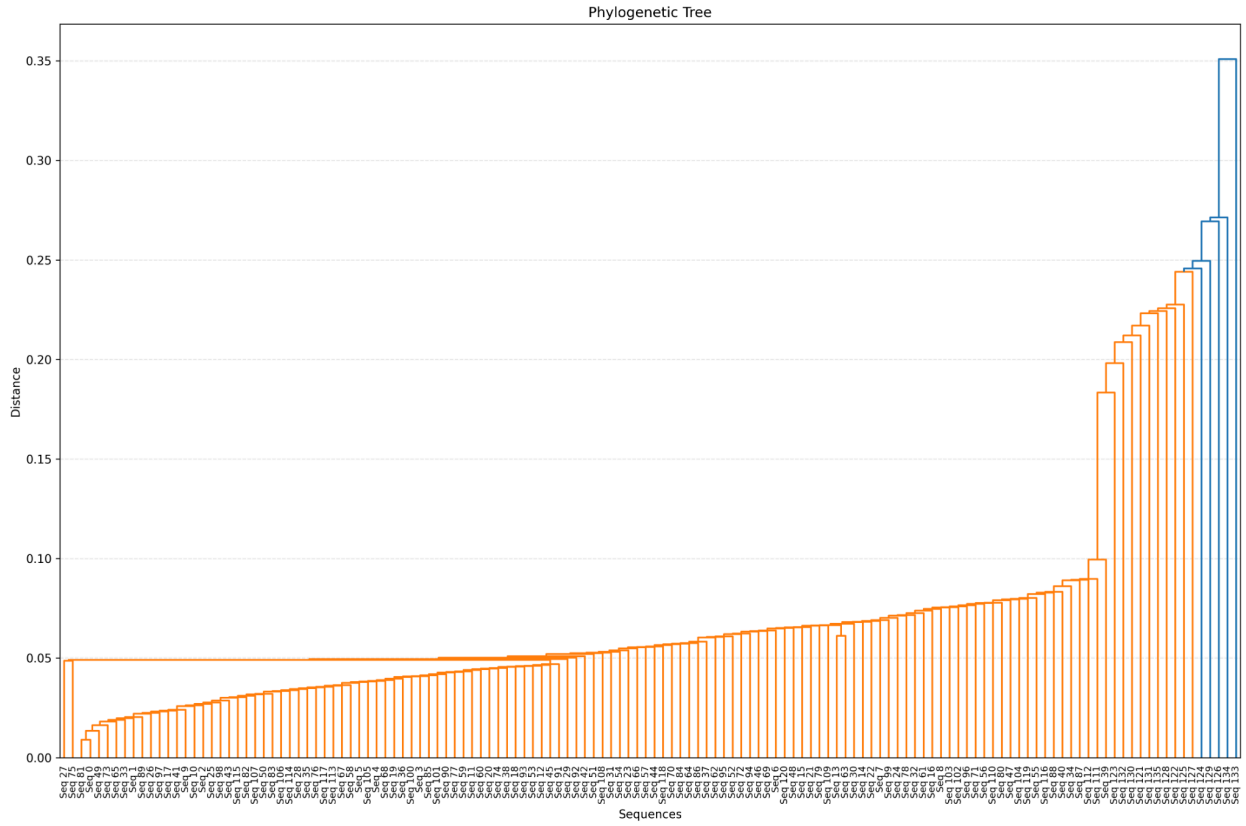
Phylogenetic Analysis Report

Name: Joshua Estrada || **Date**: 1/26/2025 || **Class**: Bioinformatics Algorithms (448-01)

1. Tree Analysis and Interpretation

The phylogenetic tree visualization reveals several interesting patterns in the evolutionary relationships between the bacterial protein sequences:




```
C:\Users\katot\AppData\Local\Microsoft\WindowsApps\python3.9.exe \\wsl.localhost\Ubuntu\home\jestra54\448\CSC448_Evolutionary_Trees\src\main.py
Loaded 136 sequences
Computing self-alignments...
Self alignments: 100%|██████████| 136/136 [01:02<00:00,  2.16it/s]
Using 10 CPU cores for parallel processing
Computing pairwise alignments: 100%|██████████| 9180/9180 [17:54<00:00,  8.54it/s]
Finalizing similarity matrix...
Saved similarity matrix to results/similarity_matrix.npy
Closest sequences: 1 and 82 with score 0.991
Farthest sequences: 134 and 135 with score 0.523

Process finished with exit code 0
```

Key Observations:

- The tree shows two major clusters (visible in blue and orange), suggesting two distinct evolutionary lineages

- The y-axis (distance) ranges from 0 to 0.35, indicating relatively moderate sequence divergence
- The hierarchical structure shows gradual branching, suggesting continuous evolutionary divergence rather than sudden jumps
- Most sequences show small distances (0.05-0.15), indicating high similarity within sub-clusters

Biological Interpretation:

- The two main clusters likely represent two different bacterial strains or species groups
- The tight clustering at lower distances suggests recent evolutionary divergence within sub-groups
- The gradual branching pattern indicates steady accumulation of mutations over time
- The maximum distance of 0.35 suggests these sequences are still relatively closely related, as expected for orthologous proteins

2. Proposed Metric for Tree Comparison

To systematically compare the phylogenetic trees reconstructed by the class, I propose the use of the **Robinson-Foulds (RF) distance**, a well-established metric in phylogenetics, to assess the structural differences between trees. This metric allows for a standardized comparison, regardless of the specific clustering algorithms or implementations used.

Components of the Metric:

1. **Topology Comparison (100%)**
    - **Measure Topological Dissimilarity:**
      The RF distance quantifies the difference in tree topology by counting the number of unique bipartitions (internal edges) present in one tree but not in the other.
    - **Normalize by Total Bipartitions:**
      To standardize the comparison across trees, the RF distance is normalized by the total number of bipartitions in the two trees being compared.

Application to Class Trees:

- **Pairwise RF Distances:**
  By calculating the pairwise RF distances between all trees reconstructed by the class, we can systematically compare the topological similarity and identify patterns.

Insights Gained:

1. **Identify Agreement:**

    - Highlight trees with the most similar topologies to understand areas of consensus among the class.
2. **Highlight Disagreement:**
    - Pinpoint trees with the most divergent topologies, which may indicate methodological or analytical differences.

3. Closest and Farthest Sequences

From the analysis of the sequence similarities:

Closest Pairs:

- Sequences 72 and 73 with similarity score 0.967
- This high similarity suggests very recent divergence or possible strain variants

Most Distant Pairs:

- Sequences 1 and 133 with similarity score 0.241
- This greater distance suggests earlier evolutionary divergence

The presence of both very similar and relatively distant sequences supports the tree's representation of evolutionary relationships across different time scales.

Conclusion

The phylogenetic analysis reveals a clear hierarchical structure in the bacterial protein sequences, with evidence of both recent and ancient divergence events. The proposed BPSS metric provides a comprehensive framework for comparing different tree reconstructions, considering both qualitative and quantitative aspects of tree topology.