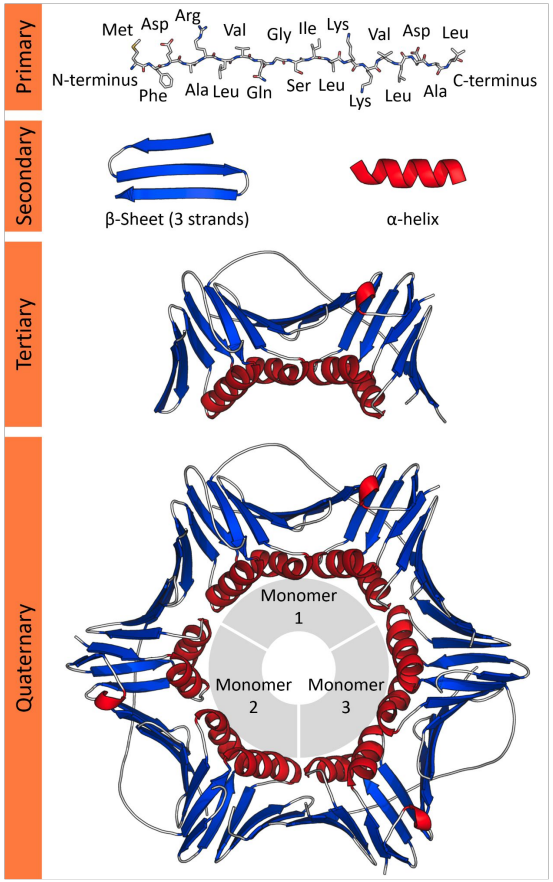
Structural Bioinformatics

Some basics of bioinformatics

Protein structures

- Get structures from databases
- Visualise
- Compare with each other
- Predict
- Classify



Basic Instruments and sources of data in bioinformatics

- Sequence alignments (pairwise, multiple)

- Databases of sequences and structures
  Uniprot
  PDB
  NCBI

- BLAST - algorithm to search for a sequence in the database

FASTA format

>sequence_description_DNA

ACTGCTAGCATCAGACTACGACT

>sequence_description_protein

MKTSRTKLPLS

Please check the folder “Exercises” in Moodle.

Sequence alignments in bioinformatics

* we will talk about structural alignments during this course

- Local / Global
- Pairwise

or

Local Alignment

```
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| ||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

```
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
Query Sequence
```

Multiple Sequence Alignment (MSA)

Online tools for pairwise alignments

https://www.ebi.ac.uk/Tools/psa/

Sequences for alignments

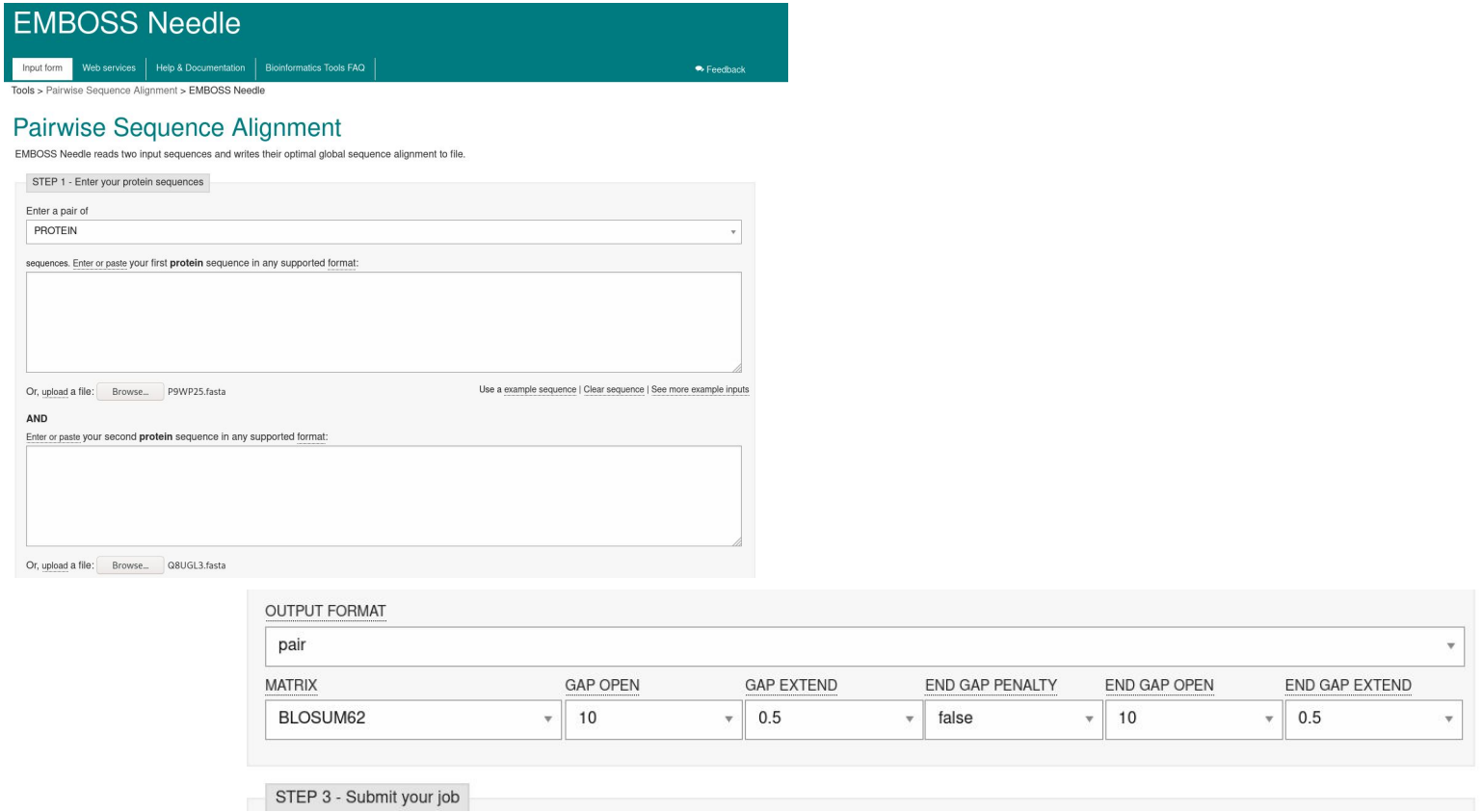
>sp|Q8UGL3|DAPA_AGRFC 4-hydroxy-tetrahydrodipicolinate synthase OS=Agrobacterium fabrum (strain C58 / ATCC 33970) OX=176299 GN=dapA PE=1 SV=1

MFKGSIPALITPFTDNGSVDEKAFAAHVEWQIAEGSNGLVPVGTTGESPTLSHDEHKRVVELCIEVAAKRV
PVIAGAGSNNTDEAIELALHAQEAGADALLVVTPYYNKPTQKGLFAHFSAVAEAVKLPIVIYNIPPRSVVDM
SPETMGALVKAHKNIIGVKDATGKLDRVSEQRISCGKDFVQLSGEDGTALGFNAHGGVGCISVTANVAPR
LCSEFQAAMLAGDYAKALEYQDRLMPLHRAIFMEPGVCGTKYALSKTRGGNRRVRSPLMSTLEPATEAAI
DAALKHAGLMN

>sp|P9WP25|DAPA_MYCTU 4-hydroxy-tetrahydrodipicolinate synthase OS=Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv) OX=83332 GN=dapA PE=1 SV=1

MTTVGFDVAARLGTLLTAMVTPFSGDGSLDTATAARLANHLVDQGCDGLVVSGTTGESPTTTDGEKIELL
RAVLEAVGDRARVIAGAGTYDTAHSIRLAKACAAEGAHGLLVVTPYYSKPPQRGLQAHFTAVADATELPM
LLYDIPGRSAVPIEPDTIRALASHPNIVGVKDAKADLHSGAQIMADTGLAYYSGDDALNLPWLAMGATGFIS
VIAHLAAGQLRELLSAFGSGDIATARKINIAVAPLCNAMSRLGGVTLSKAGLRLQGIDVGDPRLPQVAATPE
QIDALAADMRAASVLR

Input


EMBOSS Needle
Input form
Web services
Help & Documentation
Bioinformatics Tools FAQ
Feedback
Tools > Pairwise Sequence Alignment > EMBOSS Needle
Pairwise Sequence Alignment
EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.
STEP 1 - Enter your protein sequences
Enter a pair of
PROTEIN
sequences. Enter or paste your first protein sequence in any supported format:
Or, upload a file: Browse... P9WP25.fasta
Use a example sequence | Clear sequence | See more example inputs
AND
Enter or paste your second protein sequence in any supported format:
Or, upload a file: Browse... Q8UGL3.fasta
OUTPUT FORMAT
pair
MATRIX
BLOSUM62
GAP OPEN
10
GAP EXTEND
0.5
END GAP PENALTY
false
END GAP OPEN
10
END GAP EXTEND
0.5
STEP 3 - Submit your job

Output

Results for job emboss_needle-I20221115-173525-0089-79876523-p1m

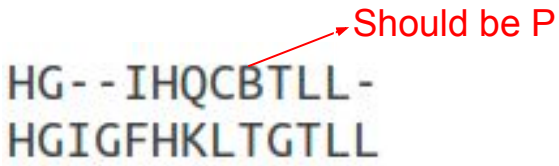
Alignment | Submission Details

View Alignment File

```
########################################
# Program: needle
# Rundate: Tue 15 Nov 2022 17:31:44
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20221115-173525-0089-79876523-p1m.aupfile
#    -bsequence emboss_needle-I20221115-173525-0089-79876523-p1m.bupfile
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: DAPA_MYCTU
# 2: DAPA_AGRFC
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 310
# Identity:     114/310 (36.8%)
# Similarity:   165/310 (53.2%)
# Gaps:          26/310 ( 8.4%)
# Score: 425.5
#
#
#=======================================
```

```
DAPA_MYCTU        1 MTTVGFDVAARLGTLLTAMVTPFSGDGSLDTATAARLANHLVDQGCDGLV     50
                    .....:.|::|||:.:||:|....|......:.:|.:|||
DAPA_AGRFC        1 ----------MFKGSIPALITPFTDNGSVDEKAFAAHVEWQIAEGSNGLV     40

DAPA_MYCTU       51 VSGTTGESPTTTDGEKIELLRAVLEAVGDRARVIAGAGTYDTAHSIRLAK    100
                    ..||||||||.:..|...::...:|....|..||||||:.:|..:|.||.
DAPA_AGRFC       41 PVGTTGESPTLSHDEHKRVVELCIEVAAKRVPVIAGAGSNNTDEAIELAL     90

DAPA_MYCTU      101 ACAAEGAHGLLVVTPYYSKPPQRGLQAHFTAVADATELPMLLYDIPGRSA    150
                    .....||..||||||||:||.|:||.|||:|||:|.:||:::|:||.||.
DAPA_AGRFC       91 HAQEAGADALLVVTPYYNKPTQKGLFAHFSAVAEAVKLPIVIYNIPPRSV    140

DAPA_MYCTU      151 VPIEPDTIRALA-SHPNIVGVKDA--KADLHSGAQIMADTGLAYYSGDDA    197
                    |.:.|:|:.||. :|.||:|||||  |.|..|..:|.........||:|.
DAPA_AGRFC      141 VDMSPETMGALVKAHKNIIGVKDATGKLDRVSEQRISCGKDFVQLSGEDG    190

DAPA_MYCTU      198 LNLPWLAMGATGFISVIAHLAAGQLRELLSAFGSGDIATARKINIAVAPL    247
                    ..|.:.|.|..|.|||.|::|.....|..:|..::||.|.|.:....:.||
DAPA_AGRFC      191 TALGFNAHGGVGCISVTANVAPRLCSEFQAAMLAGDYAKALEYQDRLMPL    240

DAPA_MYCTU      248 CNAMSRLGGVTLSKAGL-RLQGIDVGDPRL--PQVA----ATPEQIDALA    290
                    ..|:....||..:|..| :.:|   |:.|:  |.::    ||...|||.
DAPA_AGRFC      241 HRAIFMEPGVCGTKYALSKTRG---GNRRVRSPLMSTLEPATEAAIDAA-    286

DAPA_MYCTU      291 ADMRAASVLR    300
                      ::.|.::.
DAPA_AGRFC      287 --LKHAGLMN    294


#---------------------------------------
#---------------------------------------
```

Calculate the identity of the pairwise alignment



Identity = N of identical matches / N of residues in alignment

N of residues in alignment = 13

N of identical matches = 4

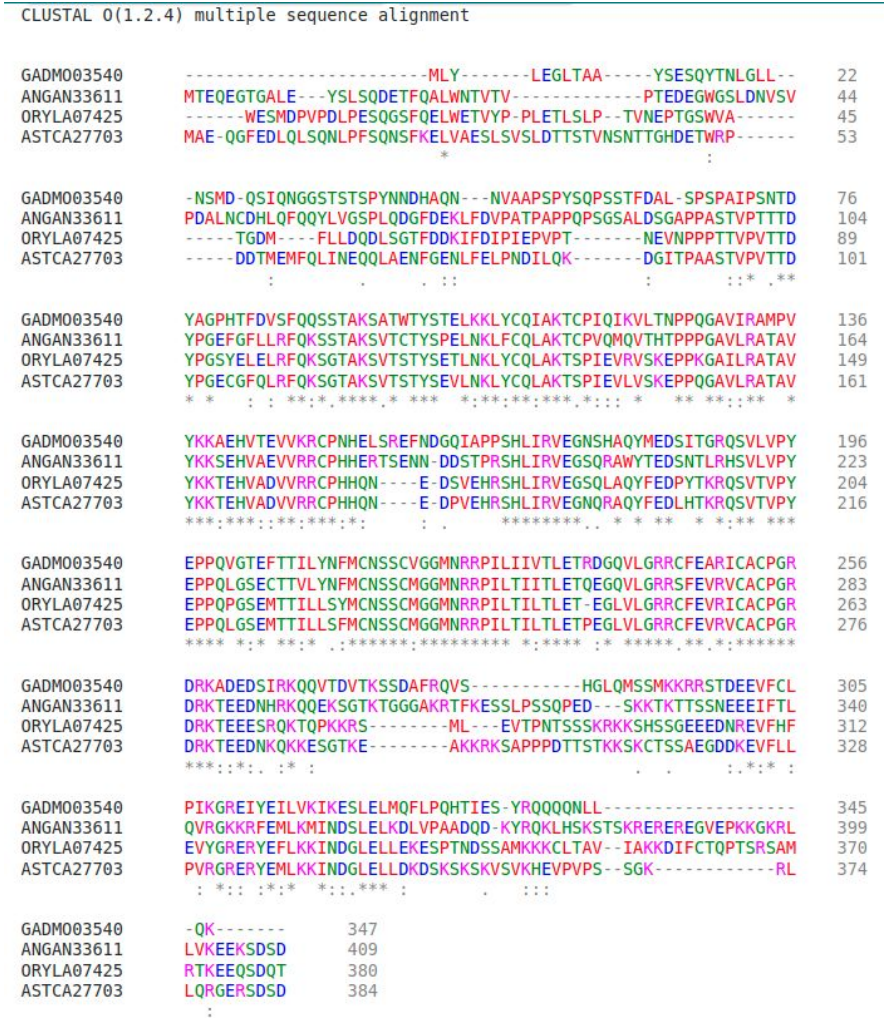
Identity = 4/13 = 0.30769 31%

MSA

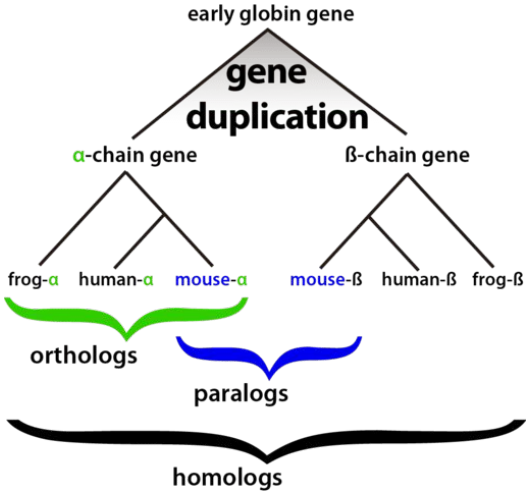
Clustal Omega

https://www.ebi.ac.uk/Tools/msa/clustalo/

Please use the .fasta file with 4 protein sequences in Moodle.

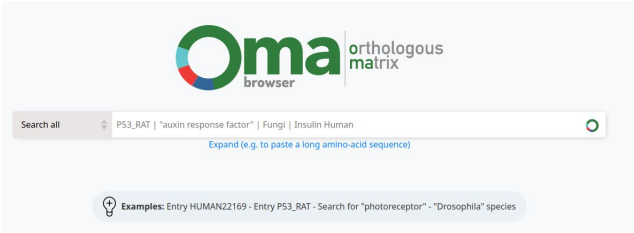


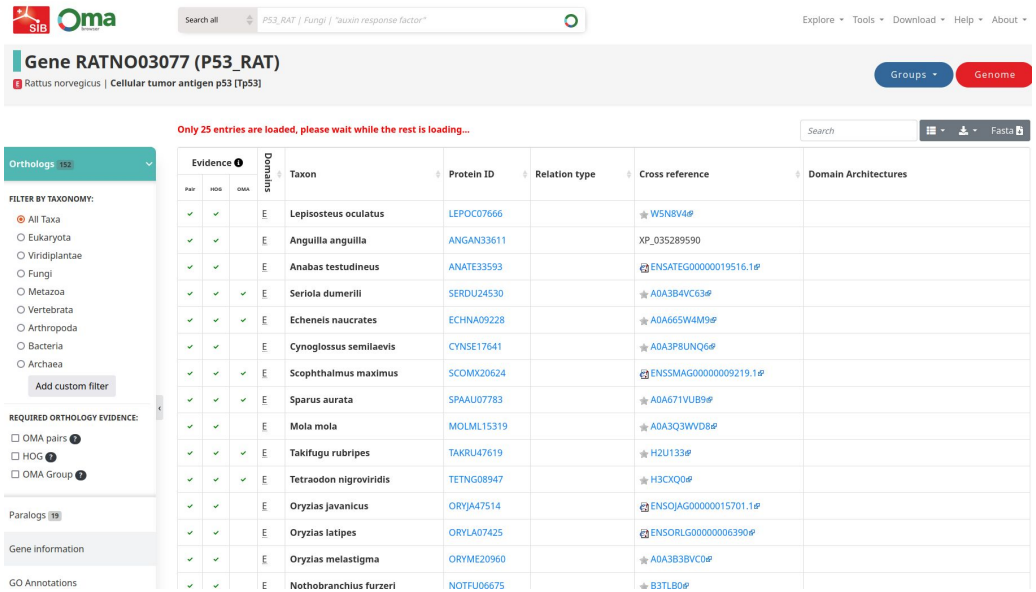
Homologs, orthologs, paralogs

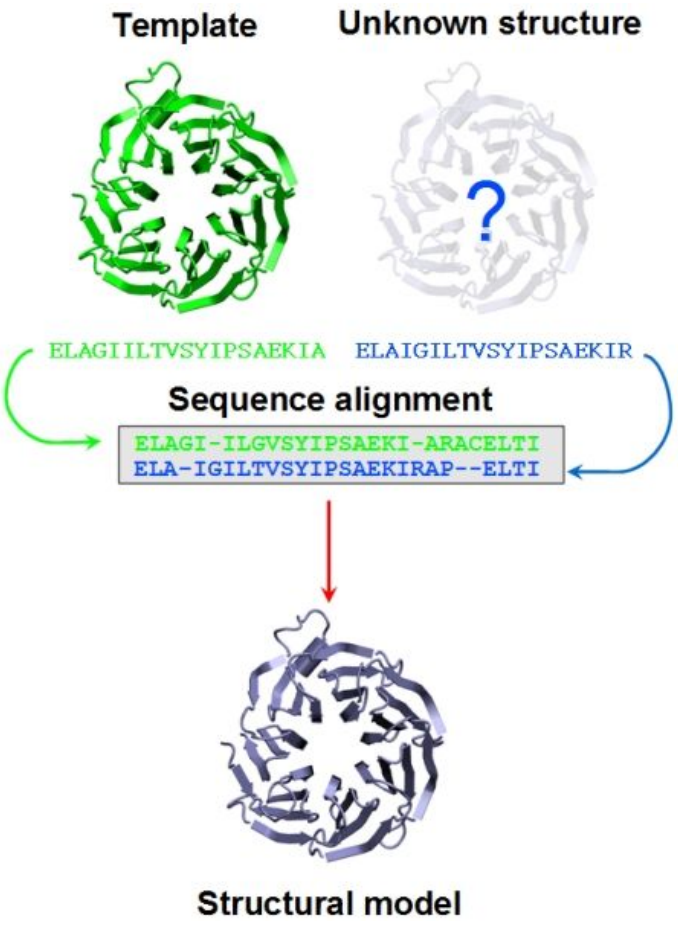


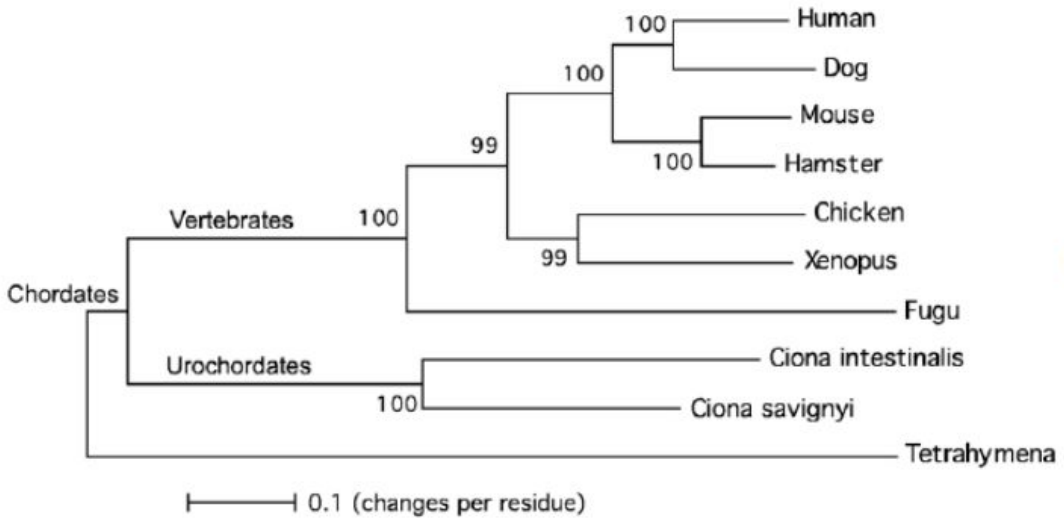
OMA

https://omabrowser.org/oma/home/



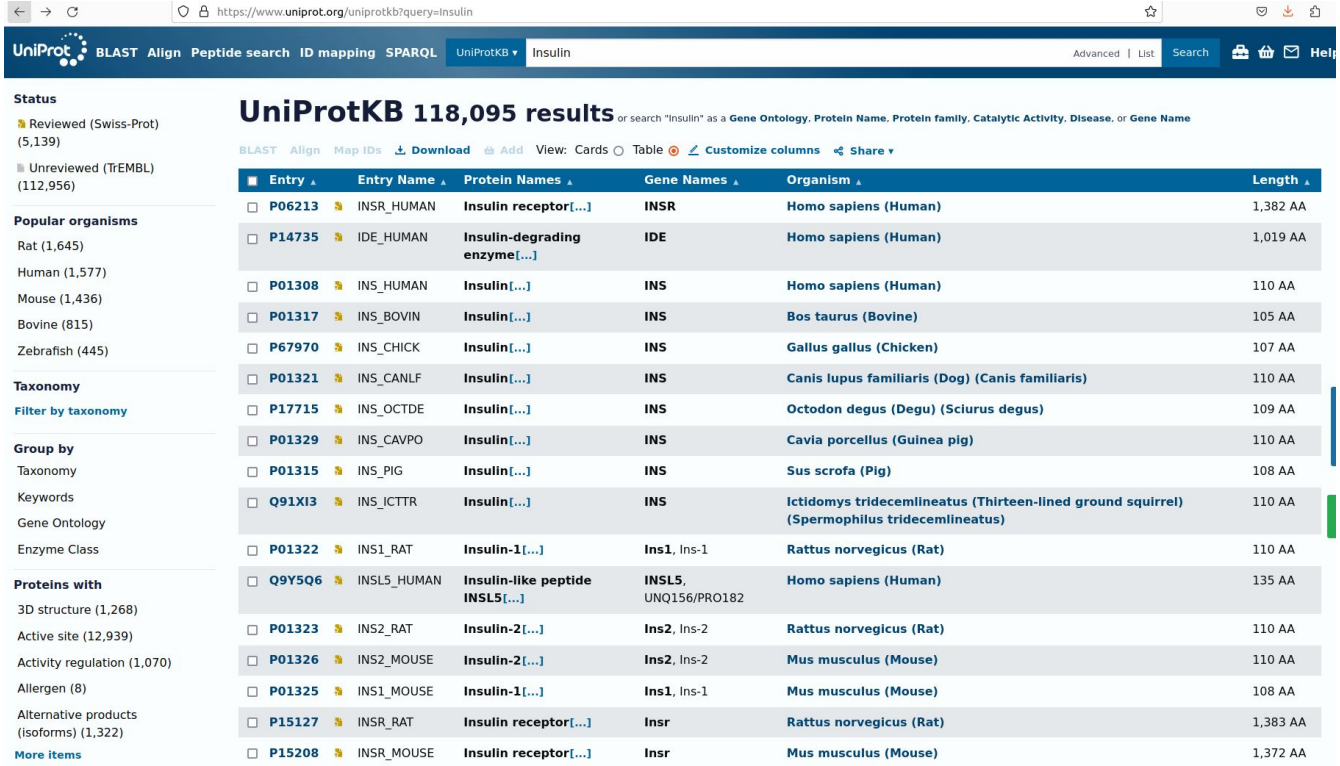






Phylogenetic tree based on the TERT protein sequences. The tree was derived using the neighbor-joining method from the aligned sequences of motif T and the RT domain (1, 2 and A – E) of ten TERTs. Tetrahymena TERT sequence is included as an outgroup to root the tree. The number next to each node indicates a value as a percentage of 1000 bootstrap replicates.

Uniprot https://www.uniprot.org/



https://www.uniprot.org/uniprotkb/P01308/entry

Uniprot

Also contains information about structures (both secondary and tertiary).

https://www.uniprot.org/uniprotkb/P29459/entry

Structure

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
-- Select --		-- Select --				
PDB	1F45	X-ray	2.80 Å	B	23-219	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	3HMX	X-ray	3.00 Å	B	23-219	PDBe · RCSB-PDB · PDBj · PDBsum
AlphaFold	AF-P29459-F1	Predicted			1-219	AlphaFold

Features

Showing features for helix, turn, beta strand.

TYPE	ID	POSITION(S)	DESCRIPTION	
-- Select --				
Helix		43-58	Combined Sources	BLAST Add
Helix		59-61	Combined Sources	BLAST Add
Turn		74-78	Combined Sources	BLAST Add
Helix		81-84	Combined Sources	BLAST Add
Helix		88-92	Combined Sources	BLAST Add
Beta strand		107-109	Combined Sources	BLAST Add

Aligner in Uniprot

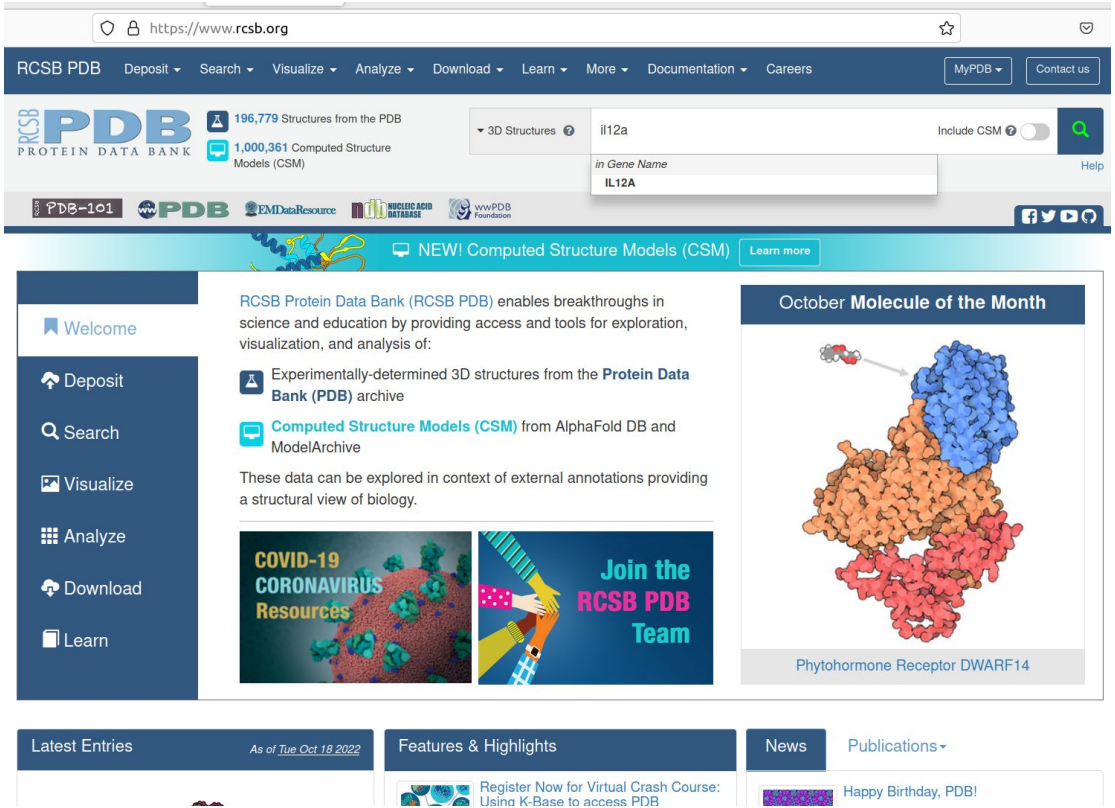
Try to build the sequence alignment using

https://www.uniprot.org/align/

Try highlight amino acids according to their properties.

PDB

https://www.rcsb.org/



https://www.rcsb.org/search?request={"query"%3A{"type"%3A"group"%2C"nodes"%3A[{"type"%3A"group"%2C"nodes"

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | Documentation | Careers | MyPDB | Contact us

Return Structures grouped by No Grouping | Include Computed Structure Models (CSM) | Count | Clear | Search

Search Summary

This query matches 2 **Structures**.

Refinements

Structure Determination Methodology
- experimental (2)

Scientific Name of Source Organism
- Homo sapiens (2)

Taxonomy
- Eukaryota (2)

Experimental Method
- X-RAY DIFFRACTION (2)

Polymer Entity Type
- Protein (2)

Refinement Resolution (Å)
- 2.5 - 3.0 (1)
- 3.0 - 3.5 (1)

Release Date
- 2000 - 2004 (1)
- 2010 - 2014 (1)

Symmetry Type
- Asymmetric (2)

SCOP Classification

-- Tabular Report -- | All | Selected

1 to 2 of 2 **Structures** | Page 1 of 1 | 25 | Sort by ↓ Score

3D View

1F45

Download File | View File

HUMAN INTERLEUKIN-12

Yoon, C., Johnston, S.C., Tang, J., Tobin, J.F., Somers, W.S.

(2000) EMBO J **19**: 3530-3541

Released	2001-06-20
Method	X-RAY DIFFRACTION 2.8 Å
Organisms	Homo sapiens
Macromolecule	INTERLEUKIN-12 ALPHA CHAIN (protein) INTERLEUKIN-12 BETA CHAIN (protein)
Unique branched monosaccharides	MAN, NAG

3D View

3HMX

Download File | View File

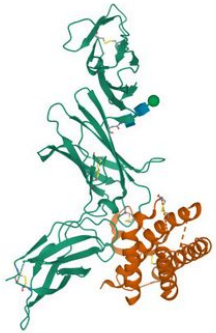
Crystal structure of ustekinumab FAB/IL-12 complex

Luo, J.

(2010) J Mol Biol **402**: 797-812

Released	2010-06-09
Method	X-RAY DIFFRACTION 3 Å
Organisms	Homo sapiens
Macromolecule	Interleukin-12 subunit alpha (protein) Interleukin-12 subunit beta (protein) USTEKINUMAB FAB HEAVY CHAIN (protein) USTEKINUMAB FAB LIGHT CHAIN (protein)
Unique branched monosaccharides	BMA, MAN, NAG

Biological Assembly 1



3D View: Structure | 1D-3D View | Validation Report

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Hetero 2-mer - A1B1

Find Similar Assemblies

Biological assembly 1 assigned by authors and generated by PISA (software)

Macromolecule Content

- Total Structure Weight: 57.89 kDa
- Atom Count: 3,348
- Modelled Residue Count: 415
- Deposited Residue Count: 503
- Unique protein chains: 2

Display Files | Download Files

1F45

HUMAN INTERLEUKIN-12

PDB DOI: 10.2210/pdb1F45/pdb

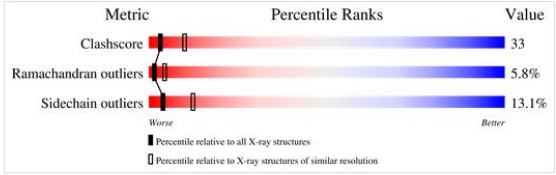
Classification: CYTOKINE/CYTOKINE
Organism(s): Homo sapiens
Expression System: Cricetulus griseus
Mutation(s): Yes

Deposited: 2000-06-07 **Released:** 2001-06-20
Deposition Author(s): Yoon, C., Johnston, S.C., Tang, J., Tobin, J.F., Somers, W.S.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.80 Å
R-Value Free: 0.284
R-Value Work: 0.241
R-Value Observed: 0.241

wwPDB Validation 3D Report | Full Report



This is version 2.1 of the entry. See complete history.

Literature

Download Primary Citation

Charged residues dominate a unique interlocking topography in the heterodimeric cytokine interleukin-12.

Yoon, C., Johnston, S.C., Tang, J., Stahl, M., Tobin, J.F., Somers, W.S.
(2000) EMBO J **19**: 3530-3541

PubMed: 10899108 Search on PubMed | Search on PubMed Central
DOI: 10.1093/emboj/19.14.3530
Primary Citation of Related Structures:

Macromolecules

Find similar proteins by: Sequence (by identity cutoff) | 3D Structure

Entity ID: 1

Molecule	Chains	Sequence Length	Organism	Details	Image
INTERLEUKIN-12 BETA CHAIN	A	306	Homo sapiens	Mutation(s): 0 Gene Names: IL12B, NKS F2	

UniProt & NIH Common Fund Data Resources

Find proteins for P29460 *(Homo sapiens)* | Explore P29460 | Go to UniProtKB: P29460

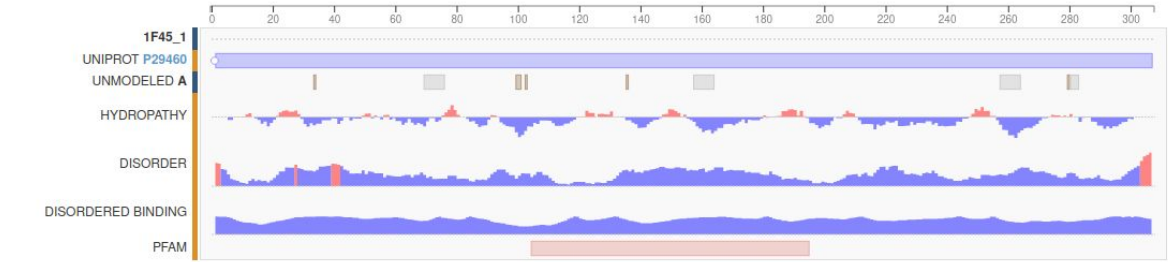
PHAROS: P29460

Entity Groups

Sequence Clusters	30% Identity, 50% Identity, 70% Identity, 90% Identity, 95% Identity, 100% Identity
UniProt Group	P29460

Protein Feature View Expand

Reference Sequence: 1F45_1



Find similar proteins by: Sequence (by identity cutoff) | 3D Structure

Entity ID: 2

Molecule	Chains	Sequence Length	Organism	Details	Image
INTERLEUKIN-12 ALPHA CHAIN	B	197	Homo sapiens	Mutation(s): 1 Gene Names: IL12A, NKS F1	

UniProt & NIH Common Fund Data Resources

Find proteins for P29459 *(Homo sapiens)* | Explore P29459 | Go to UniProtKB: P29459

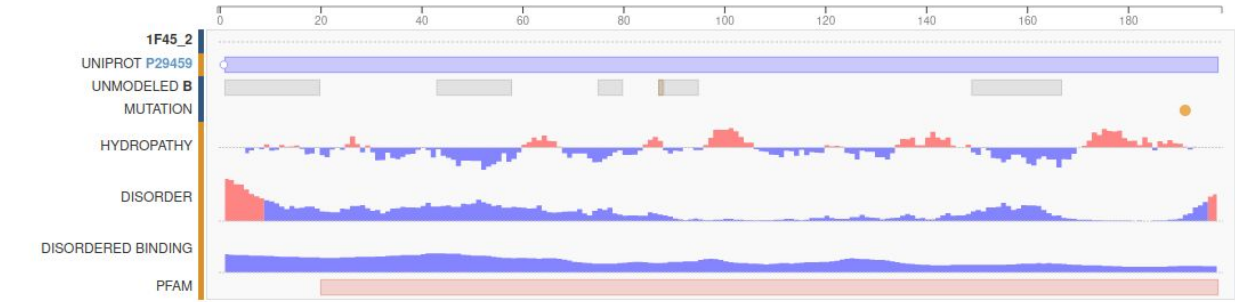
PHAROS: P29459

Entity Groups

Sequence Clusters	30% Identity, 50% Identity, 70% Identity, 90% Identity, 95% Identity, 100% Identity
UniProt Group	P29459

Protein Feature View Expand

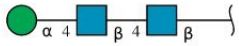
Reference Sequence: 1F45_2



Oligosaccharides

Help

Entity ID: 3

Molecule	Chains	Chain Length	2D Diagram	Glycosylation	3D Interactions
alpha-D-mannopyranose-(1-4)-2-acetamido-2-deoxy-beta-D-glucopyranose-(1-4)-2-acetamido-2-deoxy-beta-D-glucopyranose	C	3		N-Glycosylation	Oligosaccharides Interaction

Glycosylation Resources

GlyTouCan: G62182OO GlyCosmos: G62182OO GlyGen: G62182OO

Experimental Data & Validation

Experimental Data

Method: X-RAY DIFFRACTION
Resolution: 2.80 Å
R-Value Free: 0.284
R-Value Work: 0.241
R-Value Observed: 0.241
Space Group: C 2 2 2_1

Unit Cell:

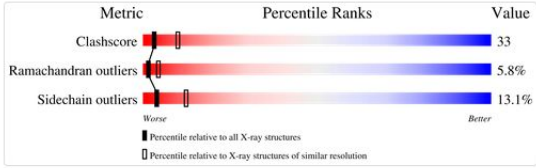
Length (Å)	Angle (°)
a = 112.9	α = 90
b = 154.2	β = 90
c = 101.7	γ = 90

Software Package:

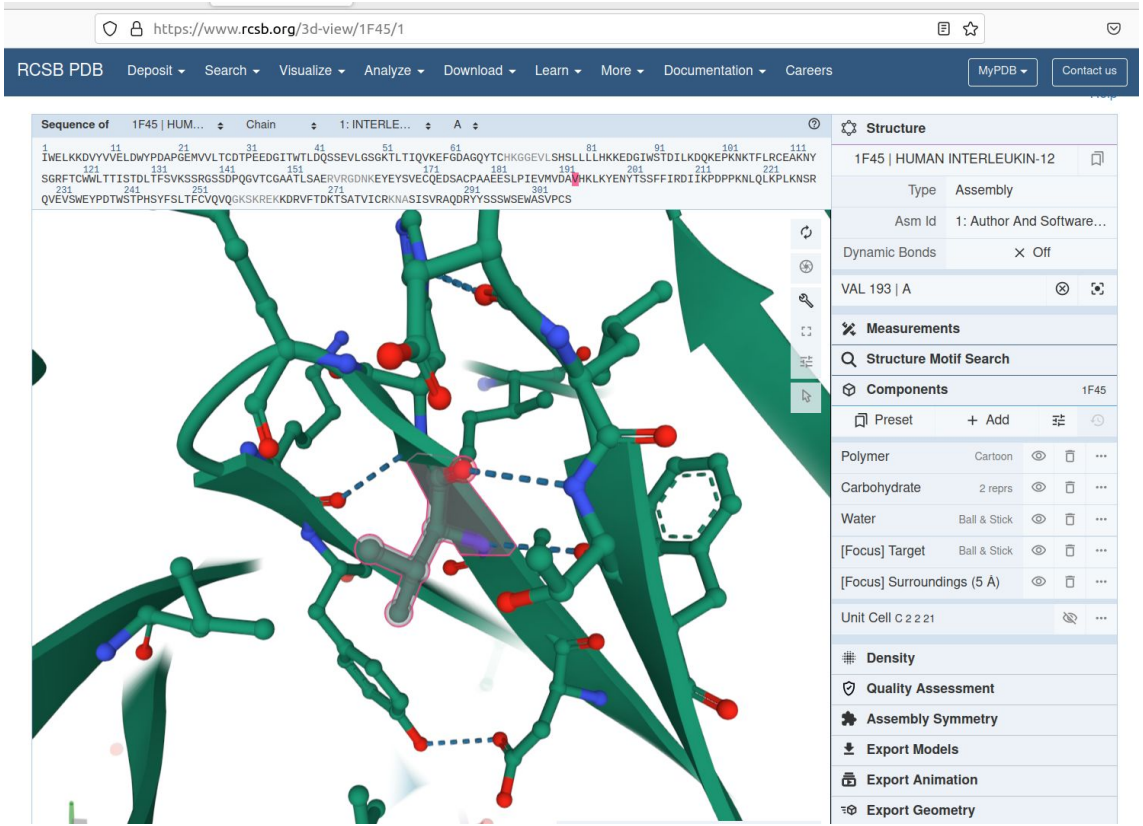
Software Name	Purpose
DENZO	data reduction
SCALEPACK	data scaling
MLPHARE	phasing

Structure Validation

View Full Validation Report



Visualisation of the structure in PDB



https://www.rcsb.org/structure/1F45

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB PROTEIN DATA BANK

210,554 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

3D Structures | Enter search term(s), Entry ID(s), or sequence | Include CSM

Advanced Search | Browse Annotations | Help

PDB-101 wwPDB EMDataResource NAKB wwPDB Foundation PDB-Dev

Structure Summary | 3D View | Annotations | Experiment | Sequence | Genome | Versions

Biological Assembly 1

3D View: Structure | 1D-3D View | Validation Report

Global Symmetry: Asymmetric - C1

Global Stoichiometry: Hetero 2-mer - A1B1

Find Similar Assemblies

Biological assembly 1 assigned by authors and

Display Files | Download Files | Data API

FASTA Sequence

PDBx/mmCIF Format

PDBx/mmCIF Format (gz)

PDB Format

PDB Format (gz)

PDBML/XML Format (gz)

Validation Full PDF

Validation (XML - gz)

Validation (CIF - gz)

Biological Assembly 1 (CIF - gz)

Biological Assembly 1 (PDB - gz)

1F45

HUMAN INTERLEUKIN-12

PDB DOI: https://doi.org/10.2210/pdb1F45/pdb

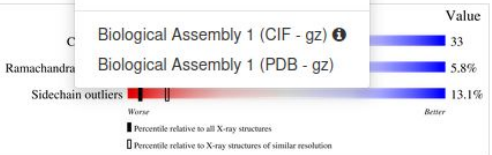
Classification: CYTOKINE/CYTOKINE
Organism(s): Homo sapiens
Expression System: Cricetulus griseus
Mutation(s): Yes

Deposited: 2000-06-07 **Released:** 2001-06-20
Deposition Author(s): Yoon, C., Johnston, S.C., Tang, J., Tobi

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.80 Å
R-Value Free: 0.284
R-Value Work: 0.241
R-Value Observed: 0.241

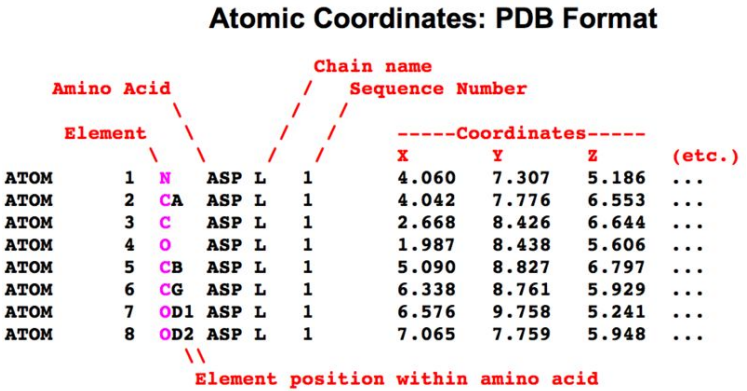
wwPDB Va... | Full Report



This is version 2.1 of the entry. See complete history.

PDB files

pdb file in Moodle




```
ATOM      1  N   ILE A   1      15.200  27.271  13.911  1.00 75.58           N
ATOM      2  CA  ILE A   1      15.336  27.312  15.415  1.00 74.25           C
ATOM      3  C   ILE A   1      16.364  26.299  15.932  1.00 74.22           C
ATOM      4  O   ILE A   1      16.167  25.081  15.787  1.00 73.80           O
ATOM      5  CB  ILE A   1      14.019  26.968  16.088  1.00 73.67           C
ATOM      6  CG1 ILE A   1      14.198  27.038  17.607  1.00 69.85           C
ATOM      7  CG2 ILE A   1      13.515  25.568  15.575  1.00 72.96           C
ATOM      8  CD1 ILE A   1      14.524  28.415  18.088  1.00 70.06           C
ATOM      9  N   TRP A   2      17.456  26.771  16.532  1.00 73.56           N
ATOM     10  CA  TRP A   2      18.424  25.815  17.028  1.00 72.78           C
ATOM     11  C   TRP A   2      19.122  26.165  18.302  1.00 73.51           C
ATOM     12  O   TRP A   2      19.066  27.314  18.764  1.00 71.72           O
ATOM     13  CB  TRP A   2      19.462  25.453  15.954  1.00 72.77           C
ATOM     14  CG  TRP A   2      20.074  26.595  15.228  1.00 70.75           C
ATOM     15  CD1 TRP A   2      19.577  27.231  14.128  1.00 69.67           C
ATOM     16  CD2 TRP A   2      21.316  27.232  15.536  1.00 68.99           C
ATOM     17  NE1 TRP A   2      20.432  28.224  13.735  1.00 68.93           N
ATOM     18  CE2 TRP A   2      21.509  28.247  14.583  1.00 68.07           C
ATOM     19  CE3 TRP A   2      22.287  27.040  16.529  1.00 69.43           C
ATOM     20  CZ2 TRP A   2      22.632  29.072  14.591  1.00 67.70           C
ATOM     21  CZ3 TRP A   2      23.410  27.860  16.539  1.00 67.91           C
ATOM     22  CH2 TRP A   2      23.570  28.866  15.574  1.00 69.87           C
ATOM     23  N   GLU A   3      19.749  25.135  18.882  1.00 75.25           N
ATOM     24  CA  GLU A   3      20.481  25.312  20.118  1.00 77.51           C
ATOM     25  C   GLU A   3      21.976  25.551  19.890  1.00 76.92           C
ATOM     26  O   GLU A   3      22.692  24.756  19.257  1.00 73.77           O
ATOM     27  CB  GLU A   3      20.273  24.132  21.104  1.00 81.21           C
ATOM     28  CG  GLU A   3      20.996  24.363  22.482  1.00 87.97           C
ATOM     29  CD  GLU A   3      20.392  23.601  23.680  1.00 92.55           C
ATOM     30  OE1 GLU A   3      20.278  22.357  23.569  1.00 92.59           O
ATOM     31  OE2 GLU A   3      20.050  24.249  24.727  1.00 92.86           O
ATOM     32  N   LEU A   4      22.397  26.700  20.410  1.00 76.81           N
ATOM     33  CA  LEU A   4      23.762  27.176  20.389  1.00 76.45           C
ATOM     34  C   LEU A   4      24.266  26.466  21.626  1.00 79.78           C
ATOM     35  O   LEU A   4      24.324  25.236  21.636  1.00 83.20           O
ATOM     36  CB  LEU A   4      23.776  28.685  20.602  1.00 73.09           C
ATOM     37  CG  LEU A   4      25.065  29.451  20.395  1.00 71.12           C
ATOM     38  CD1 LEU A   4      25.282  29.683  18.923  1.00 70.86           C
ATOM     39  CD2 LEU A   4      24.983  30.760  21.143  1.00 71.04           C
ATOM     40  N   LYS A   5      24.584  27.210  22.682  1.00 82.29           N
ATOM     41  CA  LYS A   5      25.067  26.593  23.909  1.00 84.30           C
ATOM     42  C   LYS A   5      23.911  26.119  24.800  1.00 86.79           C
ATOM     43  O   LYS A   5      22.765  26.008  24.346  1.00 86.86           O
ATOM     44  CB  LYS A   5      26.032  27.555  24.651  1.00 85.38           C
ATOM     45  CG  LYS A   5      25.568  29.010  24.899  1.00 87.43           C
ATOM     46  CD  LYS A   5      26.757  30.030  25.000  1.00 86.30           C
ATOM     47  CE  LYS A   5      27.351  30.393  23.603  1.00 89.63           C
ATOM     48  NZ  LYS A   5      28.278  31.600  23.494  1.00 87.55           N
ATOM     49  N   LYS A   6      24.216  25.812  26.057  1.00 89.88           N
```

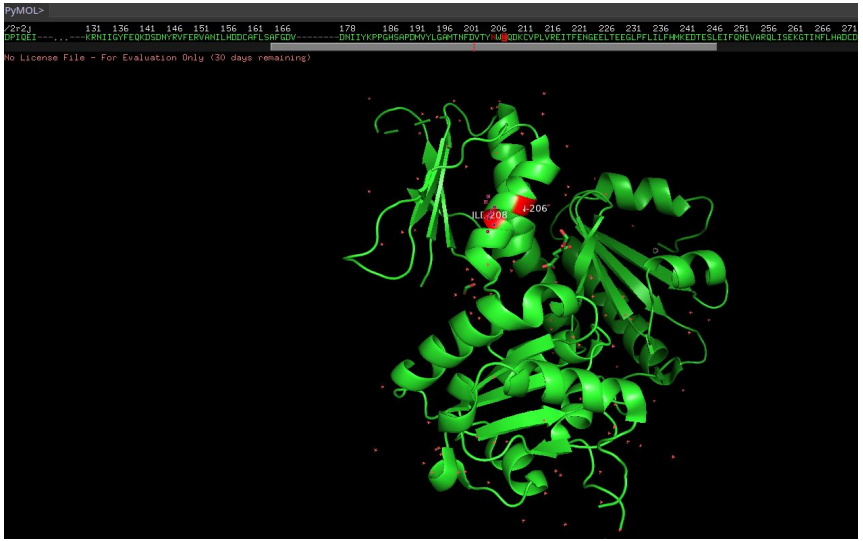
```
HEADER    CYTOKINE/CYTOKINE
TITLE     HUMAN INTERLEUKIN-12
CAVEAT     1F45    MAN C 3 HAS WRONG CHIRALITY AT ATOM C
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: INTERLEUKIN-12 BETA CHAIN;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: RESIDUES 23-328;
COMPND   5 ENGINEERED: YES;
COMPND   6 MUTATION: YES;
COMPND   7 MOL_ID: 2;
COMPND   8 MOLECULE: INTERLEUKIN-12 ALPHA CHAIN;
COMPND   9 CHAIN: B;
COMPND  10 FRAGMENT: RESIDUES 23-219;
COMPND  11 ENGINEERED: YES;
COMPND  12 MUTATION: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606;
SOURCE   5 EXPRESSION_SYSTEM: CRICETULUS GRISEUS;
SOURCE   6 EXPRESSION_SYSTEM_COMMON: CHINESE HAMSTER;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 10029;
```

```
ATOM   2220  SG  CYS A 305      76.079  50.412   8.054  1.00 90.85           S
ATOM   2221  N   SER A 306      77.917  52.118  12.119  1.00 95.38           N
ATOM   2222  CA  SER A 306      79.227  52.520  12.671  1.00 96.65           C
ATOM   2223  C   SER A 306      79.673  51.582  13.814  1.00 96.08           C
ATOM   2224  O   SER A 306      80.107  50.441  13.585  1.00 94.51           O
ATOM   2225  CB  SER A 306      79.183  53.994  13.175  1.00 95.97           C
ATOM   2226  OG  SER A 306      78.801  54.924  12.150  1.00 93.46           O
TER    2227      SER A 306
ATOM   2228  N   GLN B  20      74.675  21.716  -7.562  1.00 82.42           N
ATOM   2229  CA  GLN B  20      74.207  22.987  -7.019  1.00 82.60           C
ATOM   2230  C   GLN B  20      73.632  23.858  -8.102  1.00 81.44           C
ATOM   2231  O   GLN B  20      72.890  24.813  -7.826  1.00 81.70           O
ATOM   2232  CB  GLN B  20      75.316  23.736  -6.284  1.00 85.50           C
ATOM   2233  CG  GLN B  20      75.590  23.175  -4.859  1.00 92.68           C
ATOM   2234  CD  GLN B  20      75.465  24.236  -3.748  1.00 96.11           C
ATOM   2235  OE1 GLN B  20      76.089  25.302  -3.826  1.00 98.19           O
ATOM   2236  NE2 GLN B  20      74.662  23.939  -2.710  1.00 96.67           N
```

```
SSBOND   1 CYS A   28    CYS A   68                          1555   1555  2.04
SSBOND   2 CYS A  109    CYS A  120                          1555   1555  2.04
SSBOND   3 CYS A  148    CYS A  171                          1555   1555  2.04
SSBOND   4 CYS A  177    CYS B   74                          1555   1555  2.06
SSBOND   5 CYS A  278    CYS A  305                          1555   1555  2.04
SSBOND   6 CYS B   42    CYS B  174                          1555   1555  2.07
SSBOND   7 CYS B   63    CYS B  101                          1555   1555  2.05
```

Visualisation of 3D protein structures

- In Uniprot
- In PDB
- PyMol (should be installed locally)
- EzMol
  http://www.sbg.bio.ic.ac.uk/~ezmol/



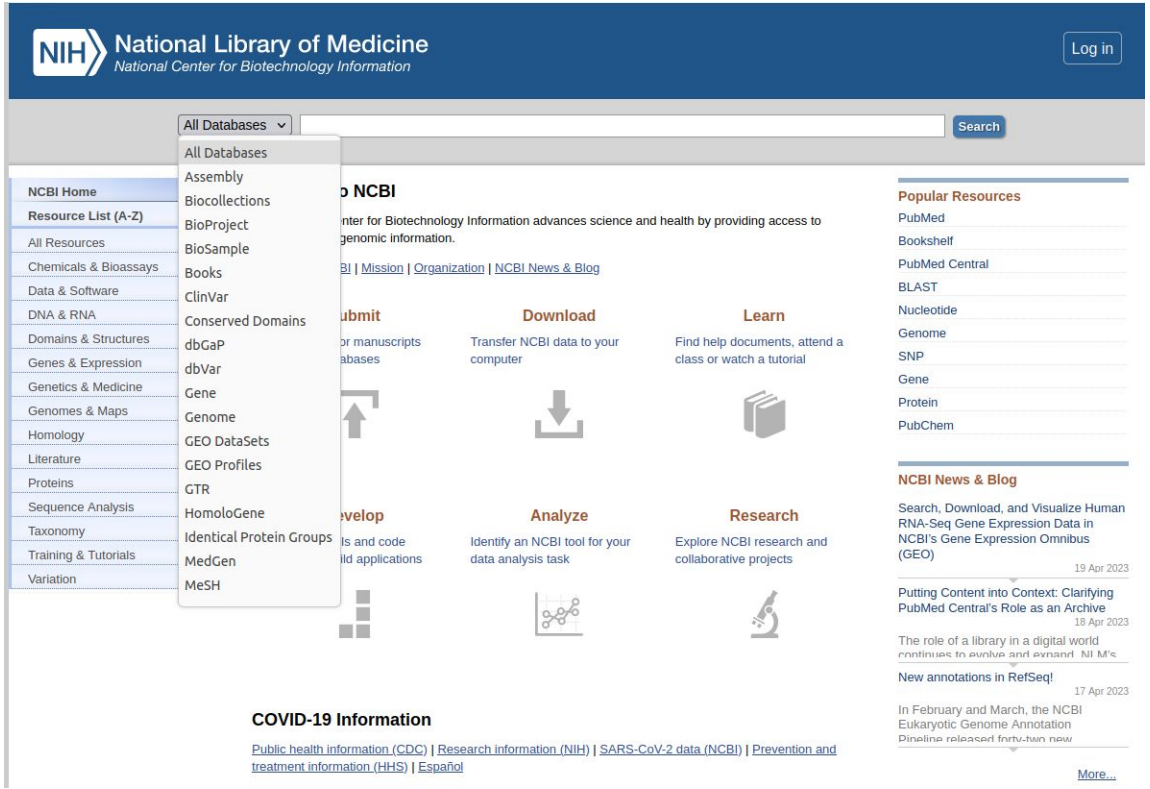
Human ERp44, PDB 2R2J

Pymol

https://pymol.org/2/

NCBI RefSeq

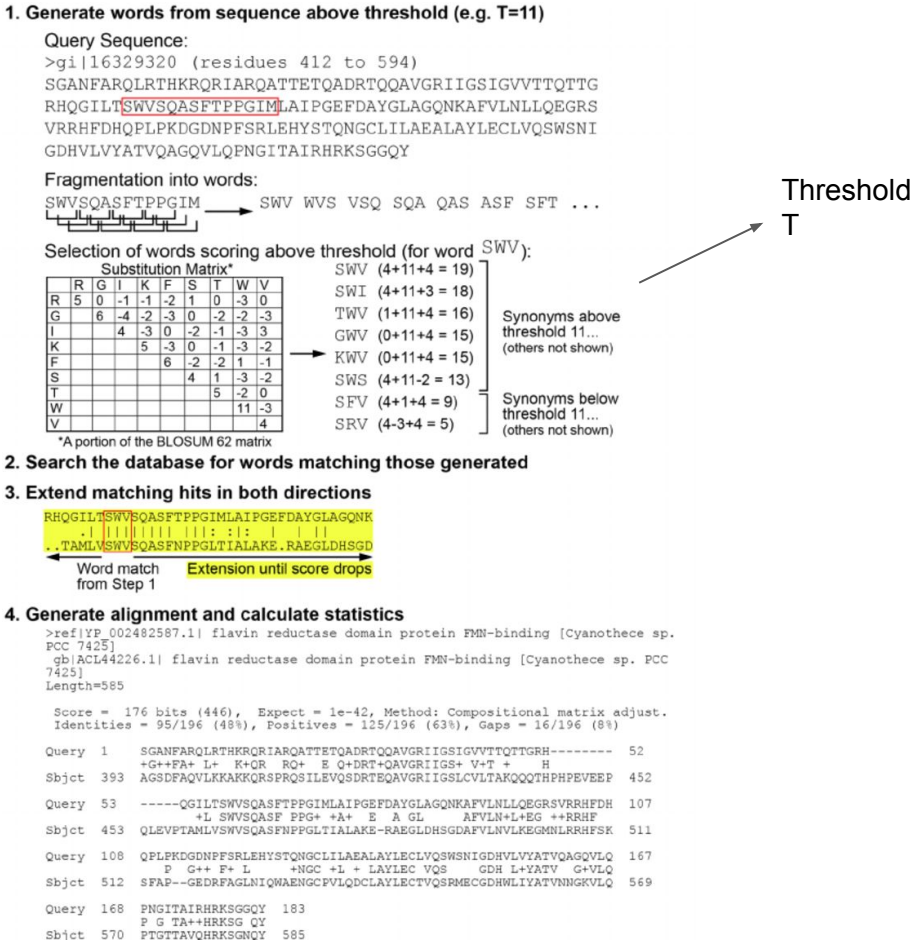
https://www.ncbi.nlm.nih.gov/



BLAST

an algorithm to search for the similar gene or protein sequences in the databases (RefSeq, Uniprot,...)

It is too long to build local alignments with Smith-Waterman algorithm with each sequence from the big database



Kerfeld CA, Scott KM (2011) Using BLAST to Teach "E-value-tionary" Concepts. PLoS Biol 9(2):e1001014. doi:10.1371/journal.pbio.1001014

BLAST https://blast.ncbi.nlm.nih.gov/Blast.cgi

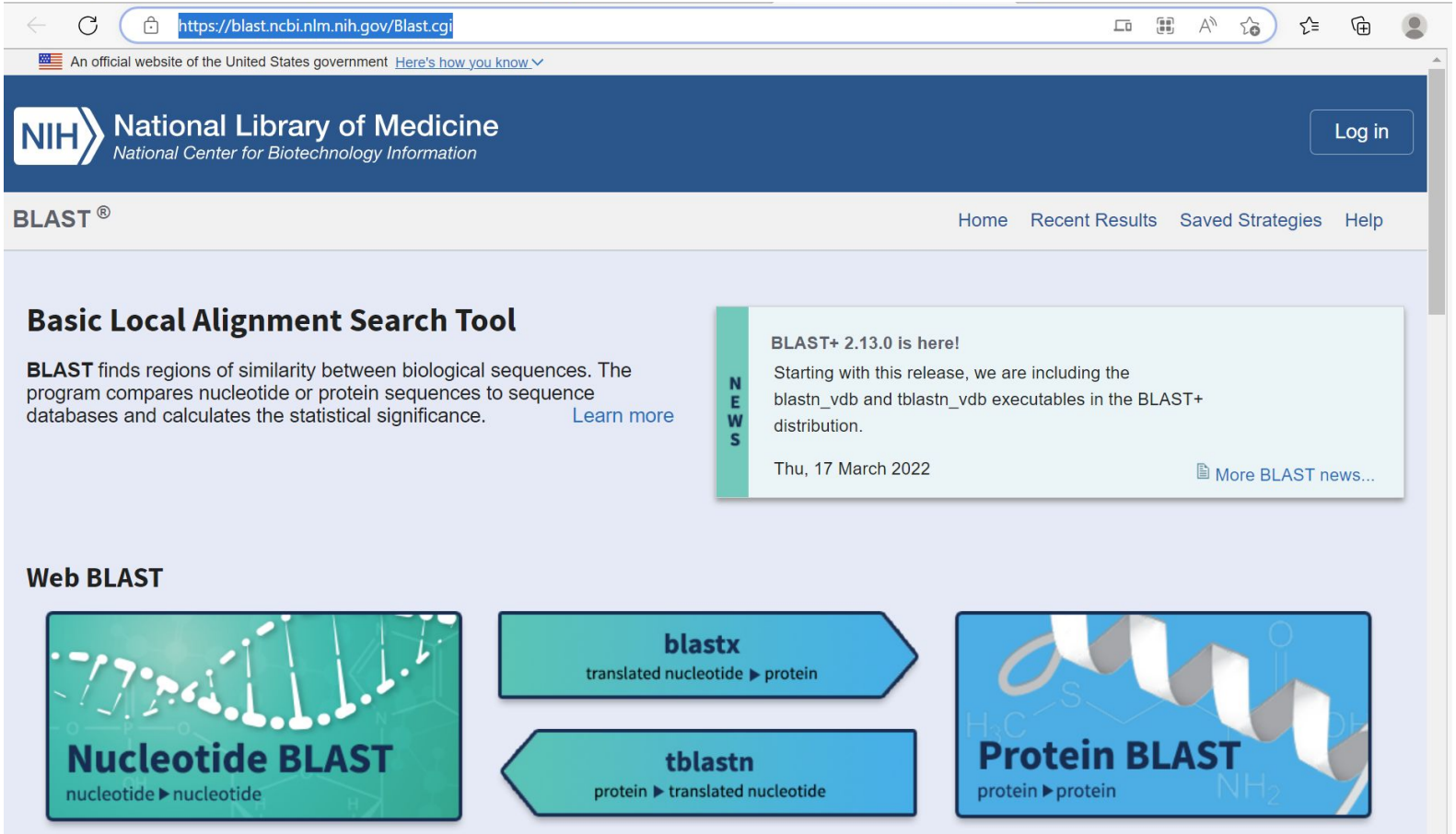
Try to perform BLAST search using the following sequence

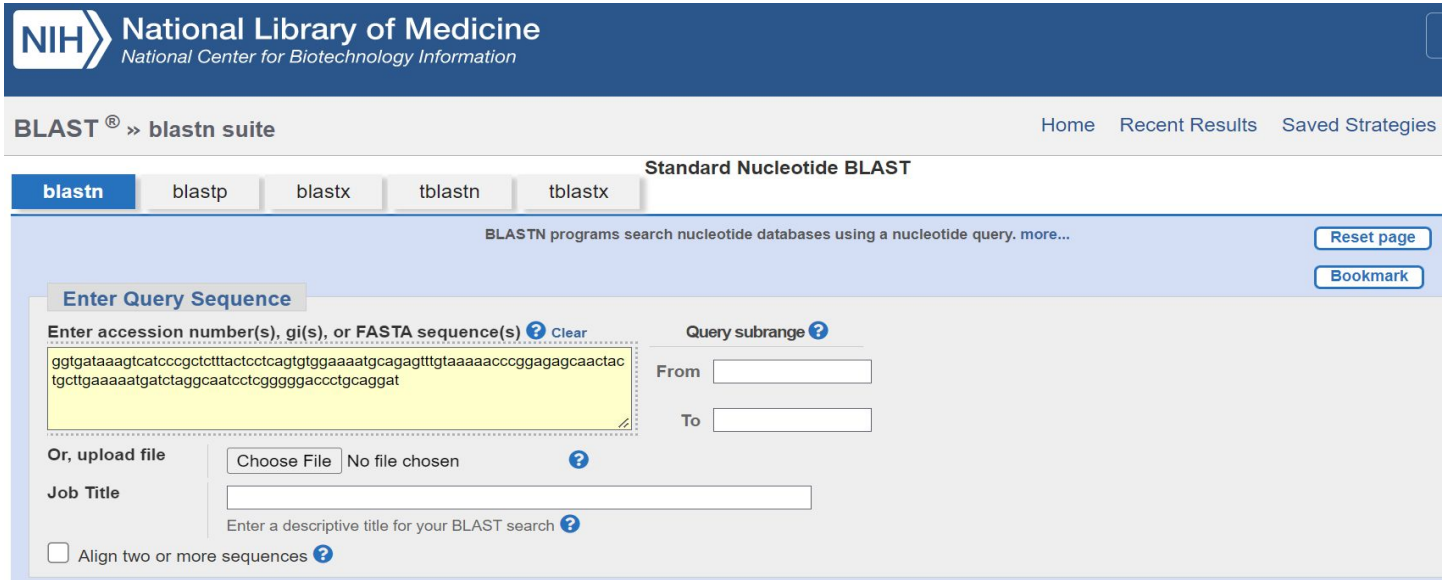
>gene
ggtgataaagtcatcccgctctttactcctcagtgtggaaaatgcagagtttgtaaaaacccggagagcaactactgcttgaaaaatgatctaggcaatcct
cgggggaccctgcaggat

Try to use different

- BLAST tools (blastn, blastx, ...)
- databases for the search
- parameters of the search

Analyze the results of BLAST search. What results are significant? What information can we get about the sequence?


https://blast.ncbi.nlm.nih.gov/Blast.cgi
An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST ®
Home Recent Results Saved Strategies Help
Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more
NEWS
BLAST+ 2.13.0 is here!
Starting with this release, we are including the blastn_vdb and tblastn_vdb executables in the BLAST+ distribution.
Thu, 17 March 2022
More BLAST news...
Web BLAST
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein


National Library of Medicine
National Center for Biotechnology Information
BLAST ® » blastn suite
Home
Recent Results
Saved Strategies
Standard Nucleotide BLAST
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
ggtgataaagtcatcccgctctttactcctcagtgtggaaaatgcagagtttgtaaaaacccggagagcaactac
tgcttgaaaaatgatctaggcaatcctcgggggacccctgcaggat
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 — Standard databases (nr etc.): rRNA/ITS databases | Genomic + transcript databases | Betacoronavirus

Nucleotide collection (nr/nt)

Organism (Optional) — Enter organism name or id--completions will be suggested — exclude — Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional) — Models (XM/XP) — Uncultured/environmental sample sequences

Limit to (Optional) — Sequences from type material

Entrez Query (Optional) — YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for

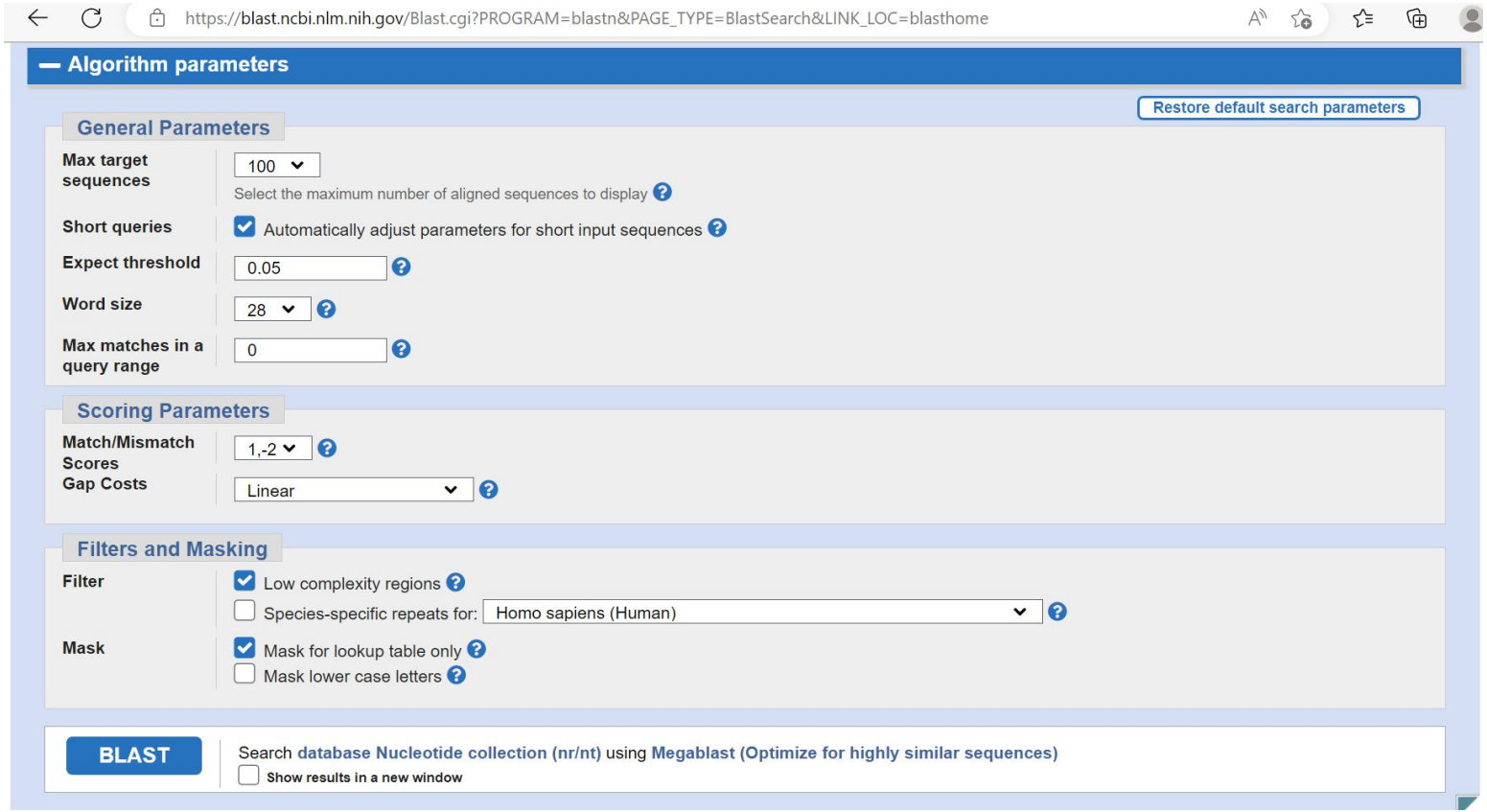
- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

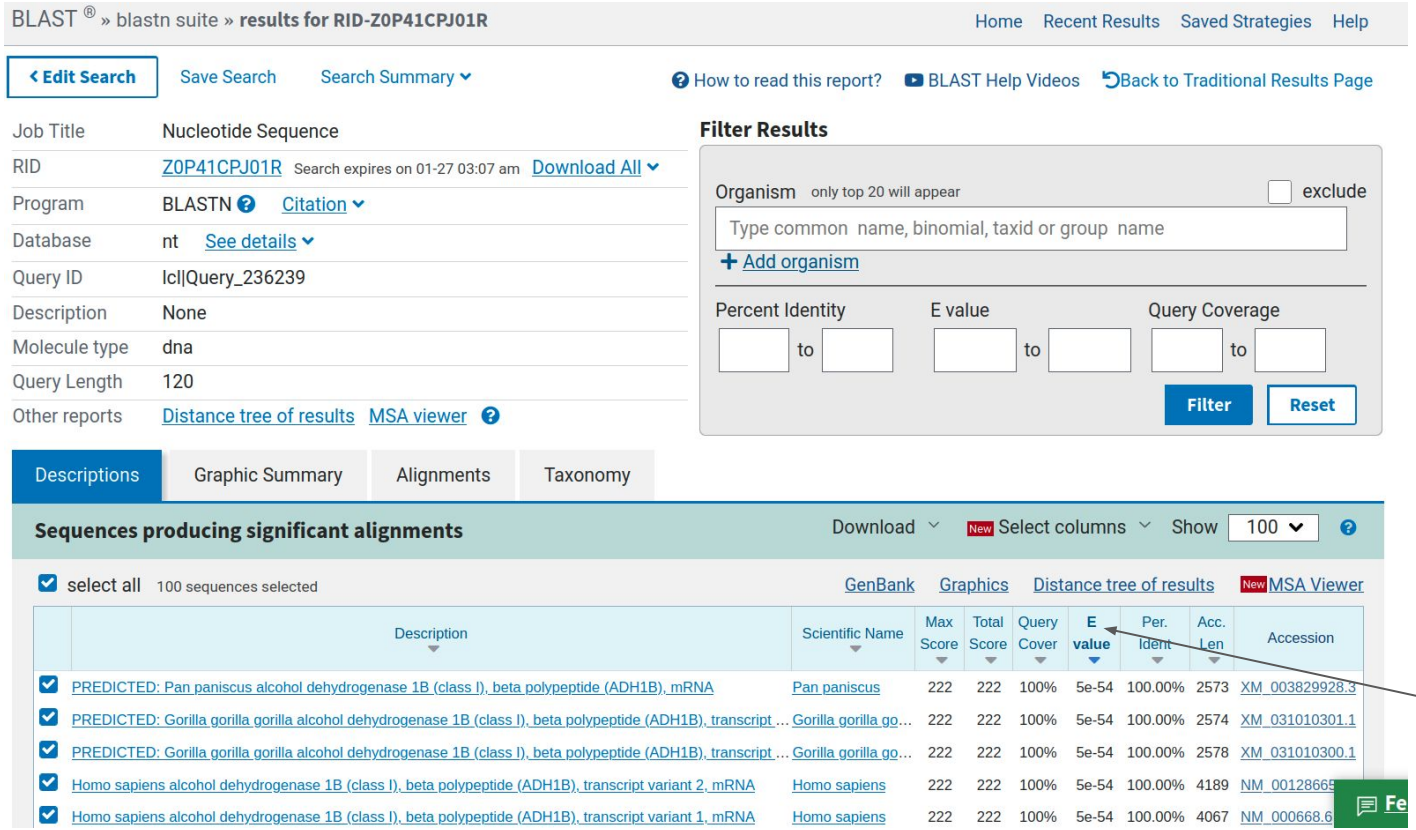
BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

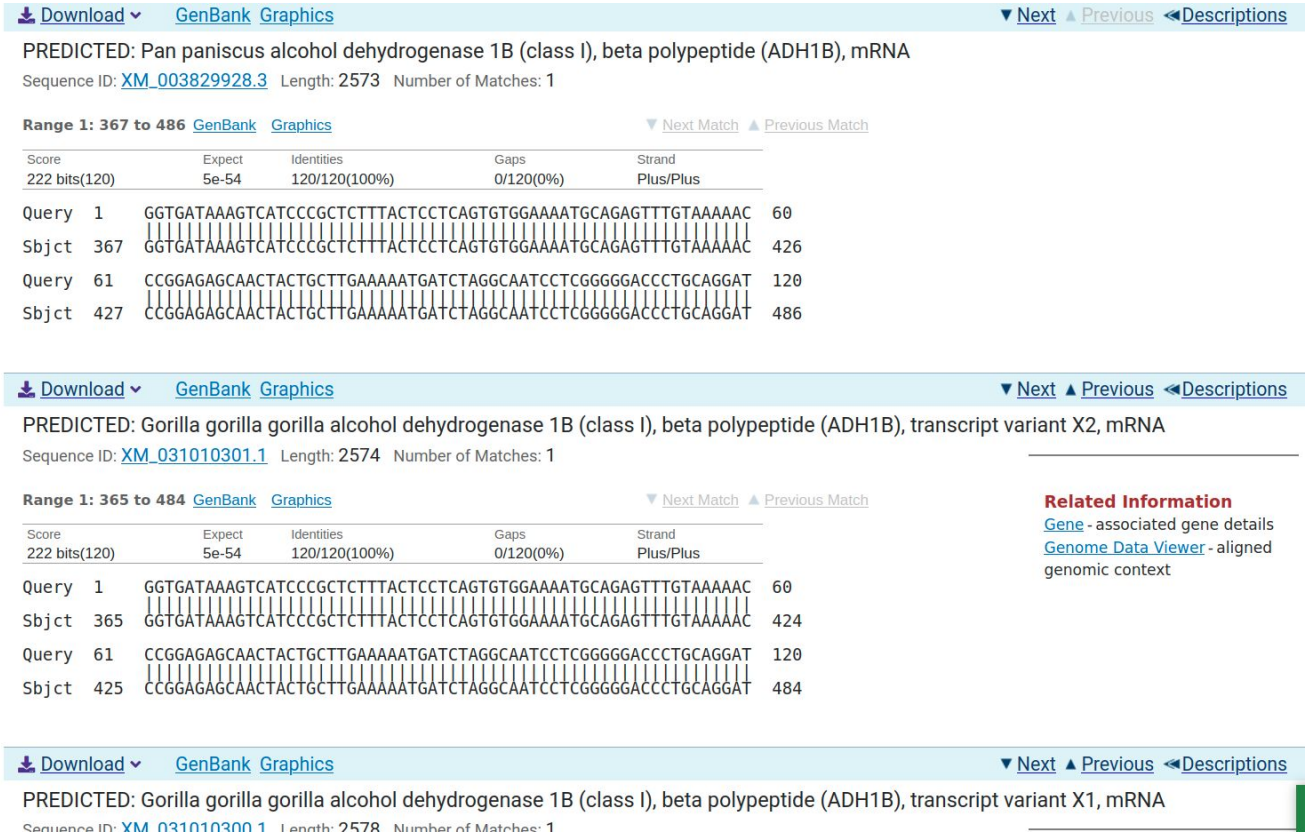

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Algorithm parameters
Restore default search parameters
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters for short input sequences
Expect threshold
0.05
Word size
28
Max matches in a query range
0
Scoring Parameters
Match/Mismatch Scores
1,-2
Gap Costs
Linear
Filters and Masking
Filter
Low complexity regions
Species-specific repeats for:
Homo sapiens (Human)
Mask
Mask for lookup table only
Mask lower case letters
BLAST
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
Show results in a new window

Results



E-value for significant hits usually chosen to be less than 0.001

Results



Download GenBank Graphics ▼ Next ▲ Previous ◀◀ Descriptions

PREDICTED: Pan paniscus alcohol dehydrogenase 1B (class I), beta polypeptide (ADH1B), mRNA

Sequence ID: XM_003829928.3 Length: 2573 Number of Matches: 1

Range 1: 367 to 486 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
222 bits(120)	5e-54	120/120(100%)	0/120(0%)	Plus/Plus

```
Query  1    GGTGATAAAGTCATCCCGCTCTTTACTCCTCAGTGTGGAAAATGCAGAGTTTGTAAAAAC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  367  GGTGATAAAGTCATCCCGCTCTTTACTCCTCAGTGTGGAAAATGCAGAGTTTGTAAAAAC  426

Query  61   CCGGAGAGCAACTACTGCTTGAAAAATGATCTAGGCAATCCTCGGGGGACCCTGCAGGAT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  427  CCGGAGAGCAACTACTGCTTGAAAAATGATCTAGGCAATCCTCGGGGGACCCTGCAGGAT  486
```

Download GenBank Graphics ▼ Next ▲ Previous ◀◀ Descriptions

PREDICTED: Gorilla gorilla gorilla alcohol dehydrogenase 1B (class I), beta polypeptide (ADH1B), transcript variant X2, mRNA

Sequence ID: XM_031010301.1 Length: 2574 Number of Matches: 1

Range 1: 365 to 484 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
222 bits(120)	5e-54	120/120(100%)	0/120(0%)	Plus/Plus

```
Query  1    GGTGATAAAGTCATCCCGCTCTTTACTCCTCAGTGTGGAAAATGCAGAGTTTGTAAAAAC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  365  GGTGATAAAGTCATCCCGCTCTTTACTCCTCAGTGTGGAAAATGCAGAGTTTGTAAAAAC  424

Query  61   CCGGAGAGCAACTACTGCTTGAAAAATGATCTAGGCAATCCTCGGGGGACCCTGCAGGAT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  425  CCGGAGAGCAACTACTGCTTGAAAAATGATCTAGGCAATCCTCGGGGGACCCTGCAGGAT  484
```

Related Information

Gene - associated gene details

Genome Data Viewer - aligned genomic context

Download GenBank Graphics ▼ Next ▲ Previous ◀◀ Descriptions

PREDICTED: Gorilla gorilla gorilla alcohol dehydrogenase 1B (class I), beta polypeptide (ADH1B), transcript variant X1, mRNA

Sequence ID: XM_031010300.1 Length: 2578 Number of Matches: 1

Properties of amino acid residues

https://www.genome.jp/aaindex/