BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 7 2014, pages 1020–1021
doi:10.1093/bioinformatics/btt729

Phylogenetics
Advance Access publication December 18, 2013

Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package

Nicolas Rodrigue[1,*] and Nicolas Lartillot[2]

[1]Department of Mathematics and Statistics, University of Calgary, 2500 University Drive NW, Calgary AB T2N 1N4, Canada and [2]UMR CNRS 5558 - LBBE, Université Lyon 1, Villeurbanne Cedex, France

Associate Editor: David Posada

ABSTRACT

Motivation: In recent years, there has been an increasing interest in the potential of codon substitution models for a variety of applications. However, the computational demands of these models have sometimes lead to the adoption of oversimplified assumptions, questionable statistical methods or a limited focus on small data sets.

Results: Here, we offer a scalable, message-passing-interface-based Bayesian implementation of site-heterogeneous codon models in the mutation-selection framework. Our software jointly infers the global mutational parameters at the nucleotide level, the branch lengths of the tree and a Dirichlet process governing across-site variation at the amino acid level. We focus on an example estimation of the distribution of selection coefficients from an alignment of several hundred sequences of the influenza PB2 gene, and highlight the site-specific characterization enabled by such a modeling approach. Finally, we discuss future potential applications of the software for conducting evolutionary inferences.

Availability and implementation: The models are implemented within the PhyloBayes-MPI package, (available at phylobayes.org) along with usage details in the accompanying manual.

Contact: nicolas.rodrigue@ucalgary.ca

Received on October 25, 2013; revised on December 10, 2013; accepted on December 12, 2013

1 INTRODUCTION

There is growing interest for the use of codon substitution models in several contexts, including phylogenetic inference (Gil *et al*., 2013), ancestral sequence reconstruction (Chang *et al*., 2012) and the characterization of the selective effects of specific nonsynonymous mutations (Tamuri *et al*., 2012). Focusing on the latter, Halpern and Bruno (1998) first showed how to devise a model that accounts for both global mutational features at the nucleotide level and site-specific selective constraints at the amino acid level. Although their approach was directed to the estimation of evolutionary distances, it was later recognized as enabling the estimation of distributions of selection coefficients from phylogenetic data (see Thorne *et al*., 2012, for a review of these developments). However, a serious issue with the Halpern and Bruno model, and some of the subsequent re-implementations (e.g.Tamuri *et al*., 2012), lies in the use of site-specific parameters optimized to maximum likelihood estimates; such an approach induces the 'infinitely many parameters trap', in which each additional observation changes the form of the overall model (see Rodrigue, 2013).

Yang and Nielsen (2008) devised some simpler homogeneous mutation-selection models, along with a likelihood ratio test aimed at evaluating the significance of codon usage bias. While statistically well-justified, the homogeneity of their mutation-selection models makes them biologically unsatisfying for the purpose of estimating distributions of selection coefficients.

Subsequently, we proposed the use of a nonparametric approach based on the Dirichlet process, providing a flexible and statistically well-founded method to accommodating across-site heterogeneity of amino acid constraints (Rodrigue *et al*., 2010). However, our proof-of-concept implementation only allowed for its application on very small data sets, and its rate-limiting Markov chain Monte Carlo (MCMC) updates on the Dirichlet process, based on a Chinese-restaurant approach, were not amendable to parallelization.

Working with nucleotide and amino acid level substitution models, we recently developed PhyloBayes-MPI, which, among other speedup strategies, uses message-passing-interface (MPI) and a truncated stick-breaking representation of the Dirichlet process for parallelized updating (Lartillot *et al*., 2013). Here, we have expanded PhyloBayes-MPI for the implementation of several types of codon substitution models, including the Dirichlet process-based site-heterogeneous mutation-selection approach. We illustrate how the software can now be used for efficient estimation of distributions of selection coefficients (scaled by the effective chromosomal population size), and discuss several future avenues that it enables.

2 METHODS

In the present application, the program is passed an alignment file of coding nucleotide sequences (of a length that is a factor of 3) and a corresponding tree topology file. The universal genetic code is assumed, but an alternative code can be specified (e.g. `-mtvert` for the vertebrate mitochondrial code). As with other models with PhyloBayes-MPI, the program uses $K > 1$ cores. At startup, the master core draws an initial model configuration from the prior, and broadcasts it using MPI to the $K-1$ compute cores. All updates are data-augmentation–based, which are several orders of magnitude faster than pruning-based updates (de Koning *et al*., 2010). Each iteration of the MCMC includes numerous updates on global parameters, performed by the master core, whereas compute cores perform several updates of a truncated stick-breaking representation of the Dirichlet process (see Lartillot *et al*., 2013, for details), and sample the data augmentations.

*To whom correspondence should be addressed.

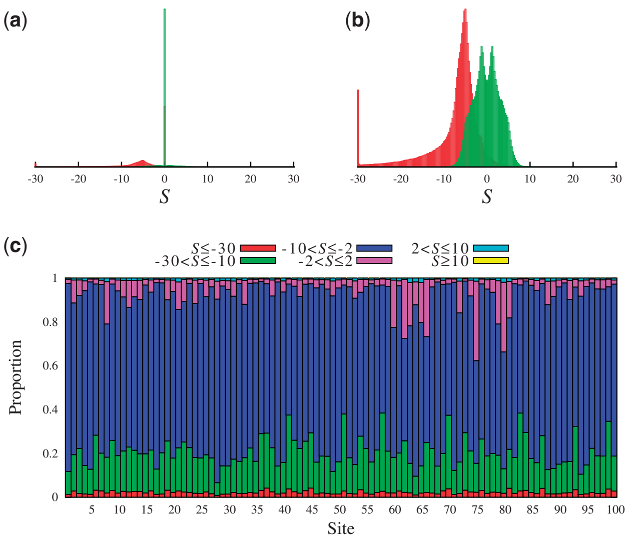

Fig. 1. Posterior mean distribution of selection coefficients (*S*) for all types of events (**a**) and for nonsynonymous events (**b**) appearing at mutation-selection balance, with red histograms for mutations (mainly deleterious, hence negative values), and green histograms for substitutions (symmetrical about 0, given the mutation-selection balance). Panel (**c**) summarizes the first 100 site-specific distributions of nonsynonymous mutations

3 EXAMPLE

We present an example that we could not run with our previous implementation, consisting of 401 sequences of the PB2 gene of influenza, taken from Tamuri *et al.* (2012). With the present implementation, we obtained a sample of 1100 draws within ~5 h, running on a 12 core (hyper-threaded), Intel i7-based workstation. Discarding the first 100 draws, we display the posterior mean distributions of selection coefficients (*S*) in Figure 1.

Looking at panel **a**, we find that the highest-valued bin is that of synonymous events ($S = 0$), most mutations (red) are deleterious ($S < 0$) and most substitutions (green) are either neutral (synonymous) or nearly neutral. Whereas the 'infinitely many parameters' approach used by Tamuri *et al.* (2012) led to the conclusion that most nonsynonymous mutations have $S < -10$, panel **b** shows that most have a selection coefficient between -10 and -2, with the mode situated at -5. The red distribution, however, seems more plausible than that obtained using a parametric site-specific approach (Rodrigue, 2013), which inferred most nonsynonymous events with *S* between 0 and -5.

Panel **c** displays a site-specific assessment of the distributions of *S* for nonsynonymous mutations, focusing on the first 100 codons of the alignment. For graphical simplicity, we have added up the values over a few sets of classes. The $-10 < S \leq -2$ class, in blue, is the most represented—as previously revealed from panel **b**—but some sites, e.g. codon 75, have almost as many nonsynonymous mutations in the nearly neutral class, $-2 < S \leq 2$, indicating that they are under less stringent evolutionary constraints than other sites.

4 FUTURE DIRECTIONS

Our implementation enables numerous potential applications. For instance, the last panel of Figure 1 suggests studies on the distributions of *S* over classes of sites. Distributions could even be generated for all possible types of nonsynonymous mutations at each site. The software could also serve in the development of approaches that explicitly incorporate structural features (e.g. Meyer and Wilke, 2013), and already includes mutation-selection models using finite mixtures, homogeneous versions and models based on a univariate factor on nonsynonymous rates (ω). Applications of these models will be the focus of future papers.

Developed within the PhyloBayes-MPI package, the models we have implemented here inherit the ability to perform a variety of types of posterior predictive model assessments, cross-validation comparisons, ancestral sequence reconstruction, as well as phylogenetic inference *per se*. Much more work is needed in these contexts, to assess what insights may be gained from the mutation-selection framework, and from codon substitution models in general. The present application should help engage such work.

ACKNOWLEDGEMENT

The authors thank Claus Wilke, and two anonymous reviewers for their comments on the manuscript.

Funding: Natural Sciences and Engineering Research Council of Canada (N.R. and N.L.), Agriculture and Agri-Food Canada (N.R.), the University of Calgary (N.R.) and l'Université de Montréal (N.L.).

Conflict of Interest: none declared.

REFERENCES

Chang,B.S.W. *et al.* (2012) The future of codon models in studies of molecular function: ancestral reconstruction and clade models of functional divergence. In: Cannarozzi,G.M. and Schneider,A. (eds) *Codon Evolution*. Oxford University Press, Oxford, pp. 145–163.

de Koning,A.P.J. *et al.* (2010) Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. *Mol. Biol. Evol.*, **27**, 249–265.

Gil,M. *et al.* (2013) CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models. *Mol. Biol. Evol.*, **30**, 1270–1280.

Halpern,A.L. and Bruno,W.J. (1998) Evolutionary distances for protein-coding sequences: modeling site- specific residue frequencies. *Mol. Biol. Evol.*, **15**, 910–917.

Lartillot,N. *et al.* (2013) PhyloBayes-MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.*, **62**, 611–615.

Meyer,A.G. and Wilke,C.O. (2013) Integrating sequence variation and protein structure to identify sites under selection. *Mol. Biol. Evol.*, **30**, 36–44.

Rodrigue,N. (2013) On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics*, **193**, 557–564.

Rodrigue,N. *et al.* (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. USA*, **107**, 4629–4634.

Tamuri,A.U. *et al.* (2012) Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, **190**, 1101–115.

Thorne,J.L. *et al.* (2012) Codon models as a vehicle for reconciling population genetics with inter-specific sequence data. In: Cannarozzi,G.M. and Schneider,A. (eds) *Codon Evolution*. Oxford University Press, pp. 97–110.

Yang,Z. and Nielsen,R. (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.*, **25**, 568–579.