MBE Advance Access published December 23, 2016

PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses

Robert Lanfear,[1,2,*] Paul B. Frandsen,[3] April M. Wright,[4] Tereza Senfeld,[2] and Brett Calcott[5]

[1]Research School of Biology, Australian National University, Canberra, ACT, Australia

[2]Department of Biological Sciences, Macquarie University, Sydney, Australia

[3]Office of Research Information Services, Office of the Chief Information Officer, Smithsonian Institution, Washington, DC

[4]Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA

[5]Department of Philosophy, University of Sydney, Sydney, NSW, Australia

[*]**Corresponding author:** E-mail: rob.lanfear@anu.edu.au

Associate editor: Michael S. Rosenberg

Abstract

PartitionFinder 2 is a program for automatically selecting best-fit partitioning schemes and models of evolution for phylogenetic analyses. PartitionFinder 2 is substantially faster and more efficient than version 1, and incorporates many new methods and features. These include the ability to analyze morphological datasets, new methods to analyze genome-scale datasets, new output formats to facilitate interoperability with downstream software, and many new models of molecular evolution. PartitionFinder 2 is freely available under an open source license and works on Windows, OSX, and Linux operating systems. It can be downloaded from www.robertlanfear.com/partitionfinder. The source code is available at https://github.com/brettc/partitionfinder.

Key words: **partitioning, AIC, BIC, AICc, model selection, molecular evolution.**

Main Text

In phylogenetic analyses it is important to account for variation in rates and patterns of evolution among sites (Yang 1996; Kumar et al. 2012). Partitioning attempts to achieve this by estimating independent models of molecular evolution for subsets of sites that are deemed to have evolved in similar ways. It can be challenging to choose a good partitioning scheme, because the number of possible schemes can be extremely large.

The original version of PartitionFinder (Lanfear et al. 2012) proposed algorithms to automate the selection of a partitioning scheme given a set of user-defined data blocks as input. By combining these algorithms with the selection of models of molecular evolution, PartitionFinder improved and simplified phylogenetic analyses for many users. However, PartitionFinder was written before the advent of phylogenomic datasets such as those produced by sequencing whole genomes (e.g., Jarvis et al. 2014) and transcriptomes (e.g., Misof et al. 2014), and remains too slow to be practical for use with these datasets. Because of this, we designed new features and re-wrote all of the methods and routines in PartitionFinder, which we present as PartitionFinder 2.

PartitionFinder 2 includes a number of new features. First, we wrote faster versions of the *k*-means, relaxed-clustering, and greedy algorithms (Lanfear et al. 2014; Frandsen et al. 2015), although we urge caution with relying on purely data-driven approaches to partitioning such as *k*-means, because we still lack evidence that they perform appropriately under a wide range of simulation conditions (Frandsen et al. 2015). Second, we included a range of new models of evolution, including important recent advances such as the LG4X and LG4M mixture models (Le et al. 2012). Third, we implemented Maximum-Likelihood (ML) starting trees for all analyses, motivated by our observation that model selection methods can be biased by the choice of starting tree (Frandsen et al. 2015). Fourth, we implemented the ability to analyze morphological datasets. Finally, we included a variety of new output formats to improve interoperability with downstream software.

In addition to new features, we also implemented a number of improvements that enable the efficient analysis of genome-scale datasets. These include: a new alignment parser; more efficient use of multiple processors; a dramatic reduction in the number of files that are written and read; and many improvements in internal and external data storage and processing. These improvements streamline analyses and help to make the best use of the available computational resources.

The net result of the new features and improvements is that PartitionFinder 2 can be dramatically faster than its predecessor, particularly for very large datasets analyzed on computers with many processors. To illustrate this, we compared the performance of version 2.0.0 to version 1.1.1 on two datasets: an insect dataset comprising 2,868 protein domains (each specified as a separate data block) and 595,033 sites from 144 taxa (Misof et al. 2014); and a vertebrate dataset of

© The Author 2016. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://mbe.oxfordjournals.org/ at University of Sydney on January 4, 2017

Brief Communication

56 genes (split into 168 codon-position data blocks) and 25,919 sites from 110 taxa (Fong et al. 2012). We used Maximum Parsimony starting trees in all analyses to enable direct comparisons of execution times. We analyzed the insect dataset on a server with fifty-six 2.6 GHz processors, using the new fast relaxed clustering (rclusterf) algorithm in version 2.0.0, and the original relaxed clustering algorithm in version 1.1.1, both with default settings. Version 2.0.0 was more than 100 times faster than version 1.1.1: it completed the analysis in 35 h, while version 1.1.1 finished less than 1% of the analysis in the same time. We analyzed the vertebrate dataset on a desktop Macintosh computer with eight 4 GHz processors, using the greedy algorithm with precisely the same settings in versions 1.1.1 and 2.0.0. Version 2.0.0 was five times faster than version 1.1.1: it completed the analysis in 108 min compared with 534 min for version 1.1.1.

PartitionFinder 2 can be installed by downloading it from the website above, or installing it via GitHub. No other programs need to be compiled, but it does require the installation of Python and a small number of dependencies. These can be managed via a point-and-click installer, following the details outlined in the manual. We hope that PartitionFinder 2 will be useful to the phylogenetics community.

Acknowledgments

RML was supported by the Australian Research Council. AMW was supported by NSF DEB-1256993. This work was supported by the Macquarie University Genes to Geoscience center.

References

Frandsen PB, Calcott B, Mayer C, Lanfear R. 2015. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. *BMC Evol Biol.* 15:13.

Fong JJ, Brown JM, Fujita MK, Boussau B. 2012. A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic lissamphibia. *PLoS One* 7(11): e48990.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol Biol Evol.* 29:457–472.

Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol Biol.* 14:82.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol.* 29:1695–1701.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol.* 29:2921–2936.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346:763–767.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol.* 11:367–372.

Downloaded from http://mbe.oxfordjournals.org/ at University of Sydney on January 4, 2017