

Multiple Sequence Alignments

BIOINFORMATICS

TOPIC

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

INDEX

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

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

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

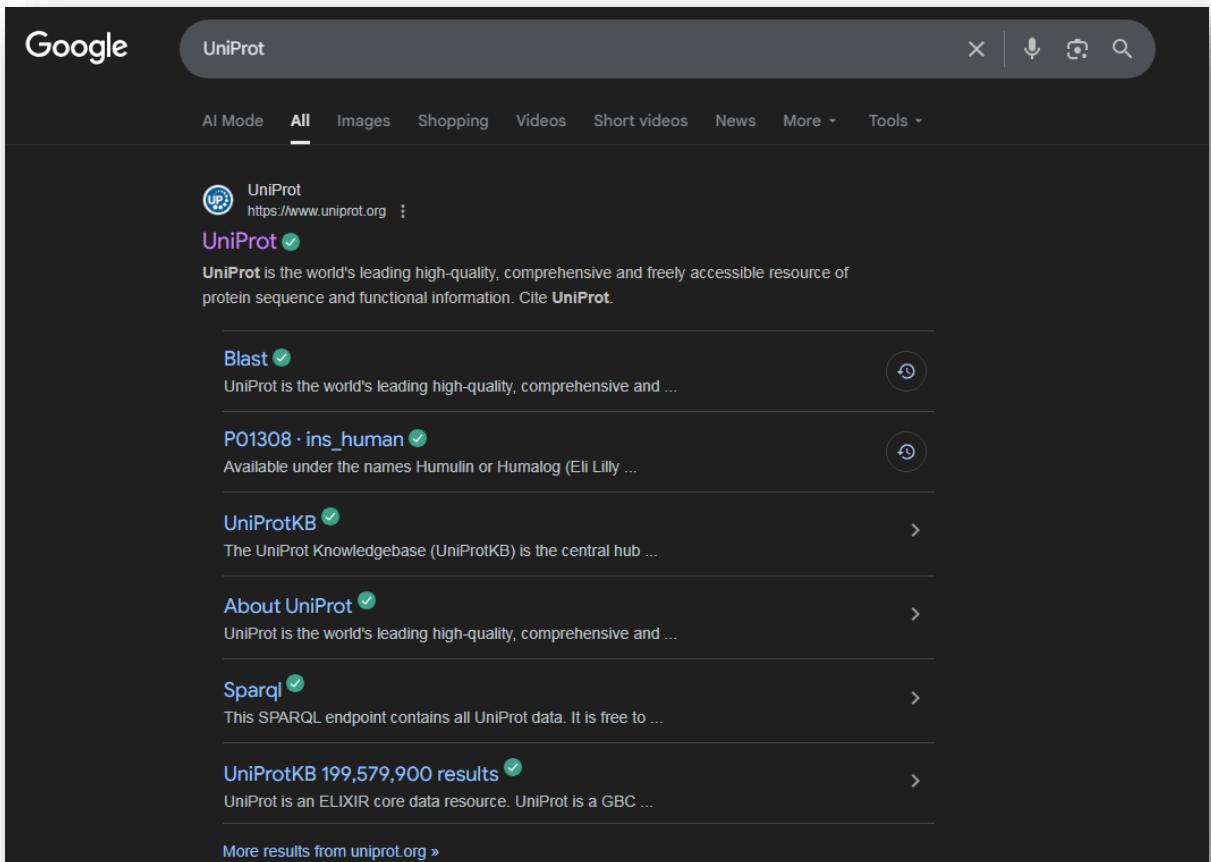
The EBI has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions.

Launch [WebPRANK](#)

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

Steps In MSA

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



A screenshot of a Google search results page for the query "UniProt". The results are displayed in a dark-themed interface. The first result is the official UniProt website, followed by links to various UniProt-related resources like Blast, P01308, UniProtKB, About UniProt, Sparql, and UniProtKB statistics. A link to "More results from uniprot.org" is also present.

- UniProt**
https://www.uniprot.org
- Blast**
- P01308 · ins_human**
- UniProtKB**
- About UniProt**
- Sparql**
- UniProtKB 199,579,900 results**
- More results from uniprot.org >

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

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

The screenshot shows the UniProtKB search results for the query 'Insulin'. The top navigation bar includes links for UniProt, BLAST, Align, Peptide search, ID mapping, and SPARQL. On the right, there are links for 'Release 2025_04 | Statistics' and 'Help'. The main search bar contains 'UniProtKB • Insulin'. Below it, a sub-bar shows 'Examples: Insulin, APP, Human, P05067, organism_id:9606'. The central area features a banner with the text: 'Our Proteomes and UniProtKB/TrEMBL resources are undergoing a significant transition. Please read our help page, view affected entries and proteomes, or contact us with any questions.' Below the banner are four colored boxes: 'Proteins UniProt Knowledgebase' (blue), 'Species Proteomes' (red), 'Protein Clusters UniRef' (orange), and 'Sequence archive UniParc' (green). A message at the bottom of the page states: '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.' A 'I agree, dismiss this banner' button is present.

3. Results will appear in this format.

The screenshot shows the detailed results for 'Insulin' on the UniProtKB site. The left sidebar includes sections for Status (Reviewed (Swiss-Prot) 5,327, Unreviewed (TrEMBL) 287,079), Popular organisms (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), and a 'More' link. The main content area lists several protein entries under the heading 'P01308 - INS_HUMAN': P01308 - INS_HUMAN, Insulin - Gene: INS - Homo sapiens (Human) - 110 amino acids - Evidence at protein level - Annotation score: 62. It also lists P06213 - INSR_HUMAN, P14735 - IDE_HUMAN, P01317 - INS_BOVIN, P01321 - INS_CANLF, and P01320 - INC_CAVIO. Each entry provides a brief description, annotation score, and a list of associated keywords. A message at the bottom of the page states: '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.' A 'I agree, dismiss this banner' button is present.

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

4. Select the required item from the list.

The screenshot shows the UniProt search results for the query 'Insulin'. The results are listed in a table with columns for ID, Name, Status, Evidence, Annotation score, and Number of entries. The first result, P01308 · INS_HUMAN, is highlighted with a red border. This entry details the protein's properties: Gene: INS · Homo sapiens (Human) · 110 amino acids · Evidence at protein level · Annotation score: 55. It also lists interactions, PTMs, and publications. Other results include P06213 · INSR_HUMAN, P14735 · IDE_HUMAN, P01317 · INS_BOVIN, and P01321 · INS_CANLF.

5. Page will appear like this.

This screenshot shows the detailed view of the UniProt entry for P01308 · INS_HUMAN. The top navigation bar includes links for BLAST, Align, Peptide search, ID mapping, SPARQL, and UniProtKB. The main content area displays the protein's name, status (Reviewed (Swiss-Prot)), and various annotations such as gene (INS), protein (Insulin), and organism (Homo sapiens (Human)). Below this, tabs for Function, Gene Ontology, and PTM/Processing are visible. The Gene Ontology section is expanded, showing a tree diagram of biological processes. A sidebar on the right provides links for Advanced, List, Search, and Help.

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

6. Click on Download.

P01308 · INS_HUMAN

Protein | Insulin
Gene | INS
Status | UniProtKB reviewed (Swiss-Prot)
Organism | Homo sapiens (Human)

Amino acids | 110 (go to sequence)
Protein existence | Evidence at protein level
Annotation score | 63

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 Ontology¹

GO annotations GO-CAM models

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

Car color indicates the number of GO terms

ASPECT	TERM	Source	Count
Molecular Function	hormone activity ¹	Source:BP UCL	1 Publication
Molecular Function	identical protein binding ¹	Source:CAZ	1 Publication
Molecular Function	insulin receptor binding ¹	Source:ChEMBL	1 Publication
Molecular Function	insulin-like growth factor receptor binding ¹	Source:BP UCL	1 Publication
Molecular Function	protease binding ¹	Source:Pro9DB	1 Publication
Biological Process	activation of protein kinase B activity ¹	Source:BP UCL	1 Publication
Biological Process	acute-phase response ¹	Source:BP UCL	1 Publication
Biological Process	alpha-beta T cell activation ¹	Source:UniProtB	1 Publication
Biological Process	cell-cell signaling ¹	Source:UniProtB	1 Publication
Biological Process	cognition ¹	Source:AU UCL	1 Publication
Biological Process	metabolic process ¹	Source:UniProtB	1 Publication

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

Dataset¹
Entry

Format
FASTA (canonical)

Download

Generate URL for API | Preview | Cancel

Swiss-Prot

Amino acids | 110 (go to sequence)
Protein existence | Evidence at protein level
Annotation score | 63

Genomic coordinates Publications External links History

(1) Add a publication Entry feedback

is cell permeability to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver.

et.

In our Privacy Notice | I agree, dismiss this banner

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

7. Click on Tools.

P01308 · INS_HUMAN

Protein | Insulin
Gene | INS
Status | UnProtKB reviewed (Swiss-Prot)
Organism | Homo sapiens (Human)

Amino acids | 110 (go to sequence)
Protein existence | Evidence at protein level
Annotation score | 0.5

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)

Cal color indicates relative number of GO terms

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

A menu will appear and then select BLAST.

P01308 · INS_HUMAN

Protein | Insulin
Gene | INS
Status | UnProtKB reviewed (Swiss-Prot)
Organism | Homo sapiens (Human)

Amino acids | 110 (go to sequence)
Protein existence | Evidence at protein level
Annotation score | 0.5

Entry Variant viewer 178 Feature viewer Genomic coordinates Publications External links History

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

BLAST (1)

Align isoforms (2)
Map ID (1)

Function'
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)

Cal color indicates relative number of GO terms

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

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

8. Scroll down and click Run BLAST.

The screenshot shows the NCBI BLAST search interface. At the top, there's a search bar for UniProt IDs and a text area for entering sequences. Below that, there's an 'OR' section for pasting sequences. A sequence for Insulin from Homo sapiens (P01308) is pasted into this area. A message below says 'Your input contains 1 sequence.' Under 'Target database', 'UniProtKB reference proteomes + Swiss-Prot' is selected. There's also a 'Restrict by taxonomy' field where you can enter taxon names or IDs. In the 'Name your BLAST job' field, 'sp | P01308 | INS_HUMAN' is entered. The 'Advanced parameters' section includes dropdowns for Sequence type (Protein), Program (blastp), E-Threshold (10), Matrix (Auto - BLOSUM62), Filter (None), Gapped (yes), and Hits (250). The 'HSPs per hit' dropdown is set to 'All'. At the bottom right of the form, there are 'Reset' and 'Run BLAST' buttons, with the 'Run BLAST' button being highlighted with a red box.

9. A page will come as shown below.

The screenshot shows the 'Tool results' page. It displays a table of tool analysis results from the last 7 days. A single row is shown for a pending BLAST job. The job details are as follows:

Job type	Name	Created	Status
BLAST	sp P01308 INS_HUMAN	Created (○)	We will notify you when your results are ready

Below the table, it says 'The server has not accepted this job yet' and 'Target database: UniProtKB reference proteomes + Swiss-Prot'. There are also small icons for file operations like download and delete.

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

Wait for few minutes and then results will appear.

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
P01308	INS_HUMAN	Insulin[...]	INS	Homo sapiens (Human)	110
Q6YK33	INS_GORGO	Insulin[...]	INS	Gorilla gorilla gorilla (Western lowland gorilla)	110
A0A2R9C3W5	A0A2R9C3W5_PANPA	Insulin		Pan paniscus (Pygmy chimpanzee) (Bonobo)	110
Q8HXV2	INS_PONPY	Insulin[...]	INS	Pongo pygmaeus (Bornean orangutan)	110
P30410	INS_PANTR	Insulin[...]	INS	Pan troglodytes (Chimpanzee)	110
P30406	INS_MACFA	Insulin[...]	INS	Macaca fascicularis (Crab-eating macaque) (Cynomolgus monkey)	110
A0A2K6CQQ5	A0A2K6CQQ5_MACNE	Insulin	INS	Macaca nemestrina (Pig-tailed macaque)	110
A0A2I3HNQ8	A0A2I3HNQ8_NOMLE	Insulin	INS	Nomascus leucogenys	110

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

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
P01308	INS_HUMAN	Insulin[...]	INS	Homo sapiens (Human)	110
Q6YK33	INS_GORGO	Insulin[...]	INS	Gorilla gorilla gorilla (Western lowland gorilla)	110
A0A2R9C3W5	A0A2R9C3W5_PANPA	Insulin		Pan paniscus (Pygmy chimpanzee) (Bonobo)	110
Q8HXV2	INS_PONPY	Insulin[...]	INS	Pongo pygmaeus (Bornean orangutan)	110
P30410	INS_PANTR	Insulin[...]	INS	Pan troglodytes (Chimpanzee)	110
P30406	INS_MACFA	Insulin[...]	INS	Macaca fascicularis (Crab-eating macaque) (Cynomolgus monkey)	110
A0A2K6CQQ5	A0A2K6CQQ5_MACNE	Insulin	INS	Macaca nemestrina (Pig-tailed macaque)	110
A0A2I3HNQ8	A0A2I3HNQ8_NOMLE	Insulin	INS	Nomascus leucogenys	110

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

11. Search Clustal Omega on google.

The screenshot shows a Google search results page with a dark theme. The search query 'clustal omega' is entered in the search bar. The top result is from EMBL-EBI, titled 'Clustal Omega < Job Dispatcher < EMBL-EBI'. Below the title, a snippet of text reads: 'Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three ... [Read more](#)'. Under the heading 'People also ask :', there are five expandable questions with arrows: 'What is Clustal Omega used for?', 'What is the difference between ClustalW and Clustal Omega?', 'How do I download Clustal Omega?', 'क्लस्टल ओमेगा किसके लिए प्रयोग किया जाता है?', and 'क्लस्टलव और क्लस्टल ओमेगा में क्या अंतर है?'. At the bottom right of the search results, there is a 'Feedback' link.

12. Open the Clustal Omega.

The screenshot shows the 'Clustal Omega' input form on the EMBL-EBI website. The header says 'Multiple Sequence Alignment (MSA)'. The main area has sections for 'Input sequence', 'Parameters', and 'Submit'. In the 'Input sequence' section, there is a text area labeled 'Paste your sequence here - or use the example sequence' with a placeholder 'Sequence type' (radio buttons for Protein, DNA, RNA, with Protein selected). Below it is a file upload field 'Choose File No file chosen' and buttons 'Use the example' and 'Clear sequence'. In the 'Parameters' section, there is a dropdown menu 'OUTPUT FORMAT' set to 'ClustalW with character counts' and a 'More options' link. In the 'Submit' section, there is a text field 'Title' containing 'Clustal Omega's job' and a blue 'Submit' button. At the bottom, there is a note about citation and terms of use.

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

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

Clustal Omega
Multiple Sequence Alignment (MSA)

Home Help & Privacy Recent Jobs Input Form Feedback

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 Protein DNA RNA

Sequence type Protein DNA RNA

Paste your sequence here - or use the example sequence

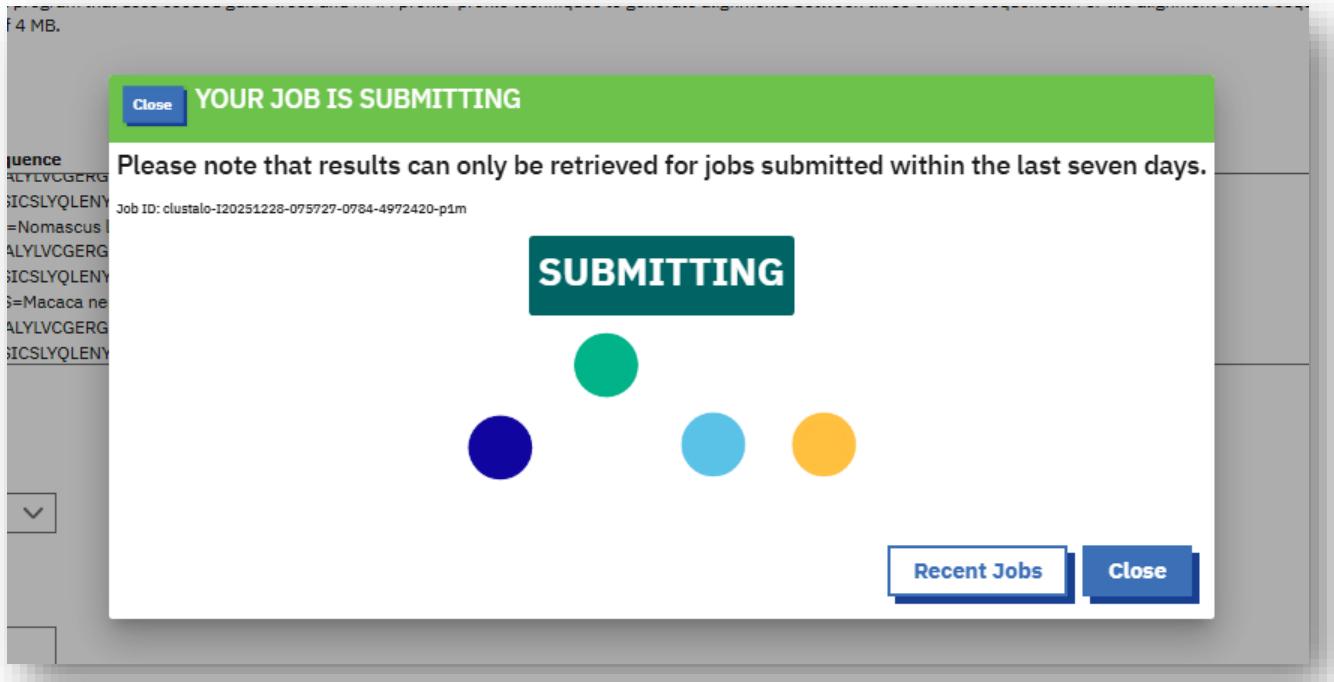
```
MALWMLRLLPALLALALVALGVPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAD
PQVGVQVELGGPGAGSLQPLALEGSLSQRGRGEVEQCCTSCISLYOLENYCN
>tr|AA213HNQ|8|AOA213HNQ8_NOMLE Insulin OS=Nomascus leucogenys OX=61853 GN=INS PE=3 SV=1
MALWMLRLLPALLALALVALGVPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAD
PQVGVQVELGGPGAGSLQPLALEGSLSQRGRGEVEQCCTSCISLYOLENYCN
>tr|AOA216C6QQ|5|AOA216C6QQ5_MACNE Insulin OS=Macaca nemestrina OX=9545 GN=INS PE=3 SV=1
MALWMLRLLPALLALALVALGVPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAD
PQVGVQVELGGPGAGSLQPLALEGSLSQRGRGEVEQCCTSCISLYOLENYCN
```

Choose File No file chosen Use the example More example inputs

Parameters OUTPUT FORMAT ClustalW with character counts

Submit Title Clustal Omega's job

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		CLUSTAL O(1.2.4) multiple sequence alignment	
Download			
tr A0A8I5TQT5 A0A8I5TQT5_PONAB	MALWMRLLPLALLALWGPDPAAFPVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	59	
sp P30410 INS_PANTR	MALWMRLLPLLVALLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
sp P01308 INS_HUMAN	MALWMRLLPLLVALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
sp Q6YK3S INS_GORGO	MALWMRLLPLLVALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A2R9C3W5 A0A2R9C3W5_PANPA	MALWMRLLPLLVALLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
sp Q8HXV2 INS_PONPY	MALWMRLLPLLVALLALWGPDPQAQFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A2K5P2L3 A0A2K5P2L3_CERAT	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A8C9LMF4 A0A8C9LMF1_9PRIM	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A8D2GB84 A0A8D2GB84_THEGE	MALWMRLLPLLVALLALWGPDPVSVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A2K6R041 A0A2K6R041_RHIRO	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
sp P30406 INS_MACFA	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr F7AUL8 F7AUL8_MACMU	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A2I3HNQ8 A0A2I3HNQ8_NOMLE	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A2K6CQ05 A0A2K6CQ05_MACNE	MALWMRLLPLLVALLALWGPDPAPAPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A096MTW9 A0A096MTW9_PAPAN	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A0D9RBQ0 A0A0D9RBQ0_CHLSB	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A2K5YKV7 A0A2K5YKV7_MANLE	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0AAJ7MUG7 A0AAJ7MUG7_RHIBE	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
tr A0A2K5JZH7 A0A2K5JZH7_COLAP	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
sp P30407 INS_CHLAE	MALWMRLLPLLVALLALWGPDPVPVFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEAD	60	
*****.***** . ***** . *****.*****			
tr A0A8I5TQT5 A0A8I5TQT5_PONAB	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 109		
sp P30410 INS_PANTR	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
sp P01308 INS_HUMAN	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
sp Q6YK3S INS_GORGO	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A2R9C3W5 A0A2R9C3W5_PANPA	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
sp Q8HXV2 INS_PONPY	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A2K5P2L3 A0A2K5P2L3_CERAT	PQVGQVELGGPGAGSLQPLSLEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A8C9LMF4 A0A8C9LMF1_9PRIM	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A8D2GB84 A0A8D2GB84_THEGE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A2K6R041 A0A2K6R041_RHIRO	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
sp P30406 INS_MACFA	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr F7AUL8 F7AUL8_MACMU	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A2I3HNQ8 A0A2I3HNQ8_NOMLE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A2K6CQ05 A0A2K6CQ05_MACNE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A096MTW9 A0A096MTW9_PAPAN	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A0D9RBQ0 A0A0D9RBQ0_CHLSB	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A2K5YKV7 A0A2K5YKV7_MANLE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0AAJ7MUG7 A0AAJ7MUG7_RHIBE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
tr A0A2K5JZH7 A0A2K5JZH7_COLAP	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
sp P30407 INS_CHLAE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCNC 110		
*****.***** ; ***** ; *****.*****			

16. Scroll down to have coloured sequences.

Coloured sequences CLUSTAL O(1.2.4) multiple sequence alignment

Hide

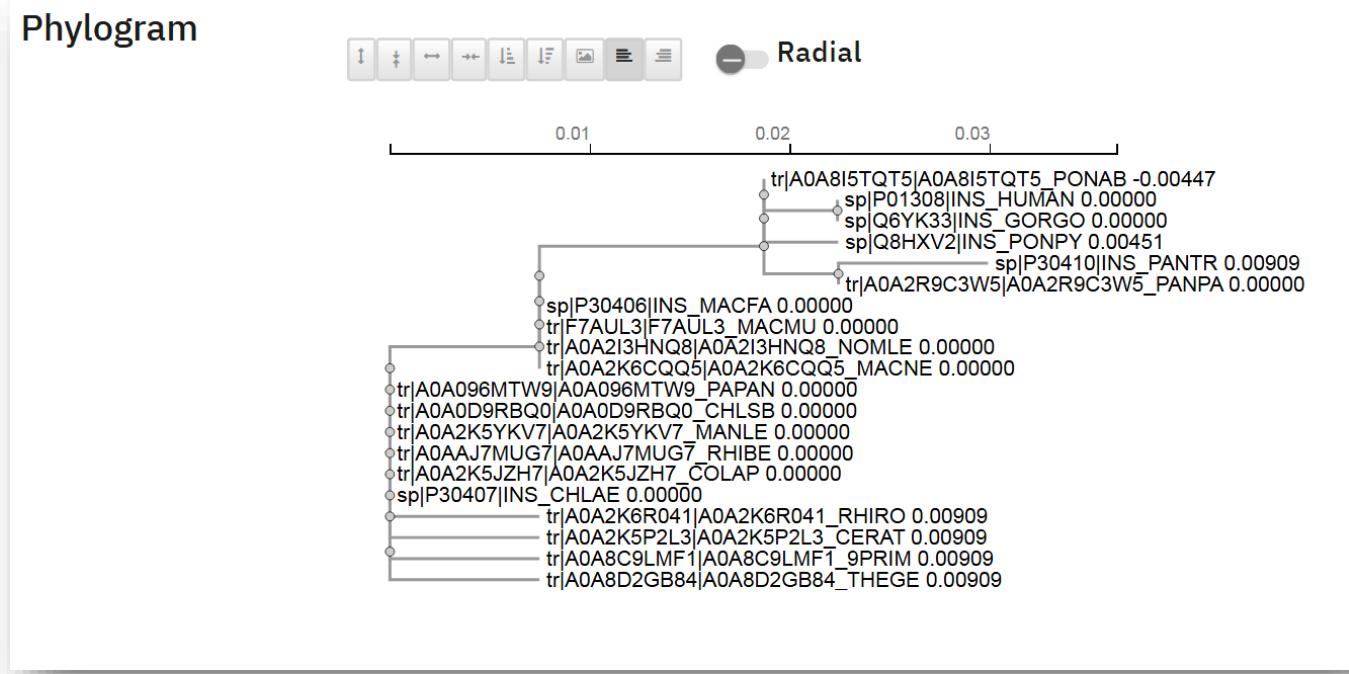
tr A0A8I5TQT5 A0A8I5TQT5_PONAB	MALWMRLLPLALLALWGPDPDA-AFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	59
sp P80410 INS_PANTR	MALWMRLLPLLVLLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
sp P01308 INS_HUMAN	MALWMRLLPLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
sp Q6YK39 INS_GORGO	MALWMRLLPLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A2R9CSW5 A0A2R9CSW5_PANPA	MALWMRLLPLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
sp Q8HXV2 INS_PONPY	MALWMRLLPLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A2K5P2L3 A0A2K5P2L3_CERAT	MALWMRLLPLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A8C9LMF1 A0A8C9LMF1_9PRIM	MALWMRLLPLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A8D2GB84 A0A8D2GB84_THEGE	MALWMRLLPLALLALWGPDSVPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A2K6R041 A0A2K6R041_RHIRO	MALWMRLLPLALLALCGDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
sp P80406 INS_MACFA	MALWMRLLPLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr F7AUL3 F7AUL3_MACMU	MALWMRLLPLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A2I3HNQ8 A0A2I3HNQ8_NOMLE	MALWMRLLPLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A2K6CQ05 A0A2K6CQ05_MACNE	MALWMRLLPLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A096MTW9 A0A096MTW9_PAPAN	MALWMRLLPLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A0D9RBQ0 A0A0D9RBQ0_CHLSB	MALWMRLLPLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A2K5YKV7 A0A2K5YKV7_MANLE	MALWMRLLPLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0AAJ7MUG7 A0AAJ7MUG7_RHIBE	MALWMRLLPLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
tr A0A2K5JZH7 A0A2K5JZH7_COLAP	MALWMRLLPLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60
sp P80407 INS_CHLAE	MALWMRLLPLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEED	60

tr A0A8I5TQT5 A0A8I5TQT5_PONAB	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 109	
sp P80410 INS_PANTR	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
sp P01308 INS_HUMAN	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
sp Q6YK39 INS_GORGO	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A2R9CSW5 A0A2R9CSW5_PANPA	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
sp Q8HXV2 INS_PONPY	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A2K5P2L3 A0A2K5P2L3_CERAT	LQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A8C9LMF1 A0A8C9LMF1_9PRIM	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A8D2GB84 A0A8D2GB84_THEGE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A2K6R041 A0A2K6R041_RHIRO	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
sp P80406 INS_MACFA	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr F7AUL3 F7AUL3_MACMU	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A2I3HNQ8 A0A2I3HNQ8_NOMLE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A2K6CQ05 A0A2K6CQ05_MACNE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A096MTW9 A0A096MTW9_PAPAN	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A0D9RBQ0 A0A0D9RBQ0_CHLSB	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A2K5YKV7 A0A2K5YKV7_MANLE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0AAJ7MUG7 A0AAJ7MUG7_RHIBE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
tr A0A2K5JZH7 A0A2K5JZH7_COLAP	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	
sp P80407 INS_CHLAE	PQVGQVELGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN 110	

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

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

18. Scroll down to have Phylogram.



Results of Multiple Sequences Alignment of 20 sequences -

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

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

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

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

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

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

>tr|A0A2K6R041|A0A2K6R041_RHIRO Insulin OS=Rhinopithecus roxellana OX=61622 GN=INS PE=3
 SV=1MALWMRLLPLLALLALCGPDPPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGPGAGSLQPLALEGSQKRGIVEQCCTSICSLYQLEN
 YCN

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

>tr|A0A8C9LMF1|A0A8C9LMF1_9PRIM Insulin OS=Piliocolobus tephrosceles OX=591936 GN=INS PE=3
 SV=1MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGPGAGSLQPLALEGSQKRGIVEQCCTSICSLYQLEN
 YCN

>tr|A0A2K5P2L3|A0A2K5P2L3_CERAT Insulin OS=Cercocebus atys OX=9531 GN=INS PE=3
 SV=1MALWMRLLPLLALLALWGPDPVPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDPQVGQVELGGPGAGSLQPLSLEGSQKRGIVEQCCTSICSLYQLEN
 YCN

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

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

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

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

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

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

>tr|A0A8I5TQT5|A0A8I5TQT5_PONAB Insulin OS=Pongo abelii OX=9601 GN=INS PE=3
 SV=1MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGPGAGSLQPLALEGSQKRGIVEQCCTSICSLYQLEN
 YCN

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

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

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

>tr|A0A2K6CQQ5|A0A2K6CQQ5_MACNE Insulin OS=Macaca nemestrina OX=9545 GN=INS PE=3
SV=1MALWMRLLPALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFYTPKTRREAEDPQVGQVELGGPGAGSLQPLALEGSQKRGIVEQCCTSICSLYQLEN
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.