Avoiding ascertainment bias in the maximum likelihood inference of phylogenies based on truncated data

Asif Tamuri[1] and Nick Goldman[†,1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton CB10 1SD, UK

This version: SNP_203.tex, compiled at 14:58 on 8 September 2017

[†]Corresponding author: goldman@ebi.ac.uk

Abstract

Some phylogenetic datasets omit data matrix positions at which all taxa share the same state. For sequence data this may be because of a focus on single nucleotide polymorphisms (SNPs) or the use of a technique such as restriction site-associated DNA sequencing (RADseq) that concentrates attention onto regions of differences. With morphological data, it is common to omit states that show no variation across the data studied. It is already known that failing to correct for the ascertainment bias of omitting constant positions can lead to overestimates of evolutionary divergence, as the lack of constant sites is explained as high divergence rather than as a deliberate data selection technique. Previous approaches to using corrections to the likelihood function in order to avoid ascertainment bias have either required knowledge of the omitted positions, or have modified the likelihood function to reflect the omitted data. In this paper we indicate that the technique used to date for this latter approach is a conditional maximum likelihood (CML) method. An alternative approach — unconditional maximum likelihood (UML) — is also possible. We investigate the performance of CML and UML and find them to have almost identical performance in the phylogenetic SNP dataset context. We also make some observations about the nucleotide frequencies observed in SNP datasets, indicating that these can differ systematically from the overall equilibrium base frequencies of the substitution process. This suggests that model parameters representing base frequencies should be estimated by maximum likelihood, and not by empirical (counting) methods.

Introduction

Leaché et al. (2015) considered likelihood methods that are available for the phylogenetic analysis of single nucleotide polymorphism (SNP) data sets, i.e. nucleotide (nt) sequence alignments in which any constant sites have been omitted. (Note that the term 'constant site' is used to mean one in which no differences are apparent amongst the sequences collected, and not to suggest that substitution cannot ever occur there.) Starting with an equivalent situation in the analysis of restriction sites (Felsenstein, 1992), then with SNP data (Kuhner et al.,

2000), morphological data (Lewis, 2001) and most recently with restriction site-associated DNA sequencing (RADseq; Baird et al., 2008; Seeb et al., 2011; Peterson et al., 2012), it has been clear that omitting such sites is a form of ascertainment bias. Analyzing only variable sites, without correction, can lead to overestimation of branch lengths and biases in phylogeny inference (Lewis, 2001; Leaché et al., 2015).

Leaché et al. (2015) investigated three likelihood methods for correcting for the omission of constant sites. The first, developed by Felsenstein (1992) and Lewis (2001) and denoted `lewis` by Leaché et al. (2015), uses a conditional likelihood and does not explicitly consider the number of constant sites missing from the data set. The second, described by Kuhner et al. (2000) and denoted `felsenstein` by Leaché et al. (2015), uses a 'reconstituted likelihood' requiring the total number of constant sites to be known, but does not consider their partitioning into constant-A, -C, -G or -T sites. The third method (`stamatakis`: Leaché et al., 2015) again uses reconstituted likelihood, now requiring knowledge of the exact numbers of each of the four different constant site patterns. The result of this approach is necessarily exactly the same as analyzing the original data set, with no sites omitted.

The `felsenstein` and `stamatakis` methods can be used in cases where data (constant sites) are omitted but the numbers of such sites are known — situations that are unlikely to arise with modern data recording techniques. Situations with unknown amounts of omitted data are more frequent, and warrant further attention. In this paper we place the `lewis` method into a more-general likelihood inferential framework and derive a new method for estimating parameters (e.g. tree topologies, branch lengths and substitution model parameters), as well as the number of omitted sites in the case that this is not known. The new method performs almost identically to the `lewis` method, and we explore the reasons for this. Lastly, we make some observations about the effect that SNP data (i.e. missing observations of constant sites) have on observed base frequencies.

Methods

Using a slightly different notation from Leaché et al. (2015) to describe the inference problem, we assume a total of n observations (SNP or non-constant sites), of which n_i correspond to pattern (possible alignment column) i; if we observe l different patterns, then $n = \sum_{i=1}^{l} n_i$. (Sums and products are generally over $i = 1, \ldots, l$ throughout this paper; for simplicity we omit these limits when there is no ambiguity.) For a SNP data set, we do not observe the constant-A, -C, -G or -T patterns; we write the unobserved number of these as m, and the total number of sites (observed and unobserved) as $n + m = N$. Such data sets are described as *truncated* (Blumenthal, 1981).

To complete the description, we need a model describing the probabilities of occurrence of both the observed and unobserved patterns. As usual in likelihood-based phylogenetics, we will assume that an underlying tree structure with branch lengths is to be estimated, along with any free parameters of a substitution model such as JC69 (Jukes and Cantor, 1969), HKY85 (Hasegawa et al., 1985) etc. Representing all the unknowns as the multidimensional parameter θ, this model defines probabilities $p_j = p_j(\theta)$ for every possible pattern j; it is these p_j that are usually calculated using Felsenstein's pruning algorithm (Felsenstein, 1973, 1981). Note in particular that p_j is defined for all possible j, including the unobserved constant patterns and any patterns that happen not to have occurred in a given data set. It is useful to write $c = c(\theta)$ for the total probability of occurrence of a constant site, i.e. $c = \sum_{j \in \mathcal{C}} p_j$, where $\mathcal{C}$ is the set containing the four constant patterns constant-A, -C, -G and -T.

The truncated data likelihood function $L_T(\theta)$ is

$$L_T(\theta) = \prod_i p_i^{n_i} \tag{1}$$

Maximizing $L_T(\theta)$ over the model parameters θ gives their maximum likelihood (ML) estimates, $\widehat{\theta}_T$, based on the truncated data. However, as shown by Lewis (2001) and Leaché et al. (2015), for SNP data sets the omission of the constant characters can cause serious estimation biases.

The problem of estimating θ (and N) in these circumstances was described by Sanathanan (1972). Following that paper, we consider the *complete* likelihood including the contribution of the m omitted constant sites:

$$L(N,\theta) = \frac{N!}{(N-n)!\prod_i n_i!} c^{N-n} \left(\prod_i p_i^{n_i} \right) \tag{2}$$

or, equivalently,

$$L(m,\theta) = \frac{(n+m)!}{m!\prod_i n_i!} c^m L_T(\theta) \tag{3}$$

Note that this retains the combinatorial component $(n+m)!/m!\prod_i n_i!$. In typical ML problems, where all the data are observed, the corresponding term can be omitted as it is a constant and plays no part in the maximization over θ (Edwards, 1972) — hence its omission from eqn. 1. However, in the truncated data case this is not true: different (inferred) values of m will cause the term to vary and its contribution to the likelihood cannot be ignored. Although it is unusual to infer the amount of (unobserved) data using ML, there is no reason why we should not be able to do so.

Notice that the likelihood can also be written as $L(m,\theta) = L_1(m,\theta)L_2(\theta)$ where

$$L_1(m,\theta) = \frac{(n+m)!}{n!m!}(1-c)^n c^m \tag{4}$$

and

$$L_2(\theta) = \frac{n!}{\prod_i n_i!} \prod_i \left(\frac{p_i}{1-c} \right)^{n_i} = \frac{n!}{(1-c)^n \prod_i n_i!} L_T(\theta) \tag{5}$$

L_1 is the likelihood based on the probability of n, and L_2 is the likelihood based on the conditional probability of the n_i given n (Sanathanan, 1972).

Conditional ML: Sanathanan (1972) describes two approaches to estimating m and θ (see also Blumenthal, 1981). The first is the method of *conditional ML* (CML), in which the conditional likelihood $L_2(\theta)$ (eqn. 5) is maximized over θ to find ML model parameter estimates $\widehat{\theta}_C$. (Note that the combinatorial term $n!/\prod_i n_i!$ is constant and does not affect the inference.) This corresponds to precisely the method of Felsenstein (1992) and Lewis (2001), and is equivalent

to maximizing $\log L_C(\theta)$ over θ, where

$$\log L_C(\theta) = \log L_T(\theta) + \delta_C(\theta) \tag{6}$$

and $\delta_C(\theta) = -n \log(1 - c(\theta))$ is the log-likelihood 'correction' term that changes the truncated data set problem into the CML problem.

In the phylogenetics context we may only be interested in the inferred tree and associated parameters $\widehat{\theta}_C$. If required, however, an estimate of the number of unobserved constant sites comes from maximizing $L_1(m, \widehat{\theta}_C)$ over m to find $\widehat{m}_C$. Sanathanan (1972) shows that $\widehat{m}_C = \lfloor n\widehat{c}_C/(1 - \widehat{c}_C) \rfloor$ where $\widehat{c}_C = c(\widehat{\theta}_C)$ is the CML estimator of c and $\lfloor . \rfloor$ indicates the floor function (i.e. $\lfloor x \rfloor$ is the greatest integer $\leq x$).

The CML version of the SNP data set phylogeny problem is implemented in `RAxML` v.8 (Stamatakis, 2014), invoked using the `--asc-corr=lewis` option (Leaché et al., 2015).

Unconditional ML: The second approach described by Sanathanan (1972) is *unconditional ML* (UML), in which the full likelihood $L(m, \theta)$ (eqn. 3) is maximized simultaneously over both m and θ. We denote the corresponding inferred values by $\widehat{m}_U$ and $\widehat{\theta}_U$.

For any fixed value θ^*, optimization of $L(m, \theta^*)$ over m (eqn. 3) is analogous to optimizing $L_1(m, \theta^*)$ over m (eqn. 4) and is achieved when $\widehat{m}^* = \lfloor nc^*/(1 - c^*) \rfloor$ (Sanathanan, 1972). Substituting $m^*(\theta) = \lfloor nc(\theta)/(1 - c(\theta)) \rfloor$ into eqn. 3 and recalling that the n_i are constant means the UML problem becomes one of maximizing $\log L_U(\theta)$ over θ, where

$$\log L_U(\theta) = \log L_T(\theta) + \delta_U(\theta) \tag{7}$$

and $\delta_U(\theta) = \log(n + m^*(\theta))! - \log m^*(\theta)! + m^*(\theta) \log c(\theta)$ is the correction term that changes the truncated problem into the UML problem.

Sanathanan (1972, 1977) proves that $\widehat{m}_U < \widehat{m}_C$ and $\widehat{c}_U < \widehat{c}_C$, and that the asymptotic distributions of $(\widehat{m}_C, \widehat{\theta}_C)$ and $(\widehat{m}_U, \widehat{\theta}_U)$ are the same. In other words, as the amount of data collected increases, the CML and UML estimators will give arbitrarily close estimates of the

numbers of unobserved constant sites and model parameters. However, the approaches do not necessarily lead to equally good estimates given *finite* amounts of data.

Results and Discussion

CML and UML both perform well: We implemented the CML and UML methods for the SNP data set phylogeny problem in order to compare their performance. We modified the `baseml` software (Yang, 2007) so that for each candidate value of θ we first compute $c(\theta)$ and $m^*(\theta)$ and then use these with the truncated log-likelihood function $\log L_T(\theta)$ to compute $\log L_C(\theta)$ and $\log L_U(\theta)$ as in eqns. 6 and 7.

We simulated sequence data on the 10-taxon topology studied by Leaché et al. (2015). To create a range of realistic simulation scenarios, we scaled the tree to various lengths (scaling factors of 0.25, 0.5, 1, 2 and 4, giving tree lengths of 0.08, 0.16, 0.31, 0.62 and 1.24, respectively) and used a variety of alignment sizes ($N = 500$, 1000, 2500, 5000) under the JC69 model. After simulation, all constant site patterns were discarded. The probability of occurrence of constant sites ranged from 93% to 30% ($c = 0.93$, 0.86, 0.74, 0.54, 0.30 for scaling factors 0.25, 0.5, 1, 2, 4, respectively). For more extreme cases, inspired by population resequencing studies, we also considered scaling factors 0.03–0.21, giving rise to tree lengths 0.009–0.065 and c from 0.99–0.94, with $N = 100000$.

Lower scale factors lead to smaller trees and thus result in fewer variable (SNP) sites on which to base inference. Our simulations cover a range, from unobserved constant sites being rare (e.g. distantly related species, or sequencing strategies such as RADseq giving strong enrichment for variable sites) to common (e.g. closely related organisms). Our most extreme scenario ($c = 0.99$) resembles what might be observed with SNP data sets from population sequencing studies.

We used the CML and UML correction methods to re-estimate model parameters using only the variable sites from the simulated datasets, assuming knowledge of the true topology. We repeated this procedure with data simulated under the HKY85 model with moderate transition/

transversion rate and nucleotide bias ($\kappa = 2$, $\pi_T = 0.1, \pi_C = 0.2, \pi_A = 0.3, \pi_G = 0.4$). In all cases, we found that estimates of branch lengths, tree length and other model parameters were almost identical with CML and UML. Figs. 1 and 2 illustrate this with a variety of results summarized from 100 simulations in each scenario; other results (not shown) all confirm these findings.

Why are CML and UML almost the same? Clearly, this is because in every case studied, the maximal values of $\log L_C(\theta)$ and $\log L_U(\theta)$ are attained at very similar values of θ. As these likelihoods differ only in the terms $\delta_C(\theta)$ and $\delta_U(\theta)$, we would like to study how these vary with θ. However, as this represents a complex multidimensional parameter (Yang et al., 1995), it is difficult to visualize likelihoods as θ varies over candidate solutions. To simplify our investigation, we give an illustration using a single JC69 simulation with the 10-taxon tree of Leaché et al. (2015) with scaling factor 1 and $N = 1000$ alignment sites simulated before removal of constant patterns. For these data, we found the truncated likelihood-optimal branch lengths $\widehat{\theta}_T$ and the CML-optimal branch lengths $\widehat{\theta}_C$, and focus attention on values of θ formed by interpolating between $\widehat{\theta}_T$ and $\widehat{\theta}_C$ and extrapolating this range beyond $\widehat{\theta}_T$ and $\widehat{\theta}_C$. This gives a one-dimensional subspace which includes both $\widehat{\theta}_T$ and $\widehat{\theta}_C$. Fig. 3 shows values of $\log L_T(\theta)$ (eqn. 1), $\log L_C(\theta)$, $\delta_C(\theta)$ (eqn. 6), $\log L_U(\theta)$ and $\delta_U(\theta)$ (eqn. 7) for these values of θ, with the x-axis simultaneously labelled by the corresponding values of $c(\theta)$ and $m^*(\theta)$.

Firstly, note the truncated likelihood (indicated by the solid black line) is maximized at a value of θ suggesting far too few omitted constant sites ($c(\widehat{\theta}_T) = 0.3$ and $m^*(\widehat{\theta}_T) = 100$, whereas the true values of c and m^* for this simulation are 0.743 and 743, respectively). This corresponds to a tree that is too divergent. The correction terms $\delta_C(\theta)$ and $\delta_U(\theta)$ (dashed blue and orange lines, respectively), while very different in absolute value, have very similar gradients. Consequently, when they are added to $\log L_T(\theta)$ to form $\log L_C(\theta)$ and $\log L_U(\theta)$ (solid blue and orange lines, respectively), these likelihoods have maxima at very similar values of θ (with $c(\widehat{\theta}_C)$ and $c(\widehat{\theta}_U)$ equal to 0.69, and $m(\widehat{\theta}_C)$ and $m(\widehat{\theta}_U) \approx 570$). In brief, the effects of the corrections $\delta_C(\theta)$ and $\delta_U(\theta)$ in the optimization of $\log L_C(\theta)$ and $\log L_U(\theta)$ are indeed very similar.

It is not the absolute values of δ_C and δ_U that are critical to the difference between $\widehat{\theta}_C$ and

$\widehat{\theta}_U$, but how they vary in the regions of the CML and UML optima. Further analysis might consider the derivatives of δ_C and δ_U with respect to θ, but in the SNP phylogeny question this is complicated by the non-standard form of the topology parameter θ (Yang et al., 1995) and by the factorial and floor functions in δ_U. However, δ_C and δ_U only depend on θ through the probability of constant patterns $c(\theta)$ and so an alternative approach is to consider the relative variation of δ_C and δ_U as c varies. In particular, if $\delta_C(c_1) - \delta_C(c_0)$ and $\delta_U(c_1) - \delta_U(c_0)$ are very similar for any $c_0 \approx c_1$ (and in particular near to $c(\widehat{\theta}_C)$ and $c(\widehat{\theta}_U)$), then the CML and UML correction terms will have similar effects on the optimizations of $\log L_C$ and $\log L_U$, leading to similar CML and UML estimates.

Indeed, it can be shown (omitted for brevity) that if c_0 and c_1 are chosen to be similar, such that the corresponding values m_0^* and m_1^* satisfy $m_1^* = m_0^* + 1$, then

$$[\delta_C(c_1) - \delta_C(c_0)] - [\delta_U(c_1) - \delta_U(c_0)] \approx \frac{(1-c_0)^2}{2nc_0} \tag{8}$$

This represents a measure of the difference between the gradients $d\delta_C/dm$ and $d\delta_U/dm$ (Fig. 3); since it scales as $1/n$, $\delta_C(\theta)$ and $\delta_U(\theta)$ will be expected to have very similar effects for reasonably large values of n, which will be the case for most phylogenetic problems.

Base frequencies can behave unexpectedly in SNP datasets: While analyzing simulated datasets as described above, we noticed the observed frequencies of nucleotides A, C, G and T in the HKY simulations did not always match the corresponding model parameters. We realized this is because in those models where different nucleotides have different substitution rates, the constant-A, -C, -G and -T site patterns have different probabilities of occurrence. As a consequence, the observed frequencies of A, C, G and T amongst the constant site patterns omitted from SNP datasets, and amongst the SNP patterns retained for analysis, will differ from the model parameter values. This is illustrated in Fig. 4, using the HKY simulation scenario described above.

As a consequence, it may be advisable to estimate base frequencies using ML rather than the simple counting (empirical) method when working with SNP data (Goldman, 1993). We have

used this approach throughout this paper.

Conclusions

In all of our analyses of simulated data using the CML and UML approaches for removing ascertainment bias from SNP datasets, we have found virtually no difference in the results obtained. We explain this by observing that the different approaches' respective correction terms δ_C and δ_U behave very similarly in their effects (Fig. 3). This has further been supported by our analysis of the gradients of δ_C and δ_U, which confirms their similarity for plausible phylogenetic scenarios and data quantities. Although we have not investigated tree topology estimation, the near-identical results of CML and UML for estimation of other parameters, including branch lengths, lead us to think it very unlikely that they could behave differently for topology estimation. We therefore conclude that it is of little importance which of these methods is used in practice in phylogenetic studies. The CML method is widely available via the `--asc-corr=lewis` option of RAxML (Stamatakis, 2014; Leaché et al., 2015).

We note in passing that the observed base frequencies in SNP datasets can differ systematically from the corresponding substitution model parameters, due to bias in the frequency with which constant-nucleotide site patterns arise and are thus omitted. A simple solution to this should be to use ML to estimate these parameters (Goldman, 1993).

Acknowledgments

We thank Joe Felsenstein and Alexis Stamatakis for discussions of an earlier (incorrect) attempt at this study, Jakub Truszkowski for helpful discussions of that attempt and then the CML and UML methods, and Melissa Ward for alerting us to issues of base frequency estimation with SNP datasets.

References

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376.

Blumenthal, S. 1981. A Survey of Estimating Distributional Parameters and Sample Sizes from Truncated Samples Pages 75–86. Springer Netherlands, Dordrecht.

Edwards, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge.

Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Biology 22:240–249.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

Felsenstein, J. 1992. Phylogenies from restriction sites: a maximum-likelihood approach. Evolution 46:159–173.

Goldman, N. 1993. Statistical tests of models of dna substitution. Journal of Molecular Evolution 36:182–198.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22:160–174.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of Protein Molecules. Pages 21–132 *in* Mammalian Protein Metabolism (H. N. Munro, ed.). Academic Press, New York.

Kuhner, M. K., P. Beerli, J. Yamato, and J. Felsenstein. 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics 156:439–447.

Leaché, A. D., B. L. Banbury, J. Felsenstein, A. Nieto-Montes de Oca, and A. Stamatakis. 2015. Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. Systematic Biology 64:1032–1047.

Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50:913–925.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS ONE 7:e37135.

Sanathanan, L. 1972. Estimating the size of a multinomial population. Annals of Mathematical Statistics 43:142–152.

Sanathanan, L. 1977. Estimating the size of a truncated sample. Journal of the American Statistical Association 72:669–672.

Seeb, J. E., G. Carvalho, L. Hauser, K. Naish, S. Roberts, and L. W. Seeb. 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Resources 11 (Suppl. 1):1–8.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24:1586–1591.

Yang, Z., N. Goldman, and A. Friday. 1995. Maximum likelihood trees from dna sequences: A peculiar statistical estimation problem. Systematic Biology 44:384–399.

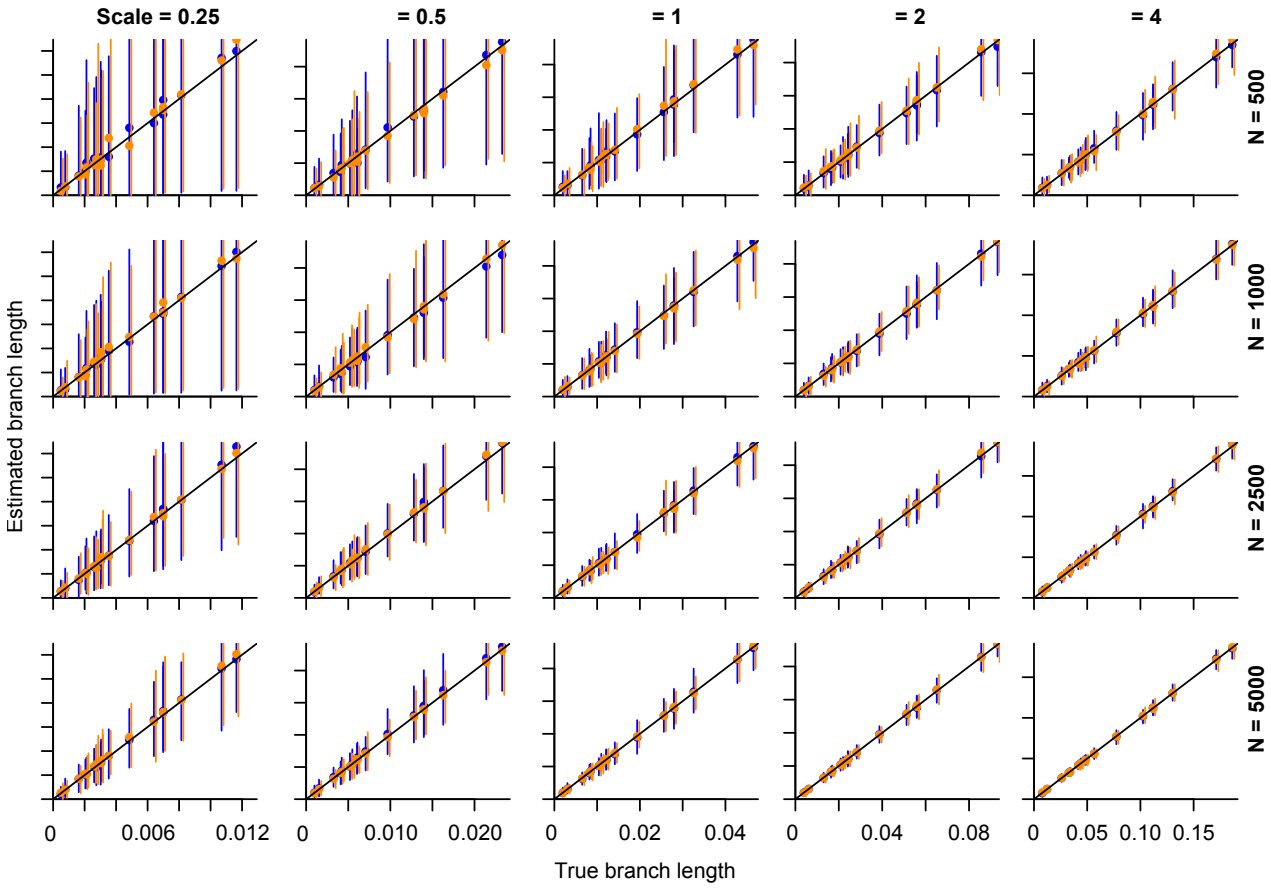

Figure 1: Branch length estimates under various JC69 simulation scenarios. Within each row, scenarios have the same number of sites simulated (N, i.e. before constant sites were removed); within columns, the same tree length scaling factor. Graphs show the mean and 5–95%-ile range for each of the 17 branch length estimates plotted against the true value, derived from 100 simulations. CML results are shown in blue; UML in orange. The two methods' results are essentially indistinguishable.

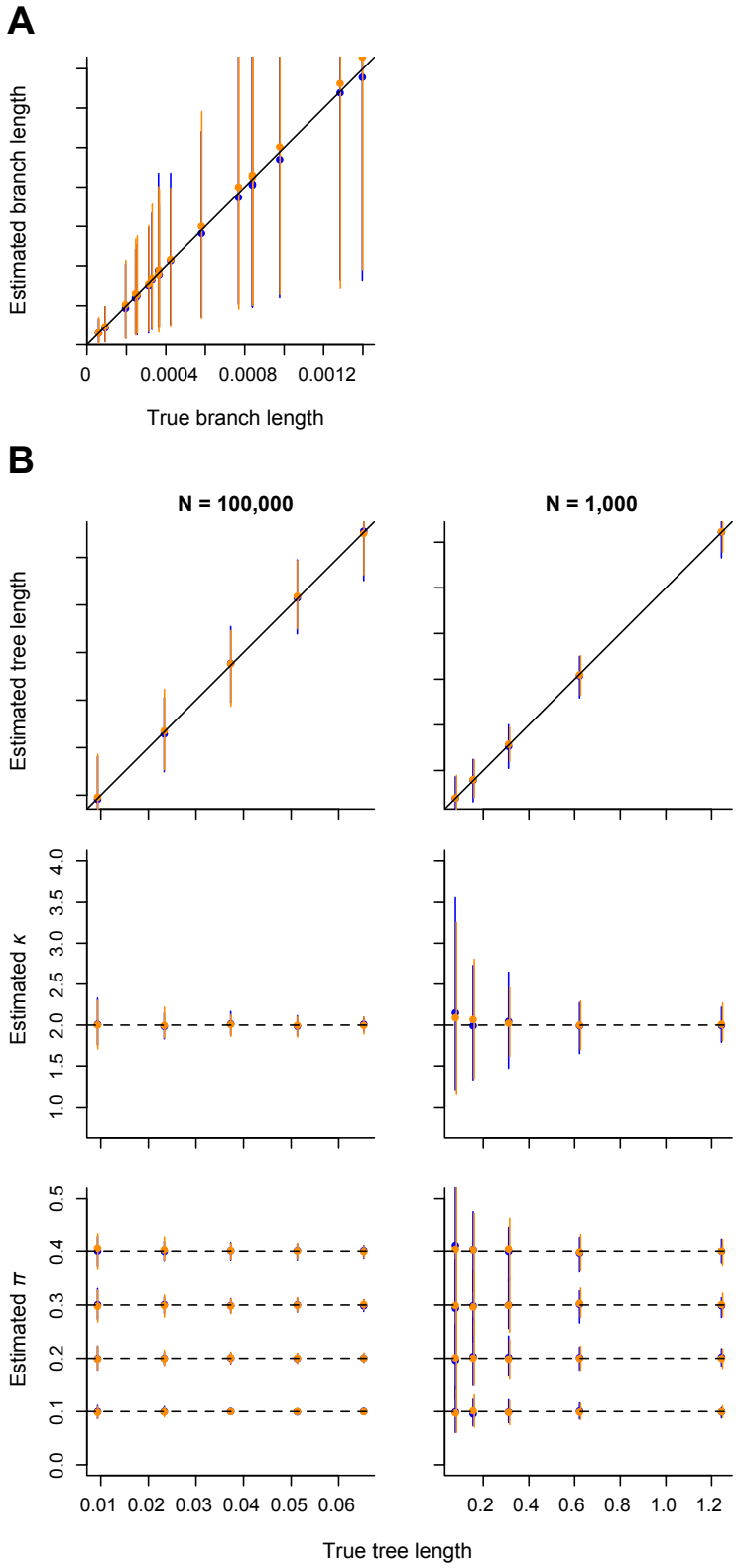

Figure 2: Parameter estimates (means and 5–95%-ile ranges from 100 simulations in each case) under various simulation scenarios. **A:** Branch length estimates from JC69 simulations with scaling factor 0.03 (tree length 0.09; $c = 0.99$) and $N = 100000$. **B:** Parameter estimates from HKY85 simulations. Left column: $N = 100000$, c from 0.99–0.94; right column: $N = 1000$, c from 0.93–0.30. Graphs show estimates of overall tree length (top), κ (middle) and nucleotide frequencies (bottom) for various tree length scaling scenarios. Colours etc. as for Fig. 1. Again, CML and UML give essentially the same results.

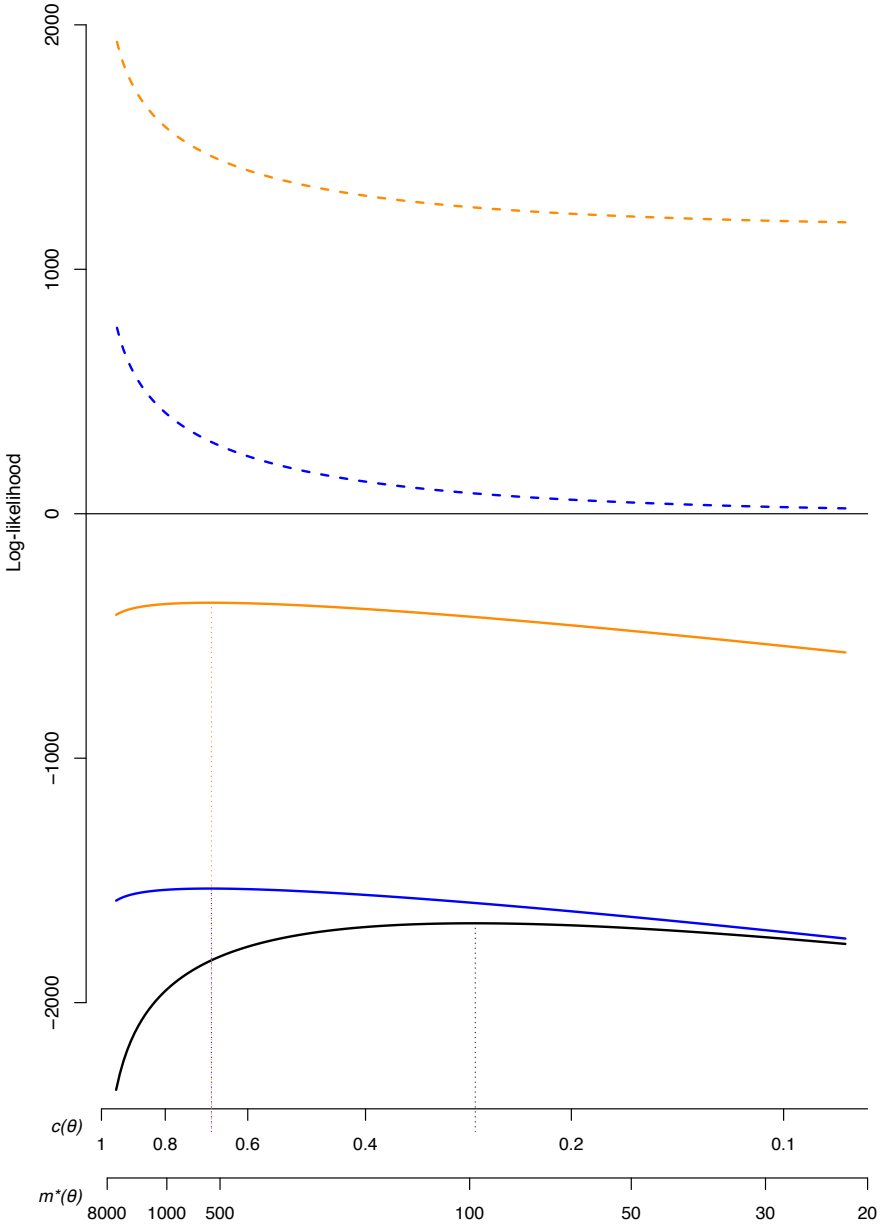

Figure 3: Likelihoods and correction terms in a single simulation case. From bottom to top, plots show $L_T(\theta)$ (solid black line), $L_C(\theta)$ (solid blue), $L_U(\theta)$ (solid orange), $\delta_C(\theta)$ (dashed blue) and $\delta_U(\theta)$ (dashed orange). The x-axis is labelled with the values of $c(\theta)$ and $m^*(\theta)$ corresponding to the range of branch length parameters θ described in the text. Locations of $\widehat{\theta}_T$, $\widehat{\theta}_C$ and $\widehat{\theta}_U$ are indicated by vertical dotted lines.

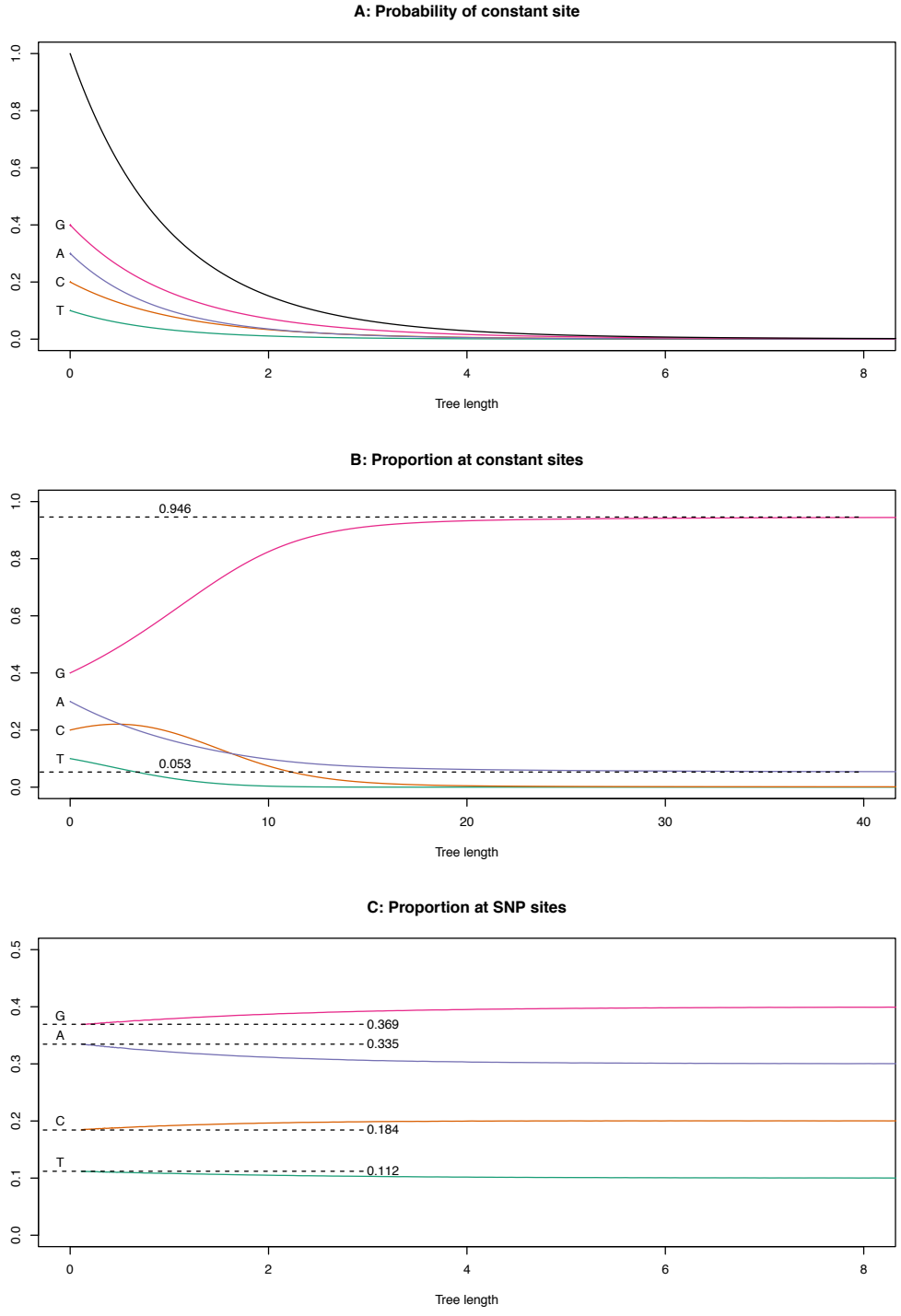

Figure 4: Base frequencies observed with SNP datasets. **A:** As the 10-taxon tree used throughout this study is scaled from a length of 0 to increasingly large branch lengths, the proportion of constant sites (top, black) falls to 0; the component proportions of constant-A, -C, -G and -T patterns falls from their equilibrium values to 0. **B**: Conditional on observing a constant pattern, the proportions of nucleotides A, C, G and T vary as the tree size increases. (Note that these proportions need not vary monotonically, as observed for the constant-C patterns in this example.) **C:** Conditional on observing a non-constant (SNP) pattern, the observed base frequencies also differ from the model parameters for shorter tree lengths.