

# morePhyML — User Guide

[Version 1.14] August 2011

by Alexis Criscuolo

`ftp://ftp.pasteur.fr/pub/GenSoft/projects/morePhyML/`  
`http://mobyale.pasteur.fr/cgi-bin/portal.py`

Please cite this paper if you use this software in your publications:

Criscuolo A (2011) morePhyML: improving the phylogenetic tree space exploration with PhyML 3. *Molecular Phylogenetics and Evolution* (in press).

**PhyML** is a widely used Maximum Likelihood (ML) phylogenetic tree inference software based on a standard hill-climbing method (Guindon and Gascuel 2003). **PhyML** first builds an initial tree that is secondly used as a starting tree in a heuristic local search in order to optimize the ML criterion. Since its version 3, **PhyML** allows using several different starting trees (*i.e.* random tree, MP tree, BioNJ tree, or user-defined tree), and two different tree swapping techniques (*i.e.* simultaneous NNIs or SPR moves; *e.g.* Swofford *et al.* 1996) to explore the tree space (see Guindon *et al.* 2010 for more details). As the accuracy of a phylogenetic tree inference based on a heuristic local search is highly dependant on both the starting tree (*e.g.* Criscuolo *et al.* 2006) and the respective size of the NNI- or SPR-based neighborhoods (*e.g.* Morrison 2007), **morePhyML** has been implemented to improve ML tree space exploration with **PhyML**.

Given a multiple sequence alignment (*i.e.* nucleotides or amino acids), the script **morePhyML** first uses **PhyML** to infer an initial tree. The user can choose among different options to perform this first ML tree search (*i.e.* starting tree, and tree swapping techniques). After this first step, **morePhyML** performs a deeper exploration of tree space. The first inferred tree is set as user tree in **PhyML** to perform a ML tree search by using NNIs on a bootstrap replicate of the multiple sequence alignment. This so-obtained 'noisy' tree is then used as starting tree in another NNI-based ML tree search. This approach, closely related to the so-called ratchet technique (Nixon 1999; see also Morrison 2007), allows the ML criterion (and the corresponding phylogenetic tree) to be improved in many cases. The script **morePhyML** repeats this ratchet procedure (*i.e.* 'noisy' tree used as a starting tree for a new ML tree search) as long as it allows trees with better log-likelihood value to be reached. Finally, when the ratchet loop stops, additional SPR-based ML tree searches are performed in order to escape from another possible local optimum.

## Installation and execution

Written in BASH shell, **morePhyML** uses only some of the standard UNIX commands (e.g. echo, cat, pwd, chmod, mv, cp, rm, grep, tr, sed). The script **morePhyML** then runs on UNIX operating systems (e.g. Linux, Mac OS X) without prior compilation. It uses the **PhyML** binary (version 3 or higher) to improve its ML tree space exploration. A **PhyML** linux 32 binary is provided in the **morePhyML** archive, but it is strongly recommended to use the latest version of **PhyML** available at the following URLs:

<http://www.atgc-montpellier.fr/phyml/binaries.php> ,  
<http://code.google.com/p/phyml/> .

The script **morePhyML** is a simple text file named **morePhyML.sh**. **Prior to any launch, first edit this file and indicate the path to the PhyML binary (version 3 minimum) installed on your computer;** this is indicated in the file **morePhyML.sh** (approximately line 100) by the following flag:

```
PHYML="/home/username/morePhyML/phyml_3.0";  
  
#####  
#####  
## <=== WRITE HERE THE PATH TO THE PHYML      ##  
##          BINARY (VERSION 3.0 MINIMUM)      ##  
#####  
#####
```

To get this (absolute) path, you just have to open a terminal, move to the directory containing the **PhyML** binary (here named **phyml\_3.0**), and use the following command:

```
pwd phyml_3.0
```

Secondly, move to the directory containing the script **morePhyML.sh**, and give the execute permission by using the following command:

```
chmod +x morePhyML.sh .
```

Given an input file **infile**, you can then launch the script **morePhyML** with the following command line:

```
./morePhyML.sh -i infile [options] .
```

During its execution, the script **morePhyML** creates different temporary files. These files are used to launch **PhyML** (**infile\_launcher.sh**), and to store intermediate ML trees and parameters (**infile\_starting\_tree.txt**, **infile\_phyml\_{tree,stats,boot\_trees,boot\_stats,rand\_trees}.txt**). All these files are deleted when **morePhyML** ends without error.

There also exist alternative versions of **PhyML** (e.g. Le *et al.* 2008a, 2008b) that are available from the **PhyML** website (see above). These alternative versions implement several new evolutionary models (e.g. CAT, EHO, EX2, EX3, UL2, UL3; for more details, see Le *et al.* 2008a, 2008b), but require some additional installation steps to be used by the script **morePhyML**. First, download the specified alternative version of **PhyML** from its website (see above). Secondly, edit the script **morePhyML** to specify the path to this new **PhyML** binary (see above). Third, complete the list of evolutionary models in the script **morePhyML** with the ones you wish to use.

For example, if you wish to use the amino acid evolutionary model CAT with 20, 30, 40 and 50 profiles (Lartillot and Philippe 2004; Le *et al.* 2008a), download the corresponding **PhyML** version from its website (i.e. <http://www.atgc-montpellier.fr/models/index.php?model=mixture>), modify the path to this new **PhyML** binary in the script **morePhyML**, and add the new models C20, C30, C40 and C50 (i.e. the model CAT with 20, 30, 40 and 50 profiles, respectively) in the list of evolutionary models accepted by the script **morePhyML** (approximately line 300):

```
#####  
##### ACCEPTED EVOLUTIONARY MODELS #####  
#####  
MODELS=(JC69 F81 K80 HKY85 F84 TN93e TN93 TPM1e K81 TPM1u TPM2e TPM2u TPM3e TPM3u TIM1e TIM1u TIM2e TIM2u TIM3e  
TIM3u TVMe TVMu SYM GTR LG WAG JTT MtREV Dayhoff DCMut RtREV CpREV VT Blosum62 MtMam MtArt HIVw HIVb  
C20 C30 C40 C50);
```

## Quick start

By default, **morePhyML** infers ML trees from alignments of nucleotide sequences by using the substitution model GTR. Given a file `infile.nt.phy` containing such an alignment in PHYLIP interleaved format, this is performed by using the following command:

```
./morePhyML.sh -i infile.nt.phy .
```

This will produce two output files, named `infile.nt.phy_morephyml_tree.txt` and `infile.nt.phy_morephyml_stats.txt`, that contain the inferred ML tree and model parameters, respectively. In complement, if the option `-x` is used:

```
./morePhyML.sh -i infile.nt.phy -x
```

then **morePhyML** will write the two files `infile.nt.phy_phyml_tree.txt` and `infile.nt.phy_phyml_stats.txt` that contain the ML tree and model parameters as inferred by **PhyML** only, respectively.

If the input file contains an alignment of amino acid sequences, then option `-d` must be set to `aa`:

```
./morePhyML.sh -i infile.aa.phy -d aa .
```

In this case, the default amino acid substitution model LG (Le and Gascuel 2008) will be used to estimate the ML tree and parameters.

If the input file contains a multiple sequence alignment in PHYLIP sequential format, then the option `-q` must be used:

```
./morePhyML.sh -i infile.phy -q .
```

Different options that allow choosing among different starting trees and substitution models are described in the next section. These different options can be displayed by using the following command:

```
./morePhyML.sh -? .
```

However, you can use the option `-z` to launch **morePhyML** with efficient ML tree search and standard model for amino acid data (*i.e.* evolutionary model  $LG+\Gamma_4+I$ ; Le and Gascuel 2008):

```
./morePhyML.sh -i infile.aa.phy -z aa ,
```

or nucleotide data (*i.e.* evolutionary model  $GTR+\Gamma_4+I+F$ ; Lanave *et al.* 1984; Tavaré 1986; Rodríguez *et al.* 1990; Yang 1994):

```
./morePhyML.sh -i infile.aa.phy -z nt .
```

By using the option `-z`, a ML tree was inferred by **morePhyML** from each of the two multiple sequence alignments available in the directory `example/`. For each data, the tree inferred by **PhyML** was saved (with the option `-x`) and is also available in the directory `example/`.

## morePhyML command line options

The script **morePhyML** can only be used with command line options. Most of these options (*i.e.* `-i`, `-d`, `-q`, `-t`, `-v`, `-c`, `-a`, `-v`, `-u`, `-p`) are identical to the **PhyML** ones. For more details about these options, please read the **PhyML** documentation available at the following URL:

<http://www.atgc-montpellier.fr/phyml/binaries.php>.

### Input file

By default, **morePhyML** reads an input file containing an alignment of nucleotide sequences in PHYLIP interleaved format. It should be stressed that **morePhyML** does not infer trees from a file containing several multiple sequence alignments. The name of this input file is set with the `-i` option. If its format is not the default one (*e.g.* not nucleotide sequences and not PHYLIP interleaved), you must use different options.

`-i <infile>`

This option allows the nucleotide or amino-acid sequence file (in PHYLIP format) to be indicated.

`-d <nt|aa>`

If the input file contains an alignment of nucleotide sequences, set the option `-d` to `nt` (default option). Otherwise (*i.e.* amino acid sequences), set this option to `aa`.

`-q`

By default, **morePhyML** reads the multiple sequence alignments in PHYLIP interleaved format. Use this option if the input file is in PHYLIP sequential format.

### Substitution models

A wide range of substitution models is available in **morePhyML**. These models are used by setting the following options.

`-m <model>`

The different available nucleotide-based models are:

- JC69 (Jukes and Cantor 1969),
- F81 (Felsenstein 1981),
- K80 (Kimura 1980),
- F84 (Kishino and Hasegawa 1989; Felsenstein and Churchill 1996),
- HKY85 (Hasegawa *et al.* 1985),
- TN93e (same as TrNe in Posada 2008),
- TN93 (Tamura and Nei 1993),
- K81 (Kimura 1981),
- TPM1e (same as K81),
- TPM1u (Posada 2008),
- TPM2e (same as TPM2 in Posada 2008),
- TPM2u (Posada 2008),
- TPM3e (same as TPM3 in Posada 2008),
- TPM3u (Posada 2008),
- TIM1e (Posada 2003; same as TIM1e in Posada 2008),
- TIM1u (Posada 2003; same as TIM1 in Posada 2008),
- TIM2e (Posada 2008),
- TIM2u (same as TIM2 in Posada 2008),
- TIM3e (Posada 2008),
- TIM3u (same as TIM3 in Posada 2008),
- TVMe (Posada 2003; same as TVMe in Posada 2008),
- TVMu (Posada 2003; same as TVM in Posada 2008),
- SYM (Zharkikh 1994),
- GTR (Lanave *et al.* 1984; Tavaré 1986; Rodríguez *et al.* 1990; Yang 1994).

For amino acid sequences, the different available substitution models are:

- Dayhoff (Dayhoff *et al.* 1978),
- Blosom62 (Henikoff and Henikoff 1992),
- JTT (Jones *et al.* 1992),
- MtREV (Adachi and Hasegawa 1996),
- MtMam (Cao *et al.* 1998),
- CpREV (Adachi *et al.* 2000),
- VT (Muller and Vingron 2000),
- WAG (Whelan and Goldman 2001),
- DCMut (Kosiol and Goldman 2005),
- RtREV (Dimmic *et al.* 2002),
- MtArt (Abascal *et al.* 2007),
- HIVw (Nickle *et al.* 2007),
- HIVb (Nickle *et al.* 2007),
- LG (Le and Gascuel 2008).

Default substitution models are GTR and LG for nucleotide and amino acid sequences, respectively.

`-f <m|e>`

The method for estimating the character state equilibrium frequencies is selected with this option. For nucleotide data (*i.e.* `-d nt`), this option is only available for models allowing unequal base frequencies (*i.e.* F81, F84, HKY85, TN93, TPM1u, TPM2u, TPM3u, TIM1u, TIM2u, TIM3u, TVMu, GTR). Nucleotide and amino acid equilibrium frequencies can be estimated by counting their occurrence in the multiple sequence alignment (*i.e.* the empirical method: option `-f e`). Otherwise, the alternative methods (*i.e.* corresponding to the option `-f m`) is the ML estimate of the nucleotide equilibrium frequencies, or the amino acid equilibrium frequencies defined by the substitution model specified with the option `-m`. When required, default option is `-f m`.

`-t <e|real>`

This option allows specifying the transition/transversion ratio (*i.e.* a positive real number). This option can be set with only nucleotide sequences (*i.e.* option `-d nt`) and the substitution models K80, F84, HKY85, and TN93. By default with these four models, the transition/transversion ratio is estimated (*i.e.* `-t e`).

`-c <integer>`

This option allows modelling the substitution rate heterogeneity across characters by using a discrete gamma distribution. The positive integer specified with this option is the number of categories of this discrete distribution. By default, **morePhyML** uses only one category (*i.e.* `-c 1`), but, in practice with real datasets, at least four categories are recommended to infer accurate phylogenetic trees (*i.e.* `-c 4`).

`-a <e|real>`

When the option `-c` is set with more than one category for the discrete gamma distribution, then the option `-a` allows the shape of this distribution to be modelled. By default, **morePhyML** performs a ML estimate of this shape (*i.e.* `-a e`). However, a positive real value can be specified.

`-v <e|real>`

By default, the proportion of invariable characters, *i.e.* the expected frequency of characters with no substitution, is fixed to zero by **morePhyML**. However, another value can be specified (*i.e.* varying from 0 to 1). It is also possible to perform a ML estimate (*i.e.* `-v e`). When the ML estimate of the proportion of invariable characters is very close to zero, it is strongly recommended to relaunch **morePhyML** without this option (or after setting `-v 0`).

## Starting tree(s)

As **PhyML**, the script **morePhyML** uses the BioNJ tree (Gascuel 1997) as starting tree in the heuristic local search. However, different options allow alternative starting trees to be used.

`-p`

This option allows using a quickly computed MP tree as starting tree. Even if **morePhyML** allows escaping from local optima, this option is recommended when the input file contains a supermatrix of characters with a large number of unknown character states (*e.g.* Criscuolo *et al.* 2006).

`-n <integer>`

This option allows performing the initial ML tree searching from a specified number of random starting

trees. For example, when set to 5 (which is often sufficient), **morePhyML** uses **PhyML** to perform ML tree searching from the BioNJ tree as well as 5 more random starting trees. Then, the best phylogenetic tree is used by **morePhyML** as a starting tree in the ratchet procedure. This option leads to longer running times.

**-u <treefile>**

This option allows inputting a file containing a user-defined starting tree in NEWICK format. This tree must be defined on the same taxon names as the multiple sequence alignment. Branch lengths and confidence values at branches are accepted.

**-s <NNI|SPR|BEST>**

The first step of **morePhyML** is to launch **PhyML** to infer a ML phylogenetic tree. This is performed from the starting tree(s) by using the two different tree swapping techniques available in **PhyML** : simultaneous NNIs (*i.e.* **-s NNI**), or SPR (*i.e.* **-s SPR**; default). A third way (*i.e.* **-s BEST**) allows performing a ML tree searching by using both methods; in this case the most likely phylogenetic tree among the two is selected.

## Output tree(s) and files

By default, given the input file `infile.phy`, **morePhyML** outputs two files, named `infile.phy_morephyml_tree.txt` and `infile.phy_morephyml_stats.txt`, that contain the ML tree and parameters, respectively. In complement, several options allow getting additional results.

**-b <0|-1|-2>**

It is not possible to directly compute bootstrap support at branches with **morePhyML** : bootstrap replicates of the initial multiple sequence alignment must be generated separately and next analyzed by **morePhyML**. By default, **morePhyML** computes SH-like branch supports (Anisimova and Gascuel 2006), but two other confidence values at branches based on approximate likelihood ratio tests can be computed: aLRT statistics (*i.e.* **-b -1**) and  $\chi^2$ -based parametric branch supports (*i.e.* **-b -2**; for more details, see Anisimova and Gascuel 2006). However, it is also possible to output the ML phylogenetic tree inferred by **morePhyML** with no confidence value at branches (*i.e.* **-b 0**).

**-x**

The first step performed by **morePhyML** is to use **PhyML** to infer a first tree, that is next used in further ML tree searches. Given the input file `infile.phy`, when the option **-x** is set, this first tree and the corresponding ML parameters are written in the two files `infile.phy_phyml_tree.txt` and `infile.phy_phyml_stats.txt`, respectively. This allows comparing the results produced by **PhyML** and **morePhyML**, respectively.

**-l**

Given the input file `infile.phy`, this option allows writing the likelihood for each character in a file named `infile.phy_morephyml_lk.txt`

**-e <string>**

By default, the filename extension of every files outputed by **morePhyML** is `txt`. However, another filename extension can be set with this option.

## All-in-one

The following option allows setting current and efficient evolutionary models for nucleotide and amino acid multiple sequence alignments.

**-z <aa|nt>**

This option allows launching **morePhyML** with a BioNJ starting tree to perform a first SPR-based ML tree search, followed by the ratchet procedure. When set to `aa`, **morePhyML** uses the evolutionary model  $\text{LG}+\Gamma_4+\text{I}$ ; when set to `nt`, it uses the model  $\text{GTR}+\Gamma_4+\text{I}+\text{F}$ .

## References

- Abascal F, Posada D, Zardoya R (2007) MtArt: a new model of amino acid replacement for Arthropoda. *Molecular Biology and Evolution*, 24(1):1-5.
- Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution*, 42(4):459-68.
- Adachi J, Martin W, Hasegawa M (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution*, 50:348-358.
- Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Systematic Biology*, 55(4):539-552.
- Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Paabo S, Hasegawa M (1998) Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *Journal of Molecular Evolution*, 47:307-322.
- Criscuolo A (2011) morePhyML: improving the phylogenetic tree space exploration with PhyML 3. *Molecular Phylogenetics and Evolution* (in press).
- Criscuolo A, Berry V, Douzery EJP, Gascuel O (2006) SDM: a fast distance-based approach for (super)tree building in phylogenomics. *Systematic Biology*, 55(5):740-755.
- Dayhoff MO, Schwartz RM, Orcutt BD (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed), *Atlas of Protein Sequence and Structure. Volume 5*. National Biomedical Research Foundation, Washington, pp. 345-352.
- Dimmic M, Rest J, Mindell D, Goldstein D (2002) rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *Journal of Molecular Evolution*, 55:65-73.
- Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368-376.
- Felsenstein J, Churchill GA (1996) A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13 93-104
- Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14:685-695.
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52:696-704.
- Guindon S, Dufayard JF, 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. *Systematic Biology*, 59:307-321.
- Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial-DNA. *Journal of Molecular Evolution*, 22:160-174
- Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Science USA*, 89:10915-10919.
- Jones D, Taylor W, Thornton J (1992) The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, 8:275-282.
- Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) *Mammalian protein metabolism*. Academic Press, New York, pp. 21-32.
- Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111-120.
- Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Science USA*, 78:454-458.

- Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, 29(2):170-179.
- Kosiol C, Goldman N (2004) Different versions of the Dayhoff rate matrix. *Molecular Biology and Evolution*, 22:193-199.
- Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86-93.
- Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21:1095-1109.
- Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25:1307-1320.
- Le SQ, Gascuel O, Lartillot N (2008a) Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, 24:2317-2323.
- Le SQ, Lartillot N, Gascuel O (2008b) Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 363:3965-3976.
- Morrison DA (2007) Increasing the efficiency of searches for the Maximum Likelihood tree in a phylogenetic analysis of up to 150 nucleotide sequences. *Systematic Biology*, 56:988-1010
- Muller T, Vingron M (2000) Modeling amino acid replacement. *Journal of Computational Biology*, 7:761-776.
- Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Kosakovsky Pond SL (2007) HIV-specific probabilistic models of protein evolution. *PloS One*, 2(6):e503.
- Nixon KC (1999) The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15:407-414.
- Posada D (2003) Using Modeltest and PAUP\* to select a model of nucleotide substitution. In: Baxevanis AD, Davison DB, Page RDM, Petsko GA, Stein LD, Stormo GD (eds) *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, pp. 6.5.1-6.5.14.
- Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, 25:1253-1256.
- Rodríguez F, Oliver JL, Marin A, Medina JR (1990) The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142:485-501.
- Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in human and chimpanzees. *Molecular Biology and Evolution*, 10:512-526
- Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM (ed) *Some Mathematical Questions in Biology: DNA Sequence Analysis, Vol. 17*. American Mathematical Society, pp. 57-86.
- Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18:691-699.
- Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution*, 39:315-329.