Virus Evolution, 2020, 6(2): veaa066

doi: 10.1093/ve/veaa066
Research Article

Efficient inference, potential, and limitations of site-specific substitution models

Vadim Puller,[1,2] Pavel Sagulenko,[3] and Richard A. Neher[1,2,*,†]

[1]Biozentrum, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland, [2]SIB Swiss Institute of Bioinformatics, Klingelbergstrasse 61, Basel, Switzerland and [3]Max Planck Institute for Developmental Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany

*Corresponding author: E-mail: richard.neher@unibas.ch

†https://orcid.org/0000-0003-2525-1407

Abstract

Natural selection imposes a complex filter on which variants persist in a population resulting in evolutionary patterns that vary greatly along the genome. Some sites evolve close to neutrally, while others are highly conserved, allow only specific states, or only change in concert with other sites. On one hand, such constraints on sequence evolution can be to infer biological function, one the other hand they need to be accounted for in phylogenetic reconstruction. Phylogenetic models often account for this complexity by partitioning sites into a small number of discrete classes with different rates and/or state preferences. Appropriate model complexity is typically determined by model selection procedures. Here, we present an efficient algorithm to estimate more complex models that allow for different preferences at every site and explore the accuracy at which such models can be estimated from simulated data. Our iterative approximate maximum likelihood scheme uses information in the data efficiently and accurately estimates site-specific preferences from large data sets with moderately diverged sequences and known topology. However, the joint estimation of site-specific rates, and site-specific preferences, and phylogenetic branch length can suffer from identifiability problems, while ignoring variation in preferences across sites results in branch length underestimates. Site-specific preferences estimated from large HIV *pol* alignments show qualitative concordance with intra-host estimates of fitness costs. Analysis of these substitution models suggests near saturation of divergence after a few hundred years. Such saturation can explain the inability to infer deep divergence times of HIV and SIVs using molecular clock approaches and time-dependent rate estimates.

Key words: phylogenetics; fitness landscapes; algorithms.

1. Introduction

Over time, genome sequences change through mutations and are reshuffled by recombination. Modifications to the genomes are filtered by selection for survival such that beneficial variants spread preferentially and those that impair function are purged. As a result, some parts of genomes change rapidly, while other are strongly conserved. In addition to variation of the evolutionary rate, different sites in a genome explore different subsets of the available states. Some positions in a protein, for example, might only allow for hydrophobic amino acids, while others require acidic side chains. Patterns of conservation and variation, possibly involving more than one site, are therefore shaped by functional constraints which in turn allows inference of biological function from genetic variation.

Similarly, phylogenetics aims at reconstructing the relationships and history of homologous sequences from the substitutions that occurred in the past. Modern phylogenetic methods describe this stochastic evolutionary process with probabilistic models of sequence evolution and aim to find phylogenies that either maximize the likelihood of observing the alignment or

sample phylogenies from a posterior probability distribution (Felsenstein 2004).

Inferring phylogenies is a computationally challenging problem since the number of phylogenies grows super-exponentially with the number of taxa and because the calculation of the likelihood is computationally costly (though linear in the number of taxa). Due to this computational complexity, the most commonly used substitution models are simple caricatures of biological complexity. The simplest substitution models assume that all sites and sequence states are equivalent and evolve at the same rate, that is, they assume an unconstrained non-functional sequence that mutates at random between the different sequence states. Such simple models are clearly inadequate and ignoring rate heterogeneity tends to result in biased estimates of divergence times or otherwise erroneous results (Yang 1996). Most commonly used models account for variation in substitution rates among sites and average properties of the substitution process such as transition/transversion bias or more frequent substitution between similar amino acids (Yang 1994; Price, Dehal, and Arkin 2010; Stamatakis 2014; Nguyen et al. 2015). To avoid over-fitting, these methods typically do not estimate a rate for each site, but treat site-specific rates as random effects that are integrated out (often using discrete approximations of a Gamma distribution (Yang 1996), mixture of multiple unimodal distributions (Mayrose, Friedman, and Pupko 2005), or a small number of fixed rates).

In addition to rate variation different sites in a protein differ in the amino acids they allow, different positions in codons experience different constraints, and secondary structure or protein binding to DNA or RNA imposes additional levels of selection. Such variation can again be accounted for by modeling multiple categories with different preferences for amino acids or nucleotides to which sites can be assigned or which can be integrated out during phylogenetic inference (Lartillot and Philippe 2004; Shapiro, Rambaut, and Drummond 2006; Kainer and Lanfear 2015).

Recent deep mutational scanning experiments have shown that site-specific preferences are mostly conserved between moderately diverged proteins (Doud, Ashenberg, and Bloom 2015). Using such experimentally inferred site-specific models in phylogenetic inference greatly increases the likelihood of the data Bloom (2014). More than two decades ago, Halpern and Bruno (1998) pointed out that ignoring that equilibrium frequencies vary from one position to another will result in underestimation of branch lengths of a phylogeny—possibly dramatically when frequencies are heavily skewed. Hilton and Bloom (2018) recently showed that experimentally measured preference not only improve the phylogenetic fit, but also results in longer branch length estimates. Models with site-specific preferences are also known as mutation–selection balance models (Bruno 1996; Yang and Nielsen 2008) reflecting the intuition that equilibrium frequencies are determined by competition of diversifying processes (like mutation) and selection for an optimal function.

Instead of partitioning the sequence into a small number of categories, we ask under what circumstances it is possible to estimate models that allow different state frequencies at every site in the alignment and what implications such variation has for phylogenetic inference. While biologically plausible, estimating such models from data exacerbates the over-fitting problem and it rarely attempted in practice. In the context of the site-specific models this issue has become known as *extensive parametrization* or even *infinitely many parameters problem* Rodrigue (2013) and Spielman and Wilke (2016). With sufficient data, however, site-specific parameters can be accurately estimated (Tamuri, dos Reis, and Goldstein 2012; Scheffler, Murrell, and Pond 2014; Spielman and Wilke 2016). Here, we implement an EM-style algorithm inspired by (Bruno 1996) to infer site-specific rates and preferences from simulated data, quantify its accuracy and the different sources of bias and noise, and show how divergence time estimates depend on the fidelity with which site-specific model parameters are known. We apply this algorithm to large HIV-1 alignments and explore the consequences for phylogenetic inference.

2. Results

2.1 Efficient inference of site-specific substitution models

Following work by Halpern and Bruno (1998), we parameterize a site-specific general time-reversible (GTR) substitution model at site a from state j to i as:

$$\begin{aligned} Q_{ij}^a &= \mu^a p_i^a W_{ij} \text{for} i \neq j, \\ Q_{ii}^a &= -\sum_k Q_{ki}^a . \end{aligned} \tag{1}$$

Here, μ^a is the substitution rate at site a, p_i^a is the equilibrium frequency of state i at site a, and W_{ij} is a symmetric substitution matrix that we assume to be the same for all sites (in what follows, superscript a will always refer to the position in the sequence, while subscript i, j, n, m refers to the state). The second equation ensures conservation of probability. In addition, we require $\sum_i p_i^a = 1$ and $\sum_{a=1}^{L} \sum_{i \neq j} W_{ij} p_i^a p_j^a = L$ to normalize the frequencies and fix the scale of W_{ij}.

Extending the approach by Bruno (1996), we show in Section 3 and the supplement that the iterative update rules

$$\begin{aligned} \mu^a &\leftarrow \mu^a \frac{c + \sum_{ij} n_{ij}^a}{\sum_{ij} \mu^a p_i W_{ij} \tau_j^a} \\ p_i^a &\leftarrow p_i^a \frac{c + \delta_{is^a} + \sum_{j \neq i} n_{ij}^a}{p_i^a (qc + 1) + \mu^a p_i^a \sum_j W_{ij} \tau_j^a} \\ W_{ij} &\leftarrow W_{ij} \frac{\sum_a (n_{ij}^a + n_{ji}^a)}{\sum_a \mu^a W_{ij} (p_i^a \tau_j^a + p_j^a \tau_i^a)} \end{aligned} \tag{2}$$

approximately maximize the likelihood of the data. Here, τ_j^a is the time site a spends in state j across the tree, n_{ij}^a is the number of transitions from state j to state i at site a, c is a pseudocount analogous to a Dirichlet prior that in the absence of data will drive the p_i^a to a flat distribution and the substitution rates to c divided by the total tree length. To make the behavior of these update rules more explicit, we have multiplied numerator and denominator with the parameter that is being updated. This 1, shows that the update rules are multiplicative and therefore ensure positivity, and 2, illustrates that each of these rules are the ratio of the *observed* number of transitions between states n_{ij}^a and the *expected* number $p_i^a W_{ij} \tau_j^a$—each appropriately summed over sites or states. These update rules are an example of non-negative factorization algorithms (Lee and Seung 2001).

2.2 Accuracy of model inferences

The accuracy of this iterative solution will depend 1, on the validity of the linear approximation, 2, the accuracy to which n_{ij}^a and τ_j^a can be estimated, and 3, the accuracy of the tree reconstruction. Furthermore, the estimation of this extensive number of parameters requires large data sets in which most positions

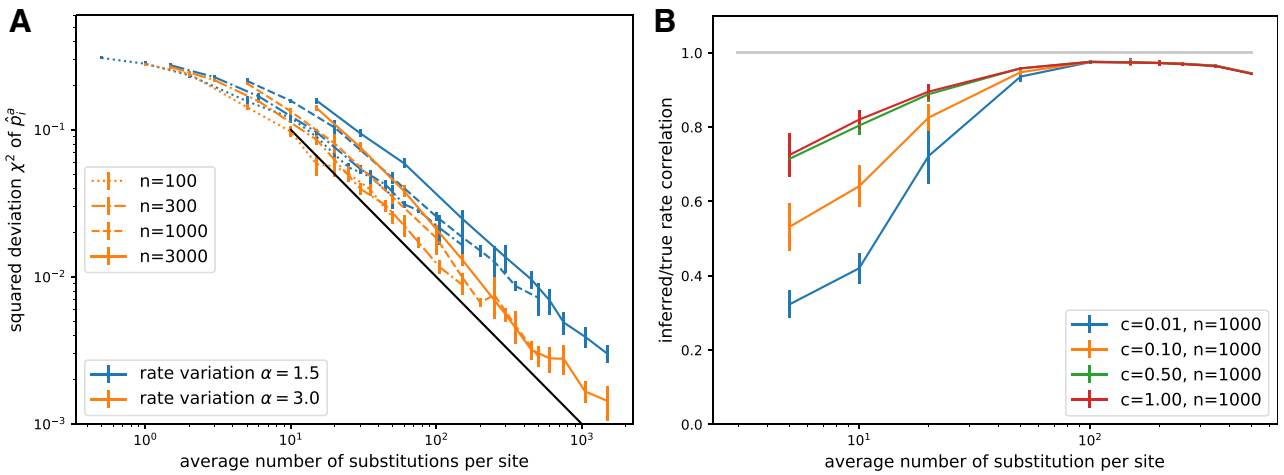

Figure 1. Accuracy of iterative estimation of site-specific GTR models as a function of the expected number of state changes along the tree. (A) Mean squared error of the inferred $\hat{p}_i^a$ scales inversely with the tree length (indicated by the black line), suggesting the accuracy is limited by the Poisson statistics of observable mutations. This scaling holds for different number of sequences per tree ($n \in [100, 300, 1,000, 3,000]$) and different degrees of rate variation. (B) The relative substitution rates are accurately inferred as soon as the typical site experiences several substitutions across the tree as quantified here as Pearson correlation coefficient between true and inferred rates. Regularization via pseudo-counts reduces over-fitting at low divergence. Analogous results for alphabets of size $q = 20$ are shown in Supplementary Fig. S1.

mutate multiple times on the tree—otherwise the model will overfit the data and result in inaccurate estimates.

To assess these sources of error and effect of data set sizes independently, we simulated sequences evolving along trees and explicitly recorded the number of transitions between states at every site (n_{ij}^a) and the time spent in each state (τ_j^a) for a range of sample sizes, levels of divergence, and models with different degrees of variation between sites, see Section 3. From these simulated data sets, we inferred the site-specific models using different aspects of the simulated data: 1, full knowledge of all transitions and ancestral states, 2, the tree and the tip sequences, or 3, only the tip sequences from which a tree was reconstructed. We quantified the accuracy of the inferences as the squared estimation error $\chi^2 = L^{-1}\sum_{a,i}(\hat{p}_i^a - p_i^a)^2$, where $\hat{p}_i^a$ and p_i^a are the inferred and true equilibrium probabilities, respectively. Lower χ^2 corresponds to more accurate estimates.

In our first analysis summarized in Fig. 1 we quantified the accuracy at which p_i^a and μ^a can be estimated if ancestral sequences are known. As expected, the average squared error χ^2 decreases with the substitution rate and the size of the data set. The error is well predicted by the average number of substitutions per site, which increases with both data set size and the substitution rate. The data in Fig. 1A are shown on double logarithmic scales, such that the approximately straight decrease in χ^2 with slope -1 implies that the squared deviation is inversely proportional to the expected number of substitutions. This scaling suggests that accuracy is limited by the inherent stochasticity of the evolutionary process and that the inference uses all available information efficiently.

The simulation data were generated by drawing site-specific rates from a Gamma distribution with $\alpha = 1.5$ (blue lines) and $\alpha = 3$ (orange lines). The scaling behavior $\chi^2 \sim (\text{tree length})^{-1}$ is more evident for $\alpha = 3$ than for $\alpha = 1.5$. In the latter case of strong rate variation, the average squared error χ^2 is dominated by a small number of sites with low rates at which frequencies are estimated poorly and the average tree length does not capture behavior at these sites.

At the largest substitution rates, branch lengths are on the order of 0.5 and the linear approximation underlying the iterative equations is not longer accurate. Nevertheless, the accuracy of the $\hat{p}_i^a$ continues to increase. While equilibrium frequencies are insensitive, the substitution rate estimates are affected by linearization and Equation (2) will consistently underestimate μ^a. This is expected as the linearization ignores cases where the same site changes twice along a branch in very much the same way as Hamming distance will underestimate branch length. The relative substitution rates, however, are accurately estimated (with Pearson correlation coefficients >0.9 as soon as the majority of sites experience several mutations across the tree), see Fig. 1B. Regularization by pseudo-counts c reduces the overfitting problem when the number of substitutions per site is small.

Above, we investigated the accuracy of model inferences when the ancestral states and the tree are known. In practice, however, ancestral sequences are unknown and need to be inferred or summed over (marginalized) which will introduce additional uncertainty. This problem was recently described as the 'Darwinian uncertainty principle' by Gascuel and Steel (2020) showing that there are fundamental limits to the accuracy at which model and ancestral states can be estimated jointly.

To test the influence of tree and ancestral state reconstruction, we reconstructed phylogenetic trees using IQ-tree (Nguyen et al. 2015) with a GTR+R10 model (simulated nucleotide sequences) or FastTree (Price, Dehal, and Arkin 2009) using the default settings (simulated amino-acid sequences). Figure 2 compares different schemes to reconstruct ancestral sequences, infer substitution models, and optimize the tree. The simplest approach is to take the inferred tree as given, reconstruct the ancestral states using a simple evolutionary model (e.g. Jukes–Cantor model) and calculate n_{ij}^a and τ_i^a from this reconstruction. This naive approach works well up to root-to-tip distances of about 0.3, beyond which estimates deteriorate, see Fig 2 ('Reconstructed ancestral sequences', red line).

Instead of reconstructing the most likely ancestral sequences, we can instead average over all possible ancestral states, which results in a modest reduction of the error ('Marginalized ancestral sequences', purple line in Fig. 2). More significant

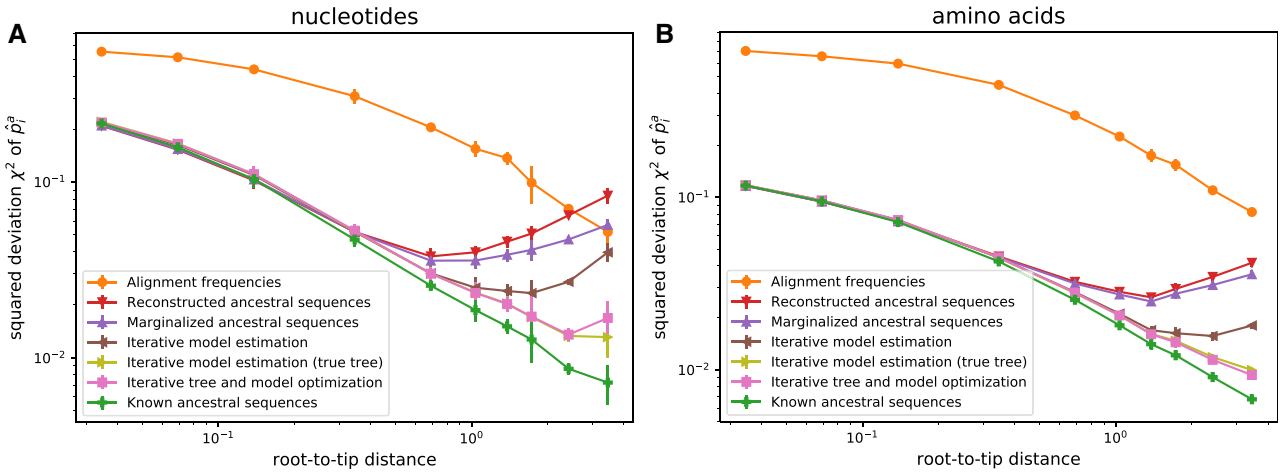

Figure 2. Quantification of errors stemming from tree inference and ancestral reconstruction. Panels A and B show the mean-squared deviation χ^2 of inferred $\hat{p}_i^a$ from the true p_i^a for four-letter and twenty-letter alphabets, respectively. Alignment frequencies are poor estimates of the p_i^a even in very diverse samples, while the different phylogeny aware methods initially improve rapidly with root-to-tip distance. At large root-to-tip distances, ancestral reconstruction becomes less and less certain and estimation of p_i^a fails (red lines). These errors are gradually eliminated by first summing over ancestral uncertainty (violet), iteratively redoing ancestral reconstruction using the inferred model (brown), and re-optimizing branch lengths using the updated models (or using the true tree, yellow/pink). Data in this figure uses were generated assuming Gamma distributed rate variation with $\alpha = 1.5$.

gains are made when iterating model inference and ancestral reconstruction using the inferred site-specific model ('Iterative model estimation', brown line). The error now continues to decrease up to root-to-tip distances of about one. At this level of divergence, tree reconstruction starts becoming problematic and branch lengths deviate substantially from their true values. Using the true tree instead of the reconstructed tree leads to continuous improvements of accuracy with increasing levels of divergence (yellow line). Similar improvements are achieved by optimizing tree branch lengths along with the model. These improvements are consistent with Gascuel and Steel (2020) in that joint estimation of ancestral states and model is possible for Yule trees in which tree length grows linearly with the number of tips while tree depth increases only logarithmically. In contrast, using the frequencies of different sequence states in the alignment as estimates for p_i^a is much less accurate due to the correlation induced by shared ancestry (orange lines in Fig. 1A).

We found qualitatively similar patterns for four-letter and twenty-letter alphabets in the overall accuracy of the estimated model and its dependence on the different approximations, compare Fig. 2A and B. For larger alphabets, the breakdown at large root-to-tip distances is less dramatic and sets in at larger values. Overall, we conclude that site-specific frequencies can be estimated accurately when the overall tree length comfortably exceeds ten such that most sites experienced multiple mutations.

2.3. Implications for branch length estimates in phylogenies

As previously observed by various authors (Halpern and Bruno 1998; Hilton and Bloom 2018), sites with heavily skewed preferences for specific states result in underestimation of branch lengths if these skews are not modeled appropriately. This is a straightforward consequence of the fact that the probability of observing the same state at random is $\sum_i (p_i^a)^2$, which is increasing sharply with more peaked preferences. Models that do not account for site-specific frequencies will take a large number of sites that agree between two sequences as evidence for their close evolutionary relationship, while it is simply a consequence of resampling the same states with high probability. In other words, this underestimation is the result of incorrect models of saturation and is closely related to the observation that incomplete purifying selection can result in erroneous and apparently time-dependent evolutionary rates Wertheim and Kosakovsky Pond (2011).

Figure 3 shows the average branch length (panel A) and the average root-to-tip distance of mid-point rooted trees (panel B) with branch lengths optimized using 1, the true model, 2, using the GTR+R10 model of IQ-tree, and 3, inferred models using different degrees of regularization. While branch lengths are accurately estimated when using the true model, they are systematically underestimated by the GTR+R10 model without site-specific preferences. This problem is particularly severe for the average root-to-tip distance which is dominated by long branches deep in the phylogenies that are prone to underestimation. When ignoring site-specific preferences, the inferred root-to-tip distance is essentially flat for distances greater than 1 (Fig. 3B). This effect is entirely due to skewed equilibrium frequencies, as branch length inference by IQ-tree is accurate if the simulated data had flat $p_i^a = q^{-1}$ with the same rate variation. The underestimation of branch lengths is less severe for larger alphabets or if frequencies are not heavily skewed, see Supplementary Fig. S2.

Surprisingly, using the inferred site-specific models only partially rectified the problem of branch length underestimation, despite the fact that these models are close to the true model in terms of low χ^2. The principle contributor to this deviation is inaccuracies in the rate estimates. Combining the true rates with inferred preferences reduces the error in branch length estimation (see Fig. 3, lines labeled ('true rates')).

2.4 The effect of model misspecification on divergence estimates

In the previous section, we have observed that model deviations, for example the error made during model inference, can result in substantial errors in branch length estimates. To investigate this effect more systematically, we constructed

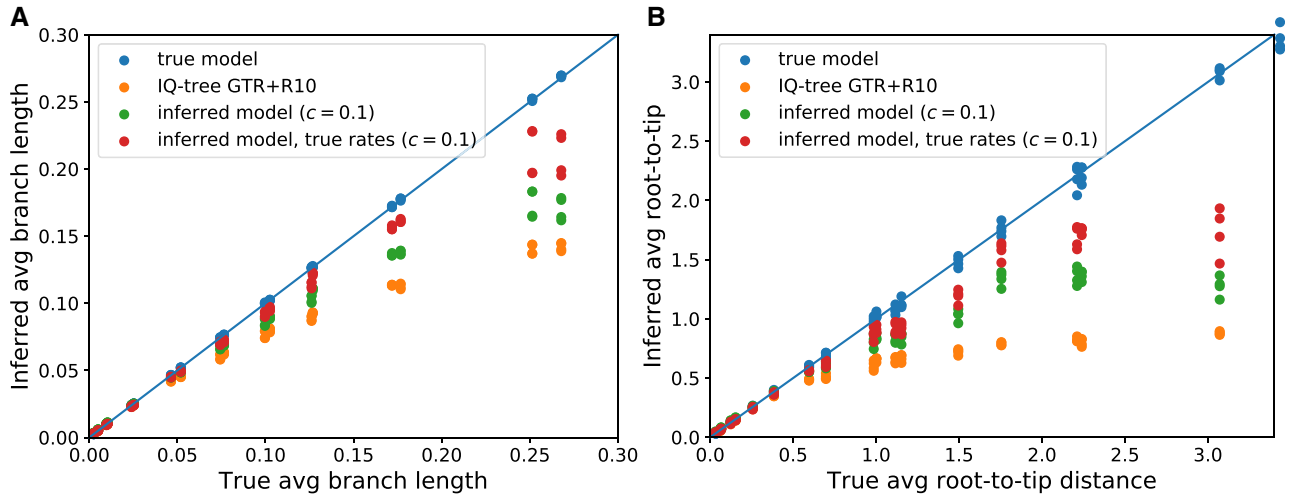

Figure 3. Skewed equilibrium concentration results in branch length underestimates. Panel A shows the inferred average branch length as estimated by IQ-tree and TreeTime as a function of the true average branch length. Panel B shows the results of the same optimization for the average root-to-tip distance which is dominated by deep long branches that are more strongly affected. While using the true model results in unbiased branch lengths, inferred site-specific models only partially ameliorate underestimation, in particular with high pseudo-counts c (see main text). Parameters: $n = 1{,}000$, $\alpha = 1.5$.

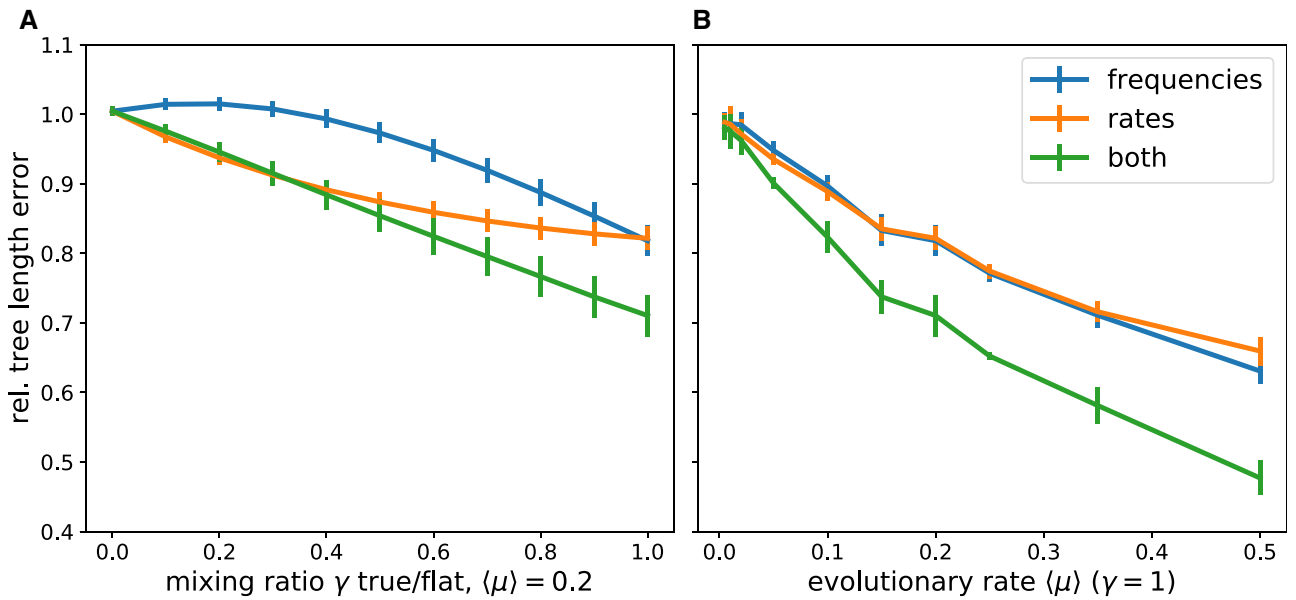

Figure 4. Sensitivity of branch length estimates on model misspecification. Panels A and B show the relative error in total tree length when using a mixture model as defined in Equation (3) for branch length inference. Panel A shows this error as a function of the mixing fraction γ for $\langle\mu\rangle = 0.2$. Panel B shows the error as a function of the evolutionary rate $\langle\mu\rangle$ for $\gamma = 1$. The mixing is applied to the equilibrium frequencies p_i^a, the rates μ^a, or both. The models assume an alphabet size $q = 4$ (nucleotides).

mixtures of the true model and a model with flat p_i^a and/or μ^a as follows: For a mixing fraction γ, we constructed a model with rates

$$\mu^a = \gamma\langle\mu^a\rangle + (1-\gamma)\mu^a_{\text{true}}, \tag{3}$$

where $\langle\mu^a\rangle$ is the average of the rate. Mixtures of site-specific frequencies are constructed analogously.

Figure 4A shows how deviation from the true model affect relative branch length estimates. Deviation in rate estimates result immediately in underestimated branch length, while substantial effects of too flat site-specific preferences only manifest themselves once deviations are of order $\gamma = 0.3$. If the same preferences are used for every site ($\gamma = 1$), however, deviations are substantial. Deviations in rate and preferences are approximately additive. These observations are consistent with the finding above that using inferred preferences and true substitution rates result in more accurate branch length estimates than inferring both rates and preferences.

Figure 4B shows the degree of mis-estimation when using flat preferences, flat rates, or both as a function of the degree of divergence. The relative error increases approximately linearly with the average evolutionary rate.

2.5 Applications to large HIV alignments

Since its zoonosis, HIV-1 has diversified into several different subtypes that differ from each other at about 20 per cent of sites

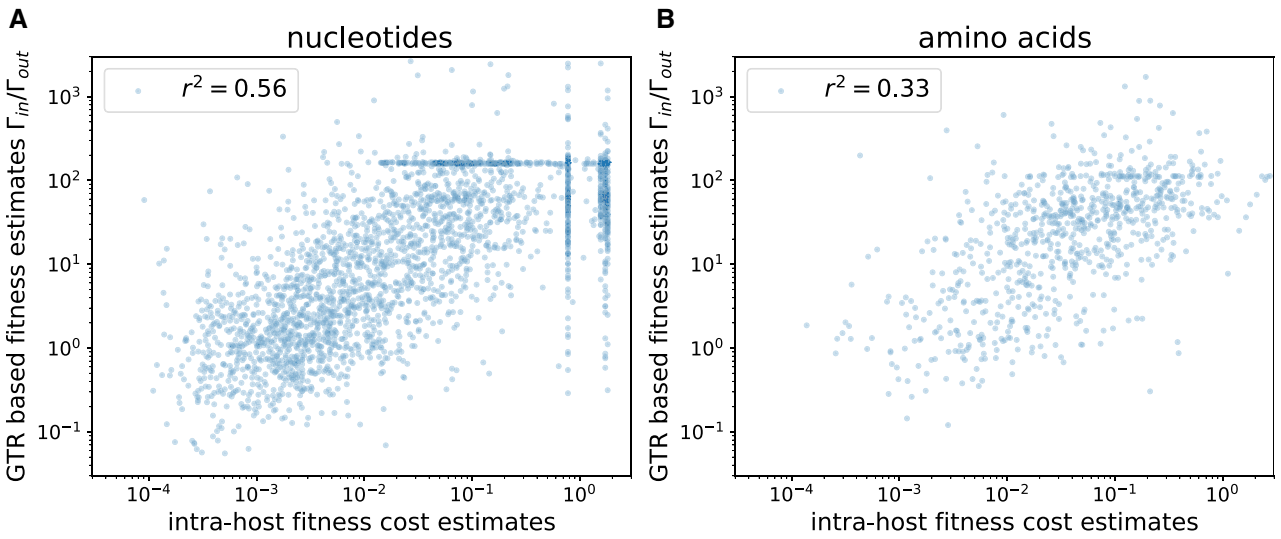

Figure 5. Intra-host vs cross-sectional mutation–selection balance. Panels A and B show the ratio of in/out rates for consensus nucleotides/amino acids along the *pol* of HIV-1 subtype B vs of fitness costs of non-consensus states estimated from with-inhost mutation–selection balance. The logarithm of the rate ratio is roughly linear in the logarithm of the fitness cost. Analogous results for the genes *gag* and *nef* are shown in Supplementary Fig. S3.

in their genome. For each subtype, thousands of sequences are available, which typically differ by about 10 per cent (Los Alamos HIV Sequence Database 2017). This large sample of moderately diverged sequences should be suited to estimate site-specific preference using the iterative inference frame work and quantify how selection shaped the evolution of HIV.

Simple population genetic models predict that the fixation probability f_{ij}^a of a mutation from state j to state i at site a should depend on the fitness difference s_{ij}^a and the effective population size N as (Kimura 1964)

$$f_{ij}^a = \frac{1 - e^{-2s_{ij}^a}}{1 - e^{-2Ns_{ij}^a}} \approx \begin{cases} 2s_{ij}^a & s_{ij}^a > 0 \\ 2s_{ij}^a e^{2Ns_{ij}^a} & s_{ij}^a < 0 \end{cases}. \quad (4)$$

On longer time scales, the fixation rates f_{ij}^a correspond to transitions rates of the site-specific GTR model Q_{ij}^a. The effective population size N plays the role of a coalescent time scale T_c and in general has little to do with census population sizes, in particular in rapidly adapting populations (Neher 2013). Nevertheless, the logarithm of the ratio $\log f_{ij}^a / f_{ji}^a \approx 2Ns_{ij}^a$ is an interpretable quantity related to the average fitness difference between states on time scales longer than the population genetic scale $N \sim T_c$. We generalize this notion to multiple states and define a fitness score of state i at position a as the ratio of rates into and out of state i

$$e^{2Ns_i^a} = \frac{\Gamma_{in}}{\Gamma_{out}} = \frac{\sum_j \pi_i^a W_{ij}}{\sum_j W_{ij} \pi_j^a}. \quad (5)$$

The logarithm of this ratio is expected to be proportional to the fitness difference between state i and the average alternative state at site a, while the common multiplicative factor $2N$ is unknown.

We downloaded alignments of HIV-1 pol sequences from the LANL data base, constructed phylogenetic trees, and inferred site-specific GTR models at the nucleotide and amino acid level, see Section 3. We compared $2Ns_i^a = \log(\Gamma_{in}/\Gamma_{out})$ to fitness costs measured using mutation–selection balance models and within host diversity data of HIV (Zanini et al. 2017). However, instead of a linear relationship between s_i^a and with-in host estimates of fitness costs, we observed a linear relationship between the logarithm of fitness effects measured with-in host and s_i^a (see Figs 5 and 6) This relationship explains about half the variance of nucleotide effects and about one-third of amino acid effects. Interestingly, this correlation between the intra-host estimates and cross-sectional estimates decreased as soon as regularization increased above 0.05 and we therefore used a weak regularization of $c = 0.01$.

The fact that the relationship of fitness proxies deviates from the expectation

$$\text{cross} - \text{sectional} \sim \text{with} - \text{in host}$$

and instead is approximately

$$\text{cross} - \text{sectional} \sim \log(\text{with} - \text{in host})$$

points toward different process that drive cross-sectional and with-in host diversification. While the order of fitness effects with-in host and cross-sectionally seem to be mostly concordant, with-in host variation is much greater than cross-sectional variation. With-in host fitness effects are masked and damped at the population level, possibly due to fluctuating selection by diverse host immune systems or epistasis (Shekhar et al. 2013; Zanini et al. 2015). The ratio of in/out rates $e^{2Ns_i^a}$ is a very good correlate of alignment diversity (see Supplementary Fig. S4).

Next, we quantified the effect of ignoring site-specific preferences on branch length estimates. We generated sequence pairs that evolved for a specific time t under the site-specific model inferred from the HIV-1 pol alignment and then estimated a maximum likelihood branch length between these two sequences. This estimate was done using a model with homogeneous equilibrium frequencies set to the average frequencies across sites while maintaining site-specific rate variation. As soon as the length of the simulated branch approaches 0.2, the length inferred by a model without site-specific preferences deviates substantially from the true value and saturates around 0.5 for larger and larger t. As expected, these deviations are even more severe when estimating branch length as simple sequence

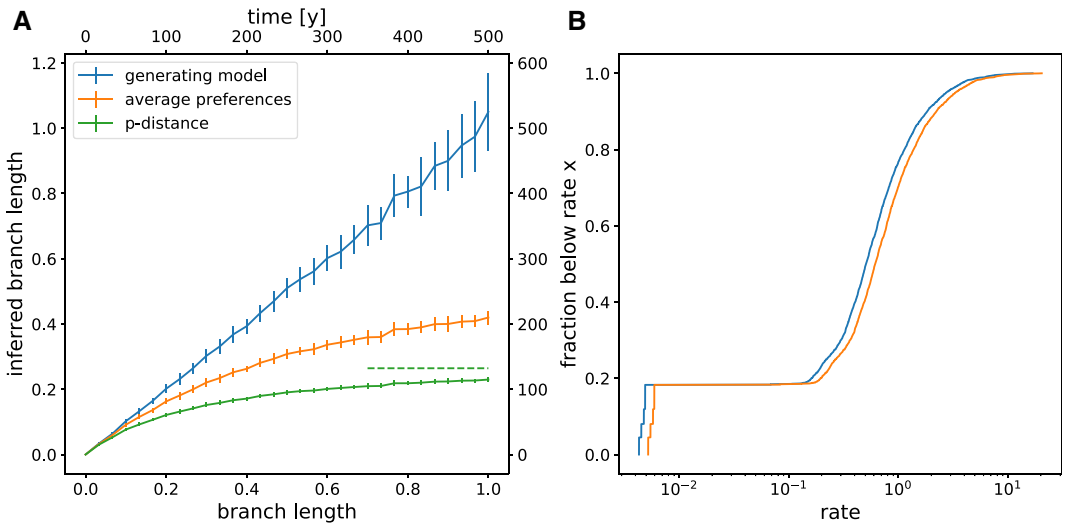

Figure 6. Underestimation of divergence in HIV. Panel A shows the estimated ML branch length for the model used to generate the sequences and a model with constant equilibrium frequencies as a function of true branch length. For comparison, the *p*-distance between the two simulated sequences is also shown. The top axis shows branch length in units of years assuming a substitution rate of 0.002/year and site. Error bars denote one standard deviation. Panel B shows the distribution of rates across sites for both models. About 20 per cent of sites are essentially invariable, while the rates of the remainder vary by at least ten-fold. The distributions differ slightly in the overall scale since rates have been rescaled such that the average substitution rate in both models is identical.

divergence (*p*-distance), while using the generating model reproduces the correct branch length.

Assuming a typical evolutionary rate of HIV of 0.002 changes per site and year, this analysis suggests that length estimates of branches longer than 100 years start to become inaccurate. Furthermore, this analysis suggests very little signal to estimate the length of branches that are longer than 300 years.

3. Discussion

Different positions in a genome sequence or a protein change at different rates and explore different subsets of the available states. These site-specific properties are evolutionary conserved and are the basis of common homology search tools. Which states are permissible at which position can be measured with high-throughput using deep mutational scanning techniques (Fowler and Fields 2014) or can be estimated from large alignments of homologous sequences. Both approaches can reveal biological function and organization of the entity coded for by the sequence. Bruno (1996) pointed out that the frequencies of states in columns of an alignment are distorted by phylogenetic correlations and how these phylogenetic correlations can be accounted for. However, estimating these frequencies accurately requires large data sets that have only recently become available.

While accurate estimates of site-specific frequencies require correction of phylogenetic correlations, accurate phylogenetic inference requires sufficiently realistic models of sequence evolution while being computationally tractable and easy to parameterize. A common compromise is to model rate heterogeneity as random effects, while assuming identical site preferences across the sequence. More complex models allow for a small number of data partitions with different preferences (Lartillot and Philippe 2004; Kainer and Lanfear 2015).

Here, we presented method to estimate site-specific preferences using an iterative EM-type approach inspired by Bruno (1996). We showed that site-specific models can be inferred with an accuracy limited by the Poisson statistics of the substitution process using an efficient iterative scheme. For short branch lengths a parsimonious ancestral reconstruction is sufficient, while for more diverged samples iterative inference of the model, the ancestral states, and the branch lengths is necessary. The computational complexity of the inference is dominated by ancestral reconstruction and branch length optimization, which is quadratic in the alphabet size and linear in sample size and sequence length (Felsenstein 2004). We have implemented the algorithm for site-specific model inference and branch length optimization in TreeTime (Sagulenko, Puller, and Neher 2018).

We further explored how model choice affects the maximum likelihood estimates of branch lengths. As has been reported before, branch lengths are underestimated when variation in rate or preferences are not fully accounted for (Halpern and Bruno 1998; Hilton and Bloom 2018). While such underestimation often has a moderate effect on the total length of a tree, it can result in substantial underestimation of root-to-tip distance that are dominated by a few long branches close to the root. While using the correct model results in correct branch length estimates, joint inference of site-specific models and branch lengths is typically unable to recover the true branch lengths and substantial underestimation remains. These points to parameter identifiability problems rooted in the similar effects that skewed equilibrium frequencies, low rates, and shorter branch length have on the likelihood of the data. These problems are related to the inability to jointly infer ancestral states and rate parameters of evolving discrete states (Gascuel and Steel 2020). Model and tree inference alone, however, does not require acurate reconstruction of ancestral states as these can be marginalized.

Branch length estimates using models constructed by mixing the true model and flat Jukes–Cantor type models showed that misspecified equilibrium frequencies result in substantial errors as soon as the model deviates from the true model by 30 per cent or more. This mirrors observations by Hilton and Bloom (2018), who found that preferences measured for influenza HA proteins of type H1 or H3 affect branch length of in the vicinity of the focal sequence, but not globally on the tree. Using site-specific models inferred from HIV alignments, we

quantified the error made when ignoring site-specific preferences. While the true models are unknown in this case, this analysis nevertheless suggests that errors are substantial as soon as branch length exceed $t = 0.2$ (~100 years) and sequence divergence saturates at levels around 0.3. This result is consistent with the discrepancy between molecular clock-based and biogeography-based estimates of the divergence times of different SIV lineages (Worobey et al. 2010): Beyond a few hundred years, there is very little signal to estimate branch length—in particular when the underlying site-specific model parameters need to be estimated themselves. This lack of information is further underscored by the average nucleotide distances between *pol* sequences of different HIV and SIV strains: HIV-1 subtypes differ at about 10 per cent of sites, distances between pol and HIV-1 and SIVcpz are on the order of 25 per cent and distances between sequences in the HIV-1/HIV-2/SIV compendium alignment are about 35 per cent (all distances ignore sites with >20% gaps). These observations are compatible with evidence for frequent reversion of HIV to a preferred sequence state following immune escape (Leslie et al. 2004; Carlson et al. 2014; Zanini et al. 2015).

This potentially large error in branch length estimates can seriously affect deep divergence time estimates. Typically, short branches close to the tips of the tree are used to calibrate molecular clock models. The deep branches, however, tend to be much longer and rate estimation and variation of site-specific preferences will result in saturation effects not accounted for by the model (Hilton and Bloom 2018). This effective variation in rate is related to the effect of transient deleterious mutations that inflate the rate on short time scales (Ho et al. 2005; Wertheim and Kosakovsky Pond 2011): the time it takes purifying selection to prune deleterious mutations is related to the relaxation time scales of GTR models with site-specific preferences. Instead of the $q-1$ degenerate eigenvalues of a Jukes–Cantor model, each site has a spectrum of eigenvalues and different eigenmodes relax at different speeds, generating apparently time-dependent rates.

Estimating site-specific models from sequence alignments faces one fundamental problem: Reliable estimates require observation of many changes at each site which requires many sufficiently diverged sequences. At the same time, preferences at individual sites are expected to change due to epistatic interactions with other sites in the sequence, as for example witnessed by the gradual divergence of experimentally determined preferences (Doud, Ashenberg, and Bloom 2015; Haddox et al. 2018). Hence the approach described in this work is largely restricted to cases like HIV where many moderately diverged sequences (~10%) are available. Outside of this limit there either is not enough data to reliably estimate the large number of coefficients, or epistatic interactions need to be taken into account—probably at the expense of ignoring phylogenetic signals (Morcos et al. 2011).

4. Materials and methods

4.1. Model and notation

Most models of sequence evolution express the probability that sequence $\vec{s}$ evolved from sequence $\vec{r}$ in time t as

$$P(\vec{s} \leftarrow \vec{r}, t, \mathbf{Q}) = \prod_{a=1}^{L} (e^{\mathbf{Q}^a t})_{s^a, r^a}, \tag{6}$$

where $\mathbf{Q}^a$ is the substitution matrix governing evolution at site a, and s^a and r^a are the sequence states at position a. The product runs over all L sites a and amounts to assuming that different sites of the sequence evolve independently.

In absence of recombination, homologous sequences are related by a tree and the likelihood of observing an alignment $\mathbf{A} = \{\vec{s_k}, k = 1 \ldots n\}$ conditional on the tree T and the substitution model $\mathbf{Q}^a$ can be written in terms of propagators defined in Equation (6). It is helpful to express this likelihood as product of sequence propagators defined in Equation (6) between sequences at the ends of each branch in the tree (implicitly assuming that evolution on different branches is independent and follows the same time reversible model). Unknown sequences of internal nodes $\{\vec{s}'\}$ need to be summed over and the likelihood can be expressed as

$$\ell(\mathbf{A}|T, \mathbf{Q}) = \sum_{\{\vec{s}'\}} \prod_{a=1}^{L} p^a_{s^a_0} \prod_{k \in T} P(\vec{s}_c \leftarrow \vec{s}_p, t, \mathbf{Q}) = \sum_{\{\vec{s}\}} e^{\ell(\{\vec{s}\}|T, \mathbf{Q})}, \tag{7}$$

where $\vec{s}_c$ and $\vec{s}_p$ are the child and parent sequences of branch k, respectively, and the factor $\prod_a p^a_{s^a_0}$ is the product of the probabilities of the root sequence s^a_0 over all positions a. The probabilities $\vec{p}^a$ are the equilibrium probabilities of the substitution model at position a. The latter ensures that the likelihood is insensitive to a particular choice of the tree root. This equation defines the log-likelihood ℓ of a particular internal node assignment $\{\vec{s}'\}$ which is given by

$$\ell(\mathbf{A}, \{\vec{s}'\}|T, \mathbf{Q}) = \sum_a [\log(p^a_{s^a_0}) + \sum_{k \in T} \log(e^{\mathbf{Q}^a t_k})_{s^a_c, s^a_p}], \tag{8}$$

where s^a_c and s^a_p are indices corresponding to the child and parent sequence of branch k.

The sum over unknown ancestral sequences can be computed efficiently using standard dynamic programming techniques. Nevertheless, it requires $\mathcal{O}(n \times L \times q^2)$ operations (where q is the size of the alphabet $\mathcal{A}$) and optimizing it with respect to a large number of parameters is costly. Our goal here is to infer site-specific substitution models using a computationally efficient iterative procedure.

Instead of inferring completely independent models for every site in the genome, we follow Halpern and Bruno (1998) and only allow for site-specific rates and equilibrium frequencies while using the same transition matrix for every site. Such a site-specific GTR model, can be parameterized as:

$$\begin{aligned} Q^a_{ij} &= \mu^a p^a_i W_{ij} \text{for} i \neq j, \\ Q^a_{ii} &= -\sum_k Q^a_{ki} \end{aligned} \tag{9}$$

where W_{ij} is a symmetric matrix with $W_{ii} = 0$ and the second equation ensures conservation of probability. In addition, we require $\sum_i p^a_i = 1$ and $\sum_{a=1}^{L} \sum_{i \neq j} W_{ij} p^a_i p^a_j = L$ to ensure that the average rate per site is μ^a.

4.2 Iterative optimization algorithm

The derivatives of ℓ with respect to μ^a, p^a_i, and W_{ij} need to vanish at the values that maximize ℓ.

$$\frac{\partial \ell(\mathbf{A}|T, \mathbf{Q})}{\partial X} = \sum_{\{\vec{s}\}} e^{\ell(\{\vec{s}\}|T, \mathbf{Q})} \frac{\partial \ell(\{\vec{s}\}|T, \mathbf{Q})}{\partial X} = 0, \tag{10}$$

where X is one of the parameters we vary.

These conditions can be solved iteratively. Here, we derive the update for μ^a and refer to the supplement for the other update rules. The derivative of the log-likelihood for a specific sequence assignment is given by

$$\frac{\partial \ell(\{\vec{s}\}|T,\mathbf{Q})}{\partial \mu^a} = \sum_{\beta \in T} \frac{t_\beta}{\mu^a} \frac{\sum_i Q^a_{s^a_c,i} (e^{Q^a t_\beta})_{i,s^a_p}}{(e^{Q^a t_\beta})_{s^a_c,s^a_p}}, \tag{11}$$

where s^a_p and s^a_c are the states at site a at the parent and child end of branch β. The individual terms in this sum behave very differently for cases where $s^a_c = s^a_p$ (no change at site a on branch β) and $s^a_c \neq k^a_p$ (at least one mutation). In the limit of short branches $\mu^a t_\beta \ll 1$, we can expand the matrix exponential $e^{Qt} = \delta_{ij} + \mathbf{Q}t + \cdots$ to obtain approximate but solvable conditions for maximum likelihood parameter estimates, see supplement. We will separate the sum over branches into those with $s^a_c = s^a_p$ and $s^a_c \neq s^a_p$. Suppressing the index a for the position in the sequence, we find

$$\begin{aligned}\frac{d}{d\mu}\ell(\vec{s})\,|T,\mathbf{Q}) \approx & -\sum_{\beta\in T, s_c=s_p} \frac{t_\beta \sum_{k\neq s_c} p_k W_{ks_c}}{1 - t_\beta \mu \sum_{k\neq s_c} p_k W_{ks_c}} + \sum_{\beta\in T, s_c\neq s_p} \frac{t_\beta p_{s_c} W_{s_c s_p}}{t_\beta \mu p_{s_c} W_{s_c s_p}} \approx \\ & - \sum_{\beta\in T, s_c=s_p} t_\beta \sum_{k\neq s_c} p_k W_{ks_c} + \sum_{\beta\in T, s_c\neq s_p} \frac{1}{\mu} \\ & = -\sum_{j,k\neq j} p_k W_{kj} \tau^a_j + \sum_{i\neq j} n^a_{ij}/\mu, \end{aligned} \tag{12}$$

where τ^a_j is the sum of all branch length along which site a is in state j and n^a_{ij} is the number of times the sequence at site a changes from j to i along branches of the tree (we have re-instantiated the position index a in the last line). Additional terms necessary for regularization and normalization are discussed in the supplement. Setting this expression to zero (and the corresponding ones in the supplement) suggests solving for μ^a at fixed τ^a_j and n^a_{ij} using the iterative update rules given in Equation (2). The quantities n^a_{ij} and τ^a_i can be averaged over unknown ancestral states.

4.3 Implementation

We extended our package TreeTime (Sagulenko, Puller, and Neher 2018) to handle site-specific GTR models as defined in Equation (1) by adding an additional class GTR_site_specific that generalizes existing the GTR class. Since these classes have an almost identical interface and can be used interchangeably in other analysis run by TreeTime. Using the new class, TreeTime can generate sequences with site-specific evolutionary models and infer these models from sequence data using the algorithms above. In the future, this could extended to also allow site-specific W_{ij} but this is not implemented as of now.

4.4 Generation of simulated data

To generate models and sequence ensembles for which the ground truth is known, we sampled binary trees with $n = 100, 300, 1,000, 3,000$ leaves and Yule tree branch length statistics using betatree (Neher, Kessinger, and Shraiman 2013) (last accessed 10 December 2019) for values of the average substitution rate given by $\langle\mu\rangle = [0.005, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.35, 0.5]$.

Given these tree, we generated multiple sets of sequences of length $L = 1{,}000$ for each tip of the tree. Specifically, we explored the effect of model choice and realization of the evolutionary process by generating data sets as follows: For each tree, we sampled two site-specific models for sequences of length $L = 1{,}000$ from following distribution: Site-specific rates μ^a were sampled iid from a Gamma distribution with parameter $\alpha = 1.5$ or 3.0. Site-specific preferences p^a_i (or equilibrium probabilities) were sampled from Dirichlet distributions with parameters $\alpha = 1$ for alphabets with $q = 4$ states (nucleotides) or $\alpha = 0.2$ and $\alpha = 0.5$ for alphabets with $q = 20$ (amino acids). These distributions correspond to an average number of effective states of $N_{\text{eff}} = 2.6$ (nucleotides) and 4.7 or 7.8 (amino acids), where the number of effective states is given as $(\sum_i (p^a_i)^2)^{-1}$. The entries of the transition matrix W_{ij} were sampled from a Dirichlet distribution with $\alpha = 2$. The average substitution rate of these models was fixed to the required $\langle\mu\rangle$. In addition, we generated one set of models ($q = 4$) with rate variation but uniform p^a_i and W_{ij}. For each combination of tree and model, two sets of sequences were evolved using the sequence generation function of TreeTime.

In total, this amounts to eighty alignments for each of the data set sizes $n = 100, 300, 1,000, 3,000$ and four different ensembles. For each alignment, we reconstructed phylogenetic trees using IQ-tree (Nguyen et al. 2015) with a GTR+R10 model or FastTree (Price, Dehal, and Arkin 2009) using the default twenty category model for nucleotide and amino-acid sequences, respectively. These trees were used to infer site-specific models from data using TreeTime, see below. The exact workflow is documented in the script src/generate_toy_data.py in the associated git repository at github.org/neherlab/2019_Puller_SiteSpecificGTR (last accessed 27 July 2020).

4.5 Model inference from simulated data

Simulated data and trees (reconstructed or true) were read in by TreeTime and models reconstructed using functions of TreeTime to infer models with the different approximations discussed in the text. The exact workflow is documented in the script src/reconstruct_toy_data.py in the associated git repository at github.org/neherlab/2019_Puller_SiteSpecificGTR (last accessed 27 July 2020).

4.6 HIV sequence analysis

HIV-1 sequences were downloaded from LANL HIV database (Los Alamos HIV Sequence Database 2017) setting filters to 'one sequence per patient', 'non-ACGT < 0.3'. Separate downloads where made for sequences the following sets: *pol*, subtype B (5 May 2019); *pol*, subtype C (13 May 2019); *gag*, *nef*, subtype B (7 June 2019).

Sequences were aligned to the HXB2 reference sequences using mafft (Katoh and Standley 2013), phylogenies were inferred using IQ-tree (Nguyen et al. 2015), and ancestral sequences were inferred using TreeTime (Sagulenko, Puller, and Neher 2018) via the nextstrain's augur pipeline (Hadfield et al. 2018). The different steps were assembled into a pipeline using the workflow manager Snakemake (Koster and Rahmann 2012).

The sequences and trees were then used for site-specific GTR inference as implemented in TreeTime (Sagulenko, Puller, and Neher 2018). The scripts detailing this analysis and producing the figures are available on GitHub in repository github.org/neherlab/2019_Puller_SiteSpecificGTR (last accessed 27 July 2020).

Acknowledgments

We are grateful to Sarah Hilton and Pierre Barrat-Charlaix for stimulating discussions and insightful comments on the manuscript.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

Funding

This work was supported by core funding by the University of Basel, the Max Planck society, and ERC Stg 260686.

References

Bloom, J. D. (2014) 'An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit', Molecular Biology and Evolution, 31: 1956–78.

Bruno, W. J. (1996) 'Modeling Residue Usage in Aligned Protein Sequences via Maximum Likelihood', Molecular Biology and Evolution, 13: 1368–74.

Carlson, J. M et al. (2014) 'Selection Bias at the Heterosexual HIV-1 Transmission Bottleneck', Science, 345: 1254031.

Doud, M. B., Ashenberg, O., and Bloom, J. D. (2015) 'Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs', Molecular Biology and Evolution, 32: 2944–60.

Felsenstein, J. (2004) Inferring Phylogenies. Sunderland, MA: Sinauer Associates.

Fowler, D. M., and Fields, S. (2014) 'Deep Mutational Scanning: A New Style of Protein Science', *Nature Methods*, 11: 801–7.

Gascuel, O., and Steel, M. (2020) 'A Darwinian Uncertainty Principle', Systematic Biology, 69: 521–9.

Haddox, H. K. et al. (2018) 'Mapping Mutational Effects along the Evolutionary Landscape of HIV Envelope', *eLife*, 7: 10.7554/eLife.34420.

Hadfield, J. et al. (2018) 'Nextstrain: Real-time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–3.

Halpern, A. L., and Bruno, W. J. (1998) 'Evolutionary Distances for Protein-Coding Sequences: modeling Site- Specific Residue Frequencies', Molecular Biology and Evolution, 15: 910–7.

Hilton, S. K., and Bloom, J. D. (2018) '*Modeling Site-Specific Amino-Acid Preferences Deepens Phylogenetic Estimates of Viral Sequence Divergence*' Virus Evolution, 4:

Ho, S. Y. W et al. (2005) 'Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times', Molecular Biology and Evolution, 22: 1561–8.

Kainer, D., and Lanfear, R. (2015) 'The Effects of Partitioning on Phylogenetic Inference', Molecular Biology and Evolution, 32: 1611–27.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', Molecular Biology and Evolution, 30: 772–80.

Kimura, M. (1964) 'Diffusion Models in Population Genetics', Journal of Applied Probability, 1: 177–232.

Koster, J., and Rahmann, S. (2012) 'Snakemake–A Scalable Bioinformatics Workflow Engine', Bioinformatics, 28: 2520–2.

Lartillot, N., and Philippe, H. (2004) 'A Bayesian Mixture Model for across-Site Heterogeneities in the Amino-Acid Replacement Process', Molecular Biology and Evolution, 21: 1095–109.

Lee, D. D., and Seung, H. S. (2001) 'Algorithms for Non-negative Matrix Factorization', in Leen, T. K., Dietterich, T. G., and Tresp V. (eds) *Advances in Neural Information Processing Systems* 13, pp 556, MIT Press

Leslie, A, J et al. (2004) 'HIV evolution: CTL escape mutation and reversion after transmission', Nature Medicine, 10: 282

Los Alamos HIV Sequence Database. http://www.hiv.lanl.gov (last accessed 2019-06-07)

Mayrose, I., Friedman, N., and Pupko, T. (2005) 'A Gamma Mixture Model Better Accounts for among Site Rate Heterogeneity', Bioinformatics, 21: ii151–8.

Morcos, F et al. (2011) 'Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families', Proceedings of the National Academy of Sciences, 108: E1293–301.

Neher, R. A. (2013) 'Genetic Draft, Selective Interference, and Population Genetics of Rapid Adaptation', Annual Review of Ecology, Evolution, and Systematics, 44: 195–215.

——, Kessinger, T. A., and Shraiman, B. I. (2013) 'Coalescence and Genetic Diversity in Sexual Populations under Selection', Proceedings of the National Academy of Sciences, 110: 15836–41.

Nguyen, L.-T et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', Molecular Biology and Evolution, 32: 268–74.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009) 'FastTree: Computing Large Minimum Evolution Trees with Profiles Instead of a Distance Matrix', *Molecular Biology and Evolution*, 26: 1641–50.

——, ——, and —— (2010) 'FastTree 2—Approximately Maximum-likelihood Trees for Large Alignments', *PLoS One*, 5: e9490.

Rodrigue, N. (2013) 'On the Statistical Interpretation of Site-specific Variables in Phylogeny-based Substitution Models', *Genetics*, 193: 557–64.

Sagulenko, P., Puller, V., and Neher, R. (2018) 'TreeTime: Maximum-likelihood Phylodynamic Analysis', *Virus Evolution*, 4: vex042.

Scheffler, K., Murrell, B., and Pond, S. L. K. (2014) 'On the Validity of Evolutionary Models with Site-specific Parameters', *PLoS One*, 9: e94534.

Shapiro, B., Rambaut, A., and Drummond, A. J. (2006) 'Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-coding Sequences', *Molecular Biology and Evolution*, 23: 7–9.

Shekhar, K. et al. (2013) 'Spin Models Inferred from Patient-derived Viral Sequence Data Faithfully Describe HIV Fitness Landscapes', *Physical Review E*, 88:062705.

Spielman, S. J., and Wilke, C. O. (2016) 'Extensively Parameterized Mutation–Selection Models Reliably Capture Site-specific Selective Constraint', *Molecular Biology and Evolution*, 33: 2990–3002.

Stamatakis, A. (2014) 'RAxML Version 8: A Tool for Phylogenetic Analysis and Post-analysis of Large Phylogenies', *Bioinformatics*, 30: 1312–13.

Tamuri, A. U., dos Reis, M., and Goldstein, R. A. (2012) 'Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation–Selection Models', *Genetics*, 190: 1101–15.

Wertheim, J. O., and Kosakovsky Pond, S. L. (2011) 'Purifying Selection Can Obscure the Ancient Age of Viral Lineages', *Molecular Biology and Evolution*, 28: 3355–65.

Worobey, M. et al. (2010) 'Island Biogeography Reveals the Deep History of SIV', *Science*, 329: 1487.

Yang, Z. (1994) 'Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods', Journal of Molecular Evolution, 39: 306–14.

—— (1996) 'Among-Site Rate Variation and Its Impact on Phylogenetic Analyses', Trends in Ecology & Evolution, 11: 367–72.

——, and Nielsen, R. (2008) 'Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage', Molecular Biology and Evolution, 25: 568–79.

Zanini, F. et al. (2015) 'Population Genomics of Intrapatient HIV-1 Evolution', *eLife*, 4: e11282.

Zanini, F et al. (2017) 'In vivo mutation rates and the landscape of fitness costs of HIV-1' Virus Evolution, 3: vex003.