BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 2 2012, pages 279–281
doi:10.1093/bioinformatics/btr623

Genome analysis

Advance Access publication November 10, 2011

BadiRate: estimating family turnover rates by likelihood-based methods

P. Librado, F. G. Vieira and J. Rozas*

Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain

Associate Editor: David Posada

ABSTRACT

Motivation: The comparative analysis of gene gain and loss rates is critical for understanding the role of natural selection and adaptation in shaping gene family sizes. Studying complete genome data from closely related species allows accurate estimation of gene family turnover rates. Current methods and software tools, however, are not well designed for dealing with certain kinds of functional elements, such as microRNAs or transcription factor binding sites.

Results: Here, we describe BadiRate, a new software tool to estimate family turnover rates, as well as the number of elements in internal phylogenetic nodes, by likelihood-based methods and parsimony. It implements two stochastic population models, which provide the appropriate statistical framework for testing hypothesis, such as lineage-specific gene family expansions or contractions. We have assessed the accuracy of BadiRate by computer simulations, and have also illustrated its functionality by analyzing a representative empirical dataset.

Availability: BadiRate software and documentation is available from http://www.ub.edu/softevol/badirate.

Contact: jrozas@ub.edu

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

Received on July 24, 2011; revised on November 4, 2011; accepted on November 7, 2011

1 INTRODUCTION

It is generally accepted that gene and genome duplications are a major evolutionary mechanism for generating functional innovation (Ohno, 1970). The increasing availability of closely related genome sequences allows an accurate analysis of gene family evolution (Hahn *et al*., 2007; Sanchez-Gracia *et al*., 2009; Vieira and Rozas, 2010). Such studies have shown that most families are highly dynamic and evolve under a birth-and-death (BD) process (Nei and Rooney, 2005). Indeed, the comprehensive analysis of gene gains and losses can provide helpful insight into the role of natural selection and adaptation in shaping gene family size variation.

The stochastic BD model (BDM) (Hahn *et al*., 2005) implemented in the programs CAFE (De Bie *et al*., 2006) and BEGFE (Liu *et al*., 2011) allows estimating the family turnover rate (λ) by maximum likelihood (ML) and by Bayesian methods, respectively. This model, nevertheless, has some drawbacks. First, it assumes equal BD rates, an assumption that may not hold. Secondly, because duplications from zero ancestral genes are not possible (zero is an absorbing state in the probabilistic BDM), it cannot handle gene families without elements in the phylogenetic root. These assumptions can therefore bias the estimates of both the number of members in internal nodes, as well as the BD rate. Two recently developed computer programs, GLOOME (Cohen *et al*., 2010) and Count (Csuros, 2010), overcome these difficulties. Nevertheless, GLOOME can only model presence/absence of phyletic patterns instead of size changes, whereas Count assumes independent turnover rates for all lineages, which precludes testing biological relevant hypothesis such as linage-specific accelerations.

Here, we describe BadiRate, a new software tool to estimate family turnover rates through the Gain-and-Death (GD) and the Birth-Death-and-Innovation (BDI, also known as Birth-Death-and-Immigration) stochastic models. The current implementation allows modeling families of diverse functional elements, such as microRNAs, *cis*-regulatory elements or coding-protein genes. Additionally, these models provide a statistical framework for hypothesis testing, such as family expansions/contractions in specific lineages.

2 DESCRIPTION

BadiRate implements methods to estimate the family turnover rates such as, gain (γ), birth (β), death (δ) and innovation (ι) rates. These rates can be classified as density-dependent (BD) and density-independent (gain and innovation). Indeed, the probability of having a death (e.g. a gene loss via deletion or pseudogenization) or a birth (e.g. a gene gain by unequal crossing-over) event is proportional to the actual family size. Conversely, the probability of having a gain or innovation [e.g. horizontal gene transfer (HGT) or by *de novo* origin of short *cis*-regulatory elements] event does not depend on the actual gene number.

We have implemented two stochastic population models in BadiRate, the BDI and the GD models (Csuros and Miklos, 2009; Hahn *et al*., 2005) (for details see Supplementary Material), to analyze the evolution of such diverse functional elements. The GD model is especially suitable for analyzing families where members might have a *de novo* origin, such as transcription factor binding sites (TFBSs) and small non-coding RNAs (miRNAs, piRNAs, etc) or even families exhibiting high HGT. In contrast, the BDI model is appropriated to study gene families whose major mechanism for the acquisition of genes is density-dependent (although it can also model scenarios with a reduced number of HGT or *de novo* origin events). BadiRate also implements a particular case of the BDI model in which BD rates are assumed to be equal, the Lambda-Innovation

*To whom correspondence should be addressed.

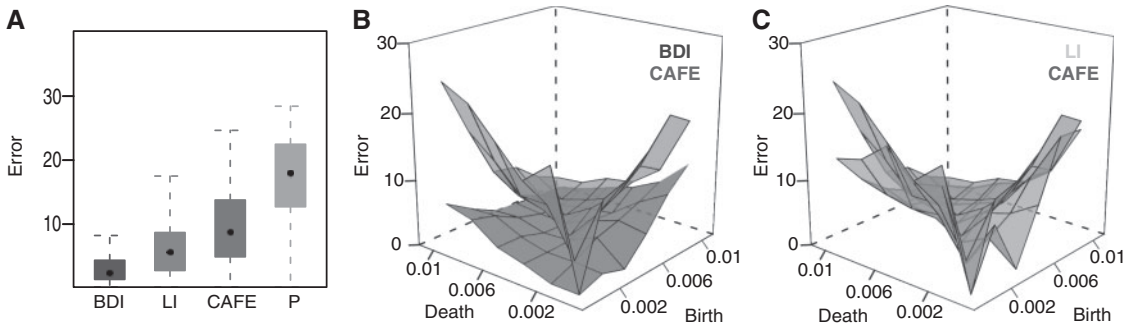

Fig. 1. Relative performance of the different methods. Computer simulations based on gene families of five members ($S = 5$) at the most internal node, and a null innovation rate ($I = 0$). (**A**) The box plot represents the error values (see Supplementary Material) distribution across tested methods. (**B** and **C**) Surface plots showing the error values of the ML estimates in scenarios simulated under a combination biologically realistic BD rates. Analysis under the BDI, LI and P models are conducted with BadiRate and are depicted in blue, green and cyan, respectively. The model implemented in CAFE is depicted in red. BD rates are measured in number of events per gene per million years (See Supplementary Material for details).

model (LI). This model is nearly equivalent to that implemented in CAFE (except for the innovation parameter), which allows the comparison between the two programs.

We have implemented three statistical frameworks to estimate family turnover rates and the number of members in internal nodes: maximum a posteriori (MAP), ML and parsimony (see Supplementary Material). Likelihood-based methods have the advantage of contrasting biological relevant scenarios, such as the identification of gene families or specific-lineages with extreme turnover rates (Johnson and Omland, 2004). Moreover, the MAP approach allows the incorporation of prior biological information without a large computational cost.

BadiRate requires as input the established species phylogenetic tree and a tab-delimited file with the family size of each species represented in the phylogeny (see BadiRate's documentation).

3 SIMULATION RESULTS

We assessed the accuracy of the turnover rates estimates by computer simulations on the well-characterized 12 *Drosophila species* phylogeny (Supplementary Material; Supplementary Figures S1, S2, S3, S4 and S5). Particularly, we benchmarked the BadiRate ML (under the BDI, LI and GD models) and parsimony estimates, as well as the CAFE (v2.2) ML estimates (under the implemented BDM model).

Our results show that, in general (and as expected), the parsimony algorithm performs worse than ML methods (Fig. 1A). Among the ML models, the BDI method outperforms all others (LI and CAFE), especially in cases where small-size families have asymmetric birth/death rates, i.e. $\beta > \delta$ or vice versa (Fig. 1B; Supplementary Figure S2). Apart from the higher accuracy of the BDI model, which mainly results from the separate estimation of BD rates, the LI model also outperforms CAFE even in scenarios with a null innovation rate (Fig. 1A and C; Supplementary Figure S3). Unlike LI and CAFE, which are particular cases of the BDI model, the performance of the GD model is not directly comparable to BDI. Still, our simulations show good performance of the GD model in the analyzed scenarios (Supplementary Figure S4).

4 EMPIRICAL RESULTS

We also illustrate the BadiRate application by analyzing the suggested miRNA expansion in the *D.willistoni* lineage (Nozawa *et al*., 2010). Since the identification of the miRNA copies in the 11 *Drosophila* species was conducted by similarity based on the available *D.melanogaster* miRNA data, the identification of the family members is less accurate for longer divergence times. To control for this effect, we contrasted two scenarios, one assuming independent turnover rates in two classes of branches (in the internal lineages leading to *D.melanogaster* and in the rest of branches) and the other incorporating a third class of turnover rates (for the *D.willistoni* lineage). The lower Akaike Information Criterion (AIC) value of the second scenario (AIC = 1509.8128) compared with the first one (AIC = 1518.3918) suggests that the *D.willistoni* lineage indeed has distinct miRNA turnover rates. We also inferred the most likely number of miRNA elements in the internal nodes of the *Drosophila* phylogeny; these figures are very similar to that estimated in (Nozawa *et al*., 2010) (Supplementary Figure S6).

ACKNOWLEDGEMENTS

We thank F.C. Almeida, M.C. Frias-Lopez, S. Guirao-Rico, A. Sanchez-Gracia and V. Soria-Carrasco, and three anonymous reviewers for their comments and suggestions on the manuscript and on the software.

Funding: This work was supported by grants from the Ministerio de Ciencia e Innovación of Spain (BFU2007-62927 and BFU2010-15484) and from the Comissió Interdepartamental de Recerca i Innovació Tecnològica of Spain (2009SGR-1287). J.R. was partially supported by ICREA Academia (Generalitat de Catalunya).

Conflict of Interest: none declared.

REFERENCES

Cohen,O. *et al*. (2010) GLOOME: gain loss mapping engine. *Bioinformatics*, **26**, 2914–2915.

Csuros,M. (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, **26**, 1910–1912.

Csuros,M. and Miklos,I. (2009) Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol. Biol. Evol.*, **26**, 2087–2095.

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

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.

Johnson,J.B. and Omland,K.S. (2004) Model selection in ecology and evolution. *Trends Ecol. Evol.*, **19**, 101–108.

Liu,L. *et al*. (2011) A Bayesian model for gene family evolution. *BMC Bioinformatics*, **12**, 426.

Nei,M. and Rooney,A.P. (2005) Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.*, **39**, 121–152.

Nozawa,M. *et al*. (2010) Origins and evolution of microRNA genes in Drosophila species. *Genome Biol. Evol.*, **2**, 180–189.

Ohno,S. (1970) *Evolution by Gene Duplication*. Springer, Berlin.

Sanchez-Gracia,A. *et al*. (2009) Molecular evolution of the major chemosensory gene families in insects. *Heredity*, **103**, 208–216.

Vieira,F.G. and Rozas,J. (2010) Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biol. Evol.*, **3**, 476–490.