

Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFMKIIQLDDYPKCFVVGADNVGSKOMQOIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--SALE	76
Q7ZUG3_BRARE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGKNTMMRKAIRGHLENN--PALE	76
RLA0_ICTPU	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_DROME	-----MVRENKAAWKAQYFIKVVLEFDEFKCFIVGADNVGSKOMQOIRMSLRGL-AVVLGKNTMMRKAIRGHLENN--PQLE	76
RLA0_DICDI	-----MSGAG-SKRKKLFIEKATKLTFTYDKMIVAEADFVGSSOLOKIRKSIRGI-GAVLMGKNTMIRKVIDRLADSK--PELD	75
Q54LP0_DICDI	-----MSGAG-SKRKNVFIEKATKLTFTYDKMIVAEADFVGSSOLOKIRKSIRGI-GAVLMGKNTMIRKVIDRLADSK--PELD	75
RLA0_PLAF8	-----MAKLSQKKQMYIEKLSLIQQYSKILIVHVDNVGSKOMASVRKSLRGK-ATILMGKNTIRIRALTAKNLQAV--PQIE	76
RLA0_SULAC	-----MIGLAVTTTKKIAKWKVDEVAELTEKLTHTKTIITIANIEGFPADKLHEIRKKLRGK-ADIKVTKNLNFNIALKNAG----YDLE	79
RLA0_SULTO	-----MRIMAVITQERKIAKWKIEVKELEKLRKYHTIITIANIEGFPADKLHDIRKKMRGM-AEIKVTKNLTFGIAAKNAG----LDVS	80
RLA0_SULSO	-----MKRLALALKQRKVASWKEEVKELTELKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNLTFKIAAKNAG----IDIE	80
RLA0_AERPE	MSVVSIVGQMYKREKPIPEWKTLMLELEELFSKHRVFLADLTGTPTFVVRVVRKLLWKK-YPMVAKKRIILRAMKAAGLE---LDDN	86
RLA0_PYRAE	MMLAIGKRRYVTRQYPARKYKIVSEATELLQKYPYVFLDLHGLSRILHEYRYRLRY-GVIKIIKPTLFKIAFTKVYGG---IPAE	85
RLA0_METAC	-----MAEERHHTHEIPQWKDEIENIKELIQSHKVFVMVIEGILATKMKIRRDLDV-AVLKVSNTLIERALNQLG----ETIP	78
RLA0_METMA	-----MAEERHHTHEIPQWKDEIENIKELIQSHKVFVMVIEGILATKMKIRRDLDV-AVLKVSNTLIERALNQLG----ESIP	78
RLA0_ARCFU	-----MAAVRGS---PPEYKYRAVEEIKRMISKPVVAIVSFRNVPAGOMQIRREFRGK-AEIKVKNLTLERDALG----GDYL	75
RLA0_METKA	MAVKAKGQPPSGYEPKVAEWKREVEKELKELMDEYENVGLVDLEGIPAPQLQEIIRAKLRERDIIIRMSNTLMIRALEEKLDER--PELE	88
RLA0_METTH	-----MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPAROLQKMRQLDS-ALIRMSKKTLSIALKAGREL--ENVY	74
RLA0_METTL	-----MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPAROLQEIIRDKIR-GTMTLKMSNTLIERAIEVAEETGNPEFA	82
RLA0_METVA	-----MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAROLQEIIRDKIR-DQMTLKMSNTLIERAIEVAEETGNPEFA	82
RLA0_METJA	-----METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIIRDKIR-DKVLRMSNTLIERALKEAAEELNPNKLA	81
RLA0_PYRAB	-----MAHVAEWKKKEVEELANLKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAGELGKPELE	77
RLA0_PYRHO	-----MAHVAEWKKKEVEELAKLKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAGELGKPELE	77
RLA0_PYRFU	-----MAHVAEWKKKEVEELANLKSYPVIALVDVSSMPAYPLSQMRRLIRENGLLRVSRNTLIELAIKKAAGELGKPELE	77
RLA0_PYRKO	-----MAHVAEWKKKEVEELANIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKKAAGELGQPELE	76
RLA0_HALMA	MSAESERKTETIPEWKQEEVDIVEMIESYESVGVVNIAGIPSRLOQDMRRDLHGT-AELRVSRNTLIERALDDVD----DGLE	79
RLA0_HALVO	MSESEVRQTEVIPQWKREEVDLVDLIESYESVGVVGAGIPSRLOQSMRRELHGS-AAVRMSNTLVNRALEVN----DGFE	79
RLA0_HALSA	MSAEEQRTTEVIPWKQREVAELVDLLETYSVGVVNTGIPSKOLOQDMRRGLHGO-AALRMSNTLLVRALEEAG----DGLD	79
RLA0_THEAC	-----MKEVSQKKELVNEITRIKASRSVAIVDTAGIRTRIQIDIRGKNRGK-INLKVIKKTLLFKALENLGD---EKLS	72
RLA0_THEVO	-----MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTROMDIRAKNRDK-VKIKVVKKTLLFKALDSIND---EKLT	72
RLA0_PICTO	-----MTEPAQWKIDFVKNLENEINSRKVAIVSIKGLRNNFQKIRNSIRDK-ARIKVSARLLRLAIENTGK---NNIV	72
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90	

# Multiple Sequence Alignments

BIOINFORMATICS

# **TOPIC**

**Identification of Conserved Regions in  
Protein Sequences Through  
Multiple Sequence Alignments.**

# INDEX

Sr No.	CONTENT	Page No
1	<b><u>Introduction to Multiple Sequence Alignment (MSA)</u></b>	1-2
2	<b><u>Types and Tools Involved in MSA</u></b>	3-4
3	<b><u>Steps of MSA</u></b>	5-19
4	<b><u>Application of MSA</u></b>	20
5	<b><u>Conclusion</u></b>	21
6	<b><u>Reference</u></b>	22

# Introduction

**Proteins** are fundamental macromolecules that perform a vast array of functions essential for life, including catalysis, structural support, signalling, and transport. The sequence of amino acids in a protein dictates its three-dimensional structure and, consequently, its function. **Across evolution, certain segments of protein sequences remain remarkably unchanged**, even among distantly related species. These conserved regions are critical to understanding protein function, stability, and evolutionary relationships.

**Multiple Sequence Alignment (MSA)** is a cornerstone technique in bioinformatics that enables researchers to compare and align three or more biological sequences simultaneously. By arranging sequences such that homologous residues are positioned in vertical columns, MSA reveals patterns of conservation and variation. These alignments are instrumental in identifying conserved regions, which often correspond to functional domains, active sites, or structural motifs essential for protein activity.

## What A Multiple Sequence Alignment mean?

In a Multiple sequence alignment, **homologous** residues among a set of sequences are aligned together in columns. '**Homologous**' is meant in both the structural and evolutionary sense.

The significance of conserved regions extends beyond mere sequence similarity. They often represent evolutionary constraints—areas where mutations are deleterious and thus eliminated by natural selection. **As noted by Mount (2004), conserved regions are key to inferring functional importance and evolutionary relationships.** In structural biology, conserved residues frequently contribute to the protein's core or active site, maintaining structural integrity and catalytic efficiency (**Lesk, 2017**).

From an evolutionary perspective, conserved regions serve as molecular fossils, providing insights into ancestral sequences and speciation events. Phylogenetic analyses based on these regions help reconstruct evolutionary trees, elucidating relationships between species and protein families. **The work of Felsenstein (2004) underscores the power of sequence alignment in evolutionary biology, enabling hypotheses about common ancestry and functional divergence.**

This project employs MSA to identify and analyse conserved regions in a set of protein sequences, aiming to bridge sequence analysis with functional and evolutionary insights. By integrating bioinformatics tools, this study seeks to demonstrate how conserved regions inform our understanding of protein biology, from molecular function to evolutionary history.

# Types

## 1. Progressive Alignment Methods

Probably the most commonly used approach to multiple sequence alignment is progressive alignment. This works by constructing a succession of pairwise alignments. Initially, two sequences are chosen and aligned by standard pairwise alignment; this alignment is fixed. Then, a third sequence is chosen and aligned to the first alignment, and this process is iterated until all sequences have been aligned. Classic example: CLUSTAL.

## 2. ITERATIVE Refinement Methods

One problem with progressive alignment algorithms is that the subalignments are 'frozen'. That is, once a group of sequences has been aligned, their alignment to each other cannot be changed at a later stage as more data arrive. Iterative refinement algorithms attempt to circumvent this problem.

# Tools Involved In MSA

## Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

Launch [Clustal Omega](#)

## EMBOSS Cons

EMBOSS Cons creates a consensus sequence from a protein or nucleotide multiple alignment

Launch [EMBOSS Cons](#)

## Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

Launch [Kalign](#)

## MAFFT

MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.

Launch [MAFFT](#)

## MUSCLE

Accurate MSA tool, especially good with proteins. Suitable for medium alignments.

Launch [MUSCLE](#)

## MUSCLE 5

Muscle v5 is an extensive re-write of the MUSCLE code based on new algorithms. Suitable for medium-large alignments.

Launch [MUSCLE 5](#)

## MView

Transform a Sequence Similarity Search result into a Multiple Sequence Alignment or reformat a Multiple Sequence Alignment using the MView program.

Launch [MView](#)

## T-CoffeE

Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.

Launch [T-Coffee](#)

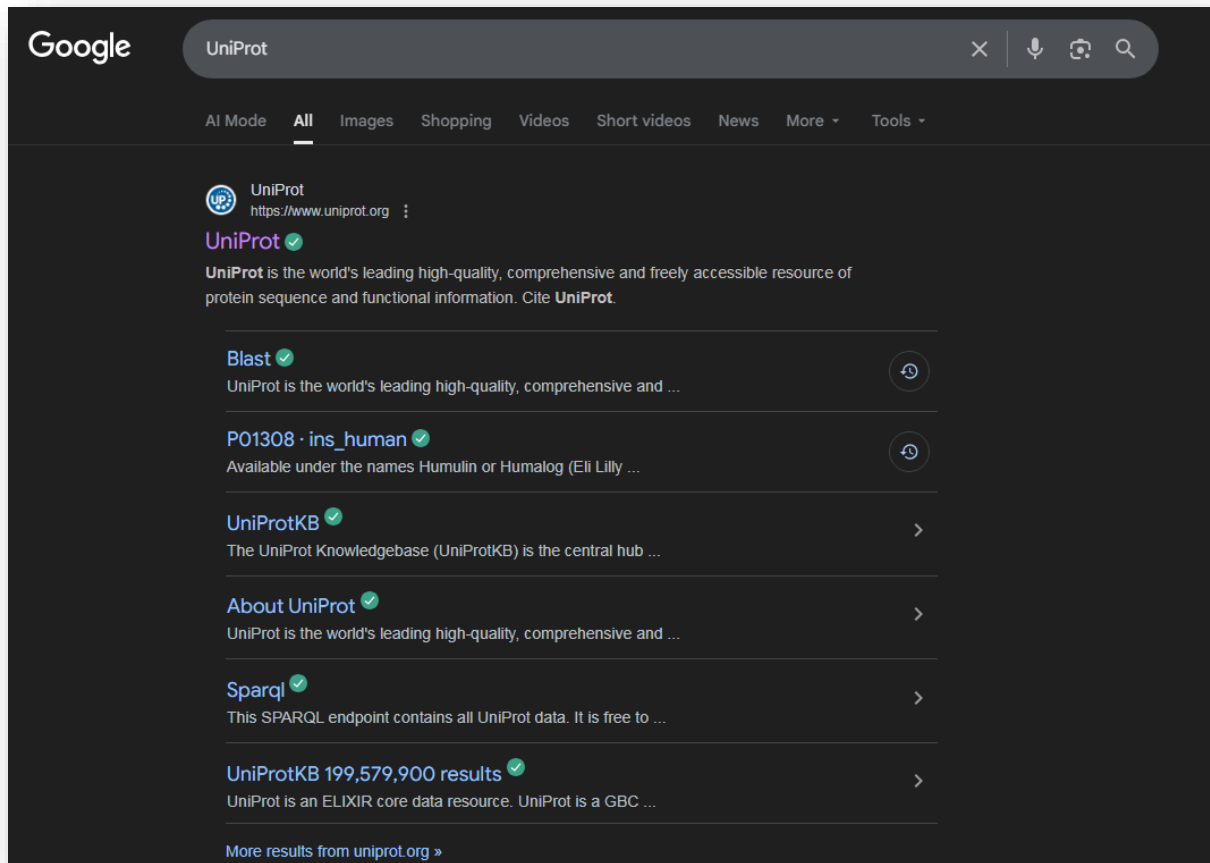
## WebPRANK

The EBI has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions.

Launch [WebPRANK](#)

# Steps In MSA

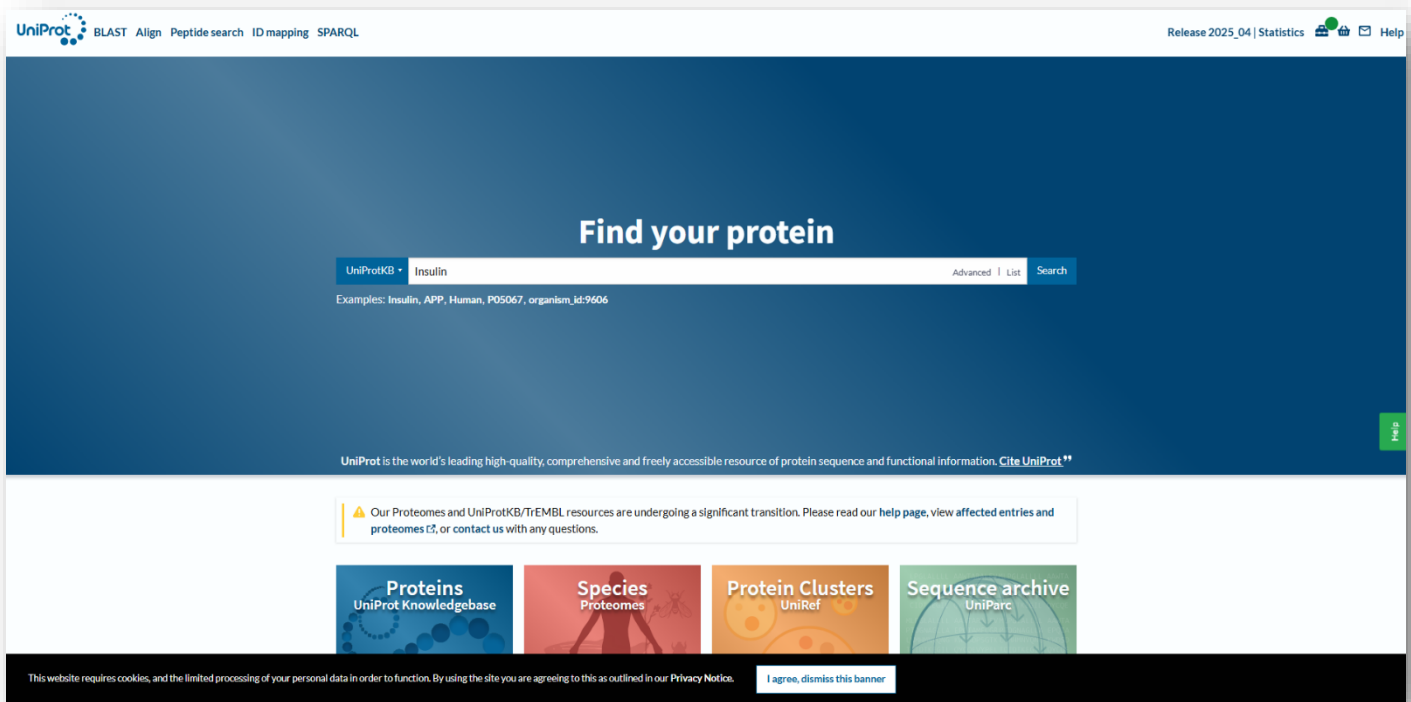
1. Search **UniPort** On Google and open it.



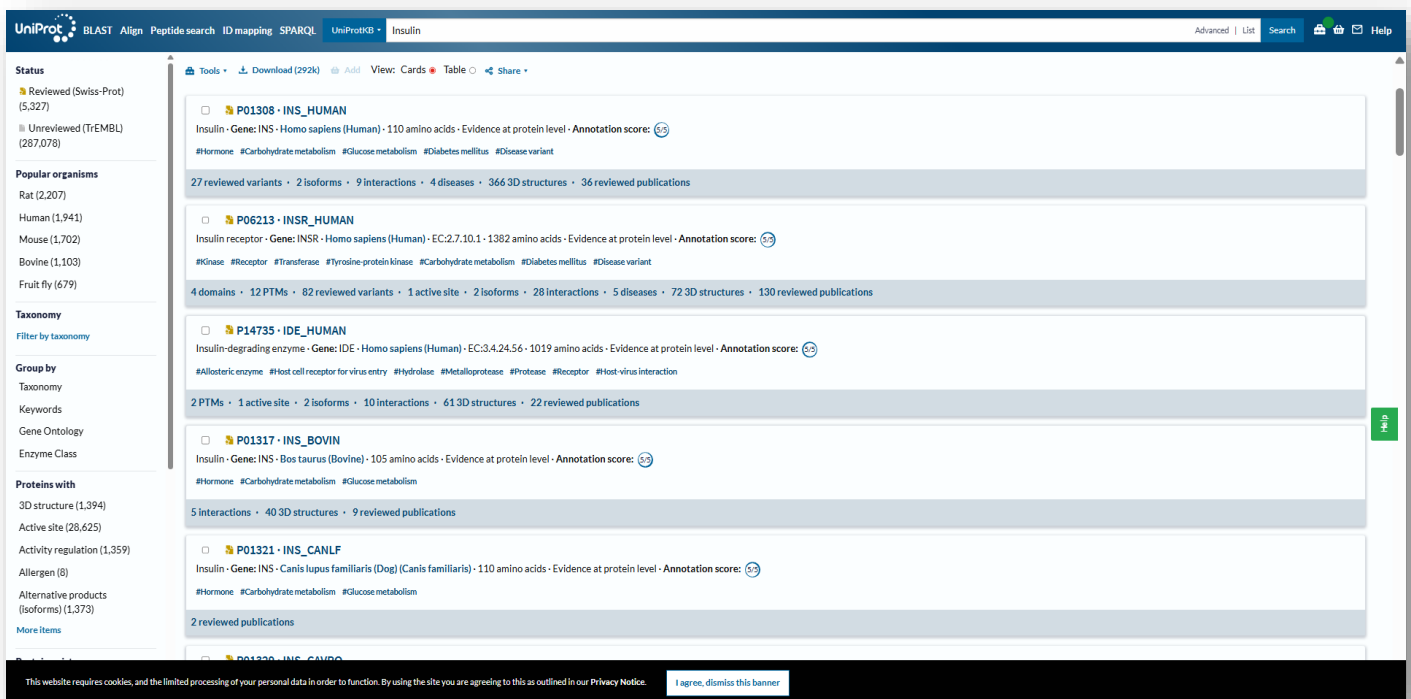


<https://github.com/code-aradhana/bioinformatics-msa-project.git>

## 2. Enter a Query sequence in search box.



## 3. Results will appear in this format.



<https://github.com/code-aradhana/bioinformatics-msa-project.git>

#### 4. Select the **required item** from the list.

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Insulin

Tools Download (292K) Add View: Cards Table Share

**Status**  
 Reviewed (Swiss-Prot) (5,327)  
 Unreviewed (TrEMBL) (287,078)

**Popular organisms**  
 Rat (2,207)  
 Human (1,941)  
 Mouse (1,702)  
 Bovine (1,103)  
 Fruit fly (679)

**Taxonomy**  
 Filter by taxonomy

**Group by**  
 Taxonomy  
 Keywords  
 Gene Ontology  
 Enzyme Class

**Proteins with**  
 3D structure (1,394)  
 Active site (28,625)  
 Activity regulation (1,359)  
 Allergen (8)  
 Alternative products (isoforms) (1,373)  
 More items

**P01308 · INS\_HUMAN**  
 Insulin - Gene: INS - Homo sapiens (Human) - 110 amino acids - Evidence at protein level - Annotation score: 65  
 #Hormone #Carbohydrate metabolism #Glucose metabolism #Diabetes mellitus #Disease variant  
 27 reviewed variants · 2 isoforms · 9 interactions · 4 diseases · 366 3D structures · 36 reviewed publications

**P06213 · INSR\_HUMAN**  
 Insulin receptor - Gene: INSR - Homo sapiens (Human) - EC:2.7.10.1 - 1382 amino acids - Evidence at protein level - Annotation score: 65  
 #Kinase #Receptor #Transferase #Tyrosine-protein kinase #Carbohydrate metabolism #Diabetes mellitus #Disease variant  
 4 domains · 12 PTMs · 82 reviewed variants · 1 active site · 2 isoforms · 28 interactions · 5 diseases · 72 3D structures · 130 reviewed publications

**P14735 · IDE\_HUMAN**  
 Insulin-degrading enzyme - Gene: IDE - Homo sapiens (Human) - EC:3.4.24.56 - 1019 amino acids - Evidence at protein level - Annotation score: 65  
 #Allosteric enzyme #Host cell receptor for virus entry #Hydrolase #Metalloprotease #Protease #Receptor #Host-virus interaction  
 2 PTMs · 1 active site · 2 isoforms · 10 interactions · 613D structures · 22 reviewed publications

**P01317 · INS\_BOVIN**  
 Insulin - Gene: INS - Bos taurus (Bovine) - 105 amino acids - Evidence at protein level - Annotation score: 65  
 #Hormone #Carbohydrate metabolism #Glucose metabolism  
 5 interactions · 40 3D structures · 9 reviewed publications

**P01321 · INS\_CANLF**  
 Insulin - Gene: INS - Canis lupus familiaris (Dog) (Canis familiaris) - 110 amino acids - Evidence at protein level - Annotation score: 65  
 #Hormone #Carbohydrate metabolism #Glucose metabolism  
 2 reviewed publications

**P01320 · INS\_CANDO**

This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Notice. [I agree, dismiss this banner](#)

#### 5. Page will appear like this.

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB

**P01308 · INS\_HUMAN**  
 Protein: Insulin  
 Gene: INS  
 Status: UniProtKB reviewed (Swiss-Prot)  
 Organism: Homo sapiens (Human)

Amino acids: 110 (gap to sequence)  
 Protein existence: Evidence at protein level  
 Annotation score: 65

Entry Variant viewer Feature viewer Genomic coordinates Publications External links History

Tools Download Add Community curated (1) Add a publication Entry feedback

**Function**  
 Insulin decreases blood glucose concentration. It increases cell permeability to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver.

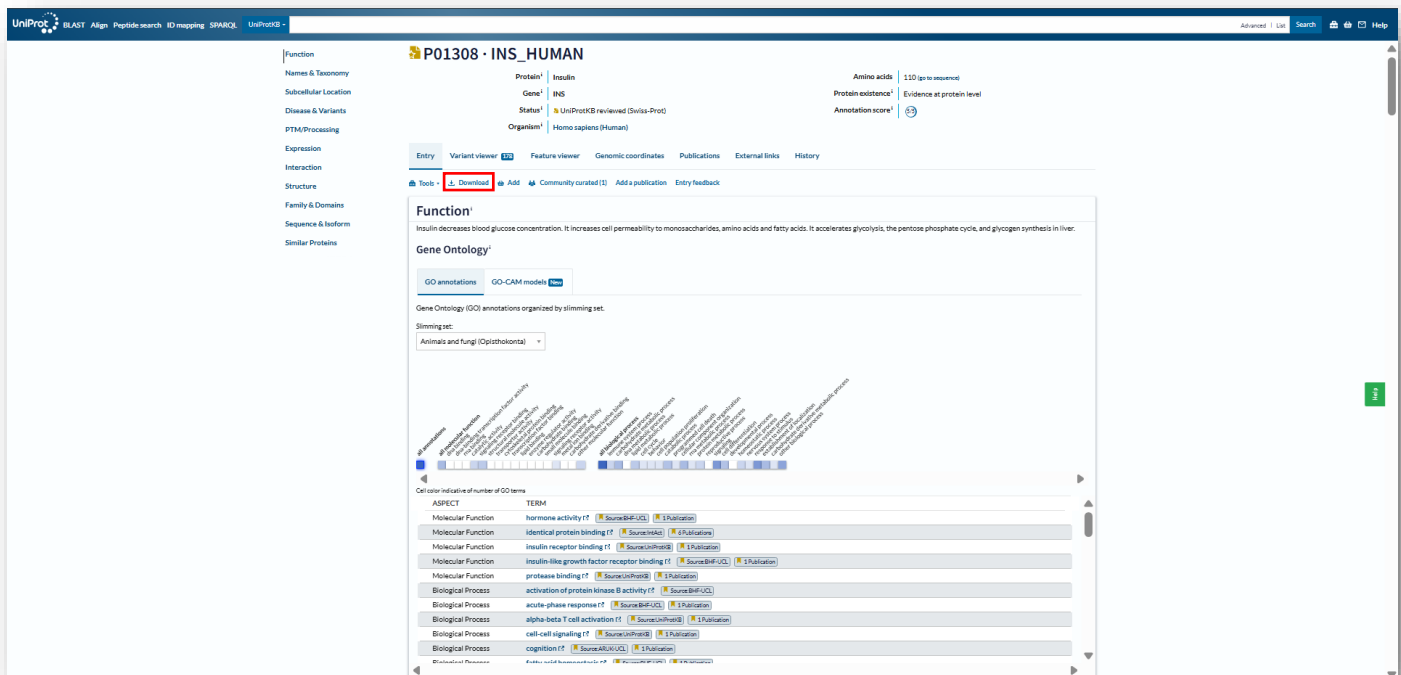
**Gene Ontology**  
 GO annotations GO-CAM models

Gene Ontology (GO) annotations organized by slimming set.  
 Slimming set: Animals and fungi (Opisthokonta)

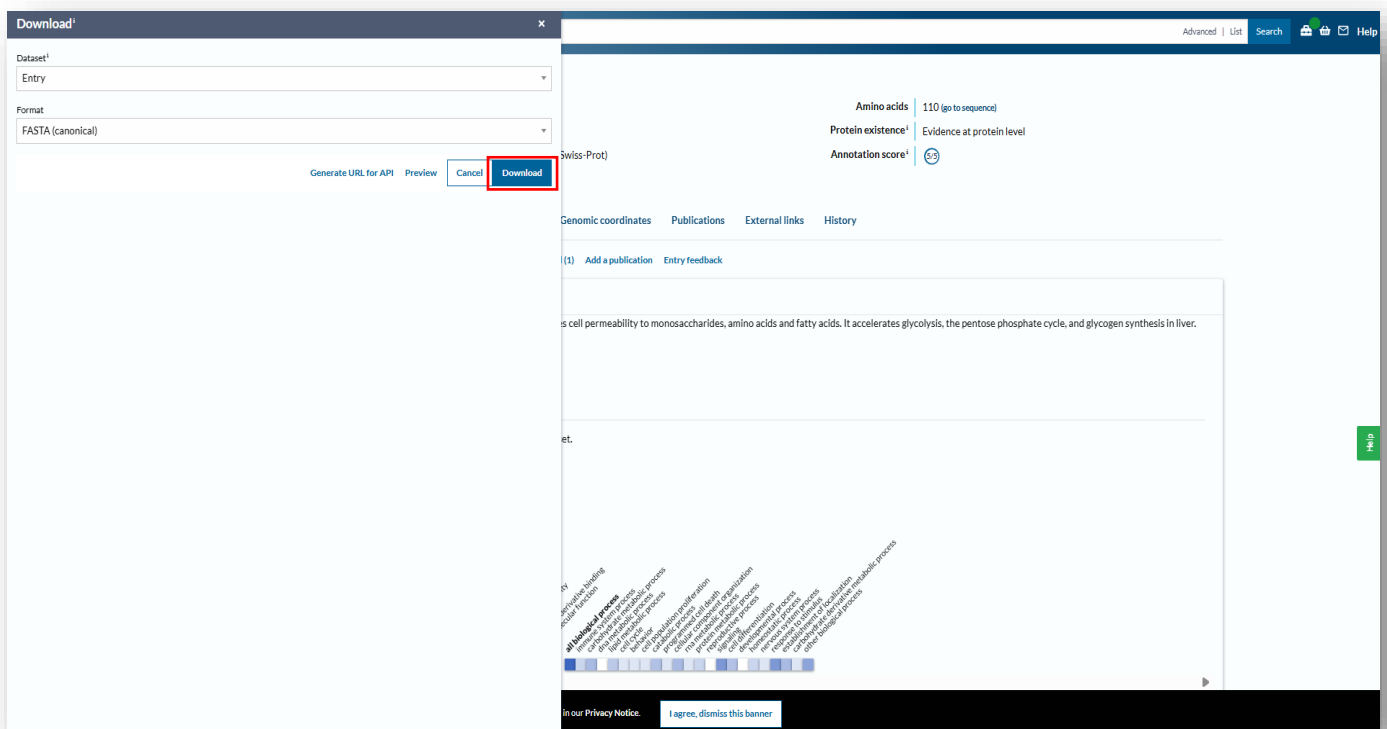
Cell color indicates number of GO terms

ASPECT	TERM	Source	Publication
Molecular Function	hormone activity	Source:UniProt	1 Publication
Molecular Function	identical protein binding	Source:UniProt	1 Publication
Molecular Function	insulin receptor binding	Source:UniProt	1 Publication
Molecular Function	insulin-like growth factor receptor binding	Source:UniProt	1 Publication
Molecular Function	protease binding	Source:UniProt	1 Publication
Biological Process	activation of protein kinase B activity	Source:UniProt	1 Publication
Biological Process	acute-phase response	Source:UniProt	1 Publication
Biological Process	alpha-beta T cell activation	Source:UniProt	1 Publication
Biological Process	cell-cell signaling	Source:UniProt	1 Publication
Biological Process	cognition	Source:UniProt	1 Publication

6. Click on **Download**.



and select **FASTA** in Format drop-down and after that click on **Download** to retrieve it's FASTA sequence.



<https://github.com/code-aradhana/bioinformatics-msa-project.git>

## 7. Click on **Tools**.

The screenshot shows the UniProt entry for P01308 (INS\_HUMAN). The 'Tools' menu is highlighted in the top navigation bar. The main content area displays the 'Function' section, which includes a description of insulin's role in decreasing blood glucose concentration and increasing cell permeability. Below this, the 'Gene Ontology' section is visible, showing a list of GO terms and their associated evidence.

A menu will appear and then select **BLAST**.

The screenshot shows the UniProt entry for P01308 (INS\_HUMAN) with the 'Tools' menu open. The 'BLAST' option is highlighted. The main content area displays the 'Function' section, which includes a description of insulin's role in decreasing blood glucose concentration and increasing cell permeability. Below this, the 'Gene Ontology' section is visible, showing a list of GO terms and their associated evidence.

## 8. Scroll down and click **Run BLAST**.

**BLAST<sup>i</sup>**

Find a protein sequence to run BLAST sequence similarity search by UniProt ID (e.g. P05067 or A4\_HUMAN or UPI0000000001).

UniProt IDs

OR

Enter one or more sequences (5 max). You may also [load from a text file](#).

```
>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens OX=9606 GN=INS PE=1 SV=1
MALWMRLLPLLALLLALWGDPDAFAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIKSLYQLENYCN
```

ⓘ Your input contains 1 sequence.

Target database: UniProtKB reference proteomes + Swiss-Prot

Restrict by taxonomy:

Name your BLAST job: sp|P01308|INS\_HUMAN

▼ **Advanced parameters**

Sequence type: Protein

Program: blastp

E-Threshold: 10

Matrix: Auto - BLOSUM62

Filter: None

Gapped: yes

Hits: 250

HSPs per hit: All

## 9. A page will come as shown below.

**Tool results**

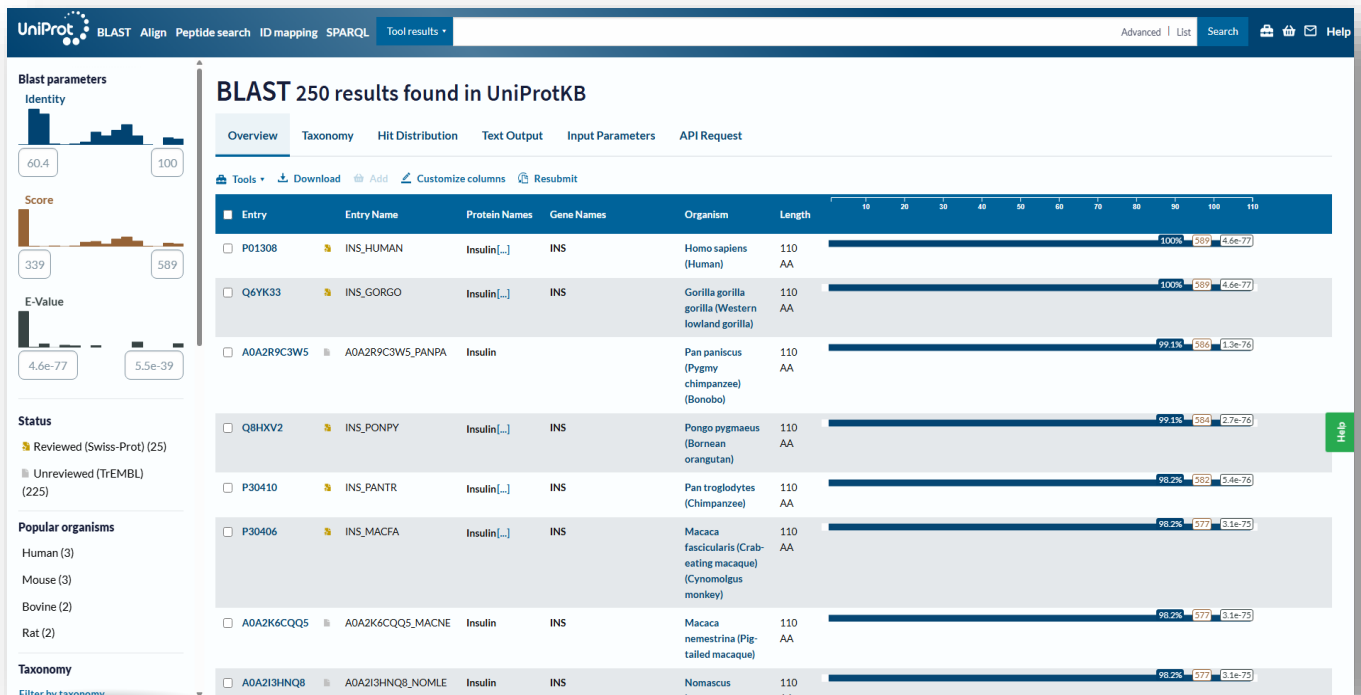
Your tool analysis results from the last 7 days are listed below. If you have tools jobs running, you can navigate away to other pages and you will be notified once the job is completed.

Job type	Name	Created	Status
BLAST	sp P01308 INS_HUMAN	Created	<div> <div></div> <div> <p>We will notify you when your results are ready</p> <p>Target database: UniProtKB reference proteomes + Swiss-Prot</p> </div> </div>

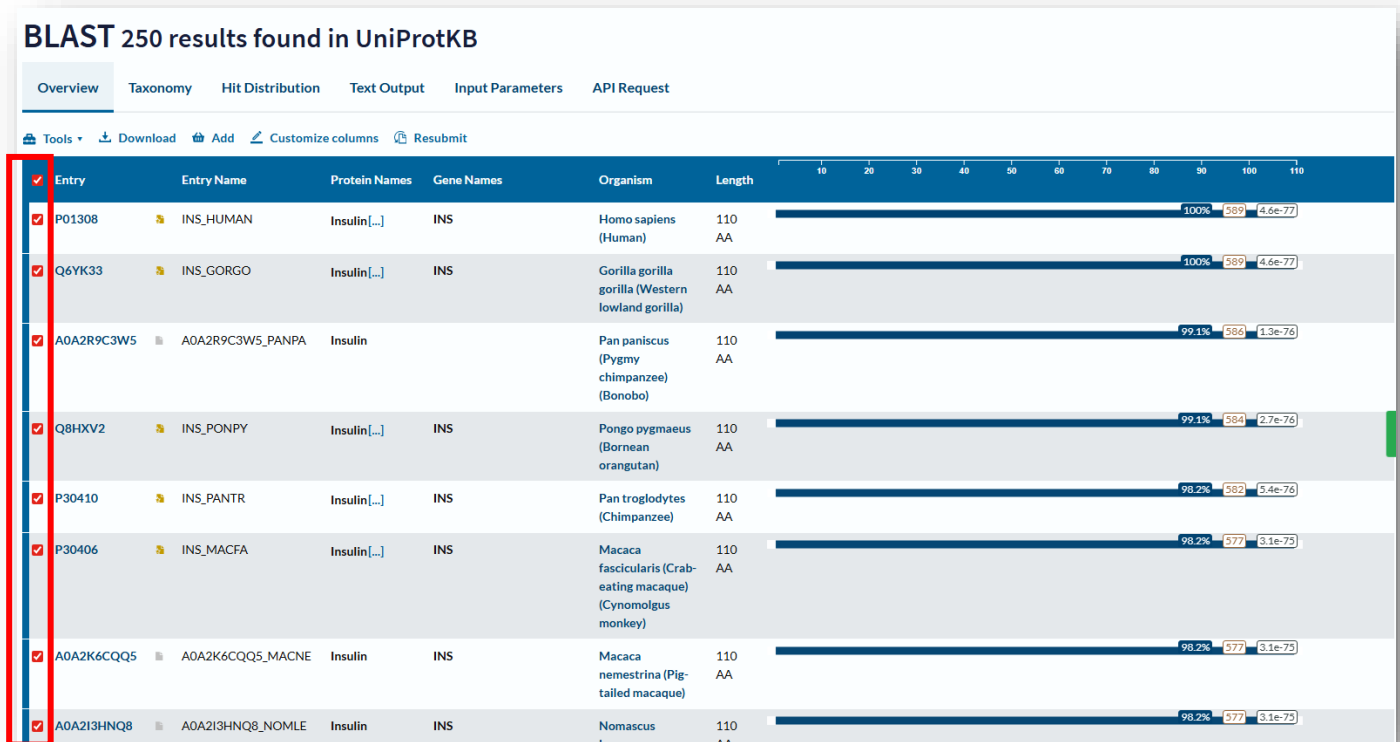
The server has not accepted this job yet

<https://github.com/code-aradhana/bioinformatics-msa-project.git>

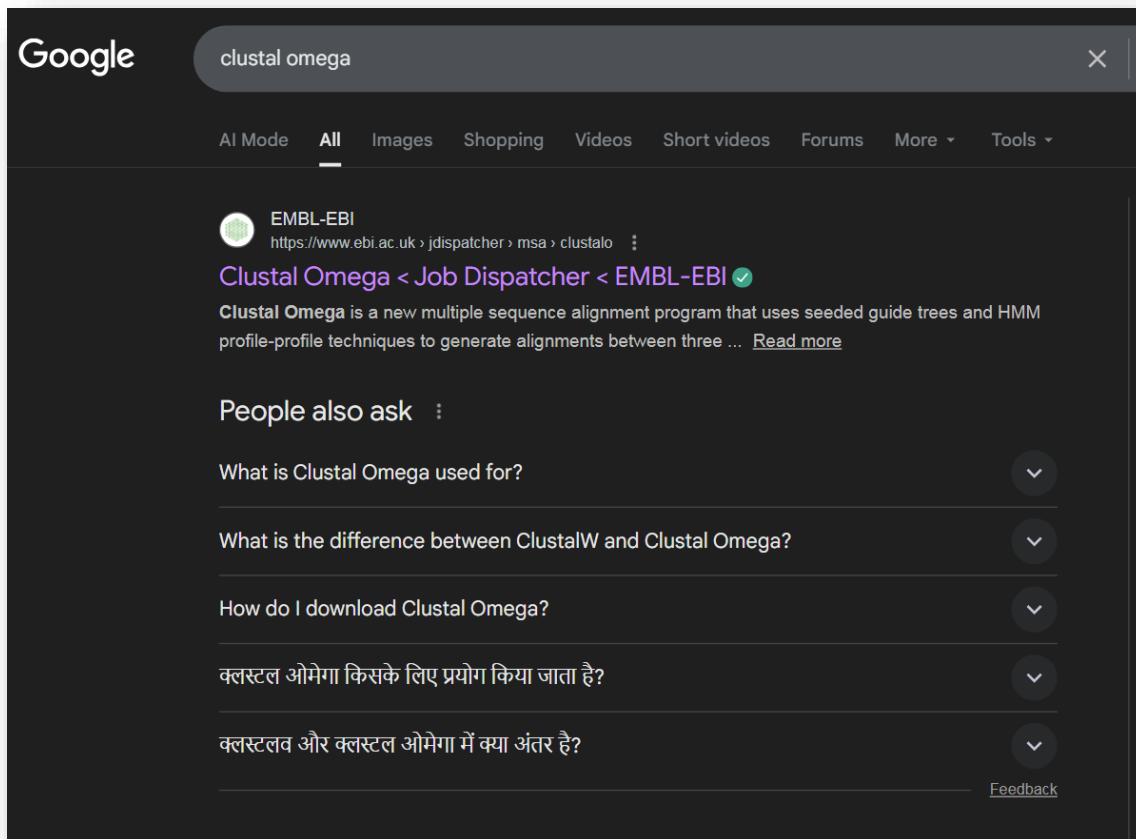
Wait for few minutes and then results will appear.



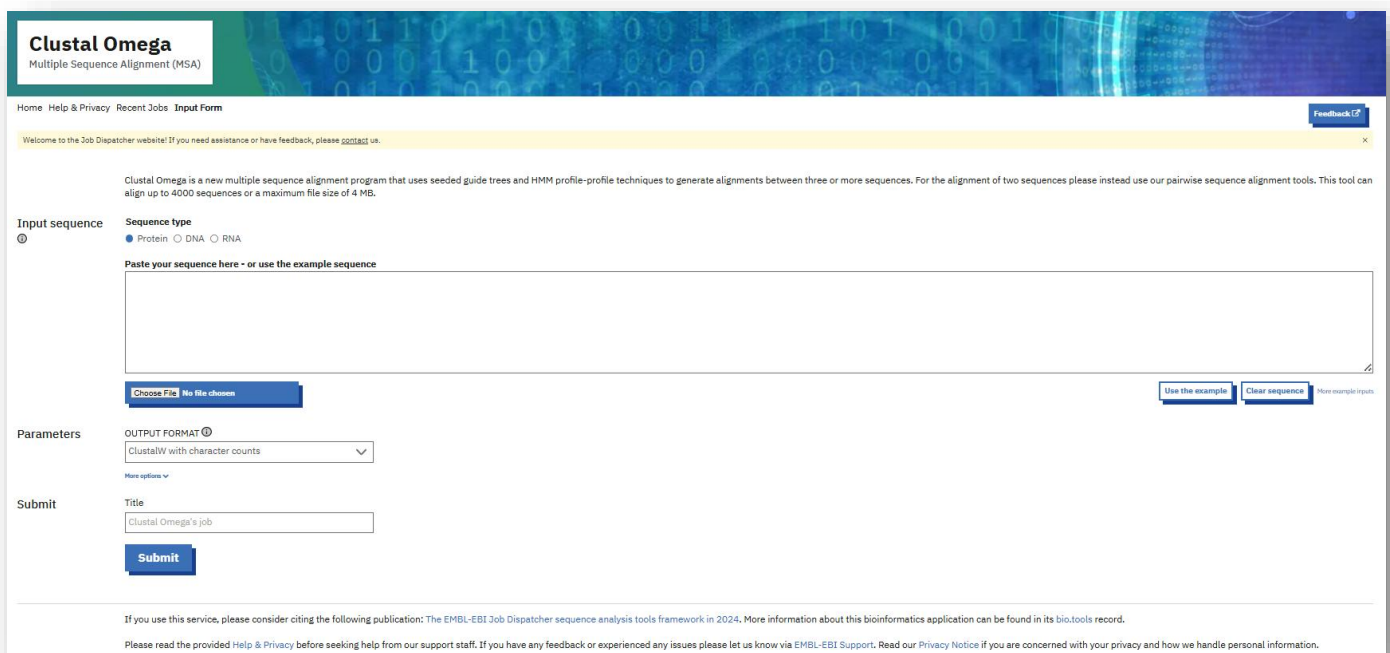
10. We are selecting **20 sequences** and **download it**. Download the Multiple Sequence Alignment.



## 11. Search **Clustal Omega** on google.



## 12. Open the **Clustal Omega**.



### 13. Paste all the 20 sequences in the query box.

**Clustal Omega**  
Multiple Sequence Alignment (MSA)

Home Help & Privacy Recent Jobs Input Form

Welcome to the Job Dispatcher website! If you need assistance or have feedback, please [contact us](#).

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools. This tool can align up to 4000 sequences or a maximum file size of 4 MB.

**Input sequence**

**Sequence type**  
☒ Protein ☐ DNA ☐ RNA

Paste your sequence here - or use the example sequence

```

MPCVYKCEPCLNGLVWGPDPAPFVNHQHLGSHLVEALYVCGRGFFYTPKTRREAED
PQVGVVGLGGGPGAGSLQPLALEGSLQKRGIVEQCTSCISLYQLENYCN
>tr|A0A2I3HNQ8|A0A2I3HNQ8_NOMLE Insulin OS=Nomascus leucogenys OX=61853 GN=INS PE=3 SV=1
MALWMRLRLPLLALLALWGPDPAPFVNHQHLGSHLVEALYVCGRGFFYTPKTRREAED
PQVGVVGLGGGPGAGSLQPLALEGSLQKRGIVEQCTSCISLYQLENYCN
>tr|A0A2K6CQ05|A0A2K6CQ05_MACNE Insulin OS=Macaca nemestrina OX=9545 GN=INS PE=3 SV=1
MALWMRLRLPLLALLALWGPDPAPFVNHQHLGSHLVEALYVCGRGFFYTPKTRREAED
PQVGVVGLGGGPGAGSLQPLALEGSLQKRGIVEQCTSCISLYQLENYCN
  
```

[Choose File](#) [No file chosen](#) [Use the example](#) [Clear sequence](#) [More example inputs](#)

**Parameters**

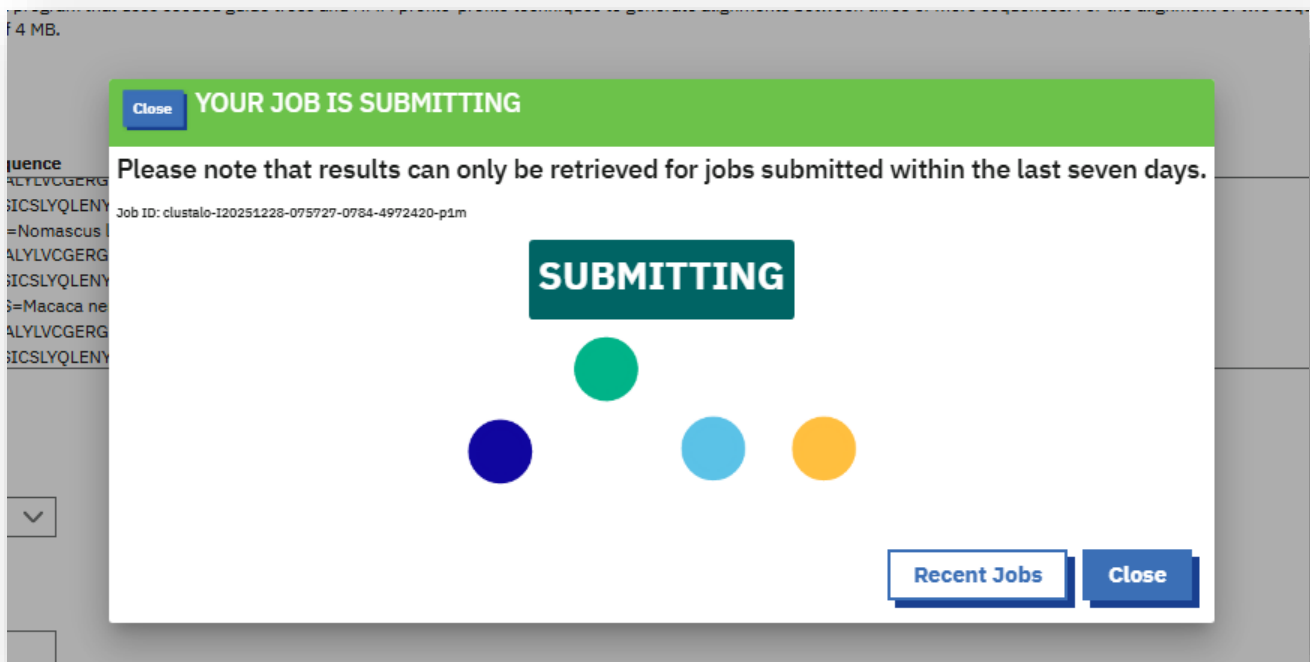
OUTPUT FORMAT [ⓘ](#)  
 ClustalW with character counts
 [More options](#)

**Submit**

Title  
 Clustal Omega's job

[Submit](#)

### 14. Again, a **waiting pop-up** appear. Wait until the results appear on the screen. Clustal omega sometime take longer time to show results.





## 15. Under **Tool Output** results will shown up.

### Tool output

[Download](#)

CLUSTAL O(1.2.4) multiple sequence alignment

```

tx|A0A8I5TQT5|A0A8I5TQT5_PONAB      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    59
sp|P3G410|INS_PANTR                  MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
sp|P01308|INS_HUMAN                  MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
sp|Q6YK33|INS_GORGO                  MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A2R9C3W5|A0A2R9C3W5_PANPA      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
sp|Q8HXV2|INS_PONPY                  MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A2K5P2L3|A0A2K5P2L3_CERAT      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A8C9LMF1|A0A8C9LMF1_9PRIM      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A8D2G8B4|A0A8D2G8B4_THEGE      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A2K6R041|A0A2K6R041_RHIRO      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
sp|P3G406|INS_MACFA                  MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|F7AUL3|F7AUL3_MACMU              MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A2I3HNQ8|A0A2I3HNQ8_NOMLE      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A2K6CQ05|A0A2K6CQ05_MACNE      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A096MTW9|A0A096MTW9_PAPAN      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A0D9RBQ0|A0A0D9RBQ0_CHLSB      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A2K5YKV7|A0A2K5YKV7_MANLE      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0AAJ7MUG7|A0AAJ7MUG7_RHIBE      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
tx|A0A2K5JZH7|A0A2K5JZH7_COLAP      MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
sp|P3G407|INS_CHLAE                  MALWMRLLPLLALLALWGPDPAA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED    60
*****

tx|A0A8I5TQT5|A0A8I5TQT5_PONAB      LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    109
sp|P3G410|INS_PANTR                  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
sp|P01308|INS_HUMAN                  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
sp|Q6YK33|INS_GORGO                  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A2R9C3W5|A0A2R9C3W5_PANPA      LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
sp|Q8HXV2|INS_PONPY                  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A2K5P2L3|A0A2K5P2L3_CERAT      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A8C9LMF1|A0A8C9LMF1_9PRIM      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A8D2G8B4|A0A8D2G8B4_THEGE      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A2K6R041|A0A2K6R041_RHIRO      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
sp|P3G406|INS_MACFA                  PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|F7AUL3|F7AUL3_MACMU              PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A2I3HNQ8|A0A2I3HNQ8_NOMLE      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A2K6CQ05|A0A2K6CQ05_MACNE      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A096MTW9|A0A096MTW9_PAPAN      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A0D9RBQ0|A0A0D9RBQ0_CHLSB      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A2K5YKV7|A0A2K5YKV7_MANLE      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0AAJ7MUG7|A0AAJ7MUG7_RHIBE      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tx|A0A2K5JZH7|A0A2K5JZH7_COLAP      PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
sp|P3G407|INS_CHLAE                  PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
*****

```

## 16. Scroll down to have coloured sequences.

### Coloured sequences

CLUSTAL O(1.2.4) multiple sequence alignment

Hide

```

tr|A0A8I5TQT5|A0A8I5TQT5_PONAB      MALWMRLLPLLALLALWGPDPAA-AFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    59
sp|P3G410|INS_PANTR                  MALWMRLLPLLALLALWGPDPAAAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
sp|P01308|INS_HUMAN                   MALWMRLLPLLALLALWGPDPAAAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
sp|Q6YK33|INS_GORGO                   MALWMRLLPLLALLALWGPDPAAAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A2R9C3W5|A0A2R9C3W5_PANPA       MALWMRLLPLLALLALWGPDPAAAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
sp|Q8HXV2|INS_PONPY                   MALWMRLLPLLALLALWGPDPAAAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A2K5P2L3|A0A2K5P2L3_CERAT        MALWMRLLPLLALLALWGPDPVPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A8C9LMF1|A0A8C9LMF1_9PRIM        MALWMRLLPLLALLALWGPDPVPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A8D2G8B4|A0A8D2G8B4_THEGE       MALWMRLLPLLALLALWGPDSVPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A2K6R041|A0A2K6R041_RHIRO       MALWMRLLPLLALLALWGPDPVPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
sp|P3G406|INS_MACFA                   MALWMRLLPLLALLALWGPDPAPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|F7AUL3|F7AUL3_MACMU                MALWMRLLPLLALLALWGPDPAPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A2I3HNQ8|A0A2I3HNQ8_NOMLE       MALWMRLLPLLALLALWGPDPAPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A2K6CQ05|A0A2K6CQ05_MACNE       MALWMRLLPLLALLALWGPDPAPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A096MTW9|A0A096MTW9_PAPAN        MALWMRLLPLLALLALWGPDPVPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A0D9RBQ0|A0A0D9RBQ0_CHLSB       MALWMRLLPLLALLALWGPDPVPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A2K5YKV7|A0A2K5YKV7_MANLE       MALWMRLLPLLALLALWGPDPVPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0AAJ7MUG7|A0AAJ7MUG7_RHIBE       MALWMRLLPLLALLALWGPDPVPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
tr|A0A2K5JZH7|A0A2K5JZH7_COLAP       MALWMRLLPLLALLALWGPDPVPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
sp|P3G407|INS_CHLAE                   MALWMRLLPLLALLALWGPDPVPAFVNQHLGSHLVEALYLVCGERGFFYTPKTRREAED    60
*****

tr|A0A8I5TQT5|A0A8I5TQT5_PONAB      LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    109
sp|P3G410|INS_PANTR                  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
sp|P01308|INS_HUMAN                   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
sp|Q6YK33|INS_GORGO                   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A2R9C3W5|A0A2R9C3W5_PANPA       LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
sp|Q8HXV2|INS_PONPY                   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A2K5P2L3|A0A2K5P2L3_CERAT        PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A8C9LMF1|A0A8C9LMF1_9PRIM        PQVGQVELGGGPGTGSQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A8D2G8B4|A0A8D2G8B4_THEGE       PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A2K6R041|A0A2K6R041_RHIRO       PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
sp|P3G406|INS_MACFA                   PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|F7AUL3|F7AUL3_MACMU                PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A2I3HNQ8|A0A2I3HNQ8_NOMLE       PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A2K6CQ05|A0A2K6CQ05_MACNE       PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A096MTW9|A0A096MTW9_PAPAN        PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A0D9RBQ0|A0A0D9RBQ0_CHLSB       PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A2K5YKV7|A0A2K5YKV7_MANLE       PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0AAJ7MUG7|A0AAJ7MUG7_RHIBE       PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
tr|A0A2K5JZH7|A0A2K5JZH7_COLAP       PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
sp|P3G407|INS_CHLAE                   PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIICSLYQLENYCN    110
*****

```

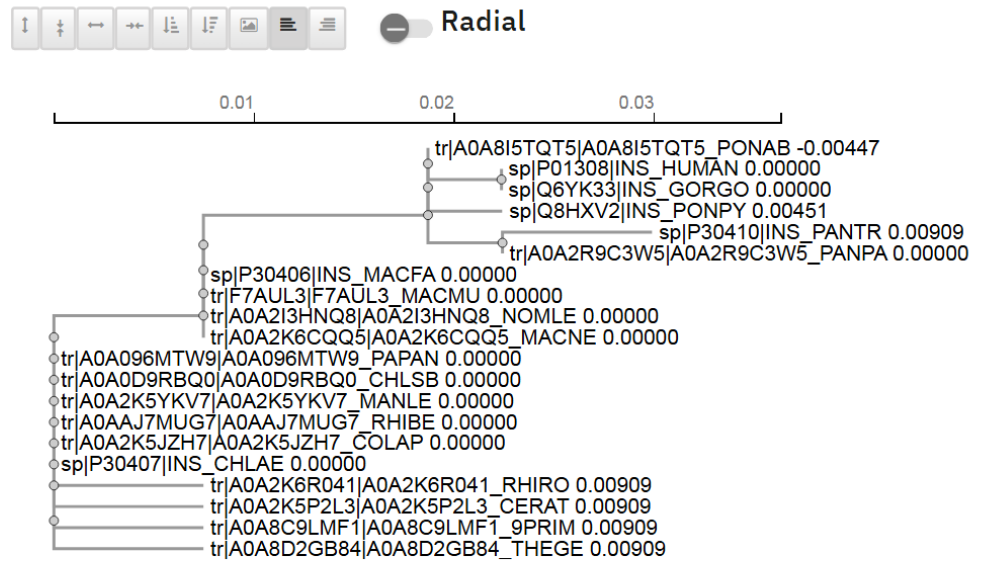
17. Click the **Phylogenetic Tree** in horizontal tabular menu bar.

## Phylogenetic Tree

[illegible]

## 18. Scroll down to have **Phylogram**.

### Phylogram



## Results of Multiple Sequences Alignment of 20 sequences -

```
>sp|P01308|INS_HUMAN Insulin OS=Homo sapiens OX=9606 GN=INS PE=1
SV=1MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>sp|Q6YK33|INS_GORGO Insulin OS=Gorilla gorilla gorilla OX=9595 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>tr|A0A2R9C3W5|A0A2R9C3W5_PANPA Insulin OS=Pan paniscus OX=9597 PE=3
SV=1MALWMRLLPLLALLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>sp|Q8HXV2|INS_PONPY Insulin OS=Pongo pygmaeus OX=9600 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>sp|P30410|INS_PANTR Insulin OS=Pan troglodytes OX=9598 GN=INS PE=1
SV=1MALWMRLLPLLALLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>sp|P30406|INS_MACFA Insulin OS=Macaca fascicularis OX=9541 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>tr|A0A2K6R041|A0A2K6R041_RHIRO Insulin OS=Rhinopithecus roxellana OX=61622 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
CN

>tr|A0A8D2GB84|A0A8D2GB84_THEGE Insulin OS=Theropithecus gelada OX=9565 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDSVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>tr|A0A8C9LMF1|A0A8C9LMF1_9PRIM Insulin OS=Ptilocolobus tephrosceles OX=591936 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>tr|A0A2K5P2L3|A0A2K5P2L3_CERAT Insulin OS=Cercopithecus atys OX=9531 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>tr|A0A096MTW9|A0A096MTW9_PAPAN Insulin OS=Papio anubis OX=9555 GN=INS PE=3
SV=2MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>tr|A0A0D9RBQ0|A0A0D9RBQ0_CHLSB Insulin OS=Chlorocebus sabaeus OX=60711 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>tr|A0A2K5YKV7|A0A2K5YKV7_MANLE Insulin OS=Mandrillus leucophaeus OX=9568 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>tr|A0AAJ7MUG7|A0AAJ7MUG7_RHIBE Insulin OS=Rhinopithecus bieti OX=61621 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>tr|A0A2K5JZH7|A0A2K5JZH7_COLAP Insulin OS=Colobus angolensis palliatus OX=336983 PE=3
SV=1MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>sp|P30407|INS_CHLAE Insulin OS=Chlorocebus aethiops OX=9534 GN=INS PE=1
SV=1MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN

>tr|A0A8I5TQT5|A0A8I5TQT5_PONAB Insulin OS=Pongo abelii OX=9601 GN=INS PE=3
SV=1MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
CN

>tr|F7AUL3|F7AUL3_MACMU Insulin OS=Macaca mulatta OX=9544 GN=INS PE=3
SV=3MALWMRLLPLLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSI
SLYQLEN
YCN
```

<https://github.com/code-aradhana/bioinformatics-msa-project.git>

```
>tr|A0A2I3HNQ8|A0A2I3HNQ8_NOMLE Insulin OS=Nomascus leucogenys OX=61853 GN=INS PE=3  
SV=1MALWMRLLPLLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLEN  
YCN
```

```
>tr|A0A2K6CQQ5|A0A2K6CQQ5_MACNE Insulin OS=Macaca nemestrina OX=9545 GN=INS PE=3  
SV=1MALWMRLLPLLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLEN  
YCN
```

# Application of Multiple Sequence Alignment

- ✚ Recombinant protein synthesis
- ✚ Drugs production
- ✚ Antibiotic production
- ✚ Functional genomics
- ✚ Determination of protein folding patterns in bioinformatics
- ✚ It plays vital role in proteomics.
- ✚ Used for the prediction of final structure, function and location of protein.
- ✚ To find out location of gene coding for that protein.
- ✚ Genetic diseases.
- ✚ Identification of sequence differences and variations such as point mutations.
- ✚ Revealing the evolution and genetic diversity of sequence and organisms.

# Conclusion

This project successfully demonstrated the fundamental bioinformatics workflow for identifying evolutionarily conserved regions in protein sequences using Multiple Sequence Alignment (MSA). The primary objective—to align homologous insulin protein sequences from various primate species and analyze patterns of conservation—was achieved through a systematic process.

The practical steps involved retrieving target sequences (Human Insulin, P01308) from the UniProt database, using BLAST to find closely related homologous sequences, and performing the alignment with the Clustal Omega tool. The resulting MSA of 20 primate insulin sequences provided a clear visual and data-driven output.

In summary, this project confirms that Multiple Sequence Alignment is an indispensable tool in computational biology. It provides a powerful lens to visualize evolution's fingerprint on protein sequences, enabling researchers to infer function, trace ancestry, and form hypotheses for experimental validation. The skills developed—from sequence retrieval and database searching to alignment execution and interpretation—form an essential foundation for any subsequent bioinformatics analysis.



# Reference

- Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids  
Authors: Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchison.
- <https://www.scribd.com/document/479739675/Multiple-Sequence-Alignment>
- <https://www.ebi.ac.uk/jdispatcher/msa>
- Mount, D. W. (2004). Bioinformatics: Sequence and Genome Analysis (2nd ed.). Cold Spring Harbor Laboratory Press.
- **Lesk, A. M. (2017).** *Introduction to Protein Science: Architecture, Function, and Genomics* (3rd ed.). Oxford University Press.
- Felsenstein, J. (2004). Inferring Phylogenies. Sinauer Associates.  
A comprehensive guide to phylogenetic methods, emphasizing the role of conserved sequences in evolutionary studies.
- **UniProt Consortium (2023).** UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Research*, 51(D1), D523–D531.