Supplementary Material for "LinguaPhylo: a probabilistic model specification language for reproducible phylogenetic analyses"

Alexei J. Drummond[1,2,3]*, Kylie Chen[1,2,3], Fábio K. Mendes[1,4], Dong Xie[1,2,3]

1 Centre for Computational Evolution, University of Auckland, Auckland, New Zealand

2 School of Biological Sciences, University of Auckland, Auckland, New Zealand

3 School of Computer Science, University of Auckland, Auckland, New Zealand

4 Department of Biology, Washington University in St. Louis, St. Louis, United States

*Corresponding author: a.drummond@auckland.ac.nz

List of Tables

Function	Description	Examples
binaryRateMatrix	Binary trait rate matrix	errorModel1.lphy, errorModel2.lphy
f81	F81 model[4]	f81Coalescent.lphy
generalTimeReversible	General time reversible rate matrix	h5n1.lphy
gtr	GTR model[17]	gtrCoalescent.lphy
hky	HKY model[5]	hkyCoalescent.lphy
jukesCantor	Jukes-Cantor model[8]	jcCoalescent.lphy
k80	K80 model[10]	
lewisMK	LewisMK model[11]	lewisMKCoalescent.lphy
migrationMatrix	Population process rate matrix	simpleStructuredCoalescent.lphy
wag	WAG model[19]	wagCoalescent.lphy

Table A: Functions for substitution models and rate matrices in LPhy.

Generative distribution	Description	Examples
MultispeciesCoalescent	Multispecies coalescent	simpleMultispeciesCoalescent.lphy, simpleMultispeciesCoalescentTaxa.lphy, twoGeneMultispeciesCoalescent.lphy
Coalescent	Kingman's coalescent [13]	RSV2.lphy
SkylineCoalescent	Skyline coalescent [2]	hcv_col.lphy
StructuredCoalescent	Structured coalescent[12]	simpleStructuredCoalescent.lphy

Table B: Coalescent tree generative distributions in LPhy.

Generative distribution	Description	Examples
BirthDeathSampling	Birth-death-sampling tree[14, 15]	birthDeathRhoSampling.lphy
BirthDeathSerialSampling	Birth-death serial sampling tree[16]	simpleBirthDeathSerial.lphy
BirthDeath	Calibrated birth-death[7]	simpleCalibratedBirthDeath.lphy, simpleExtantBirthDeath.lphy
FossilBirthDeathTree	Fossilized birth-death process[6]	simFossilsCompact.lphy
FullBirthDeath	Birth-death tree[9]	simpleFullBirthDeath.lphy
RhoSampleTree	Birth-death tree sampled from a larger tree	
SimBDReverse	Birth-death tree with extant and extinct species	simFossils.lphy
SimFBDAge	Birth-death tree with extant and extinct species sampled through time	simFBDAge.lphy
SimFossilsPoisson	Tree with fossils added to given tree at rate psi	simFossils.lphy
Yule	Yule tree[18]	simpleYule.lphy, yuleRelaxed.lphy

Table C: Birth-death tree generative distributions in LPhy.

Generative distribution	Description	Examples
PhyloBrownian	Brownian motion process[3]	simplePhyloOU.lphy
PhyloCTMC	Continuous time Markov process[4]	simpleBModelTest.lphy
PhyloMultivariateBrownian	Multivariate Brownian motion	simplePhyloMultivariateBrownian.lphy
PhyloOU	Ornstein-Ulhenbeck process[3]	simplePhyloBrownian.lphy

Table D: Phylogenetic likelihood distributions in LPhy.

Generative distribution	Description	Examples
Bernoulli	Bernoulli distribution	simpleRandomLocalClock.lphy, simpleBModelTest.lphy
Beta	Beta distribution	birthDeathRhoSampling.lphy, simpleBModelTest.lphy
Cauchy	Cauchy distribution	
Dirichlet	Dirichlet distribution	birthDeathRhoSampling.lphy, dirichlet.lphy
DiscreteUniform	Discrete-uniform distribution	simpleBModelTest.lphy, simpleBModelTest2.lphy
DiscretizeGamma	Discretize-gamma distribution	gtrGammaCoalescent.lphy, simpleBModelTest.lphy
Exp	Exponential distribution	birthDeathRhoSampling.lphy, yuleRelaxed.lphy
ExpMarkovChain	Smoothing distribution [2]	skylineCoalescent.lphy
Gamma	Gamma distribution	covidDPG.lphy
Geometric	Geometric distribution	
InverseGamma	Inverse-gamma distribution	totalEvidence.lphy
LogNormal	Log-normal distribution	hkyCoalescent.lphy, errorModel1.lphy
Normal	Normal distribution	simplePhyloBrownian.lphy, simplePhyloOU.lphy
NormalGamma	Normal-gamma distribution	simplePhyloBrownian.lphy, simplePhyloOU.lphy
Poisson	Poisson distribution	expression4.lphy, simpleRandomLocalClock2.lphy
RandomBooleanArray	Samples a random boolean array	simpleRandomLocalClock2.lphy
RandomComposition	Samples a random k-tuple of positive integers that sum to n	skylineCoalescent.lphy
Uniform	Uniform distribution	simFossilsCompact.lphy
Weibull	Weibull distribution	
WeightedDirichlet	Weighted dirichlet distribution	totalEvidence.lphy, weightedDirichlet.lphy

Table E: Parametric distributions in LPhy.

Function	Description	Examples
aminoAcids	Amino acid data type	wagCoalescent.lphy
binaryDataType	Binary data type	
nucleotides	Nucleotide data type	primates2.lphy
standard	Standard data type	totalEvidence.lphy

Table F: Alignment data types in LPhy.

Function	Description	Examples
nucleotideModel	bModelTest[1] rate matrix	simpleBModelTest.lphy, simpleBModelTest2.lphy
bModelSet	bModelTest model set	simpleBModelTest.lphy
bSiteRates	Site rates for the given bModelTest parameters	simpleBModelTest2.lphy
bSiteModel	bModelTest site model	simpleBModelTest.lphy

Table G: Bayesian phylogenetic site model averaging in LPhy.

References

[1] Bouckaert, R., and Drummond, A. bmodeltest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evolutionary Biology 17*, 42 (2017).

[2] Drummond, A., Rambaut, A., Shapiro, B., and Pybus, O. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution 22* (2005), 1185–1192.

[3] Felsenstein, J. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics 25*, 5 (1973), 471.

[4] Felsenstein, J. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol 17*, 6 (1981), 368–76.

[5] Hasegawa, M., Kishino, H., and Yano, T.-a. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution 22*, 2 (1985), 160–174.

[6] Heath, T. A., Huelsenbeck, J. P., and Stadler, T. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences 111*, 29 (2014), E2957–E2966.

[7] Heled, J., and Drummond, A. J. Calibrated birth-death phylogenetic time-tree priors for bayesian inference. *Systematic Biology 64*, 3 (2015), 369–383.

[8] Jukes, T. H., Cantor, C. R., et al. Evolution of protein molecules. *Mammalian protein metabolism 3* (1969), 21–132.

[9] Kendall, D. G. On the generalized" birth-and-death" process. *The annals of mathematical statistics 19*, 1 (1948), 1–15.

[10] Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution 16*, 2 (1980), 111–120.

[11] Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology 50*, 6 (2001), 913–925.

[12] Müller, N. F., Rasmussen, D. A., and Stadler, T. The structured coalescent and its approximations. *Molecular biology and evolution 34*, 11 (2017), 2970–2981.

[13] Rodrigo, A. G., and Felsenstein, J. Coalescent approaches to hiv population genetics. In *The Evolution of HIV*, C. K., Ed. Johns Hopkins Univ. Press, Baltimore, 1999, ch. Chapter 8.

[14] Stadler, T. Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences 108*, 15 (2011), 6187–6192.

[15] Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., Günthard, H., Drummond, A., Bonhoeffer, S., and Swiss HIV Cohort Study. Estimating the basic reproductive number from viral sequence data. *Molecular biology and evolution 29* (2012), 347–357.

[16] Stadler, T., and Yang, Z. Dating phylogenies with sequentially sampled tips. *Syst Biol 62*, 5 (Sep 2013), 674–88.

[17] Tarvaré, S. Some probabilistic and statistical problems in the analysis of dna sequences. *Some mathematical question in biology-DNA sequence analysis* (1986).

[18] Udny Yule, G. A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs. *Philosophical Transactions of the Royal Society of London Series B 213* (1925), 21–87.

[19] Whelan, S., and Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution 18*, 5 (2001), 691–699.