Bioinformatics, 31(14), 2015, 2391–2393
doi: 10.1093/bioinformatics/btv126
Advance Access Publication Date: 27 February 2015
Applications Note

Phylogenetics

GeLL: a generalized likelihood library for phylogenetic models

Daniel Money[1,2,*] **and Simon Whelan**[3]

[1]Department of Plant and Animal Sciences, Faculty of Agriculture, Dalhousie University, Truro, B2N 5E3 Canada, [2]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, 66045, USA and [3]Department of Evolutionary Biology, Uppsala University, Uppsala, 75236, Sweden

*To whom correspondence should be addressed.

Associate Editor: David Posada

Received on October 15, 2014; revised on February 20, 2015; accepted on February 23, 2015

Abstract

Summary: Phylogenetic models are an important tool in molecular evolution allowing us to study the pattern and rate of sequence change. The recent influx of new sequence data in the biosciences means that to address evolutionary questions, we need a means for rapid and easy model development and implementation. Here we present GeLL, a Java library that lets users use text to quickly and efficiently define novel forms of discrete data and create new substitution models that describe how those data change on a phylogeny. GeLL allows users to define general substitution models and data structures in a way that is not possible in other existing libraries, including mixture models and non-reversible models. Classes are provided for calculating likelihoods, optimizing model parameters and branch lengths, ancestral reconstruction and sequence simulation.
Availability and implementation: http://phylo.bio.ku.edu/GeLL under a GPL v3 license.
Contact: daniel.money@dal.ca
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Many studies in molecular evolution use phylogenetic substitution models to infer patterns of change between homologous characters using maximum likelihood (ML). These models allow us to infer many valuable quantities, such as how the rate evolution varies among sites or the selective pressures acting on codons (Yang, 2006). The growing availability of genome sequences has led to new forms of data, such as gene content and the presence or absence of promoters (e.g. ENCODE Project Consortium, 2012), and their study requires statistically rigorous inferential tools. Nearly all phylogenetic substitution models are continuous time Markov chains with discrete and finite state spaces, which require the same underlying computational machinery. General phylogenetic libraries, such as BEAST (Drummond *et al.*, 2012), PLL (Flouri *et al.*, 2015) and Bio++ (Guéguen *et al.*, 2013), take advantage of these similarities and allow programmers to build phylogenetic models and are often the basis of existing packages. Here, we present GeLL, a Generalised Likelihood Library, which provides non-expert users a text interface to define data and models, allowing users to create data structures and models parameterizations that reflect the properties of their data. To aid in characterization of these models, GeLL also allows users to perform ML inference, simulate data and perform ancestral reconstruction.

2 Capabilities

GeLL is designed around the idea that phylogenetic models are defined by three key components. The first component is a finite and discrete character space used to define a data matrix, such as a nucleotide sequence, counts of occurrences of gene family members in genome or presence or absence of morphological characters in a species. The second component is the instantaneous rate matrix of a continuous time Markov process that describes the relative rates of transition between characters in the model state space. These two components are related by a function that maps each single element of the character space on to one or more elements in the model state space. The final component defines output, which allows users to access information derived from the other two components. Figure 1 shows the major components of GeLL and how they link together to perform phylogenetic analysis.

2.1 Data

GeLL allows users to define their own character space, which can be any set of discrete characters for which users are confident of assigning one-to-one homology. The user expresses these homology statements through the equivalent of a multiple sequence alignment conditioned on that character space, which in turn is input into GeLL for computation. Data may be commonly used sequence character spaces, such as nucleotides or amino acids, or more unusual characters, including genome size, gene family copy number or ploidy number. A state may also contain groups of 'characters', such as the three nucleotides in a codon. GeLL comes with classes to read a traditional alignment (in relaxed Phylip and FASTA formats) or an 'alignment' of numbers (such as gene family size). The sites in an 'alignment' can be partitioned, so that each partition has its own model and the parameters of that model may be shared between all or subsets of partitions. The relationship between observed elements of the character space and the elements used in the model space may be straight forward, such as the direct mapping of 'A' in a nucleotide character-space to an 'A' in the model state space. One may wish, however, to include characters that represent uncertainty about the observed characters, such as the use of 'W' to represent the weakly bonding nucleotides A:T or work with covarion-style models (Whelan, 2008), where an observed 'A' may map to both an 'ON' state or an 'OFF' state in the model. To accommodate these forms of models, GeLL allows users to define the relationship between the observed character space and the state space used in the model.

2.2 Models

The most important component of GeLL for many users will be the Model class. Central to the definition of phylogenetic substitution models (Yang, 2006) is the rate matrix, **Q**, that uses a set of parameters to define the instantaneous rate of change between any two elements in the model state space. In Figure 1, for example, the nucleotide HKY model is (Yang, 2006) defined through parameters describing the relative rate of change of transitions to transversions (k) and the stationary distribution of the nucleotides. GeLL allows users to define this matrix through a two-dimensional array of strings, with each string representing an equation consisting of parameters and simple mathematical operators. Models may be both reversible and non-reversible, with the maximum size of the model state space limited only by the computational resources available. Models may also have multiple rate matrices, allowing users to create complex mixture models, including models with multiple rate classes (Yang, 2006), temporal hidden Markov models (Whelan, 2008) and structure-based mixture models (Le *et al.*, 2008).

To enable non-stationary models, GeLL allows users to define the distribution of states at the root of the tree in one of four ways: (i) through a string of character-states, such as a sequence; (ii) through the stationary distribution of the rate matrix; (iii) through the quasi-stationary distribution for models with a sink-state (Darroch and Seneta, 1967) or (iv) through the method of FitzJohn *et al.* (2009), where root frequencies are proportional to the probability of that root assignment generating the data. We note that the quasi-stationary distribution is new to phylogenetics, although it has been used for modelling populations (Ovaskainen, 2002) and endemics (Nasell, 1996). To our knowledge, no other library allows users to define such a broad range of complex non-stationary and non-reversible models.

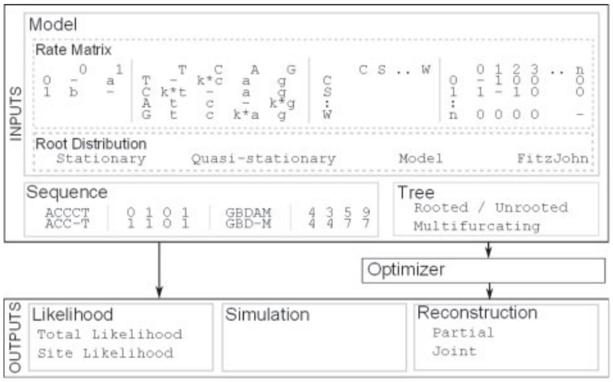

Fig. 1. A schematic of the major components of GeLL, including examples of models that can be implemented in GeLL such as, from left to right, a two-state binary character model, the HKY nucleotide model, amino acid GTR models and birth–death models

2.3 Output

GeLL provides a number of ways to study both a model and its application to data. At the simplest level, GeLL allows the computation of the likelihood of an alignment given a parameterized substitution model and a rooted or unrooted tree with branch lengths using the pruning algorithm (Felsenstein, 1981). Model parameters and branch lengths can also be estimated from sequence data using ML. Given a set of model parameters and data, GeLL can be used to simulated data or perform ancestral reconstruction using either marginal (Yang *et al.*, 1995) or, in the case of non-mixture models, joint reconstruction (Pupko *et al.*, 2000, 2002). To aid the statistical comparison of models, there are also classes for comparing models through the likelihood ratio test statistic (Goldman, 1993), the Akaike information criteria (Akaike, 1974) and the Bayesian information criteria (Schwarz, 1978). Classes are also included that allow the comparison of models with different state spaces using the method of Whelan *et al.* (2015).

2.4 Using GeLL

GeLL offers considerably different functionality than existing libraries, such as Bio++ (Guéguen *et al.*, 2013) or packages, such as HyPhy (Pond and Muse, 2005) and PAML (Yang, 1997). By allowing users to define the character space of their data, the state space of their model and how the two spaces interact, it allows complex model specification outside the norms of phylogenetic inference, in contrast to existing tools that are strongly linked to existing modelling approaches and structures. The generality of the model, including the specification of non-reversible models and the root distributions, also allows users to create novel and complex models that cannot be implemented in existing packages. GeLL includes a driver which allows many simple investigations to be completed without the need for any programming. To further demonstrate the functionality and utility of GeLL, we include as an example the models described in Mayrose *et al.* (2010) to describe the evolution of chromosome number. This implementation required fewer than 100 lines of code and around half an hour of programming. A second example comes from our previous work, where GeLL was used to model gene family evolution through a birth–death-innovation model with an upper bound on family size (Ames *et al.*, 2012). The Supplementary Material includes many examples of how to code new models to help users to become familiar with GeLL's syntax. It also includes a broad comparison of GeLL's features and performance with other phylogenetic libraries and software packages (Supplementary Tables S2 and S3), the numerical optimization methods available in GeLL (Supplementary Table S4) and the tests used to verify GeLL's calculations (Supplementary Table S5).

Acknowledgements

We thank Simon Lovell and Ryan Ames for testing GeLL and Mark Holder for supporting the project and hosting the library.

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council, UK, and National Science Foundation [DEB-0732920 to Mark Holder].

Conflict of Interest: none declared.

References

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automatic Control,* **19**, 716–723.

Ames,R.M. *et al.* (2012) Determining the evolutionary history of gene families. *Bioinformatics,* **28**, 48–55.

Darroch,J.N. and Seneta,E. (1967) On quasi-stationary distributions in absorbing continuous-time finite Markov chains. *J. Appl. Probability,* **4**, 192–196.

Drummond,A.J. *et al.* (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol,* **29**, 1969–1973.

ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature,* **489**, 57–74.

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.,* **17**, 368–376.

FitzJohn,R.G. *et al.* (2009) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst. Biol.,* **58**, 595–611.

Flouri,T. *et al.* (2015) The phylogenetic likelihood library. *Syst. Biol,* **64**, 356–362.

Goldman,N. (1993) Statistical tests of models of DNA substitution. *J. Mol. Evol.,* **36**, 182–198.

Guéguen,L. *et al.* (2013) Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.,* **30**, 1745–1750.

Le,S.Q. *et al.* (2008) Phylogenetic mixture models for proteins. *Philos. Trans. R. Soc. B Biol. Sci.,* **363**, 3965–3976.

Mayrose,I. *et al.* (2010) Probabilistic models of chromosome number evolution and the inference of polyploidy. *Syst. Biol.,* **59**, 132–144.

Nasell,I. (1996) The quasi-stationary distribution of the closed endemic SIS model. *Adv. Appl. Probability,* **28**, 895–932.

Ovaskainen,O. (2002) The effective size of a metapopulation living in a heterogeneous patch network. *Am. Nat.,* **160**, 612–628.

Pond,S.L.K. and Muse,S.V. (2005) HyPhy: hypothesis testing using phylogenies. In: Nielsen,R. (ed.) *Statistical Methods in Molecular Evolution, Statistics for Biology and Health*. Springer, New York, pp. 125–181.

Pupko,T. *et al.* (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.,* **17**, 890–896.

Pupko,T. *et al.* (2002) A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: application to the evolution of five gene families. *Bioinformatics,* **18**, 1116–1123.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.,* **6**, 461–464.

Whelan,S. (2008) Spatial and temporal heterogeneity in nucleotide sequence evolution. *Mol. Biol. Evol.,* **25**, 1683–1694.

Whelan,S. *et al.* (2015) ModelOMatic: fast and automated model selection between RY, nucleotide, amino acid, and codon substitution models. *Syst. Biol.,* **64**, 42–55.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.,* **13**, 555–556.

Yang,Z. (2006) *Computational Molecular Evolution.* Oxford University Press, Oxford, UK.

Yang,Z. *et al.* (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics,* **141**, 1641–1650.