Exploring Phylogenetic Tree Construction:

A Framework for Initial Algorithm Selection

Jay Annadurai, Noah Hamilton, Tejaswini Repala, Prasanth Kumar Thuthika

Luddy School of Informatics, Indianapolis University - Indianapolis

Abstract

The study investigated the selection of an initial phylogenetic tree construction algorithm that balances accurate inference of evolutionary history with acceptable computational efficiency. Artificial reference trees were generated using the Phylosim R package under two evolutionary models: a simple, uniform-substitution model (JC69) and a more complex model incorporating unequal substitution rates, invariant sites, and rate heterogeneity (GTR + I + R). Each model was examined with varying numbers of taxa (10, 25, 50) and sequence lengths (500, 1000, 1500), producing 18 simulated evolutionary histories.

Reconstructed trees were generated using distance-based methods—Neighbor-Joining, Minimum Evolution, UPGMA—and character-based methods—Maximum Likelihood, and Maximum Parsimony algorithms—within MEGA11. The reconstructed trees are then evaluated against the reference trees using Maximum Agreement Subtree , CopheneticL2 Weighted, and Weighted RFCluster metrics. Time and memory usage were also recorded.

Results showed that increasing sequence length and taxa number improved reconstruction accuracy across most algorithms. Under the simpler JC69 model, distance-based methods (Neighbor-Joining, Minimum Evolution, UPGMA) performed almost identically, delivering acceptable accuracy with low computational overhead. The complex GTR + R + I model demonstrated Maximum Parsimony's inability to capture complex evolutionary histories and Maximum Likelihood's ability to do so at the cost of extremely high computational time. Space complexity was relatively insignificant within the scale of the data.

Overall, distance-based methods suffice when evolutionary processes are relatively uniform, while Maximum Likelihood is preferable for capturing intricate evolutionary patterns, given sufficient computational resources. Maximum Parsimony is not recommended for complex evolutionary contexts.

The Phylogenetic Tree Construction Algorithm Problem

Definition

Phylogenetic analysis provides insight into evolutionary relationships between genes, organisms, and species from analysis of genetic sequence data (Saitou & Nei, 1987). Phylogenetic analysis quantify the genetic inheritance, evolutionary divergence, and the similarity between entities with a graphical representation known as a phylogenetic tree. Phylogenetic tree analysis provides insights into evolutionary lineage, enabling more accurate predictions of gene function and facilitating the identification of evolutionary patterns critical for biological research. With the advent of high-throughput sequencing methods, phylogenetic analyses have greatly increased access to genetic data but when working with larger datasets, it is imperative to correctly choose an algorithm that can appropriately handle the dataset within computational constraints (Minh et al., 2020; Stamatakis, 2014).

Tree Terminology

Phylogenetic trees are characterized by two primary features: tree topology and branch length. Tree topology refers to the arrangement of branches and nodes, illustrating the relationships and ancestry among the entities being studied, regardless of species, genes, or organisms. Topology defines how groups are connected but does not convey quantitative information about the magnitude of evolutionary change. Branch length is the quantitative measure of evolutionary divergence, corresponding to the genetic differences, mutation rates, or time since the last common ancestor. If branch length is uniform, it most likely suggests the tree was generated with the Molecular Clock hypothesis; that is, divergence can be assumed to occur at a fixed rate. Together, tree topology and branch length offer both a qualitative and quantitative framework to interpret

evolutionary history, revealing evolutionary relationships and the extent or timing of their divergence. Other terms of note are clades, which refers to groups of taxa, and 'rooted' or 'unrooted' trees which respectively refers to if the tree has an assumed, common ancestral point. Subsequent discussion only refers to rooted trees.

Construction Complexities

The complexity of the phylogenetic tree construction problem stems from the intricate nature of evolutionary processes, which introduce challenges in resolving relationships among entities (Felsenstein, 2004). Gene duplication, for instance, can create paralogous sequences within a genome, complicating the distinction between shared ancestry and functional divergence. Similarly, horizontal gene transfer blurs lineage boundaries by allowing genetic material to move across unrelated species, creating evolutionary patterns that traditional tree-like structures may struggle to represent. Variable mutation rates, influenced by factors such as genomic hotspots, selective pressures, and environmental conditions, further obscure the evolutionary distances between sequences, leading to inconsistent phylogenetic signals.

Overall, a single algorithm cannot universally excel in capture of sophisticated biological contexts while remaining computationally efficient. To validate performance across the various evolutionary contexts, there would ideally be a reference set or model set of histories but . As such, genuine experimental data potentially suffers from incomplete lineage sorting, convergent evolution, and limited sampling, which can limit applicability of algorithm performance testing to other contexts; must instead rely on simulated datasets which mimic realistic scenarios but inherently are not exact.

Standard Phylogenetic Tree Construction Algorithms

Algorithms

Phylogenetic tree construction methods that align genetic sequences fall into two broad categories: distance-based and character-based—though they may be further sub-divided as seen in Figure 1.

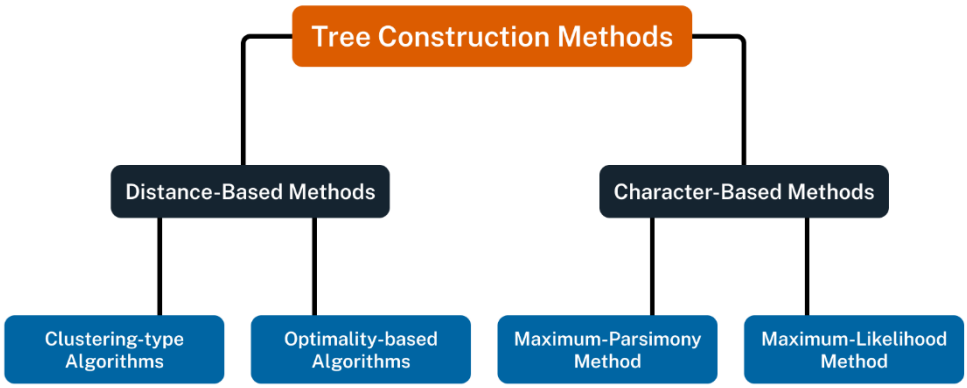


Figure 1: Categorization of Different Phylogenetic Analysis Algorithms

As indicated by their namesake, distance-based algorithms operate on a pair-wise distance matrix of nodes that is statistically derived from a sequence alignment. Common distance based-algorithms are Neighbor-Joining, Minimum Evolution, and Unweighted Pair Group with Arithmetic Mean (UPGMA).

Distance-Based Tree Construction Algorithm	
Neighbor Joining	Iterative Minimization of Branch Lengths Relatively Quick & Efficient ~ $O(n^3)$ Quick; Accurate for Simple Evolutions
Minimum Evolution	Selects Tree of Minimum Total Branch Length Moderately Efficient, Can be Optimized Balanced for both Quickness & Accuracy
Unweighted Pair Group **UPGMA** with Arithmetic Mean	Simple Hierarchal Clustering with Assumption Highly Efficient ~ $O(n^2)$ Optimized only with Molecular Clock data*

Figure 2: Overview of Distance-Based Tree Construction Algorithms

Character-based algorithms rather directly operate on the aligned sequence data, such as Maximum Likelihood, Maximum Parsimony, and Bayesian Inference (Zou et al., 2024). These methods are typically more computationally intensive as they use infer complex evolutionary characteristics not seen within distance-based models.

Character-Based Tree Construction Algorithm	
Maximum Likelihood	Maximize & Check Observation Probability Factorial Scaling, relatively Intensive Accounts for Complex Evolution*
Maximum Parsimony	Scores Trees with Minimal Evolution Events NP-Complete, Intensive at Scale Best for small Datasets with Simple Evolution
Bayesian Inference*	Probabilistic Models with MCMC Simulations Markov Chain Monte Carlo is Very Inefficient Excellent at Handling Complex Evolution

Figure 3: Overview of Character-Based Tree Construction Algorithms

Despite their overall categorization, each algorithm operates with different computational complexities and is optimized for a specific context (See Appendix A). Five of the six algorithms—ML, NJ, ME, UPGMA, and MP—are integrated into the Molecular Evolutionary Genetics Analysis (MEGA) 11 software suite, offering a unified environment for phylogenetic analysis (Tamura et al., 2021). Bayesian Inference, while a powerful approach, is not available in MEGA and typically requires specialized tools like MrBayes which can introduce lurking variables for computational complexity measurements and therefore BI is not included in the final analysis (Ronquist et al., 2012). The integration of multiple algorithms within a single tool ensures consistency by eliminating biases introduced by differences in software ideology, optimization strategies, and parameter handling.

Algorithm Analysis Methodology

Analysis Process

The research goal is to develop a framework for selecting the best initial tree generation algorithm for surveying evolutionary relationships within a dataset. Given the inherent inconsistencies of using real-world experimental data as a control, the analysis relied exclusively on simulated data generated via the R package Phylosim (Sipos et al., 2011). PhyloSim is supplied with a phylogenetic tree structure in the form of a newick file which represents the evolutionary relationships, final network topology, and branch lengths that will be simulated as a reference. With the simulated data as a reference, the variants themselves are exported into MEGA11 to generate a phylogenetic tree with one of the aforementioned analysis algorithms. The tree is then compared against the reference tree using the online tool, Visual TreeCmp, which generates comparison metrics (Goluch et al., 2020).

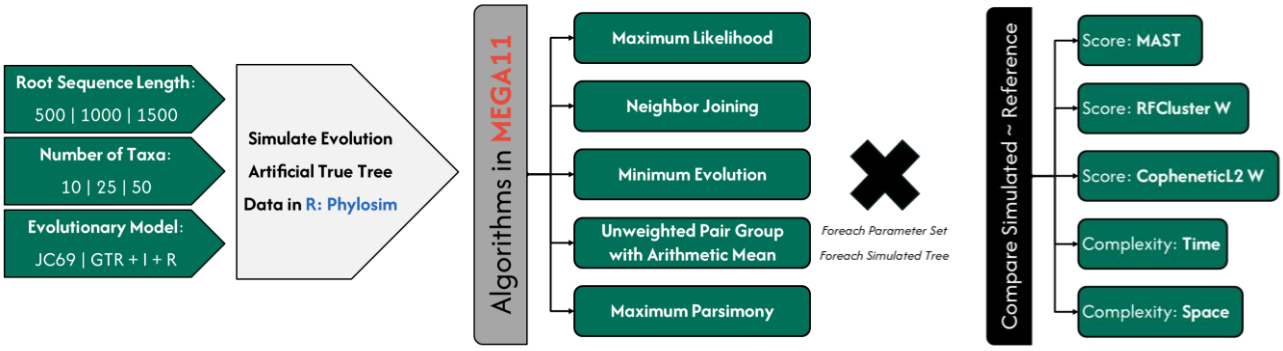


Figure 4: Algorithm Comparison Pipeline

Reference Tree Generation

Reference Tree simulation offers flexibility in evolutionary divergence with control of three core parameters: the generative model for evolutionary processes, the number of taxa or comparative entities, and the reference sequence that acts as the root and therefore ancestral sequence (Sipos & Massingham, 2023). The root sequence undergoes simulated

evolutionary divergence with the use of a generative model. Model selection impacts the mutations or evolutionary divergence of the emulated nodes. JC69 is a simple model assuming a uniform substation probability of 25% across all nucleotides, with no rate variation across sites and thus satisfying the molecular clock hypothesis (Jukes & Cantor, 1969). GTR + I + R is a sophisticated model incorporating generalized, but unequal, substitution rates (GTR), invariant sites (I), and rate heterogeneity among sites (R), making it more realistic in capturing complex speciation events and evolutionary dynamics. While there are additional models with differing evolutionary dynamics such as HKY85 and TN93, they were not utilized. Three taxa quantities, $T_n = \{10, 25, 50\}$, and three sequence lengths, $S_L = \{500, 1000, 1500\}$ are used and modeled from contemporary literature. The reference parameters are modulated independently to change the reference tree and therefore the evolutionary context the simulated phylogenetic tree is trying to recreate. Across 2 models, 3 taxa quantities, and 3 root sequence lengths, there exists 18 variants of the reference tree.

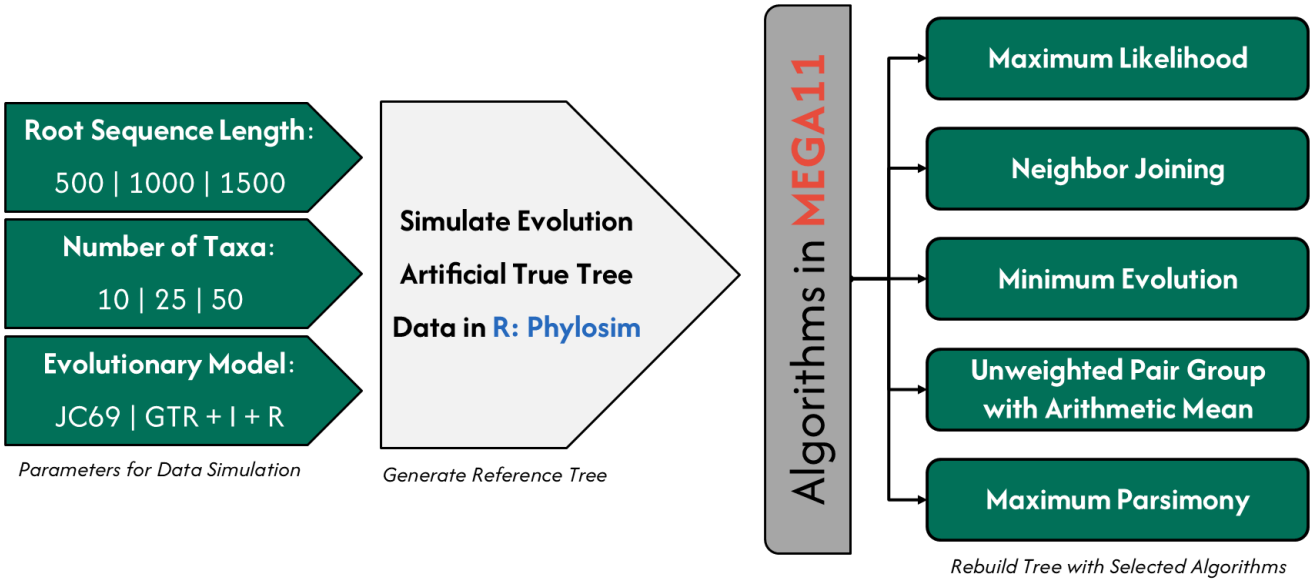


Figure 4a: Reference Tree Generation & Algorithmic Tree Comparison

Evaluation Metrics

Computational efficiency of algorithms is relatively easy to capture with runtime measurement of tree generation to capture dynamic memory used and time required but the biological assessment of the accuracy and validity of trees generated with the algorithms requires specialized metrics: Maximum Agreement Subtree (MAST) focuses on the largest subtree common to both trees. MAST excels in identifying large clusters and topological networks but not with differences spread throughout the topology. CopheneticL2 (CL2) measures differences in pairwise distance between taxa across trees, accounting for branch lengths. CL2 provides detailed information on evolutionary distances, but is overly sensitive to small variations and may be less accurate for assessing tree topology. A widely used measure is RFCluster, which measures the differences in tree topology by counting the number of differing splits between reference and generated trees. The weighted variant accounts for branch lengths, which allows the RFCluster Weighted to provide a balanced comparison of tree accuracy and validity in both topology and evolutionary divergence. Other metrics such as NodalSplitted, Triples, MatchingCluster, MatchingPair, and GeoRooted were assessed but ultimately not selected for representation of tree validity and accuracy.

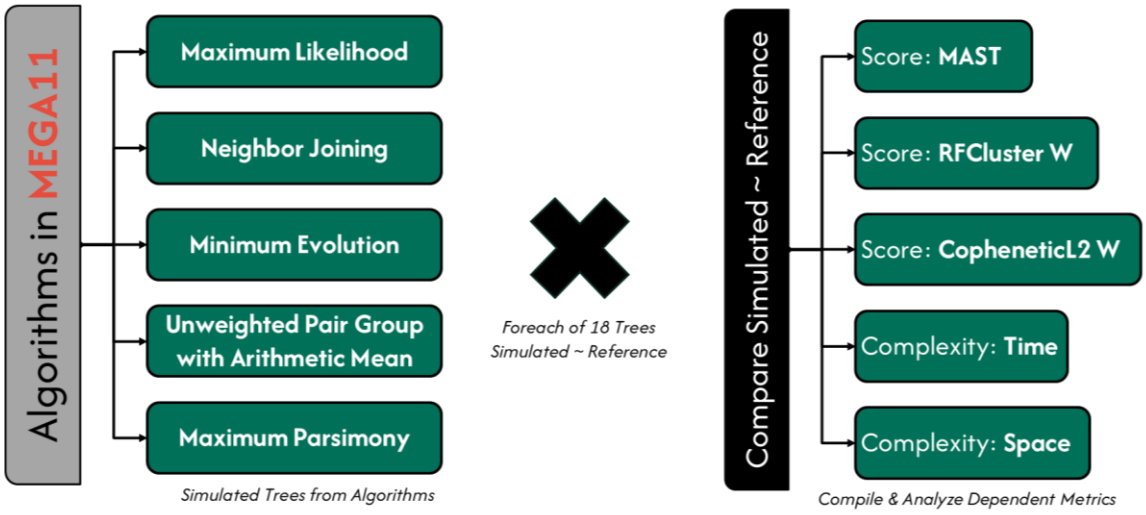


Figure 4b: Application of Evaluation Metrics to Tree Comparisons

Results

Observations & Trends

The observed results, seen in Appendix B, indicate and align with expectations that increasing sequence length and taxa number increases the accuracy of phylogenetic tree analysis. The phylogenetic tree generated by the selected algorithm, is henceforth referred to as the reconstructed tree, and the artificial representation of an evolutionary history from the reference tree. A notable anomaly was the near-indistinguishability of performance among the distance-based methods, NJ, ME, UPGMA, in both simple and moderately complex scenarios. Although literature often suggests subtle differences due to how these algorithms treat pairwise distances, the controlled simulation environment and limited parameter space may have minimized these discrepancies. A further dive into the metrics, MAST, CopheneticL2 Weighted, and RFCluster0.5 Weighted offers performance insights.

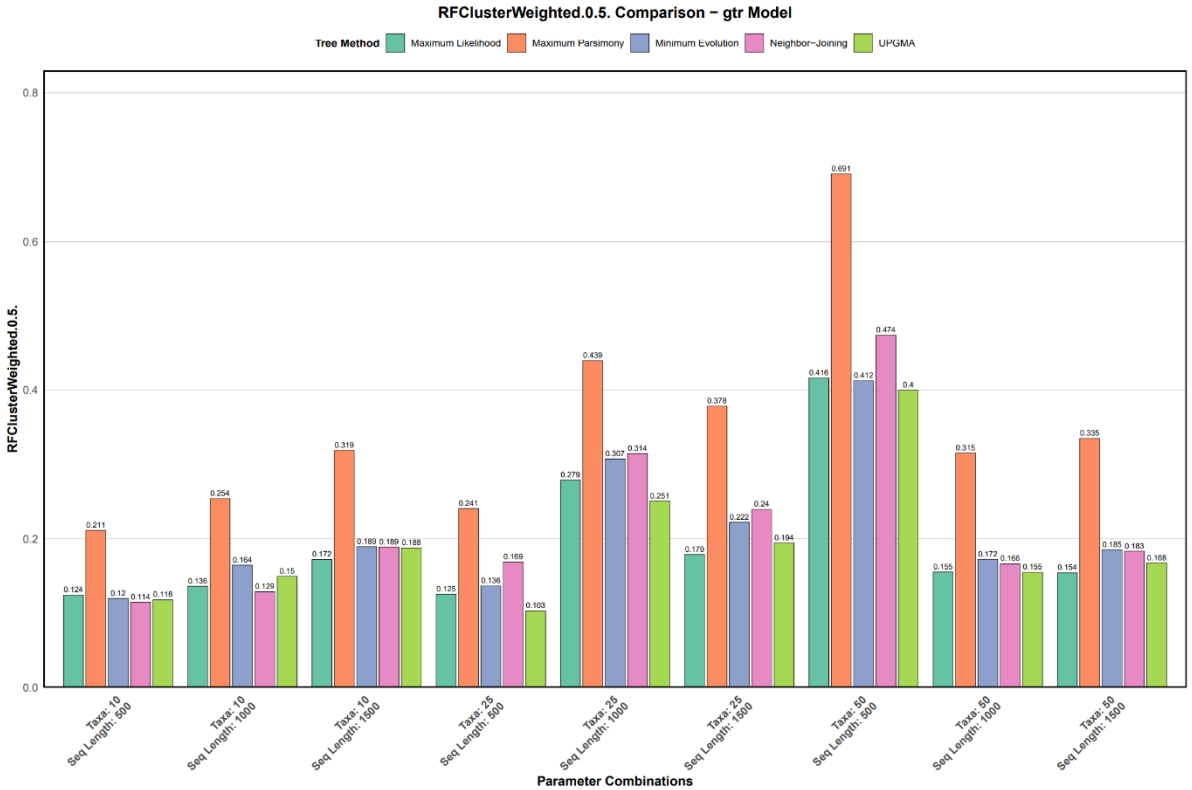


Figure 5: Representative Plot of Complex Evolution with GTR scored with RFCluster W

MAST (Maximum Agreement Subtree). MAST values improved with larger datasets, reflecting more substantial agreement on large, well-supported clades between the reconstructed and reference trees. MAST scores reinforce the idea that as sequence length and taxa number increase, the prediction of the core phylogenetic structure is more accurate. Increasing data volume via sequence length allowed ML to capture topological features more consistently. Conversely, MP displayed less improvement in MAST scores under the GTR model, reinforcing its inability to generate complex evolutionary history.

CopheneticL2 Weighted. CopheneticL2 Weighted values, which measure differences in pairwise evolutionary distances, demonstrated the sensitivity of methods to subtle variations in branch lengths. Under the simpler JC69 model, algorithms like Neighbor-Joining (NJ), Minimum Evolution (ME), and UPGMA produced relatively uniform results, suggesting that their reliance on distance matrices is well-suited to straightforward substitution patterns. However, with the GTR model, ML showed a relatively stable increase in scores as data volume increased, highlighting its ability to capture nuanced rate variations. In contrast, MP's CopheneticL2 Weighted values spiked with GTR and data complexity.

RFCluster0.5 Weighted. RFCluster Weighted scores, which assess topological differences while accounting for branch length, generally decreased with increased data volume for most algorithms, indicating stronger representation of the reference tree by the between reconstructed tree. RFCluster's balanced sensitivity to both topology and branch length underscores the advantage of model-based methods like ML under complex evolutionary simulations. Distance-based methods also showed notable improvements with more data, but their starting accuracy under GTR conditions was not as robust as ML. MP's relatively poor RFCluster scores in complex scenarios reinforce the minimization approach's poor performance with non-uniform substitution rates and sophisticated speciation.

Conclusion

Selection Framework

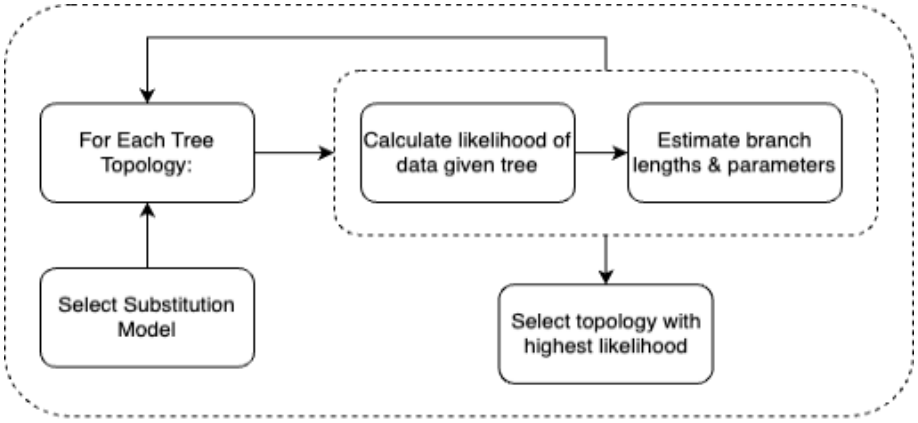
The results collectively inform a selection framework for selecting a tree-building algorithm for surveying data as based on biological and computational complexity. For simpler, clock-like substitution models where evolutionary dynamics are uniform, the choice among NJ, ME, or UPGMA is relatively inconsequential in terms of outcome quality. All distance-based methods were similarly effective, and MP did not dramatically underperform. In these scenarios, prioritizing speed and ease may lead one to pick a computationally efficient distance-based algorithm, knowing that topological and branch-length accuracy will likely be sufficient. As evolutionary scenarios become more intricate, incorporating invariant sites and heterogenous substitution rates, ML emerges as the superior choice for biologically realistic inference. The improved performance in MAST and RFCluster Weighted metrics with larger datasets clearly shows that ML can capture nuanced evolutionary patterns. However, ML's accuracy comes at a substantial computational cost; ML is exponentially time-constrained and requires more computational effort, making it less ideal for large-scale exploratory analyses. The other algorithms are relatively similar in time complexity and especially in memory usage within the scale of the selected parameter multiset.

Limitations

Future work should expand on this framework by incorporating additional substitution models, larger parameter ranges, and parallelized computational environments. Exploring Bayesian Inference or alternative scoring metrics may also yield deeper insights into algorithm performance across an even broader array of evolutionary scenarios. The framework should also be validated on experimental data; despite an inability to provide conclusive quantitative results, the results may be verifiable by literature.

Appendix A: Algorithm Details for Phylogenetic Tree Algorithms

A1. Maximum Likelihood

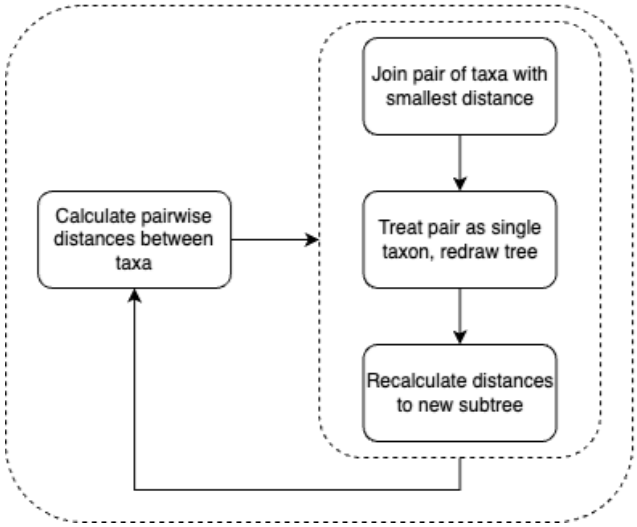


Method: ML estimates the tree topology and branch lengths that maximize the probability of observing the given data under a specific model of sequence evolution, such as GTR or HKY85 (Felsenstein, 2004; Stamatakis, 2014). It evaluates all possible trees and scores them using substitution models to predict sequence changes over time.

Complexity: The method is computationally intensive, often scaling factorially with the number of sequences (O(n!)), requiring heuristic optimization techniques for practical application.

Use: Suitable for datasets with varying mutation rates and complex evolutionary histories, providing high accuracy when computational resources are available (Minh et al., 2020).

A2. Neighbor-Joining

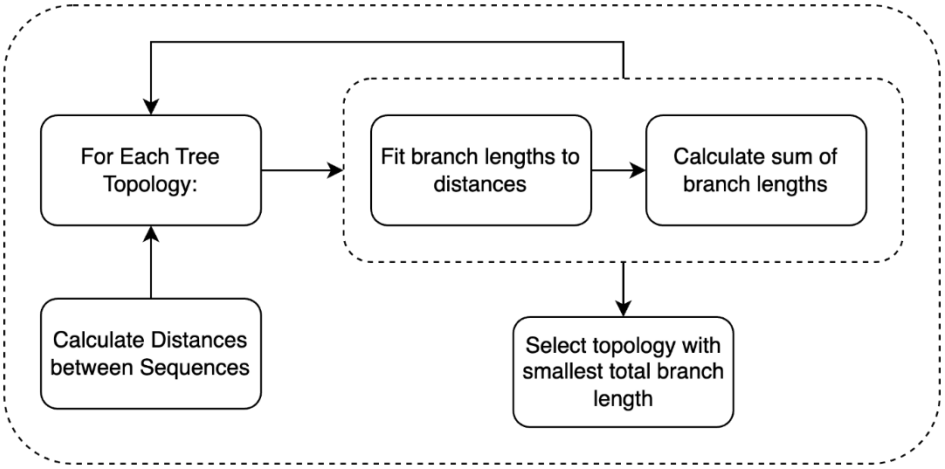


Method: NJ is a distance-based algorithm that constructs a tree by minimizing total branch lengths through an iterative process using a pairwise distance matrix (Saitou & Nei, 1987). Distances are often calculated with metrics like Jukes-Cantor or Kimura 2P models.

Complexity: Scales as $O(n^3)$, making it more efficient than likelihood-based methods while maintaining reasonable accuracy for datasets with moderate divergence.

Use: Ideal for large datasets or as a quick initial analysis when sequence divergence is not extreme (Cunningham et al., 2022).

A3. Minimum Evolution

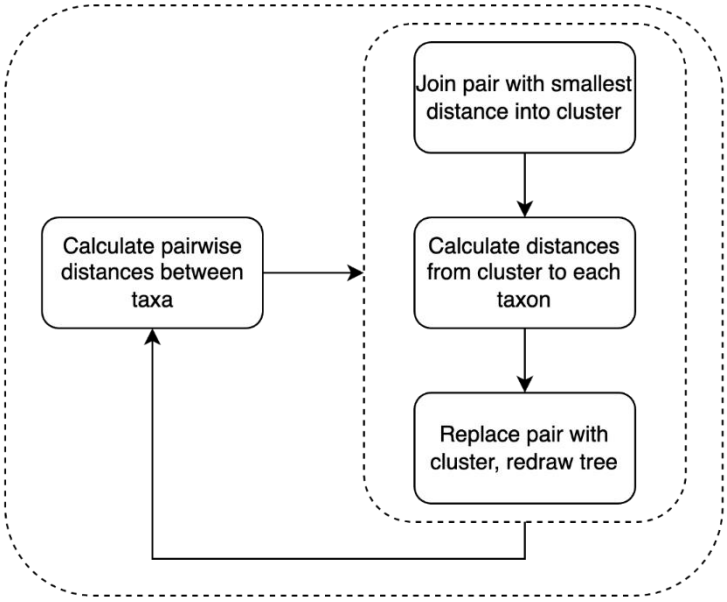


Method: ME selects tree topologies with the smallest total branch lengths, assuming that shorter trees represent more parsimonious evolutionary paths. ME relies on distance metrics and heuristic searches to identify the optimal tree (Desper & Gascuel, 2005).

Complexity: Moderate, as heuristic methods like branch-and-bound are often used to reduce computational demand.

Use: Balances computational efficiency and accuracy, making it a middle-ground option when ML is infeasible (Felsenstein, 2004).

A4. Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

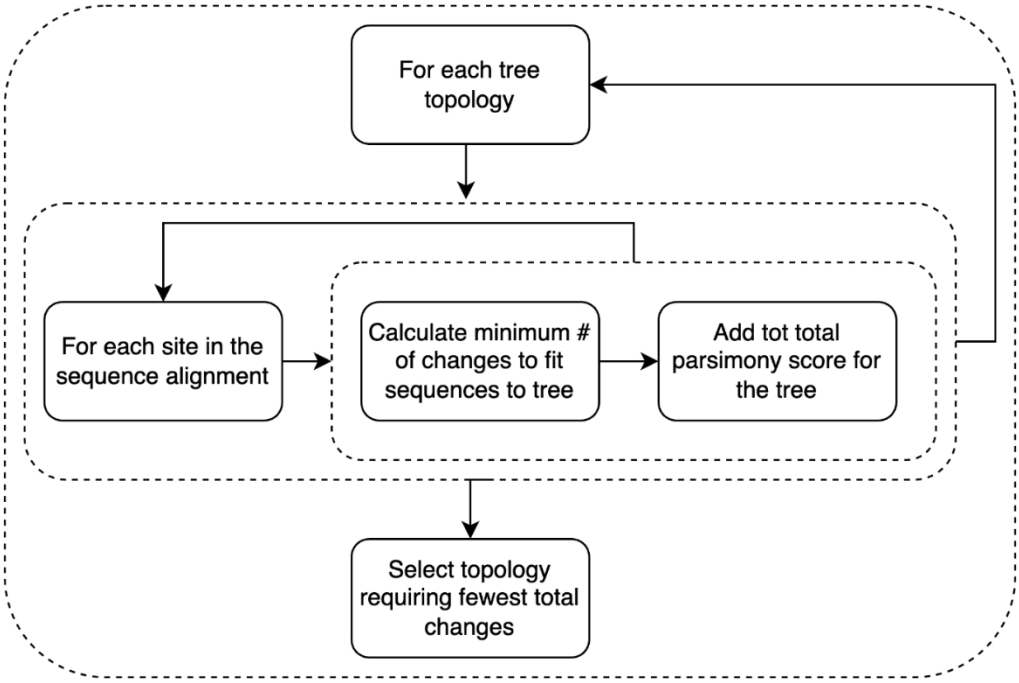


Method: UPGMA creates a tree through hierarchical clustering, assuming a molecular clock, a constant rate of evolution, and averaging distances between groups at each clustering step (Hall, 2005).

Complexity: Efficient, scaling as $O(n^2)$, due to its simple clustering mechanism.

Use: Best suited for ultrametric datasets where the molecular clock assumption holds, though it performs poorly with rate heterogeneity (Fletcher & Yang, 2009).

A5. Maximum Parsimony



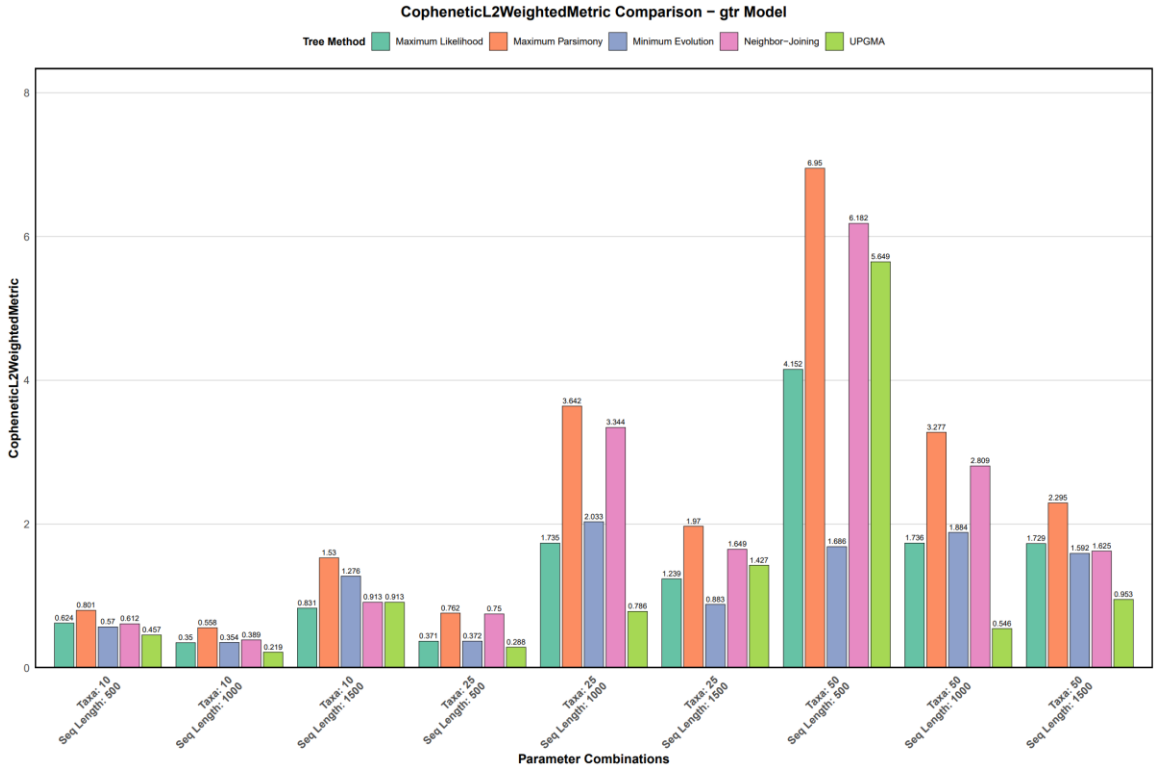
Method: MP identifies the tree requiring the fewest evolutionary changes to explain the data. It evaluates potential trees using a scoring system based on character states (Kannan & Wheeler, 2012).

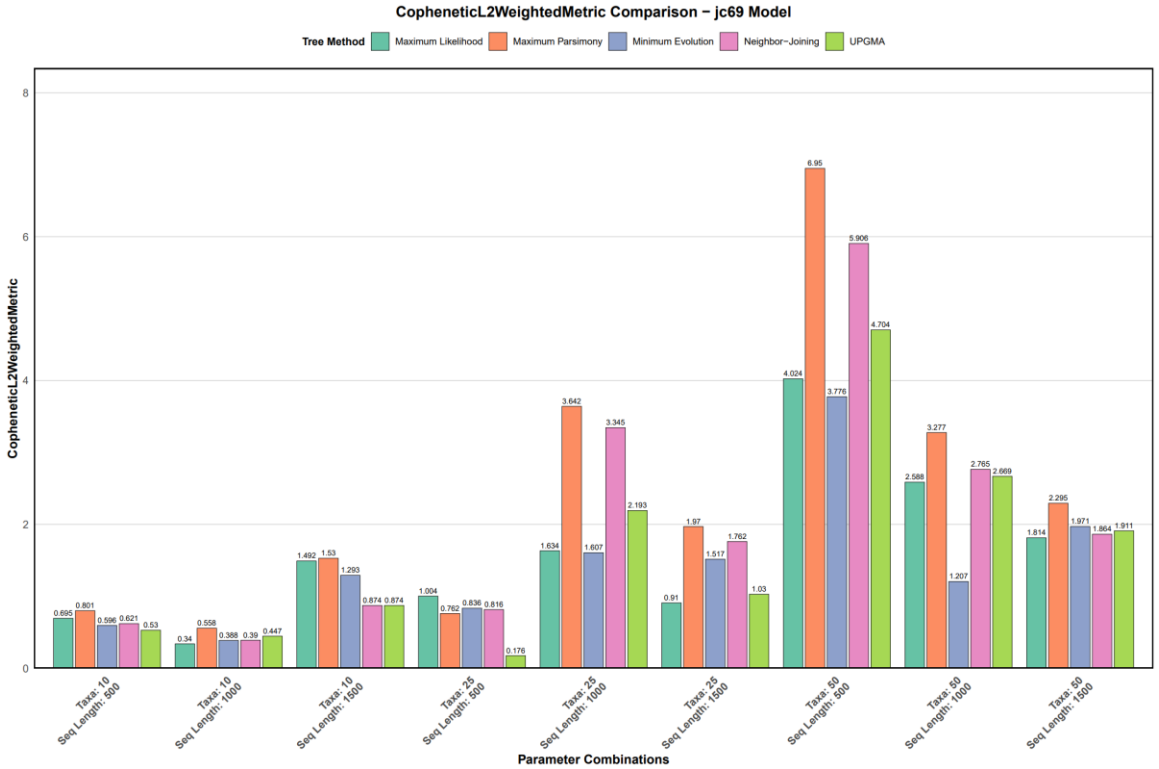
Complexity: NP-complete, making it computationally challenging for large datasets. Optimization techniques like branch-and-bound or subtree pruning are often employed.

Use: Effective for smaller datasets with low levels of homoplasy but struggles with complex evolutionary scenarios (Pyron et al., 2015).

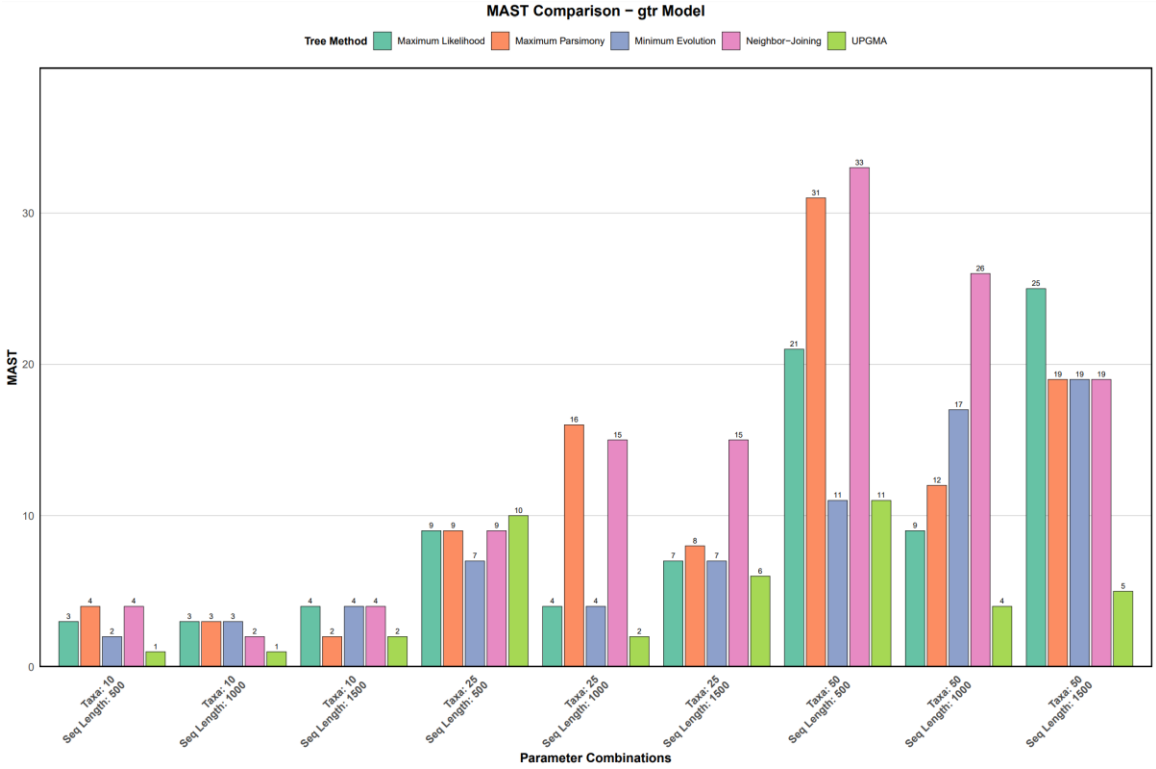
Appendix B: Selected Clustered Bar Plots of Tree Construction Algorithm Scores

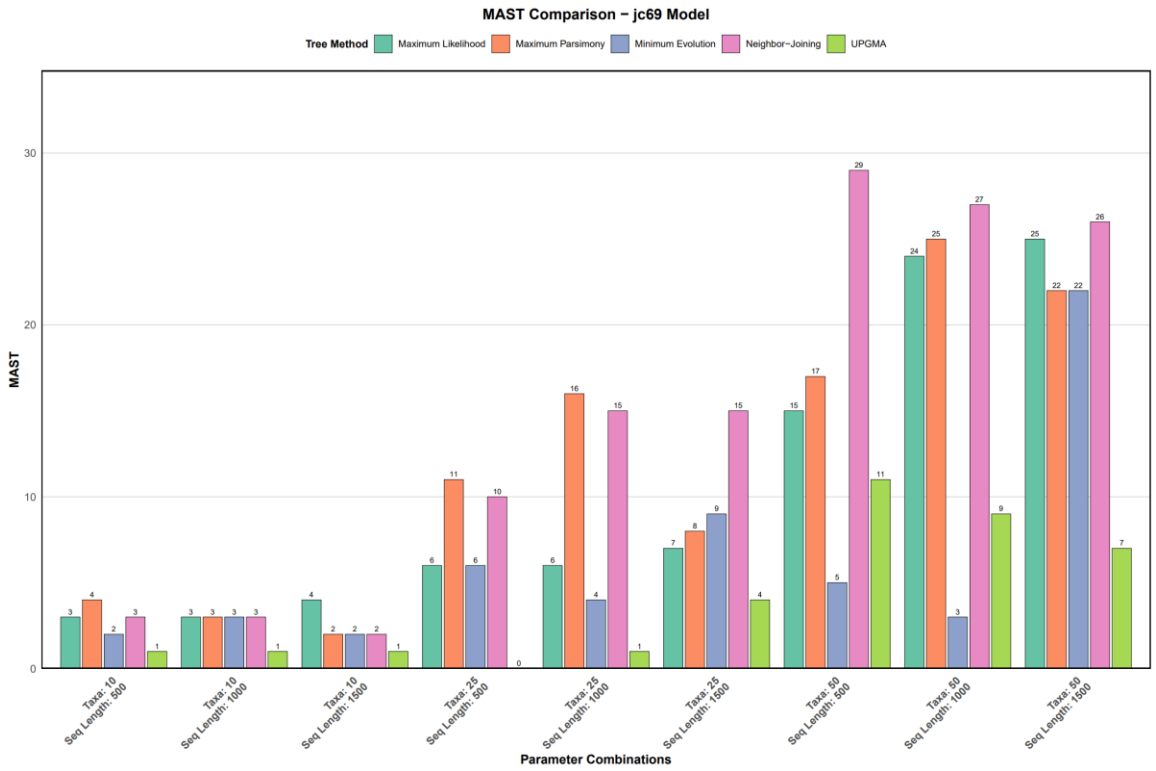
B1. CopheneticL2 Weighted Assessment



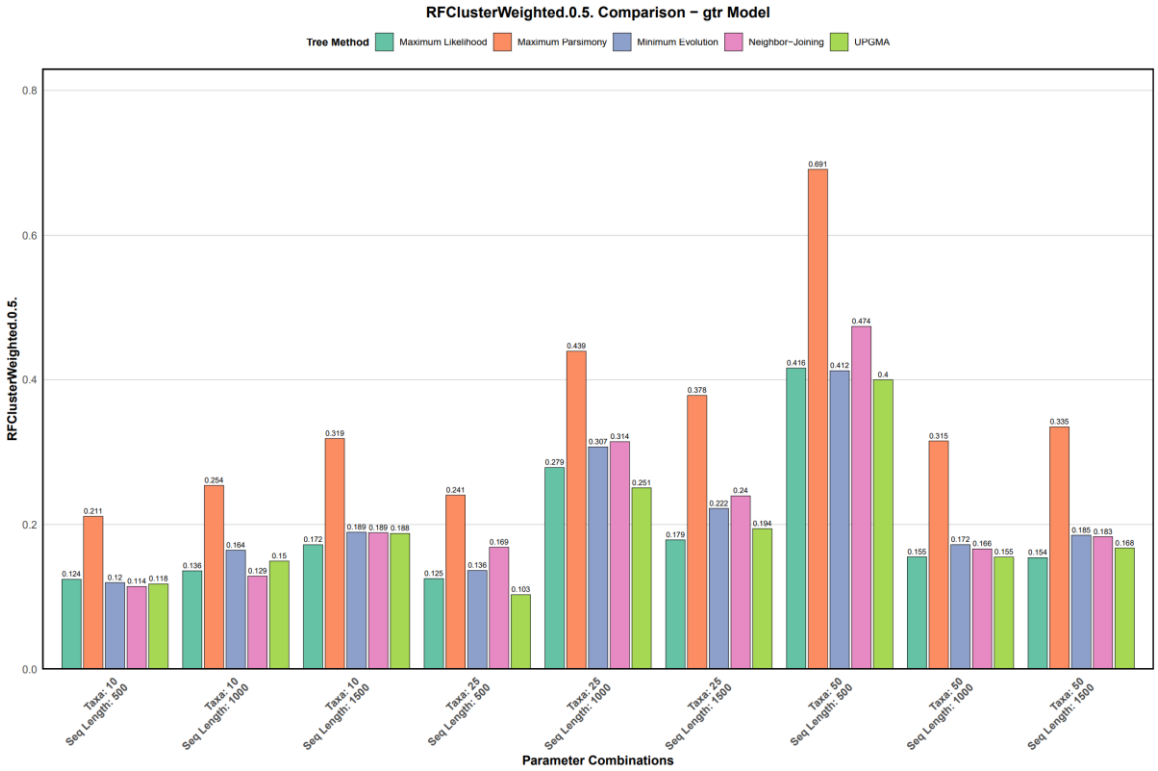


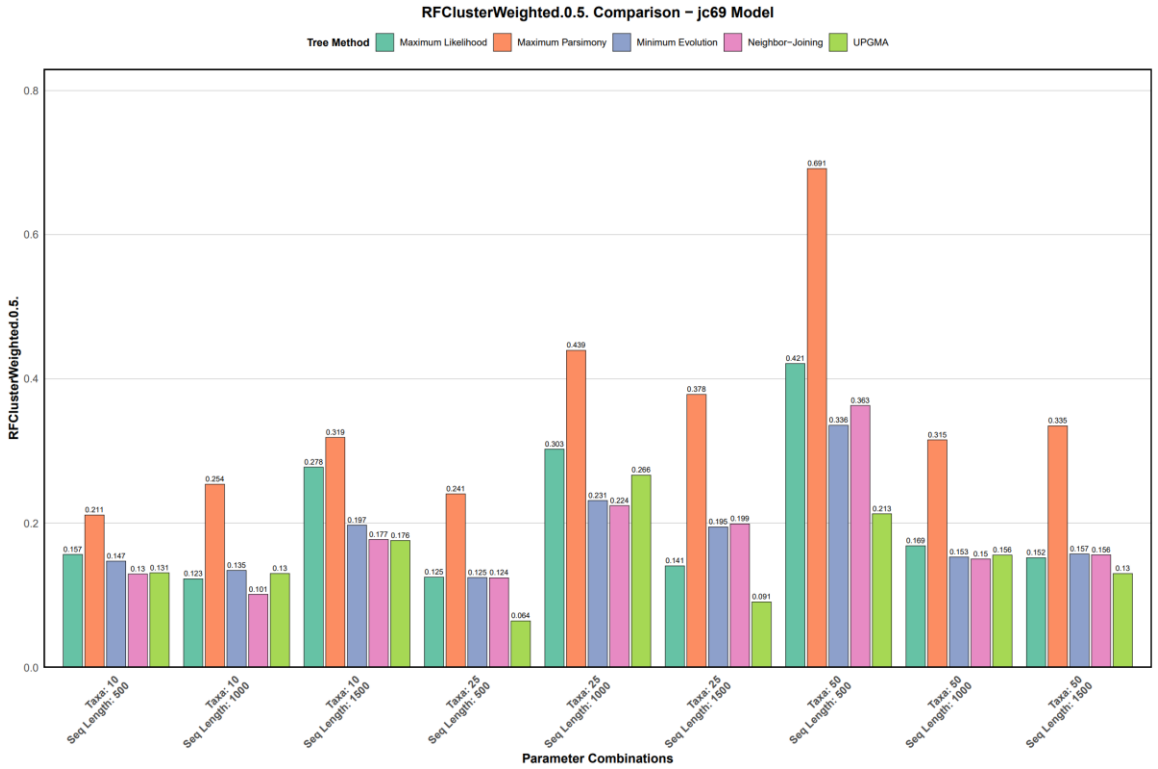
B2. MAST Assessment



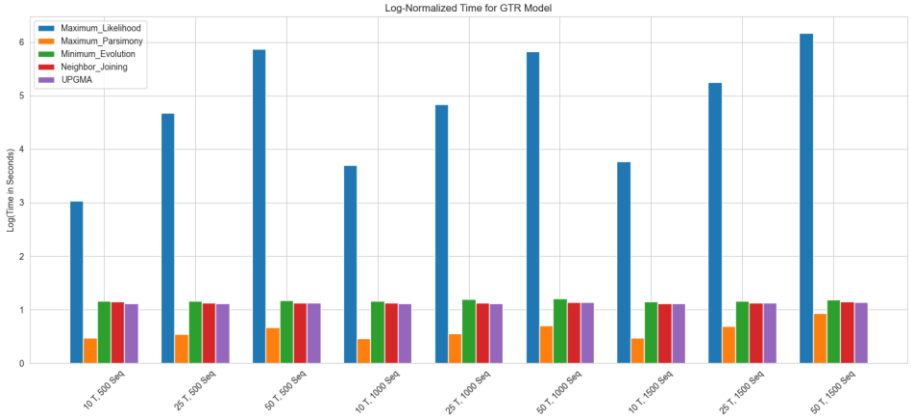


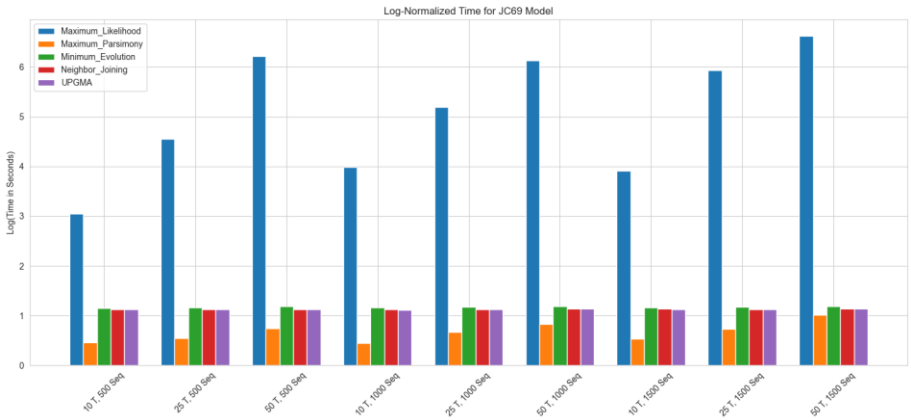
B3. RFCluster0.5 Weighted Assessment





B4. Computational Efficiency: Log(Time) & Memory





References

Allio, R., Delsuc, F., Belkhir, K., Douzery, E. J. P., Ranwez, V., & Scornavacca, C. (2024). OrthoMaM v12: A database of curated single-copy ortholog alignments and trees to study mammalian evolutionary genomics. *Nucleic Acids Research*, *52*(D1), D529–D535. https://doi.org/10.1093/nar/gkad834

Church, S. H., Mah, J. L., & Dunn, C. W. (2024). Integrating phylogenies into single-cell RNA sequencing analysis allows comparisons across species, genes, and cells. *PLOS Biology*, *22*(5), e3002633. https://doi.org/10.1371/journal.pbio.3002633

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

Desper, R., & Gascuel, O. (2005). *The Minimum Evolution Distance-Based Approach to Phylogenetic Inference*.

Dias, B. C., & Nery, M. F. (2020). Analyses of RAG1 and RAG2 genes suggest different evolutionary rates in the Cetacea lineage. *Molecular Immunology*, *117*, 131–138. https://doi.org/10.1016/j.molimm.2019.10.014

Everson, K. M., Goodman, S. M., & Olson, L. E. (2020). Speciation and gene flow in two sympatric small mammals from Madagascar, Microgale fotsifotsy and M. soricoides (Mammalia: Tenrecidae). *Molecular Ecology*, *29*(9), 1717–1729. https://doi.org/10.1111/mec.15433

Everson, K. M., Olson, L. E., & Goodman, S. M. (2020). *Data from: Speciation and gene flow in two sympatric small mammals from Madagascar, Microgale fotsifotsy and M. soricoides (Mammalia: Tenrecidae)* (Version 4, p. 145565207 bytes) [Dataset]. Dryad. https://doi.org/10.5061/DRYAD.9P8CZ8WC1

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376. https://doi.org/10.1007/BF01734359

Felsenstein, J. (with Internet Archive). (2004). *Inferring phylogenies*. Sunderland, Mass. : Sinauer Associates. http://archive.org/details/inferringphyloge0000fels

Fletcher, W., & Yang, Z. (2009). INDELible: A Flexible Simulator of Biological Sequence Evolution. *Molecular Biology and Evolution*, *26*(8), 1879–1888. https://doi.org/10.1093/molbev/msp098

Goluch, T., Bogdanowicz, D., & Giaro, K. (2020). Visual TreeCmp: Comprehensive Comparison of Phylogenetic Trees on the Web. *Methods in Ecology and Evolution*, *11*(4), 494–499. https://doi.org/10.1111/2041-210X.13358

Hall, B. G. (2005). Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Molecular Biology and Evolution*, *22*(3), 792–802. https://doi.org/10.1093/molbev/msi066

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Kannan, L., & Wheeler, W. C. (2012). Maximum Parsimony on Phylogenetic networks. *Algorithms for Molecular Biology*, *7*(1), 9. https://doi.org/10.1186/1748-7188-7-9

Ly-Trong, N., Naser-Khdour, S., Lanfear, R., & Minh, B. Q. (2022). AliSim: A Fast and Versatile Phylogenetic Sequence Simulator for the Genomic Era. *Molecular Biology and Evolution*, *39*(5), msac092. https://doi.org/10.1093/molbev/msac092

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Moody, E. R. R., Mahendrarajah, T. A., Dombrowski, N., Clark, J. W., Petitjean, C., Offre, P., Szöllősi, G. J., Spang, A., & Williams, T. A. (2022). An estimate of the deepest branches of the tree of life from ancient vertically evolving genes. *eLife*, *11*, e66695. https://doi.org/10.7554/eLife.66695

National Center for Biotechnology Information (NCBI). (1988). *National Center for Biotechnology Information (NCBI)*. https://www.ncbi.nlm.nih.gov/

Near, T. J., & Kim, D. (2021). Phylogeny and time scale of diversification in the fossil-rich sunfishes and black basses (Teleostei: Percomorpha: Centrarchidae). *Molecular Phylogenetics and Evolution*, *161*, 107156. https://doi.org/10.1016/j.ympev.2021.107156

Near, T., & Kim, D. (2021). *Data from: Phylogeny and time scale of diversification in the fossil-rich Sunfishes and Black Basses (Teleostei: Percomorpha: Centrarchidae)* (Version 2, p. 68160025 bytes) [Dataset]. Dryad. https://doi.org/10.5061/DRYAD.KPRR4XH45

Phylogeny—Taxonomy, Classification, Systematics | Britannica. (2024, November 27). https://www.britannica.com/science/phylogeny

Pinna, C., Piron-Prunier, F., & Elias, M. (2021). *Data from: Mimicry can drive convergence in structural and light transmission features of transparent wings in Lepidoptera: Alignement and phylogenetic tree of 106 Lepidoptera* (Version 3, p. 1411858 bytes) [Dataset]. Dryad. https://doi.org/10.5061/DRYAD.C2FQZ617S

Pinna, C. S., Vilbert, M., Borensztajn, S., Daney de Marcillac, W., Piron-Prunier, F., Pomerantz, A., Patel, N. H., Berthier, S., Andraud, C., Gomez, D., & Elias, M. (2021). Mimicry can drive convergence in structural and light transmission features of transparent wings in Lepidoptera. *eLife*, *10*, e69080. https://doi.org/10.7554/eLife.69080

Pyron, R. A., Hendry, C. R., Chou, V. M., Lemmon, E. M., Lemmon, A. R., & Burbrink, F. T. (2014). Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). *Molecular Phylogenetics and Evolution*, *81*, 221–231. https://doi.org/10.1016/j.ympev.2014.08.023

Pyron, R. A., Hendry, C. R., Chou, V. M., Lemmon, E. M., Lemmon, A. R., & Burbrink, F. T. (2015). *Data from: Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia)* (Version 1, p. 12283913 bytes) [Dataset]. Dryad. https://doi.org/10.5061/DRYAD.RB5NC

Rambaut, A., & Grass, N. C. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, *13*(3), 235–238. https://doi.org/10.1093/bioinformatics/13.3.235

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*, *61*(3), 539–542. https://doi.org/10.1093/sysbio/sys029

Sadasivan, H., Ross, L., Chang, C.-Y., & Attanayake, K. U. (2020). Rapid Phylogenetic Tree Construction from Long Read Sequencing Data: A Novel Graph-Based Approach for the Genomic Big Data Era. *Journal of Engineering and Technology*, *2*(1), Article 1.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sipos, B., & Massingham, T. (2023, October 12). *The Phylosim Package*. https://www.ebi.ac.uk/research/goldman/software/phylosim/

Sipos, B., Massingham, T., Jordan, G. E., & Goldman, N. (2011). PhyloSim—Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC Bioinformatics*, *12*(1), 104. https://doi.org/10.1186/1471-2105-12-104

Soltis, D. E., & Soltis, P. S. (2016a). Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. *Plant Diversity*, *38*(6), 264–270. https://doi.org/10.1016/j.pld.2016.12.001

Soltis, D. E., & Soltis, P. S. (2016b). Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. *Plant Diversity*, *38*(6), 264–270. https://doi.org/10.1016/j.pld.2016.12.001

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027. https://doi.org/10.1093/molbev/msab120

Zou, Y., Zhang, Z., Zeng, Y., Hu, H., Hao, Y., Huang, S., & Li, B. (2024). Common Methods for Phylogenetic Tree Construction and Their Implementation in R. *Bioengineering*, *11*(5), 480. https://doi.org/10.3390/bioengineering11050480