Bioinformatics, 36(22-23), 2020, 5516–5518
doi: 10.1093/bioinformatics/btaa1022
Advance Access Publication Date: 16 December 2020
Applications Note

Genome analysis

CAFE 5 models variation in evolutionary rates among gene families

Fábio K. Mendes[1,2,†], Dan Vanderpool [1,*,†], Ben Fulton[1,3,†] and Matthew W. Hahn[1,4]

[1]Department of Biology, Indiana University, Bloomington, IN 47405, USA, [2]Department of Computer Science, The University of Auckland, 1010 Auckland, New Zealand, [3]Department of Computer Science, University Information Technology Services, Indiana University, Bloomington, IN 47405, USA and [4]Department of Computer Science, Indiana University, Bloomington, IN 47405, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Peter Robinson

Received on June 2, 2020; revised on November 12, 2020; editorial decision on November 25, 2020; accepted on November 30, 2020

Abstract

Motivation: Genome sequencing projects have revealed frequent gains and losses of genes between species. Previous versions of our software, Computational Analysis of gene Family Evolution (CAFE), have allowed researchers to estimate parameters of gene gain and loss across a phylogenetic tree. However, the underlying model assumed that all gene families had the same rate of evolution, despite evidence suggesting a large amount of variation in rates among families.

Results: Here, we present CAFE 5, a completely re-written software package with numerous performance and user-interface enhancements over previous versions. These include improved support for multithreading, the explicit modeling of rate variation among families using gamma-distributed rate categories, and command-line arguments that preclude the use of accessory scripts.

Availability and implementation: CAFE 5 source code, documentation, test data and a detailed manual with examples are freely available at https://github.com/hahnlab/CAFE5/releases.

Contact: danvand@indiana.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The earliest eukaryotic genome sequencing projects revealed large and frequent changes between species in the size of gene families (Gibbs *et al.*, 2007; Rubin et al., 2000; Waterston *et al.*, 2002). Variation in family size results from the gain or loss of genes, either of which may be advantageous, deleterious or neutral. To enable the rigorous study of changes in gene family size, we previously proposed a statistical framework that would allow for inferences regarding gene family evolution among species along a phylogenetic tree (Hahn et al., 2005). We showed that this model can be used for hypothesis testing, inference of ancestral states and estimation of gene duplication and loss rates. Since its release, the software implementing this model, Computational Analysis of gene Family Evolution (CAFE), has been steadily improved to accommodate growing genomic resources (De Bie et al., 2006; Han et al., 2013). CAFE continues to be widely used in comparative genomics. However, in order to fully exploit the benefits of the rapidly growing number of sequenced genomic datasets, improvements such as multi-core parallelization and more sophisticated models of gene family evolution must be implemented.

CAFE models rates of change among gene families with a birth-death distribution having a mean rate (λ) of gain and loss common to all families. In reality, individual families can evolve at very different rates, with the most rapidly evolving families in terms of gain and loss (e.g. sex and reproduction-related, immunity) being the same as those observed to be evolving most rapidly at the sequence level (Demuth et al., 2006; Hahn et al., 2007). Furthermore, there appears to be a class of genes that are extremely resistant to duplication or loss, a trait that can be used to assess genome assembly quality (Waterhouse et al., 2013). Recent studies have confirmed variation in rates among families to be true in many different taxa (Casola and Lawing, 2019). Both DupliPHY (Ames et al., 2012) and Badirate (Librado et al., 2012), older programs designed for gene family analysis, employ measures to account for rate variation among families in some way.

Here, we present CAFE 5, an upgrade that explicitly models rate-variation among families in a manner directly analogous to similar models used for nucleotides and amino acids. Below we describe the implementation and testing of this model, as well as the other new features and improvements included in CAFE 5.

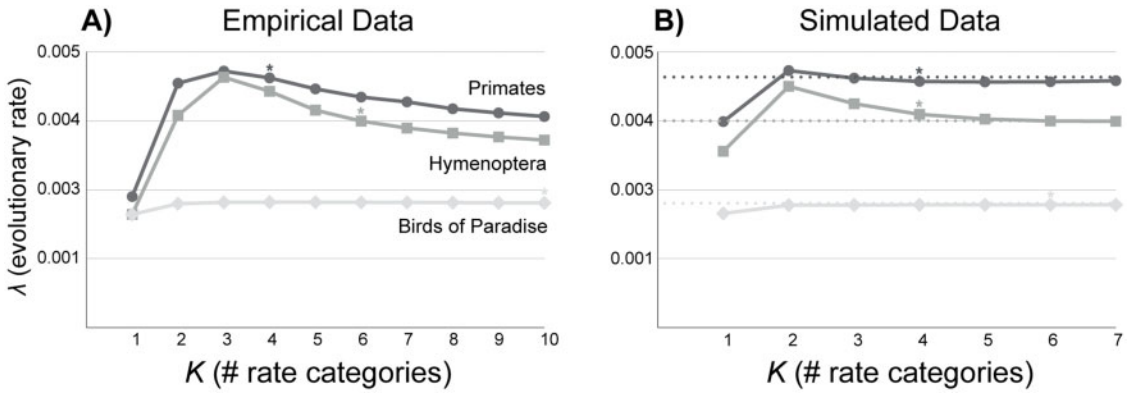

Fig. 1. Evolutionary rate (λ) estimated for empirical and simulated data under an increasing number of discrete rate categories (K). (**A**) λ estimated for three empirical datasets under increasing values of K. (**B**) Data simulated using the maximum likelihood values estimated for each dataset in A. The simulated values of λ are indicated by the dashed lines. The * indicates the value of λ for which the highest likelihood was estimated, in both panels

2 Materials and methods

2.1 Improved performance, stability and usability of CAFE

CAFE 5 introduces a more modular style of programming with a re-write from C into C++. The current version employs powerful compilers and matrix multiplication libraries and is able to take advantage of multiple cores, with noticeable performance increases up to at least 64 cores (Supplementary Fig. S1). To provide a simpler experience for the user, the script-based paradigm of earlier versions has been discarded in favor of a strictly command-line interface. Output files have been reconfigured to minimize post-processing, with trees written to a Nexus file for easy viewing.

2.2 New features in CAFE 5

A common approach used in molecular phylogenetics to model variation in rates among nucleotide or amino acid sites is to use a discrete approximation of the gamma (Γ) distribution (Yang, 1994). CAFE 5 uses this same approach (as does DupliPHY, Ames *et al.*, 2012), with the number of discrete rate categories, K, specified *a priori* and each category assumed equi-probable ($1/K$). The Γ distribution is scaled such that the mean rate across categories is 1, with shape parameter α ($=\beta$) estimated from the data. The shape of the Γ distribution determines a unique rate for each category, under which a gene family has its probability (given a set of parameter values) calculated. CAFE 5 then uses an empirical Bayes approach to estimate the posterior probability of a family belonging to a rate category, which in turn enables down-stream analyses of 'slow' or 'fast' families.

In addition, ancestral state reconstruction is now performed using the algorithm of Pupko et al. (2000), resulting in run times that scale linearly with the number of taxa in the tree.

3 Results

We used CAFE 5 to analyze three published datasets consisting of gene families from primates (Thomas et al., 2020b), birds of paradise (Prost et al., 2019) and Hymenoptera (Thomas et al., 2020a). For each dataset rates were estimated using increasing values for K (Fig. 1a).For primates, the highest likelihood was found using K=4 rate categories, with λ=0.00453 and α=0.62. The birds of paradise dataset had the highest likelihood using K=10 rate categories, with λ=0.00226 and α=0.98. The Hymenoptera had the highest likelihood using K=6 rate categories, with λ=0.00375 and α=0.373. As expected (Gillespie, 1986; Golding, 1984; Yang, 1996) single-rate models consistently underestimate λ (Fig. 1a), highlighting the need to model rate variation. Although $K > 1$ always results in higher likelihoods, the maximum likelihood value does not always have the largest value of K considered; indeed, models with $K = 2$–3 often overestimate λ. This latter effect is likely due to the bifurcation of the data into one rate category for families that undergo little or no change and the other category accounting for all other families that change across the tree.

To assess the accuracy of the software and these results, we simulated three datasets (see Supplementary Material for simulation conditions) intended to match the distributions inferred from the empirical data. In all cases, CAFE 5 accurately estimated the maximum likelihood value of λ and of K (Fig. 1b). As in the empirical datasets, we also see that λ is consistently underestimated with K=1, and slightly overestimated for $K = 2$–3.

4 Summary

- CAFE 5 is now written in C++, a modular style of programming facilitating future development.
- Support for powerful compilers, parallelization and matrix multiplication allow CAFE 5 to take advantage of high-performance computing clusters.
- Accurate and fast joint ancestral state reconstruction is now available.
- Variation in the evolutionary rate among gene families is accounted for using a discrete approximation of the gamma distribution.

Accounting for rate variation among families using a discrete gamma approximation results in a better model fit and more accurate rate estimates. While this can be accomplished with as few as K=2 rate categories, we recommend testing K=3–4 categories with real data.

Acknowledgements

The authors thank Simon Whelan, Robert Henschel and Gregg Thomas for help throughout this project.

Funding

This work was supported by National Science Foundation [DBI-1564611 to M.W.H.].

Conflict of Interest: none declared.

References

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

Casola,C. and Lawing,A.M. (2019) The nonrandom evolution of gene families. *Am. J. Bot.*, **106**, 14–17.

De Bie,T. *et al.* (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, 1269–1271.

Demuth,J.P. *et al.* (2006) The evolution of mammalian gene families. *PLoS One*, **1**, e85.

Gibbs,R.A. *et al.*; Rhesus Macaque Genome Sequencing and Analysis Consortium. (2007) Evolutionary and biomedical insights from the Rhesus Macaque Genome. *Science*, **316**, 222–234.

Gillespie,J.H. (1986) Variability of evolutionary rates of DNA. *Genetics*, **113**, 1077–1091.

Golding,G.B. (1984) Estimates of DNA and protein sequence divergence: an examination of some assumptions. *Mol. Biol. Evol.*, **1**, 125–142.

Hahn,M.W. *et al.* (2005) Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.*, **15**, 1153–1160.

Hahn,M.W. *et al.* (2007) Gene family evolution across 12 Drosophila genomes. *PLoS Genet.*, **3**, e197.

Han,M.V. *et al.* (2013) Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.*, **30**, 1987–1997.

Librado,P. *et al.* (2012) BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics*, **28**, 279–281.

Prost,S. *et al.* (2019) Comparative analyses identify genomic features potentially involved in the evolution of birds-of-paradise. *GigaScience*, **8**, giz003.

Pupko,T. *et al.* (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.*, **17**, 890–896. https://doi.org/10.1093/oxfordjournals.molbev.a026369

Rubin,G.M. *et al.* (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.

Thomas,G.W.C. *et al.* (2020a) Gene content evolution in the arthropods. *Genome Biol.*, **21**, 15.

Thomas,G.W.C. *et al.* (2020b) Origins and long-term patterns of copy-number variation in rhesus macaques. *Mol. Biol. and Evol.*, *msaa303*. https://doi.org/10.1093/molbev/msaa303

Waterhouse,R.M. *et al.* (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.*, **41**, D358–D365. https://doi.org/10.1093/nar/gks1116

Waterston,R.H. *et al.*; Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.

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