Documentation. The main content area is titled "Pairwise Sequence Alignment" and states: "LALIGN finds internal duplications by calculating non-intersecting local alignments of **protein** or **nucleotide** sequences." The first step, "STEP 1 - Enter your protein sequences", contains two text input fields. The first field contains the sequence: >gi|189599|gb|AAA60019.1l mucin [Homo sapiens] MTPGTQSPFFLLLLLTVVTGSGHASSTPGEKEKTSATQRSSVPSSTEKNAVSMTSSVLSSHSPGS... The second field also contains the same sequence. Below each field is a "Choose File" button and a "No file chosen" message. Step 2, "STEP 2 - Set your pairwise alignment options", has a note: "The default settings will fulfill the needs of most users and, for that reason, are not visible." A "More options..." link is provided. Step 3, "STEP 3 - Submit your job", includes a checkbox for "Be notified by email" and a "Submit" button. A large red arrow points to the "Submit" button.

EMBL-EBI LALIGN

Protein alignment Nucleotide alignment Web services Help & Documentation Share Feedback

Tools > Pairwise Sequence Alignment > LALIGN

Pairwise Sequence Alignment

LALIGN finds internal duplications by calculating non-intersecting local alignments of **protein** or **nucleotide** sequences.

STEP 1 - Enter your protein sequences

Enter or paste your first **protein** sequence in any supported format:

```
>gi|189599|gb|AAA60019.1l mucin [Homo sapiens]
MTPGTQSPFFLLLLLTVVTGSGHASSTPGEKEKTSATQRSSVPSSTEKNAVSMTSSVLSSHSPGS...
STTQGQDVTLAPATEPASGSAAATWQGDVTSVPVTRALGSTTPPAHDVTSAPDNKPAPGSTAPPAGVTS
APDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGS
TAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTS
APDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGS
TAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTS
```

Or, upload a file: No file chosen

AND

Enter or paste your second **protein** sequence in any supported format:

```
>gi|189599|gb|AAA60019.1l mucin [Homo sapiens]
MTPGTQSPFFLLLLLTVVTGSGHASSTPGEKEKTSATQRSSVPSSTEKNAVSMTSSVLSSHSPGS...
STTQGQDVTLAPATEPASGSAAATWQGDVTSVPVTRALGSTTPPAHDVTSAPDNKPAPGSTAPPAGVTS
APDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGS
TAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTS
APDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGS
TAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTSAPDTRPAPGSTAPPAGVTS
```

Or, upload a file: No file chosen

STEP 2 - Set your pairwise alignment options

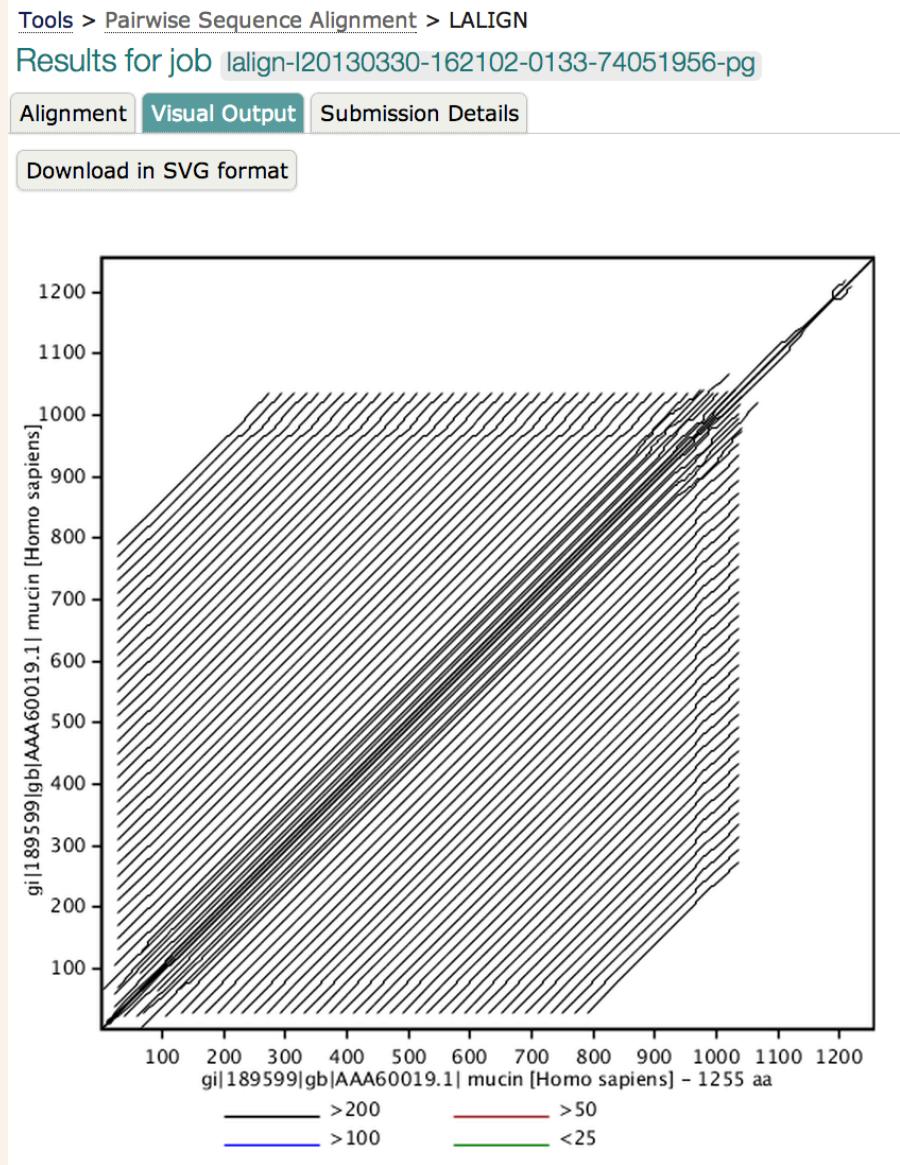
The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

Dotplot for the human mucin compared to itself



What does it tell us about the protein sequence?

This sequence contains repeats!

Dotplots are useful to identify repeats!

Where to find information about repeats?

- NCBI doesn't provide information on repeats.
- Let's see what UniProt provides.
- Since there is no direct link between the two and IDs do not match, the only option to search UniProt by sequence.

The screenshot shows a web browser window with the URL www.ncbi.nlm.nih.gov/protein/189599?report=genpept. The page is titled "mucin [Homo sapiens]" and displays the following information:

Protein [Help](#)

Display Settings: GenPept **Send to:**

mucin [Homo sapiens]

GenBank: AAA60019.1 [FASTA](#) [Graphics](#)

Go to:

LOCUS AAA60019 **DEFINITION** mucin [Homo sapiens]. **ACCESSION** AAA60019 **VERSION** AAA60019.1 GI:189599 **DBSOURCE** locus HUMPANMU accession [J05582.1](#) **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 1255) **AUTHORS** Lan,M.S., Batra,S.K., Qi,W.N., Metzgar,R.S., and Hollingsworth,M.A.

Analyze this sequence

- [Run BLAST](#)
- [Identify Conserved Domains](#)
- [Highlight Sequence Features](#)
- [Find in this Sequence](#)

Protein 3D Structure

Crystal Structure Of The Tumor Specific Antibody Sm3 PDB: 1SM3

Go to <http://web.expasy.org/blast/>
paste the sequence, select blastp, Homo Sapience and search for
curated sequences to speed up the search

The screenshot shows the ExPASy BLAST interface. At the top, there are several tabs: 'ExPASy BLAST for', 'Mucin-1 precursor', 'Exercises', 'Dot plots', 'MAO: a Multiple', 'bips.u-strasbg.fr', 'The Open Biologic', 'Protein BLAST: Ali', 'mucin (Homo sapiens)', and 'LALIGN - Alignme'. Below the tabs, the URL 'web.expasy.org/blast/' is visible.

The main page title is 'BLAST' and it says 'SIB BLAST Network Service'. It states that this is a NCBI BLAST2 service maintained by the Swiss Institute of Bioinformatics. A note says to click on the ? icons for help.

The sequence input area contains a multi-line protein sequence starting with >gi|1895999|gb|AAA60019.1| mucin [Homo sapiens]. The sequence continues with various amino acid residues. To the right of the sequence is a dropdown menu labeled 'Output format: HTML'.

Below the sequence input are two buttons: 'Run BLAST' (highlighted with a red box) and 'Reset Form'.

The next section is titled 'Choose the appropriate BLAST ? program and ? database:' and shows a selected option: 'blastp - query against the UniProt Knowledgebase (Swiss-Prot + TrEMBL)'.

The 'Taxonomic groups' section has a dropdown menu set to 'Homo sapiens' (also highlighted with a red box). There is also a text input field for specifying a taxonomic group and a dropdown for selecting a database subsection.

A large red arrow points to the bottom-left corner of the taxonomic groups section, indicating where to click to proceed.

At the bottom right, there is a note: 'Non-redundant Swiss-Prot+TrEMBL complete proteome sets, see the HAMAP pages.'

There are five hits with E-value 0.0

Welcome to the SIB BLAST Network Service
If results of this search are reported or published, please mention that the computation was performed at the SIB using the BLAST network service. The SIB BLAST network service uses a server developed at SIB and the NCBI BLAST 2 software.
In case of problems, please read the [online BLAST help](#). If your question is not covered, please [contact us](#).

NCBI BLAST program reference [PMID:[9254694](#)]:
Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402(1997).

Query: gi|189599|gb|AAA60019.1|; 1255 AA (of which 5% low-complexity regions filtered out)
Date run: 2013-03-30 18:07:28 UTC+0100 on blast02.vital-it.ch
Program: NCBI BLASTP 2.2.17 [Aug-26-2007]
Database: UniProtKB_Swiss-Prot_HUMAN
38,199 sequences; 22,191,929 total letters
UniProt Knowledgebase Release 2013_03 consists of:
UniProtKB/Swiss-Prot Release 2013_03 of 06-Mar-13: 539616 entries
UniProtKB/TrEMBL Release 2013_03 of 06-Mar-13: 32153798 entries

[Taxonomic view](#) [NiceBlast view](#) [Printable view](#)

List of potentially matching sequences
Send selected sequences to [Clustal W \(multiple alignment\)](#) [Submit](#) [Select up to...](#)

Include query sequence

Db	AC	Description	Score	E-value
<input type="checkbox"/>	sp_P15941	MUC1_HUMAN Mucin-1 precursor (MUC-1) (Breast carcinoma...)	2441	0.0
<input type="checkbox"/>	sp_vs_P15941-3	Isoform 3 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	2437	0.0
<input type="checkbox"/>	sp_vs_P15941-2	Isoform 2 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	2434	0.0
<input type="checkbox"/>	sp_vs_P15941-4	Isoform 4 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	2415	0.0
<input type="checkbox"/>	sp_vs_P15941-5	Isoform 5 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	2132	0.0
<input type="checkbox"/>	sp_vs_P15941-7	Isoform 7 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	363	e-100
<input type="checkbox"/>	sp_vs_P15941-6	Isoform 6 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	275	2e-73
<input type="checkbox"/>	sp_Q02817	MUC2_HUMAN Mucin-2 precursor (MUC-2) (Intestinal mucin...)	264	4e-70
<input type="checkbox"/>	sp_P98088	MUC5A_HUMAN Mucin-5AC precursor (MUC-5AC) (Gastric muc...)	217	6e-56
<input type="checkbox"/>	sp_Q9UKN1	MUC12_HUMAN Mucin-12 precursor (MUC-12) (Mucin-11) (MU...)	209	2e-53
<input type="checkbox"/>	sp_Q685J3	MUC17_HUMAN Mucin-17 precursor (MUC-17) (Small intesti...)	198	3e-50
<input type="checkbox"/>	sp_vs_Q685J3-2	Isoform 2 of Mucin-17 OS=Homo sapiens GN=MUC17 [MUC1...]	198	3e-50
<input type="checkbox"/>	sp_vs_P15941-10	Isoform 10 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1...]	190	8e-48
<input type="checkbox"/>	sp_Q6W4X9	MUC6_HUMAN Mucin-6 precursor (MUC-6) (Gastric mucin-6)...	186	1e-46

We are looking for the protein of 1255 aa, to see the length, change view to “Nice Blast view”

The screenshot shows the SIB BLAST Network Service interface. At the top, there are several tabs: ExPASy BLAST2 In, Mucin-1 precursor, Exercises, Dot plots, MAO: a Multiple, bips.u-strasbg.fr, The Open Biologist, Protein BLAST: All, mucin | Homo sapiens, and LALIGN - Alignme. Below the tabs, the main header reads "ExPASy" and "BLAST". A message at the top says "Welcome to the SIB BLAST Network Service". It includes a note about reporting results and mentions the SIB BLAST network service uses NCBI BLAST 2 software. A link to online help and contact information is provided. Below this, a section about the NCBI BLAST program reference [PMID:9254694] is shown, mentioning Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402(1997). The search query is "gi|189599|gb|AAA60019.1"; 1255 AA (of which 5% low-complexity regions filtered out). The search was run on 2013-03-30 18:07:28 UTC+0100 on blast02.vital-it.ch. The program used was NCBI BLASTP 2.2.17 [Aug-26-2007]. The database used was UniProtKB_Swiss-Prot_HUMAN, containing 38,199 sequences; 22,191,929 total letters. UniProt Knowledgebase Release 2013_03 consists of: UniProtKB/Swiss-Prot Release 2013_03 of 06-Mar-13: 539616 entries, UniProtKB/TrEMBL Release 2013_03 of 06-Mar-13: 32153798 entries. At the bottom of this section, there are three buttons: "Taxonomic view", "NiceBlast view" (which is highlighted with a red box), and "Printable view". Below this, a section titled "List of potentially matching sequences" is shown. It includes a dropdown menu "Send selected sequences to" set to "Clustal W (multiple alignment)", a "Submit" button, and a "Select up to..." button. There is also a checkbox "Include query sequence". A table follows, with columns "Db", "AC", "Description", "Score", and "E-value". The table lists various entries, mostly from the UniProtKB/Swiss-Prot database, such as P15941 (MUC1_HUMAN Mucin-1 precursor) and P02817 (MUC2_HUMAN Mucin-2 precursor).

Db	AC	Description	Score	E-value
sp_vs	P15941	MUC1_HUMAN Mucin-1 precursor (MUC-1) (Breast carcinoma...)	2441	0.0
sp_vs	P15941-3	Isoform 3 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	2437	0.0
sp_vs	P15941-2	Isoform 2 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	2434	0.0
sp_vs	P15941-4	Isoform 4 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	2415	0.0
sp_vs	P15941-5	Isoform 5 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	2132	0.0
sp_vs	P15941-8	Isoform 8 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	405	e-112
sp_vs	P15941-7	Isoform 7 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	363	e-100
sp_vs	P15941-6	Isoform 6 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1] ...	275	2e-73
sp	Q02817	MUC2_HUMAN Mucin-2 precursor (MUC-2) (Intestinal mucin...)	264	4e-70
sp	P98088	MUC5A_HUMAN Mucin-5AC precursor (MUC-5AC) (Gastric muc...)	217	6e-56
sp	Q9UKN1	MUC12_HUMAN Mucin-12 precursor (MUC-12) (Mucin-11) (MU...	209	2e-53
sp	Q685J3	MUC17_HUMAN Mucin-17 precursor (MUC-17) (Small intesti...	198	3e-50
sp_vs	Q685J3-2	Isoform 2 of Mucin-17 OS=Homo sapiens GN=MUC17 [MUC1...	198	3e-50
sp_vs	P15941-10	Isoform 10 of Mucin-1 OS=Homo sapiens GN=MUC1 [MUC1...	190	8e-48
sp	Q6W4X9	MUC6_HUMAN Mucin-6 precursor (MUC-6) (Gastric mucin-6)...	186	1e-46

Our protein has UniProt Accession number P15941. Click on it!

NiceBlast

If results of this search are reported or published, please mention that the computation was performed at the SIB using the BLAST network service. The SIB BLAST network service uses a server developed at SIB and the NCBI BLAST 2 software.

Program: NCBI BLASTP 2.2.17 [Aug-26-2007]
 Databases: UniProtKB 32,727,302 sequences; 10,543,978,207 total letters
 Query: gil189599gblAA60019.1; 1255 Amino acids Date run: 2013-03-30 17:58:04 UTC+0100

Taxonomic view HTML view Printable view

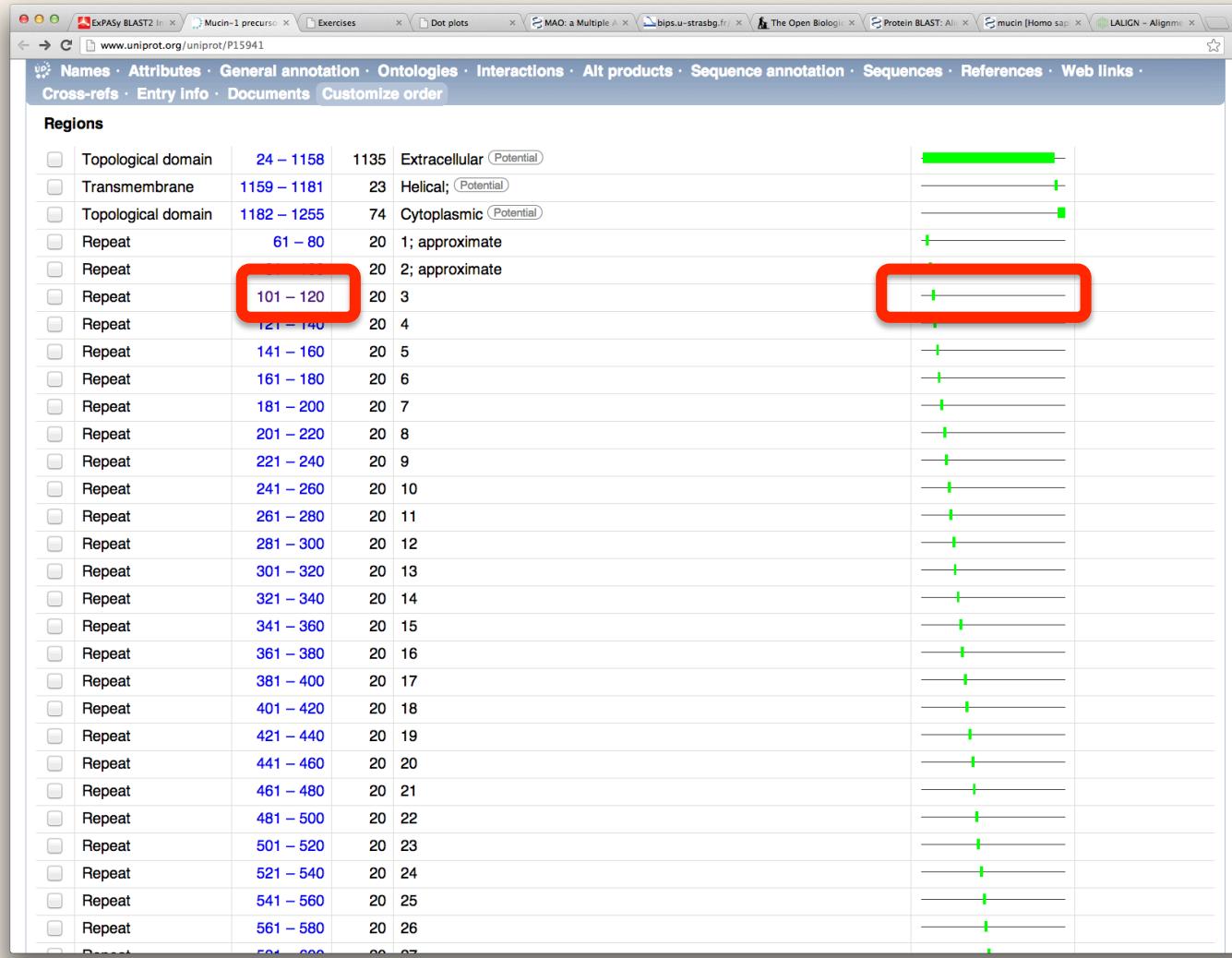
Hit **BLAST** to perform a BLAST search of one of the results with the same parameters

Send selected sequences to

Include query sequence

Score	E-value	Accession number	Entry name	Database	Length
		P15941	BLAST MUC1_HUMAN	sp	1255 Amino acids
2441	0.0		Mucin-1 precursor (MUC-1) (Breast carcinoma-associated antigen DF3) (Cancer antigen 15-3) (CA 15-3) (Carcinoma-associated mucin) (Episialin) (H23AG) (Krebs von den Lungen-6) (KL-6) (PEMT) (Peanut-reactive urinary mucin) (PUM) (Polymorphic epithelial membrane mucin) (PEM) (Tumor-associated epithelial membrane antigen) (EMA) (Tumor-associated mucin) (CD227 antigen) [Contains: Mucin-1 subunit alpha (MUC1-NT) (MUC1-alpha); Mucin-1 subunit beta (MUC1-beta) (MUC1-CT)] [Gene: MUC1 OR PUM] - Homo sapiens (Human) .		
2437	0.0	P15941-3	BLAST	sp_vs	1252 Amino acids
2434	0.0		Isoform 3 of Mucin-1 OS=Homo sapiens Mucin-1 precursor (MUC-1) (Breast carcinoma-associated antigen DF3) (Cancer antigen 15-3) (CA 15-3) (Carcinoma-associated mucin) (Episialin) (H23AG) (Krebs von den Lungen-6) (KL-6) (PEMT) (Peanut-reactive urinary mucin) (PUM) (Polymorphic epithelial mucin) (PEM) (Tumor-associated epithelial membrane antigen) (EMA) (Tumor-associated mucin) (CD227 antigen) [Contains: Mucin-1 subunit alpha (MUC1-NT) (MUC1-alpha); Mucin-1 subunit beta (MUC1-beta) (MUC1-CT)] [Gene: MUC1 OR PUM] - Homo sapiens (Human) .		
2415	0.0	P15941-2	BLAST	sp_vs	1264 Amino acids
2132	0.0		Isoform 2 of Mucin-1 OS=Homo sapiens Mucin-1 precursor (MUC-1) (Breast carcinoma-associated antigen DF3) (Cancer antigen 15-3) (CA 15-3) (Carcinoma-associated mucin) (Episialin) (H23AG) (Krebs von den Lungen-6) (KL-6) (PEMT) (Peanut-reactive urinary mucin) (PUM) (Polymorphic epithelial mucin) (PEM) (Tumor-associated epithelial membrane antigen) (EMA) (Tumor-associated mucin) (CD227 antigen) [Contains: Mucin-1 subunit alpha (MUC1-NT) (MUC1-alpha); Mucin-1 subunit beta (MUC1-beta) (MUC1-CT)] [Gene: MUC1 OR PUM] - Homo sapiens (Human) .		
		P15941-4	BLAST	sp_vs	1243 Amino acids
			Isoform 4 of Mucin-1 OS=Homo sapiens Mucin-1 precursor (MUC-1) (Breast carcinoma-associated antigen DF3) (Cancer antigen 15-3) (CA 15-3) (Carcinoma-associated mucin) (Episialin) (H23AG) (Krebs von den Lungen-6) (KL-6) (PEMT) (Peanut-reactive urinary mucin) (PUM) (Polymorphic epithelial mucin) (PEM) (Tumor-associated epithelial membrane antigen) (EMA) (Tumor-associated mucin) (CD227 antigen) [Contains: Mucin-1 subunit alpha (MUC1-NT) (MUC1-alpha); Mucin-1 subunit beta (MUC1-beta) (MUC1-CT)] [Gene: MUC1 OR PUM] - Homo sapiens (Human) .		
		P15941-5	BLAST	sp_vs	1087 Amino acids
			Isoform 5 of Mucin-1 OS=Homo sapiens Mucin-1 precursor (MUC-1) (Breast carcinoma-associated antigen DF3) (Cancer antigen 15-3) (CA 15-3) (Carcinoma-associated mucin) (Episialin) (H23AG) (Krebs von den Lungen-6) (KL-6) (PEMT) (Peanut-reactive urinary mucin) (PUM) (Polymorphic epithelial mucin) (PEM) (Tumor-associated epithelial membrane antigen) (EMA) (Tumor-associated mucin) (CD227 antigen) [Contains: Mucin-1 subunit alpha (MUC1-NT) (MUC1-alpha); Mucin-1 subunit beta (MUC1-beta) (MUC1-CT)] [Gene: MUC1 OR PUM] - Homo sapiens (Human) .		

Clicking on any repeat or region highlights it on the sequence



ExPASy BLAST2 In X P15941[101-120] Exercises Dot plots MAO: a Multiple A X bips.u-strasbg.fr/ The Open Biologi X Protein BLAST: All mucin [Homo sapiens] LALIGN - Alignme X

www.uniprot.org/blast/?about=P15941[101-120]

UniProt Jobs Downloads · Contact · Documentation/Help

Search Blast * Align Retrieve ID Mapping

Sequence or UniProt identifier

>sp|P15941|101-120
PVTRPALGSTTPAHDVTS

Blast Clear « Options

Help
For a sequence similarity search, enter:
 • a protein or nucleotide sequence
 • a UniProt identifier, e.g.
 P00750 or A4_HUMAN or UPI0000000001
[More...](#)

Database Threshold Matrix Filtering Gapped Hits

UniProtKB 10 Auto None yes 250

P15941[101-120], Mucin-1, Homo sapiens

10	20	30	40	50	60
MTPGTQSPFF LLLLLTVLTV VTGSGHASST PGGEKETSAT QRSSVPSSSTE KNAVSMTSSV					
70	80	90	100	110	120
LSHSHPGS S TTQGQDVTL APATEPASGS AATWGQDVTS PVTRPALGS TPPAHDVTS					
130	140	150	160	170	180
APDNKPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS					
190	200	210	220	230	240
APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS					
250	260	270	280	290	300
APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS					
310	320	330	340	350	360
APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS					
370	380	390	400	410	420
APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS					
430	440	450	460	470	480
APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS					
490	500	510	520	530	540
APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS					
550	560	570	580	590	600
APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS					
610	620	630	640	650	660
APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS APDTRPAPGS T APPAHGVTS					

Scrolling down shows summary on repeats. Any region can be sent to blast by selecting it! If more than one region selected they can be aligned!

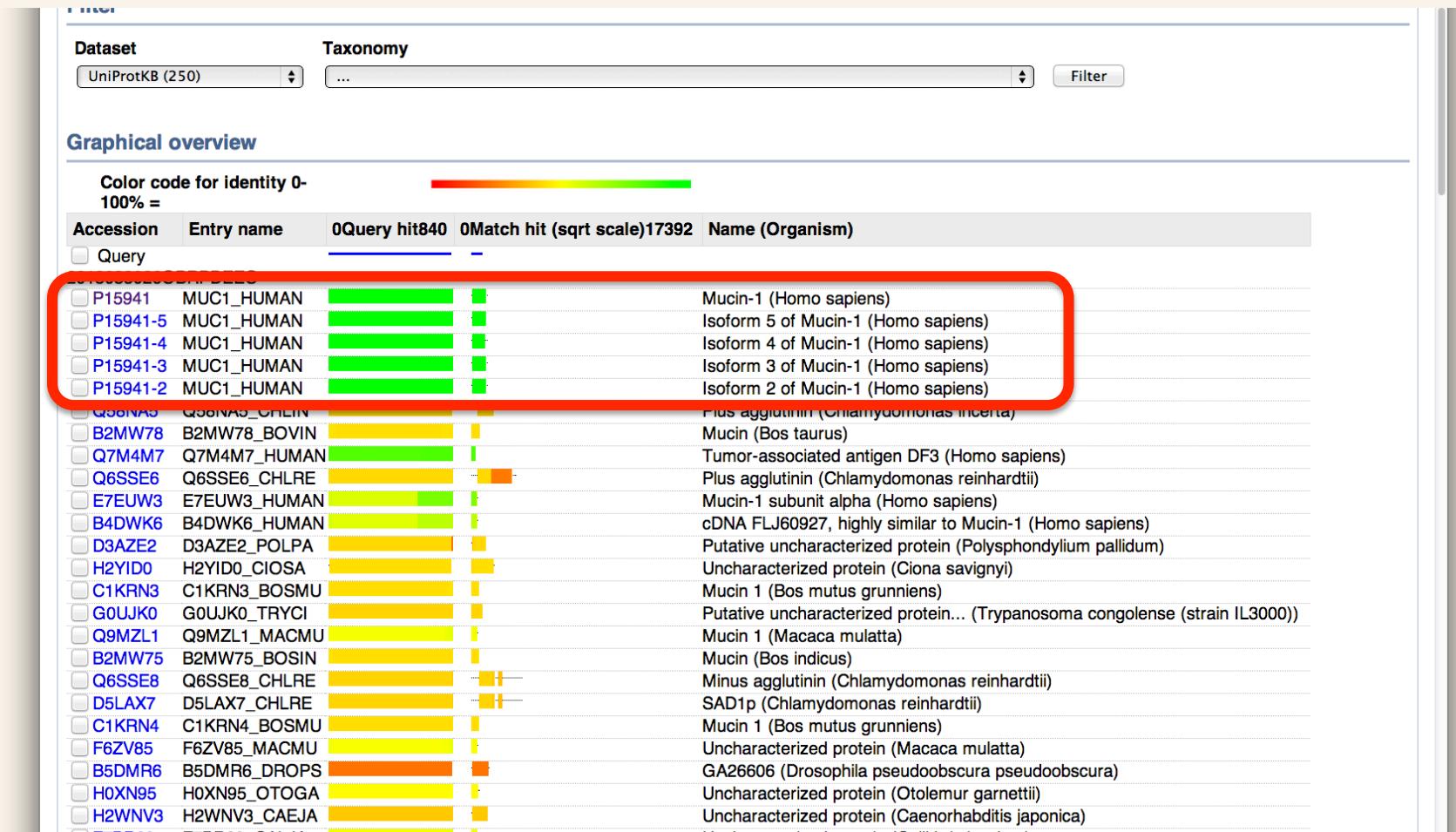
The screenshot shows a web browser window displaying the Uniprot protein details page for P15941. The page has a header with various tabs like Names, Attributes, General annotation, Ontologies, Interactions, Alt products, Sequence annotation, Sequences, References, and Web links. Below the header is a navigation bar with links for Cross-refs, Entry info, Documents, and Customize order.

The main content is a table listing protein features. The columns are labeled 'Repeat' (checkbox), 'Range' (e.g., 601 – 620, 20 28), and 'Count' (e.g., 20, 29). To the right of each row is a horizontal bar with green segments indicating the location of each repeat within the protein sequence. A red box highlights the first row, which corresponds to the 'Region' 126 – 965, described as having 42 X 20 AA approximate tandem repeats of the sequence P-A-P-G-S-T-A-P-P-A-H-G-V-T-S-A-P-D-T-R.

Repeat	Range	Count	Description
<input type="checkbox"/>	Repeat 601 – 620	20 28	
<input type="checkbox"/>	Repeat 641 – 660	20 30	
<input type="checkbox"/>	Repeat 661 – 680	20 31	
<input type="checkbox"/>	Repeat 681 – 700	20 32	
<input type="checkbox"/>	Repeat 701 – 720	20 33	
<input type="checkbox"/>	Repeat 721 – 740	20 34	
<input type="checkbox"/>	Repeat 741 – 760	20 35	
<input type="checkbox"/>	Repeat 761 – 780	20 36	
<input type="checkbox"/>	Repeat 781 – 800	20 37	
<input type="checkbox"/>	Repeat 801 – 820	20 38	
<input type="checkbox"/>	Repeat 821 – 840	20 39	
<input type="checkbox"/>	Repeat 841 – 860	20 40	
<input type="checkbox"/>	Repeat 861 – 880	20 41	
<input type="checkbox"/>	Repeat 881 – 900	20 42	
<input type="checkbox"/>	Repeat 901 – 920	20 43	
<input type="checkbox"/>	Repeat 921 – 940	20 44	
<input type="checkbox"/>	Repeat 941 – 960	20 45	
<input type="checkbox"/>	Repeat 961 – 980	20 46; approximate	
<input type="checkbox"/>	Repeat 981 – 1000	20 47; approximate	
<input type="checkbox"/>	Repeat 1001 – 1020	20 48; approximate	
<input type="checkbox"/>	Domain 1021 – 1151	118 SFA	
<input checked="" type="checkbox"/>	Region 126 – 965	840	42 X 20 AA approximate tandem repeats of P-A-P-G-S-T-A-P-P-A-H-G-V-T-S-A-P-D-T-R
<input type="checkbox"/>	Region 1223 – 1230	8	Required for interaction with GSK3B
<input type="checkbox"/>	Region 1233 – 1241	9	Required for interaction with beta- and gamma-catenins
<input type="checkbox"/>	Motif 1203 – 1206	4	Interaction with GRB2
<input type="checkbox"/>	Motif 1229 – 1232	4	Interaction with SRC and ESR1
<input type="checkbox"/>	Motif 1243 – 1246	4	Required for interaction with AP1S2

At the bottom left, a green bar indicates '1 selected: P15941[126-965]'. At the bottom right are buttons for Retrieve, Align, Blast, and Clear.

For example, from the blast results it can be seen that this region of 42 tandem repeats is present in isoforms 2, 3, 4, and 5 of human mucin, but not in other reported in the UniProt entry isoforms 6, 7, 8, 9, 10, without looking at the description of each isoform



Let's generate a dotplot, using ALIGN
for P69193 comparing it against itself

uniprot.org P69193

Screenshot of the UniProt website search results for P69193.

The search query "P69193" is highlighted in a red box in the search bar.

WELCOME
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none">★ Swiss-Prot, which is manually annotated and reviewed.★ TrEMBL, which is automatically annotated and is not reviewed. Includes complete and reference proteome sets .
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords, subcellular locations, cross-referenced databases and more .

Getting started

- [Text search](#)
- [Sequence similarity searches \(BLAST\)](#)
- [Sequence alignments](#)
- [Batch retrieval](#)
- [Database identifier mapping \(ID Mapping\)](#)

NEWS

UniProt release 2013_03 - Mar 6, 2013
Latest from the prokaryotic world: bacterial Cas9, a new tool for genome engineering | Cross-references to ChiTARS and SABIO-RK | Removal of cross-references to 8 2D gel databases and AGD

› Statistics for UniProtKB:
[Swiss-Prot](#) · [TrEMBL](#)
› Forthcoming changes
› News archives
[Follow @uniprot](#) 552 followers

SITE TOUR



Learn how to make best use of the tools and data on this site.

PROTEIN SPOTLIGHT
the silence within
March 2013

1 selected: P15941[126-965] ↗

Downloads · Contact · Documentation/Help

Search Blast Align Retrieve ID Mapping

Protein Knowledgebase (UniProtKB) Query P69193 Search Advanced Search Clear

Retrieves Align Blast Clear

Get FASTA sequence, copy it and paste in ALIGN

P69193 [UniParc]. **FASTA** 989 111,095 Blast go

Last modified February 15, 2005. Version 1.
Checksum: E732960770C15F52

```
10      20      30      40      50      60
MKSYISLFFI LCVIFKNKVI KCTGESQTGN TGGGQAGNTV GDQAGSTGGS PQGSTGASQP

70      80      90      100     110     120
GSSEPSNPVS SGHSVSTVSV SQTSTSSEKQ DTIQVKSALL KDYMLGLKVTG PCNENFIMFL

130     140     150     160     170     180
VPHIYIDVDT EDTNIELRTT LKETNNNAISF ESNSGSLEKK KYVKLPSNGT TGEQGSSTGT

190     200     210     220     230     240
VRGDTEPISD SSSSSSSSSS SSSSSSSSSS SSSSSSSSSS SSSSESLSLP NGPDSPPTVKP

250     260     270     280     290     300
PRNLQNICET GKNFKLVVYI KENTLIIKWK VYGETKDTTE NNKVDVRKYL INEKETPFTS

310     320     330     340     350     360
ILIHAYKEHN GTNLIESKNY ALGSDIPEKC DTLASNCFLS GNFNIEKCFQ CALLVEKENK

370     380     390     400     410     420
NDVCYKYLSE DIVSNFKEIK AETEDDDDED YTEYKLTESI DNILVKMFKT NENNDKSELI

430     440     450     460     470     480
KLEEVDDSLK LELMNYCSLL KDVDTTGTLN NYGMGNEMDI FNNLKRLLIY HSEENINTLK

490     500     510     520     530     540
NKFRNAAVCL KNVDDWIVNK RGLVLPELNY DLEYFNEHLY NDKNSPEDKD NKGKGVVHVD

550     560     570     580     590     600
TTLERKEDTLS YDNSDNMFCN KEYCNRLKDE NNCISNLQVE DQGNCDTSWI FASKYHLETI

610     620     630     640     650     660
RCMKGYEPTK ISALYVANCY KGEHKDRDCDE GSSPMEFQLQI IEDYGFPLAE SNYPYNVVKV

670     680     690     700     710     720
GEQCPKVEDH WMNLWDNGKI LHNKNEPNSL DGKGYTAYES ERFHDNMADF VKIIKTEVMN

730     740     750     760     770     780
KGSVIAYIKA ENVMGYEFSG KKVNQNLCGDD TADHAVNIVG YGNVYVNSEGE KKSYWIVRNS
```

www.uniprot Mucin-1 pre... Exercises Dot plots W BLAST - Wiki http://myhits Low-complex bips.u-strasb The Open Bio NCBI Blast:sp human sox-1 LALIGN | Pair

www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=lalign&context=protein

EMBL-EBI Services Research Training Industry About us

LALIGN

Protein alignment Nucleotide alignment Web services Help & Documentation Share Feedback

Tools > Pairwise Sequence Alignment > LALIGN

Pairwise Sequence Alignment

LALIGN finds internal duplications by calculating non-intersecting local alignments of **protein** or **nucleotide** sequences.

STEP 1 - Enter your protein sequences

Enter or paste your first **protein** sequence in any supported format:

```
RCMKGYEPTKISALYVANCYKGEHKDRCDEGSSPMELQIETYGLPAESNPYPNYVKV  
GEQCPKVEDHWMNLWDNGKILHNKNEPNSLDGKGTYATESERFHNDNMDAFKIIKTEVMN  
KGSVIAVYKAENVMGYEFSGKVCNLCGDDTADHAVNVGYNVNSEGEKKSYWIRVNS  
WGPYWGDEGYFKVDMYGPTHCHNFVHSVVFNVDLMNNKTTKESKIVDYYLKASPEF  
YHNLYFKNPNVGKKNLFSEKEDNENNKKLGNNYIIFGQDTAGSGQSGKESNTALESAGTS  
NEVSERVHILKHDKGKIRMGRKYIDTQDVNKHSCTRSYAFNPNEYKCVNLCNV  
NWKTCEEKTSPLGLCLSKLDTNNECYFCYV
```

Or, upload a file: No file chosen

AND

Enter or paste your second **protein** sequence in any supported format:

```
>sp|P69193|SERA_PLA6D Serine-repeat antigen protein OS=Plasmodium falciparum (isolate CDC / Honduras) GN=SERA PE=1 SV=1  
MKSYSISLFFILCVIFNKNVIKCTGESQTGNTGGQQAGNTVGDQAGSTGGSPQGSTGASQP  
GSSEPSNPVSSGHSVSTVSQSQTSTSSEKQDTIQVKSALLKDYMGLKVTGPCNENFIMFL  
VPHIYDVDTEDTNIELRTTLKETNNNAIFESNSGSLEKKKYVLPSNGTTGEQQSSTGT  
VRGDTEPISDSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSPANGPDSPTVKP  
PRNLQNICKETGKFNFKLVVYIKENTLIWKVYGETKDTTENKVDVRKYLINEKETPFTS  
ILIHAYKEHNGTNLIESKNYALGSDIPEKCCTLASNCFLSGNFQIEKCFQCALLVEKENK
```

Or, upload a file: No file chosen

STEP 2 - Set your pairwise alignment options

The default settings will fulfill the needs of most users and, for that reason, are not visible.

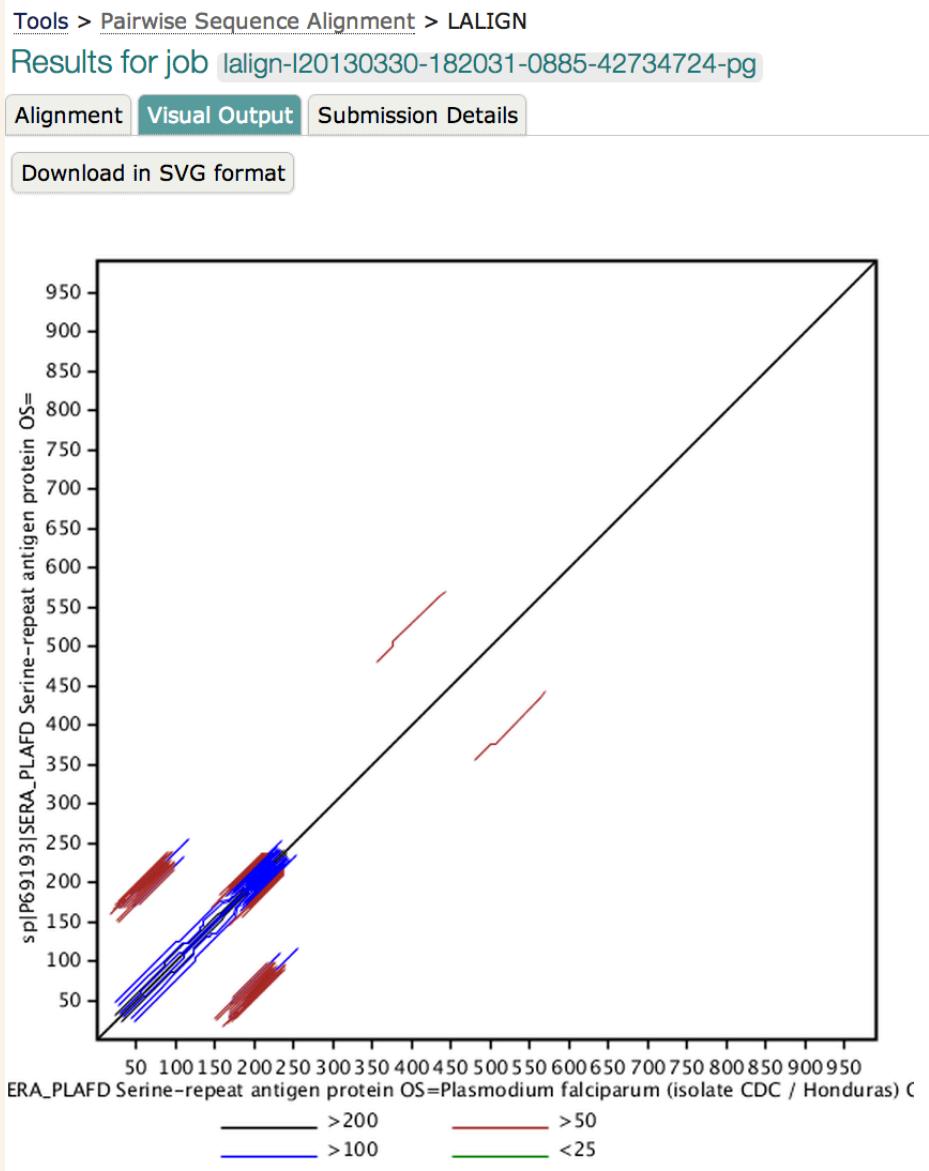
(Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

If you plan to use these services during a course please [contact us](#).

Dotplot for P69193 compared to itself



What does it tell us about the protein sequence?

This sequence contains a low-complexity region around position 200!

Dotplots are useful to identify low-complexity regions!

- "**Low-complexity region**" means a region of compositional bias; that is a sequence composed of few kinds of elements:
 - homopolymeric runs,
 - short-period repeats,
 - overrepresentation of several residues

Names and origin

P69193

Protein names

Recommended name:
Serine-repeat antigen protein

Alternative name(s):
111 kDa antigen
p126

Gene names

Name:**SERA**

Organism

Plasmodium falciparum (isolate CDC / Honduras)

Regions

<input type="checkbox"/>	Region	571 – 989	419	Thiol-protease-like	
<input type="checkbox"/>	Compositional bias	27 – 227	201	Ser-rich	
<input type="checkbox"/>	Compositional bias	191 – 225	35	Poly-Ser	

10	20	30	40	50	60
MKSYISLFFI	LCVIFNKNVI	KCTGESQTGN	TGGGQAGNTV	GDQAGSTGGS	PQGSTGASQP
70	80	90	100	110	120
GSSEPSNPVS	SGHSVSTVSV	SQTSTSSEKQ	DTIQVKKSALL	KDYMGLKVTG	PCNENFIMFL
130	140	150	160	170	180
VPHIYIDVDT	EDTNIELRTT	LKETNNNAISF	ESNSGSLEKK	KYVKLPSNGT	TGEQGSSTGT
190	200	210	220	230	240
VRGDTEPISD	SSSSSSSSSS	SSSSSSSSSS	SSSSSSSSSS	SSSSSESLPA	NGPDSPTVKP
250	260	270	280	290	300
PRNLQNICET	GKNFKLVVYI	KENTLIIKWK	VYGETKDTTE	NNKVDVRKYL	INEKETPFTS

P50553

Names and origin

Protein names	<i>Recommended name:</i> Achaete-scute homolog 1 Short name=ASH-1 Short name=hASH1 <i>Alternative name(s):</i> Class A basic helix-loop-helix protein 46 Short name=bHLHa46
Gene names	Name: ASCL1 Synonyms:ASH1, BHLHA46, HASH1
Organism	Homo sapiens (Human) [Reference proteome]

Regions

<input type="checkbox"/> Domain	118 – 170	53	bHLH	
<input type="checkbox"/> Compositional bias	33 – 47	15	Poly-Ala	
<input type="checkbox"/> Compositional bias	51 – 62	12	Poly-Gln	

10 20 30 40 50 60
MESSAKMESG GAGQQPQPQP QQPFLPPAAC FIAATAAAAAA AAAAAAAQSA QQQQQQQQQQ

70 80 90 100 110 120
QQAPQLRPAA DGQPSGGGHK SAPKQVKRQR SSSPELMRCK RRLNFSGFGY SLPQQQPAAV

130 140 150 160 170 180
ARRNERERNR VKLVNLGFAT LREHVPNGAA NKKMSKVETL RSAVEYIRAL QQLLDEHDADV

190 200 210 220 230
SAAFQAGVLS PTISPNTYSND LNSMAGSPVVS SYSSDEGSYD PLSPEEQELL DFTNWF

Let's try to align two nucleotide sequences
with LALIGN which implements a modified
Smith-Waterman algorithm of 1981

- ✓ Scoring of the alignment
- ✓ Match/mismatch
- ✓ Penalty on a gap opening
- ✓ Penalty on a gap extension
- ✓ E-value

LALIGN at EBI

www.ebi.ac.uk/Tools/psa/lalign/nucleotide.html

The screenshot shows the LALIGN web interface. At the top, there's a navigation bar with links for Services, Research, Training, Industry, and About. Below that is a main header with the LALIGN logo and a sub-header with links for Protein alignment, Nucleotide alignment (which is selected), Web services, and Help & Documentation. A breadcrumb trail indicates the user is in Tools > Pairwise Sequence Alignment > LALIGN. The main content area is titled "Pairwise Sequence Alignment" and explains that LALIGN finds internal duplications by calculating non-intersecting local alignments of nucleotide or protein sequences. The interface is divided into three main sections: STEP 1 - Enter your nucleotide sequences, STEP 2 - Set your pairwise alignment options, and STEP 3 - Submit your job. In the first section, there are two input fields. The top field contains the sequence "acctgagagg" and has a red box drawn around it. The bottom field contains the sequence "acgtggcagg" and also has a red box drawn around it. Both fields have placeholder text "Enter or paste your first nucleotide sequence in any supported format:" above them. Below each field is a "Choose File" button and a message indicating "no file selected". The second section, "STEP 2 - Set your pairwise alignment options", contains a note that default settings will fulfill most users' needs and a "More options..." link. The third section, "STEP 3 - Submit your job", contains a checkbox labeled "Be notified by email" with the sub-instruction "(Tick this box if you want to be notified by email when the results are available)" and a "Submit" button.

EMBL-EBI

Services Research Training Industry About

LALIGN

Protein alignment Nucleotide alignment Web services Help & Documentation

Tools > Pairwise Sequence Alignment > LALIGN

Pairwise Sequence Alignment

LALIGN finds internal duplications by calculating non-intersecting local alignments of **nucleotide** or **protein** sequences.

STEP 1 - Enter your nucleotide sequences

Enter or paste your first nucleotide sequence in any supported format:

acctgagagg

Or, upload a file: Choose File no file selected

AND

Enter or paste your second nucleotide sequence in any supported format:

acgtggcagg

Or, upload a file: Choose File no file selected

STEP 2 - Set your pairwise alignment options

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... (Click here, if you want to view or change the default settings.)

STEP 3 - Submit your job

Be notified by email (Tick this box if you want to be notified by email when the results are available)

Submit

```
>>EMBOSS_001 (10 nt)
Waterman-Eggert score: 23; 9.3 bits; E(1) < 0.15
70.0% identity (70.0% similar) in 10 nt overlap (1-10:1-10)

      10
EMBOSS ACCTGAGAGG
      :: :: :::
EMBOSS ACGTGGCAGG
      10
```

Let's look at parameters

Pairwise Sequence Alignment

LALIGN finds internal duplications by calculating non-intersecting local alignments of **nucleotide** or **protein** sequences.

STEP 1 - Enter your nucleotide sequences

Enter or paste your first **nucleotide** sequence in any supported format:

```
acctgagagg
```

Or, upload a file: no file selected

AND

Enter or paste your second **nucleotide** sequence in any supported format:

```
acgtggcagg
```

Or, upload a file: no file selected

STEP 2 - Set your pairwise alignment options

MATRIX	GAP OPEN	GAP EXTEND	E() THRESHOLD	OUTPUT FORMAT	GRAPHICS
+5/-4	-12	-4	10.0	MARKX 0	yes

STEP 3 - Submit your job

Be notified by email (*Tick this box if you want to be notified by email when the results are available*)

Default parameters: score 5 for match, -4 for mismatch,
-12 for gap opening, -4 for a gap extension

```
>>EMBOSS_001                                     (10 nt)
Waterman-Eggert score: 23;  9.3 bits; E(1) <  0.15
70.0% identity (70.0% similar) in 10 nt overlap (1-10:1-10)

          10
EMBOSS ACCTGAGAGG
      :: :: :::
EMBOSS ACGTGGCAGG
          10
```

Score = $5 * \{\text{number of matches}\} - 4 * \{\text{number of mismatches}\} - 12 * \{\text{number of gap openings}\} - 4 * \{\text{number of positions gaps extended}\}$ →

Score is 23 = $5 * 7 - 4 * 3 - 12 * 0 - 4 * 0$

Let's “not punish” the appearance of a gap

Pairwise Sequence Alignment

LALIGN finds internal duplications by calculating non-intersecting local alignments of **nucleotide** or **protein** sequences.

STEP 1 - Enter your nucleotide sequences

Enter or paste your first **nucleotide** sequence in any supported format:
acctgagagg

Or, upload a file: no file selected

AND

Enter or paste your second **nucleotide** sequence in any supported format:
acgtggcagg

Or, upload a file: no file selected

STEP 2 - Set your pairwise alignment options

MATRIX +5/-4	GAP OPEN <input type="text" value="0"/>	GAP EXTEND -4	E() THRESHOLD 10.0	OUTPUT FORMAT MARKX 0	GRAPHICS yes
-----------------	--	------------------	-----------------------	--------------------------	-----------------

STEP 3 - Submit your job

Be notified by email (*Tick this box if you want to be notified by email when the results are available*)

```
>>EMBOSS_001                                     (10 nt)
      Waterman-Eggert score: 28; 10.2 bits; E(1) < 0.08
      72.7% identity (72.7% similar) in 11 nt overlap (1-10:1-10)

          10
EMBOSS ACCTGAG-AGG
      :: :: : :::
EMBOSS ACGTG-GCAGG
          10
```

Score = $5 * \{\text{number of matches}\} - 4 * \{\text{number of mismatches}\} - 12 * \{\text{number of gap openings}\} - 4 * \{\text{number of positions gaps extended}\}$ →

Score is 28 = $5 * 8 - 4 * 3 - 0 * \underline{2} - 4 * 0$

Which alignment is “better”?

```
>>EMBOSS_001 (10 nt)
Waterman-Eggert score: 23; 9.3 bits; E(1) < 0.15
70.0% identity (70.0% similar) in 10 nt overlap (1-10:1-10)

      10
EMBOSS ACCTGAGAGG
      :: :: :::
EMBOSS ACGTGGCAGG
      10
```

```
>>EMBOSS_001 (10 nt)
Waterman-Eggert score: 28; 10.2 bits; E(1) < 0.08
72.7% identity (72.7% similar) in 11 nt overlap (1-10:1-10)

      10
EMBOSS ACCTGAG-AGG
      :: :: : :::
EMBOSS ACGTG-GCAGG
      10
```

The second one: it has higher score and lower E-value, although it was obtained with changed from default parameters. While both E-values are not much significant as they above 1e-6.

What is E-value?

- The E-value of the alignment is the **expected number of sequences** that give the same **Z-score** or better if the database is probed with a random sequence.
- **Z-score** of the score S of the alignment = $\{(S - \text{mean})/SD\}$.
- **The E-value is not the probability!** If P is the probability to obtain by chance an alignment with the same Z-score or better, then P and E (E-value) are related as $P = \{1 - e^{\wedge}(-E)\}$.
- **The E-value depends on the size of the database** and it ranges from 0 to the number of sequences in the databases.

Interpretation of E-value

- In LALIGN, E-values have been calculated only with the default parameters on a set of 500 randomly selected sequences. Therefore it is much less meaningful than the E-value reported by BLAST search of a database.
- A rough interpretation of E-value:
 - ✓ $E > 1$ – This match can be expected by chance
 - ✓ $1 > E > 0.001$ - Homology cannot be ruled out
 - ✓ $E < 0.001$ ($1e-3$) – Probable homology, while different researches reported different E-values for different algorithms and databases to be significant ranging from $1e-3$ to $1e-6$
- The E-value is not the only measure of the alignment significance. For proteins (as a lot of research has been done): If two proteins share 25% and above of sequence identity at 150 aa or 40% at 70 aa intervals they might be homologous.

Local vs Global alignment

- Global alignment compares two sequences along the entire length.
- Local alignment searches for the most similar regions in the two sequences.

Let's try longer sequences in LALIGN

The screenshot shows the LALIGN web application interface. At the top, there is a dark teal header with the word "LALIGN" in large white letters. Below the header is a navigation bar with four tabs: "Protein alignment", "Nucleotide alignment" (which is highlighted in black), "Web services", and "Help & Doc". Underneath the navigation bar, the path "Tools > Pairwise Sequence Alignment > LALIGN" is displayed. The main title "Pairwise Sequence Alignment" is in a large, light blue font. A descriptive subtitle below it states "LALIGN finds internal duplications by calculating non-intersecting local alignments". The first input section is titled "STEP 1 - Enter your nucleotide sequences". It contains a text area with placeholder text "Enter or paste your first nucleotide sequence in any supported format:" followed by a sample sequence "tccCAGTTATGTCAGgggacacgagcatgcagagac". Below this, there is a file upload field with the text "Or, upload a file: Choose File no file selected". The second input section is titled "AND". It contains a text area with placeholder text "Enter or paste your second nucleotide sequence in any supported format:" followed by a sample sequence "aattgccgcgcgtcgtttcagCAGTTATGTCAGatc".

LALIGN result

```
>>EMBOSS_001                                     (36 nt)
Waterman-Eggert score: 60;  21.0 bits; E(1) <  0.00063
100.0% identity (100.0% similar) in 12 nt overlap (4-15:22-33)

          10
EMBOSS CAGTTATGTCAG
      ::::::::::::::
EMBOSS CAGTTATGTCAG
          30
```

Go back to the input page of LALIGN and keep it open.

NCBI BLAST <http://blast.ncbi.nlm.nih.gov/Blast.cgi>

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

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](#).

<input type="checkbox"/> Human	<input type="checkbox"/> Oryza sativa	<input type="checkbox"/> Gallus gallus
<input type="checkbox"/> Mouse	<input type="checkbox"/> Bos taurus	<input type="checkbox"/> Pan troglodytes
<input type="checkbox"/> Rat	<input type="checkbox"/> Danio rerio	<input type="checkbox"/> Microbes
<input type="checkbox"/> Arabidopsis thaliana	<input type="checkbox"/> Drosophila melanogaster	<input type="checkbox"/> Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontiguous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> 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)
- Search [immunoglobulins](#) (IgBLAST)
- Search using [SNP flanks](#)
- Screen sequence for [vector contamination](#) (vecscren)
- Align two (or more) sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search SRA [transcript and genomic libraries](#)
- Constraint Based Protein [Multiple Alignment Tool](#)
- Needleman-Wunsch [Global Sequence Alignment Tool](#)
- Search [RefSeqGene](#)



Take a break! Walk! Do...
33 minutes ago Registered

News

Improved BLAST statistics described in BMC Research Notes.

BLAST calculates expect values that describe the significance of a match, with a lower expect value indicating a more significant match.

Fri, 11 Jan 2013 10:00:00 EST

[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.

[More tips...](#)

Needlman-Wunsch global alignment for nucleotide sequences: Let's run it for our two sequences

Global Alignment

Home Recent Results Saved Strategies My NCBI [Sign In] [Register]

NCBI/ BLAST/ Global Alignment Needleman-Wunsch Global Align Nucleotide Sequences

Nucleotide Protein

Enter Query Sequence

Needleman-Wunsch alignment of two nucleotide sequences [?](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) Clear

tccCAGTTATGTCAGggacacgagcatgcagagac

Query subrange [?](#)

From To

Or, upload file [Choose File](#) No file chosen [?](#)

Job Title
Enter a descriptive title for your BLAST search [?](#)

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence [?](#) Clear

aattgccggcgtctttcagCAGTTATGTCAGatc

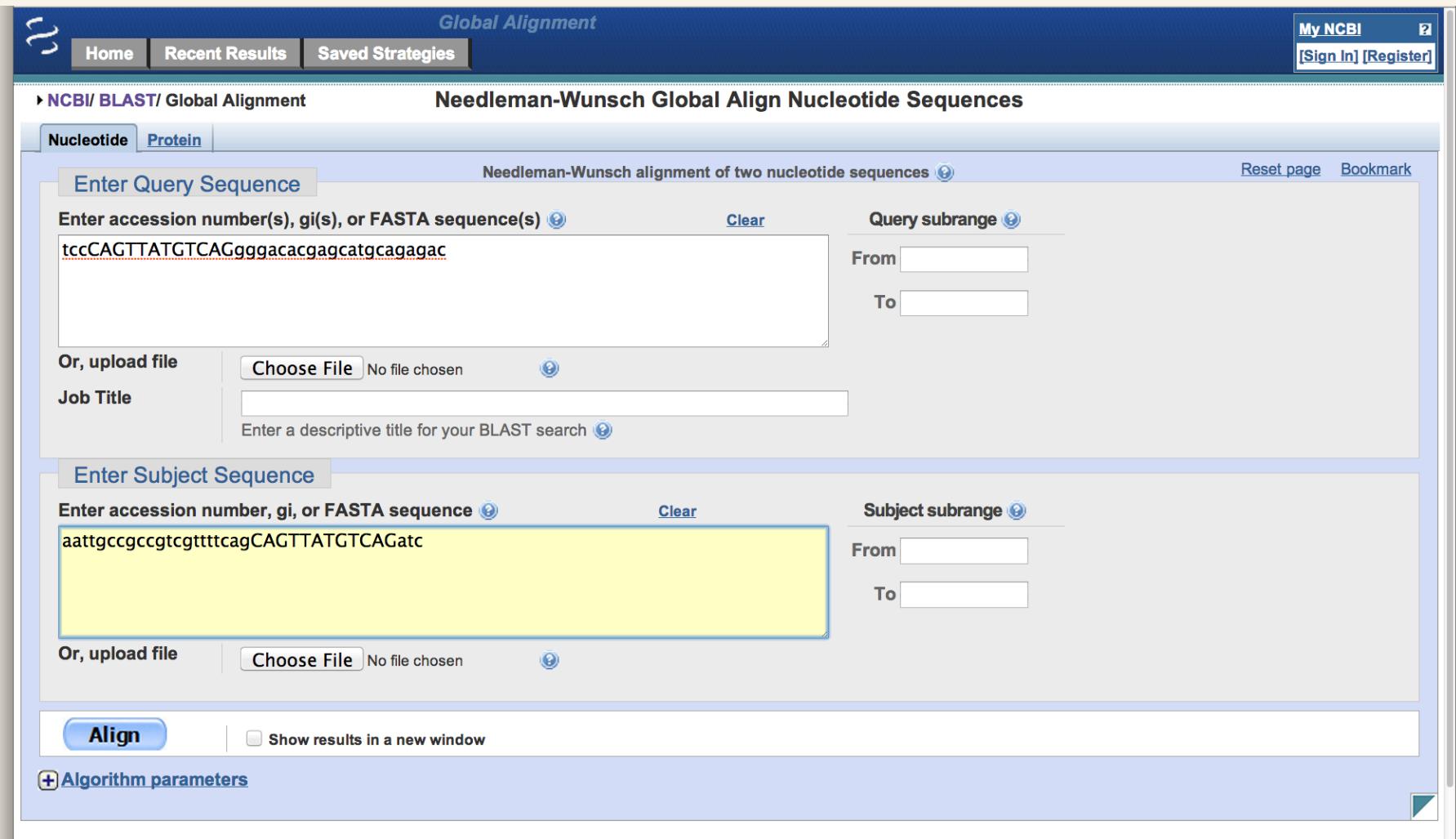
Subject subrange [?](#)

From To

Or, upload file [Choose File](#) No file chosen [?](#)

Align Show results in a new window

[+] Algorithm parameters



Needlman-Wunsch global alignment

NW Score	Identities	Gaps
-47	19/40(48%)	6/40(15%)
Query 1	TCCCAGTTATGTCAGGGGACACGAGC-ATG-CAGAGAC	36
Sbjct 1	AATTGCCGCC--GTC-GTTTCAGCAGTTATGTCAGAT-C	36

Local vs Global alignment

- **Global alignment** compares two sequences along the entire length. It is therefore **best for highly similar sequences of approximately the same length**.

```
TCCCAGTTATGTCAGGGGACACGAGC-ATG-CAGAGAC
| | | | | | | | | | | | | | | | | | | | | |
AATTGCCGCC--GTC-GTTTCAGCAGTTATGTCAGAT-C
```

- **Local alignment** searches for the most similar regions in the two sequences. It can produce more than one alignment. It is best for sequences that share some degree of similarity or of different lengths. **Best for finding conserved elements**.

```
tccCAGTTATGTCAGggacacgagcatgcagagac
||||||| |||||  
aattgccgcgtcgtttcagCAGTTATGTCAGatc
```

BLAST

- BLAST algorithms belongs to the class of word or k-tulip methods for local alignment (Altschuls et al., 1991). This parameter, the word size is adjustable.
- It is used to search NCBI databases and UniProt: NCBI BLAST server, Entrez, ExPASy, UniProt web site.

Let's use Bl2seq, the implementation of BLAST for the pairwise alignment, to see the result of masking low-complexity regions and to introduce scoring matrices: PAM, BLOSUM

NCBI BLAST <http://blast.ncbi.nlm.nih.gov/Blast.cgi>

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

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](#).

<input type="checkbox"/> Human	<input type="checkbox"/> Oryza sativa	<input type="checkbox"/> Gallus gallus
<input type="checkbox"/> Mouse	<input type="checkbox"/> Bos taurus	<input type="checkbox"/> Pan troglodytes
<input type="checkbox"/> Rat	<input type="checkbox"/> Danio rerio	<input type="checkbox"/> Microbes
<input type="checkbox"/> Arabidopsis thaliana	<input type="checkbox"/> Drosophila melanogaster	<input type="checkbox"/> Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms:</i> blastn, megablast, discontiguous megablast
protein blast	Search protein database using a protein query <i>Algorithms:</i> 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)
- Search [immunoglobulins](#) (IgBLAST)
- Search using [SNP flanks](#)
- Screen sequence for [vector contamination](#) (vecscren)
- Align two (or more) sequences using BLAST (bl2seq) 
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search SRA [transcript and genomic libraries](#)
- Constraint Based Protein [Multiple Alignment Tool](#)
- Needleman-Wunsch [Global Sequence Alignment Tool](#)
- Search [RefSeqGene](#)

Take a break! Walk! Do...
33 minutes ago [\[Logout\]](#)

News

Improved BLAST statistics described in [BMC Research Notes](#).

BLAST calculates expect values that describe the significance of a match, with a lower expect value indicating a more significant match.

Fri, 11 Jan 2013 10:00:00 EST

[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.

[More tips...](#)

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite Align Sequences Protein BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

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

Query subrange From To

Or, upload file Choose File No file chosen

Job Title Enter a descriptive title for your BLAST search

Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence Clear

Subject subrange From To

Or, upload file Choose File No file chosen

Program Selection

Algorithm blastp (protein-protein BLAST)
Choose a BLAST algorithm

BLAST Search protein sequence using Blastp (protein-protein BLAST)
 Show results in a new window

Search for “human major prion protein precursor” in protein NCBI databases

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

Protein human major prion protein precursor Search

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

Genotypes and Phenotypes

Data from Genome Wide Association studies that link genes and diseases. See study variables, protocols, and analysis.

II 1 2 3 4 5 6 7 8

Popular Resources

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

NCBI Announcements

Now Available: NCBI Insights Blog! 28 Jan 2013

NCBI has just released a new blog called *NCBI Insights*. Blog posts will provide an insider's perspective to help

Come to the NCBI Discovery Workshops on February 4&5! 16 Jan 2013

Spaces are still available for the free, 2-day Discovery Workshops to be held on

New version of Genome Workbench available 06 Sep 2012

An integrated, downloadable application for viewing and analyzing sequence

Get FASTA sequence of human and mouse proteins and paste them into bl2seq

BLAST® Basic Local Alignment Search Tool My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite Align Sequences Protein BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

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

MKHMAGAAAAGAVVGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHD
CVNITIKQHTVTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPVILLI
SFLI
FLIVG

Or, upload file No file chosen

Job Title Enter a descriptive title for your BLAST search

Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence Subject subrange

SWGQPHGGSGWPQPHGGGWGQGGGTHNQWNKPSKPCTNLKHVAGAAAAGAVVGLGGYMLGSAMSRPMIHF
GNDWEDRYYRENMYRPNQVYYRPVDQYSNQNNFVHDCVNITIKQHTVTTTKGENFTETDVKMMERVVE
QMCVTQYQKESQAYYDGRRSSSTVLFSSPPVILLISFLIFLIVG

Or, upload file No file chosen

Program Selection

Algorithm blastp (protein-protein BLAST)
 Choose a BLAST algorithm

BLAST

Search protein sequence using Blastp (protein-protein BLAST) Show results in a new window

Score	Expect	Method	Identities	Positives	Gaps
367 bits(941)	4e-132	Compositional matrix adjust.	227/287(79%)	243/287(84%)	35/287(12%)
Query 1	MANLGCWMLVLFVATWSDLGLCKRKPKPGWNTGGSRYPGQGSPGGNRYPQGGGWGQP				60
	MANLG W+L LFV W+D+GLCKRKPKPGWNTGGSRYPGQGSPGGNRYPQ GG WGQP				
Sbjct 1	MANLGYWLLALFVTMWTDVGLCKRKPKPGWNTGGSRYPGQGSPGGNRYPQ-GGTWGQP				59
Query 61	HGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGG				120
	HGGGWGQPHGG WGQP				HGG WGQPHGGG
Sbjct 60	HGGGWGQPHGGSWGQP-----				HGGSWGQPHGGG 87
Query 121	WGQGGGTHSQWNKPSKPCTNMKHMAGAAAAGAVVGGLGGYMLGSAMSRPIIHFGSDYEDR				180
	WGQGGGTH+QWNKPSKPCTN+KH+AGAAAAGAVVGGLGGYMLGSAMSRP+IHFG+D+EDR				
Sbjct 88	WGQGGGTHNQWNKPSKPCTNLKHVAGAAAAGAVVGGLGGYMLGSAMSRPMIHFGNDWEDR				147
Query 181	YYRENMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTTKGENFTETDVKMMER				240
	YYREN+RYPNQVYYRP+D+YSNQNNFVHDCVNITIKQHTVTTTKGENFTETDVKMMER				
Sbjct 148	YYRENMYRYPNQVYYRPVDQYSNQNNFVHDCVNITIKQHTVTTTKGENFTETDVKMMER				207
Query 241	VVEQMCITQYERESQAYY--QRGSSMVLFSSPPVILLISFLIFLIVG			285	
	VVEQMC+TQY++ESQAYY +R SS VLFSSPPVILLISFLIFLIVG				
Sbjct 208	VVEQMCVTQYQKESQAYYDGRRSSSTVLFSSPPVILLISFLIFLIVG			254	

Let's mask low-complexity regions

The screenshot shows the NCBI BLAST search interface with the following configuration:

- Algorithm parameters** section:
 - General Parameters**:
 - Max target sequences: 100
 - 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** button
- Search protein sequence using **Blastp (protein-protein BLAST)**
- Show results in a new window checkbox

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

BLAST is a registered trademark of the National Library of Medicine.

Score	Expect	Method	Identities	Positives	Gaps
182 bits(462)	6e-60	Compositional matrix adjust.	127/144(88%)	140/144(97%)	2/144(1%)
Query 127	THSQWNKPSKPCTNMKHM	agaaaagavvgglggymlgsamsRPIIHFGSDYEDRYYRENM TH+QWNKPSKPCTN+KH+AGAAAAGAVVGGGLGGYMLGSAMSRP+IHFG+D+EDRYYRENM		186	
Sbjct 94	THNQWNKPSKPCTNLKHV	AGAAAAGAVVGGGLGGYMLGSAMSRPMIHFGNDWEDRYYRENM		153	
Query 187	HRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQH	tvtttkgenftetDVKM MERVVEQMC +RYPNQVYYRP+D+YSNQNNFVHDCVNITIKQHTVTTTKGENFTETDVKM MERVVEQMC		246	
Sbjct 154	YRYPNQVYYRPVDQYSNQNNFVHDCVNITIKQH	TVTTTKGENFTETDVKM MERVVEQMC		213	
Query 247	ITQYERESQAYY--QRGSSMVLFS	268			
Sbjct 214	+TQY++ESQAYY +R SS VLFS				
Sbjct 214	VTQYQKESQAYYDGRSSSTVLFS	237			

Range 2: 1 to 49 [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#) [▲ First Match](#)

Score	Expect	Method	Identities	Positives	Gaps
98.2 bits(243)	1e-28	Compositional matrix adjust.	42/49(86%)	45/49(91%)	0/49(0%)
Query 1	MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNR	Y	49		
Sbjct 1	MANLG W+L LFV W+D+GLCKKRPKPGGWNTGGSRYPGQGSPGGNR	Y	49		

Score	Expect	Method	Identities	Positives	Gaps
367 bits(941)	4e-132	Compositional matrix adjust	227/287(79%)	243/287(84%)	35/287(12%)
Query 1	MANLGCWMLVLFVATWSDLGLCKRKPKPGWNTGGSRYPGQGSPGGNRY	60			
Sbjct 1	MANLG W+L LFV W+D+GLCKRKPKPGWNTGGSRYPGQGSPGGNRY	59			
Query 61	HGGGWGQPHGGGWQPHGGGWQPHGGGWQPHGGGWQPHGGGWQPHGGG	120			
Sbjct 60	HGGGWGQPHGGSWQPHGGSWQPHGGSWQPHGGSWQPHGGG-----	87			
Query 121	WGQGGGTHSQWNKPSKPCTNMKHAGAAAAGAVVGLGGYMLGSAMS	180			
Sbjct 88	RPIIHFGSDYEDR+QWNKPSKPCTN+KH+AGAAAAGAVVGLGGYMLGSAMSRP+IHFG+D+EDR	147			
Query 181	YYRENMRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTTKGENFTETDVKM	240			
Sbjct 148	YYRENMRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTTKGENFTETDVKM	207			
Query 241	VVEQMCITQYERESQAYY--QRGSSMVLFS	285			
Sbjct 208	+TQY++ESQAYY +R SS VLFS	254			

182 bits(462)	6e-60	Compositional matrix adjust	127/144(88%)	140/144(97%)	2/144(1%)
Query 127	THSQWNKPSKPCTNMKHMagaaaagavvglggymlgamsRPIIHFGSDYEDRYYREN	186			
Sbjct 94	M+QWNKPSKPCTN+KH+AGAAAAGAVVGLGGYMLGSAMSRP+IHFG+D+EDRYYREN	153			
Query 187	HRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTTKGENFTETDVKM	246			
Sbjct 154	MRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTTKGENFTETDVKM	213			
Query 247	ITQYERESQAYY--QRGSSMVLFS	268			
Sbjct 214	+TQY++ESQAYY +R SS VLFS	237			

Range 2: 1 to 49 Graphics

▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
98.2 bits(243)	1e-28	Compositional matrix adjust	42/49(86%)	45/49(91%)	0/49(0%)
Query 1	MANLGCWMLVLFVATWSDLGLCKRKPKPGWNTGGSRYPGQGSPGGNRY	49			
Sbjct 1	MANLG W+L LFV W+D+GLCKRKPKPGWNTGGSRYPGQGSPGGNRY	49			

PAM30

PAM70

PAM250

BLOSUM80

✓ BLOSUM62

BLOSUM45

BLOSUM50

BLOSUM90

Protein similarity scoring matrices

- **PAM** (point accepted mutations) matrices (**Dayhoff et al., 1978**) depict amino acid substitution (1572 changes) patterns in 71 groups of closely related (sequence identity > 85%) proteins.
- Major assumptions:
 - Substitutions are independent!
 - All positions are equally mutable!
 - In 1978 most of known proteins were small and globular!
 - Speed of evolution and forces are constant!
- PAM1 corresponds to 1 aa change per 100 aa (1% divergence).
- PAM100 – 100 changes per 100 aa; made from PAM1
- PAM250 – 250 changes per 100 aa; made from PAM1

All PAM and BLOSUM matrices are log odds

$$S_{ij} = \log \frac{P_{ij}}{p_i p_j}$$

- S_{ij} is the log odd ratio of two probabilities: the probability that two residue types, i and j, are aligned by evolutionary descent and the probability that they are aligned by chance
- p_{ij} are the frequencies that residue i and j are observed to align in sequences known to be related.
- p_i and p_j are the frequencies of occurrence of residue i and j in all protein sequences (their abundance; or background probability)

Positive values mean that the substitution is observed to occur more often than would by chance, negative - the substitution is not observed to occur frequently and may arise more often than not by chance

PAM 100 Matrix

```
#  
# This matrix was produced by "pam" Version 1.0.7 [13-Aug-03]  
#  
# PAM 100 substitution matrix, scale = ln(2)/2 = 0.346574  
#  
# Expected score = -1.99, Entropy = 1.18 bits  
#  
# Lowest score = -9, Highest score = 12  
#  
A R N D C Q E G H I L K M F P S T W Y V B Z X *  
A 4 -3 -1 -1 -3 -2 0 1 -3 -2 -3 -3 -2 -5 1 1 1 -7 -4 0 -1 -1 -1 -9  
R -3 7 -2 -4 -5 1 -3 -5 1 -3 -5 2 -1 -6 -1 -1 -3 1 -6 -4 -3 -1 -2 -9  
N -1 -2 5 3 -5 -1 1 -1 2 -3 -4 1 -4 -5 -2 1 0 -5 -2 -3 4 0 -1 -9  
D -1 -4 3 5 -7 0 4 -1 -1 -4 -6 -1 -5 -8 -3 -1 -2 -9 -6 -4 4 3 -2 -9  
C -3 -5 -5 -7 9 -8 -8 -5 -4 -3 -8 -8 -7 -7 -4 -1 -9 1 -3 -6 -8 -5 -9  
Q -2 1 -1 0 -8 6 2 -3 3 -4 -2 0 -2 -7 -1 -2 -2 -7 -6 -3 0 5 -2 -9  
E 0 -3 1 4 -8 2 5 -1 -1 -3 -5 -1 -4 -8 -2 -1 -2 -9 -5 -3 3 4 -2 -9  
G 1 -5 -1 -1 -5 -3 -1 5 -4 -5 -6 -3 -4 -6 -2 0 -2 -9 -7 -3 -1 -2 -2 -9  
H -3 1 2 -1 -4 3 -1 -4 7 -4 -3 -2 -4 -3 -1 -2 -3 -4 -1 -3 1 1 -2 -9  
I -2 -3 -3 -4 -3 -4 -3 -5 -4 6 1 -3 1 0 -4 -3 0 -7 -3 3 -3 -3 -2 -9  
L -3 -5 -4 -6 -8 -2 -5 -6 -3 1 6 -4 3 0 -4 -4 -3 -3 0 -5 -4 -3 -9  
K -3 2 1 -1 -8 0 -1 -3 -2 -3 -4 5 0 -7 -3 -1 -1 -6 -6 -4 0 -1 -2 -9  
M -2 -1 -4 -5 -7 -2 -4 -4 -4 1 3 0 9 -1 -4 -3 -1 -6 -5 1 -4 -2 -2 -9  
F -5 -6 -5 -8 -7 -7 -8 -6 -3 0 0 -7 -1 8 -6 -4 -5 -1 4 -3 -6 -7 -4 -9  
P 1 -1 -2 -3 -4 -1 -2 -2 -1 -4 -4 -3 -4 -6 7 0 -1 -7 -7 -3 -3 -1 -2 -9  
S 1 -1 1 -1 -1 -2 -1 0 -2 -3 -4 -1 -3 -4 0 4 2 -3 -4 -2 0 -2 -1 -9  
T 1 -3 0 -2 -4 -2 -2 -2 -3 0 -3 -1 -1 -5 -1 2 5 -7 -4 0 -1 -2 -1 -9  
W -7 1 -5 -9 -9 -7 -9 -9 -4 -7 -3 -6 -6 -1 -7 -3 -1 12 2 -9 -6 -8 -6 -9  
Y -4 -6 -2 -6 -1 -6 -5 -7 -1 -3 -3 -6 -5 4 -7 -4 -4 -2 9 -4 -4 -6 -4 -9  
V 0 -4 -3 -4 -3 -3 -3 -3 -3 3 0 -4 1 -3 -3 -2 0 -9 -4 5 -4 -3 -2 -9  
B -1 -3 4 4 -6 0 3 -1 1 -3 -5 0 -4 -6 -3 0 -1 -6 -4 -4 4 2 -2 -9  
Z -1 -1 0 3 -8 5 4 -2 1 -3 -4 -1 -2 -7 -1 -2 -2 -8 -6 -3 2 5 -2 -9  
X -1 -2 -1 -2 -5 -2 -2 -2 -2 -3 -2 -2 -4 -2 -1 -1 -6 -4 -2 -2 -2 -2 -9  
* -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 -9 1
```

PAM 250 Matrix

```
#  
# This matrix was produced by "pam" Version 1.0.7 [13-Aug-03]  
#  
# PAM 250 substitution matrix, scale = ln(2)/3 = 0.231049  
#  
# Expected score = -0.844, Entropy = 0.354 bits  
#  
# Lowest score = -8, Highest score = 17  
#  
A R N D C Q E G H I L K M F P S T W Y V B Z X *  
A 2 -2 0 0 -2 0 0 1 -1 -1 -2 -1 -1 -3 1 1 1 -6 -3 0 0 0 0 0 0 0 0 -8  
R -2 6 0 -1 -4 1 -1 -3 2 -2 -3 3 0 -4 0 0 -1 2 -4 -2 -1 0 -1 -8 0 -1 -8  
N 0 0 2 2 -4 1 1 0 2 -2 -3 1 -2 -3 0 1 0 -4 -2 -2 2 1 0 -8  
D 0 -1 2 4 -5 2 3 1 1 -2 -4 0 -3 -6 -1 0 0 -7 -4 -2 3 3 -1 -8  
C -2 -4 -4 -5 12 -5 -5 -3 -3 -2 -6 -5 -5 -4 -3 0 -8 0 -2 -4 -5 -3 -8  
Q 0 1 1 2 -5 4 2 -1 3 -2 -2 1 -1 -5 0 -1 -1 -5 -4 -2 1 3 -1 -8  
E 0 -1 1 3 -5 2 4 0 1 -2 -3 0 -2 -5 -1 0 0 -7 -4 -2 3 3 -1 -8  
G 1 -3 0 1 -3 -1 0 5 -2 -3 -4 -2 -3 -5 0 1 0 -7 -5 -1 0 0 -1 -8  
H -1 2 2 1 -3 3 1 -2 6 -2 -2 0 -2 -2 0 -1 -1 -3 0 -2 1 2 -1 -8  
I -1 -2 -2 -2 -2 -2 -3 -2 5 2 -2 -2 1 -1 -5 0 -5 -1 4 -2 -2 -1 -8  
L -2 -3 -3 -4 -6 -2 -3 -4 -2 2 6 -3 4 2 -3 -3 -2 -2 -1 2 -3 -3 -1 -8  
K -1 3 1 0 -5 1 0 -2 0 -2 -3 5 0 -5 -1 0 0 -3 -4 -2 1 0 -1 -8  
M -1 0 -2 -3 -5 -1 -2 -3 -2 2 4 0 6 0 -2 -2 -1 -4 -2 2 -2 -2 -1 -8  
F -3 -4 -3 -6 -4 -5 -5 -2 1 2 -5 0 9 -5 -3 -3 0 7 -1 -4 -5 -2 -8  
P 1 0 0 -1 -3 0 -1 0 0 -2 -3 -1 -2 -5 6 1 0 -6 -5 -1 -1 0 -1 -8  
S 1 0 1 0 0 -1 0 1 -1 -1 -3 0 -2 -3 1 2 1 -2 -3 -1 0 0 0 -8  
T 1 -1 0 0 -2 -1 0 0 -1 0 -2 0 -1 -3 0 1 3 -5 -3 0 0 -1 0 -8  
W -6 2 -4 -7 -8 -5 -7 -3 -5 -2 -3 -4 0 -6 -2 -1 17 0 -6 -5 -6 -4 -8  
Y -3 -4 -2 -4 0 -4 -4 -5 0 -1 -1 -4 -2 7 -5 -3 -3 0 10 -2 -3 -4 -2 -8  
V 0 -2 -2 -2 -2 -2 -2 -1 -2 4 2 -2 2 -1 -1 0 -6 -2 4 -2 -2 -1 -8  
B 0 -1 2 3 -4 1 3 0 1 -2 -3 1 -2 -4 -1 0 0 -5 -3 -2 3 2 -1 -8  
Z 0 0 1 3 -5 3 3 0 2 -2 -3 0 -2 -5 0 0 -1 -6 -4 -2 2 3 -1 -8  
X 0 -1 0 -1 -3 -1 -1 -1 -1 -1 -2 -1 0 0 -4 -2 -1 -1 -1 -1 -1 -1 -8  
* -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -8 -1
```

PAM30

PAM70

PAM250

BLOSUM80

✓ BLOSUM62

BLOSUM45

BLOSUM50

BLOSUM90

Protein similarity scoring matrices

- **BLOSUM** (blocks substitution matrices) (**henikoff & Henikoff, 1992**) depict amino acid substitution in 2,000 blocks (corresponding to structural or functional motifs) in 500 groups of related proteins.
- Same major assumptions as for PAM matrices:
 - Substitutions are independent!
 - All positions are equally mutable!
 - Speed of evolution and forces are constant!
- In contrast to PAM matrices BLOSUM matrices are not derived but calculated directly.
- BLOSUM**62** represent the conservation level of sequences used to derive the matrix (seq. identity >62%)
- BLOSUM**90** ~ 90%, BLOSUM**45** ~ 45%, etc.

BLOSUM62 – the most often used matrix

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

N D E Q

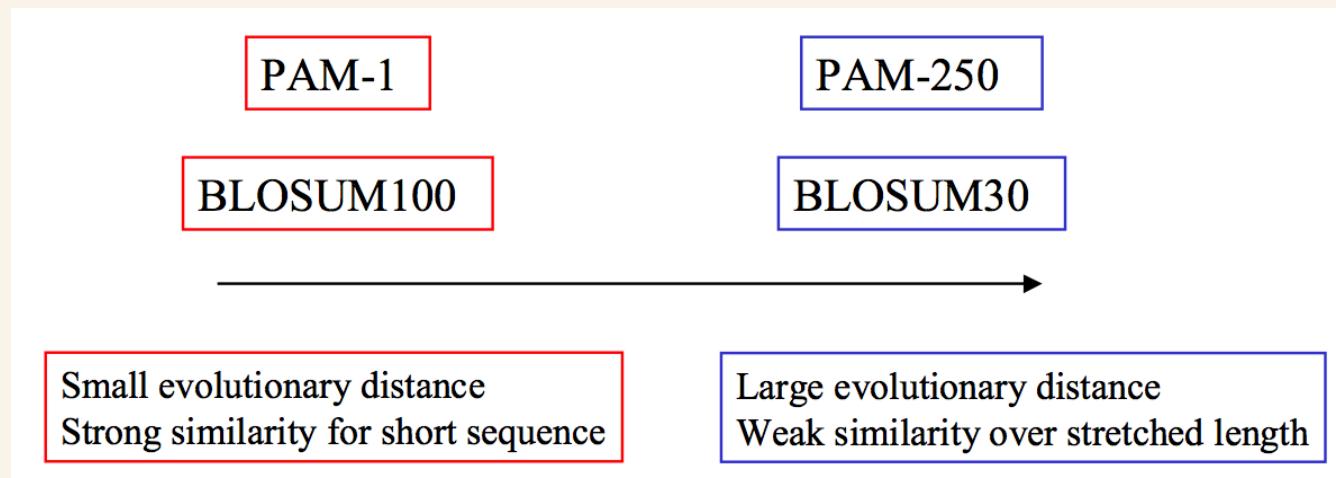
N T W Q

Score 6-1-3+5 = 7

PAM30
PAM70
PAM250
BLOSUM80
✓ BLOSUM62
BLOSUM45
BLOSUM50
BLOSUM90

What matrices should be used when?

Wheeler, 2003



- PAM40 & BLOSUM90 – Short highly significant alignments (70-90% similarity)
- PAM160 (~BLOSUM62) - Detecting members of a protein family (50-60%)
- BLOSUM80 (~PAM120) – Detecting members of a protein family (50-60%)
- **BLOSUM62 (~PAM160)** – Effective in finding all potential similarities (30-40%)
- PAM250 & BLOSUM30 – Longer alignments of divergent sequences (<30%)

The sequences we aligned using bl2seq have similarity >85%

Score	Expect	Method	Identities	Positives	Gaps
182 bits(462)	6e-60	Compositional matrix adjust.	127/144(88%)	140/144(97%)	2/144(1%)
Query 127	THSQWNKPSKPKNMKH	magaaaaagavvgglggymlgsamsRPIIHFGSDYEDRYYRENM TH+QWNKPSKPKN+KH+AGAAAAGAVVGGGLGGYMLGSAMSRP+IHFG+D+EDRYYRENM		186	
Sbjct 94	THNQWNKPSKPKNLKH	VAGAAAAGAVVGGGLGGYMLGSAMSRPMIHFGNEDRYYRENM		153	
Query 187	HRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQH	tvttttkgenftetDVKM MERVVEQMC +RYPNQVYYRP+D+YSNQNNFVHDCVNITIKQHTVTTTKGENFTETDVKM MERVVEQMC		246	
Sbjct 154	YRYPNQVYYRPVDQYSNQNNFVHDCVNITIKQH	TVTTTKGENFTETDVKM MERVVEQMC		213	
Query 247	ITQYERESQAYY--QRGSSMVLFS	268			
Sbjct 214	+TQY++ESQAYY +R SS VLFS				
Sbjct 214	VTQYQKESQAYYDGRSSSTVLFS	237			
Range 2: 1 to 49 Graphics			▼ Next Match	▲ Previous Match	▲ First Match
Score	Expect	Method	Identities	Positives	Gaps
98.2 bits(243)	1e-28	Compositional matrix adjust.	42/49(86%)	45/49(91%)	0/49(0%)
Query 1	MANLCWMLVLFVATWSDLGLCKRKPKPGGWNTGGSRYPGQGSPGGNR	Y	49		
Sbjct 1	MANLG W+L LFV W+D+GLCKRKPKPGGWNTGGSRYPGQGSPGGNR	Y	49		

And which matrix we used (default)?

BLAST Search protein sequence 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 protein sequence using Blastp (protein-protein BLAST)
 Show results in a new window

Let's try BLOSUM80 for higher sequence similarity

- The result is the same because there is no gaps in the alignment
- But let's try another sequences

NP_000566 vs NP_000567

BLAST® Basic Local Alignment Search Tool My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite Align Sequences Protein BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

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

Query subrange From
 To

Or, upload file No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence Clear

Subject subrange From
 To

Or, upload file No file chosen

Program Selection

Algorithm blastp (protein-protein BLAST)
Choose a BLAST algorithm

BLAST Search protein sequence using Blastp (protein-protein BLAST)
 Show results in a new window

BLOSUM62

Score	Expect	Method	Identities	Positives	Gaps
42.4 bits(98)	3e-09	Compositional matrix adjust.	71/286(25%)	118/286(41%)	42/286(14%)
Query 1	MAKVPDMFEDLKNCYSENEED-SSSIDHLSLNQKSFYHVSYGPLHEGCMQSVSLSISET MA+VP++ ++ YS NE+D D + SF + PL D + L IS+				59
Sbjct 1	MAEVPELASEMMAYSGNEDDLFFEADGPKQMCKCSFQDLDLCPL-----DGGIQLRISDH				55
Query 60	SKTSKLTFKESMVVATNGVLKKRRLSLSQSITDDDL-----EAIANDSEEEII + F+++ VV K L+K + Q+ ++DL E I D+ +				109
Sbjct 56	HYSK--GFRQAASVVVAMDK-LRKMLVPCPQTQENDLSTFFPFIFEEEPIFFDTWDNEA				112
Query 110	KPRSAPFSFLSNVKYNFMRIIKYEFILNDALNQSIIRANDQYLTAAL--HNLDEAVKFD AP L+ L D+ +S++ + L A L +++++ V F				167
Sbjct 113	YVHDAPVRSLNCT-----LRDSQQKSLVMSGPYELKALHLQGQDMEQQVVFS				159
Query 168	MGAYKSSKDDAKITVILRISKTLQYVTA-QDEDQPVLLKEMPEIPKTITGSETNLLFFWE M + + + KI V L + + LY++ +D+P L E + PK + F +				226
Sbjct 160	MSFVQGEESNDKIPVALGLKEKNLYLSCVLKDDKPTLQLESVD-PKNYPKKMEKRFVFN				218
Query 227	THGTKNY--FTSVAHPNLFIAVKQ--DYWVCLAG--GPPSITDFQI 266 N F S PN +I+T Q + V L G G ITDF +				
Sbjct 219	KIEINNKLEFESAQFPNWYISTSQAENMPVFLGGTKGGQDITDFTM 264				

BLOSUM45

Score	Expect	Method	Identities	Positives	Gaps
37.7 bits(113)	8e-08	Compositional matrix adjust.	74/276(27%)	117/276(42%)	26/276(9%)
Query 1	MAKVPDMFEDLKNCYSENEED-	SSSIDHLSLNQKSFYHVSYGPLHEGCMQSVSLSISET		59	
Sbjct 1	MA+VP++ ++ YS NE+D D + SF + PL D + L IS+	GNEDDLFFEADGPKQMCKCSFQDLDLCPL-----DGGIQLRISDH		55	
Query 60	SKTSKLTFKESMVVVATNGKVLKKRRLSLSQSITDDDLEAI-	ANDSEEEIIKPRSAPFSF		118	
Sbjct 56	+ F+++ VV K L+K + Q+ ++DL EEE I F	HYSK--GFRQAASVVVAMDK-LRKMLVPCPQTQENDLSTFFPFIFEEEPIF-----FDT		107	
Query 119	LSNVKYNFMRIIK-YEFILNDALNQSIIRANDQYLTAALHNLD--	EAVKFDMGAYKSSK		175	
Sbjct 108	N Y ++ L D+ +S++ + L A L+ D + V F M + +	WDNEAYVHDAPVRSLNCTLRDSQQKSLVMSGPYELKALHLQGQDMEQQVVFMSFVQGEE		167	
Query 176	DDAKITVILRISKTQLYVT-AQDEDQPVLLKEMPEIPKTITGS--	ETNLLFFWETHGTKN		232	
Sbjct 168	+ KI V L + + LY++ +D+P L E + PK E ++F K	SNDKIPVALGLKEKNLYLSCVLKDDKPTLQLESVD-PKNYPKKMEKRFVFNKIEINNKL		226	
Query 233	YFTSVAHPNLFIATKQ--DYWVCLAG--GPPSITDF	264			
Sbjct 227	F S PN +I+T Q + V L G G ITDF	EFESAQFPNWYISTSQAENMPVFLGGTKGGQDITDF	262		

BLOSUM90

Score	Expect	Method	Identities	Positives	Gaps
45.2 bits(99)	8e-10	Compositional matrix adjust.	74/278(27%)	112/278(40%)	26/278(9%)
Query 1	MAKVPDMFEDLKNCYSENEED-SSSIDHLSLNQKSFYHVSYGPLHEGCMQSVSLSISET				59
Sbjct 1	MAEVPELASEMMAYYSGNEDDLFFEADGPQKQMKCSFQDLDLCPLDGG-----IQLRISD-				54
Query 60	SKTSKLTFKESMVVVATNGKVLKKRRLSLSQSITDDDLEAI-ANDSEEEIIKPRSAAPFSF				118
Sbjct 55	HHYSKGFRQAASVVVAMDK--LRKMLVPCPQTFQENDLSTFFPFIFEEEPIF-----FDT				107
Query 119	LSNVKYNFMRIIK-YEFILNDALNQSIIRANDQYLTAALH--NLDEAVKFDMGAYKSSK				175
Sbjct 108	N Y ++ L D+ +S++ + L A L+ +++++ V F M + WDNEAYVHDAPVRSLNCTLRDSQQKSLVMSGPYELKALHLQGQDMEQQVVFSMSFVQGEE				167
Query 176	DDAKITVILRISKTQLYVTA-QDEDQPVLLKEMPEIPKTITGSETNLLFWETHGTNY-				233
Sbjct 168	+ KI V L + LY + +D+P L E + PK F N SNDKIPVALGLKEKNLYLSCVLKDDKPTLQLESVD-PKNYPKKMEKRFVFNKIEINNKL				226
Query 234	-FTSVAHPNLFIATKQ--DYWVCLAG--GPPSITDFQI	266			
Sbjct 227	F S PN +I+T Q + V L G G ITDF + EFESAQFPNWYISTSQAENMPVFLGGTKGGQDITDFTM	264			

PAM250

Score	Expect	Method	Identities	Positives	Gaps
31.9 bits(101)	3e-06	Compositional matrix adjust.	45/209(22%)	107/209(51%)	19/209(9%)
Query 1	MAKVPDMFEDLKNCYSENEEDSSSIDHLSLNQKSFYHVSYGPLHEGCMDQSVSLSISETS				60
Sbjct 1	MAEVPELASEMMAYSGNEDDLFFEADGPKQMKC----SFQDLDLCPLDGGIQLRISDH				56
Query 61	KTSKLTFKESMVVATNGKVLKKRRLSLSQSITDDDLEAIAN---DSEEEIIKPRSAPFS				117
Sbjct 57	+ F+++ VV + +K L+K ++ +Q++ ++DL ++ + E ++ + + YSK--GFRQAASVVVAMD				112
Query 118	L RDSQQKSLVMSGPYELKALHLQGQDMEQQVVFSMSFVQGEE				175
Sbjct 113	YVHDAP---VRSLNC--TLRDSQQKSLVMSGPYELKALHLQGQDMEQQVVFSMSFVQGEE				167
Query 176	DDAKITVILRISKTQLYVT-AQDEDQPVL	203			
Sbjct 168	+ KI V L + +LY++ +D+P L SNDKIPVALGLKEKNLYLSCVLKDDKPTL	196			

PAM30

Score	Expect	Method	Identities	Positives	Gaps
52.4 bits(116)	9e-12	Compositional matrix adjust.	74/277(27%)	96/277(34%)	28/277(10%)
Query 1	MAKVPDMFEDLKNCYSENEEDSSSIDHLSLNQKSFYHVSYGPLHEGCMDQSVSLSISETS				60
	MA VP++ ++ YS NE+D	K S+ L +D + L IS+			
Sbjct 1	MAEVPELASEMMAYSGNEDDLFFEADGPKQMKC----	SFQDLDLCPLDGGIQLRISD-H			55
Query 61	KTSKLTFKESMVVATNGKVLKKRRLSLSQSITDDDLEA-IANDSEEEIIKPRSAPFSFL				119
	SK F + VV K L K Q ++DL EEE I F				
Sbjct 56	HYSK-GFRQAASVVVAMDK-LRKMLVPCPQTFQENDLSTFFPFIFEEEPI-----FFDTW				108
Query 120	SNVKYNF---MRIIKYEFILNDALNQSIIRANDQYLTAAL--HNLDEAVKFDMGAYKSS				174
	N Y R L D S + L A L +++++ V F M				
Sbjct 109	DNEAYVHDAPVRSLNC--TLRDSQQKSLVMSGPYELKALHLQGQDMEQQQVVFMSFVQGE				166
Query 175	KDDAKITVILRISKTLQYVT-AQDEDQPVLLKEMPEIPKTITGS--ETNLLFFWETHGTK				231
	+ KI V L LY +D P L E + PK E F K				
Sbjct 167	ESNDKIPVALGLKEKNLYLSCVLKDDKPTLQLESVD-PKNYPKKMEKRFVFNKIEINNK				225
Query 232	NYFTSVAHPNLFIATKQDYW--VCLAG--GPPSITDF	264			
	F S PN +I T Q V L G G ITDF				
Sbjct 226	LEFESAQFPNWYISTSQAENMPVFLGGTKGGQDITDF	262			

PAM250 (<30%) → BLOSUM45 → BLOSUM62 (30-40%) →
 BLOSUM90 (70-90%) → PAM30 (80-100%)

Score	Expect	Method	Identities	Positives	Gaps
31.9 bits(101)	3e-06	Compositional matrix adjust.	45/209(22%)	107/209(51%)	19/209(9%)
Score	Expect	Method	Identities	Positives	Gaps
37.7 bits(113)	8e-08	Compositional matrix adjust.	74/276(27%)	117/276(42%)	26/276(9%)
Score	Expect	Method	Identities	Positives	Gaps
42.4 bits(98)	3e-09	Compositional matrix adjust.	71/286(25%)	118/286(41%)	42/286(14%)
Score	Expect	Method	Identities	Positives	Gaps
45.2 bits(99)	8e-10	Compositional matrix adjust.	74/278(27%)	112/278(40%)	26/278(9%)
Score	Expect	Method	Identities	Positives	Gaps
52.4 bits(116)	9e-12	Compositional matrix adjust.	74/277(27%)	96/277(34%)	28/277(10%)

- In **BLAST** the suggested cut-offs to infer biologically significant similarity:
 - ✓ For proteins: sequence identity $\geq 25\%$, E-value $\leq 1e-3$
 - ✓ For nucleotides: sequence identity $\geq 70\%$, E-value $\leq 1e-6$
- Do not use those cut-offs blindly!