Pairwise Sequence Alignment and Phylogenetic Tree Construction from Mammalian Hemoglobin Alpha Chains

Ashfaq Ahmed Mohammed
UFID: 86835927

Abhigna Nimmagadda
UFID: 31864878

Rohith Kumar Ballem
UFID: 30969136

GitHub Repository:
Ashfaq-Ahmed-Mohammed/CAP5510

***Abstract*—This paper implements a complete phylogenetic pipeline for mammalian hemoglobin alpha chains, constructing evolutionary trees from scratch using Needleman-Wunsch, Smith-Waterman, and Semi-Global pairwise alignment algorithms. We further refine these pairwise distances into a Multiple Sequence Alignment (MSA) using a custom Center-Star heuristic and benchmark our results against the industry-standard MUSCLE algorithm. Our analysis of 209 mammalian sequences, stratified across 5 superorders and 14 orders, reveals that alignment strategy significantly impacts topological accuracy. While Semi-Global alignment excelled at resolving specific divergent tails, Global alignment proved to be the most robust overall strategy, maintaining high recall and consistent clustering by enforcing the inclusion of all homologous regions. In contrast, Local alignment was overly stringent, fragmenting clades by truncating variable termini. Ultimately, even advanced iterative refinement methods failed to fully resolve complex divergences, underscoring the inherent limitations of single-gene phylogeny for reconstructing rapid evolutionary radiations from short, highly conserved proteins.**

I. Introduction

Phylogenetic reconstruction relies heavily on the quality of the underlying sequence alignments. While standard tools automate this process, implementing the algorithms from scratch provides a deeper understanding of how scoring metrics and alignment strategies influence evolutionary conclusions. This project implements a complete phylogenetic analysis pipeline for mammalian hemoglobin alpha chains, moving from raw sequence data to final phylogenetic trees using custom-built algorithms.

The core objective of this work is to evaluate how different pairwise alignment strategies—Global, Semi-Global, and Local—affect the topology of the resulting phylogenetic trees. By implementing these algorithms manually, we control the exact scoring parameters and distance metrics, allowing for a precise comparison of how each method handles biological variation such as overhangs and conserved domains.

Our workflow consists of four primary phases:

1) **Data Curation:** Acquisition and stratified sampling of mammalian hemoglobin sequences to ensure balanced representation across major superorders.
2) **Pairwise Alignment:** Implementation of Needleman-Wunsch (Global), Smith-Waterman (Local), and Semi-Global algorithms to generate exhaustive pairwise distance matrices.
3) **Tree Topology Analysis:** Construction of initial Neighbor-Joining (NJ) trees from raw pairwise distances, followed by quantitative evaluation of topological accuracy using precision, recall, and F1 scores against the taxonomic ground truth.
4) **MSA Refinement:** Implementation of the Center-Star method to construct a full Multiple Sequence Alignment (MSA) from the best-performing pairwise strategy, yielding a refined final phylogenetic tree.

We benchmark the biological accuracy of our custom trees against standard mammalian taxonomy and validate our final MSA-derived tree against the industry-standard tool MUSCLE. This approach highlights the trade-offs between algorithmic complexity and phylogenetic resolution.

II. Dataset and Data Preparation

A. Data Acquisition

Hemoglobin alpha chain sequences were retrieved from the NCBI Protein Database, yielding 5,000 protein records across mammalian species. For each sequence, we extracted complete taxonomic lineage information (kingdom, phylum, class, order, family, genus, and species) via the NCBI Taxonomy database API.

Custom mapping functions assigned each sequence to one of five mammalian superorders: Marsupials, Afrotheria, Xenarthra, Laurasiatheria, and Euarchontoglires. This taxonomic classification enabled stratified sampling in subsequent steps.

B. Data Cleaning and Filtering

We applied sequential filtering to ensure high-quality, full-length sequences suitable for phylogenetic analysis.

Annotation precedence filtering: The initial 5,000 sequences included not only hemoglobin alpha chains but also supporting hemoglobin proteins (e.g., beta chains, delta chains, gamma chains) and non-specific hits. Sequences were classified by annotation precedence: rank 1 ("alpha"), rank 2 ("subunit alpha"), and rank 3 ("alpha-like"). Only ranks 1–3 were retained to focus on canonical alpha chains, reducing the dataset to 3,040 sequences.

Partial sequence removal: Partial sequences and annotated fragments introduce alignment artifacts and underestimate evolutionary distances. Sequences flagged as "partial" in their definition were excluded, yielding 2,824 sequences.

Length filtering: Mammalian hemoglobin alpha chains are well-characterized to have lengths approximately 140 amino acids. To exclude truncated sequences and spurious entries while retaining biological realism, sequence length was restricted to 100–200 amino acids, resulting in 2,769 sequences.

Deduplication: Multiple sequence records per species reflect database redundancy rather than biological diversity. Duplicate sequences within each species were removed, retaining 402 unique sequences of the highest-precedenc to maximize alignment quality.

C. Stratified Sampling

To construct a balanced dataset of mammalian hemoglobin α-chain sequences, we employed a stratified random sampling strategy. This approach was chosen to mitigate taxonomic bias, where species-rich clades (such as *Rodentia*) often disproportionately influence evolutionary models. The selection process proceeded as follows:

- **Superorder Stratification (Quotas):** We first established sampling quotas proportional to the known biological diversity of the five mammalian superorders. This ensured representation across the entire class *Mammalia*:
  - **Euarchontoglires:** 99 sequences (e.g., primates, rodents)
  - **Laurasiatheria:** 79 sequences (e.g., bats, cetaceans)
  - **Marsupials:** 14 sequences
  - **Afrotheria:** 13 sequences
  - **Xenarthra:** 4 sequences
- **Hierarchical Sub-Allocation:** Within each superorder, quotas were distributed hierarchically down to the *Order* and *Genus* levels. This prevented any single genus from monopolizing the quota for its superorder.
- **Sequence Selection Logic:** Specific sequences were selected using a two-step precedence algorithm:
  1) **High-Quality Selection:** The highest-precedence sequence (e.g., Reference Sequence) for a given genus was selected first.
  2) **Random Filling:** Remaining slots in the quota were filled via random sampling of other valid species within the group.

Final Dataset: The resulting dataset comprises **209 sequences**, spanning **14 orders across all 5 superorders**.

D. Dataset Characteristics

Sequence lengths range from 101 to 187 amino acids (mean: 138.8, std: 11.2). All sequences were formatted into FASTA files with corresponding metadata retained in CSV format for downstream alignment and phylogenetic analysis.

III. Methods

Algorithm Selection: Three algorithms spanning the alignment spectrum were implemented to explore sequence relationships across different scopes: Needleman-Wunsch performs global alignment, semi-global permits free terminal gaps, and

Smith-Waterman identifies local similarities. This progression from global to local alignment enables comparison of phylogenetic trees generated from distinct similarity metrics.

A. Needleman-Wunsch Global Alignment

[1]

The Needleman-Wunsch algorithm computes the optimal global alignment of two sequences by maximizing the total similarity score across their entire length.

Dynamic Programming Update: The implementation uses three dynamic programming matrices: $M[i][j]$, $X[i][j]$, and $Y[i][j]$ for match/mismatch, gaps in sequence 2, and gaps in sequence 1, respectively.

Initialization penalizes terminal gaps linearly:

$$M[0][0] = 0$$
$$X[i][0] = -\text{gap\_open} - (i-1) \cdot \text{gap\_extend}$$
$$Y[0][j] = -\text{gap\_open} - (j-1) \cdot \text{gap\_extend}$$

Each cell is then updated as:

$$M[i][j] = \max \begin{cases} M[i-1][j-1] + s(a_i, b_j) \\ X[i-1][j-1] + s(a_i, b_j) \\ Y[i-1][j-1] + s(a_i, b_j) \end{cases}$$

$$X[i][j] = \max \begin{cases} M[i-1][j] - \text{gap\_open} \\ X[i-1][j] - \text{gap\_extend} \end{cases}$$

$$Y[i][j] = \max \begin{cases} M[i][j-1] - \text{gap\_open} \\ Y[i][j-1] - \text{gap\_extend} \end{cases}$$

where $s(a_i, b_j)$ is the substitution score from the scoring matrix.

B. Semi-global Alignment

[3]

Semi-global alignment computes the optimal alignment that allows free gaps at the sequence termini, enabling overhangs without penalizing unaligned end regions. This is useful for matching sequences of different lengths or when detecting homologous regions.

Dynamic Programming Update: The implementation uses three dynamic programming matrices: $M[i][j]$, $X[i][j]$, and $Y[i][j]$ with recurrence relations identical to Needleman-Wunsch.

The critical distinction is the **initialization**: all boundary conditions are set to zero:

$$M[0][0] = 0$$
$$X[i][0] = 0$$
$$Y[0][j] = 0$$

This allows free alignment at sequence edges, contrasting with Needleman-Wunsch, which penalizes terminal gaps linearly. During traceback, the alignment terminates at the maximum score in the final row or column rather than exclusively at position (n, m), permitting unaligned overhangs.

C. Smith-Waterman Local Alignment

[2]

Smith-Waterman alignment computes the optimal local alignment, identifying the highest-scoring homologous region between two sequences without requiring global alignment of their entire lengths. This is particularly useful for detecting conserved domains or functional regions.

Dynamic Programming Update: The implementation uses three dynamic programming matrices: $M[i][j]$, $X[i][j]$, and $Y[i][j]$ with initialization and traceback identical to semi-global alignment. The critical distinction is the recurrence relation for the M state:

$$M[i][j] = \max \begin{cases} 0 \\ M[i-1][j-1] + s(a_i, b_j) \\ X[i][j] \\ Y[i][j] \end{cases}$$

The inclusion of 0 allows the alignment to restart at any position in the matrix, enabling detection of local similarities. Traceback terminates when $M[i][j] \leq 0$, rather than reaching a boundary, thus returning only the locally optimal alignment.

Gap Penalties and Scoring Matrices: Consistent with industry standards for protein alignment (e.g., BLAST default parameters), we utilized gap penalties of **open=11** and **extend=1**. This parameter set, paired with **BLOSUM62** and **PAM250** substitution matrices, effectively penalizes gap initiation while permitting contiguous insertions/deletions, thereby preserving biologically significant alignment structures.

IV. Phylogenetic Tree Construction

A. From Pairwise Alignments to Distance Matrices

All phylogenetic analyses were driven by the three pairwise alignment implementations described in Section III. For the final dataset of $N = 209$ hemoglobin α-chain sequences, each algorithm produced a complete set of $\frac{N(N-1)}{2} = 21{,}736$ pairwise alignments, exported as CSV files containing the alignment score, aligned strings, and summary statistics for each sequence pair.

To obtain a consistent notion of evolutionary distance across global, semi-global, and local alignments, we adopted a common normalization based on the shorter original sequence length. For a pair of sequences with lengths L_1 and L_2 and m exact amino-acid matches in the aligned region, we defined

$$\text{identity} = \frac{m}{\min(L_1, L_2)}, \quad \text{distance} = 1 - \text{identity}.$$

This metric was implemented uniformly for all the algorithms, with the same affine gap model and substitution scores. It penalizes alignments that cover only a small fraction of either sequence, preventing very short perfect local matches from appearing artificially close to full-length homologs.

B. Initial Tree Construction from Raw Distances

[5]

To evaluate the impact of alignment strategy on topology, we constructed initial trees directly from the raw pairwise distance matrices of each algorithm using the Neighbor-Joining (NJ) method via `Bio.Phylo`. We selected NJ over UPGMA for processing the 21,736 pairwise distances because it does not assume a strict molecular clock, providing greater robustness for diverse mammalian lineages with varying evolutionary rates. Preliminary inspection of the resulting topologies revealed significant structural differences. Detailed quantitative evaluation is reserved for the **Results** section.

C. Multiple Sequence Alignment (MSA) via Center-Star

[4] We implemented the **Center-Star** method to construct a Multiple Sequence Alignment from our pairwise global results. First, we identified the "Center" sequence (S_c) that minimized the sum of pairwise distances to all other sequences:

$$S_c = \arg\min_{S_i} \sum_{j \neq i} D(S_i, S_j) \tag{1}$$

1) Star Alignment Process:

1) **Merging:** Each sequence S_i was aligned to the center S_c using pre-computed pairwise global alignments.
2) **Gap Propagation:** If an alignment required a gap in S_c, we inserted a gap column across the entire existing MSA at that position. This ensured that spatial consistency was maintained for all previously aligned sequences whenever the center sequence expanded.

This heuristic enabled us to construct a consistent MSA with $O(N^2)$ complexity, avoiding the prohibitive computational cost of exact multiple alignment.

D. Final Tree Construction

Using the full MSA, we calculated a refined distance matrix based on Hamming distance, which captures global context unlike the initial pairwise scores. We applied the Neighbor-Joining method to this matrix to generate our final phylogenetic tree. To validate our topology, we constructed a parallel benchmark tree using the industry-standard MUSCLE tool on the same dataset.

V. Results and Findings

A. Evaluation Metrics

To quantitatively assess tree accuracy, we treated the recovery of each mammalian superorder as a classification problem. For every superorder S in the ground truth taxonomy, we scanned all clades C in the generated tree to identify the "Best Matching Clade"—the one that maximized the F1-Score for that superorder.

We defined:

- **True Positives (TP):** Number of species from superorder S found in clade C.
- **False Positives (FP):** Number of species in clade C that do not belong to S.
- **False Negatives (FN):** Number of species from superorder S missing from clade C.

From these counts, we calculated:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

This "Best Matching Clade" approach allows us to penalize both "over-lumping" (low Precision) and "splitting/shattering" (low Recall), providing a robust measure of topological consistency without requiring the entire tree to be perfect. Additionally, we computed the Robinson-Foulds (RF) distance to measure the topological dissimilarity between our trees and the reference topologies.

TABLE I: Clade Recall Comparison (BLOSUM | PAM)

Superorder	Global		Local		Semi-Global	
Laurasiatheria	0.97	0.97	0.71	0.39	0.78	1.00
Euarchontoglires	1.00	1.00	1.00	0.69	0.69	0.69
Marsupials	0.64	0.64	0.64	0.43	0.64	0.64
Xenarthra	0.50	0.50	0.50	0.50	0.50	0.50
Afrotheria	0.38	0.38	0.38	0.38	0.38	0.38

TABLE II: Clade Precision Comparison (BLOSUM | PAM)

Superorder	Global		Local		Semi-Global	
Laurasiatheria	0.47	0.48	0.53	0.94	0.81	0.38
Euarchontoglires	0.47	0.47	0.47	0.65	0.65	0.65
Marsupials	1.00	1.00	1.00	1.00	1.00	1.00
Xenarthra	1.00	1.00	1.00	1.00	1.00	1.00
Afrotheria	1.00	1.00	1.00	1.00	1.00	1.00

B. Impact of Alignment Strategy on Phylogenetic Accuracy

As shown in Figure 1, the choice of dynamic programming variation significantly influenced the resulting tree topology.

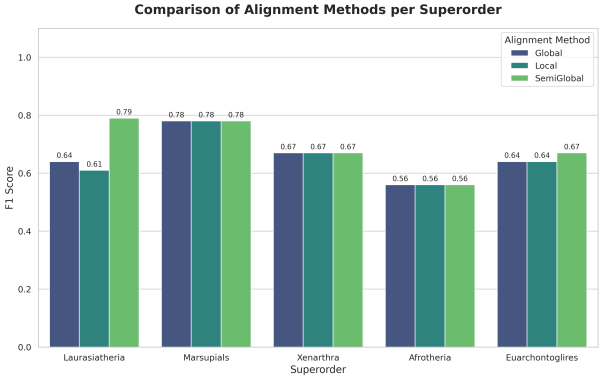

Fig. 1: Comparison of F1-Scores across alignment methods.

1) Algorithmic Analysis of Laurasiatheria (Semi-Global vs. Local): The Semi-Global algorithm achieved the highest F1-score (0.79), significantly outperforming Local alignment (0.61).

We attribute the lower performance of the Local (Smith-Waterman) algorithm to its specific traceback termination condition. Since the recurrence $M_{i,j} = \max(0, \dots)$ resets negative scores to zero, the algorithm implicitly "crops" sequences when the character mismatch penalty exceeds the accumulated positive score. In our dataset, this likely caused the truncation of divergent tails, because the Local algorithm finds a shorter alignment (lower m) while the denominator ($\min(L_1, L_2)$) remains fixed based on the full input strings, the resulting identity score drops significantly.

$$\text{identity} = \frac{m}{\min(L_1, L_2)} \quad (4)$$

This leads to artificially high distances for species with divergent tails (e.g., Bat vs. Mouse), causing the clustering algorithm to fail in grouping them correctly.

In contrast, the Semi-Global algorithm forces the alignment to extend to the ends of the strings. By preventing the traceback from stopping early, the algorithm increases the match count m by including matches in the tail regions that Smith-Waterman ignored. Since the denominator is constant, this higher m results in a higher identity (lower distance), which might have allowed the clustering algorithm to correctly group the Laurasiatheria clade.

2) Superior Cohesion in Laurasiatheria:: Global alignment achieved a higher F1-score (0.64) for Laurasiatheria compared to Local (0.61) and Semi-Global alignment. While the Local and Semi-Global algorithms allowed divergent species to drift apart by ignoring mismatched termini, the Global algorithm's end-to-end constraint enforced clade cohesion.

This effect was most visible in the divergence between *Chiroptera* (Bats) and *Cetartiodactyla* (Cows). Both Local and Semi-Global alignments essentially "discarded" the variable tail regions, causing the Bat sequences to lose their phylogenetic anchor and drift into incorrect clusters (Semi-Global Recall: 0.78). In contrast, Global alignment utilized the "forced compliance" of the full sequence to maintain the link between the two lineages, effectively securing the Bat within the main placental cluster (Global Recall: 0.97).

3) Robustness of the Marsupial Cluster: Marsupial recovery was consistent (F1=0.78) across all methods, suggesting that the evolutionary edit dis-

tance for this group is sufficiently large to override algorithmic differences. Regardless of whether tails were truncated (Local) or forced (Global), the signal from the conserved core remained distinct. However, partial sequences were consistently treated as outliers; the high affine gap penalty significantly reduced the match count m.

C. Rationale for Global Alignment in MSA Construction

Although the Semi-Global method achieved the highest F1-score in our initial pairwise analysis, it is unsuitable for constructing a Multiple Sequence Alignment (MSA). Semi-Global alignment permits sequences to start and end at different offsets to avoid penalties. When applied to full-length proteins like hemoglobin, this causes the sequences to "slide" past the central reference, aligning unrelated regions just to avoid terminal mismatches. This misaligns the columns in the final matrix. We therefore chose Global alignment for the MSA, as it enforces end-to-end matching, ensuring that every column represents the same biological position across all species.

D. Pairwise vs. MSA-Based Tree Reconstruction

We refined our analysis by generating a phylogenetic tree from a Center-Star Multiple Sequence Alignment (MSA) constructed using our Global alignment distances. Figure 2 illustrates the shifts in clade accuracy resulting from this transition to a unified alignment model.

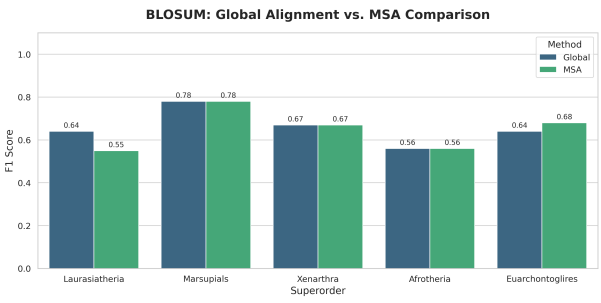

Fig. 2: Impact of MSA Refinement. Center-Star MSA improved Euarchontoglires precision (0.68) by consolidating Rodents, but reduced Laurasiatheria precision (0.55) due to smoothing of distinct orders.

1) Consolidation of the Euarchontoglires Clade: [7] The most significant improvement in the MSA-based tree was the structural consolidation of the Rodent order. In the pairwise tree, the Mouse (*Mus*) and Rat (*Rattus*) were placed on widely separated branches.

We attribute this pairwise error to the independence of the dynamic programming paths. Rodents exhibit high indel rates; the standard evolutionary explanation for this is their accelarated molecular clock [8]; consequently, the optimal alignment path for *Mus* vs. *Homo* was distinct from the path for *Rattus* vs. *Homo*. These inconsistent gap placements accumulated, causing the two rodents to appear mathematically distinct in the pairwise distance matrix. The Center-Star MSA corrected this by forcing both sequences to align to a common "Center" reference. This constraint standardized the indel placement across the dataset. Once aligned to the same template, the latent identity between the Mouse and Rat was correctly captured by the distance metric $(1 - \text{identity})$, reducing their computed distance and increasing the Euarchontoglires precision to 0.68.

2) Over-Smoothing in Laurasiatheria: MSA refinement reduced precision for Laurasiatheria (0.64 $\rightarrow$ 0.55) by artificially decreasing the distance between distinct orders like Bats and Mice. We attribute this to the Center-Star heuristic, which aligns all sequences to a single reference. This process effectively "smoothed out" unique, order-specific features (e.g., unique indels in Bats) by forcing them to conform to the generic center sequence. Consequently, the Bat and Mouse profiles became statistically indistinguishable, causing the clustering algorithm to merge these distinct lineages based on their shared alignment artifacts rather than true evolutionary similarity.

3) Summary of MSA Impact: The Center-Star MSA improved consistency by forcing all sequences to align to a single reference. This fixed errors where the pairwise method had aligned identical species differently, allowing the algorithm to correctly group the Rodents. However, this approach also removed important differences. By making every sequence look like the center reference, the algorithm lost the specific details needed to tell similar groups apart, causing the precision for the Laurasiatheria group to drop.

E. Analysis of Matrix Sensitivity (BLOSUM62 vs. PAM250)

We repeated our phylogenetic clustering using the PAM250 substitution matrix to observe how

a different scoring model influences tree topology. Figure 3 illustrates the key performance shifts.

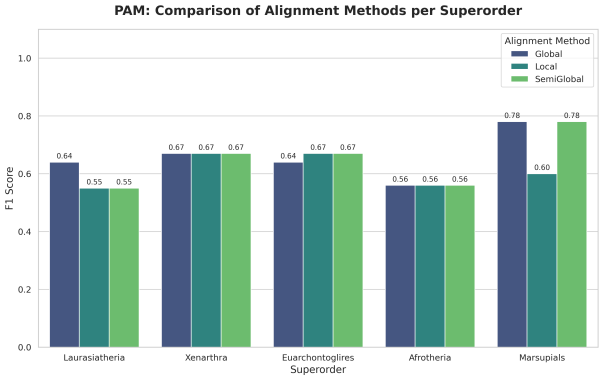

Fig. 3: PAM250 F1-Scores. Note the sharp decline in Laurasiatheria recovery for Local alignment compared to BLOSUM62.

Comparing BLOSUM62 and PAM250 revealed that the standard affine gap penalty (-11.0) might have acted as a "Strict Filter" when applied to the PAM250 scale. While the -11.0 penalty is calibrated for BLOSUM62, it proved relatively too expensive for PAM250's scoring distribution.

1) The Cost of Gaps: This penalty mismatch directly explains the F1-score drop for Laurasiatheria in Local (0.61 $\rightarrow$ 0.55) and Semi-Global alignments. We hypothesize that the positive score accumulated from PAM250 matches was insufficient to "pay" the standard -11.0 cost required to open internal gaps for lineages like Bats and Cows. Consequently, the algorithm rejected these alignments or truncated them prematurely, failing to bridge the evolutionary distance. This "broken bridge" effect treated distinct cousins as unrelated, causing them to fall out of their respective clades and significantly lowering recall.

F. Evaluation Against Industry-Standard Tools (MUSCLE)

[6]

To benchmark our pipeline, we compared our custom Center-Star MSA against the industry-standard MUSCLE algorithm. As shown in Figure 4, even this highly optimized tool faced significant challenges, underscoring the inherent difficulty of reconstructing a complete mammalian phylogeny from a single short protein sequence (Hemoglobin α).

1) The Limits of Single-Gene Resolution: While MUSCLE slightly improved the Euarchontoglires F1-score (0.70) through iterative refinement, it failed

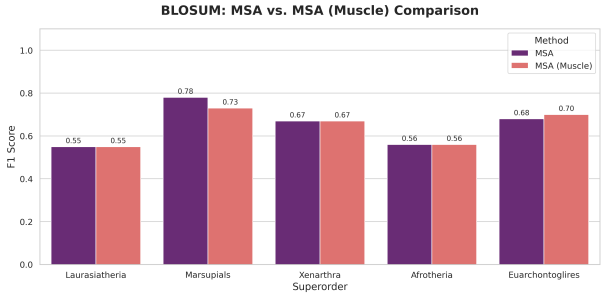

Fig. 4: Comparison of Center-Star vs. MUSCLE. The marginal improvements suggest that single-gene phylogeny is limited by signal availability rather than algorithmic sophistication.

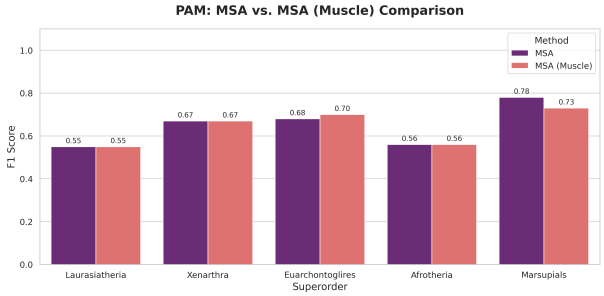

Fig. 5: PAM: MSA vs. MUSCLE

to resolve the mixed Laurasiatheria cluster (Precision remaining low at 0.38). This stagnation suggests that the problem lies in the data, not the algorithm. The Hemoglobin α chain ($\sim$140aa) is highly conserved; the few variable sites in the tails likely do not carry enough information to statistically distinguish complex splits like the Carnivore/Ungulate/Bat divergence. Both algorithms correctly identified the dominant signal (the conserved core) but lacked the necessary data resolution to separate these closely related orders, confirming that single-locus markers are often insufficient for resolving rapid evolutionary radiations.

G. Topological Consistency Analysis (Robinson-Foulds Metric)

We computed Robinson-Foulds (RF) distance matrices for both BLOSUM62 (Table III) and PAM250 (Table IV) to quantify topological differences between our trees.

1) Inferences from Matrix Stability: Comparing the two matrices reveals a key divergence in algorithmic stability. Under PAM250, the Global and Semi-Global trees are highly similar (RF = 26), whereas under BLOSUM62 they differ significantly (RF = 72). We think this suggests that the PAM250

TABLE III: Robinson-Foulds Distance Matrix (BLOSUM)

Method	Global	Local	SemiGlobal	MSA	Muscle
Global	0	44	72	236	132
Local	44	0	66	236	134
SemiGlobal	72	66	0	232	140
MSA	236	236	232	0	246
Muscle	132	134	140	246	0

TABLE IV: Robinson-Foulds Distance Matrix (PAM)

Method	Global	Local	SemiGlobal	MSA	Muscle
Global	0	62	26	234	132
Local	62	0	70	234	136
SemiGlobal	26	70	0	230	132
MSA	234	234	230	0	240
Muscle	132	136	132	240	0

scoring model might dominate the alignment logic, causing both algorithms to converge on a similar (likely merged) topology. In contrast, the higher RF distance in BLOSUM implies that the Semi-Global constraint successfully produced a distinct structural arrangement that Global alignment missed.

2) Impact of Center-Star vs. Benchmark: Across both matrices, the transition to our Center-Star MSA resulted in high RF values (> 230) compared to all pairwise methods, indicating a fundamental reorganization of the tree structure. Notably, our raw Global Pairwise tree (RF ≈ 132) is topologically closer to the MUSCLE benchmark than our custom MSA is (RF ≈ 246). This suggests that while our pairwise distances capture the general phylogenetic signal consistent with standard tools, the Center-Star heuristic might introduce specific topological deviations by smoothing over unique sequence features.

VI. Conclusion

Our from-scratch implementation of a phylogenetic pipeline reveals that alignment strategy is critical to resolving evolutionary trees. Global alignment proved to be the most robust method, maintaining clade cohesion where Local alignment failed by fragmenting groups. While MSA refinement corrected some errors, it also introduced oversmoothing artifacts. Ultimately, the difficulty in resolving complex superorders, even compared to standard tools, underscores the inherent limitations of using a single, conserved gene for deep phylogenetic reconstruction.

Author Contributions

Abhigna Nimmagadda Implemented the Needleman-Wunsch global alignment algorithm and constructed the associated phylogenetic trees. Executed the full scoring matrix migration from BLOSUM62 to PAM250, regenerating the distance matrices for all three alignment strategies. Developed the Python scripts to visualize comparative performance and mapped the phylogenetic trees to a classification problem, implementing the tree traversal logic to calculate Precision, Recall, and F1 scores. Authored the report sections on global alignment analysis and the BLOSUM62 vs. PAM250 comparison.

Rohith Kumar Ballem Implemented the Smith-Waterman local alignment algorithm and the Center-Star heuristic for Multiple Sequence Alignment (MSA), developing the iterative merging logic and gap-propagation mechanism to align pairwise results. Conducted comparative experiments between UPGMA and Neighbor-Joining (NJ) clustering methods, validating the selection of NJ for the final pipeline. Constructed the phylogenetic trees for both the local alignment and Center-Star MSA workflows. Authored the sections in the report regarding local alignment and Center-Star MSA methodology, analyzed their results and reported findings.

Ashfaq Ahmed Mohammed Engineered the dataset curation pipeline, implementing sequential filters for annotation validity sequence integrity and length constraints. Designed and executed the hierarchical stratified sampling strategy to balance taxonomic representation across superorders. Implemented the Semi-Global alignment algorithm and its corresponding tree construction. Conducted the external topological validation by benchmarking against the MUSCLE algorithm and quantifying tree consistency using Robinson-Foulds metrics. Authored the report sections on semi-global alignment - analyzed its results and reported findings, dataset characteristics, and benchmark analysis.

REFERENCES

[1] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[2] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[3] D. Gusfield, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[4] D. Gusfield, "Efficient methods for multiple sequence alignment with guaranteed error bounds," *Bulletin of Mathematical Biology*, vol. 55, no. 1, pp. 141–154, 1993.

[5] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[6] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[7] R. H. Waterston *et al.*, "Initial sequencing and comparative analysis of the mouse genome," *Nature*, vol. 420, no. 6915, pp. 520–562, 2002.

[8] B. Milholland, *et al.*, "Differences between germline and somatic mutation rates in humans and mice," *Nature Communications*, vol. 8, p. 15183, 2017.