**BIN 506: Protein & DNA Sequence Analysis**

Assignment 03

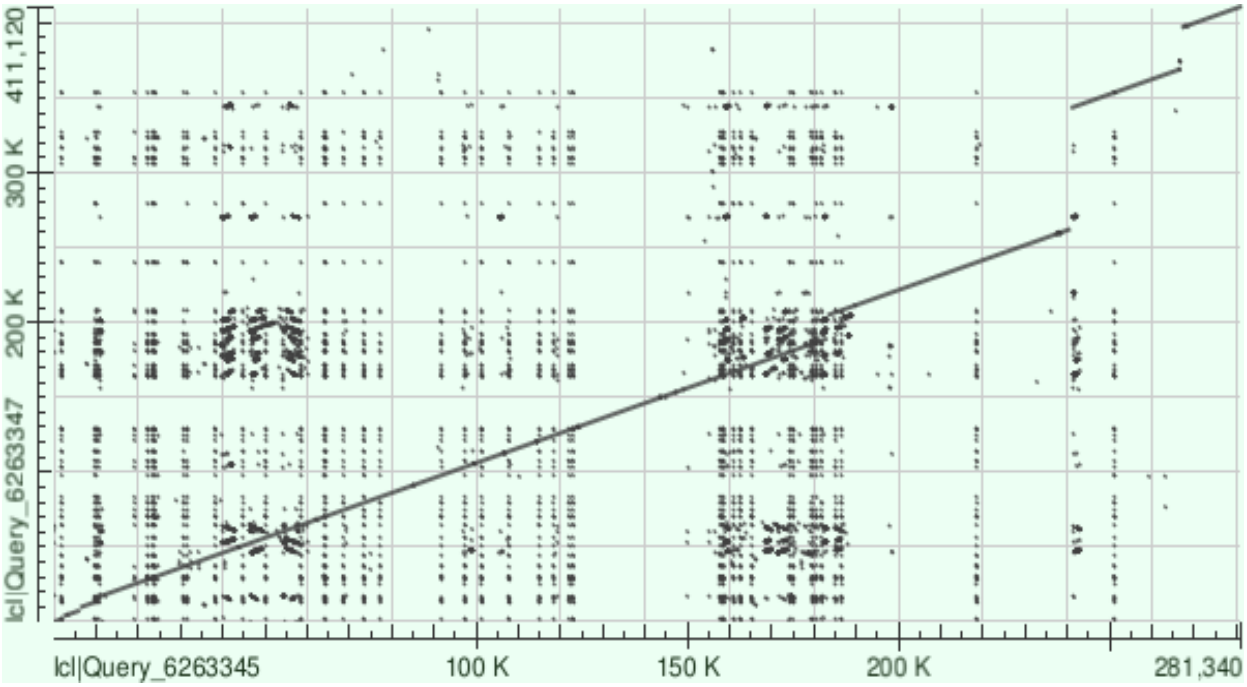
By: Taha Ahmad

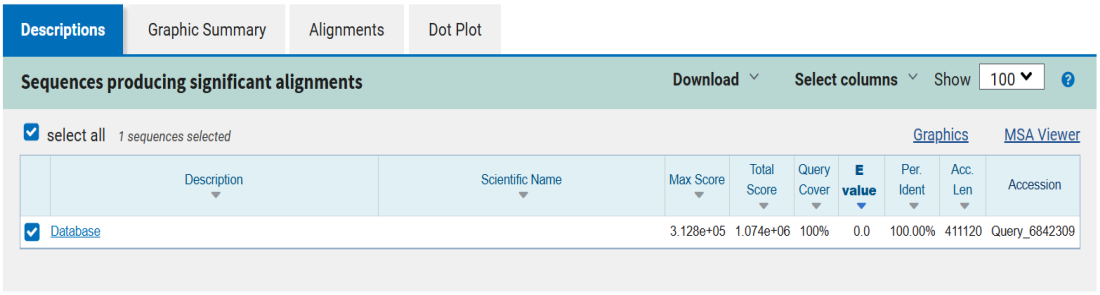
Student ID: 2546125

Instructor: Dr Yesim Aydin Son

**1) You will see 5 sequences (sequence_#.fasta) attached to this assignment. Use blastn to align them to 'sequence_DB.fasta' and answer the following questions. Add screenshots to your assignment.**

**a)** Align sequence_1 to sequence_DB, comment on the scores and the dot-plot

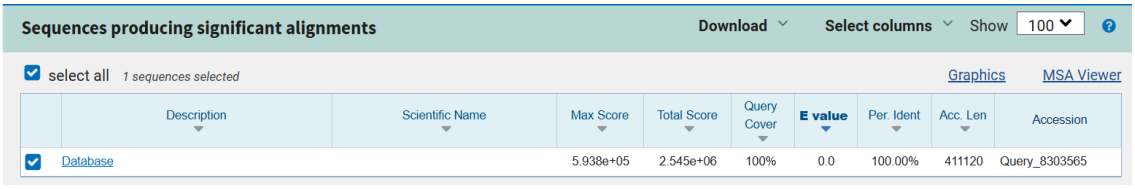


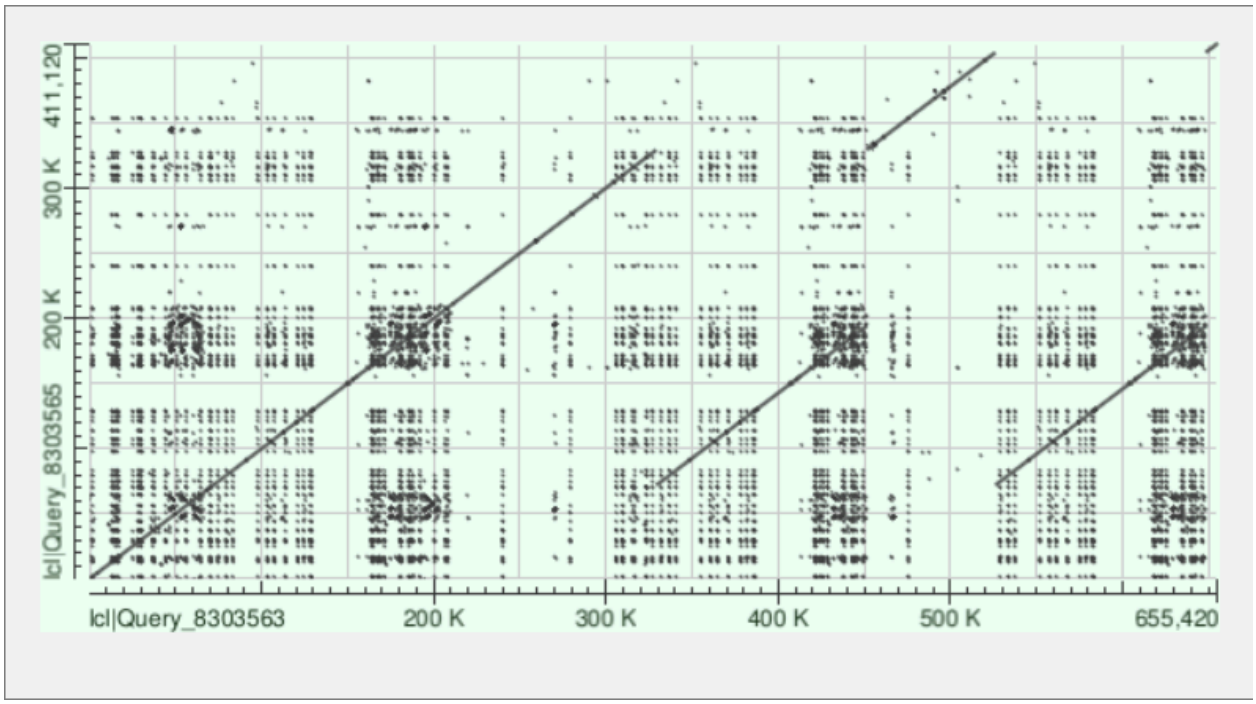



**Answer:** The BLAST alignment between our query sequence and the database shows a high quality sequence matching with a total score of 1.074e+06, the stronger the total score the more significant the matches are. The max value indicates the highest score for the best individual alignment segment thus representing the most significant local alignment , our max score is 3.128e+05. The E Value is the expected number of chance matches in the database in terms of probability. The E value of 0.0 which we got means that the matches are extremely unlikely to occur by chance. A query cover of 100% means an exact match percentage between query and database sequence, a perfect match like ours indicates identical sequences.

The dotplot spans approximately 281,340 base pairs. We can observe a long diagonal line ranging from the origin to approximately 175k representing a near perfect sequence alignment; the jump between the next diagonal indicates an indel (insertion or deletion). Another diagonal dominates the graph till 250k representing perfect seq alignment and then another indel jump occurring 2 times at the end. The dot density remains strongly uniform throughout the plot, with dots tightly clustered along the primary alignment axis. These densely packed dots symbolize base-by-base sequence correspondence, creating a visual representation of genetic similarity which complements the statistical BLAST metrics

**b)** Align sequence_2 to sequence_DB, comment on the scores and the dot-plot.






**Answer :** The BLAST results reveal a remarkably high-quality sequence match. The total score of 2.545e+06 indicates an extremely significant alignment, with the max score of 5.938e+05 highlighting the most robust local alignment segment. The E-value of 0.0 underscores the statistical improbability of these matches occurring by random chance, while the 100% query coverage and 100% percent identity confirm an essentially perfect sequence correspondence.

The dot plot spans approximately 655,420 base pairs. The plot is characterized by multiple diagonal segments that demonstrate near-perfect sequence alignment, integrated within the graph are strategic indel (insertion/deletion) events. Unlike the previous sequence, this dot plot exhibits more complex structural variations, with several distinct diagonal segments interrupted by precise genomic variations.The primary diagonal segments extend across significant genomic regions - notable stretches around 200k, 300k, 400k, and 500k base pairs. Each of these segments represents exceptional sequence homology, with dense, uniform dot clustering indicating base-by-base genetic correspondence. The vertical and horizontal deviations between these diagonal segments represent localized genetic variations, potentially reflecting subtle evolutionary mutations or structural rearrangements. the indel appears more pronounced compared to the previous

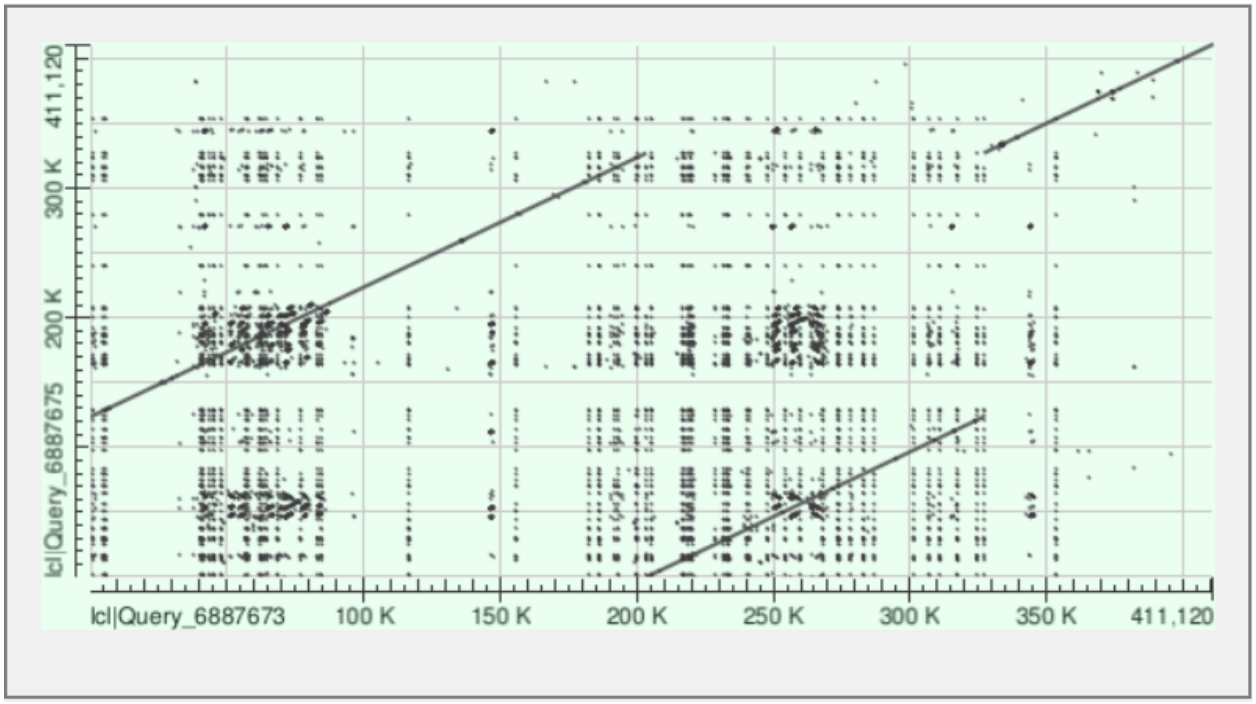
sequence. These discontinuities in the diagonal alignment suggest complex genomic events potentially representing insertional or deletional mutations that have occurred during the sequence's evolutionary history.

**c)** Align sequence_3 to sequence_DB, comment on the scores and the dot-plot.

**Sequences producing significant alignments** Download Select columns Show 100

select all *0 sequences selected* Graphics MSA Viewer

| Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|
| Database | | 3.670e+05 | 1.470e+06 | 100% | 0.0 | 100.00% | 411120 | Query_6887675 |




**Answer:** The BLAST alignment for this sequence reveals a total score of 1.470e+06 signifying an extraordinarily significant sequence match, with a max score of 3.670e+05 indicating the best local alignment segment. The E-value of 0.0 emphatically demonstrates the statistical impossibility of these matches occurring by chance, while the 100% query coverage and percent identity validate a virtually perfect sequence correspondence.

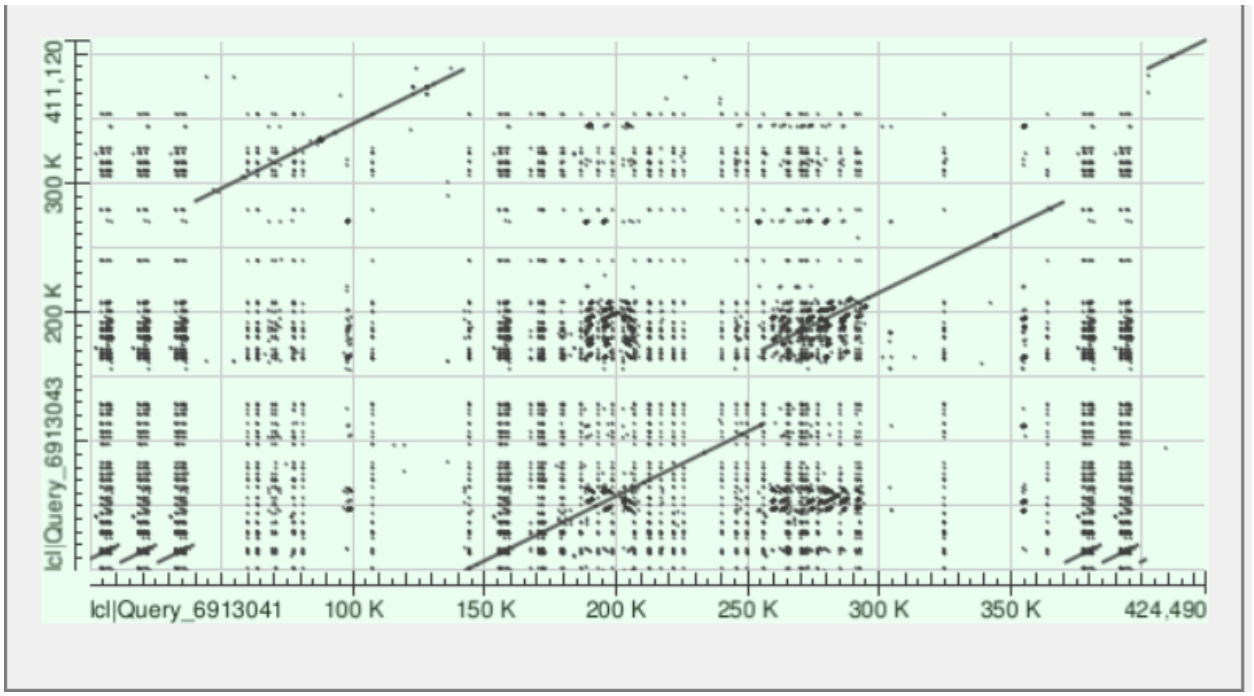
The dot plot spans approximately 411,120 base pairs. Unlike previous analyses, this visualization demonstrates a more intricate alignment pattern characterized by

multiple discrete diagonal segments. These segments are not uniformly continuous but strategically interrupted by precise genomic variations, suggesting a complex evolutionary history. Notably, the plot features several distinct diagonal regions - prominent alignments around 100k, 200k, and 300k base pairs. Each diagonal segment represents an area of homology, with densely clustered dots symbolizing base-level genetic concordance. The vertical and horizontal deviations between these segments represent localized genetic variations, potentially indicative of subtle mutational events or structural rearrangements.The indel events in this sequence appear more strategically positioned compared to previous observations. These diagonal interruptions suggest nuanced genomic modifications - potentially representing insertional or deletional mutations that have subtly reshaped the sequence's molecular architecture.

**d)** Align sequence_4 to sequence_DB, comment on the scores and the dot-plot.

select all *1 sequences selected* Graphics MSA Viewer

| Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|
| Database | | 2.079e+05 | 1.657e+06 | 100% | 0.0 | 100.00% | 411120 | Query_6913043 |



**Answer:** The BLAST alignment for this sequence gives a total score of 1.657e+06, signifying an extraordinarily significant sequence match. The max score of 2.079e+05 pinpoints the best local alignment segment, while the E-value of 0.0 emphatically demonstrates the statistical impossibility of these matches occurring by chance. The 100% query coverage and percent identity validate a virtually perfect sequence correspondence.

The dot plot spans approximately 424,490 base pairs. The plot's initial segment features three microscopic diagonal alignments, each representing localized regions of sequence homology. These initial short diagonals are followed by a dramatic vertical jump to a longer diagonal spanning from approximately 40k to 150k base pairs. This significant displacement suggests a substantial genomic rearrangement or structural variation. A second prominent diagonal emerges around 250k, again preceded by a notable vertical jump. This segment demonstrates another zone of exceptional sequence similarity. Towards the plot's terminus, two additional minute diagonal segments appear, mirroring the plot's initial configuration. The final diagonal represents a remarkable vertical leap at around 420k.These complex diagonal interruptions and strategic jumps suggest intricate molecular events - potentially representing large-scale insertional or deletional mutations, chromosomal rearrangements, or sophisticated evolutionary genomic modifications.

**2) Use the msa.txt file attached to the assignment to answer the following questions.**

**a) Can you figure out the origins of these sequences from the identification line?**

The accession prefixes (emb, ref, sp, gb) denote the source databases (EMBL, RefSeq, UniProt, GenBank) using that information and the accession (or ID) numbers on each sequence header we find out that:

1. Curvularia (P49053): A fungal genus (likely Ascomycota)
2. Embellisia (emb|CAA72344.1): Another fungal genus.

3. Drechslera (emb|CAA72008.1): A fungal genus (plant pathogen).
4. Nostoc (ref|NP_484716.1): A cyanobacterium (photosynthetic bacteria).
5. Deinococcus (ref|NP_294738.1): A radiation-resistant bacterium.
6. Ascophyllum (sp|P81701): A brown algae (seaweed).
7. Corallina (gb|AAM46061.1): A red algae.
8. Fucus (gb|AAC35279.1): Another brown algae (e.g., bladderwrack).

**b) Use T-Coffee and Clustal-O to perform MSA for the given sequences. Compare the results.**

1. Main Differences in the Alignment Results

Aligned Regions and Gap Lengths:

T-Coffee introduces more gaps in some sequences in order to maintain alignment consistency across homologous regions. For example in the T-Coffee output Curvularia and Embellisia show extended gaps in several places, such as around the conserved sequence regions (e.g between positions 50–150).

```
Ascophyllum     SDDADDPTPPNERDDEAFASRV--AAAKRELEGTGTVCQI-
NNGE-----
Corallina       ----DN---LQSRAKASFDTRV--SAAELAL-ARGVVPSL-
ANGEELLYR
Curvularia      ----------NISDNAYAQLGLVLDRSVLEA-PGGVDRES--
AS------
Deinococcus     --------------------------------------------------
-
Drechslera      --------------------------------------------------
-
Embellisia      ----------PISHNAYAQLQHVLDISVTKA-PAACDPAS--
SS------
Fucus           SDDALDPTAPNRRDNVAFASRR--DAARRERDGTGTVCQI-
TNGE-----
Nostoc          ----------KAKNNSFFETER--
DKAIEELVSSGVSVQSIGDG------
```

Clustal-O generally produces a more compact alignment with fewer gaps in some regions. For instance, in the Clustal-O output, e.g. in the output Fucus and Ascophyllum maintain a more continuous sequence with minimal interruptions.

The introduction of gaps in T-Coffee is more frequent and strategic, whereas Clustal-O often aligns without introducing excessive gaps.

```
Ascophyllum      NGETDLAAKFH-KSLPHDDLGQV-
DADAFAALEDCILNGDLSICEDVPVGNSEGD-----  97
Fucus            NGETDLATMFH-KSLPHDELGQV-
TADDFAILEDCILNGDFSICEDVPA----GD-----  216


Embellisia       L-
SPRALGMLQLAVHDAYFAIHPSAGFTTFLTPGAEDGAYRLPDPSYAKDARQAVAGAAI  108
Curvularia       L-
SARALGMLHLAIHDAYFSICPPTDFTTFLSPDTENAAYRLPSPNGANDARQAVAGAAL  107
Drechslera       ----------------------------------------------
------------    OO
Nostoc           TWISRTGAILHSAIYDA---VNS-------
IEKKYNPYLEIIPANPGASP-------EAA   66
```

Length of Aligned Segments:

Clustal-O maintains longer continuous segments of alignment. For example, in the region around residues 100–200, sequences such as Corallina and Deinococcus appear more contiguous.
T-Coffee, on the other hand, introduces interruptions to maximize local similarities, which leads to some sequences appearing more fragmented.

2. Reasons for These Differences:

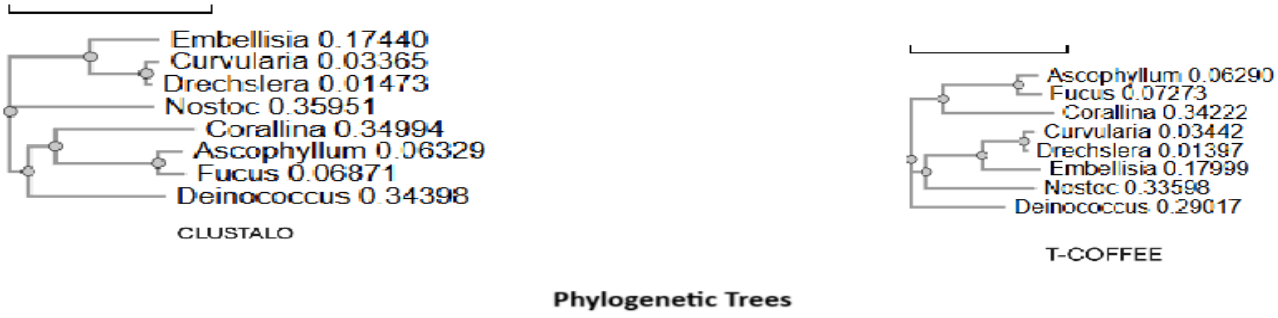
T-Coffee utilizes a consistency-based scoring system, which integrates multiple alignments (including CLUSTAL and MUSCLE) to enhance accuracy. This results in alignments that prioritize sequence homology and conservation, sometimes at the cost of introducing more gaps.

Clustal-O employs progressive alignment, optimizing for speed and efficiency. This method can sometimes produce slightly different cluster arrangements compared to T-Coffee, particularly in regions with insertions or deletions.

While both tools aim to align homologous sequences accurately, T-Coffee places a stronger emphasis on maintaining evolutionary conservation at a broader level, whereas Clustal-O focuses on a more direct sequence comparison.

3. Comparison of Phylogenetic Trees

In both phylogenetic trees, the clusters remain remarkably consistent. The organisms are grouped in the same pattern, with Embellisia, Curvularia, and Drechslera forming one distinct cluster, while Nostoc, Corallina, and Ascophyllum form another closely related cluster. Fucus and Deinococcus maintain their respective positions in both trees. Despite being generated by different tools (T-Coffee and Clustal-O), the clustering patterns are nearly identical, suggesting a strong and reliable representation of the evolutionary relationships between these organisms.




**Phylogenetic Trees**

**3) Answer the following questions about Clustal-O.**

**a) What type of MSA method does Clustal-O use?**

Clustal-O uses a progressive alignment method which builds the multiple sequence alignment (MSA) in steps by first aligning the most similar sequences and then progressively adding less similar sequences based on a guide tree. The process follows these steps:

1. Compute a distance matrix between sequences.

2. Construct a guide tree using neighbor-joining.

3. Align sequences progressively following the tree structure.

Unlike T-Coffee which integrates multiple alignment strategies for consistency, Clustal Omega focuses on efficiency and scalability therefore making it more suitable for large datasets.

**b) What is the substitution matrix used in Clustal-O?**

Clustal-O primarily uses the Gonnet substitution matrix by default which is derived from actual sequence alignments and is similar to PAM and BLOSUM but it is dynamically adjusted based on evolutionary distances.

**c) What are the default gap opening and extension penalties for Clustal-O?**

Gap opening penalty is 6.0 and Gap extension penalty is 1.0

**d) What do the (*), (:) and (.) symbols at the alignment positions indicate? What does it mean if there are no symbols?**

In Clustal-O:

- ( * ) represents a fully conserved position (same amino acid or nucleotide in all sequences at that position).

- ( : ) represents strong conservation (amino acids with similar properties, such as hydrophobicity, are aligned).

- ( . ) represents weak conservation (amino acids with somewhat similar properties appear in that column).

- No symbols mean No conservation (alignment at that position is highly variable, meaning the sequences do not share similarity at that site).

Bonus Question : https://chatgpt.com/share/67ef93e8-5eb0-800f-ad0a-48c469490da2