DATE: 12-11-21

WEBLEM 13

Introduction to MSA

Multiple sequence alignment, which is to align multiple related sequences to achieve optimal matching of the sequences. Related sequences are identified through the database similarity searching. As the process generates multiple matching sequence pairs, it is often necessary to convert the numerous pairwise alignments into a single alignment, which arranges sequences in such a way that evolutionarily equivalent positions across all sequences are matched.

It is theoretically possible to use dynamic programming to align any number of sequences as for pairwise alignment. However, the amount of computing time and memory it requires increases exponentially as the number of sequences increases. As a consequence, full dynamic programming cannot be applied for datasets of more than ten sequences. In practice, heuristic approaches are most often used.

There is a unique advantage of multiple sequence alignment because it reveals more biological information than many pairwise alignments can. For example, it allows the identification of conserved sequence patterns and motifs in the whole sequence family, which are not obvious to detect by comparing only two sequences. Many conserved and functionally critical amino acid residues can be identified in a protein multiple alignment. Multiple sequence alignment is also an essential prerequisite to carrying out phylogenetic analysis of sequence families and prediction of protein secondary and tertiary structures. Multiple sequence alignment also has applications in designing degenerate polymerase chain reaction (PCR) primers based on multiple related sequences.

SCORING FUNCTION:

Multiple sequence alignment is to arrange sequences in such a way that a maximum number of residues from each sequence are matched up according to a particular scoring function. The scoring function for multiple sequence alignment is based on the concept of sum of pairs (SP). As the name suggests, it is the sum of the scores of all possible pairs of sequences in a multiple alignment based on a particular scoring matrix. In calculating the SP scores, each column is scored by summing the scores for all possible pairwise matches, mismatches and gap costs. The score of the entire alignment is the sum of all of the column scores. The purpose of most multiple sequence alignment algorithms is to achieve maximum SP scores.

Progressive Alignment Method:

Progressive alignment depends on the stepwise assembly of multiple alignment and is heuristic in nature. It speeds up the alignment of multiple sequences through a multistep process. It first conducts pairwise alignments for each possible pair of sequences using the Needleman–Wunsch global alignment method and records these similarity scores from the pairwise comparisons. The scores can either be percent identity or similarity scores based on a particular substitution matrix. Both scores correlate with the evolutionary distances between sequences. The scores are then converted into evolutionary distances to generate a distance matrix for all the sequences involved. A simple phylogenetic analysis is then performed based on the distance matrix to group sequences based on pairwise distance scores. As a result,

a phylogenetic tree is generated using the neighbor-joining method. The tree reflects evolutionary proximity among all the sequences.

In the next step, the next closest sequence based on the guide tree is aligned with the consensus sequence using dynamic programming. More distant sequences or sequence profiles are subsequently added one at a time in accordance with their relative positions on the guide tree. After realignment with a new sequence using dynamic programming, a new consensus is derived, which is then used for the next round of alignment. The process is repeated until all the sequences are aligned.

Two of the popular progressive alignment methods used at present are:

1. Clustal Omega
2. T-Coffee

Clustal Omega:

It is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more homologous nucleotide or amino acid sequences. This has been one of the most widely used procedures in bioinformatics for decades, as it is an essential prerequisite for most phylogenetic or comparative analyses of homologous genes or proteins.

One or two iterations are usually found to have a small beneficial effect on alignment quality. One of the most important features of this program is the flexibility of using substitution matrices. Clustal does not rely on a single substitution matrix. Instead, it applies different scoring matrices when aligning sequences, depending on degrees of similarity.

Another feature of Clustal is the use of adjustable gap penalties that allow more insertions and deletions in regions that are outside the conserved domains, but fewer in conserved regions. For example, a gap near a series of hydrophobic residues carries more penalties than the one next to a series of hydrophilic or glycine residues, which are common in loop regions. In addition, gaps that are too close to one another can be penalized more than gaps occurring in isolated loci.

The program also applies a weighting scheme to increase the reliability of aligning divergent sequences (sequences with less than 25% identity). This is done by down weighting redundant and closely related groups of sequences in the alignment by a certain factor. This scheme is useful in preventing similar sequences from dominating the alignment. The weight factor for each sequence is determined by its branch length on the guide tree. The branch lengths are normalized by how many times sequences share a basal branch from the root of the tree.

One feature of Clustal Omega that is especially useful is the ability to use an external HMM to help guide an alignment. This is referred to as External Profile Alignment (EPA) in this unit. It consists of taking each sequence to be aligned and aligning it against the HMM to help its alignment with other sequences in the dataset. It can be used with publicly available HMMs from PFAM, for example, or from an expert alignment that the user has built up. This procedure can also be used to iteratively generate an HMM from an output MSA and to input this to realign the input sequences.

What do the consensus symbols mean in the alignment?

A "-" character is used to pad sequences with gaps, which are necessary to align the sequences.

An * (asterisk) indicates positions which have a single, fully conserved residue.
A : (colon) indicates conservation between groups of strongly similar properties as below - roughly equivalent to scoring > 0.5 in the Gonnet PAM 250 matrix.

A . (period) indicates conservation between groups of weakly similar properties as below - roughly equivalent to scoring =< 0.5 and > 0 in the Gonnet PAM 250 matrix.

What do the colours mean when I show them on protein alignments?

This protein-only option colours the residues according to their physicochemical properties:

Residue	Colour	Property
AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RK	MAGENTA	Basic - H
STYHCNGQ	GREEN	Hydroxyl + sulfhydryl + amine + G
Others	Grey	Unusual amino/imino acids etc

References:

1. Xiong, J. (2006). Multiple Sequence Alignment. *Essential Bioinformatics* (1st ed.). Cambridge University Press, 63-71.
2. Sievers, F., & Higgins, D. G. (2014). Clustal Omega. Current Protocols in Bioinformatics, 3.13.1–3.13.16. doi:10.1002/0471250953.bi0313s48
3. Help - Clustal Omega FAQ - Tools Help & Documentation - EMBL-EBI. (n.d.). EMBL-EBI. Retrieved October 30, 2021, from https://www.ebi.ac.uk/seqdb/confluence/display/THD/Help+-+Clustal+Omega+FAQ
4. Clustal Omega < Multiple Sequence Alignment < EMBL-EBI. (n.d.). EMBL-EBI. Retrieved October 30, 2021, from https://www.ebi.ac.uk/Tools/msa/clustalo/

DATE: 11-11-21

WEBLEM 13/a

Clustal Omega

(URL: https://www.ebi.ac.uk/Tools/msa/clustalo/**)**

Aim:

To study rhodopsin sequnces similarity using Clustal Omega

Introduction:

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 homologous nucleotide or amino acid sequences. This has been one of the most widely used procedures in bioinformatics for decades, as it is an essential prerequisite for most phylogenetic or comparative analyses of homologous genes or proteins. One or two iterations are usually found to have a small beneficial effect on alignment quality. One of the most important features of this program is the flexibility of using substitution matrices. Clustal does not rely on a single substitution matrix. Instead, it applies different scoring matrices when aligning sequences, depending on degrees of similarity.The choice of a matrix depends on the evolutionary distances measured from the guide tree. For example, for closely related sequences that are aligned in the initial steps, Clustal automatically uses the BLOSUM62 or PAM120 matrix. When more divergent sequences are aligned in later steps of the progressive alignment, the BLOSUM45 or PAM250 matrices may be used instead.

Rhodopsin, also called visual purple, pigment-containing sensory protein that converts light into an electrical signal. Rhodopsin is found in a wide range of organisms, from vertebrates to bacteria. In many seeing animals, including humans, it is required for vision in dim light and is located in the retina of the eye specifically, within the tightly packed disks that make up the outer segment of the retina's photoreceptive rod cells, which are specially adapted for vision under low-light conditions.
Rhodopsin was discovered in 1876 by German physiologist Franz Christian Boll, who observed that the normally reddish purple frog retina turned pale in bright light. The fading of colour was later attributed to the destruction of rhodopsin, via a process known as bleaching. Bleaching and the subsequent regeneration of rhodopsin are major steps in the visual cycle—the series of biochemical reactions that is critical for vision in low light.

Methodology:

1. Open homepage for Clustal Omega (https://www.ebi.ac.uk/Tools/msa/culstalo/)
2. Select multiple sequences for query "Rhodopsin" from UniProt database and retrieve FASTA sequences for the same.
3. Ener FASTA sequences in Clustal Omega search page.
4. Submit query.
5. Observe and interpret the results.

Observation:

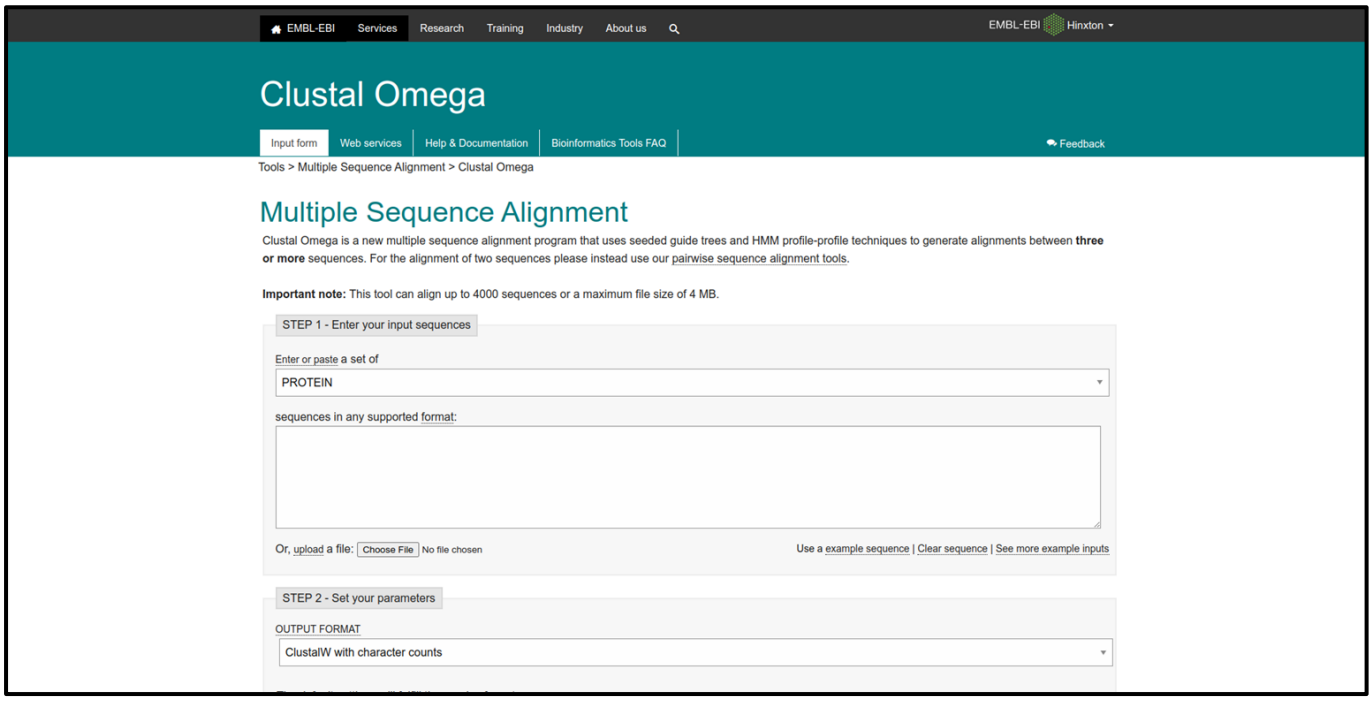

Fig1. Homepage of Clustal Omega

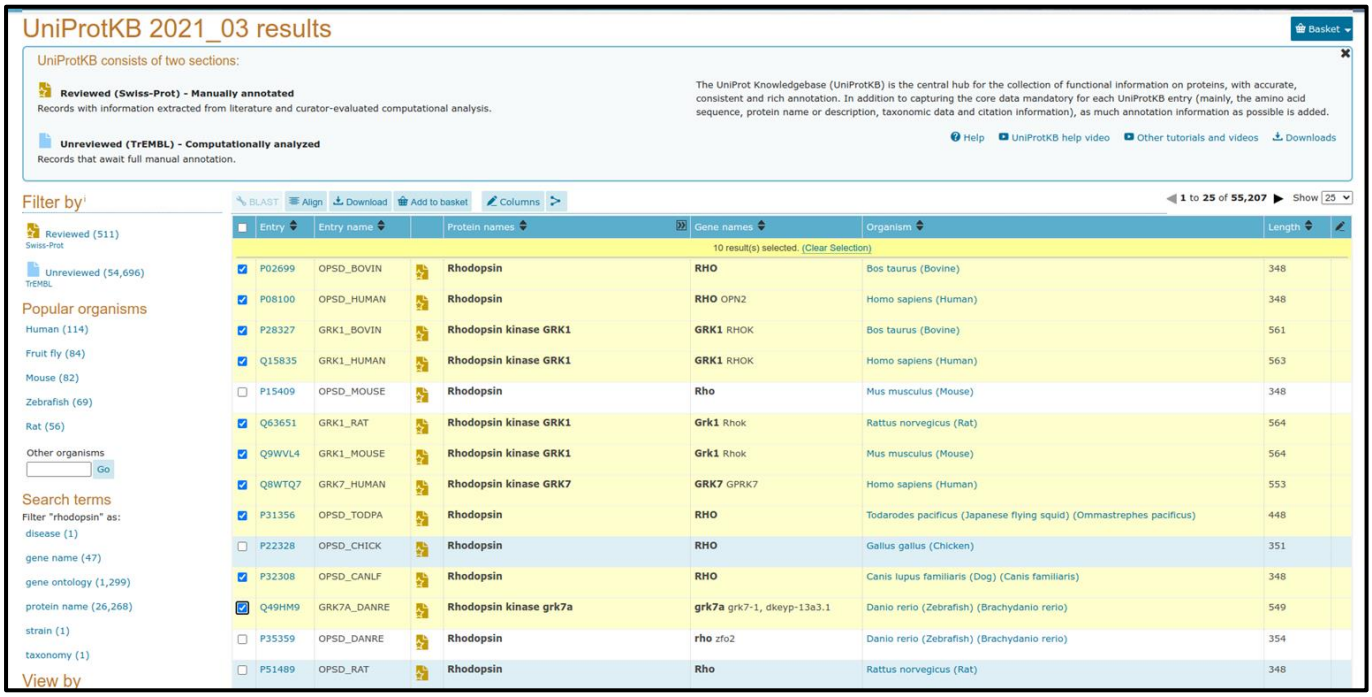

Fig2. Selected sequences for query "Rhodopsin" in UniProt database.

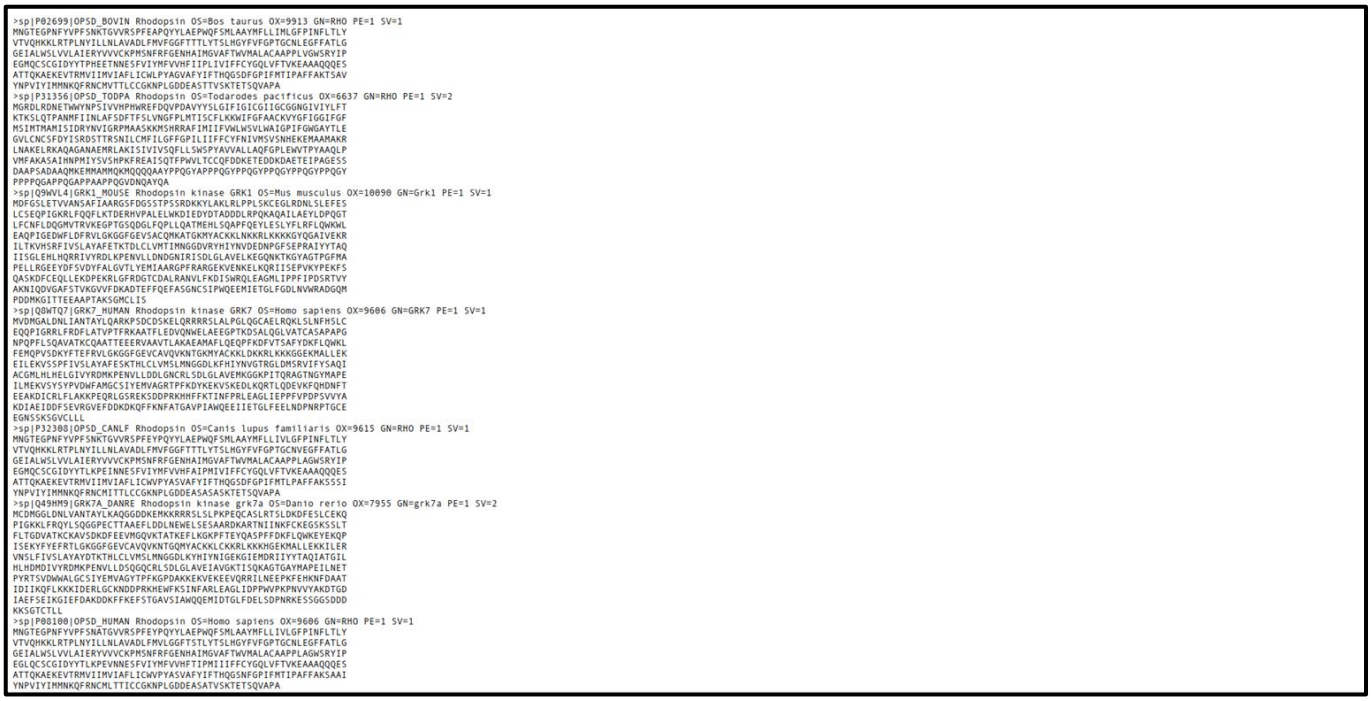

Fig3. FASTA sequences for the selected sequences for query Rhodopsin

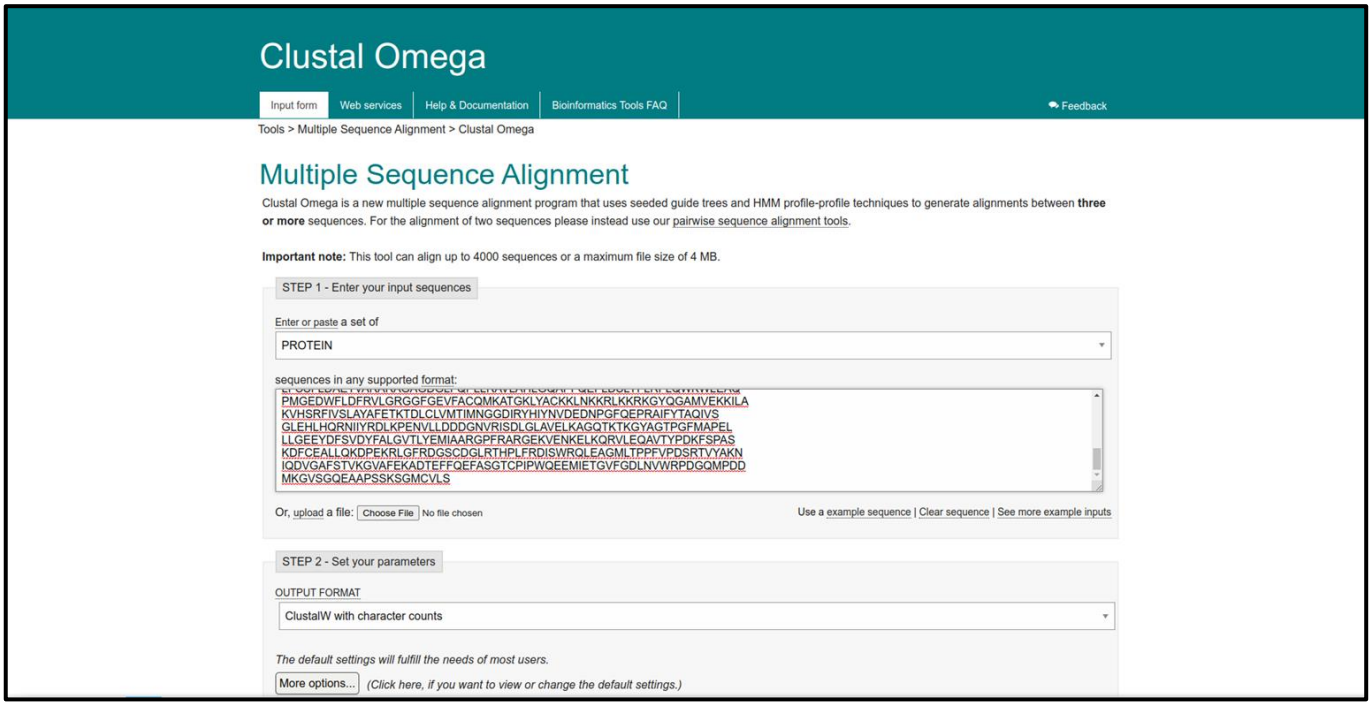

Fig4. Search page for query Rhodopsin in Clustal Omega

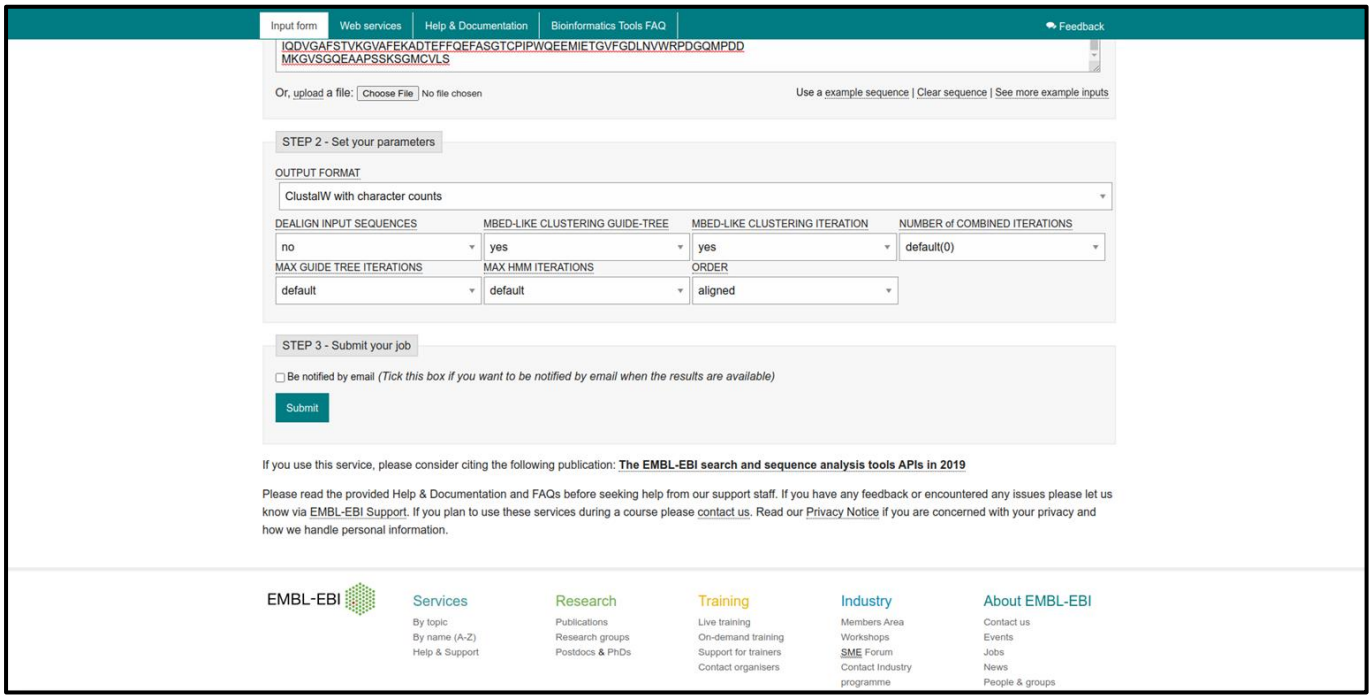

Fig5. Parameters used for query Rhodopsin

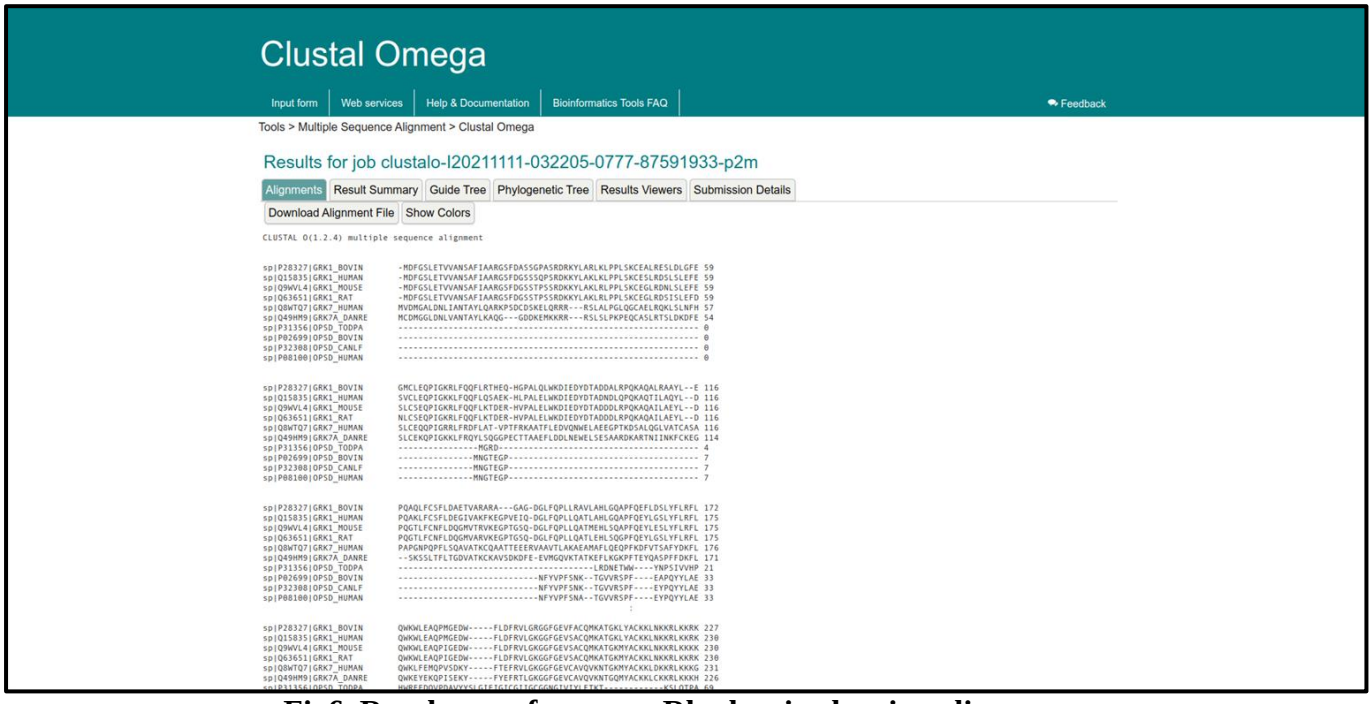

Fig6. Result page for query Rhodopsin showing alignments

Fig7. Result page for query rhodopsin showing alignments with colors

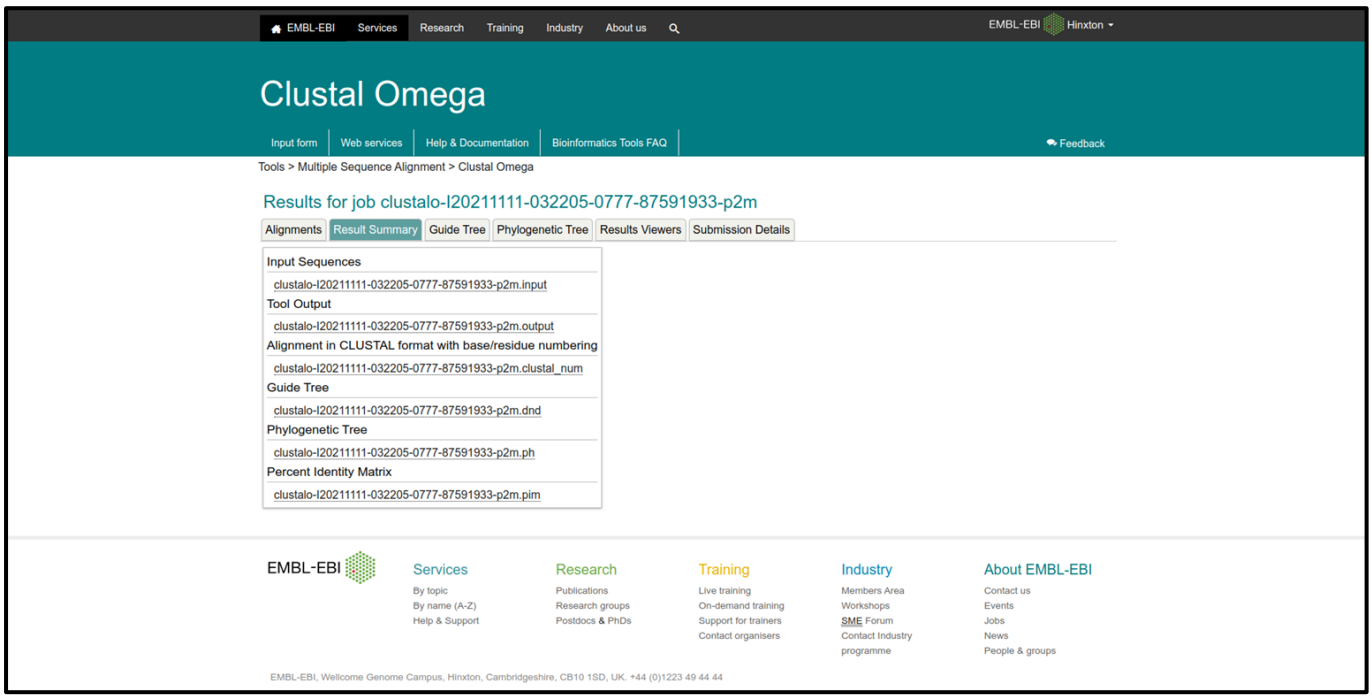

Fig8. Result page for query Rhodopsin showing result summary

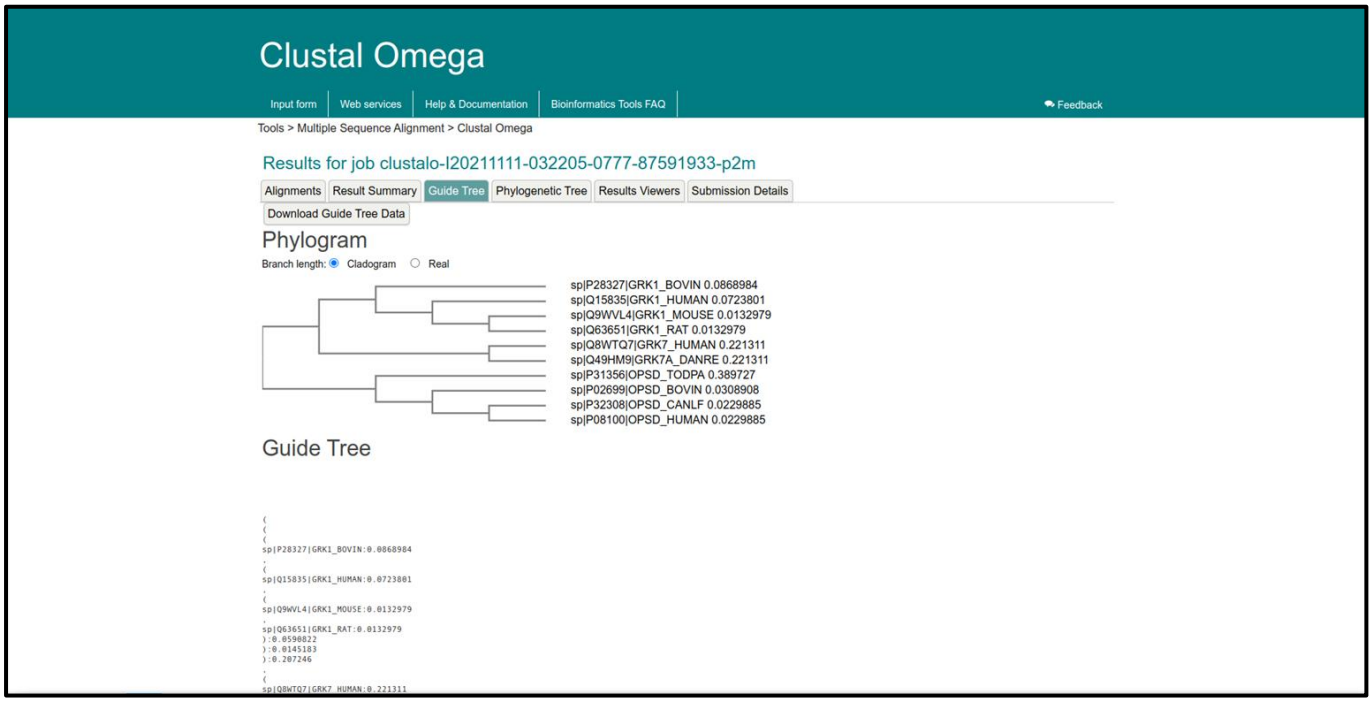

Fig9. Result page for query Rhodopsin showing Phylogram and guide tree information

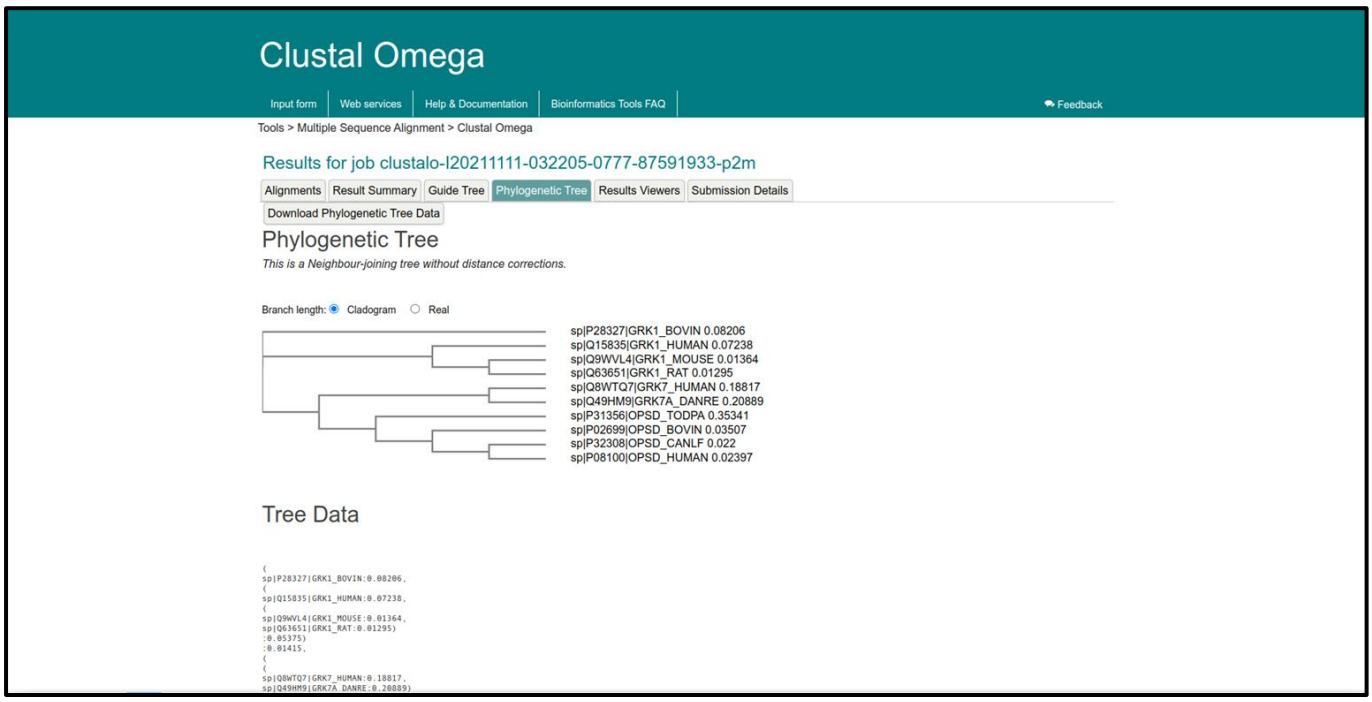

Fig10. Resukt page for query Rhodopsin showing Phylogenetic tree and tree data

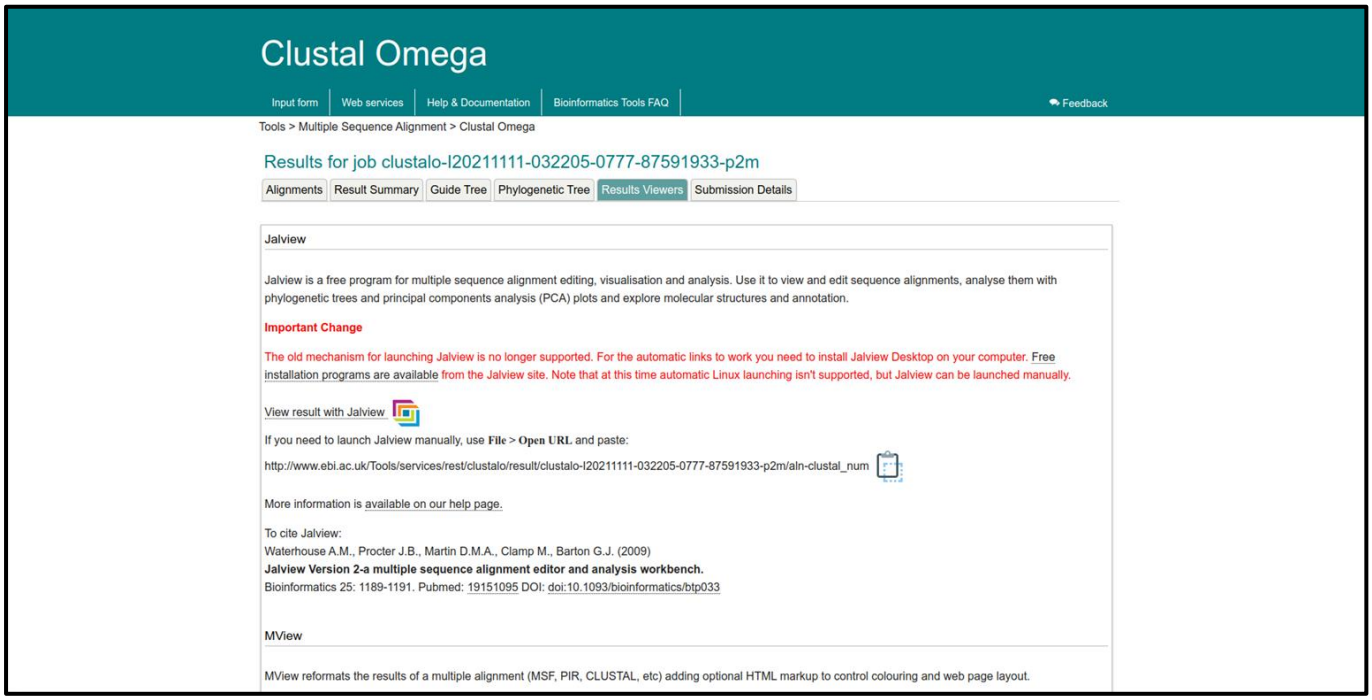

Fig11. Result page for query Rhodopsin showing result viewers section to open the results in external software

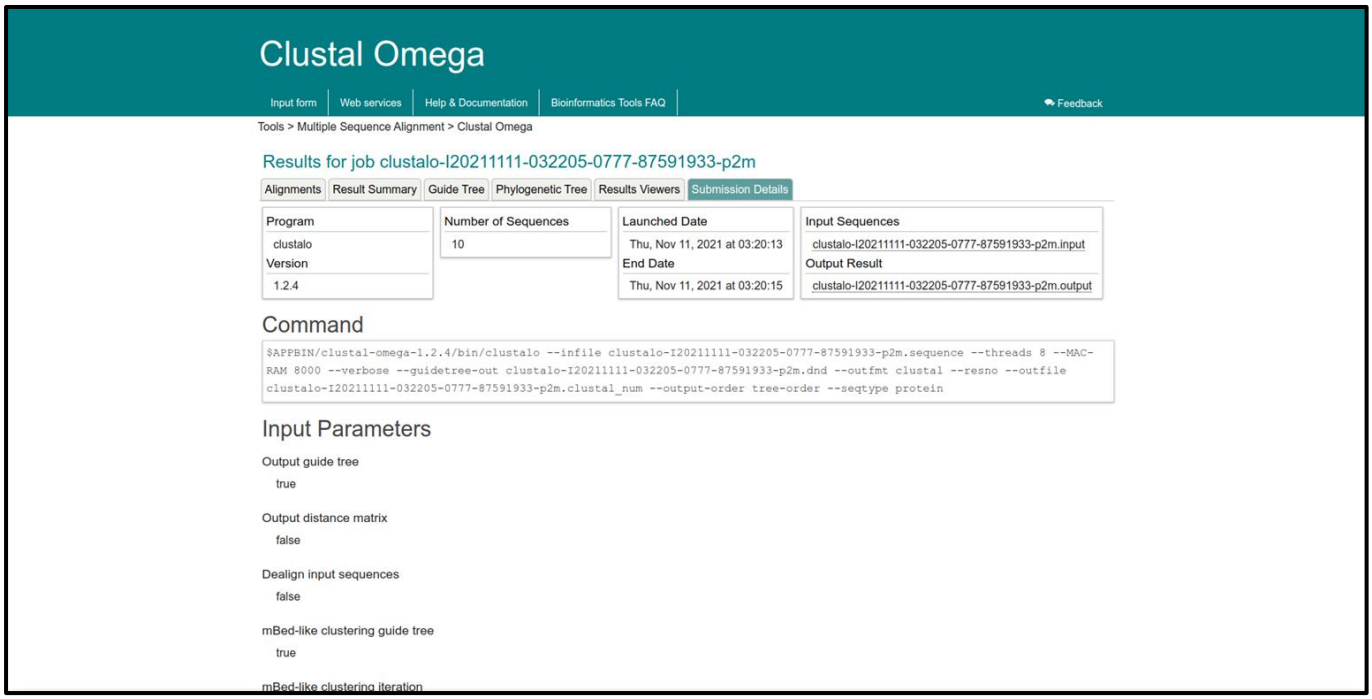

Fig12. Result page for query Rhodopsin showing submission details giving info about our search parameters

Result:

Protein sequences for query “Rhodopsin” from various species was used to perform multiple sequence alignment and alignments, result summary, guide tree, phylogenetic tree, result viewers and submission details information was observed for the same.

Conclusion:

Clustal omega which is a heuristic algorithm working on progressive alignment method allows users to achieve quick multiple sequence alignment results. It is useful identification of conserved sequence patterns and motifs, conserved and functionally critical amino acid residues can be identified in a protein, phylogenetic analysis of sequence families and prediction of protein secondary and tertiary structures.

References:

1. Xiong, J. (2006). Multiple Sequence Alignment. *Essential Bioinformatics* (1st ed.). Cambridge University Press, 63-71.
2. Moman, R. N., & Varacallo, M. (2018). Physiology, Albumin. Nih.gov; StatPearls Publishing. Retrieved November 11th, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK459198/
3. (n.d.). Retrieved November 11th, 2021, from https://www.uniprot.org/uniprot/?query=id:P02699 OR id:P08100 OR id:P28327 OR id:Q15835 OR id:Q63651 OR id:Q9WVL4 OR id:Q8WTQ7 OR id:P31356 OR id:P32308 OR id:Q49HM9&format=fasta&sort=score
4. (n.d.). Retrieved November 11th, 2021, from https://www.uniprot.org/uniprot/?query=Rhodopsin&sort=score
5. (n.d.). Retreived November 11th, 2021, from https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20211111-032205-0777-87591933-p2m&analysis=alignments
6. Haplochromis burtoni rhodopsin-like (rhodopsin), mRNA - Nucleotide - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/nuccore/NM_001287832.1?report=fasta

DATE: 11-11-21

WEBLEM 13/b

Tree view

Aim:

To view query Rhodopsin phylogenetic tree data produced in Clustal Omega in tree view

Introduction:

TreeView provides a simple way to view the phylogenetic trees produced by a range of programs, such as PAUP* and PHYLIP, TREE-PUZZLE, and ClustalX. While some phylogenetic programs (such as the Macintosh version of PAUP*) have excellent tree-printing facilities, many programs do not have the ability to generate publication-quality trees. TreeView addresses this need. The program can read and write a range of tree file formats, display trees in a variety of styles, print trees, and save the tree as a graphic file.

Rhodopsin, also called visual purple, pigment-containing sensory protein that converts light into an electrical signal. Rhodopsin is found in a wide range of organisms, from vertebrates to bacteria. In many seeing animals, including humans, it is required for vision in dim light and is located in the retina of the eye specifically, within the tightly packed disks that make up the outer segment of the retina's photoreceptive rod cells, which are specially adapted for vision under low-light conditions. Rhodopsin was discovered in 1876 by German physiologist Franz Christian Boll, who observed that the normally reddish purple frog retina turned pale in bright light. The fading of colour was later attributed to the destruction of rhodopsin, via a process known as bleaching. Bleaching and the subsequent regeneration of rhodopsin are major steps in the visual cycle—the series of biochemical reactions that is critical for vision in low light.

Methodology:

1. Retrieve phylogenetic tree data for query Rhodopsin from MSA performed on Vlustal Omega.
2. Save data on notepad with .php as file extension.
3. Open file in tree view and make required changes.

Observation:

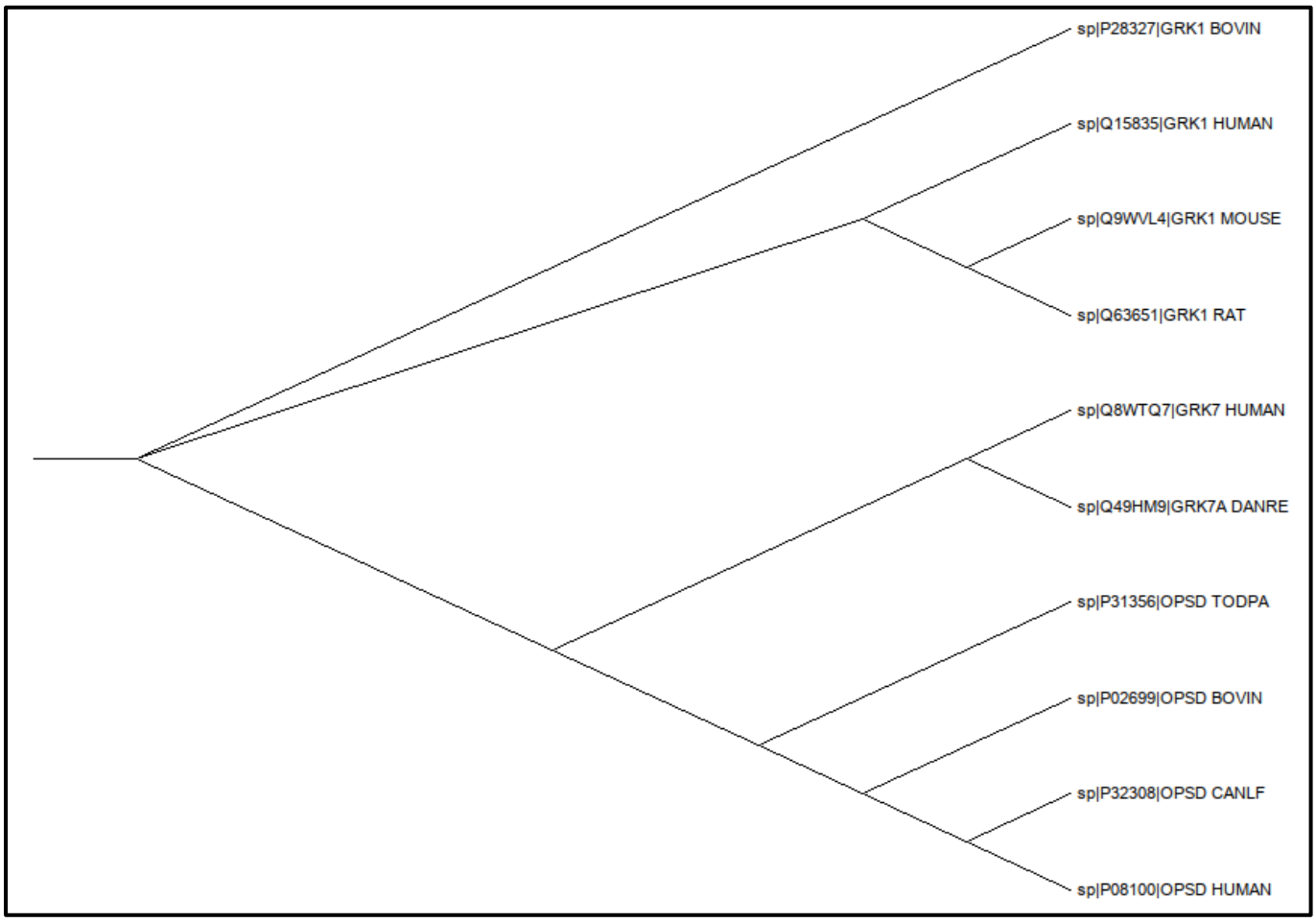

Fig1. Phylogenetic tree data for query Rhodopsin in tree view

Result:

Phylogenetic tree data produced in Clustal Omega for query Rhodopsin was viewed on tree view

Conclusion:

Tree view provides users various display options, such as viewing internal node labels. Choosing the style in which the tree is drawn and desginating an outgroup. It also allows the user to define default display preferences.

References:

1. (n.d.). Retrieved on November 11, 2021, from https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalo-I20211111-032205-0777-87591933-p2m&analysis=phylotree
2. Haplochromis burtoni rhodopsin-like (rhodopsin), mRNA - Nucleotide - NCBI. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/nuccore/NM_001287832.1?report=fasta