FastME – Manual

Version 2.1.5

June 2016

http://www.atgc-montpellier.fr/fastme

Contents

1. Citations

- « FastME 2.0: a comprehensive, accurate and fast distance-based phylogeny inference program » Lefort V, Desper R., Gascuel O. 2015, *Molecular Biology and Evolution*, **32**(10):2798-800

1.1. Related papers

- « Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle » Desper R., Gascuel O. 2002, *Journal of Computational Biology*, **9**(5):687-705
- « Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting » Desper R., Gascuel O. 2004, *Molecular Biology and Evolution*, **21**(3):587-598
- « The neighbor-joining method: a new method for reconstructing phylogenetic trees » Saitou N., Nei M. 1987, *Molecular Biology and Evolution,* **4**(4):406-25
- « BioNJ: an improved version of the NJ algorithm based on a simple model of sequence data » Gascuel O. 1997, *Molecular Biology and Evolution,* **14**(7):685-695
- « Concerning the NJ algorithm and its unweighted version, UNJ » Gascuel O. 1997 Pp. 149-170 in Mathematical Hierarchies and Biology (B. Mirkin, F.R. McMorris, F.S. Roberts and A. Rzhetsky, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Amer. Math. Soc., Providence, RI.

2. Availability

- Binaries and source code are available from : http://www.atgc-montpellier.fr/fastme

3. Authors

- Rick Desper and Olivier Gascuel conceived the original FastME algorithm.
- Rick Desper, Olivier Gascuel and Vincent Lefort implemented FastME.
- Vincent Lefort created the benchmark and implemented the tools that are used to check FastME accuracy and performance.
- Vincent Lefort conceived and implemented FastME web server.
- Vincent Lefort and Olivier Gascuel wrote this document (re-using in several places PhyML user guide, written by S. Guindon).

4. Overview

FastME is a software package whose main task is to estimate phylogenies using distance methods from nucleotide or amino acid multiple sequences alignments (MSA). It provides a wide range of options that were designed to ease standard phylogenetic analyses. One of the main strength of FastME lies in the availability of several distance algorithms and optimization principles (OLS and Balanced Minimum Evolution, iterative taxon addition, NJ, UNJ, BioNJ) for tree estimation coupled with various options to search the space of phylogenetic tree topologies (NNIs, SPRs). It also provides a parallelized implementation of the non-parametric bootstrap method to evaluate branch supports.

All algorithms in FastME are fast, with time complexity of $O(n^3)$ or less. The “comfort zone” generally lies around 2,000-10,000 sequences with standard computers. Actually, the bottleneck is in computing and storing the distance matrix, which requires $O(n^2)$ memory space (it takes about 10 minutes and 900 Gb to compute distances for a DNA MSA of 3800 taxa and 1400 sites),

5. Bug report

If you ever come across an issue, please feel free to send an email to lefort@lirmm.fr. Do not forget to mention the version of FastME, the program options you are using and attach the input data.

6. Installing FastME

6.1. Pre-compiled binaries

FastME pre-compiled binaries for the most standard OS (Linux 32 and 64 bits, macOS and Window 7) are freely distributed and available from the FastME web site.

6.2. Sources, compilation and installation

The sources are distributed free of charge in accordance with the GPL terms.

The compilation on UNIX-like systems is fairly standard. It is described in the INSTALL file that comes with the sources. In a command-line window, go to the directory that contains the sources and type :

```
./configure
make
```

By default, FastME will be compiled with optimization flags turned on. It is possible to generate a binary that can run through debugging tools (such as `gdb`) using the following instructions :

```
./configure –enable-debug
make
```

By default, FastME will be compiled with parallelization flags turned on. It is possible to generate a mono-threaded binary using the following instructions :

```
./configure –disable-OpenMP
make
```

If you wish to install FastME on your operating system type :

```
./configure
make install
```

7. Program usage

The downloadable version of FastME has two distinct user-interfaces : a PHYLIP-like text interface that makes the choice of the options self-explanatory and a standard CLI which is well-suited for people that are familiar with the command line or for running FastME in batch mode. FastME can also be launched online (http://www.atgc-montpellier.fr/fastme/) with the same options as the PHYLIP-like interface.

7.1. PHYLIP-like interface

The default is to use the PHYLIP-like text interface by simply running the FastME binary in a command line window. After entering the name of the input file, a menu of options is displayed to set up the analysis. Each option can be configured by typing the corresponding letter on the left column of the menu. Once the options have been defined, typing 'Y' (or 'y') launches the computation. The meaning of some options may not be obvious to users that are not familiar with phylogenetics. In such situation, we strongly recommend to use the default options. As long as the format of the input file is correctly specified ('Input data type' option), the safest option for non-expert users is to use the default settings. The different options are described in what follows.

7.1.1. Input data type

```
I          Input data type (distance matrix or sequence alignment)
```

Type of data in the input file. It can be either DNA or amino-acid MSA in PHYLIP format, or a distance matrix in PHYLIP format. If the input data type is a MSA, the menu will update to display evolutionary model options required to compute a distance matrix. Type `I` to change settings.

7.1.2. Evolutionary model

```
E                                                 DNA evolutionary model
```

FastME implements a wide range of substitution models for DNA : p-distance, RY symmetric, RY, JC69, K2P, F81, F84, TN93, LogDet. F84 is the default option and is recommended in most cases. Select the model by typing `E`.

E	Protein evolutionary model

FastME implements a wide range of substitution models for proteins : p-distance, F81-like, LG, WAG, JTT, Dayhoff, DCMut, CpREV, MtREV, RtREV, HIVb, HIVw and FLU. LG is the default option and is recommended in most cases. Select the model by typing `E`.

7.1.3. Gamma distribution

G	Gamma distributed rates across sites

Rates of evolution often vary from site to site. This heterogeneity is modelled using a gamma distribution. Type `G` to switch this option on or off. If switched on, this option will add the gamma shape parameter option to the menu.

7.1.4. Gamma shape parameter

A	Gamma rate variation parameter (alpha)

The shape of the gamma distribution determines the range of rate variation across sites. Small values, typically in the [0.1, 1.0] range, correspond to large variability. Larger values correspond to moderate to low rate heterogeneity. With distance methods, it is often preferable to use relatively large (biased upward) values of this parameter, and 1.0 is default. Type `A` to set this option.

7.1.5. Remove sites whith gaps

R	Remove sites whith gaps

By default, FastME does pairwise deletion of gaps when computing the distance matrix. Every site containing gap can be removed by switching on this option. It must be used with caution, especially if the input MSA contains many gaps, as then very few sites are conserved and used to estimate pairwise distances. In such situation trimming first the input alignment with Gblocks or BMGE is preferable. Type `R` to switch this option on.

7.1.6. Output calculated distance matrix

O	Output calculated distance matrix

By default, the distance matrix computed from MSA is not displayed. It can be written into an output file by switching on this option. Type `O` to switch this option on.

7.1.7. Number of datasets

D	Number of datasets

If the input file contains several data sets, FastME can analyze each of them in a single run of the program. Type `D` to change settings.

7.1.8. Initial tree

M	Initial tree: build method

Algorithm used to compute a tree from a distance matrix. It can be iterative taxon addition (optimizing the BalME criterion for 'TaxAdd_BalME' or the OLSME criterion for 'TaxAdd_OLSME'), Neighbor Joining (NJ), the unweighted version of NJ (UNJ) or an improved version of NJ based on a simple model of sequence data (BioNJ). BioNJ is recommended with sequence alignments, and UNJ with non-sequence data (e.g. expression data, to obtain clusters rather than a phylogeny). The user can also input his own tree. This tree should be in Newick format. Type `M` to select among these initial tree methods.

7.1.9. NNI postprocessing

N	NNI postprocessing

By default, FastME does not improve the initial tree topology. Select this option to use nearest-neighbor interchange (NNI) to explore the topologies space. It optimizes the balanced version of minimum evolution (BalME). Type `N` to set this option.

7.1.10. SPR postprocessing

S	SPR postprocessing

FastME can also perform subtree pruning and regrafting (SPR) with BalME. It generally finds better tree topologies compared to NNI but tends to be slower. Type `S` to switch this option on.

7.1.11. Bootstrap

B	Bootstrap: number of replicates

The support of the data for each internal branch of the phylogeny can be estimated using non-parametric bootstrap. By default, this option is switched off. Typing `B` switches on the bootstrap analysis. The user is then prompted for a number of bootstrap replicates. The largest this number the more precise the bootstrap supports are. However, for each bootstrap replicate a phylogeny is estimated. Hence, the time needed to analyze *N* bootstrap replicates corresponds to *N*-times the time spent on the analysis of the original data set. *N* = 100 is generally considered as a minimum number of replicates; as FastME is fast, we recommend using *N* = 1,000, except for the very large data sets.

7.2. Command Line Interface

The alternative to the PHYLIP-like interface is the command line interface (CLI). It provides access to more parameters and options. However, you must be aware that some combinations are highly questionnable (e.g. building an initial first tree with OLSME and then performing BalME topological moves). Users that do not need to modify the default parameters can launch the program with the command :

```
fastme -i [aln_file_name]
```

The list of all command line arguments and how to use them is given in the ‘Help’ section which is displayed after entering the command :

```
fastme -h
```

The options are also described in what follows.

- **-i** input_data_file**, --input_data=**input_data_file

  The input_data_file contains MSA or a distance matrix(ces).

- **-u** input_user_tree_file**, --user_tree=**input_user_tree_file

  FastME may use an existing tree topology available in the input_user_tree_file which corresponds to the input dataset. Multiple input trees in input_user_tree_file may be used providing there are as much datasets in input_data_file (see '-i' option) as input trees. This tree should be in Newick format.

- **-o** output_tree_file**, --output_tree=**output_tree_file

  FastME will write the infered tree into the output_tree_file.

- **-O** output_matrix_file**, --output_matrix=**output_matrix_file

  Use this option if you want FastME to write the distance matrix computed from the input MSA in the output_matrix_file.

- **-I** output_info_file**, --output_info=**output_info_file

  Use this option if you want FastME to write information about its execution in the output_info_file.

- **-B** output_boot_file**, --output_boot=**output_boot_file

  Use this option if you want FastME to write bootstrap pseudo-trees in the output_boot_file.

- **-a, --append**

  Use this option to append results to existing output files (if any). By default output files will be overwritten.

- **-m** method**, --method=**method

  Algorithm used to compute a tree from a distance matrix. It can be iterative taxon addition (optimizing the BalME criterion for 'TaxAdd_BalME' or the OLSME criterion for 'TaxAdd_OLSME'), Neighbor Joining (NJ), the unweighted version of NJ (UNJ) or an improved version of NJ based on a simple model of sequence data (BioNJ). BioNJ is recommended with sequence alignments, and UNJ with non-sequence data (e.g. expression data, to obtain clusters rather than a phylogeny).

- **-d**[model]**, --dna=**[model]

  FastME implements a wide range of substitution models for DNA : p-distance, RY symmetric, RY, JC69, K2P, F81, F84, TN93, LogDet. F84 is the default option and is recommended in most cases.

- **-p**[model]**, --protein=**[model]

  FastME implements a wide range of substitution models for proteins : p-distance, F81-like, LG, WAG, JTT, Dayhoff, DCMut, CpREV, MtREV, RtREV, HIVb, HIVw and FLU. LG is the default option and is recommended in most cases.

- **-r, --remove_gap**

  By default, FastME does pairwise deletion of gaps when computing the distance matrix. Every site containing gap can be removed by switching on this option. It must be used with caution, especially if the input MSA contains many gaps, as then very few sites are conserved and used to estimate pairwise distances. In such

situation trimming first the input alignment with Gblocks or BMGE is preferable.

- **-e, --equilibrium**

  The equilibrium frequencies for DNA are always estimated using the nucleotide frequencies in the MSA.

  For amino-acid sequences, the equilibrium frequencies are estimated using the frequencies defined by the substitution model. Use this option if you wish to estimate the amino-acid equilibrium distribution using their frequencies in the MSA.

- **-g**[alpha]**, --gamma=**[alpha]

  Rates of evolution often vary from site to site. This heterogeneity is modelled using a gamma distribution. The shape of the gamma distribution ([alpha]) determines the range of rate variation across sites. Small values, typically in the [0.1, 1.0] range, correspond to large variability. Larger values correspond to moderate to low rate heterogeneity. With distance methods, it is often preferable to use relatively large (biased upward) values of [alpha], and 1.0 is default.

- **-n**[NNI]**, --nni=**[NNI]

  By default, FastME does not improve the initial tree topology. Select this option to use nearest-neighbor interchange (NNI) to explore the topologies space. The user can choose to optimize the balanced or ordinary least-square versions of minimum evolution (BalME/OLSME).

- **-s, --spr**

  FastME can also perform subtree pruning and regrafting (SPR) with BalME. It generally finds better tree topologies compared to NNI but tends to be slower.

- **-w** branch**, --branch_length=**branch

  The Minimum Evolution algorithms (balanced and OLS) implemented in FastME compute the tree topology and the branch lengths separately. Thus, it is required to define how FastME will compute the branch lengths. By default, FastME will compute the tree topology and the branch lengths within the same framework (i.e. balanced ME for the topology with balanced branch lengths or OLSME with OLS branch lengths). However, even if we recommend not to do so, it is possible for FastME to compute a balanced minimum evolution tree topology and to assign OLS branch lengths to that tree (and conversely). If the tree is computed by a NJ-like algorithm (NJ, UNJ or BioNJ), FastME can keep the inferred branch lengths ('none' value of the option) or assign balanced or OLS branch lengths.

  Note that this option is only available when not doing any (NNI or SPR) tree swapping improvement. Moreover, this option can be used to assign branch lengths to any user-defined input tree.

  The user may choose the branch value from: BalLS (default), OLS or none.

- **-D** datasets**, --datasets=**datasets

  If the input file contains several data sets, FastME can analyze each of them in a single run of the program. Default value is 1.

- **-b** replicates**, --bootstrap=**replicates

  The support of the data for each internal branch of the phylogeny can be estimated using non-parametric bootstrap. The largest the replicates number the more precise the bootstrap supports are. However, for each bootstrap replicate a

phylogeny is estimated. Hence, the time needed to analyze *N* bootstrap replicates corresponds to *N*-times the time spent on the analysis of the original data set. *N* = 100 is generally considered as a minimum number of replicates; as FastME is fast, we recommend using *N* = 1,000, except for the very large data sets.

- **-z** seed**, --seed=**seed

  Use this option to initialize randomization with seed value. Only helpful when bootstrapping.

- **-c**

  Use this option if you want FastME to only compute distance matrix. Only helpful when the input data file contains MSA.

- **-f** number_of_digits

  Use this option to set the number of digits after the dot to use on output. Default precision is 12.

- **-T** number_of_threads**, --nb_threads=**number_of_threads

  Use this option to set the number_of_threads to use. This option is only available if FastME was compiled with the parallel flag. Default number_of_threads is the number of available CPU cores.

- **-v** value**, --verbose=**value

  Sets the verbose level to value [0-3]. Default value is 0.

- **-V, --version**

  Prints the FastME version.

- **-h, --help**

  Display this usage.

7.3. Parallelization

The computation of distances is a highly parallelizable task. Indeed, the distances between each sequence can be computed independently. Modern computers often have two or more CPU cores and each CPU core can be used to compute distances separately. Using this parallel strategy, the computation time is in theory divided by the number of CPU cores. In practice, the parallel implementation requires some additional instructions in particular for memory allocation. Thus, the speedup is not strictly proportional to the number of CPU cores but it permits substantial time savings.

FastME sources must be compiled with specific options to turn on the parallel option (see Section 6.2). Once the binary file has been generated, FastME will automatically distribute the distances calculations on available CPU cores.

Bootstraping is also a highly parallelizable task. Bootstrap replicates are independent one from the other and can be analyzed separately. Using the parallel binary file, running a bootstrap analysis with, say 100 replicates on 2 CPU cores, can be done by typing the following command :

```
fastme -i input.aln -D K2P -b 100 -T 2
```

8. Inputs / Outputs

FastME reads data from standard text files, without the need for any particular file name extension.

```
9
Aurora        0.0   0.1   0.13  0.12  0.57  0.22  0.86  0.89  0.97
Boylii        0.1   0.0   0.7   0.7   0.5   0.9   0.65  0.67  0.72
Cascadae      0.13  0.7   0.0   0.7   0.4   0.11  0.54  0.66  0.79
Muscosa       0.12  0.7   0.7   0.0   0.45  0.15  0.48  0.49  0.67
Temporaria    0.57  0.5   0.4   0.4   0.0   0.48  0.85  0.83  0.107
Pretiosa      0.22  0.9   0.11  0.15  0.48  0.0   0.54  0.55  0.6
Catesbaiana   0.86  0.65  0.54  0.48  0.85  0.54  0.0   0.54  0.59
Pipiens       0.89  0.67  0.66  0.49  0.83  0.55  0.54  0.0   0.48
Tarahumarae   0.97  0.72  0.79  0.67  0.107 0.6   0.59  0.48  0.0
```

Figure 1. **PHYLIP distance matrix format.**

8.1. Distance matrix format

Input distance matrix file must be in PHYLIP format (Figure 1). That is, it must contain on its first line the number of taxa and each taxon starting on a new line with the taxon name, followed by the distances to all taxa in order. The distance matrix is square with zero distances on the diagonal. One slight difference with PHYLIP format lies in taxon name length. While PHYLIP format limits this length to ten characters, FastME can read up to 64 character long taxon names, followed by a blank character. Blanks and the symbols “(),:” are not allowed within sequence names because the Newick tree format makes special use of these symbols.

A PHYLIP input distance matrix file may also contain more than a single data set. Each of these data sets must be in PHYLIP format and two successive distance matrices must be separated by an empty line. Processing multiple data sets requires to toggle the '`D`' option in the PHYLIP-like menu or use the '`-d`' command line option and enter the number of data sets to analyze. The multiple data set option can be used to process re-sampled data that were generated using a non-parametric procedure such as cross-validation or jackknife. It may also correspond to distance matrices computed from different genes in phylogenomic studies.

PHYLIP sequential

```
5 50
tax1 ATTGCCCTAG TAACGGCGAG TGAAGCGGCA ACAGCTCAAA T-TTGAAATC
tax2 ATTGCCTCAG TAACGGCGAG TGAAGCGGCA ACAGCTCAAA T-TTGAAATC
tax3 ?????????? TAACGGCGAG TGAAGCGGCA ACAGCTCAAA T-TTGAAATC
tax4 ATTGCCCTAG TAACGGCGAG TGAAGCGGCA ACAGCTCAAA T-TTGAAATC
tax5 ATTGCCCTAG TAACGGCGAG TGAAGCGGCA ATAGCTCAAA T-TTGAAAGC
```

PHYLIP interleaved

```
5 60
tax1   ATTGCCCTAG TAACGGCGAG TGAAGCGGCA ACAGCTCAAA T-TTGAAATC
tax2   ATTGCCTCAG TAACGGCGAG TGAAGCGGCA ACAGCTCAAA T-TTGAAATC
tax3   ?????????? TAACGGCGAG TGAAGCGGCA ACAGCTCAAA T-TTGAAATC
tax4   ATTGCCCTAG TAACGGCGAG TGAAGCGGCA ACAGCTCAAA T-TTGAAATC
tax5   ATTGCCCTAG TAACGGCGAG TGAAGCGGCA ATAGCTCAAA T-TTGAAAGC

TGG-C---CC
TGG-C-TCAC
TG----TGAT
TGG-C-T-CC
TGG-C-----
```

Figure 2. **PHYLIP sequential and interleaved formats.**

8.2. Sequences formats

MSA of DNA or protein sequences must be in PHYLIP sequential or interleaved format (Figure 2). For PHYLIP formated MSA, the first line of the input file contains the number of taxa and the number of characters, in free format, separated by blank character(s). One slight difference with PHYLIP format deals with taxon name lengths. While PHYLIP format limits this length to ten characters, FastME can read up to 64 character long taxon names. Blanks and the symbols “(),:” are not allowed within taxon names because the Newick tree format makes special use of these symbols. Another slight difference with PHYLIP format is that actual sequences must be separated from their names by at least one blank character.

A PHYLIP input sequence file may also display more than a single data set. Each of these data sets must be in PHYLIP format and two successive MSA must be separated by an empty line. Processing multiple data sets requires to toggle the 'D' option in the PHYLIP-like menu or use the ‘-d’ command line option and enter the number of data sets to analyze. The multiple data set option can be used to process re-sampled data that were generated using a non-parametric procedure such as cross-validation or jackknife. This option is also useful in multiple gene studies, even if fitting the same substitution model to all data sets may not be suitable.

8.2.1. Gaps and ambiguous characters

Gaps correspond to the '-' symbol and gappy sites are removed (globally or pairwise, see above). Table 1 and Table 2 give the list of valid characters/symbols and the corresponding nucleotides or amino acids. Any character which is not displayed in theses Tables is treated as an unknown character.

Character	Nucleotide
A	Adenosine
C	Cytidine
G	Guanosine
T	Thymidine
N or *X* or *?*	unknown

Table 1. **List of valid characters in DNA sequences and the corresponding nucleotides.**

Character	Amino-Acid	Character	Amino-Acid
A	Alanine	*K*	Lysine
R	Arginine	*M*	Methionine
N or *B*	Asparagine	*F*	Phenylalanine
D	Aspartic acid	*P*	Proline
C	Cysteine	*S*	Serine
Q or *Z*	Glutamine	*T*	Threonine
E	Glutamiq acid	*W*	Tryptophan
G	Glycine	*Y*	Tyrosine
H	Histidine	*V*	Valine
I	Isoleucine	*X* or * or *?*	unknown
L	Leucine		

Table 2. **List of valid characters in protein sequences and the corresponding amino acids.**

8.3. Tree format

Input and output trees are recorded in the Newick format (Figure 3). FastME can read one or several phylogenetic trees from an input file. This option is accessible through the 'u' option in the PHYLIP-like menu or the '`-u`' command line option. Input trees are generally used as initial estimates to be subsequently adjusted by the tree searching algorithm. Trees can be either rooted or unrooted. Taxon names must, of course, match the corresponding sequence names.

```
((seq1:0.3,seq2:0.1):0.4,(seq3:0.1,(seq4:2,seq5:0.5):2):0.1);
((seq3,seq2),seq1,(seq4,seq5));
```

Figure 3. **Input trees.** The first tree (top) is rooted and has branch lengths. The second tree (bottom) is unrooted and does not have branch lengths.

8.4. Output files

Output file name	Content
*_fastme_tree.nwk	Inferred tree
*_fastme_stat.txt	Information about FastME execution
*_fastme_mat.txt	Computed distance matrix
*_fastme_boot.txt	Pseudo-trees from bootstrap replicates

Table 3. **Standard output files**

Table 3 presents the list of files resulting from a FastME analysis. Basically, each output file name can be divided into three parts. The first part is the MSA file name, the second part corresponds to the 'fastme' extension and the third part is related to the file content. When launched with the default options, FastME only generates two files: the tree file and the information file. The inferred distance tree is in standard Newick format (see Figure 3). The information file displays the FastME computation steps. Two additional output files can be created : the distance matrix file and the pseudo-trees file. The distance matrix file can be used as input for any tree distance method (including FastME). The pseudo-trees file can be used to build a consensus tree for instance.