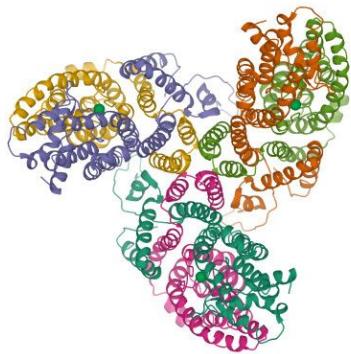


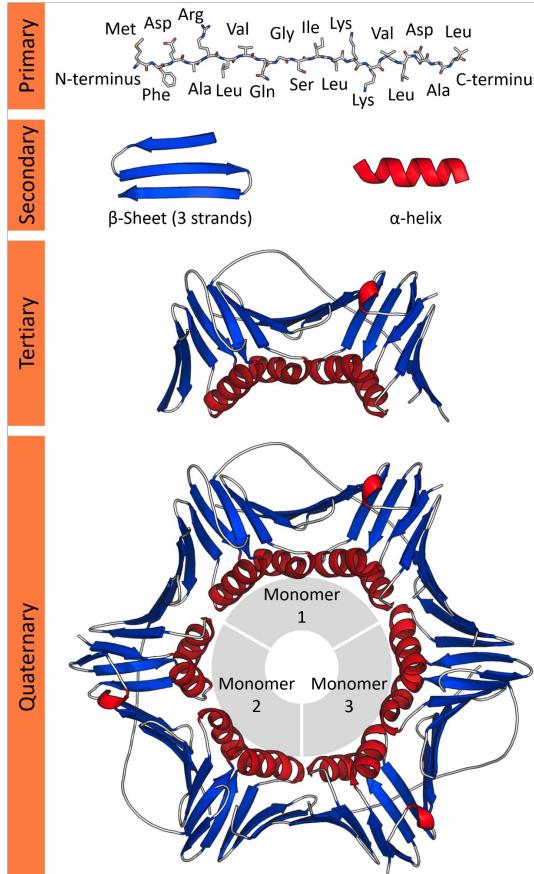
# Structural Bioinformatics

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



## Online tools for pairwise alignments

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

## Local Alignment

Target Sequence  
5' ACTACTAGATTACTACGGATCAGGTACTTAGAGGCTTCAACCA 3'  
||||| ||||| ||||| ||||| |||||  
Query Sequence 5' TACTCACGGATGAGGTACTTAGAGGC 3'

## Global Alignment

Target Sequence  
5' ACTACTAGATTACTACGGATCAGGTACTTAGAGGCTTCAACCA 3'  
||||| ||||| ||||| ||||| |||||  
5' ACTACTAGATT----ACGGATC--GTACTTAGAGGCTAGCAACCA 3'  
Query Sequence

## Multiple Sequence Alignment (MSA)

Hsa\_TMEM66      ALTLHYDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 103  
Ptr\_TMEM66      ALTLHYDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 103  
Ppy\_TMEM66      ALTLHYDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 103  
Mml\_TMEM66      ALTLHYDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 103  
Mfa\_TMEM66      ALTLHYDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 103  
Mne\_TMEM66      ALTLHYDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 103  
Ssc\_TMEM66      ALTLHYDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 103  
Bta\_TMEM66      ALTLHYDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 103  
Cfa\_TMEM66      ALTLHYDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 103  
Mmu\_TMEM66      ALTLYSDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 104  
Rno\_TMEM66      ALTLYSDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 104  
Ocu\_TMEM66      ALTLHYDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 97  
Laf\_TMEM66      ALTLHYNDRYTTSSRLDPIPQLKCVGGTAGCDSYTPKVIQCNKGWDGYDVQWECKTDLDI 89  
Mdo\_TMEM66      ALTLHRDRYTTTARRTAPIPQLQCLGGSAGCPAHIEPIEVQCRNKGWDGFDVQWECKEKAELDT 119  
Gga\_TMEM66      VLTLHRGRYTTTARRTAAPVQLOCIGGSAAGCS-DIPEVVQQCYNRGWGDGYDVQWQCKADLEN 94  
Xla\_TMEM66      TITLYADRYTNARRSAPVPQLKCIIGGNAGCHAMVPQVQQCHNRGWGLDVQWECKRVDMDN 93  
Xtr\_TMEM66      AITLYADRYTNARRSAPVPQLKCIIGGSAGCHTMVPQVQQCHNRGWDFDVQWECKVDMND 93  
Dre\_TMEM66      VLTLYGRYTTTARRSSPPVQLQCIGGSAAGCSFTPEVQQCYNRGSDGIDAQWECKADMDN 93  
Ssa\_TMEM66      VLTLYKGKYTTTARRSSAQPVQLQCVGGSAGCGSFPIPEVQQCKNKKGWDGVDAQWECKTDMDN 93  
Tru\_TMEM66      VLTLYGRYTTTARRSSPPVQLQCVGGSAGCGSFPIPEVPEVQQCONKGWDGMDI1QWECKRTDMND 99  
Tri\_TMEM66      TLTLYGRYTTTARRSSPPVQLRCVGGSAGCGQAFVPEVQQCONRGWGDGVDVQWECKTDMDN 89  
Gac\_TMEM66      ALTLTKRNYTTTARRASPPVQLQCVGGSAGCGQAFVPEVQQCONKGWDGVDVQWECKRTDMND 92  
Ppr\_TMEM66      VLTLYKGRYTTTARRSSPPVQLQCVGGSAGCGSFPEVQQCYNRGSDGIDTQWECKADMDN 93  
Cel\_TMEM66      AITLHKGKMTTGRRVSPTFQLKCVGG-SAKGAFTPKVWQQCANQGFDGSDDVQWRCDADLPH 96  
Cre\_TMEM66      AITLNGKGMTTGRRVAPTLQQLKCVGG-SAKGAFTPKVWQQCSNQGFDGSDDVQWRCDADLPH 96  
Cbr\_TMEM66      AITLHKGKMTTGRRVAPALQQLKCVGG-SAKGQFSPKVWQQCANQGFDGSDDVQWRCDADLPH 96

# 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  
PVIAGAGSNNTDEAIELALHAQEAGADALLVTPYYNKPTQKGLFAHFSAVAЕAVKLPIVIYNIPPRSVVDM  
SPETMGALVKAHKNIIGVKDATGKLDKVSEQRISCGKDFVQLSGEDGTALGFNAHGGVGCISVTANVAPR  
LCSEFQAAMLAGDYAKALEYQDRLMPLHRAIFMEPGVCGTKYALKSTRGGNRRVRSPLMSTLEPATEAAI  
DAALKHAGLMN

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

MTTVGFDVAARLGTLTAMVTPFSGDGS LTD TATAARLANHLVDQGCDGLVVSGTTGESPTTDGEKIELL  
RAVLEAVGDRARVIAGAGTYDTAHSIRLAKACAAEGAHHGLLVTPYYSKPPQRGLQAHFTAVADATELPM  
LLYDIPGRSAVPIEPDTIRALASHPNIVGVVKDAKLHSGAQIMADTGLAYYSGDDALNLPWLAMGATGFIS  
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:  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:  Q8UGI3.fasta

OUTPUT FORMAT

pair

MATRIX	GAP OPEN	GAP EXTEND	END GAP PENALTY	END GAP OPEN	END GAP EXTEND
BLOSUM62	10	0.5	false	10	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
# -db
# -stdout
# -asequence emboss_needle-I20221115-173525-0089-79876523-p1m.aupfile
# -bsequence emboss_needle-I20221115-173525-0089-79876523-p1m.bupfile
# -datafile EBLOSUM62
# -gapopen 0.0
# -gapextend 0.5
# -endopen 10.0
# -endextend 0.5
# -eformat3 pair
# -sprotein1
# -sprotein2
# Alignments: 1
# 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 MTTVGFDAARLGLLTAMTPFSGDGSLDTATAARLANHLVDQGCDGLV	50
DAPA_AGRFC	1 -----MFKGSIPALITPFTDNGSDEKAFAAHVEWQIAEGSNGLV	40
DAPA_MYCTU	51 VSGTTGESPTTDGEKIELLRAVLEAVGDRARVIAGAGTYDTAHSIRLAK	100
DAPA_AGRFC	41 PVGTTGESPTLSHDEHKRVVELCIEAAKRVPVIAAGAGSNNTDEAIELAL	90
DAPA_MYCTU	101 ACAAEAGHGLVVTPYYSKPPQRGLQAHFTAVALIDATELPMLLYDIPGRSA	150
DAPA_AGRFC	91 HAQEAGADALLVVTPTYNKPTQKGLFAHFSAVAEAVKLPIVIYNIPPRSV	140
DAPA_MYCTU	151 VPIEPDTIRALA-SHPNIVGVKDA--KADLHSGAQIMADTGLAYSGDDA	197
DAPA_AGRFC	141 VDMSPETM GALVKAHKNIIGVKDATGKLD RVSEQRISCGKDFVQLSGEDG	190
DAPA_MYCTU	198 LNLPWLAMGATGFISVIAHLLAAGQLRELLSAFGSGDIATARKINIAVAPL	247
DAPA_AGRFC	191 TALGFAHGGVGVCISVTANVAPRLCSEFQAAMLAGDYAKALEYQDRLMPL	240
DAPA_MYCTU	248 CNAMSRLGGVTLSKAGL-RLOQIDVGDPRL--PQVA---ATPEQIDALA	290
DAPA_AGRFC	241 HRAIFMEPGVCGTKYALKSTRG--GNRRVRSPLMSTLEPATEAAIDAA-	286
DAPA_MYCTU	291 ADMRAASVLR 300 ::: ...:	
DAPA_AGRFC	287 --LKHAGLMN 294	

#-----  
#-----

# Calculate the identity of the pairwise alignment

Should be P

HG - - IHQCBTLL -  
HGIGFHKLTGTLL

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

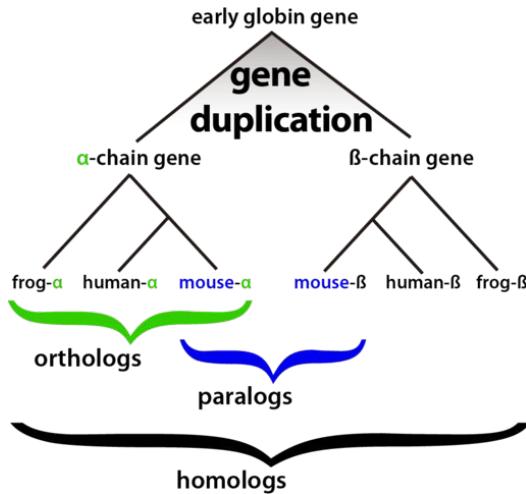
<https://www.ebi.ac.uk/Tools/msa/clustalo/>

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

### CLUSTAL 0(1.2.4) multiple sequence alignment

GADM003540	------MLY-----LEGLTAA-----YSSEQYTNLGLL--	22
ANGAN33611	MTEQEGTGALE---YSLSQDETFOQALWNTVTV-----PTEDEGWGSLDNVSV	44
ORYLA07425	-----WESMDPVPDFLPESOGSFGQELWETVYP--PLELTSLP--TVNEPTGSWA-----	45
ASTCA27703	MAE-QGFEDLQLSQNLPSQNSFKELVAESLSVSLDTTSTVNSNTGHDETWRP-----*	53
GADM003540	-NSMD-QS1QNGGSTSTSPTYNNDHAQN--NVAAPSPYSQPSSTFDAL-SPSPAIPSNTD	76
ANGAN33611	PDALNCDFHLQFQQYLVGSPQPLQDGDFEKLFDVPATPAPPQPSGSALDGGAPPASTVPTTTD	104
ORYLA07425	-----TGDM---FLLDQDLSGTFDDKIFDPIEPVPT-----NEVNPPPTTVPVTTD	89
ASTCA27703	-----DDTMEMFQLINEQQLAENFGENLFEPLPNILQK-----DGITPAASTVPTVTTD	101
GADM003540	: : . :	136
ANGAN33611	YAGPHTFDVSFQOSSTAKSATWTYSTELKKLYCQIAKTCPIQIKVLTNPQQGAVIRAMPV	164
ORYLA07425	YPGEFGFLRLRFQKSSTAKSVCTCTYSPLENKLFCLAKTCVPMQVTHTPPPPQAVLRAV	149
ASTCA27703	YPGSCYEELRFQKGTAKSCTVSTSETLRNLKLYCQAKTSPIEVRSKEPPKGAILRATV	161
GADM003540	YPGECGFQLRFQKSQTAKSCTVSTYEVLNKLYCQAKTSPIEVLVSKEPPQGAVLRAV	136
ANGAN33611	* * :	164
ORYLA07425	YKKAEHVTEVVKRCPNHELSREFNDGQIAPPSHLIRVEGNSHQAQYMEDSITGRQSVLVPY	204
ASTCA27703	YKKSEHVAEVVRCPHERTSENN-DDSTPRSHLIRVEGSQRRAWYTEDSNTLRHSVLPY	223
GADM003540	YKKTEHVADVVRCPHHQN---E-DSVHRSHLIRVEGSQLAQYFEDPYTKRQSVTVPY	204
ANGAN33611	YKKTEHVADVVRCPHHQN---E-DPVEHRSHLIRVEGNQRQYFEDLHTRKQSVTVPY	216
ORYLA07425	*****:****:****:****:****:****:****:****:****:****:****:****:	216
ASTCA27703	*****:****:****:****:****:****:****:****:****:****:****:****:****:	196
GADM003540	EPPQVGTEFTTILYNFMCNSSCVGGMNRPIIIVTLETRDGQVLGRRCFEARICACPGR	256
ANGAN33611	EPPQLGSECTTVLYNFMCNSSCMGGMNRRPILTILETQEGOVLGRRSFEVRCACPGR	283
ORYLA07425	EPPQPQGSEMTILLSYMCNSSCMGGMNRRPILTILLET-EGLVLGRRCFEVRICACPGR	263
ASTCA27703	EPPQLLGSEMTILLSFMCNSSCMGGMNRRPILTILLETPEGLVLGRRCFEVRVACPGR	276
GADM003540	*****:****:****:****:****:****:****:****:****:****:****:****:****:	256
ANGAN33611	DRKADEDSIRKQQVTDVTKSSDAFRQVS-----HGLQMSMKKRRTSTDEEVFCL	305
ORYLA07425	DRKTEEDNHRKQQEKGTKTGGAKRTFKESSLPSSQPED---SKKTKTTSNEEEIFTL	340
ASTCA27703	DRKTEEEERSQKTOPPKRS-----ML---EVTPNTSSSKRKKSNSHSGEEEDNREVHF	312
GADM003540	DRKTEEDNQKKESGTKE-----AKKRKSAPPDTTSTKKSCTSSAEGDDKEVFLL	328
ANGAN33611	*****:****:****:****:****:****:****:****:****:****:****:****:****:	305
ORYLA07425	DRKTEEDNQKKESGTKE-----AKKRKSAPPDTTSTKKSCTSSAEGDDKEVFLL	340
ASTCA27703	DRKTEEDNQKKESGTKE-----AKKRKSAPPDTTSTKKSCTSSAEGDDKEVFLL	312
GADM003540	PIKGREIYEIILVKIKESLELMQFLPQHTIES-YRQQQQNLL-----	374
ANGAN33611	QVRGKKRFEMLKMINDSLEKLDPVAAQD-KYRQLHKSSTSKREREREGVEPKKGKRL	345
ORYLA07425	EVYGRERYEFLKKINDGLELLEKEPSNTDSSAMKKKCLAV--IAKKDIFCTQPTSRSAM	399
ASTCA27703	PVRGRERYEMLKKINDGLELLDKDSKSKVSVKHEVPVPS--SGK-----RL	370
GADM003540	: :	374
ANGAN33611	-QK-----347	345
ORYLA07425	LVKEEKSDSD 409	399
ASTCA27703	RTKEEQSDQT 380	370
GADM003540	LQRGERSDSD 384	374

# Homologs, orthologs, paralogs



# OMA

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

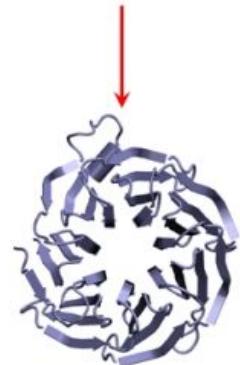
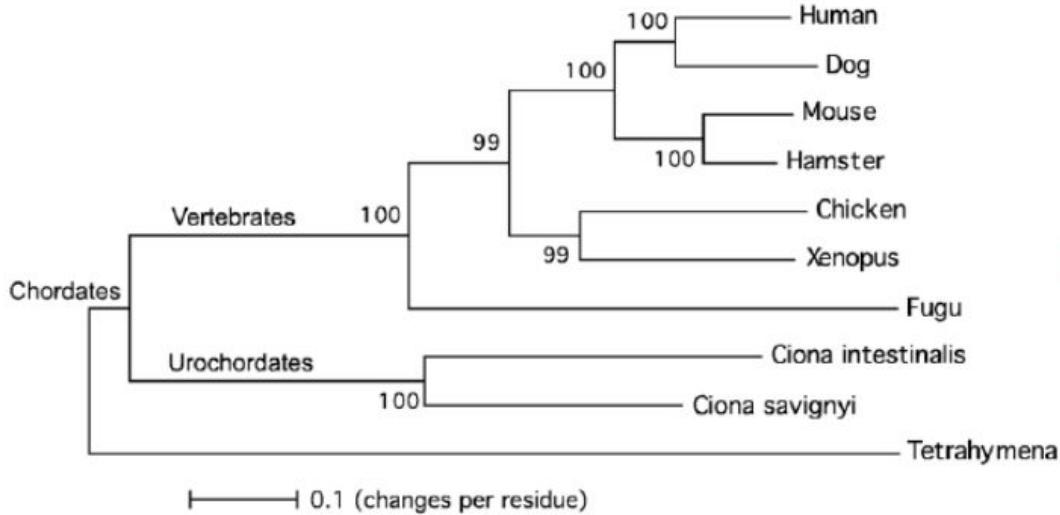
Orthologs 152	
FILTER BY TAXONOMY:	
<input checked="" type="checkbox"/> All Taxa	
<input type="radio"/>	Eukaryota
<input type="radio"/>	Viridiplantae
<input type="radio"/>	Fungi
<input type="radio"/>	Metazoa
<input type="radio"/>	Vertebrata
<input type="radio"/>	Arthropoda
<input type="radio"/>	Bacteria
<input type="radio"/>	Archaea
<a href="#">Add custom filter</a>	
REQUIRED ORTHOLOGY EVIDENCE:	
<input type="checkbox"/>	OMA pairs
<input type="checkbox"/>	HOG
<input type="checkbox"/>	OMA Group
Paralogs 19	
Gene information	
GO Annotations	

Evidence	Taxon	Protein ID	Relation type	Cross reference	Domain Architectures
✓ ✓	Lepisosteus oculatus	LEPOC07666	★ WSNBV4#		
✓ ✓	Anguilla anguilla	ANGAN33611		XP_035289590	
✓ ✓	Anabas testudineus	ANATE33593	★ ENSATEG000000019516.1#		
✓ ✓ ✓	Seriola dumerillii	SERDU24530	★ A0A3B4VC63#		
✓ ✓ ✓	Echeneis naucrates	ECHNA09228	★ A0A665W4M9#		
✓ ✓ ✓	Cynoglossus semilaevis	CYNSE17641	★ A0A3PBUNQ6#		
✓ ✓ ✓	Scorpaenichthys maximus	SCOM20524	★ ENSMAG000000009219.1#		
✓ ✓ ✓	Sparus aurata	SPAUAU07783	★ A0A671VUB9#		
✓ ✓	Mola mola	MOLML15319	★ A0A3Q3WVD8#		
✓ ✓ ✓	Takifugu rubripes	TAKRU47619	★ H2U133#		
✓ ✓ ✓	Tetraodon nigroviridis	TETNG08947	★ H3CXQ0#		
✓ ✓ ✓	Oryzias javanicus	ORYJA47514	★ ENSOJAG000000015701.1#		
✓ ✓ ✓	Oryzias latipes	ORYLA07425	★ ENSORLG000000006390#		
✓ ✓ ✓	Oryzias melastigma	ORYME20960	★ A0A3B3BVC0#		
✓ ✓ ✓	Nothobranchius furzeri	NOTFU06675	★ B3TLB0#		

**Template****Unknown structure****Sequence alignment**

ELAGI ILTVSYIPSAEKIA    ELAIGILTIVSYIPSAEKIR

ELAGI -ILGVSYIPSAEKI -ARACELTI
E LA -IGILTIVSYIPSAEKIRAP --ELTI

**Structural model**

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.

https://www.uniprot.org/uniprotkb?query=Insulin

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB ▾ Insulin Advanced | List Search Help

Status  
Reviewed (Swiss-Prot) (5,139)  
Unreviewed (TrEMBL) (112,956)

Popular organisms  
Rat (1,645)  
Human (1,577)  
Mouse (1,436)  
Bovine (815)  
Zebrafish (445)

Taxonomy  
Filter by taxonomy

Group by  
Taxonomy  
Keywords  
Gene Ontology  
Enzyme Class

Proteins with  
3D structure (1,268)  
Active site (12,939)  
Activity regulation (1,070)  
Allergen (8)  
Alternative products (isoforms) (1,322)  
More items

## UniProtKB 118,095 results

or search "Insulin" as a Gene Ontology, Protein Name, Protein family, Catalytic Activity, Disease, or Gene Name

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
P06213	INSR_HUMAN	Insulin receptor[...]	INSR	Homo sapiens (Human)	1,382 AA
P14735	IDE_HUMAN	Insulin-degrading enzyme[...]	IDE	Homo sapiens (Human)	1,019 AA
P01308	INS_HUMAN	Insulin[...]	INS	Homo sapiens (Human)	110 AA
P01317	INS_BOVIN	Insulin[...]	INS	Bos taurus (Bovine)	105 AA
P67970	INS_CHICK	Insulin[...]	INS	Gallus gallus (Chicken)	107 AA
P01321	INS_CANLF	Insulin[...]	INS	Canis lupus familiaris (Dog) (Canis familiaris)	110 AA
P17715	INS_OCTDE	Insulin[...]	INS	Octodon degus (Degu) (Sciurus degus)	109 AA
P01329	INS_CAVPO	Insulin[...]	INS	Cavia porcellus (Guinea pig)	110 AA
P01315	INS_PIG	Insulin[...]	INS	Sus scrofa (Pig)	108 AA
Q91X13	INS_ICTTR	Insulin[...]	INS	Ictidomys tridecemlineatus (Thirteen-lined ground squirrel) (Spermophilus tridecemlineatus)	110 AA
P01322	INS1_RAT	Insulin-1[...]	Ins1, Ins-1	Rattus norvegicus (Rat)	110 AA
Q9Y5Q6	INSL5_HUMAN	Insulin-like peptide INSLS5[...]	INSL5, UNQ156/PRO182	Homo sapiens (Human)	135 AA
P01323	INS2_RAT	Insulin-2[...]	Ins2, Ins-2	Rattus norvegicus (Rat)	110 AA
P01326	INS2_MOUSE	Insulin-2[...]	Ins2, Ins-2	Mus musculus (Mouse)	110 AA
P01325	INS1_MOUSE	Insulin-1[...]	Ins1, Ins-1	Mus musculus (Mouse)	108 AA
P15127	INSR_RAT	Insulin receptor[...]	Insr	Rattus norvegicus (Rat)	1,383 AA
P15208	INSR_MOUSE	Insulin receptor[...]	Insr	Mus musculus (Mouse)	1,372 AA

<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<sup>1</sup>

Structure<sup>1</sup>



Source: PDB

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
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<sup>1</sup>, turn<sup>1</sup>, beta strand<sup>1</sup>.



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

# 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/

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

RCSB PDB PROTEIN DATA BANK 196,779 Structures from the PDB 1,000,361 Computed Structure Models (CSM)

3D Structures il12a in Gene Name IL12A Include CSM Help

PDB-101 PDB EMD DataResource NUCLEIC ACID DATABASE wwPDB Foundation NEW! Computed Structure Models (CSM) Learn more

Welcome Deposit Search Visualize Analyze Download Learn

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:  
Experimentally-determined 3D structures from the Protein Data Bank (PDB) archive  
Computed Structure Models (CSM) from AlphaFold DB and ModelArchive  
These data can be explored in context of external annotations providing a structural view of biology.

October Molecule of the Month Phytohormone Receptor DWARF14

COVID-19 CORONAVIRUS Resources Join the RCSB PDB Team

Latest Entries As of Tue Oct 18 2022 Features & Highlights News Publications

Register Now for Virtual Crash Course: Using K-Base to access PDB Happy Birthday, PDB!



Return Structures

grouped by No Grouping



Include Computed Structure Models (CSM)

Count Clear

Search

Search Summary

This query matches 2 Structures.

Refinements



-- Tabular Report --

All  Selected



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 (A)

2.5 - 3.0 (1)

3.0 - 3.5 (1)

Release Date

2000 - 2004 (1)

2010 - 2014 (1)

Symmetry Type

Asymmetric (2)

SCOP Classification

1 to 2 of 2 Structures

Page 1 of 1

25



Sort by ↓ Score



3D View



1F45

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

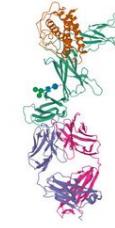
**Organisms** Homo sapiens

**Macromolecule** INTERLEUKIN-12 ALPHA CHAIN (protein)

INTERLEUKIN-12 BETA CHAIN (protein)

**Unique branched monosaccharides** MAN, NAG

Download File  View File



3D View



3HMX

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)

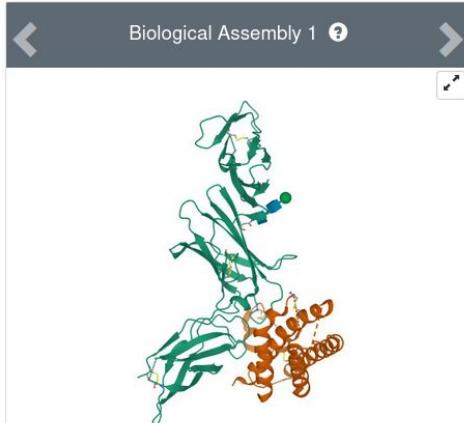
Interleukin-12 subunit beta (protein)

USTEKINUIMAB FAB HEAVY CHAIN (protein)

USTEKINUIMAB FAB LIGHT CHAIN (protein)

**Unique branched monosaccharides** BMA, MAN, NAG

Download File  View File



[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

# 1F45

## HUMAN INTERLEUKIN-12

PDB DOI: [10.2210/pdb1F45/pdb](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., Tobin, J.F., Somers, W.S.

#### Experimental Data Snapshot

Method: X-RAY DIFFRACTION

Resolution: 2.80 Å

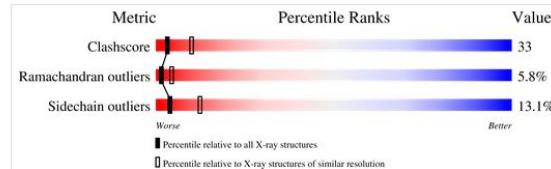
R-Value Free: 0.284

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](https://pubmed.ncbi.nlm.nih.gov/10899108/)

[Search on PubMed](#)

[Search on PubMed Central](#)

DOI: [10.1093/emboj/19.14.3530](https://doi.org/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	<a href="#">Homo sapiens</a>	Mutation(s): 0 Gene Names: <a href="#">IL12B</a> , <a href="#">NKS</a> <a href="#">F2</a>	

### 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

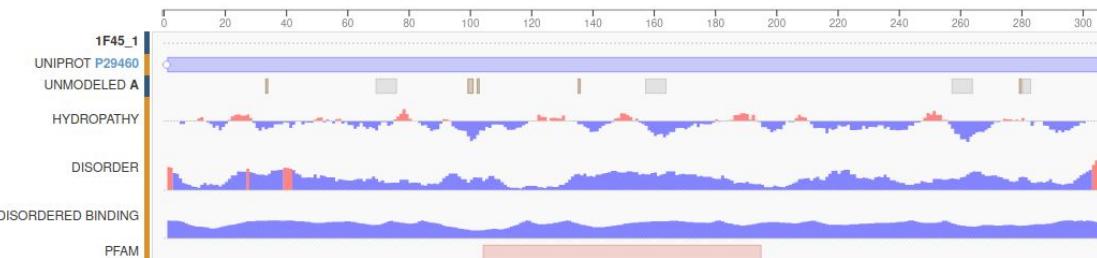
[P29460](#)

### Protein Feature View

[Expand](#)

Reference Sequence

1F45\_1



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

## Entity ID: 2

Molecule	Chains <small>i</small>	Sequence Length	Organism	Details	Image
INTERLEUKIN-12 ALPHA CHAIN	B	197	<a href="#">Homo sapiens</a>	<b>Mutation(s):</b> 1 <small>i</small> <b>Gene Names:</b> <a href="#">IL12A</a> , <a href="#">NKS</a> <a href="#">F1</a>	

## UniProt & NIH Common Fund Data Resources

Find proteins for [P29459](#) (*Homo sapiens*)

Explore [P29459](#) i

Go to UniProtKB: [P29459](#)

PHAROS: [P29459](#)

## Entity Groups i

Sequence Clusters

[30% Identity](#)  [50% Identity](#)  [70% Identity](#)  [90% Identity](#)  [95% Identity](#)  [100% Identity](#)

UniProt Group

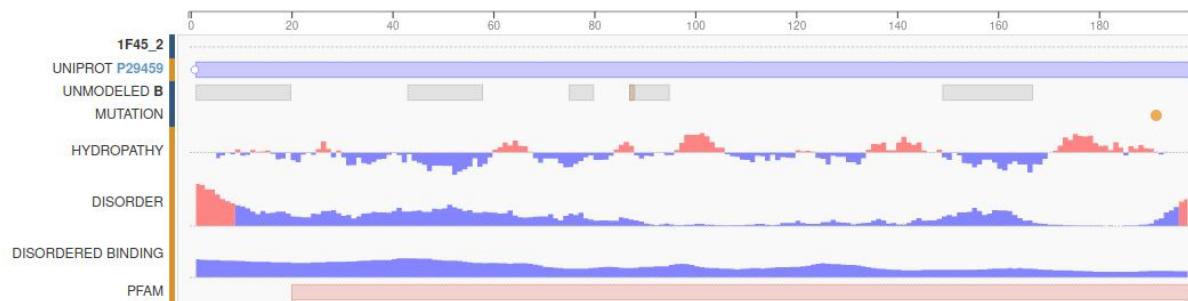
[P29459](#)

## 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	

## Glycosylation Resources

GlyToCan: [G6218200](#)GlyCosmos: [G6218200](#)GlyGen: [G6218200](#)

## Experimental Data &amp; Validation

## Experimental Data

Method: X-RAY DIFFRACTION

Resolution: 2.80 Å

R-Value Free: 0.284

R-Value Work: 0.241

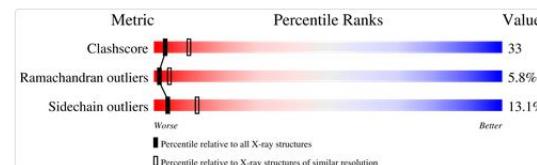
R-Value Observed: 0.241

Space Group: [C 2 2 2<sub>1</sub>](#)

## Unit Cell:

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

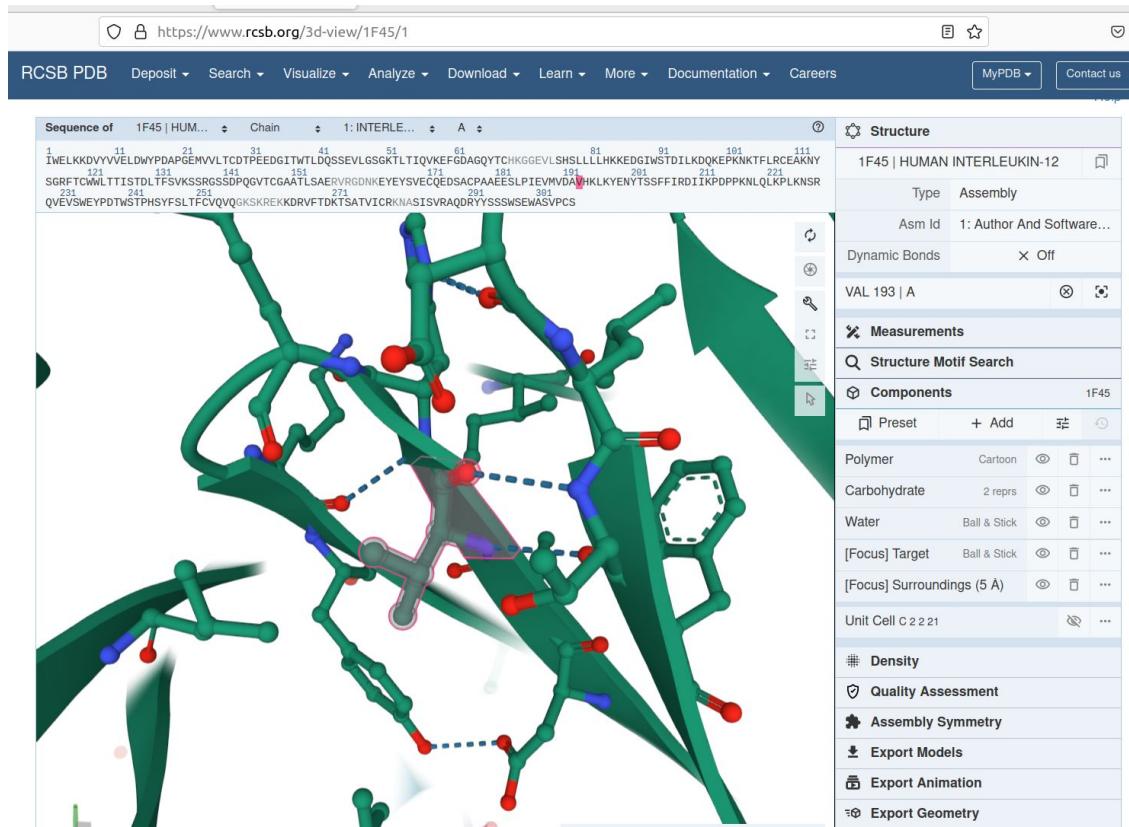
## Structure Validation

[View Full Validation Report](#)

## Software Package:

Software Name	Purpose
DENZO	data reduction
SCALEPACK	data scaling
MLPHARE	phasing

# Visualisation of the structure in PDB



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

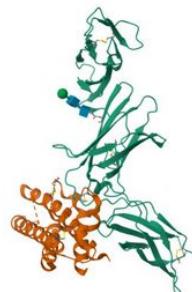
**RCSB PDB** PROTEIN DATA BANK 210,554 Structures from the PDB 1,068,577 Computed Structure Models (CSM)

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



1F45

HUMAN INTERLEUKIN-12

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

Classification: CYTOKINE/CYTOKINE

Organism(s): Homo sapiens

Expression System: Cricetus 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

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

Global Symmetry: Asymmetric - C1

Global Stoichiometry: Hetero 2-mer - A1B1

Find Similar Assemblies

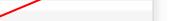
Biological assembly 1 assigned by authors and

Display Files Download Files Data API

FASTA Sequence

PDB/mmCIF Format

PDB/mmCIF Format (gz)

PDB Format 

PDB Format (gz)

PDBML/XML Format (gz)

Validation Full PDF

Validation (XML - gz)

Validation (CIF - gz)

wwPDB Validation Report Full Report

Biological Assembly 1 (CIF - gz) 33

Biological Assembly 1 (PDB - gz) 5.8%

Sidechain outliers 13.1%

Value Worse Better

Percentile relative to all X-ray structures

Percentile relative to X-ray structures of similar resolution

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

# PDB files

pdb file in Moodle

## Atomic Coordinates: PDB Format

Amino Acid	Chain name			Sequence Number	Coordinates			
	Element	X	Y	Z	(etc.)			
ATOM	1	N	ASP	L	1	4.060	7.307	5.186
ATOM	2	CA	ASP	L	1	4.042	7.776	6.553
ATOM	3	C	ASP	L	1	2.668	8.426	6.644
ATOM	4	O	ASP	L	1	1.987	8.438	5.606
ATOM	5	CB	ASP	L	1	5.090	8.827	6.797
ATOM	6	CG	ASP	L	1	6.338	8.761	5.929
ATOM	7	OD1	ASP	L	1	6.576	9.758	5.241
ATOM	8	OD2	ASP	L	1	7.065	7.759	5.948

\\" Element position within amino acid

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;

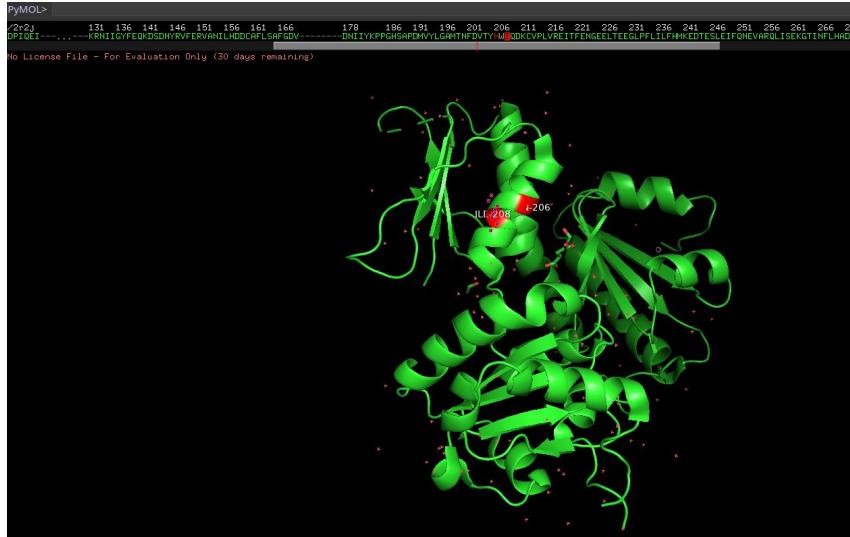
ATOM	2220	SG	CYS	A	305	76.079	50.412	8.054	1.00	90.85	S
ATOM	2221	N	SER	A	306	77.917	52.118	12.119	1.00	95.38	N
ATOM	2222	CA	SER	A	306	79.227	52.520	12.671	1.00	96.65	C
ATOM	2223	C	SER	A	306	79.673	51.582	13.814	1.00	96.08	C
ATOM	2224	O	SER	A	306	80.107	50.441	13.585	1.00	94.51	O
ATOM	2225	CB	SER	A	306	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/>



Pymol

<https://pymol.org/2/>

Human ERp44, PDB 2R2J

# NCBI RefSeq

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

National Library of Medicine  
National Center for Biotechnology Information

All Databases

All Databases

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

Assembly

Biocollections

BioProject

BioSample

Books

ClinVar

Conserved Domains

dbGaP

dbVar

Gene

Genome

GEO DataSets

GEO Profiles

GTR

HomoloGene

Identical Protein Groups

MedGen

MeSH

Submit

Transfer NCBI data to your computer

Download

Find help documents, attend a class or watch a tutorial

Learn

Develop

Identify an NCBI tool for your data analysis task

Analyze

Research

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Popular Resources

PubMed

Bookshelf

PubMed Central

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

NCBI News & Blog

Search, Download, and Visualize Human RNA-Seq Gene Expression Data in NCBI's Gene Expression Omnibus (GEO)

19 Apr 2023

Putting Content into Context: Clarifying PubMed Central's Role as an Archive

18 Apr 2023

The role of a library in a digital world continues to evolve and expand. NI M's

New annotations in RefSeq!

In February and March, the NCBI Eukaryotic Genome Annotation Pipeline released many new

More...

# 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

## 1. Generate words from sequence above threshold (e.g. T=11)

Query Sequence:

>gi|16329320 (residues 412 to 594)  
SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTQTG  
RHQGILT**SWVSQASFTPPGIM**LAIPEGFDAYGLAGQNKAFLVNL  
VRRHFDHQPLPKDGDNPFSRLEHYSTQNCGLILAEALAYLECLVQ  
GDHVLYVATVQAGQVLQPNQGITAIRHRKSGGQY

Fragmentation into words:

**SWVSQASFTPPGIM** → SWV WVS VSQ SQA QAS ASF SFT ...

## Selection of words scoring above threshold (for word SWV):

Substitution Matrix\*

R	G	I	K	F	S	T	W	V
R	5	0	-1	-2	1	0	-3	0
G	6	-4	-2	-3	0	-2	-2	-3
I	4	-3	0	-2	-1	-3	3	
K	5	-3	0	-1	-3	-2		
F		6	-2	-2	1	-1		
S			4	1	-3	-2		
T				5	-2	0		
W					11	-3		
V						4		

\*A portion of the BLOSUM 62 matrix

SWV (4+11+4 = 19)
SWI (4+11+3 = 18)
TWV (1+11+4 = 16)
GWV (0+11+4 = 15)
KWV (0+11+4 = 15)
SWS (4+11-2 = 13)
SFV (4+1+4 = 9)
SRV (4-3+4 = 5)

Synonyms above threshold 11...  
(others not shown)

Synonyms below threshold 11...  
(others not shown)

## 2. Search the database for words matching those generated

## 3. Extend matching hits in both directions

**RHOGIL**ISWVSQASFTPPGIM**LAIPEGFDAYGLAGQN**  
..**TAML**SWVSQASFNPPGILTIALAKE**,RAEGLDHS**GD  
Word match      Extension until score drops

## 4. Generate alignment and calculate statistics

>ref|NP\_002482587.1| flavin reductase domain protein FMN-binding [Cyanothece sp. PCC 7425]  
gb|RCL4226.1| flavin reductase domain protein FMN-binding [Cyanothece sp. PCC 7425]  
Length=585

Score = 176 bits (446), Expect = 1e-42, Method: Compositional matrix adjust.  
Identities = 95/196 (48%), Positives = 125/196 (63%), Gaps = 16/196 (8%)

Query 1    SGANFARQLRTHKRQRIARQATTETQADRTQQAVGRIIGSIGVVTQTGGRH----- 52  
+G++FA+ L+ K+R R+ E O+DRT+QAVGRIIGSIGVVTQTGGRH----- 52  
Sbjct 393 AGSDFAVPLKKAKKKQRSPRSQSLIEVQSDRTEQAVGRIIGSLCVLTAKQQQTHPHPEVEEP 452

Query 53 -----QGIIT**SWVSQASFTPPGIM**LAIPEGFDAYGLAGQNKAFLVNL  
+L SWVSQASFTPPGIM+A+E A GL AFVNL+L+E EG ++RRHF 107  
Sbjct 453 QLEVPTAMLVSWVSQASFNPPGILTIALAKE,RAEGLDHSGD  
AFVNL+L+E EG ++RRHF 107

Query 108 QPLFKDGDNPFSSRLHYSTQNCGLILAEALAYLECLVQGSWSNIDGHVLYATVQAGQVLQ 167  
P G++ F+ L+ +NGC + 1 + LAYLE VQS GDH L+YATV G+VVLQ 167  
Sbjct 512 SFAP+-GEDRFAGLNIQWAENGCPVLFQDCLAYLECTVQSRMECGDHWLIIYATVNNNGKVLQ 569

Query 168 PINGITAIRHRKSGGQY 183

P G TA++HRKSG QY

Sbjct 570 PTGTTAVQHRKSGNQY 585

Threshold  
T

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

Try to perform BLAST search using the following sequence

```
>gene  
ggtgataaagtcatccgctttactcctcagtgtggaaaatgcagagttgtaaaaaccggagagcaactactgctgaaaaatgatctaggcaatcct  
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



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](#)

```
ggtgataaagtcatcccgctttaactcctcagttgtggaaaatgcagagtttgtaaaaacccggagagcaactac  
tgcgtaaaaatgtataggcaatcctcggggaccctgcaggat
```

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 [?](#)

### 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

[YouTube](#) Create custom database

**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



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

Select the maximum number of aligned sequences to display [?](#)

Short queries  Automatically adjust parameters for short input sequences [?](#)

Expect threshold  [?](#)

Word size  [?](#)

Max matches in a query range  [?](#)

### Scoring Parameters

Match/Mismatch  [?](#)

Scores

Gap Costs  [?](#)

### Filters and Masking

Filter  Low complexity regions [?](#)

Species-specific repeats for:  [?](#)

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

BLAST® » blastn suite » results for RID-Z0P41CPJ01R

Home Recent Results Saved Strategies Help

◀ Edit Search Save Search Search Summary ▾

How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title Nucleotide Sequence

RID Z0P41CPJ01R Search expires on 01-27 03:07 am Download All ▾

Program BLASTN ? Citation ▾

Database nt See details ▾

Query ID Icl|Query\_236239

Description None

Molecule type dna

Query Length 120

Other reports Distance tree of results MSA viewer ?

**Filter Results**

Organism only top 20 will appear  exclude  
Type common name, binomial, taxid or group name  
+ Add organism

Percent Identity E value Query Coverage

[ ] to [ ] [ ] to [ ] [ ] to [ ]

Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Download New Select columns Show 100 ?

Sequences producing significant alignments

select all 100 sequences selected

GenBank Graphics Distance tree of results New MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident.	Acc. Len	Accession
<input checked="" type="checkbox"/>	PREDICTED: Pan paniscus alcohol dehydrogenase 1B (class I), beta polypeptide (ADH1B), mRNA	Pan paniscus	222	222	100%	5e-54	100.00%	2573	XM_003829928.3
<input checked="" type="checkbox"/>	PREDICTED: Gorilla gorilla gorilla alcohol dehydrogenase 1B (class I), beta polypeptide (ADH1B), transcript ...	Gorilla gorilla go...	222	222	100%	5e-54	100.00%	2574	XM_031010301.1
<input checked="" type="checkbox"/>	PREDICTED: Gorilla gorilla gorilla alcohol dehydrogenase 1B (class I), beta polypeptide (ADH1B), transcript ...	Gorilla gorilla go...	222	222	100%	5e-54	100.00%	2578	XM_031010300.1
<input checked="" type="checkbox"/>	Homo sapiens alcohol dehydrogenase 1B (class I), beta polypeptide (ADH1B), transcript variant 2, mRNA	Homo sapiens	222	222	100%	5e-54	100.00%	4189	NM_00128669
<input checked="" type="checkbox"/>	Homo sapiens alcohol dehydrogenase 1B (class I), beta polypeptide (ADH1B), transcript variant 1, mRNA	Homo sapiens	222	222	100%	5e-54	100.00%	4067	NM_000668.6

Filter

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 222 bits(120)	Expect 5e-54	Identities 120/120(100%)	Gaps 0/120(0%)	Strand Plus/Plus	
Query 1	GGTGATAAAGTCATCCCGCTCTTACTCCTCAGTGTGGAAAATGCAGAGTTGTAAAAAC	60			
Sbjct 367	GGTGATAAAGTCATCCCGCTCTTACTCCTCAGTGTGGAAAATGCAGAGTTGTAAAAAC	426			
Query 61	CCGGAGAGCAACTACTGCTTGAAAATGATCTAGGCAATCCTCGGGGGACCTGCAGGAT	120			
Sbjct 427	CCGGAGAGCAACTACTGCTTGAAAATGATCTAGGCAATCCTCGGGGGACCTGCAGGAT	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 222 bits(120)	Expect 5e-54	Identities 120/120(100%)	Gaps 0/120(0%)	Strand Plus/Plus	
Query 1	GGTGATAAAGTCATCCCGCTCTTACTCCTCAGTGTGGAAAATGCAGAGTTGTAAAAAC	60			
Sbjct 365	GGTGATAAAGTCATCCCGCTCTTACTCCTCAGTGTGGAAAATGCAGAGTTGTAAAAAC	424			
Query 61	CCGGAGAGCAACTACTGCTTGAAAATGATCTAGGCAATCCTCGGGGGACCTGCAGGAT	120			
Sbjct 425	CCGGAGAGCAACTACTGCTTGAAAATGATCTAGGCAATCCTCGGGGGACCTGCAGGAT	484			

## Related Information

[Gene](#) - associated gene details

[Genome Data Viewer](#) - aligned genomic context

[Download](#) ▾ [GenBank](#) [Graphics](#)

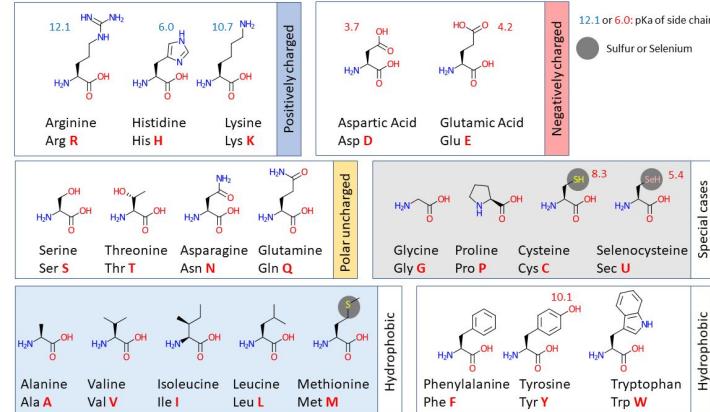
[▼ Next](#) [▲ Previous](#) [◀ Descriptions](#)

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/>



Thomas Ryckmans 2021