BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 10 2006, pages 1269–1271
doi:10.1093/bioinformatics/btl097

Genome analysis

CAFE: a computational tool for the study of gene family evolution

Tijl De Bie[1], Nello Cristianini[2], Jeffery P. Demuth[3] and Matthew W. Hahn[3,*]

[1]K.U.Leuven, OKP Research Group, Tiensestraat 102, 3000 Leuven, Belgium, [2]Department of Statistics, U.C. Davis, 360 Kerr Hall, One Shields Avenue, Davis CA 95616, USA and [3]Department of Biology and School of Informatics, Indiana University, 1001 E. Third Street Bloomington, IN 47405, USA

Received on January 20, 2006; revised on March 9, 2006; accepted on March 10, 2006
Advance Access publication March 16, 2006
Associate Editor: Chris Stoeckert

ABSTRACT

Summary: We present CAFE (Computational Analysis of gene Family Evolution), a tool for the statistical analysis of the evolution of the size of gene families. It uses a stochastic birth and death process to model the evolution of gene family sizes over a phylogeny. For a specified phylogenetic tree, and given the gene family sizes in the extant species, CAFE can estimate the global birth and death rate of gene families, infer the most likely gene family size at all internal nodes, identify gene families that have accelerated rates of gain and loss (quantified by a *p*-value) and identify which branches cause the *p*-value to be small for significant families.

Availability: Software is available from http://www.bio.indiana.edu/~hahnlab/Software.html

Contact: mwh@indiana.edu

1 INTRODUCTION

The analysis of both whole genomes and single gene families has revealed an enormous amount of change in the size of families, even between closely related organisms (Tatusov *et al*., 1997). There is much interest in these changes, as even the gain or loss of single genes have been implicated in adaptive divergence between species (Olson, 1999). In addition, large contractions and expansions of gene families are generally attributed to natural selection, without a statistical basis for these claims [discussed in Hahn *et al*. (2005)]. In order to make inferences regarding both the direction and size of changes in gene family size, as well as whether large changes in size are truly evolutionarily significant, we introduce the program CAFE (Computational Analysis of gene Family Evolution): a tool for analyzing gene family size changes in a phylogenetic context.

The probabilistic model adopted in CAFE was introduced by Hahn *et al*. (2005); it uses a random birth and death process to model gene gain and loss along each lineage of a phylogenetic tree. In order to make inferences over a whole phylogeny, a probabilistic graphical model (Lauritzen, 1996; M. I. Jordan, manuscript in preparation) is used to calculate the probability of transitions in gene family size from parent to child nodes in the phylogeny. Using the graphical models machinery, one can draw inferences on the gene family size for all ancestral species. In particular, the specification of gene family sizes in the extant taxa in the tree is sufficient to estimate the ancestral gene family sizes (and therefore the direction of change on each branch), as well as to identify unusually evolving gene families, and to pinpoint the lineages along which the model has been violated for a specific gene family. CAFE can be run either through an easy-to-use graphical interface or a command-line version that both allow researchers to specify the analyses they wish to run (Fig. 1). A complete user's manual can be found at http://www.bio.indiana.edu/~hahnlab/Software.html

2 DESCRIPTION

CAFE's main inputs are a Newick description of a rooted and bifurcating phylogenetic tree (including branch lengths in units of time) and a data file containing the gene family sizes for the extant taxa. The data file may consist of data on one family or up to thousands of families for the specified tree. The first line of the data file should contain the extant species' names (as used in the Newick tree description), tab-delimited in no particular order. Subsequent lines each correspond to a gene family and contain tab-delimited family sizes for these extant species. Columns in the data file whose header does not correspond to any of the names in the Newick tree description, which may provide additional information about the gene families, are ignored and simply copied in the output file. Examples of Newick tree descriptions and corresponding data files can be found in the online user's manual.

A third input is λ, which is the probability of both gene gain and loss per gene per unit time in the phylogeny [CAFE assumes that gene birth and death are equally probable, see Hahn *et al*. (2005)]. The user can either specify λ by entering a single numerical value, or have CAFE find the maximum likelihood value given the gene families in the data file. It should be noted that the estimated λ is the globally most probable value across all families in the input file; separate estimates of λ for individual families can be calculated by running each family by itself. To estimate the maximum likelihood value, CAFE computes the likelihood of the data for 11 equidistant values of this parameter between two numerical values provided by the user (in the command-line version the number of equidistant values used can be changed). Given this initial range for λ, the value leading to the largest likelihood is used in the subsequent analysis; alternatively, one can ask CAFE to stop the analysis after computing this maximum likelihood estimate of λ. If the most likely value is one of the two most extreme values queried by CAFE, or if the initial range given is quite wide, it is recommended that estimation of λ be run again in an interval around the previous most probable value. A logfile output by CAFE contains the likelihood values for all queried values of λ.

*To whom correspondence should be addressed.

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

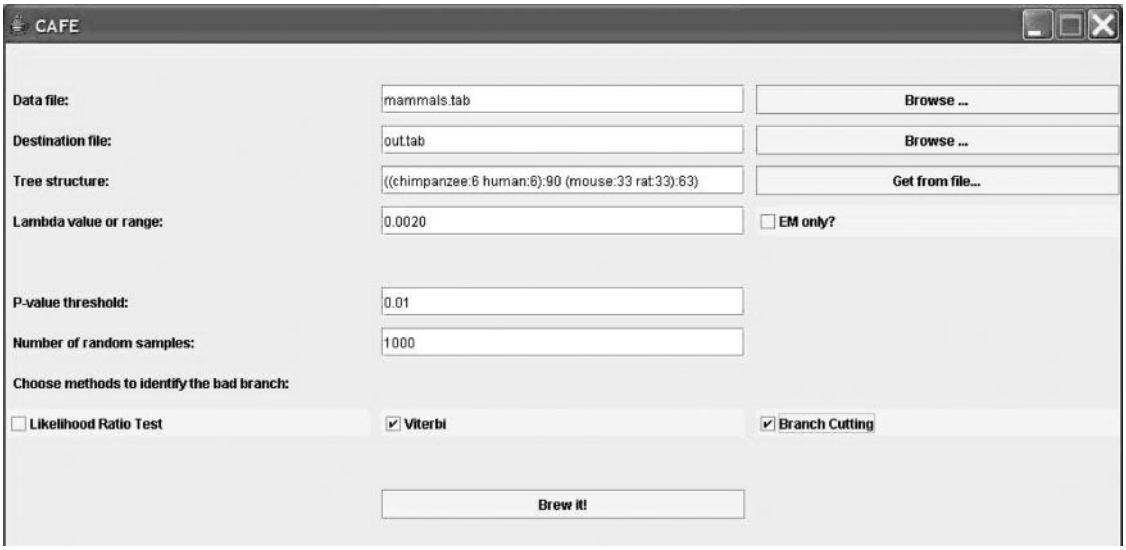

Fig. 1. The upper part of CAFE's graphical user interface. Required inputs are (see main text for a more detailed description of the inputs): the input data file; the output file; the Newick tree structure of the phylogeny; either a single value of λ to use or a range of values that λ can be optimized over; a checkbox which allows one to stop after the EM step ('EM only?'); the *p*-value threshold for families below which the branch-identification methods are run; the number of samples used in the Monte Carlo sampling steps; and three check boxes to decide which methods to use to determine the implicated branches for the gene families with *p*-values below the specified threshold. Not shown in the figure are the progress bars for the different steps of the analysis.

Given a phylogenetic tree, the gene family sizes in the extant species (data file), and the value for λ, a graphical model can be used to calculate the most likely family size in the ancestral species, and this for each family [see Hahn *et al*. (2005) for details]. CAFE calculates these so-called Viterbi assignments, and a comparison of these estimated sizes at all parent and descendant nodes allows one to infer the direction and size of change in gene family sizes along each branch. The Viterbi assignments are reported in the main output file. The average size of expansions along each branch of the tree (where negative values indicate an average contraction among all families), as well as the number of families that have no change, expand, or contract on each branch of the tree are reported in the logfile.

For each of the gene families in the data file, CAFE computes a *p*-value associated with the gene family sizes in the extant species given our model of gene family evolution. Families with a large variance in size, especially among closely related species, are likely to have low *p*-values. Gene families with low *p*-values are interesting, as large contractions or expansions may be associated with natural selection or with large duplications or deletions of stretches of chromosome containing multiple, related genes.

For those gene families with a small *p*-value, it is of interest to identify the branches of the tree where the largest changes have taken place (and hence where the model has been violated). CAFE incorporates three methods to identify these branches. While each of these methods is different in nature, it is our experience that in most cases they agree with each other. The first method ('Viterbi') uses the Viterbi assignments to the ancestral nodes, and subsequently computes a *p*-value for the transition from parent to child node along each branch of the tree. Branches with low *p*-values represent unusually large changes, either contractions or expansions. The second method ('branch cutting') calculates whether the overall *p*-value associated with a gene family increases if we cut one of the branches of the tree. By 'cutting' a branch we mean removing the probabilistic coupling between the parent and child family sizes for that branch. A *p*-value is then computed for the gene family given the tree with one branch removed as a model (and this is done for each branch separately). If the *p*-value increases considerably after cutting a branch, this branch may be held responsible for the overall low *p*-value of the complete model. The third method ('likelihood ratio test') maximizes the likelihood of the gene family by estimating a separate value for the evolutionary rate parameter, λ, along the branch under investigation. The ratio of the likelihood under the model with two parameters to the likelihood with just a single parameter can be used to assess the need for an extra parameter along individual branches. High values therefore indicate branches along which there has been a larger-than-expected amount of evolutionary change.

3 IMPLEMENTATION AND COMPUTATIONAL ISSUES

Both the GUI and command-line versions of CAFE are implemented in Java and operate as stand-alone tools. CAFE can be used on any Mac OSX, Windows or Linux machine running at least Java Virtual Machine 1.5. Output files include a record of the settings used ('userprofile.txt', only for the GUI version), a logfile of the analyses conducted ('logfile.txt'), and a specified main output file.

In order to calculate *p*-values for each gene family, Monte Carlo sampling must be used for computational feasibility. The number of samples used can be specified in CAFE, and 1000 is generally a sufficiently accurate and still computationally viable choice. The unavoidable use of Monte Carlo sampling means, however, that the birth and death model probabilities need to be computed many times with the same parent and child family sizes and evolutionary distance between them. Because this is time consuming, it is generally

beneficial to first cache the birth and death probabilities for all possible pairs of parent and child family sizes, and for each of the branch lengths in the phylogenetic tree. As it is impossible to cache the transition probabilities for unboundedly large gene family sizes, an upper bound needs to be chosen, which should be larger than the size of the largest gene family in all species in the phylogeny and this for all gene families in the data file. To this end, CAFE picks the largest gene family size among all extant species, and uses 1.2 times this value (or, 50 plus this value if this is larger) as the gene family size upper bound.

Since a birth and death transition probability is cached for each pair of parent and child family sizes and for each of the branch lengths, the total number of cached values is equal to the square of the family size upper bound times the number of differently sized branches. This makes the maximum likelihood estimation of λ (which involves a caching step for each value of λ investigated) and the caching step itself computationally the most demanding steps if the data file contains large gene families (maximization of λ for 10 000 families from 5 taxa, with a largest family of size 507, on a Macintosh 2.7 GHz Dual PowerPC with 6 GB of RAM took $\sim$10 h). Calculations scale linearly with the number of species used (Hahn *et al*., 2005). Estimation of the most probable ancestral states and calculation of the *p*-values for individual branches using both the Viterbi and branch-cutting methods are generally fast and are never computational bottlenecks. Part of the reason for this is that the demanding work has already been carried out in previous steps. Finally, the calculations for the likelihood ratio test are usually time consuming, as the most probable value of λ must be computed for each branch and for each family considered. To this end, additional caching for different values of λ is necessary (calculation of the likelihood ratio test for 150 families with *p*-values <0.0001 from the above dataset on the same machine took $\sim$5.5 h). To avoid unnecessary calculations, users can specify which of the various methods to identify significant branches should be used, as well as a minimum *p*-value above which families are not considered for further branch-specific analyses.

ACKNOWLEDGEMENTS

The authors acknowledge the contributions of Chi Nguyen in the development of a previous version of the algorithms, and Jason Stajich for input into the development of the software. T.D.B. acknowledges support from the CoE EF/05/007 SymBioSys, and from GOA/2005/04, both from the Research Council K.U.Leuven. N.C. acknowledges support from NIH grant R33HG003070-01. M.W.H. and J.P.D. are supported by NSF grant MCB-0528465 and by the METACyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc.

Conflict of Interest: none declared.

REFERENCES

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

Lauritzen,S.L. (1996) *Graphical Models*. Clarendon Press, Oxford, UK.

Olson,M.V. (1999) When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet*., **64**, 18–23.

Tatusov,R.L. *et al*. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.