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

Jaime Huerta-Cepas, François Serra, and Peer Bork

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.

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).
Comparing Trees

ETE v3 provides three measures to compute distances between trees, namely the Robinson–Foulds distance (Robinson and Foulds 1981), a branch congruence measure (%) and the TreeKO Speciation distance (Marcet-Houben and Gabaldón 2011). 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 coincidences 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 heterogenous 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


