πBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios

Filip Bielejec*[1], Philippe Lemey[1], Luiz Max Carvalho[2], Guy Baele[1], Andrew Rambaut[3] and Marc A. Suchard[4,5]

[1]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium
[2]Program for Scientific Computing (PROCC), Fundação Oswaldo Cruz, Rio de Janeiro, Brazil
[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom
[4]Departments of Biomathematics and Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA, 90095, USA
[5]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, 90095, USA

Email: Filip Bielejec*- filip.bielejec(at)rega.kuleuven.be;

*Corresponding author

Abstract

Background: Simulated nucleotide or amino acid sequences are frequently used to assess the performance of phylogenetic reconstruction methods. BEAST, a Bayesian statistical framework that focuses on reconstructing time-calibrated molecular evolutionary processes, supports a wide array of evolutionary models, but lacked matching machinery for simulation of character evolution along phylogenies.

Results: We present a flexible Monte Carlo simulation tool, called πBUSS, that employs the BEAGLE high performance library for phylogenetic computations within BEAST to rapidly generate large sequence alignments under complex evolutionary models. πBUSS sports a user-friendly graphical user interface (GUI) that allows combining a rich array of models across an arbitrary number of partitions. A command-line interface mirrors the options available through the GUI and facilitates scripting in large-scale simulation studies. Analogous to BEAST model and analysis setup, more advanced simulation options are supported through an extensible markup language (XML) specification, which in addition to generating sequence output, also allows users to combine simulation and analysis in a single BEAST run.

Conclusions: πBUSS offers a unique combination of flexibility and ease-of-use for sequence simulation under realistic evolutionary scenarios. Through different interfaces, πBUSS supports simulation studies ranging from modest endeavors for illustrative purposes to complex and large-scale assessments of evolutionary inference procedures. The software aims at implementing new models and data types that are continuously being developed as part of BEAST/BEAGLE.

(Keywords: Simulation, Phylogenetics, BEAST, BEAGLE, evolution, Monte Carlo)

Background

Recent decades have seen extensive development in phylogenetic inference, resulting in a myriad of techniques, each with specific properties concerning evolutionary model complexity, inference procedures and performance both in terms of speed of execution and estimation accuracy. With the development of such phylogenetic inference methods comes the need to synthesize evolutionary data in order to compare estimator performance and to characterize strengths and weaknesses of different approaches. Whereas the true underlying evolutionary relationships between biological sequences are generally unknown, Monte Carlo simulation allows generating test scenarios while controlling for the evolutionary history as well as the tempo and mode of evolution. This has been frequently used to compare the performance of tree topology estimation (e.g. [1]), but it also applies to evolutionary parameter estimation and ancestral reconstruction problems (e.g. [2]). In addition, Monte Carlo sequence simulation has proven useful for assessing model adequacy (e.g. [3]) and for testing competing evolutionary hypotheses (e.g. [4]). It is therefore not surprising that several general sequence simulation programs have been developed (e.g. Seq-Gen [5]), but also inference packages that do not primarily focus on tree reconstruction, such as PAML [6] and HyPhy [7], maintain code to simulate sequence data under the models they implement.

As a major application of phylogenetics, estimating divergence times from molecular sequences requires an assumption of roughly constant substitution rates throughout evolutionary history [8]. Despite the restrictive nature of this molecular clock assumption, its application in a phylogenetic context has profoundly influenced modern views on the timing of many important events in evolutionary history [9]. Following a long history of applying molecular clock models on fixed tree topologies, the Bayesian Evolutionary Analysis by Sampling Trees (BEAST) package [10] fully integrates these models, including more realistic relaxed clock models [11, 12], in a phylogenetic inference framework. By focusing on time-measured evolutionary histories, BEAST has grown increasingly popular as a general framework for historical inference of evolutionary and population genetic processes. The software integrates substitution models, molecular clock models, tree-generative (coalescent or birth-death) models and trait evolutionary models in a modular fashion, allowing the user to combine different parameterizations for each module. The ability to perform joint inference over a flexible combination of models and data types overcomes the pitfall of traditional stepwise inference procedures that purge intermediate estimates (e.g. a tree topology) of their uncertainty, but greatly complicates performance assessment. Here, we address this by introducing a parallel BEAST/BEAGLE utility for sequence simulation (πBUSS).

Implementation

We develop several sequence simulation procedures that balance between ease-of-use and accessibility to model complexity. On the one hand, the core simulation routine is implemented in the BEAST software package [10] and simulations can be performed following specifications in an extensible markup language (XML) input file (Figure 1 A). This procedure provides the most comprehensive access to the BEAST arsenal of models, but may require custom XML editing. On the other hand, πBUSS also represents a stand-alone package that conveniently wraps the simulation routines in a user friendly graphical user interface (GUI), allowing users to set up and run simulations by loading input, selecting models from drop-down lists, setting their parameter values, and generating output in different formats (Figure 1 B). To facilitate scripting, πBUSS is further accessible through a command-line interface (CLI), with options that mirror the GUI. The simulation routines are implemented in Java and interface with the Broad-platform Evolutionary Analysis General Likelihood Evaluator (BEAGLE) high-performance library [13] through its application programming interface (API) for computationally intensive tasks.

[Figure 1 about here.]

Program input

Similar to standard evolutionary analyses using BEAST, the core implementation of the software can be invoked by loading an XML file with simulation settings. The simulation procedure requires a user-specified tree topology or a set of taxa with their heights (inversely proportional to their sampling time) for which a tree topology can be simulated using a coalescent model. Setting all heights to 0 would be equivalent to contemporaneously-sampled taxa. In πBUSS, such a tree can be loaded in NEXUS or NEWICK format, or a taxa list can be set-up in the Data panel for subsequent coalescent simulation of the genealogy. Creating the latter is further facilitated by the ability to load a tab-delimited file with a set of taxa and their corresponding heights. The input tree or taxon list can also be specified through the command-line interface of πBUSS.

Program output

πBUSS generates sequence output in FASTA format but it also supports XML output of the simulation settings. Not only does the XML provide a record of the settings, but the file can be loaded in BEAST for sequence simulation. In this sense, πBUSS is analogous to the BEAUti tool that generates XML input for BEAST analysis [10], and a πBUSS-generated XML file specifying standard models can serve as a template for editing more complex simulations. Moreover, the XML file can also be edited in order to directly analyze the generated sequence data, which avoids writing the sequence data to a file.

Models of evolution

πBUSS is capable of generating trees from a list of taxa using simple coalescent models, including a constant population size or exponential growth model. The software supports simulation of nucleotide, amino acid and codon data along the simulated or user-specified phylogeny using standard substitution models. For nucleotide data, the Hasegawa, Kishino and Yano model (HKY; [14]), the Tamura Nei model (TN93; [15]) and the general time-reversible model (GTR; [16]) can be selected from a drop-down list, and more restrictive continuous-time Markov chain (CTMC) models can be specified by tailoring parameters values. Coding sequences can be simulated using Goldman and Yang's model of codon evolution (GY94; [17]), which is parameterized in terms of a non-synonymous and synonymous substitution rate ratio (dN/dS or ω) and a transition/transversion rate ratio (κ). Several empirical amino acid substitution models are implemented, including the Dayhoff [18], JTT [19], BLOSUM62 [20], WAG [21] and LG [22] model. Equilibrium frequencies can be specified for all substitution models as well as among-site rate heterogeneity through the widely-used discrete-gamma distribution [23] and proportion of invariant sites [24].

An important feature of πBUSS is the ability to set up an arbitrary number of partitions for the sequence data and associate independent substitution models to them. Such settings may reflect codon position-specific evolutionary patterns or approximate genome architecture with separate substitution patterns for coding and non-coding regions. Partitions may also be set to evolve along different phylogenies, which could be used, for example, to investigate the impact of recombination or to assess the performance of recombination detection programs in specific cases. Finally, partitions do not need to share the exact same taxa (e.g. reflecting differential taxon sampling), and in partitions where a particular taxon is not represented the relevant sequence will be padded with gaps.

Inspired by the BEAST framework, πBUSS is equipped with the ability to simulate evolutionary processes on trees calibrated in time units. Under the strict clock assumption, this is achieved by specifying an evolutionary rate parameter that scales each branch from time units into substitution units. πBUSS also supports branch-specific scalers drawn independently and identically from an underlying distribution (e.g. log normal or inverse Gaussian distributions), modeling an uncorrelated relaxed clock process [11]. Simulations do not need to accommodate an explicit temporal dimension and input trees with branch lengths in substitution units will maintain these units with the default clock rate of 1 (substitution/per site/per time unit).

The data types and models described above are available through the πBUSS GUI or CLI, but additional data types and more complex models can be specified directly in an XML file. This allows, for example, simulating any discrete trait, e.g. representing phylogeographic locations, under reversible and nonreversible models [25, 26], with potentially sparse CTMC matrices [25], as well as simulating a combination of sequence data and such traits. As an example of available model extensions is the ability to specify different CTMC matrices over different time intervals of the evolutionary history, allowing for example to model changing selective constraints through different codon model parameterizations or seasonal migration processes for viral phylogeographic traits [27].

Results and Discussion

We have developed a new simulation tool, called πBUSS, that we consider to be a rejuvenation of Seq-Gen [5], with several extensions to better integrate with the BEAST inference framework. Compared to Seq-Gen and other simulation software (Table 1), πBUSS covers a relatively wide range of models while, similar to Mesquite, offering a cross-platform, user-friendly GUI. πBUSS is implemented in the Java programming language, and therefore requires a Java runtime environment, and depends on the BEAGLE library. Although speed is unlikely to be an impeding factor in some simulation efforts, the core implementation using the BEAGLE library provides substantial increases in speed for large-scale simulations, in particular when invoking multi-core architecture to produce highly partitioned synthetic sequence data.

Program validation

We validate πBUSS in several ways. First, we compare the expected site probabilities, as calculated using tree pruning recursion [34], with the observed counts resulting from πBUSS simulations. To this purpose, we calculate the probabilities for all 4^3 possible nucleotide site patterns observed at the tips of a particular 3-taxon topology using an HKY model with a discrete gamma distribution to model rate variation among sites. We then compare these probabilities to long-run ($n = 100,000$) site pattern frequencies simulated under this model and observe good correspondence in distribution (Pearson's χ^2 test, $p = 0.42$).

We also perform simulations over larger trees and estimate substitution parameters (e.g. κ in the HKY model) using BEAST for a large number of replicates. Not only do the posterior mean estimates agree very well with the simulated values, but we also find close to nominal coverage, and relatively small bias and variance (mean squared error). These good performance measures have also recently been demonstrated for more complex substitution processes [27].

Example application

We illustrate the use of simulating sequence data along time-calibrated phylogenies to explore the limitations of estimating old divergence times for rapidly-evolving viruses. Wertheim and Kosakovsky Pond [35] examine the evolutionary history of Ebola virus from sequences sampled over the span of three decades. Although maintaining remarkable amino acid conservation, the authors estimate nucleotide substitution rates on the order of 10^{-3} substitutions/per site/per year and a correspondingly recent time to most recent common ancestor (tMRCA) of about 1,000 years. These estimates suggest a strong action of purifying selection to preserve amino acid residues over longer evolutionary time scales, which may not be accommodated by standard nucleotide substitution models. The authors demonstrate that accounting for variable selective pressure using codon models can result in substantially older origins in such cases.

Here, we explore the effect of temporally varying selection pressure throughout evolutionary history on estimates of the tMRCA using nucleotide substitution models. In particular, we model a process that is characterized by increasingly stronger purifying selection as we go further back in to time. To this purpose, we set up an 'epoch model' that specifies different GY94 codon substitution processes along the evolutionary history [27], and parameterize them according to a log-linear relationship between time and ω. Specifically, we let the process transition from $\omega =$ 1.0, 0.2, 0.1, 0.02, 0.01, 0.002, and 0.001 at time = 10, 50, 100, 500, 1000 and 5000 years in the past, respectively. We simulate a constant population size genealogy of 50 taxa, sampled evenly during a time interval of 25 years, and simulate sequences according to the time-heterogeneous codon substitution process with a constant clock rate of 3×10^{-3} codon substitutions/codon site/year. By producing 100 replicates for various population sizes (1, 5, 10, 50, 100, 500 and 1000), the sequences are simulated over genealogies with different tMRCAs. We note that under this model, trees with tMRCAs of about 10,000 years still result in sequences with a noticeable degree of homology (resulting in a mean amino acid distance of about 0.5, which is in the same range of the mean amino acid distance for sequences representative of the primate immunodeficiency virus diversity). Using a constant ω of 0.5 on the other hand results in fairly randomized sequences. We subsequently analyze the replicate data using a codon position partitioned nucleotide substitution model in BEAST and plot the correspondence between simulated and estimated tMRCAs in Figure 2.

[Figure 2 about here.]

Our simulation exercise shows that a linear relationship between simulated and estimated tMRCAs only holds for 100 to 200 years in the past, and estimates quickly level off after about 1000 years in the past. This can be explained by the unaccounted decline in amino acid substitutions and saturation of the synonymous substitutions as we go further back in time. Although we are not claiming that evolution occurs quantitatively or even qualitatively according to the particular process we simulate under, and we ignore other confounding factors (such as potential selective constraints on non-neutral synonymous sites), this simulation does conceptualize some of the limitations to estimating ancient origins for rapidly evolving viruses that experience strong purifying selection over longer evolutionary time scales.

Conclusion

πBUSS provides simulation procedures under many evolutionary models or combinations of models available in the BEAST framework. This feature facilitates the evaluation of estimator performance during the development of novel inference techniques and the generation of predictive distributions under a wide range of evolutionary scenarios that remain critical for testing competing evolutionary hypotheses. Combinations of different evolutionary models can be accessed through a GUI or CLI, and further extensions can be specified in XML format with a syntax familiar to the BEAST user community. Analogous to the continuing effort to support model set-up for BEAST in BEAUti, future releases of πBUSS aim to provide simulation counterparts to the BEAST inference tools, both in terms of data types and models. Interesting targets include discrete traits, which can already be simulated through XML specification, continuously-valued phenotype data [36] and indel models. Finally, πBUSS provides opportunities to pursue further computational efficiency through parallelization on advancing computing technology. We therefore hope that πBUSS will further stimulate the development of sequence and trait evolutionary models and contribute to advancement of our knowledge about evolutionary processes.

Availability and requirements

The parallel BEAST/BEAGLE utility for sequence simulation is licensed under the GNU Lesser GPL and its source code is freely available as part of the BEAST GoogleCode repository:
http://code.google.com/p/beast-mcmc/
Compiled, runnable package targeting all major platforms along with a tutorial and supplementary data are hosted at: http://rega.kuleuven.be/cev/ecv/software/pibuss.
πBUSS requires Java runtime environment version 1.5 or later to run its executables and the BEAGLE library to be present and configured to fully utilize its capabilities.

Authors' contributions

FB designed and implemented the software. Large portions of code extend or are based on BEAST interfaces designed and developed by AR and MAS. PL and GB conceived the original idea and helped with the design of the software. PL designed the simulation study employing the software. LMC wrote the software tutorial, tested and commented on the software. All authors contributed to the writing of this manuscript.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864, the Welcome Trust (grant no. 092807) and the National Institutes of Health (R01 GM086887 and R01 HG006139) and the National Science Foundation (IIS 1251151 and DMS 1264153). The National Evolutionary Synthesis Center (NESCent) catalyzed this collaboration through a working group (NSF EF-0423641). LMC would like to thank the Program for Scientific Computing staff for operational support.

References

1. Stamatakis A: **An efficient program for phylogenetic inference using simulated annealing**. In *Parallel and Distributed Processing Symposium, 2005. Proceedings. 19th IEEE International* 2005.

2. Blanchette M, Diallo AB, Green ED, Miller W, Haussler D: **Computational reconstruction of ancestral DNA sequences**. *Methods Mol Biol* 2008, **422**:171–84.

3. Brown JM, ElDabaje R: **PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy**. *Bioinformatics* 2009, **25**(4):537–8.

4. Goldman N: **Statistical tests of models of DNA substitution**. *J Mol Evol* 1993, **36**(2):182–98.

5. Rambaut A, Grass NC: **Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees**. *Computer applications in the biosciences: CABIOS* 1997, **13**(3):235–238.

6. Yang Z: **PAML 4: Phylogenetic Analysis by Maximum Likelihood**. *Mol. Biol. Evol.* 2007, **24**(8):1586–1591.

7. Pond SLK, Frost SDW, V MS: **HyPhy: hypothesis testing using phylogenies**. *Bioinformatics* 2005, **21**(5):676–679.

8. Zuckerkandl E, Pauling LB: *Molecular disease, evolution, and genetic heterogeneity*, New York: Academic Press 1962 :189–225.

9. Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB: **Estimating divergence times from molecular data on phylogenetic and population genetic timescales**. *Annu. Rev. Ecol. Syst.* 2002, **33**:707–740.

10. Drummond AJ, Suchard MA, Xie D, Rambaut A: **Bayesian phylogenetics with BEAUti and the BEAST 1.7**. *Molecular Biology and Evolution* 2012, **29**(8):1969–1973.

11. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A: **Relaxed phylogenetics and dating with confidence.** *PLoS Biol.* 2006, **4**(5):e88.

12. Drummond A, Suchard M: **Bayesian random local clocks, or one rate to rule them all**. *BMC Biology* 2010, **8**:114.

13. Ayres D, Darling A, Zwickl D, Beerli P, Holder M, Lewis P, Huelsenbeck J, Ronquist F, Swofford D, Cummings M, et al.: **BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics**. *Systematic Biology* 2012, **61**:170–173.

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

15. Tamura K, Nei M: **Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.** *Molecular Biology and Evolution* 1993, **10**(3):512–526.

16. Tavaré S: **Some probabilistic and statistical problems in the analysis of DNA sequences.** *Lectures on Mathematics in the Life Sciences (American Mathematical Society)* 1986, **17**:57–86.

17. Goldman N, Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Molecular Biology and Evolution* 1994, **11**(5):725–736.

18. Dayhoff MO, Schwartz RM: **A model of evolutionary change in proteins**. In *In Atlas of protein sequence and structure*, Citeseer 1978.

19. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences**. *Comput. Appl. Biosci.* 1992, **8**(3):275–282.

20. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks**. *Proc. Natl. Acad. Sci. U.S.A.* 1992, **89**(22):10915–10919.

21. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach**. *Mol. Biol. Evol.* 2001, **18**(5):691–699.

22. Le SQ, Gascuel O: **An improved general amino acid replacement matrix**. *Mol. Biol. Evol.* 2008, **25**(7):1307–1320.

23. Yang Z: **Among-site rate variation and its impact on phylogenetic analyses.** *Trends in Ecology and Evolution* 1996, **11**(9):367–372.

24. Gu X, Fu YX, Li WH: **Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites.** *Molecular Biology and Evolution* 1995, **12**(4):546–557.

25. Lemey P, Rambaut A, Drummond AJ, Suchard MA: **Bayesian Phylogeography Finds Its Roots**. *PLoS Computational Biology* 2009, **5**(9):e1000520.

26. Edwards C, Suchard M, Lemey P, Welch J, Barnes I, Fulton T, Barnett R, O'Connell T, Coxon P, Monaghan N, et al.: **Ancient hybridization and an Irish origin for the modern polar bear matriline**. *Current Biology* 2011, **21**:1251–1258.

27. Bielejec F, Lemey P, Baele G, Rambaut A, Suchard MA: **Inferring Heterogeneous Evolutionary Processes Through Time: from sequence substitution to phylogeography**. *ArXiv e-prints* 2013.

28. Fletcher W, Yang Z: **INDELible: a flexible simulator of biological sequence evolution**. *Mol. Biol. Evol.* 2009, **26**(8):1879–1888.

29. Cartwright RA: **DNA assembly with gaps (Dawg): simulating sequence evolution**. *Bioinformatics* 2005, **21 Suppl 3**:i31–38.

30. Maddison WP, Maddison D: **Mesquite: a modular system for evolutionary analysis** 2011, [http://mesquiteproject.org].

31. Sanderson MJ: **r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock**. *Bioinformatics* 2003, **19**(2):301–302.

32. Stoye J, Evers D, Meyer F: **Rose: generating sequence families**. *Bioinformatics* 1998, **14**(2):157–163.

33. Strope CL, Scott SD, Moriyama EN: **indel-Seq-Gen: a new protein family simulator incorporating domains, motifs, and indels**. *Mol. Biol. Evol.* 2007, **24**(3):640–649.

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

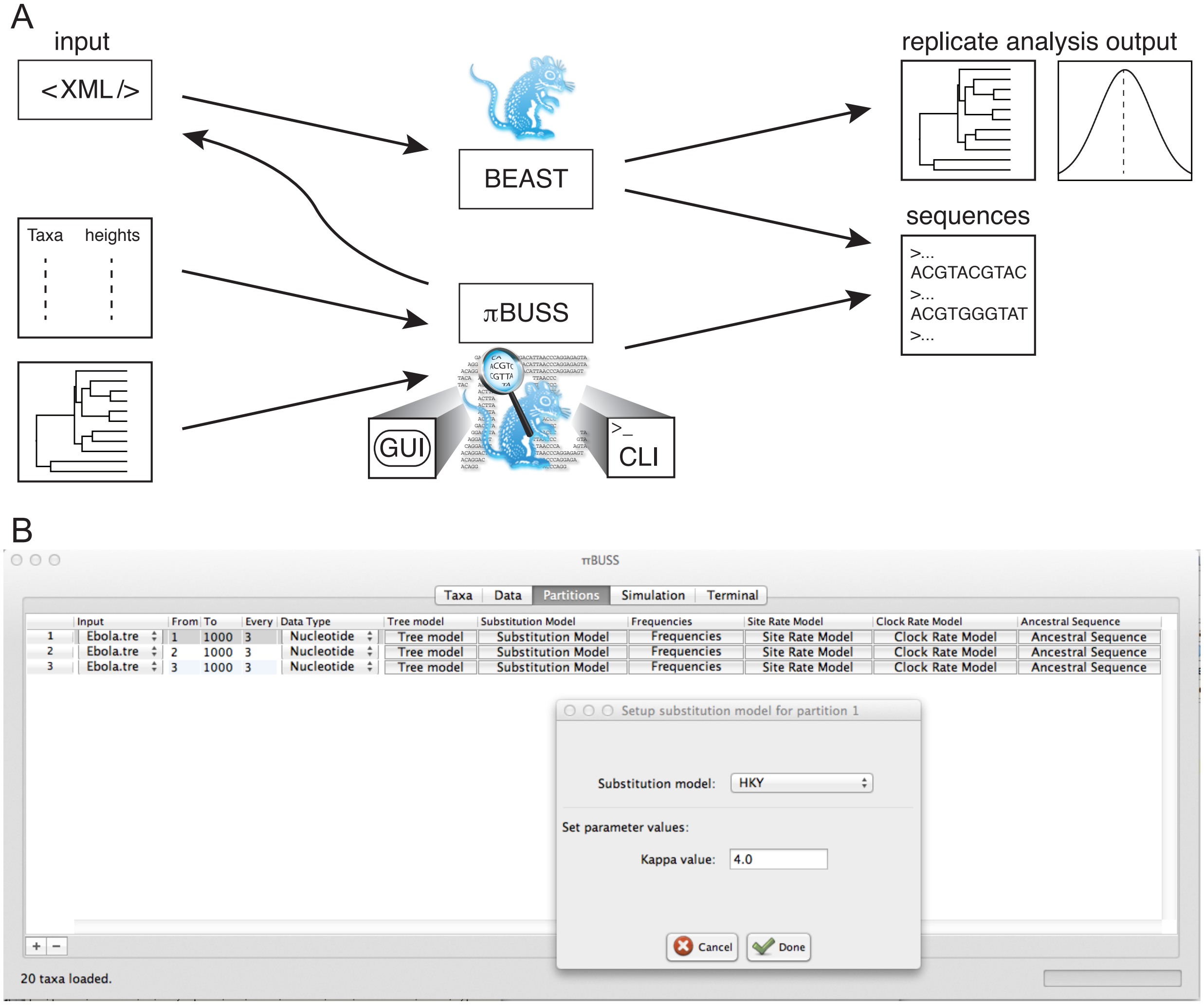
35. Wertheim JO, Kosakovsky Pond SL: **Purifying selection can obscure the ancient age of viral lineages**. *Molecular Biology and Evolution* 2011, **28**(12):3355–3365.

36. Lemey P, Rambaut A, Welch JJ, Suchard MA: **Phylogeography takes a relaxed random walk in continuous space and time**. *Mol Biol Evol* 2010, **27**(8):1877–85.

Table 1: **Comparison between a selection of sequence simulation packages.** We compared the availability of evolutionary modeling options and software implementation aspects of eight commonly used sequence simulation packages. 'X' indicates presence of the feature.

	Feature/Package	πBUSS	Seq-Gen	PhyloSim	Recodon	Indelible [28]	DAWG [29]	Mesquite [30]	r8tes [31]	Rose [32]	Evolver [6]
Evolutionary modeling	**Codon models**	X	X		X			X			X
	Amino acid models	X		X		X		X		X	X
	Indel models		X[1]		X		X	X		X	
	Partitions	X	X	X		X		X			X
	Relaxed clock models	X							X		
	Ancestral sequences[2]			X	X	X					X
	Coalescent simulation	X		X			X	X			X
	Recombination	X	X		X						
Implementation	**Graphical interface**	X						X			
	Multicore	X		X							
	Cross-platform	X	X	X		X	X	X		X	X

[1] Seq-Gen does not include indel simulation but the modified indel-Seq-Gen does [33].
[2] The ability to annotate ancestral characters at internal nodes.



	Input	From	To	Every	Data Type	Tree model	Substitution Model	Frequencies	Site Rate Model	Clock Rate Model	Ancestral Sequence
1	Ebola.tre	1	1000	3	Nucleotide	Tree model	Substitution Model	Frequencies	Site Rate Model	Clock Rate Model	Ancestral Sequence
2	Ebola.tre	2	1000	3	Nucleotide	Tree model	Substitution Model	Frequencies	Site Rate Model	Clock Rate Model	Ancestral Sequence
3	Ebola.tre	3	1000	3	Nucleotide	Tree model	Substitution Model	Frequencies	Site Rate Model	Clock Rate Model	Ancestral Sequence

Setup substitution model for partition 1

Substitution model: HKY

Set parameter values:

Kappa value: 4.0

20 taxa loaded.

Figure 1: **Overview of the πBUSS simulation procedures and GUI screenshot.** A. Schematic representation of the different ways to employ the πBUSS simulation software. Based on an XML input file, simulations can be performed using the core implementation in BEAST. BEAST can generate sequence data as well as analyze the replicate data in a single run. Using both the GUI or CLI, πBUSS can run simulations based on an input tree or a list of taxa and their heights. The software can also write the simulation settings to an XML file for simulation in BEAST. B. The screenshot example shows the set-up of a codon position partitioned simulation in the Partitions panel of the graphical user interface. The Hasegawa, Kishino and Yano (HKY) model is being set as the substitution model for partition 1, with a κ (the transition-transversion bias) parameter value of 4.0.

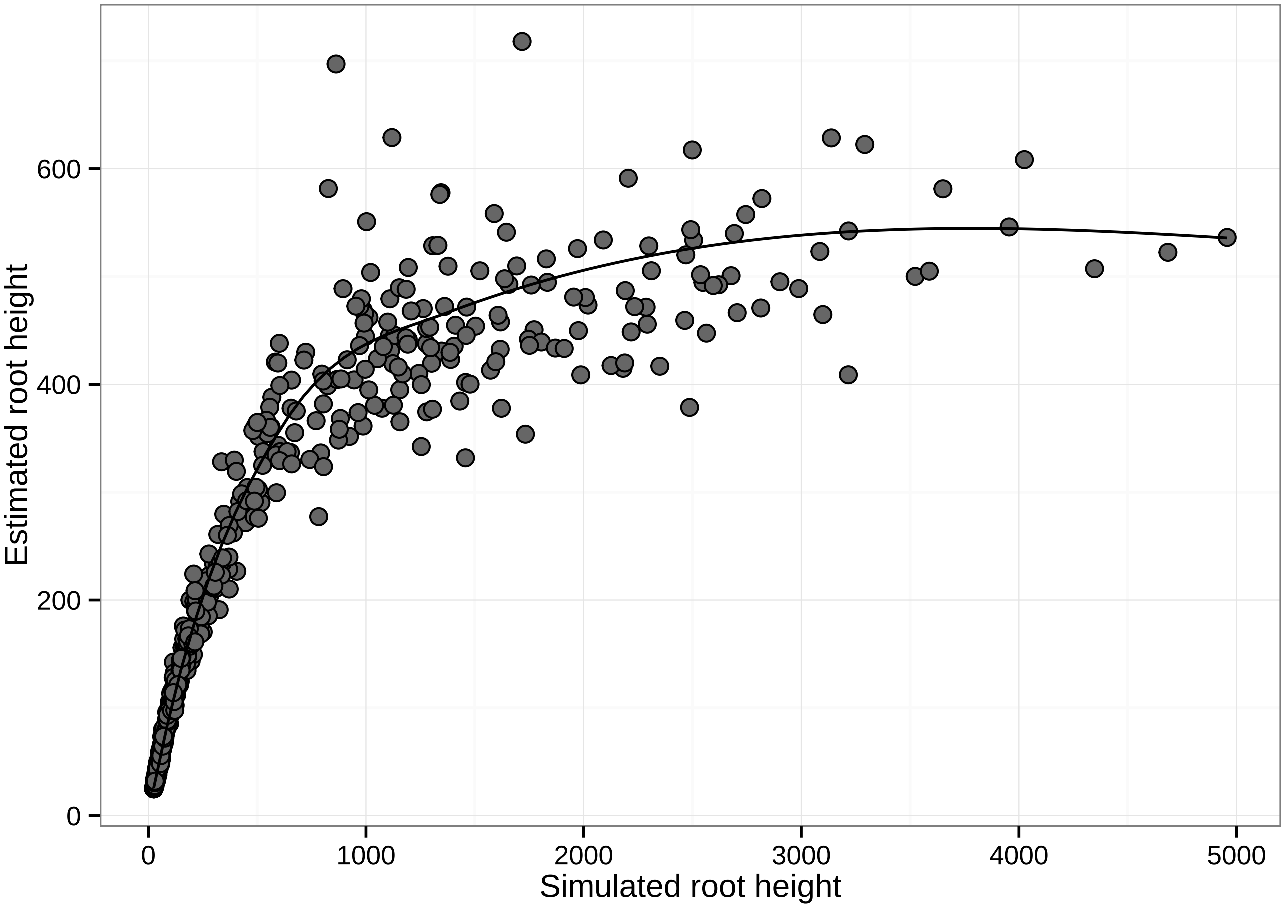


Figure 2: **Correspondence between simulated and estimated tMRCAs when purifying selection increases back in time in simulated data sets.**