The Impact of Multiple Sequence Alignment Error on Phylogenetic Estimation under Variable-Across-Phylogeny Substitution Models

Rei Doko[1] and Kevin J. Liu[1,2,3*]

[1] *Department of Computer Science and Engineering*
[2] *Ecology, Evolution, and Behavior Program*
[3] *Genetics and Genome Sciences*
Michigan State University, East Lansing, MI, USA
**Email:* kjl@msu.edu

Abstract

In many different species, it has been observed that nucleotide compositions are not identical on the genic and even genomic scale. This observation contradicts a commonly held assumption in most maximum likelihood based phylogenetic estimation methods - that the process governing DNA evolution is identical across lineages. We show that when DNA evolution is nonhomogeneous, topological estimation and continuous parameter estimation are impacted both by alignment quality and model misspecification due to the homogeneity-across-lineages assumption.

1 Introduction

Nucleotide composition biases can be found in the genomes of a variety of organisms, such as grasses [**?**], insects [**?**], and birds [**?**]. Knowing when and where these compositional biases arise in the evolutionary history of these organisms is of interest since G+C bias is hypothesized to have significance in biological processes. Computational methods can be applied to more widely available genomic data to provide a better idea of this history. Molecular phylogenetics is used to reconstruct evolutionary relationships between organisms using biomolecular sequence data such as DNA.

Maximum likelihood based phylogenetic estimation uses a stochastic model of sequence evolution to evaluate the probability of a tree topology given the observed sequence data. A common simplifying assumption is that the base composition and relative substitution rates are identical throughout the tree. However, the observation that nucleotide composition can vary across lineages demonstrates that these assumptions do not hold in biological data. Nonhomogeneous substitution models relax this assumption and allow for base composi-

tion and relative substitution rates to vary across the phylogeny, and have been implemented before in nhPhyML [?] and PAML [?].

Alignment error has been shown to impact downstream phylogenetic inference and estimation [?]. However, how alignment quality impacts estimation when sequence evolution is nonhomogeneous is not well studied. While it is likely that alignment quality will have an impact on estimation in this more complicated model, the question of what does this mean for empirical data remains. For example, how are estimates of base composition and substitution rates when using nonhomogeneous substitution models for maximum likelihood estimation? Furthermore, do more sophisticated models of DNA evolution accounting for nonhomogeneity improve estimates?

2 Materials and Methods

The objective of this study is to characterize the effect of alignment quality and model misspecification in the problem of phylogenetic estimation when evolution is nonhomogeneous and nonstationary.

Data Availability Statement Data and scripts used are available at `https://gitlab.msu.edu/liulab/nonhomogeneous-substitution-model-study-data-scripts`.

2.1 Methods for MSA and phylogenetic estimation

Preliminaries. Let $T = (V, E)$ be a rooted tree with labeled leaves $X \subset V$ and root $\rho \in V$. Each edge $e = (u, v) \in E$ where $u, v \in V$ has a length $d(e)$. An edge (u, v) is a leaf edge if either u or v is a leaf, otherwise it is an internal edge. Deleting an edge e from a tree T gives two subtrees $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$. The vertex sets V_1 and V_2 are disjoint, and $V_1 \cup V_2 = V$. The same can be said for their respective leaf sets, so $\{X_1, X_2\}$ is a bipartition of X. Let this be denoted as $b(e) = \{X_1, X_2\}$.

Multiple sequence alignment. There are a variety of multiple sequence alignment methods available. For this study, we selected a range of commonly used methods. We aligned simulated and empirical datasets using MAFFT [?] version 7.475, MUSCLE [?] version 5.0.1428, Clustal Omega [?] version 1.2.4, Clustal W [?] version 2.1, and FSA version [?] 1.15.9. Each method was run using their respective default settings.

Phylogenetic estimation. We use the general time reversible (GTR) model for phylogenetic estimation. The GTR model specifies that there are separate base frequencies $\pi_T, \pi_C, \pi_A, \pi_G$ which sum to 1, as well as rate parameters a, b, c, d, e, f. We use the same conventions as used by **?**. That is to say, a corresponds to $T \leftrightarrow C$, b corresponds to $T \leftrightarrow A$, c to $T \leftrightarrow G$, d to $C \leftrightarrow A$, e

to $C \leftrightarrow G$, and f to $G \leftrightarrow A$. f is fixed to 1 and the remaining rate parameters are relative to f. The rate matrix Q is as follows:

$$Q = \begin{bmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_c & \cdot & f\pi_G \\ c\pi_T & e\pi_c & f\pi_A & \cdot \end{bmatrix}$$

With the diagonals set to $Q_{ii} = -\sum_{i \neq j} Q_{ij}$. The transition probability matrix is given by $P(t) = \exp(-Qt)$ and is used to calculate likelihoods for a phylogenetic tree. Typically in phylogenetic estimation using Markov models of substitution, the rate matrix is assumed to be constant over the whole tree. We refer to models under this assumption as homogeneous, or having no shifts.

However, nucleotide composition biases have been observed in biological data. To account for rate and composition differences across lineages, each edge e has an associated set of parameters $\theta(e)$ that define the rate matrix for that edge. We use a GTR model for the branch models. The traditional homogeneous model is the case where $\theta(e)$ is fixed, i.e. $\theta(e_i) = \theta(e_j)$ for all $e_i, e_j \in E$. For heterogeneous models, we considered two different classes: which we refer to as single-shift and all-shift. For all-shift, $\theta(e)$ is independent for each edge. For single-shift, there are exactly two sets of parameters, θ_{shift} and $\theta_{\text{background}}$ and some restrictions on which edges they apply to. There is a shift edge, $e_{\text{shift}} \in E$, and all edges descending from it all have $\theta(e) = \theta_{\text{shift}}$. Any remaining edges are $\theta_{\text{background}}$. We will also refer to homogeneous models as no-shift interchangeably. In nonhomogeneous models, rooting can impact likelihood values since these models are not time-reversible, so rooted trees are used.

RAxML [**?**] version 8.2.12 was used to perform maximum likelihood estimation under a homogeneous GTR model. PAML [**?**] version 4.9j was used to perform maximum likelihood estimation under fixed tree topologies using a branch model. Since PAML does not support tree search under nonhomogeneous models, we wrote a wrapper script to perform local tree topology search, using PAML to evaluate the log likelihood of a topology.

Single-shift search. While a fully nonhomogeneous branch model can account for very general nucleotide substitution processes, it is highly parameterized. Also, it approaches the no common mechanism model, which is known to be statistically inconsistent [**?**]. In single-shift, there are exactly two separate sets of substitution model parameters estimated, and instead is search for a placement of the different rate matrix on the phylogeny. We use a brute-force approach of local search to determine which assignment of the shift model maximizes likelihood.

Performance assessments. We use Robinson-Foulds distance [**?**] to assess topological difference between trees. Let $S(T) = \{b(e) | e \in \mathring{E}, e \text{ is an internal edge}\}$. The Robinson-Foulds distance between two unrooted trees T and T' is the symmetric difference of $S(T)$ and $S(T')$. A way to extend RF distance to rooted

trees is to consider the bipartition representation for labeled nodes (i.e. $X \cup \{\rho\}$). So for the two subtrees T_1, T_2 induced by deleting an edge $e \in E$, if $\rho \in V_1$ then the edge representation becomes $b'(e) = \{X_1 \cup \{\rho\}, X_2\}$ and vice versa if $\rho \in V_2$. For identifying root placement, we say two trees T and T' have identical roots, ρ and ρ' respectively, if the leaf sets of the subtrees induced by deleting the respective root nodes are identical.

We take the L1 norm of the relative errors for substitution model parameters to assess model parameter estimation performance in the simulation study. For the base frequencies, this would be $\sum_{i \in ACGT} \left| \frac{\pi_i - \hat{\pi}_i}{\pi_i} \right|$.

To assess how well the shift subtree is being predicted in the single-shift model, we use the size of the maximum agreement subtree (MAST) between the true and estimated shift subtrees. The MAST problem is to find a subtree given a set of trees $\mathscr{T}$ with the largest subset of leaves that also agrees with all the trees in $\mathscr{T}$.

For evaluating alignment quality, we use sum-of-pairs false positive and false negative rates, denoted SP-FP and SP-FN respectively. SP-FP is calculated as the proportion of homologies in the estimated alignment and not in the true alignment. SP-FN is the same, but the other way around.

2.2 Simulation study

Model tree generation. Model trees were sampled using INDELible [**?**] under a birth-death process. Non-ultrametricity was introduced using a procedure described in **?** with deviation factor $c = 2$.

1. Generate a rooted model tree using INDELible with the default settings
2. For every branch:
   2.1 Choose $x \sim U(-\ln(2), \ln(2))$
   2.2 Scale the branch length by $\exp(x)$
3. Let L be the maximum root-to-tip distance for the tree and H be the desired height.
4. Scale each branch length by H/L
5. Select a subtree to evolve under the shifted substitution model containing as close to half of the leaves.

Table 1: GTR model parameters used for evolving sequences in the simulation study.

Parameter	T	C	A	G	C↔T	A↔T	G↔T	A↔C	C↔G	A↔G
Shift	0.216	0.237	0.317	0.230	5.847	3.186	1.214	3.437	1.307	1.0
Background	0.183	0.226	0.058	0.534	1.505	0.367	0.141	0.412	0.094	1.0

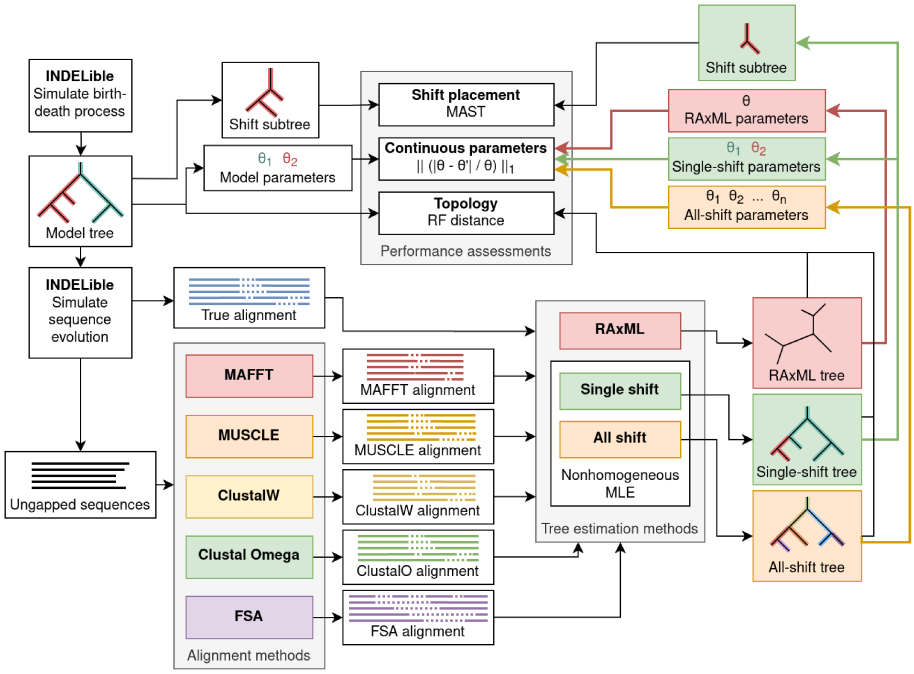


Figure 1: Flowchart of simulation study steps. A model tree is first generated under a birth-death process using INDELible, and then sequences are generated under that model tree with a nonhomogeneous model. The ungapped sequences are aligned using MAFFT, MUSCLE, ClustalW, Clustal Omega, and FSA. Every estimated alignment, as well as the true alignment, is then used to perform tree estimation. Tree estimation is done using three different classes of model: 0-shift, or homogeneous, is done using RAxML. Single-shift and all-shift, both nonhomogeneous across lineages, are performed using PAML to calculate likelihood scores and continuous parameter optimization. Finally, the resulting trees from each alignment and tree estimation method pair is compared against the model tree. For the single shift model, shift placement is evaluated by computing the size of the maximum agreement subtree (MAST) for the true and estimated shift subtrees.

Simulating sequence evolution. Model conditions were the same as those used in **?** to include a range of sequence divergence in the simulation study. INDELible was used to generate sequences under a GTR-based branch model using the phylogenies generated as described earlier. **?** found GC content variation in the avian phylogeny. The GTR model parameters were empirically estimated using single-copy orthologs from **?** for the subset of species (*Calypte anna*, *Alligator mississippiensis*, *Melopsittacus undulatus*, *Corvus brachyrhynchos*, and *Manacus vitellinus*) included in **?**'s study. To estimate these parameters, we aligned the single-copy orthologs using MAFFT with the default settings. Then, we used a single-shift model to estimate parameters on each individual aligned sequence. Then, we looked at the two sets of substitution rates estimated, and observed that the ratio between them was bimodal. The first peak ranged from a 1- to 10-fold difference, and the second ranged from a 10000- to 100000-fold difference. We chose GTR model parameters based on the estimated parameters in the first mode.

2.3 Empirical study

Grass dataset. The distribution of GC content in monocots is bimodal [**?**], which is not the case for other plants. This pattern is notably strong in rice. We applied nonhomogeneous substitution model based phylogenetic tree estimation to a set of 8 taxa: Oryza sativa japonica [**?**], Sorghum bicolor [**?**], Carex cristatella, Carex scoparia, Juncus effusus, Juncus inflexus [**?**], Ananas comosus [**?**], and Musa balbisiana [**?**]. We identified 1900 single-copy orthologs using orthofinder [**?**] with the default settings. We aligned the sequences individually using MAFFT, MUSCLE, Clustal Omega, Clustal W, and FSA with the default settings. We performed phylogenetic estimation under a single-shift model for every individual gene. We also concatenated the aligned gene sequences and ran the same analysis.

3 Results

3.1 Simulation study

Impact of alignment accuracy. In both topology estimation and substitution model parameter estimation, the true alignments perform the best, as would be expected. Across all methods and levels of sequence divergence, topological inference using estimated alignments yields significantly more error. In figure **??**, we see a correlation between topological error and alignment accuracy in more divergent model conditions. We also see this trend is maintained in the 20-taxa model conditions.

Impact of model misspecification. Nonstationary nucleotide composition and nonhomogeneous substitution rates can reflect an evolutionary adaptation.

Table 2: Model conditions and summary statistics for ground truth and estimated alignments.

Model condition	# Taxa	Tree height	Indel probability	Length	ANHD	Gappiness	Alignment	SP-FP	SP-FN	Estimated length	Estimated NHD	Estimated gappiness
10.A	10	0.47	0.13	2123.8	0.306	0.528	MAFFT	0.572	0.512	1478.6	0.389	0.326
							MUSCLE	0.579	0.508	1518.5	0.413	0.342
							CLUSTALW	0.746	0.683	1191.4	0.466	0.165
							CLUSTALO	0.734	0.682	1247.1	0.484	0.202
							FSA	0.217	0.645	3609.3	0.280	0.715
10.B	10	0.7	0.1	2315.8	0.364	0.564	MAFFT	0.683	0.629	1477.1	0.435	0.321
							MUSCLE	0.667	0.602	1570.1	0.449	0.361
							CLUSTALW	0.786	0.724	1186.0	0.496	0.155
							CLUSTALO	0.781	0.732	1248.7	0.504	0.198
							FSA	0.236	0.603	4471.2	0.319	0.770
10.C	10	1.2	0.06	2313.2	0.465	0.566	MAFFT	0.752	0.711	1484.5	0.484	0.328
							MUSCLE	0.729	0.679	1573.1	0.499	0.364
							CLUSTALW	0.822	0.768	1170.6	0.537	0.148
							CLUSTALO	0.823	0.780	1237.6	0.541	0.194
							FSA	0.272	0.783	4992.4	0.377	0.795
10.D	10	2	0.031	2202.8	0.553	0.538	MAFFT	0.828	0.807	1461.8	0.528	0.310
							MUSCLE	0.794	0.766	1561.9	0.542	0.353
							CLUSTALW	0.865	0.830	1143.0	0.573	0.119
							CLUSTALO	0.858	0.831	1222.9	0.565	0.177
							FSA	0.384	0.864	5729.6	0.420	0.820
10.E	10	4.4	0.013	2063.0	0.649	0.510	MAFFT	0.879	0.871	1529.2	0.569	0.342
							MUSCLE	0.846	0.831	1590.5	0.589	0.366
							CLUSTALW	0.897	0.872	1146.5	0.614	0.124
							CLUSTALO	0.884	0.865	1234.0	0.602	0.187
							FSA	0.497	0.912	6196.2	0.477	0.836
20.A	20	0.47	0.13	2410.3	0.301	0.581	MAFFT	0.420	0.388	1643.7	0.368	0.390
							MUSCLE	0.404	0.351	1790.2	0.380	0.439
							CLUSTALW	0.654	0.612	1278.6	0.449	0.216
							CLUSTALO	0.654	0.625	1306.3	0.471	0.233
							FSA	0.129	0.536	4394.5	0.289	0.763
20.B	20	0.7	0.1	2585.1	0.374	0.606	MAFFT	0.554	0.526	1670.7	0.433	0.394
							MUSCLE	0.521	0.470	1863.0	0.445	0.456
							CLUSTALW	0.753	0.713	1261.2	0.509	0.198
							CLUSTALO	0.738	0.708	1318.2	0.518	0.232
							FSA	0.144	0.681	5959.0	0.349	0.819
20.C	20	1.2	0.06	2895.8	0.484	0.649	MAFFT	0.771	0.752	1683.0	0.516	0.400
							MUSCLE	0.720	0.684	1907.7	0.530	0.470
							CLUSTALW	0.856	0.822	1227.8	0.573	0.178
							CLUSTALO	0.848	0.821	1303.2	0.570	0.226
							FSA	0.215	0.838	8231.8	0.429	0.874
20.D	20	2	0.031	2696.2	0.581	0.622	MAFFT	0.863	0.860	1691.4	0.569	0.403
							MUSCLE	0.815	0.800	1890.0	0.585	0.466
							CLUSTALW	0.897	0.876	1181.6	0.615	0.148
							CLUSTALO	0.890	0.875	1287.4	0.602	0.218
							FSA	0.328	0.908	10177.9	0.486	0.899
20.E	20	4.4	0.013	2723.6	0.667	0.629	MAFFT	0.939	0.940	1804.3	0.608	0.444
							MUSCLE	0.904	0.898	1979.5	0.630	0.493
							CLUSTALW	0.946	0.933	1162.8	0.652	0.140
							CLUSTALO	0.938	0.928	1283.8	0.637	0.221
							FSA	0.432	0.942	11566.5	0.525	0.913

Estimates for base frequencies and substitution rates can be used to characterize such adaptations. For substitution model parameter estimation, using the single-shift model yields the closest parameter estimates across all alignment types. In figure **??** we can see that for topological estimation, depending on the alignment and level of divergence, the single-shift model performs as well as, but usually better than the underspecified model used with RAxML. The overspecified model performs about as well as well as the homogeneous model until the 20E model condition.

3.2 Empirical study

Grass dataset For the concatenated analysis, the estimated topology was identical for all alignment types except ClustalW. Furthermore, the placement of the shift was identical for all alignment types except ClustalW. The estimated

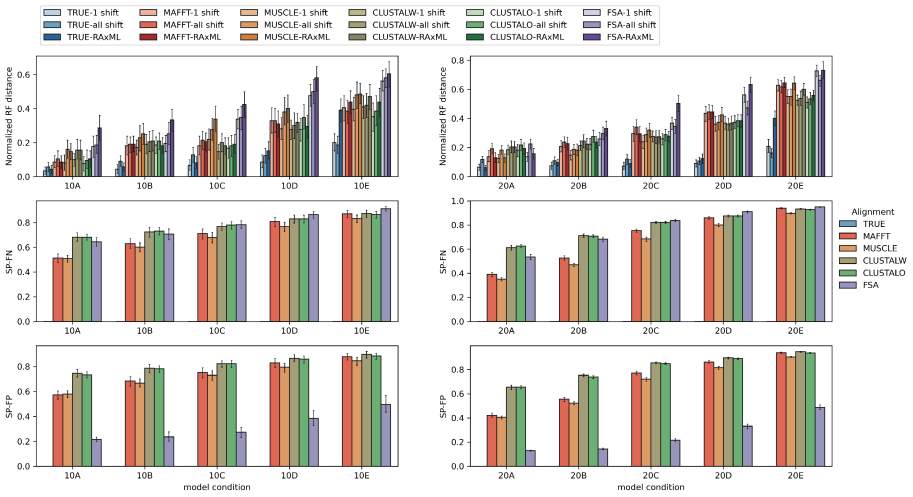


Figure 2: Topological error for every combination of alignment method and maximum likelihood tree estimation method. Topological error is measured with the normalized Robinson-Foulds distance of the model tree and the tree estimated using a single-shift nonhomogeneous substitution model. Alignment SP-FN and SP-FP are also reported.

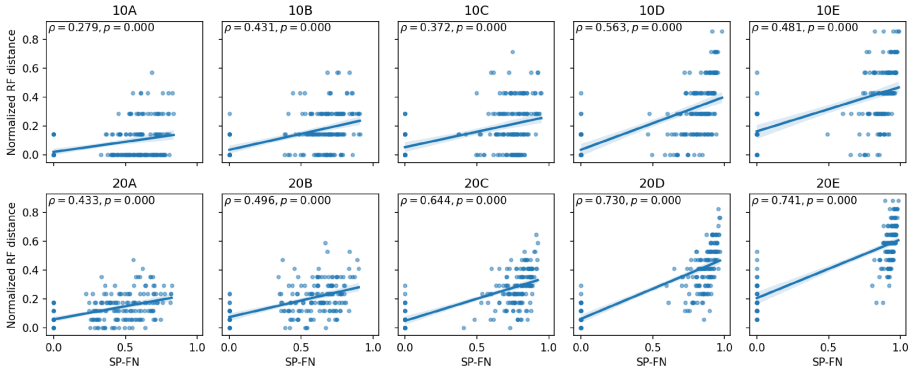


Figure 3: Alignment error (sum-of-pairs false negative rate) vs topological error (normalized Robinson-Foulds distance). Results are aggregated for all alignment types.

tree using a no homogeneous model with the concatenated MAFFT alignment identifies Ananas comosus as more closely related to Oryza sativa than Sorghum bicolor, which would be a very unconvential result. The tree estimated using a homogeneous model on the MAFFT alignment does not make this placement, and is in consensus with Clustal Omega, ClustalW, and Muscle estimated tree topologies.

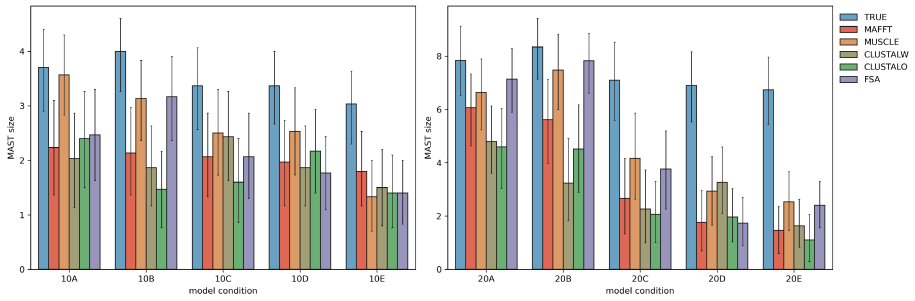


Figure 4: Size of the MAST of the predicted shift subtree and the true subtree.

Model condition	MLE method	Correct root rate					
		TRUE	MAFFT	MUSCLE	CLUSTALW	CLUSTALO	FSA
10A	single shift	56.7%	23.3%	43.3%	26.7%	26.7%	23.3%
	all shift	23.3%	20.0%	20.0%	23.3%	6.7%	20.0%
10B	single shift	60.0%	16.7%	23.3%	13.3%	6.7%	20.0%
	all shift	30.0%	6.7%	10.0%	3.3%	6.7%	0.0%
10C	single shift	60.0%	16.7%	16.7%	10.0%	10.0%	16.7%
	all shift	30.0%	13.3%	16.7%	3.3%	0.0%	6.7%
10D	single shift	40.0%	20.0%	26.7%	6.7%	16.7%	6.7%
	all shift	23.3%	3.3%	16.7%	0.0%	6.7%	3.3%
10E	single shift	30.0%	10.0%	10.0%	6.7%	10.0%	3.3%
	all shift	6.7%	3.3%	3.3%	3.3%	6.7%	0.0%
20A	single shift	33.3%	20.0%	30.0%	13.3%	16.7%	36.7%
	all shift	31.0%	6.9%	7.1%	6.7%	16.7%	17.2%
20B	single shift	37.9%	17.2%	20.7%	6.9%	10.3%	24.1%
	all shift	24.1%	13.8%	10.3%	0.0%	7.1%	10.3%
20C	single shift	46.7%	20.0%	13.3%	13.3%	3.3%	0.0%
	all shift	13.3%	6.7%	0.0%	6.7%	0.0%	3.3%
20D	single shift	50.0%	13.3%	6.7%	13.3%	6.7%	3.3%
	all shift	10.3%	3.3%	0.0%	3.3%	0.0%	0.0%
20E	single shift	46.7%	16.7%	3.3%	6.7%	6.7%	0.0%
	all shift	6.7%	0.0%	3.3%	3.3%	0.0%	0.0%

Table 3: Proportion of correct root placements by alignment and model condition.

For the per-gene analyses, figure **??** shows that on average, the topologies estimated from different alignment methods have some disagreement. For continuous parameter estimates table **??** shows that for all alignments, estimated shift and background models usually showed some difference in base frequencies and substitution rate estimates.

Figure **??** shows that the topology estimated was identical across alignments. The rooting for ClustalW was different, as was the shift placement.

4 Discussion

Alignment quality Our results from the simulation study supports that alignment accuracy can have an impact on tree topology estimation. Substitution models accounting for nonhomogeneous and nonstationary sequence

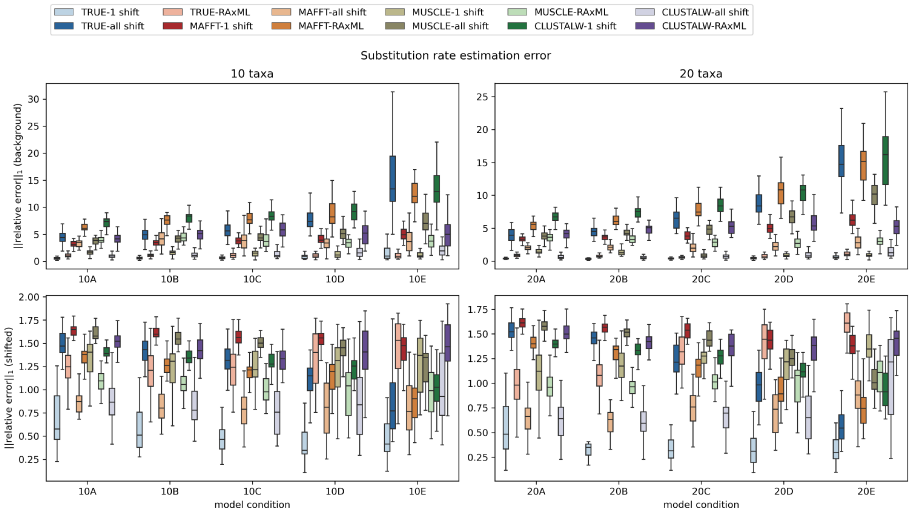


Figure 5: Substitution rate estimation error as measured by L1-norm of the relative errors for each exchangeability parameter.

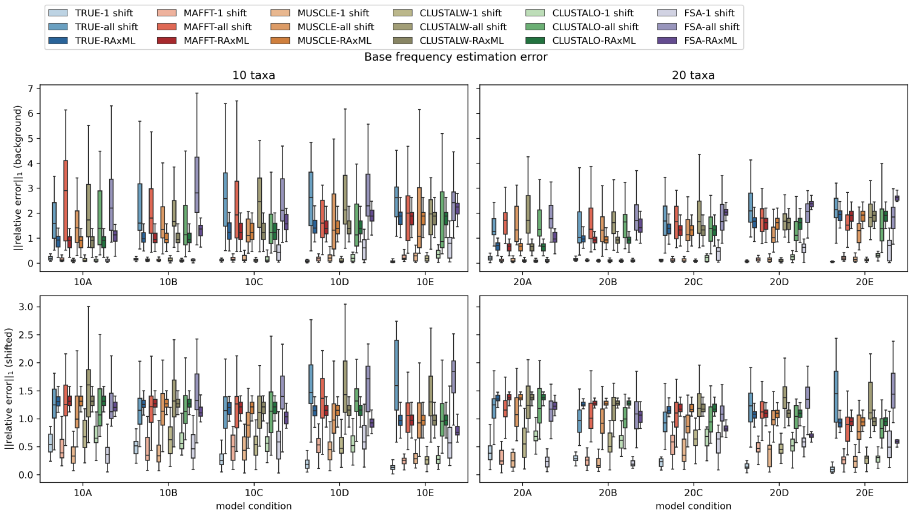


Figure 6: Base frequency estimation error as measured by the L1 norm of the relative errors for each base frequency.

evolution can be used to study the origins and consequences of substitution rate variation across lineages. From both the simulation and empirical study, downstream estimates of substitution rates can vary between different alignment methods.

FSA exhibits the most underalignment, as shown in table **??**. This seems

	muscle		clustalo		clustalw		fsa	
nRF	Mean	Std	Mean	Std	Mean	Std	Mean	Std
mafft	0.044	0.088	0.052	0.097	0.059	0.106	0.041	0.087
muscle			0.053	0.100	0.058	0.108	0.044	0.090
clustalo					0.059	0.105	0.052	0.095
clustalw							0.061	0.103

	muscle		clustalo		clustalw		fsa	
nRF (rooted)	Mean	Std	Mean	Std	Mean	Std	Mean	Std
mafft	0.171	0.191	0.188	0.200	0.191	0.198	0.191	0.205
muscle			0.192	0.204	0.191	0.203	0.198	0.206
clustalo					0.192	0.204	0.213	0.205
clustalw							0.217	0.207

Table 4: Aggregated statistics of Robinson-Foulds distance of gene trees estimated using a single-shift model.

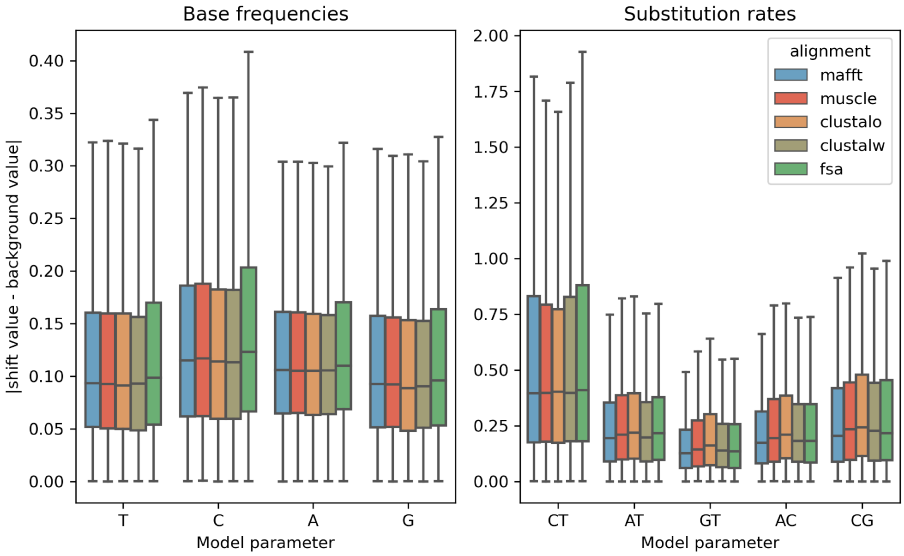
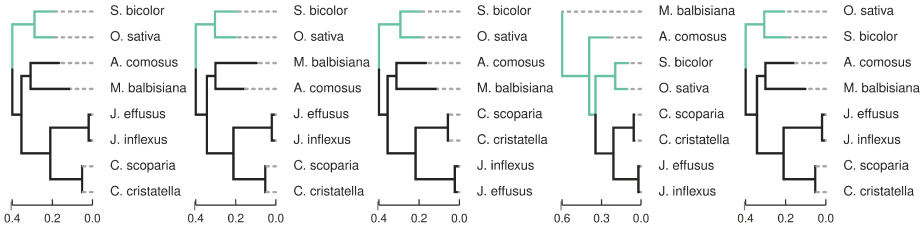


Figure 7: Box plot of the difference between background and shift model parameters estimated for each of the single copy orthologs. Median and interquartile range is represented, whiskers are 1.5IQR and values outside of the whiskers are not shown.

to negatively impact tree topology estimation far more than substitution model parameter estimation. Conversely, MUSCLE and Clustal Omega perform relatively well in the task of topology estimation among the selected alignment methods, but poorly estimate background model parameters. However, there isn't an obvious pattern from the alignment summary statistics that would point



(a) MAFFT (b) MUSCLE (c) ClustalO (d) ClustalW (e) FSA

Figure 8: Tree topologies for the grass dataset estimated using concatenated MAFFT, MUSCLE, Clustal Omega, ClustalW, and FSA alignments respectively. Branches that are predicted to have evolved with elevated G+C base frequencies relative to the rest of the tree are highlighted in green.

Alignment	Which model	T	C	A	G	C↔ T	A↔ T	G↔ T	A↔ C	C↔ G
mafft	shift	0.294	0.176	0.313	0.217	1.245	0.344	0.336	0.511	0.600
mafft	background	0.234	0.254	0.243	0.269	1.087	0.420	0.313	0.432	0.585
muscle	shift	0.294	0.178	0.311	0.217	1.233	0.389	0.380	0.580	0.676
muscle	background	0.234	0.254	0.243	0.269	1.082	0.450	0.349	0.482	0.675
clustalo	shift	0.295	0.179	0.309	0.218	1.194	0.416	0.406	0.627	0.722
clustalo	background	0.234	0.253	0.243	0.269	1.071	0.460	0.359	0.503	0.705
clustalw	shift	0.287	0.169	0.331	0.213	1.343	0.389	0.410	0.585	0.615
clustalw	background	0.246	0.241	0.251	0.262	1.111	0.450	0.341	0.479	0.575
fsa	shift	0.291	0.178	0.316	0.215	1.251	0.350	0.341	0.513	0.605
fsa	background	0.231	0.259	0.242	0.268	1.075	0.427	0.314	0.427	0.579

Table 5: Model parameter estimates from the concatenated analysis

to some quality of the alignments resulting from these methods that leads to this difference.

In the concatenated empirical study, there was consensus between all MSA methods except ClustalW topologically as well as for shift placement. Even when averaged across all loci, substitution rate estimates were highly variable between alignment methods. The simulation study results suggest that it is more difficult to estimate model parameters for the more basal substitution model. From the standard deviation columns in **??**, we see there's less variance across model parameter estimates for the model with a higher estimated G+C base frequencies across loci. The model with elevated G+C typically corresponds to the subtree containing Oryza sativa and Sorghum bicolor.

Branch model of substitution In the simulation study, phylogenetic estimation using a branch model matching the number of shifts that the sequence data was evolved under gives the best performance, as expected. Interestingly, the all-shift model performs very closely to the no-shift model for topology estimation, except in the 20E model condition, where it performs slightly better than the no-shift model. This difference in performance may indicate that the improvement of these models is greater when there are more sequences being

studied and they are more divergent. Though in most cases this seems to suggest that the most general overparameterization doesn't provide a significant improvement in topological estimation over model misspecification.

A generalization of single-shift branch model, which we'll call k-shift, can be described with the aim to achieve the simplest explanation. This model would make the tradeoff of having less continuous parameter estimation in exchange for having a larger search space to explore, but signals of strong shifts might be useful for narrowing this search space.

Limitations This study looks at a very specific case where exactly one change has occurred in the phylogeny. We demonstrate that even in this simplified case, all aspects of phylogenetic estimation are impacted by both alignment quality and model misspecification. In this scenario, we assumed that there were exactly two sets of substitution model parameters, and that one of those sets applied to all the descending edges from a starting edge. As it is, the search space of branch model assignments for each tree topology is $O(n)$ where n is the number of taxa. Natural extensions would be to allow for more sets of substitution model parameters as well as less restrictions on what assignment of these models to the branches are considered. These relaxations drastically increase the size of the search space for model assignments to the tree topology, as well as the number of continuous parameters to optimize for with the former. Because of this, this study did not look at how a k-shift model would perform in the case that there were potentially more sets of substitution model parameters.

A k-shift model is a natural extension, but would pose several challenges for estimation on empirical data as well. Furthermore, these would also limit our ability to look at realistic model conditions. One issue is in the availability of empirical data with novel observations of multiple compositional shifts to gather estimates from. Most studies only make note of two categories, usually high GC and low GC. Another arises from how quickly the search space increases, as finding multiple significant changes in internal branches would necessitate more taxa to study.

5 Conclusions

In both the simulation and empirical studies, we looked at how both alignment accuracy and model misspecification had an impact on downstream phylogenetic inference and estimation. In our simulation study, we looked at a scenario where there's exactly one change in the substitution model that occurs in the simulated tree, and that it occurs such that all descending edges also evolve with that model. We showed that even in this simple case, MSA quality affected all aspects of downstream phylogenetic estimation using a nonhomogeneous model, from tree topology to continuous parameter estimation. Furthermore, we found using a nonhomogeneous substitution model for maximum likelihood estimation yielded closer to ground truth results than using a homogeneous substitution model.

In our empirical study, we observed an impact in tree topology estimation when using a nonhomogeneous model versus a homogeneous substitution model, supporting that the homogeneity-across-lineages assumption can affect estimation even when dealing with large concatenated alignments. Furthermore, we found that estimates using different alignments had a fair amount of disagreement between their estimated gene tree topologies, and estimated continuous parameters were even more sensitive to the alignment method used.

Acknowledgements

We thank Kevin Childs and Yu-ya Liang for help with the grass dataset. This work is supported in part by the National Science Foundation Research Traineeship Program (DGE-1828149) to Rei Doko and through computational resources and services provided by the Institute for Cyber-Enabled Research at Michigan State University.

References

B. Boussau and M. Gouy. Efficient Likelihood Computations with Nonreversible Models of Evolution. *Systematic Biology*, 55(5):756–768, Oct. 2006. ISSN 1063-5157. doi: 10.1080/10635150600975218.

R. K. Bradley, A. Roberts, M. Smoot, S. Juvekar, J. Do, C. Dewey, I. Holmes, and L. Pachter. Fast Statistical Alignment. *PLOS Computational Biology*, 5 (5):e1000392, May 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000392.

S. L. Cameron. Insect Mitochondrial Genomics: Implications for Evolution and Phylogeny. *Annual Review of Entomology*, 59(1):95–117, 2014. doi: 10.1146/annurev-ento-011613-162007.

Y. Clément, M.-A. Fustier, B. Nabholz, and S. Glémin. The Bimodal Distribution of Genic GC Content Is Ancestral to Monocot Species. *Genome Biology and Evolution*, 7(1):336–348, Jan. 2015. ISSN 1759-6653. doi: 10.1093/gbe/evu278.

R. C. Edgar. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, Mar. 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh340.

D. M. Emms and S. Kelly. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):238, Nov. 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1832-y.

W. Fletcher and Z. Yang. INDELible: A Flexible Simulator of Biological Sequence Evolution. *Molecular Biology and Evolution*, 26(8):1879–1888, Aug. 2009. ISSN 0737-4038. doi: 10.1093/molbev/msp098.

E. D. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, J. Li, F. Zhang, H. Li, L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, M. S. Bayzid, V. Zavidovych, S. Subramanian, T. Gabaldón, S. Capella-Gutiérrez, J. Huerta-Cepas, B. Rekepalli, K. Munch, M. Schierup, B. Lindow, W. C. Warren, D. Ray, R. E. Green, M. W. Bruford, X. Zhan, A. Dixon, S. Li, N. Li, Y. Huang, E. P. Derryberry, M. F. Bertelsen, F. H. Sheldon, R. T. Brumfield, C. V. Mello, P. V. Lovell, M. Wirthlin, M. P. C. Schneider, F. Prosdocimi, J. A. Samaniego, A. M. V. Velazquez, A. Alfaro-Núñez, P. F. Campos, B. Petersen, T. Sicheritz-Ponten, A. Pas, T. Bailey, P. Scofield, M. Bunce, D. M. Lambert, Q. Zhou, P. Perelman, A. C. Driskell, B. Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B. Liu, K. Wu, J. Xiao, X. Yinqi, Q. Zheng, Y. Zhang, H. Yang, J. Wang, L. Smeds, F. E. Rheindt, M. Braun, J. Fjeldsa, L. Orlando, F. K. Barker, K. A. Jønsson, W. Johnson, K.-P. Koepfli, S. O'Brien, D. Haussler, O. A. Ryder, C. Rahbek, E. Willerslev, G. R. Graves, T. C. Glenn, J. McCormack, D. Burt, H. Ellegren, P. Alström, S. V. Edwards, A. Stamatakis, D. P. Mindell, J. Cracraft, E. L. Braun, T. Warnow,

W. Jun, M. T. P. Gilbert, and G. Zhang. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, Dec. 2014. doi: 10.1126/science.1253451.

K. Katoh and D. M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, Apr. 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst010.

M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, J. Thompson, T. Gibson, and D. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, Nov. 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm404.

K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, and T. Warnow. Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees. *Science*, 324(5934):1561–1564, June 2009. doi: 10.1126/science.1171243.

R. F. McCormick, S. K. Truong, A. Sreedasyam, J. Jenkins, S. Shu, D. Sims, M. Kennedy, M. Amirebrahimi, B. D. Weers, B. McKinley, A. Mattison, D. T. Morishige, J. Grimwood, J. Schmutz, and J. E. Mullet. The Sorghum bicolor reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, 93(2):338–354, 2018. ISSN 1365-313X. doi: 10.1111/tpj.13781.

R. Ming, R. VanBuren, C. M. Wai, H. Tang, M. C. Schatz, J. E. Bowers, E. Lyons, M.-L. Wang, J. Chen, E. Biggers, J. Zhang, L. Huang, L. Zhang, W. Miao, J. Zhang, Z. Ye, C. Miao, Z. Lin, H. Wang, H. Zhou, W. C. Yim, H. D. Priest, C. Zheng, M. Woodhouse, P. P. Edger, R. Guyot, H.-B. Guo, H. Guo, G. Zheng, R. Singh, A. Sharma, X. Min, Y. Zheng, H. Lee, J. Gurtowski, F. J. Sedlazeck, A. Harkess, M. R. McKain, Z. Liao, J. Fang, J. Liu, X. Zhang, Q. Zhang, W. Hu, Y. Qin, K. Wang, L.-Y. Chen, N. Shirley, Y.-R. Lin, L.-Y. Liu, A. G. Hernandez, C. L. Wright, V. Bulone, G. A. Tuskan, K. Heath, F. Zee, P. H. Moore, R. Sunkar, J. H. Leebens-Mack, T. Mockler, J. L. Bennetzen, M. Freeling, D. Sankoff, A. H. Paterson, X. Zhu, X. Yang, J. A. C. Smith, J. C. Cushman, R. E. Paull, and Q. Yu. The pineapple genome and the evolution of CAM photosynthesis. *Nature Genetics*, 47(12):1435–1442, Dec. 2015. ISSN 1546-1718. doi: 10.1038/ng.3435.

B. Nabholz, A. Künstner, R. Wang, E. D. Jarvis, and H. Ellegren. Dynamic Evolution of Base Composition: Causes and Consequences in Avian Phylogenomics. *Molecular Biology and Evolution*, 28(8):2197–2210, Aug. 2011. ISSN 0737-4038. doi: 10.1093/molbev/msr047.

L. Nakhleh, B. M. E. Moret, U. Roshan, K. St. John, J. Sun, and T. Warnow. The Accuracy of Fast Phylogenetic Methods for Large Datasets. In *Bio-*

computing 2002, pages 211–222. WORLD SCIENTIFIC, Dec. 2001. ISBN 978-981-02-4777-5. doi: 10.1142/9789812799623_0020.

S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R. L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman, and C. R. Buell. The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Research*, 35(suppl_1):D883–D887, Jan. 2007. ISSN 0305-1048. doi: 10.1093/nar/gkl976.

J. Planta, Y.-Y. Liang, H. Xin, M. T. Chansler, L. A. Prather, N. Jiang, J. Jiang, and K. L. Childs. Chromosome-scale genome assemblies and annotations for Poales species Carex cristatella, Carex scoparia, Juncus effusus, and Juncus inflexus. *G3 (Bethesda, Md.)*, 12(10):jkac211, Sept. 2022. ISSN 2160-1836. doi: 10.1093/g3journal/jkac211.

D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, Feb. 1981. ISSN 0025-5564. doi: 10.1016/0025-5564(81)90043-2.

F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539, Jan. 2011. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb.2011.75.

A. Stamatakis. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, May 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu033.

C. Tuffley and M. Steel. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59(3):581–607, May 1997. ISSN 0092-8240. doi: 10.1016/S0092-8240(97)00001-3.

W. Wang, A. Hejasebazzi, J. Zheng, and K. J. Liu. Build a better bootstrap and the RAWR shall beat a random path to your door: Phylogenetic support estimation revisited. *Bioinformatics*, 37(Supplement_1):i111–i119, July 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab263.

Z. Wang, H. Miao, J. Liu, B. Xu, X. Yao, C. Xu, S. Zhao, X. Fang, C. Jia, J. Wang, J. Zhang, J. Li, Y. Xu, J. Wang, W. Ma, Z. Wu, L. Yu, Y. Yang, C. Liu, Y. Guo, S. Sun, F.-C. Baurens, G. Martin, F. Salmon, O. Garsmeur, N. Yahiaoui, C. Hervouet, M. Rouard, N. Laboureau, R. Habas, S. Ricci, M. Peng, A. Guo, J. Xie, Y. Li, Z. Ding, Y. Yan, W. Tie, A. D'Hont, W. Hu, and Z. Jin. Musa balbisiana genome reveals subgenome evolution and functional divergence. *Nature Plants*, 5(8):810–821, Aug. 2019. ISSN 2055-0278. doi: 10.1038/s41477-019-0452-6.

Z. Yang. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39(1):105–111, July 1994. ISSN 1432-1432. doi: 10.1007/BF00178256.

Z. Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, Aug. 2007. ISSN 0737-4038. doi: 10.1093/molbev/msm088.

Supplementary Online Materials

1 Supplementary Methods

Software commands used. INDELible [Fletcher and Yang, 2009] version 1.03 was run using the following settings to simulate model tree evolution

```
[TYPE] NUCLEOTIDE 1
[MODEL] mymodel
    [submodel] JC
[TREE] mytree
    [rooted] <# taxa>
[PARTITIONS] mypartition [mytree mymodel 1]
[EVOLVE] mypartition <# replicates> output
```

The trees are rescaled to have nonultrametric branch lengths and the following control settings to simulate sequence evolution:

```
[TYPE] NUCLEOTIDE 1
[SETTINGS]
  [output] PHYLIP
[MODEL] background
  [submodel] GTR <CT> <AT> <GT> <AC> <CG>
  [statefreq] <T> <C> <A> <G>
  [indelmodel] USER <path to indel distribution>
  [indelrate] <indel rate>
[MODEL] shift
  [submodel] GTR <CT> <AT> <GT> <AC> <CG>
  [statefreq] <T> <C> <A> <G>
  [indelmodel] USER <path to indel distribution>
  [indelrate] <indel rate>
[TREE] mytree <tree>
  [treedepth] <specified tree height>
[BRANCHES] mymodel <tree with model placements defined>
[PARTITIONS] mypartition [mytree mymodel <sequence length>]
[EVOLVE] mypartition 1 sequence
```

The following command was used to perform MSA estimation with MAFFT [Katoh and Standley, 2013] version 7.475

```
mafft <input sequence> <output alignment>
```

MUSCLE [Edgar, 2004] version 5.0.1428 was run with the following:
`muscle -align <input sequence> -output <output alignment>`
Clustal Omega [Sievers et al., 2011] version 1.2.4 was run with the following
`clustalo -i <input sequence> -t DNA --threads 1 ><output alignment>`
ClustalW [Larkin et al., 2007] version 2.1 was run with the following
`clustalw2 <input sequence> -type=DNA -outfile=<output alignment>`
FSA [Sievers et al., 2011] version 1.15.9 was run with the following
`fsa <input sequence --maxram 8192 ><output alignment>`
For single-shift search, PAML [Yang, 2007] was run with the following control file:

```
     seqfile = <sequence_path>
    treefile = <tree path>
     outfile = <result output path>
       noisy = 3
     verbose = 3
     runmode = 0
       model = 7   * GTR model
       Mgene = 0
       ndata = 1
       nhomo = 5
   fix_kappa = 2
       clock = 0
   fix_alpha = 1
       alpha = 0.
       getSE = 0
RateAncestor = 0
   cleandata = 0
      method = 0
 fix_blength = 0
```

The control file for the all-shift model is identical, except:

```
       nhomo = 3
   fix_kappa = 0
```

RAxML [Stamatakis, 2014] was run with the following command:
`raxml -s <msa path> -n <name> -m GTRCAT -V -p <random number>`

Grass dataset processing. To obtain single copy orthologs for the grass dataset, we ran orthofinder using the following command:
`orthofinder -f <path containing sequence files>`

2 Supplementary Results and Discussion.

Simulation study runtime and memory usage. Memory usage for the nonhomogeneous substitution model did not exceed 1 GB.

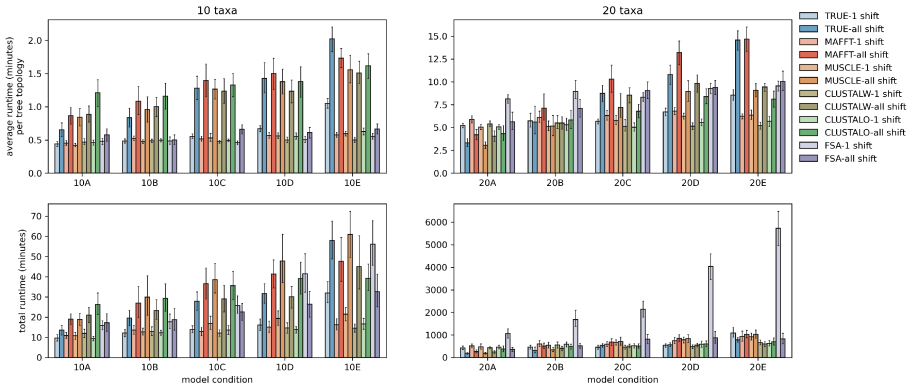


Figure 1: Runtime for nonhomogeneous tree search. The first row represents how long it takes for PAML to do continuous parameter optimization and likelihood calculation, and the second row represents the total time the wrapper script takes.

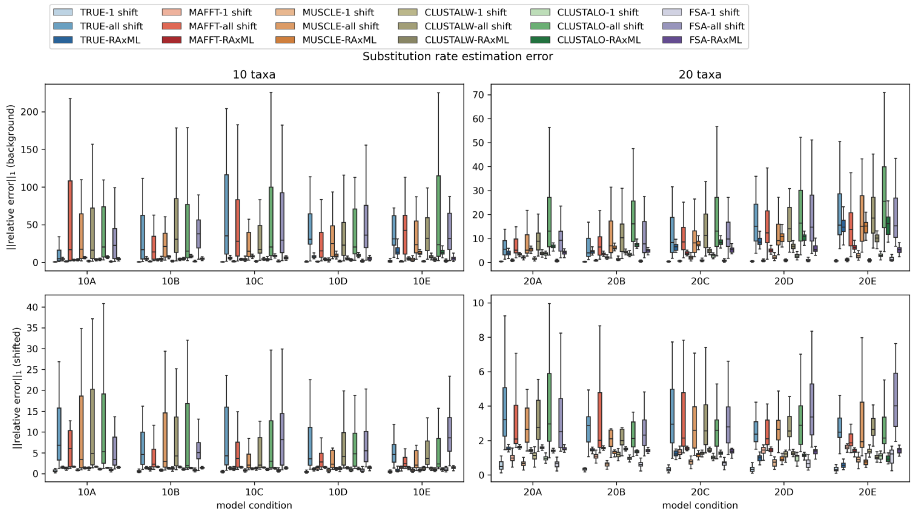


Figure 2: Substitution rate error.

Estimation error For branch lengths, there wasn't a clean way to quantify error, in part because tree estimation and branch length estimation are heavily intertwined. Comparing leaf-edge distances only captures roughly half of the estimated branch lengths, but comparing pairwise distances does not consider the topology in any way. Kuhner-Felsenstein distance [Kuhner and Felsenstein, 1994], shown in the middle panels in 3, takes topology into consideration, but is difficult to interpret. For this reason we cannot draw any meaningful conclusions about branch length estimation.

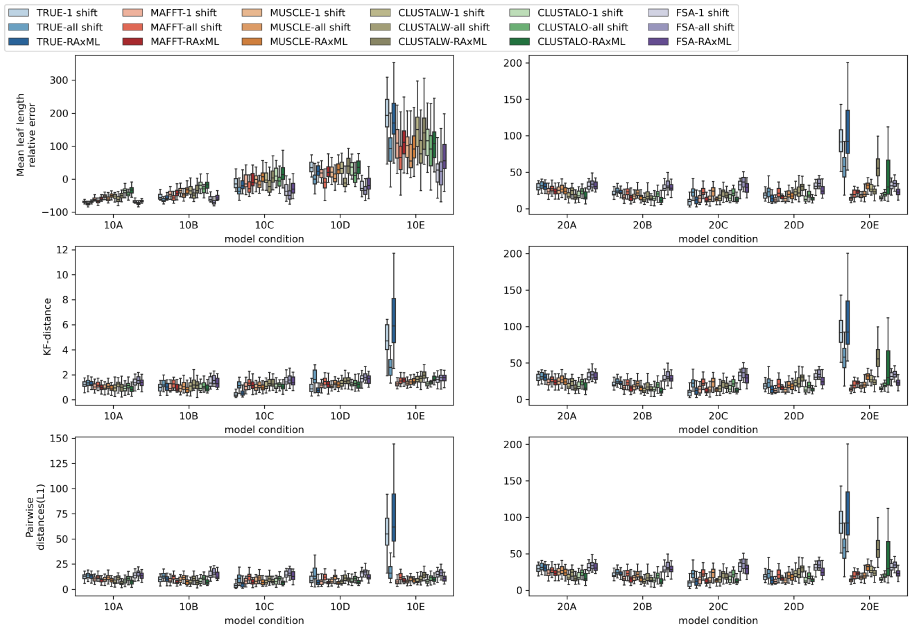


Figure 3: Various measures for branch length estimation error. KF stands for Kuhner and Felsenstein [1994].

References

R. C. Edgar. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, Mar. 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh340.

W. Fletcher and Z. Yang. INDELible: A Flexible Simulator of Biological Sequence Evolution. *Molecular Biology and Evolution*, 26(8):1879–1888, Aug. 2009. ISSN 0737-4038. doi: 10.1093/molbev/msp098.

K. Katoh and D. M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, Apr. 2013. ISSN 0737-4038. doi: 10.1093/molbev/mst010.

M. K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3):459–468, May 1994. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040126.

M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, J. Thompson, T. Gibson, and D. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, Nov. 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm404.

F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539, Jan. 2011. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb.2011.75.

A. Stamatakis. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, May 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu033.

Z. Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, Aug. 2007. ISSN 0737-4038. doi: 10.1093/molbev/msm088.