Bioinformatics Advance Access published September 12, 2012

Phylogenetics, Likelihood, Evolution and Complexity (*PLEX*)

A.P. Jason de Koning[1,2], Wanjun Gu[3], Todd A. Castoe[1,4] and David D. Pollock[1*]

[1]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, USA

[2]Department of Biochemistry and Molecular Biology, University of Calgary, Faculty of Medicine, Canada

[3]Key Laboratory of Child Development and Learning Science, Southeast University, China

[4]Department of Biology, The University of Texas at Arlington, USA

Associate Editor: David Posada

ABSTRACT

Summary: *PLEX* is a flexible and fast Bayesian MCMC software program for large-scale analysis of nucleotide and amino acid data using complex evolutionary models in a phylogenetic framework. The program gains large speed improvements over standard approaches by implementing 'partial sampling of substitution histories', a data augmentation approach that can reduce data analysis times from months to minutes on large comparative datasets. A variety of nucleotide and amino-acid substitution models are currently implemented, including non-reversible and site-heterogeneous mixture models. Due to efficient algorithms that scale well with data size and model complexity, *PLEX* can be used to make inferences from hundreds to thousands of taxa in only minutes on a desktop computer. It also performs probabilistic ancestral sequence reconstruction. Future versions will support detection of co-evolutionary interactions between sites, probabilistic tests of convergent evolution, and rigorous testing of evolutionary hypotheses in a Bayesian framework.

Availability and Implementation: *PLEX* v1.0 is licensed under GPL. Source code and documentation will be available for download at www.evolutionarygenomics.com/ProgramsData/PLEX. *PLEX* is implemented in C++ and supported on Linux, Mac OS X, and other platforms supporting standard C++ compilers. Example data, control files, documentation and accessory Perl scripts are available from the website.

***Contact:** David.Pollock@UCDenver.edu

Supplementary Information: Supplemental results file

1 INTRODUCTION

Comparative genomic and phylogenetic datasets are growing dramatically in size, thanks to rapid and inexpensive next-generation DNA sequencing technologies. The wealth of information present in large datasets, however, is difficult to exploit due to the non-linearly increasing computational burden imposed by increasing data size and model complexity (de Koning, et al., 2010). For example, it is not uncommon for phylogenetic analysis of large datasets to now incur compute times on the order of months, even with models that are fairly simple from a mechanistic, biochemical perspective (*e.g.*, Rodrigue, et al., 2010).

To help overcome these challenges, we developed *PLEX*, which implements rapid data augmentation methods using partial sampling of substitution histories (de Koning, et al., 2010), and derivative methods. *PLEX* is an active research platform and will support the future development of new models and algorithms. The capabilities of *PLEX* will therefore expand over time. For example, the current version now incorporates more models than previously, as well as implementations of conjugate Gibbs sampling (Lartillot, 2006) and slice sampling (Neal, 2003).

2 APPROACH

PLEX performs Bayesian Markov chain Monte Carlo (MCMC) analysis using novel data augmentation strategies that can afford it great speed advantages over MCMC packages that use fully integrated likelihood calculations (*e.g., MrBayes,* Ronquist and Huelsenbeck, 2003 and *BEAST,* Drummond and Rambaut, 2007). Speed advantages are obtained by 'partially sampling' substitution events to the nearest branch region, and rapidly integrating over the timing of events within regions when evaluating the likelihood. This approach can be much faster than sampling the exact timing of substitution events, but appears to retain the computational advantages of doing so (de Koning, et al., 2010). As a result, *PLEX* is faster than programs that augment sequences with fully specified substitution histories (*e.g.*, *PhyloBayes*; Lartillot, et al., 2009).

The program samples branch-lengths, model parameters, and 'missing' unobserved data (*e.g.,* ancestral states) from their posterior distributions, and provides posterior summaries to facilitate inferences about evolutionary mechanisms. The distribution also includes the ability to calculate posterior statistics of substitution histories, which can be an effective approach for making complex inferences from comparative data without fully modeling the processes of interest (*e.g.*, for detecting co-evolution; de Koning et al., unpubl.). Posterior-predictive simulation is included for some model checking applications, and Bayes factor calculation can be made using thermodynamic integration (Lartillot and Philippe, 2006; Rodrigue and Aris-Brosou, 2011).

Importantly, the speed advantages enabled by *PLEX*'s data augmentation approach help to overcome computational challenges of large-scale phylogenomic inference in general, and are thus applicable to virtually any evolutionary model, or phylogenetic analysis task (de Koning, et al., 2010). Furthermore, the approach was designed to substantially reduce the scaling of time complexity with

model complexity (Supplementary Fig. 1), so that model realism can be more effectively explored without the need to greatly over-simplify models for computational convenience.

For example, the times to achieve an average of 1,000 effectively-independent posterior samples on a large dataset are shown for a variety of amino-acid substitution models (**Table 1**). Remarkably, the most parameter-rich site-homogeneous model in this set is also the easiest to sample from (and therefore the fastest), because Conjugate Gibbs sampling mechanisms are possible for all parameters in this case (general non-reversible). Thus, not only do these approaches reduce the need to simplify models for computational convenience, in some cases they facilitate an inversion of incentives, making complex models more computationally convenient than simpler ones. It should be noted that similar contrasts can be constructed when exact substitution timings are augmented (Lartillot, 2006). For performance comparisons between *PLEX* and several popular phylogenetics programs, see supplemental results.

Table 1. Speed and efficiency of sampling amino-acid rate matrix parameters (224 mammalian cyt-*b* sequences).

Method	Num. Params.	Gen. time	Efficiency*	Time§
General Non-Rev. model (Conjugate Gibbs)	827	0.0012s	78.0% ± 31	144s
General Reversible model (Slice sampling)	636	0.0023s	57.6% ± 41	406s
General Non-Rev. + 4-cat. Discrete Gamma (Slice + MH sampling)	828	0.0044s	72.4% ± 32	602s

*average effective sample size per generation (mixing efficiency); §time to achieve average of 1,000 effectively independent samples

When Gibbs update mechanisms are unavailable for a given model, *PLEX* uses slice sampling or Metropolis Hastings (MH) updates. The simple form of *PLEX*'s likelihood function often makes likelihood evaluations for these mechanisms extremely rapid, and sometimes even $O(1)$ (de Koning, et al., 2010).

The *PLEX* code is written in C/C++ and has no dependencies on external libraries. It should therefore compile on any platform that supports standard C/C++.

3 APPLICATIONS

PLEX is run at the command-line using a control file interface that allows customization of analyses, including locations of input sequence alignments and tree topologies. Example control files are included that demonstrate how to run most types of supported analyses. Any position in the alignment possessing a gap can be excluded, or optionally treated as missing.

PLEX currently has several main uses, including:

(1) Rate matrix and branch-length estimation from hundreds to thousands of taxa in minutes on a desktop computer;

(2) Inference of site heterogeneity in substitution patterns;

(3) Probabilistic inference of ensembles of ancestral states;

(4) Calculation of posterior statistics of substitution histories (*e.g.,* for tests of co-evolution; de Koning et al., unpubl.);

(5) Hypothesis testing and model evaluation using posterior-predictive statistics (scripts to automate thermodynamic integration will be made available in a future release).

PLEX is an active research platform for the development of new models and algorithms. As such, it will be updated frequently and will grow in capabilities over time. We therefore expect that *PLEX* will continue to evolve into an increasingly capable and comprehensive package for studying molecular evolutionary mechanisms.

4 CONCLUSION

PLEX is a software package for performing rapid and accurate Bayesian MCMC analysis on large phylogenomic datasets that remains under active development. Future updates will include models of temporal heterogeneity, spatial substitution gradients (Faith and Pollock, 2003), and codon-substitution. It is likely that parallelization and tuning of our slice sampling algorithms will further substantially improve speed for certain complex analysis tasks, and will thus be pursued in future releases.

Funding: This work was supported by the National Institutes of Health (NIGMS R01 GM083127 to DDP) and the National Natural Science Foundation of China (30900836 to WG).

Conflicts of Interest: None declared.

REFERENCES

de Koning, A.P.J., Gu, W. and Pollock, D.D. (2010) Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories, *Mol Biol Evol*, **27**, 249-265.

Drummond, A.J. and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees, *BMC Evol Biol*, **7**, 214.

Faith, J.J. and Pollock, D.D. (2003) Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes, *Genetics*, **165**, 735-745.

Lartillot, N. (2006) Conjugate Gibbs sampling for Bayesian phylogenetic models, *J Comput Biol*, **13**, 1701-1722.

Lartillot, N., Lepage, T. and Blanquart, S. (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating, *Bioinformatics*, **25**, 2286-2288.

Lartillot, N. and Philippe, H. (2006) Computing Bayes factors using thermodynamic integration, *Syst Biol*, **55**, 195-207.

Neal, R.M. (2003) Slice Sampling, *The Annals of Statistics*, **31**, 705-741.

Rodrigue, N. and Aris-Brosou, S. (2011) Fast Bayesian choice of phylogenetic models: prospecting data augmentation-based thermodynamic integration, *Syst Biol*, **60**, 881-887.

Rodrigue, N., Philippe, H. and Lartillot, N. (2010) Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles, *Proc Natl Acad Sci U S A*, **107**, 4629-4634.

Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models, *Bioinformatics*, **19**, 1572-1574.