Version dated: April 28, 2015

Supplement: TODO insert final title

Adam D. Leaché[1,2,*], Barbara L. Banbury[1], Joseph Felsenstein[1,3], Adrián Nieto-Montes de Oca[4], and Alexandros Stamatakis[5,6]

[1] *Department of Biology, University of Washington, Seattle, WA 98195, USA;*

[2] *Burke Museum of Natural History and Culture, University of Washington, Seattle, WA 98195, USA;*

[3] *Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA;*

[4] *Depto. de Biología Evolutiva, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad Universitaria, México 04510, Distrito Federal, México;*

[5] *Scientific Computing Group, Heidelberg Institute for Theoretical Studies (HITS gGmbH), Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany;*

[6] *Institute for Theoretical Informatics, Department of Informatics, Karlsruhe Institute of Technology, Am Fasanengarten 5, 76131 Karlsruhe, Germany.*

Abstract.— In this supplement we describe the ascertainment bias correction models we developed in RAxML in more detail. Apart from describing the equations we also provide some implementation details and hints such that it can easily be integrated into other likelihood-based (ML and Bayesian) tools.

Initially, we outline the subtle difference between invariable and invariant sites. As invariable sites, we denote those sites that simply and truly do not vary. As invariant sites we denote those sites that are observed not to vary, but could potentially also be variable. Hence the set of invariable sites is a subset of the set of invariant sites.

First, we describe the standard ascertainment bias correction (conditional likelihood correction; RAxML flag `--asc-corr=lewis`) as described by Lewis (2001) which we extend for the DNA alphabet.

The log likelihood for an alignment with $i = 1...n$ sites under this model is corrected as follows:

$\ln(L) = \sum \ln(L_i) - n \cdot \ln(1.0 - c)$.

Here, $\sum \ln(L_i)$ is simply the standard phylogenetic log likelihood without ascertainment bias correction. The expression $-n \cdot \ln(1.0 - c)$ is the correction factor. For DNA data, c is defined as follows: $c = L(A) + L(C) + L(G) + L(T)$. Here, $L(A), L(C), ...$ represent the likelihood (not the log likelihood!) of so-called dummy sites consisting entirely of As, Cs, etc. for the tree, branch lengths, and model parameters on which we intend to compute $\ln(L)$. The above equation is obtained by calculating $L_i/(1.0 - c)$ for each SNP site. This corresponds to the probability of the pattern observed at site i, conditional on that site *not* being invariant.

Note that, for mathematical reasons, all sites $i = 1...n$ need to be variable. That is, if the alignment on which we compute $\ln(L)$ also contains some invariant sites, the above equation may return positive log likelihood values. We wish to thank Derrick Zwickl for pointing this out. Thus, we recommend that programs implementing an ascertainment bias correction first verify that all sites of an alignment (or partition) for which the user intends to use the above correction are indeed variable.

In addition, this standard correction does not allow to use values $n' > n$ in the correction term (i.e., $-n' \cdot \ln(1.0 - c)$) because the log likelihood score will become positive for some choice of $n' > n$. Furthermore, it is counter-intuitive to include additional $(1.0 - c)$ correction denominators than there are SNP sites. Also note that, this correction does not allow to explicitly incorporate the observed number of invariant sites when this number is known as for our empirical data. To achieve this, we need to deploy a different correction, the reconstituted DNA approach.

The first correction that allows for this is the reconstituted likelihood approach (Kuhner et al. (2000), McGill et al. (2013); RAxML flag `--asc-corr=felsenstein`) and is calculated as follows:

$$\ln(L) = \sum \ln(L_i) + w \cdot \ln(c)$$

Here, the correction c is defined as above and w can be any integer value > 0 (in particular $w > n$ *is* allowed) that corresponds to the known/true number of invariant sites in the data. Once again, the data used for computing $\sum \ln(L_i)$ should not contain any invariant sites.

This correction does not take into account that the count $cnt(A), cnt(C), ...$ of invariant sites in the input alignment consisting entirely of As, Cs, etc. may, in fact, be known as well. When it *is* known, we can slightly refine the reconstituted likelihood correction (RAxML flag `--asc-corr=stamatakis`) as follows:

$$\ln(L) = \sum \ln(L_i) + cnt(A) \cdot \ln(L(A)) + cnt(C) \cdot \ln(L(C)) + cnt(G) \cdot \ln(L(G)) + cnt(T) \cdot \ln(L(T))$$

Note that, this is essentially equivalent to the calculations we would conduct on a full alignment (including invariant sites) when using the standard alignment site pattern compression technique. In this case, $cnt()$ would just correspond to the occurrence count of invariant sites in the input alignment. However, all of these corrections assume that the correction likelihoods $(L(A), L(C), ...)$ are calculated on sites consisting entirely of As, Cs, etc. That is, they do not take into account that some, essentially, invariant sites have missing data, e.g., sites such as `AA??` or `AAA?`. An extension of the reconstituted likelihood corrections that takes into account missing data patterns is subject of future work.

For details on using the three aforementioned corrections please refer to the RAxML manual.

The implementation of the corrections requires modifications of the likelihood functions for calculating (i) conditional probability vectors at nodes, (ii) the overall log likelihood at the virtual root of the tree, and (iii) first and second derivatives of the likelihood used for the Newton-Raphson procedure to optimize branch lengths. Thus, the implementation can be inspected in the following three RAxML sources files: `newviewGenericSpecial.c`, `evaluateGenericSpecial.c`, and `makenewzGenericSpecial.c`. The relevant parts of the code can be identified by applying `grep ascBias`.

$*$

References

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.

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

McGill, J. R., E. A. Walkup, and M. K. Kuhner. 2013. Correcting coalescent analyses for panel-based SNP ascertainment. Genetics 193:1185–1196.