ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data

Jaime Huerta-Cepas,*1 François Serra,2 and Peer Bork1,3,4

1Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany
2Centro Nacional de Análisis Genómico (CNAG-CRG), Center for Genomic Regulation, Universitat Pompeu Fabra (UPF), 08028 Barcelona, Spain
3Germany Molecular Medicine Partnership Unit (MMPU), University Hospital Heidelberg and European Molecular Biology Laboratory, 69117 Heidelberg, Germany
4Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany

*Corresponding author: E-mail: huerta@embl.de
Associate editor: Tal Pupko

Abstract

The Environment for Tree Exploration (ETE) is a computational framework that simplifies the reconstruction, analysis, and visualization of phylogenetic trees and multiple sequence alignments. Here, we present ETE v3, featuring numerous improvements in the underlying library of methods, and providing a novel set of standalone tools to perform common tasks in comparative genomics and phylogenetics. The new features include (i) building gene-based and supermatrix-based phylogenies using a single command, (ii) testing and visualizing evolutionary models, (iii) calculating distances between trees of different size or including duplications, and (iv) providing seamless integration with the NCBI taxonomy database. ETE is freely available at http://etetoolkit.org

Key words: phylogenomics, tree visualization, tree comparison, NCBI taxonomy, hypothesis testing, phylogenetics.

The Environment for Tree Exploration (ETE) is a toolkit developed to facilitate the computation, analysis and visualization of phylogenetic data. ETE provides a comprehensive Python programming library (API) that allows researchers to automate common tasks in comparative genomics. Since its first release (Huerta-Cepas et al. 2010), ETE has been widely used as a computational framework to perform numerous phylogenomic analyses, including characterizing newly sequenced genomes (Richards et al. 2010; Wang et al. 2014), extracting information from large sets of phylogenetic trees (Derele and Lang 2012; Chiapello et al. 2015; Marcet-Houben and Gabaldón 2015) and developing third party tools and databases (Zhang et al. 2013; Huerta-Cepas et al. 2014; Szitzenberg et al. 2015). Here, we describe the latest version of the software (ETE v3), featuring a significantly improved API library and a novel collection of standalone tools. While the API continues to offer full programmatic control on data analysis and visualization, the new standalone tools facilitate the use of common phylogenetic methods at the genomic scale. We here describe the most notable additions.

Tree Building

The ete-build tool provides a unified interface to wrap the execution of reproducible phylogenetic workflows, comprising the reconstruction of gene–trees and supermatrix-based species trees. To do so, ETE relies on a versioned collection of external tools that are transparently installed and executed upon request. A single command is used to configure and launch complex phylogenetic pipelines, covering sequence alignment, trimming, substitution-model testing, tree inference, and image rendering (fig. 1A). In addition, the supermatrix-based reconstruction mode permits to build and concatenate multiple sequence alignments with ease, simplifying the inference of species trees based on multiple genes. Advanced options allow to automatically switch from amino-acid to nucleotide alignments based on sequence identity, resuming the execution of workflows, or even testing multiple strategies in parallel. As an example, a single command line can be used to test several alignment methodologies or phylogenetic inference programs simultaneously, making the tool particularly suitable to run phylogenomic pipelines. Notably, ETE-build was recently used to compute over one million phylogenetic trees for the EggNOG v4.5 database (Huerta-Cepas et al. 2016).

Testing Evolutionary Hypotheses

Measuring selective pressures on molecular sequences is a common task in evolutionary biology. Softwares such as CodeML (Yang 2007) or SLR (Massingham and Goldman 2005) provide the statistical and computational framework to perform these analyses. However, the use of such tools at the genomic scale requires substantial work on data preparation, on experimental design, and on results interpretation. To aid in these tasks, the ete-evol tool automates CodeML/SLR-based analyses by using pre-configured evolutionary models and directly producing a graphical representation of
the results. These pre-configured models include site (Yang et al. 2000; Massingham and Goldman 2005), branch (Yang and Nielsen 2002), branch-site (Zhang et al. 2005), and clade (Yang and Nielsen 2002; Bielawski and Yang 2004) models. For instance, ete-evol can test, in parallel, and with a single call, the differential selective pressures along each branch in a given phylogeny. Importantly, fitted models are compared using a built-in likelihood ratio test. Evolutionary measures from the best-fitting models are then plotted (or interactively visualized) by mapping the predicted selective pressures acting on sites and branches into the tested topology, as well as on the multiple sequence alignment (fig. 1B). For
Many common tasks in phylogenomics for both expert and
tools, ETE has significantly broadened its scope, simplifying
data while providing unique visualization capabilities.

In contrast to existing software (Felsenstein 2005; Soria-Carrasco et al. 2007), ete-compare calculates all three distances at the same time; it accepts trees varying in size and containing duplication events; it allows filtering branches with low support; and it is optimized for comparing large datasets. In addition, ete-compare can provide a detailed list of the differences and co-incidences among the compared trees for further analysis. Conveniently, the TreeKO method for splitting gene trees into duplication-free subtrees has been optimized and integrated into ETE’s API library, thereby enabling its use for other tests. For instance, ETE allows summarizing the phylogenetic signal (i.e., gene tree support) from an heterogeneous sample of gene trees using a species tree topology as reference (fig. 1C).

Taxonomy Databases

Efficient queries to the NCBI-taxonomy database (Benson et al. 2014) are now available through the ete-ncbiquery tool or the relevant methods in the API. Extracting pruned subtrees, converting NCBI taxids into their corresponding scientific names, obtaining full lineage tracks, or annotating user-trees with taxonomic data are common tasks that can be easily performed with the ete-ncbiquery tool. Importantly, all queries are carried out locally, avoiding unnecessary lags and permitting the integration of the tool into genomic and metagenomic pipelines.

Finally, other ETE-tools and methods are available that aid in routine tasks such as format conversion, topology manipulation, and custom visualization of trees linked to multiple sequence alignments (fig. 1D).

Conclusions

Although several software packages are available for the standalone exploration of trees (Letunic and Bork 2007; Huson and Scornavacca 2012; Asnicar et al. 2015) and the programmatic manipulation of data (Paradis et al. 2004; Knight et al. 2007; Sukumaran and Holder 2010; Vos et al. 2011; Talevich et al. 2012), ETE offers a unified framework to compute and analyze genome-wide collections of evolutionary data while providing unique visualization capabilities. Moreover, with the recent addition of the command line tools, ETE has significantly broadened its scope, simplifying many common tasks in phylogenomics for both expert and casual users.

Acknowledgments

We wish to thank Toni Gabaldón, Renato Alves, Falk Hildebrand, and Gabriela Aguilera for valuable feedback, as well as the GitHub community for contributions. This study was supported by the European Molecular Biology Laboratory (EMBL). Funding for open access charge: European Molecular Biology Laboratory (EMBL).

References

Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. 2015. Compact graphical representation of phylogenetic data and metadata with GraPhAn. Peer J. 3:e1029.
Benson DA, Clark K, Karchs-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2014. GenBank. Nucleic Acids Res. 42:D32–D37.
Bielawski JP, Yang Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J. Mol. Evol. 59:121–132.
Carrasco V, Talavera G, Igea J, Castresana J. 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23:2954–2956.
Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. 2015. Compact graphical representation of phylogenetic data and metadata with GraPhAn. Peer J. 3:e1029.
Benson DA, Clark K, Karchs-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2014. GenBank. Nucleic Acids Res. 42:D32–D37.
Bielawski JP, Yang Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J. Mol. Evol. 59:121–132.
Carrasco V, Talavera G, Igea J, Castresana J. 2007. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23:2954–2956.
Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.
Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.
Talevich E, Invergo B, Cock P, Chapman B. 2012. Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics* 13:209.

Vos RA, Caravas J, Hartmann K, Jensen MA, Miller C. 2011. BIO::Phylo-phyloinformatic analysis using perl. *BMC Bioinformatics* 12:63.

Wang W, Haberer G, Gundlach H, Gläßer C, Nussbaumer T, Luo MC, Lomsadze A, Borodovsky M, Kerstetter RA, Shanklin J, et al. 2014. The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat. Commun.* 5:3311.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19:908–917.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

Zhang J, Kapli P, Pavlidis P, Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* 29:2869–2876.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22:2472–2479.

Huerta-Cepas et al. · doi:10.1093/molbev/msw046

MBE