# FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program

Vincent Lefort,[1] Richard Desper,[1] and Olivier Gascuel*,[1]

[1]Institut de Biologie Computationnelle, LIRMM, UMR 5506: CNRS & Université de Montpellier, France

***Corresponding author:** E-mail: gascuel@lirmm.fr.

**Associate editor:** Michael Rosenberg

## Abstract

**FastME provides distance algorithms to infer phylogenies. FastME is based on balanced minimum evolution, which is the very principle of Neighbor Joining (NJ). FastME improves over NJ by performing topological moves using fast, sophisticated algorithms. The first version of FastME only included Nearest Neighbor Interchange. The new 2.0 version also includes Subtree Pruning and Regrafting, while remaining as fast as NJ and providing a number of facilities: Distance estimation for DNA and proteins with various models and options, bootstrapping, and parallel computations. FastME is available using several interfaces: Command-line (to be integrated in pipelines), PHYLIP-like, and a Web server (http://www.atgc-montpellier.fr/fastme/).**

***Key words:*** **phylogeny inference, distance-based, fast algorithms, (balanced) minimum evolution, NNI and SPR topological moves.**

Distance algorithms infer phylogenies from matrices of pairwise distances among taxa. These algorithms are fast and have been shown to be fairly accurate using both real and simulated data (e.g., Kuhner and Felsenstein 1994). Moreover, they account for probabilistic modeling of substitutions while estimating evolutionary distances. Even if they are not as accurate as likelihood-based methods, these algorithms are still widely used due to their speed and simplicity, as assessed by the high number of citations for Neighbor Joining (NJ, Saitou and Nei 1987; see also Studier and Keppler 1988): Approximately 2,000 in 2014 (Web of Science).

NJ is a greedy algorithm that builds trees by iterative agglomeration of taxa. Gascuel and Steel (2006) showed that the criterion being minimized by NJ is the balanced version of minimum evolution (BME), which estimates the tree length using Pauplin's formula (2000). We proposed fast, BME-based algorithms (Desper and Gascuel 2002, 2004) to 1) construct an initial tree using greedy taxon insertion and 2) perform topological moves, namely Nearest Neighbor Interchanges (NNIs), to improve an initial (e.g., NJ) tree. These algorithms were implemented in FastME 1.0 and were shown to improve accuracy substantially in comparison to NJ's (e.g., Vinh and von Haeseler 2005), while having a similar computational cost. A related NNI-based approach, using profiles of ancestral sequences instead of a distance matrix, was proposed by Price et al. (2009) and implemented in FastTree1. FastME has been developed over the past several years:

- Subtree Pruning and Regrafting (SPR) topological moves are available in FastME 2.0. SPR consists of removing a subtree from the initial tree and reinserting this subtree by dividing any of the remaining branches in the initial tree. We thus have $O(n^2)$ alternative trees to improve the initial tree, where $n$ is the number of taxa. The best SPR is selected and the procedure is iterated until no more improving SPR is found. SPRs are more powerful than NNIs (with $O(n)$ alternative trees) and have been shown to be useful in a number of contexts and studies (e.g., with maximum-likelihood [ML]-based tree building; Guindon et al. 2010). Our algorithm first precomputes the average distance between every pair of subtrees of the initial topology; this can be achieved in $O(n^2)$ time. Then, the criterion value for any new tree obtained by SPR is computed in constant time, meaning that the total cost of the SPR-based tree search is $O(kn^2)$, where $k$ is the number of iterations. As $k$ is usually smaller than $n$, the computational cost is similar to that of NJ, that is, $O(n^3)$. Experiments with real data (both DNA and proteins) show that a substantial gain is obtained, compared with NJ and NJ+NNIs; the best alternative is FastTree1, which (quickly) infers trees that are less fitted than NJ+SPR's regarding minimum evolution, but have similar likelihood value with DNA sequences. Details on our SPR algorithm and these experiments are provided in Supplementary Material online.
- A number of tree-building algorithms have been added, to infer an initial tree or to improve that tree (or any input tree) with topological moves. These algorithms seek to optimize BME, but also the Ordinary Least Square version of minimum evolution (OLSME; Rzhetsky and Nei 1993), which may be relevant with nonsequence data. These algorithms and their properties are summarized in table 1.
- The calculation of evolutionary distance matrices from DNA and protein sequences is also available. For DNA, most models having an analytical solution (e.g., TN93) have been implemented. For protein sequences, we use

**Table 1.** Substitution Models and Algorithms Available in FastME 2.0.

| Models | | | |
|---|---|---|---|
| | | **Target** | **Method** |
| **DNA** | p-distance | General | Analytical formula |
| | RY symmetric | | |
| | RY | | |
| | JC69 (Jukes, *Mam. Prot. Metab.*, 1969) | | |
| | K2P (Kimura, *J. Mol. Evol.*, 1980) | | |
| | F81 (Felsenstein, *J. Mol. Evol.*, 1981) | | |
| | F84 (Felsenstein, *Evolution*, 1984) | | |
| | TN93 (Tamura, *MBE*, 1993) | | |
| | LogDet (Lockhart, *MBE*, 1994) | | |
| **Protein** | p-distance | General | Analytical formula |
| | F81-like | General | Analytical formula |
| | LG (Le, *MBE*, 2008) | General | ML estimation |
| | WAG (Whelan, *MBE*, 2001) | General | ML estimation |
| | JTT (Jones, *CABIOS*, 1992) | General | ML estimation |
| | Dayhoff (Dayhoff, *A. Prot. Seq. Struct.*, 1978) | General | ML estimation |
| | DCMut (Kosiol, *MBE*, 2004) | General | ML estimation |
| | CpRev (Adachi, *J. Mol. Evol.*, 2000) | Chloroplast | ML estimation |
| | MtREV (Adachi, *J. Mol. Evol.*, 1996) | Mitochondria | ML estimation |
| | RtREV (Dimmic, *J. Mol. Evol.*, 2002) | Retrovirus | ML estimation |
| | HIVb/w (Nickle, *PLoS One*, 2007) | HIV | ML estimation |
| | FLU (Dang et al., *BMC Evol. Biol.*, 2010) | Flu | ML estimation |
| **Algorithms** | | | |
| | | **Optimization Criterion** | **Method and Complexity** |
| **First tree** | BME (Desper, *J. Comp. Biol.*, 2002) | BME | Taxon addition $O(n^2)$ |
| | GME (Desper, *J. Comp. Biol.*, 2002) | OLSME | Taxon addition $O(n^2)$ |
| | NJ (Saitou, *MBE*, 1987) | BME | Agglomerative $O(n^3)$ |
| | UNJ (Gascuel, *Math. Hierarchies & Biol.*, 1997) | OLSME | Agglomerative $O(n^3)$ |
| | BioNJ (Gascuel, *MBE*, 1997) | — | Agglomerative $O(n^3)$ |
| **Topo. moves** | BNNI (Desper, *J. Comp. Biol.*, 2002) | BME | NNI $O(kn^2)$ |
| | FASTNNI (Desper, *J. Comp. Biol.*, 2002) | OLSME | NNI $O(kn^2)$ |
| | SPR | BME | SPR $O(kn^2)$ |

NOTE.—All models (except p-distance and LogDet) can be used with a continuous gamma distribution of rates across sites with user-defined parameter (typically 1.0). We distinguish models where a fast analytical formula is available to estimate evolutionary distances, from those (slower) requiring maximization of the likelihood function. For algorithms, we distinguish 1) the criterion being optimized (BME or OLSME) and 2) the construction of a first tree (using iterative taxon addition, or the agglomerative [NJ] scheme) versus the improvement of this initial tree using topological moves (NNIs or SPRs). We display worst case time complexities (as usual); $n$ is the number of taxa and $k$ the number of iterations. With NNIs, $k$ is usually similar to $n$. With SPRs, $k$ is usually much smaller than $n$.

standard ML-based estimations, combined with a number of rate matrices (e.g., JTT [Jones, Taylor, and Thorton]) to accommodate various data sets (mitochondria, virus, etc.). In both cases, distances can be estimated assuming a continuous gamma distribution of rates across sites with user-defined parameter. Models and options are summarized in table 1.

- Bootstrapping and analysis of multiple data sets can be performed within a single run. FastME 2.0 implements Felsenstein's bootstrap, where pseudo trees are built from resampled alignments and compared with the original tree obtained from the input alignment. Users can also submit a unique file containing multiple alignments (e.g., corresponding to different genes in phylogenomics studies) and launch tree construction for all of them using the same program options.
- Bootstrapping is a highly parallelizable task. The same holds for distance estimations. FastME 2.0 provides parallel computing for these two tasks using the OpenMP API. When compiling FastME, users can choose to obtain a mono-thread or a parallel binary. They may then set, on the command line, the number of cores to be used.
- FastME 2.0 includes a menu-driven PHYLIP-like interface, and a command-line interface, to be typically integrated in phylogenomics pipelines. A Web server is also available for occasional users. FastME is an open-source C program, with binaries available for the three main operating systems.

FastME 2.0 is thus a comprehensive program, including all required tools (numerous algorithms, distance estimation with various models, bootstrapping) to infer phylogenies using a distance approach. Source code, binaries, Web server, user guide, examples, benchmark data sets, etc., are available from http://www.atgc-montpellier.fr/fastme/ (last accessed July 14, 2015).

## Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgment

This research was supported by the Institut Français de Bioinformatique (RENABI-IFB, Investissements d’Avenir).

## References

Desper R, Gascuel O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comp Biol.* 9:687–705.

Desper R, Gascuel O. 2004. Theoretical foundations of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol Biol Evol.* 21:587-598.

Gascuel O, Steel M. 2006 Neighbor-joining revealed. *Mol Biol Evol.* 23:1997–2000.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.

Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol.* 11:459-468.

Pauplin Y. 2000. Direct calculation of a tree length using a distance matrix. *J Mol Evol.* 51:41–47.

Price MN, Dehal PS, Arkin AP. 2009 FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 26:1641-1650.

Rzhetsky A, Nei M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol Biol Evol.* 10:1073-1095.

Saitou N, Nei M. 1987 The neighbor-joining method: a new method for reconstruction of phylogenetic trees. *Mol Biol Evol.* 4:406-425.

Studier JA, Keppler KJ. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol.* 5:729-731.

Vinh LS, von Haeseler A. 2005. Shortest triplet clustering: reconstructing large phylogenies using representative sets. *BMC Bioinformatics* 6:92.