Experiment 6: Multiple Sequence Alignment and Phylogenetic Tree Construction

AIM

1. To perform multiple sequence alignment (MSA) of protein/DNA sequences using different alignment algorithms (Clustal Omega, MUSCLE, MAFFT, and T-COFFEE) in Jalview
2. To compare and analyze the alignment results from different algorithms
3. To construct a phylogenetic tree using MEGA software based on the aligned sequences
4. To interpret the evolutionary relationships among the sequences

THEORY

Multiple Sequence Alignment (MSA)

Multiple Sequence Alignment is a fundamental technique in bioinformatics that aligns three or more biological sequences (DNA, RNA, or protein) to identify regions of similarity. These similarities may indicate functional, structural, or evolutionary relationships among sequences.

Key Alignment Algorithms:

1. **Clustal Omega**: Uses HMM (Hidden Markov Model) profiles and produces accurate alignments for large datasets. It employs a progressive alignment approach with improved scalability.
2. **MUSCLE (Multiple Sequence Comparison by Log-Expectation)**: Achieves better accuracy and speed through iterative refinement. Uses k-mer counting for distance estimation and progressive alignment followed by refinement.
3. **MAFFT (Multiple Alignment using Fast Fourier Transform)**: Employs FFT for rapid detection of homologous regions. Offers various strategies (L-INS-i, G-INS-i, E-INS-i) for different scenarios.
4. **T-COFFEE (Tree-based Consistency Objective Function For Alignment Evaluation)**: Uses a consistency-based approach that combines information from all pairwise alignments. Generally, more accurate but computationally intensive.

Phylogenetic Analysis

Phylogenetic trees represent evolutionary relationships among organisms or genes. The branch lengths often correspond to evolutionary time or genetic distance. Trees can be:

- **Rooted**: Shows common ancestor and evolutionary direction
- **Unrooted**: Shows relationships without indicating ancestry direction

Common Tree Construction Methods:

- Neighbor-Joining (NJ): Distance-based method, fast and suitable for large datasets
- Maximum Likelihood (ML): Statistical method, computationally intensive but accurate
- Maximum Parsimony (MP): Finds tree requiring the minimum evolutionary changes

REQUIREMENTS

Software Prerequisites:

- Jalview (version 2.11 or later)
- MEGA X (version 10 or later)
- Internet connection for web services

Sample Dataset:

Use cytochrome c protein sequences from 8 different organisms (download from the Lab's GitHub Page):

- Human (*Homo sapiens*)
- Chimpanzee (*Pan troglodytes*)
- Mouse (*Mus musculus*)
- Chicken (*Gallus gallus*)
- Zebrafish (*Danio rerio*)
- Fruit fly (*Drosophila melanogaster*)
- Yeast (*Saccharomyces cerevisiae*)
- Arabidopsis (*Arabidopsis thaliana*)

PROCEDURE

Part A: Multiple Sequence Alignment in Jalview (30 minutes)

Step 1: Data Preparation (5 minutes)

1. Open Jalview software
2. Click **File → Input Alignment → From File**
3. Load the provided FASTA file containing 8 cytochrome c sequences
4. Verify all sequences are loaded correctly in the alignment window

Step 2: Perform Alignment with Clustal Omega (5 minutes)

1. Select **Web Service → Alignment → Clustal Omega**
2. Keep default parameters
3. Click **OK** to submit job

4. Wait for alignment completion
5. Save alignment: **File → Save As →** "ClustalO_alignment.aln"

Step 3: Perform Alignment with MUSCLE (5 minutes)

1. Reload original unaligned sequences
2. Select **Web Service → Alignment → MUSCLE**
3. Use default parameters
4. Submit and wait for completion
5. Save as "MUSCLE_alignment.aln"

Step 4: Perform Alignment with MAFFT (5 minutes)

1. Reload original sequences
2. Select **Web Service → Alignment → MAFFT**
3. Choose "auto" strategy
4. Submit and save as "MAFFT_alignment.aln"

Step 5: Perform Alignment with T-COFFEE (5 minutes)

1. Reload original sequences
2. Select **Web Service → Alignment → T-COFFEE**
3. Use default mode
4. Submit and save as "TCOFFEE_alignment.aln"

Step 6: Compare Alignments (5 minutes)

1. Open each saved alignment in separate Jalview windows
2. Observe differences in:
    - Gap positions and distributions
    - Conserved regions (highlighted columns)
    - Overall alignment length
3. Export one best alignment in FASTA format for phylogenetic analysis

Part B: Phylogenetic Tree Construction in MEGA (25 minutes)

Step 7: Import Alignment into MEGA (5 minutes)

1. Open MEGA X software
2. Click **DATA → Open a File/Session**
3. Select the exported FASTA alignment file (Let's choose the "MUSCLE_alignment.aln", as the protein sequences show clear conservation patterns)
4. Choose **Protein sequences** when prompted
5. Click **Analyze** for phylogenetic analysis

Step 8: Construct Neighbor-Joining Tree (10 minutes)

1. Select **PHYLOGENY → Construct/Test Neighbor-Joining Tree**
2. Set parameters:
    - Test of Phylogeny: Bootstrap method
    - No. of Bootstrap Replications: 100
    - Model: p-distance
    - Gaps/Missing Data: Pairwise deletion
3. Click **Compute**
4. View and save the tree

Step 9: Tree Visualization and Annotation (5 minutes)

1. In the tree viewer:
    - Click **View → Topology Only** for clearer view
    - Show bootstrap values: **View → Show/Hide → Bootstrap Values**
    - Root the tree: Right-click on outgroup branch → **Root**
2. Identify clusters of closely related organisms
3. Note bootstrap support values (>70% indicates strong support)

Step 10: Export Results (5 minutes)

1. Export tree image: **Image → Save as PNG/PDF**
2. Save tree file: **File → Save Session**
3. Export distance matrix: **Data → Export Distances**

OBSERVATIONS AND RESULTS

1. Record your observations in the following table:

Algorithm	**Alignment Length**	**Number of Gaps**	**Highly Conserved Regions**
Clustal Omega			
MUSCLE			
MAFFT			
T-COFFEE			

2. Record the sequence of the region with the longest stretch and highest degree of conservation.

3. Record the CXXCH motif location and its conservation across algorithms

4. Where do you observe most gaps? Are there differences in the way gaps are introduced between different alignments?

5. Based on the phylogenetic tree analysis, which pair of species was the most closely related? Does this match known evolutionary relationships and timeline?

6. Why do bootstrap values matter while performing phylogenetic analysis?

7. Discuss the difference you observed between a rooted and unrooted tree.