PhyML+M3L User Manual

Victor Hanson-Smith[1,2],
Bryan Kolaczkowski[2],
John St. John[1,2],
and Joseph W. Thornton[2,3]

[1]Department of Computer and Information Science,
[2]Center for Ecology and Evolutionary Biology,
and [3]Howard Hughes Medical Institute,
University of Oregon, Eugene, OR 97403 USA

Correspondence to:
Victor Hanson-Smith,
victorhs@cs.uoregon.edu

December 11, 2009

Copyright Disclosure

PhyML+M3L is an extension of the publicly-available source code for PhyML version 3.0 [Guindon and Gascuel (2003)]. The original authors of PhyML were not involved in the creation of PhyML+M3L. Please do not send them questions about the +M3L extensions. To learn more about PhyML, go here:

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

To learn more about PhyML+M3L, go here:

http://phylo.uoregon.edu/software/phyml+m3l

Contents

1 Hello Friends

We hope you find PhyML+M3L useful. However, be aware that we do not have a large team of software developers. We are providing this software as a resource to the research community, but without the promise of support. If you have questions, or find software bugs (!), please do not hesitate to contact us at the email address listed on the front page. We plan to release a new version of PhyML+M3L in the very near future, including MPI-based parallelism (and some other cool speedups).

2 What is PhyML+M3L?

PhyML+M3L is an extension of the publicly-available source code for PhyML version 3.0. PhyML is a “simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood” [Guindon and Gascuel (2003)]. We extended the PhyML source code to include several useful features: an evolutionary model incorporating mixed branch length heterotachy, an empirical Bayesian MCMC sampler to estimate clade support as posterior probabilities, OpenMP multiprocessor parallelization to greatly improve computational performance, and an experimental heuristic to optimize phylogenies using simulated thermal annealing. These features are described in more detail below.

2.1 A mixed branch length model of heterotachy

The mixed branch length model calculates the likelihood of phylogenies at each site in a given sequence alignment as a weighted sum over multiple independent branch length sets; weights and branch lengths can be inferred from the given sequence data [Kolaczkowski and Thornton (2008)]. Under most conditions, the mixed branch length model improves phylogenetic accuracy compared to other homotachous and heterotachous models. This model should not be confused with other heterotachous models, such as the gamma model [Yang (1994)] or the covarion model [Penny et al. (2001)]. Unlike those models, the mixed branch length model relaxes the assumption that the ratio of branch lengths remains constant across sites.

2.2 An empirical Bayes MCMC sampler to estimate posterior probabilities of clades

Posterior probability (PP) can be a useful metric to estimate the statistical support for the existence of a phylogenetic clade. However, simulation studies have shown that when PPs are estimated using a Bayesian MCMC strategy that integrates over branch length uncertainty, PPs can significantly diverge from their expected values had the branch lengths been known in advance. Alternatively, an empirical Bayesian strategy that fixes branch lengths at their maximum likelihood values is more accurate at estimating the posterior probability of clades [Kolaczkowski and Thornton (2007), Kolaczkowski and Thornton (2009)].

Phylogenetic practicioners traditionally use the software package Mr. Bayes to perform MCMC sampling and compute posterior probabilities

[Huelsenbeck and Ronquist (2001), Ronquist and Huelsenbeck (2003)]. Unfortunately, Mr. Bayes does not support a sampling scheme in which we can calculate the ML value of branch lengths while integrating over uncertainty about other parameters. Out of necessity, we implemented such an empirical Bayes strategy in PhyML+M3L. Combined with the previous features of PhyML, you can now use PhyML+M3L as a single tool to estimate bootstrap values, approximate likelihood ratio test (aLRT) values, and posterior probability values.

2.3 Multicore parallelization

The basic likelihood algorithm calculates the likelihood of a proposed phylogeny, given a sequence alignment, an evolutionary model, and a set of parameter values for the model [Felsenstein (1981)]. We assume the sites evolve independently, and therefore the the likelihood of a phylogeny is calculated as the product of likelihoods at each site in the alignment. These per-site likelihoods can be calculated in any order, as long as they are all combined together as a product. The likelihood algorithm is "embarrassingly parallel" over sites because we can directly parallelize the per-site likelihood calculations to independent processors. If your CPU contains eight independent cores, you can therefore delegate each core to calculate likelihoods for one-eighth of the sites.

In order to speedup the likelihood calculation, we use methods from the OpenMP library to dispatch per-site likelihood calculations to parallel CPU cores within a shared-memory architecture. If your computer's CPU contains multiple cores (for example, Intel dual-core or quad-core Mac products), the OpenMP parallelization will improve the runtime of PhyML+M3L. The efficacy of the parallelization depends on the number of available cores; a CPU with 8 cores will yield a greater speedup than a CPU of that same architecture with only 2 cores. If your CPU is single-core, or if you would rather not use OpenMP, you can run PhyML+M3L without the OpenMP features. Consult the section named 'Installation' for more information.

2.4 Optimization by simulated thermal annealing

Traditional hill-climbing optimization algorithms can struggle to escape local optima when searching over extremely rugged multi-parameterized likelihood landscapes. By default, PhyML uses a hill-climbing algorithm based on "Brent's Method" [Brent (1972)]. Although Brent's method seems to generally yield good results, the mixed branch length model can create rugged conditions in which hill-climbing seems to be ineffective.

As an alternative to hill-climbing methods, PhyML+M3L provides a method to optimize the topology, branch lengths, and model parameters using simulated thermal annealing (STA) [Kirkpatrick et al. (1983), Kirkpatrick (1984), Kolaczkowski and Thornton (2008)]. Although STA can yield extremely excellent results, STA is computationally demanding and can require hours, days, (or longer!) to infer a likelihood maxima. STA is provided here for experimental purposes.

3 Installation

You can run PhyML+M3L as a precompiled binary, or you can download the source code and build the application yourself.

3.1 Precompiled binaries

PhyML+M3L can be downloaded as a precompiled binary for Intel architectures running OSX 10.4 and higher:

```
http://phylo.uoregon.edu/software/phyml+m3l
```

The precompiled binary is available in two flavors: with OpenMP enabled and with OpenMP disabled. If you are using a multicore CPU (for example, an Intel dual-core Mac), then we suggest you download the version with OpenMP enabled. If you are using a single-core CPU, or if you'd rather not run OpenMP, then download the version with OpenMP disabled. Precompiled binaries require no installation. They should be executable upon download.

3.2 Source code

PhyML+M3L is written in C. you can download a packaged release as a ZIP file here:

```
http://phylo.uoregon.edu/software/phyml+m3l
```

Or, you can checkout (i.e. download) the latest version of source code from a Google Code repository:

```
http://code.google.com/p/m3l/
```

We tested the source code using the following software tools: **gcc** version 4.2.1 (Apple Inc. build 5574), with hardware target = i686-apple-darwin9. **aclocal** version 1.10, **GNU Make** version 3.81, and **GNU Autoconf** version 2.61.

To install from PhyML+M3L from source code on a Unix-based machine, follow these instructions:

1. If you downloaded a ZIP package, unzip the archive:

   ```
   \%> gzip -d phyml+m3l_12.10.2009.zip
   ```

   This will create a folder named phyml+m3l_12.10.2009. You can rename this folder whatever you desire, and you can move this folder to the installation location of your choice. If you downloaded the code from our Google Code repository, then you will need to create this folder yourself.

2. Navigate inside the folder from the previous step, and type the following commands:

   ```
   aclocal
   autoconf -f
   automake -f
   ./configure
   make
   ```

3. Et voila! You can start PhyML+M3L by executing 'phyml'.

Be aware the installation can be customized; see the file named `INSTALL` for more information. Finally, if you desire to use OpenMP parallelization, please read the next section of this manual.

3.3 Enabling OpenMP multiprocessor parallelization

If you are building PhyML+M3L from source code and you wish to enable OpenMP parallelization, follow these steps:

1. Open the file named `Makefile.am` and uncomment this line:

   ```
   #AM_CFLAGS=-O3 -g -funroll-loops -Wall
                -ftree-vectorize -ffast-math -fopenmp
   ```

   In other words, remove the '`#`' character at the beginning of the line.

2. Open the file name `utilities.h` and uncomment this line:

   ```
   //#define USE_OPENMP 1
   ```

   In other words, removed the '//' characters at the beginning of the line.

3. Finally, re-build the application, following the previous instructions in the section titled "Source Code". By the way, these instructions are also pasted into the file named `INSTALL`.

4 A note about model-fitting

When comparing phylogenies inferred under two different models of evolution, it is tempting to think the phylogeny with the higher likelihood is the better phylogeny. In some cases, higher likelihood values are incorrectly achieved by using over-parameterized evolutionary models. For example, suppose we use a simple model with no among-site rate variation to infer the phylogeny of a set of given sequences. Suppose we also infer the phylogeny using a complex model with four gamma-distributed evolutionary rates. In this example, the gamma model includes four parameters that are absent from the simple model. The likelihood of the tree inferred using the complex model will always be higher than the likelihood of the tree inferred using the simple model. If the increased likelihood from the complex model is not proportionate to the increased number of model parameters, then the complex model is said to *overfit* the data.

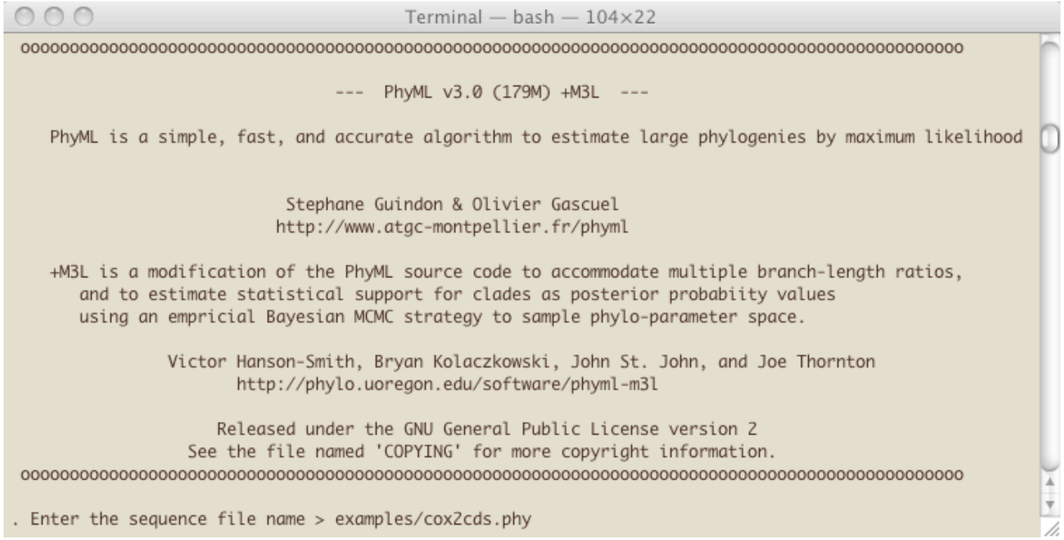
PhyML+M3L implements a mixed branch length model of heterotachy. This model uses more parameters than a simpler model that assumes no branch length heterotachy. The heterotachous model might not be appropriate for your data. We strongly encourage you to find the *best-fitting* evolutionary model by applying widely-used statistical tests such as the likelihood ratio test [Felsenstein (1981), Huelsenbeck and Crandall (1997), and Huelsenbeck and Rannala (1997)] or the Akaike Information Criterion [Akaike (1973)].

5 Examples

In this section, we demonstrate how the use the features of PhyML+M3L. These examples use an alignment file named *cox2cds.phy*, which contains 51 nucleotide sequences encoding for cyclooxygenase-II proteins. This file is included with PhyML+M3L download package.

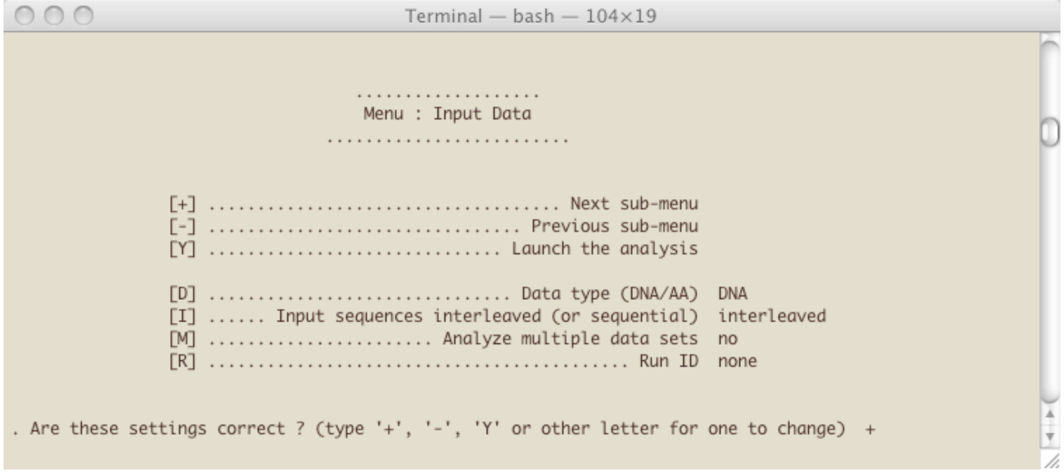
5.1 Using the mixed branch-length model

Begin by navigating to the folder that contains the PhyML+M3L program. This folder should contain a subfolder named *examples*, in which the *cox2cds.phy* alignment is located.



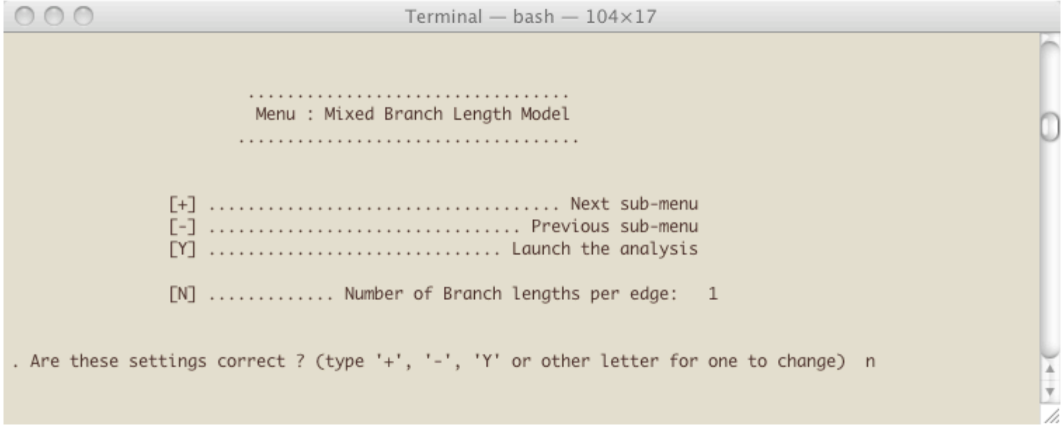
```
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo

                                ---  PhyML v3.0 (179M) +M3L  ---

   PhyML is a simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood


                          Stephane Guindon & Olivier Gascuel
                         http://www.atgc-montpellier.fr/phyml

   +M3L is a modification of the PhyML source code to accommodate multiple branch-length ratios,
      and to estimate statistical support for clades as posterior probabiity values
      using an empricial Bayesian MCMC strategy to sample phylo-parameter space.

                 Victor Hanson-Smith, Bryan Kolaczkowski, John St. John, and Joe Thornton
                        http://phylo.uoregon.edu/software/phyml-m3l

                      Released under the GNU General Public License version 2
                   See the file named 'COPYING' for more copyright information.
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo

. Enter the sequence file name > examples/cox2cds.phy
```

Figure 1: After PhyML+M3L loads, enter the filepath of the *cox2cds.phy* alignment.



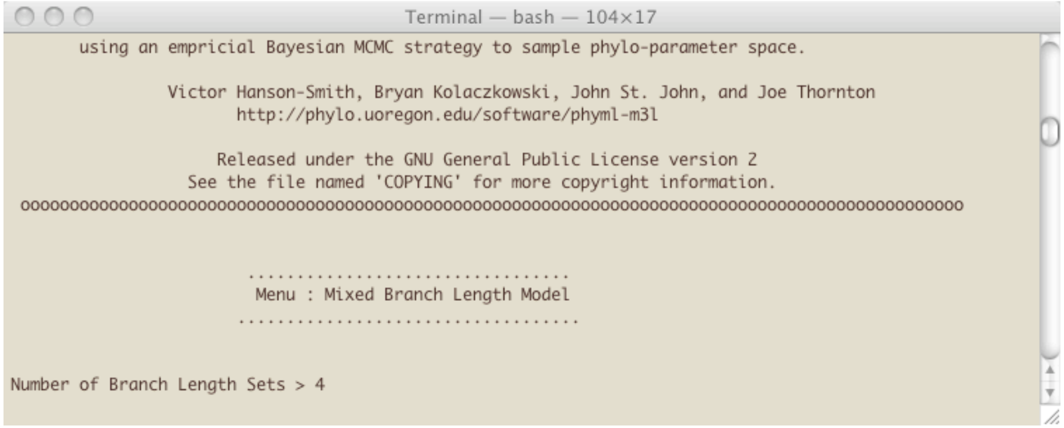
```
                                   ...................
                                   Menu : Input Data
                                .........................


                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [D] ............................... Data type (DNA/AA)  DNA
                [I] ...... Input sequences interleaved (or sequential)  interleaved
                [M] ....................... Analyze multiple data sets  no
                [R] ........................................... Run ID  none


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  +
```

Figure 2: The Cox-2 data is DNA (nucleotide) data, so no changes need to be made on this menu page. Enter '+' to advance to the next menu page.



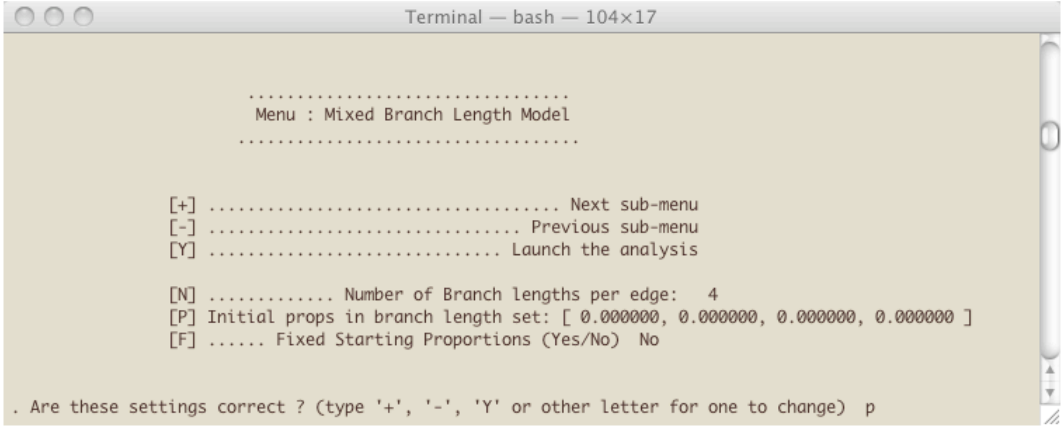
```
                         .................................
                          Menu : Mixed Branch Length Model
                        ...................................


                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [N] ............. Number of Branch lengths per edge:   1


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  n
```

Figure 3: The mixed branch length model is disabled by default. Enter ‘n’ to enable the model and specify multiple branch lengths per edge.



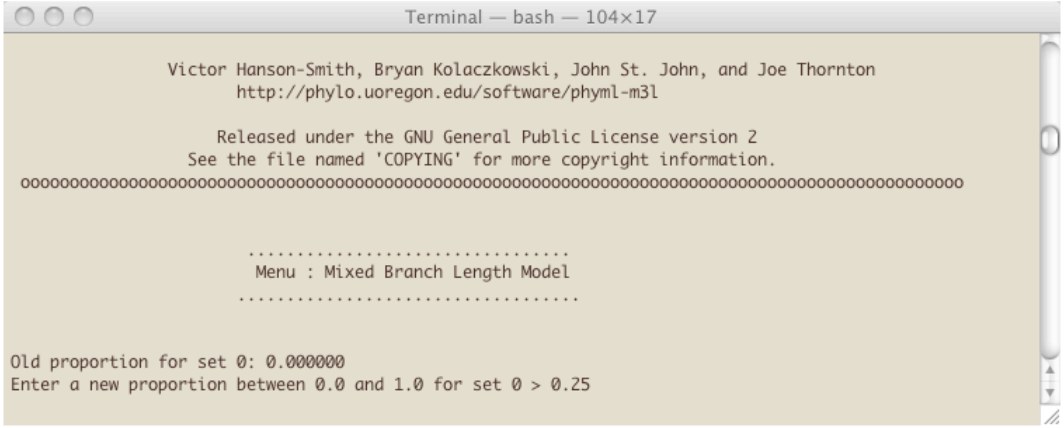
```
       using an empricial Bayesian MCMC strategy to sample phylo-parameter space.

                Victor Hanson-Smith, Bryan Kolaczkowski, John St. John, and Joe Thornton
                       http://phylo.uoregon.edu/software/phyml-m3l

                     Released under the GNU General Public License version 2
                  See the file named 'COPYING' for more copyright information.
 oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo


                        .................................
                         Menu : Mixed Branch Length Model
                       ...................................


Number of Branch Length Sets > 4
```

Figure 4: Enter ‘4’ for the number of branch length sets.



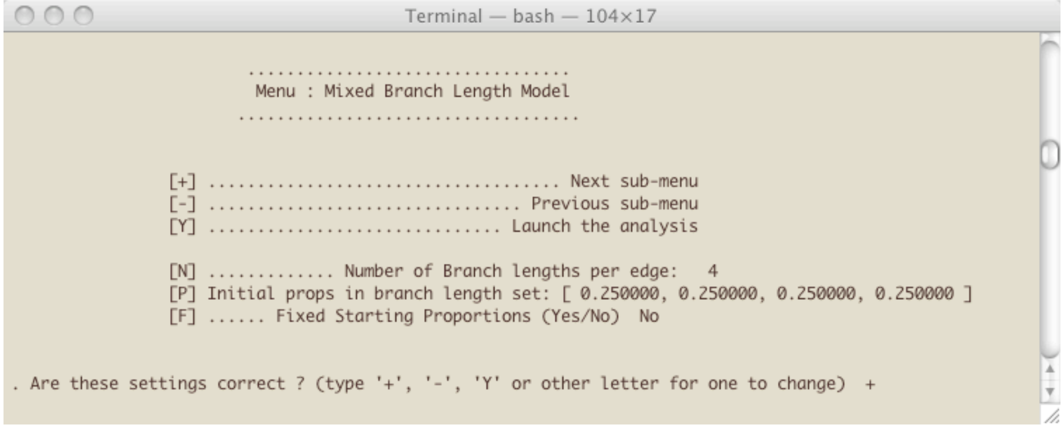
```
                      .................................
                       Menu : Mixed Branch Length Model
                     ...................................


              [+] .................................... Next sub-menu
              [-] ................................ Previous sub-menu
              [Y] .............................. Launch the analysis

              [N] ............ Number of Branch lengths per edge:   4
              [P] Initial props in branch length set: [ 0.000000, 0.000000, 0.000000, 0.000000 ]
              [F] ...... Fixed Starting Proportions (Yes/No)  No


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  p
```

Figure 5: After you specify the number of branch length sets, you will automatically return to the menu. Notice that some additional options now appear. Next we'll specify the proportion of sites in each branch length category. Enter 'p'.



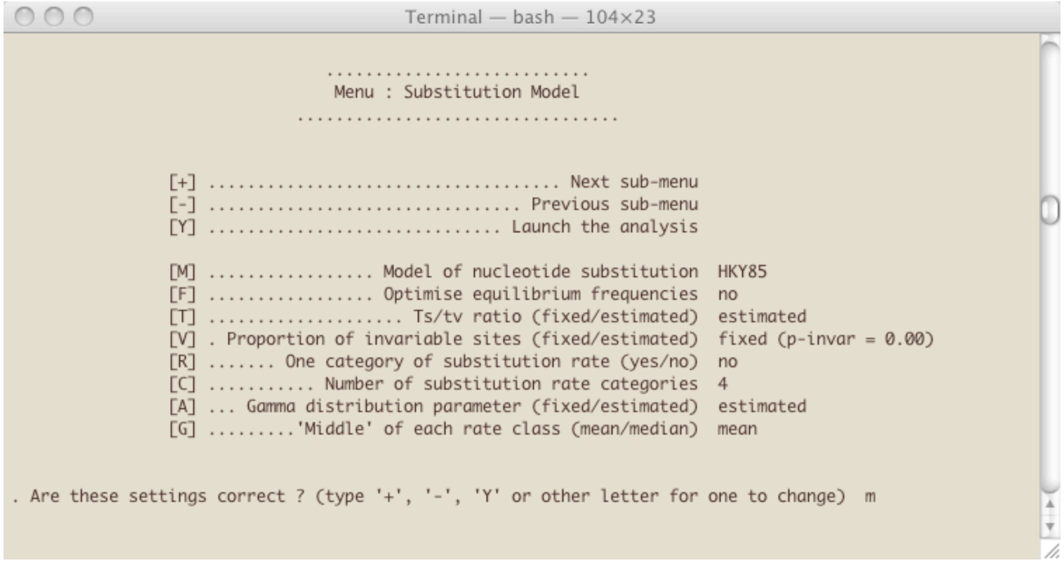
```
                 Victor Hanson-Smith, Bryan Kolaczkowski, John St. John, and Joe Thornton
                        http://phylo.uoregon.edu/software/phyml-m3l

                      Released under the GNU General Public License version 2
                   See the file named 'COPYING' for more copyright information.
 oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo



                         .................................
                          Menu : Mixed Branch Length Model
                        ...................................


Old proportion for set 0: 0.000000
Enter a new proportion between 0.0 and 1.0 for set 0 > 0.25
```

Figure 6: Enter ‘0.25’ for each branch length proportion.



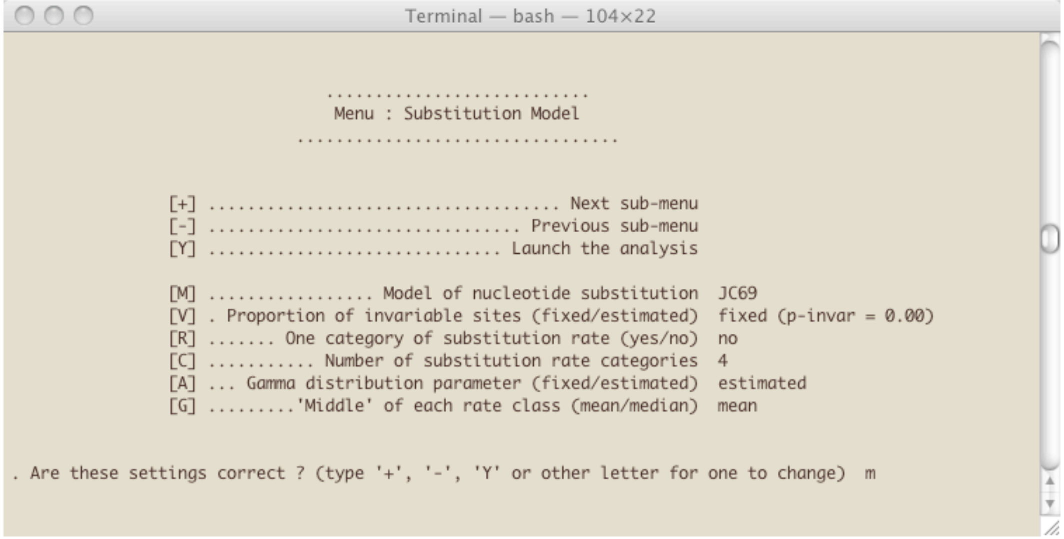
```
                         ..................................
                          Menu : Mixed Branch Length Model
                        ....................................


                [+] ..................................... Next sub-menu
                [-] ................................. Previous sub-menu
                [Y] ............................... Launch the analysis

                [N] ............. Number of Branch lengths per edge:   4
                [P] Initial props in branch length set: [ 0.250000, 0.250000, 0.250000, 0.250000 ]
                [F] ...... Fixed Starting Proportions (Yes/No)  No


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  +
```

Figure 7: After specifying branch length proportions, you will automatically return to the menu. Notice that your proportion specifications are now shown in the menu. Enter '+' to advance to the next menu.



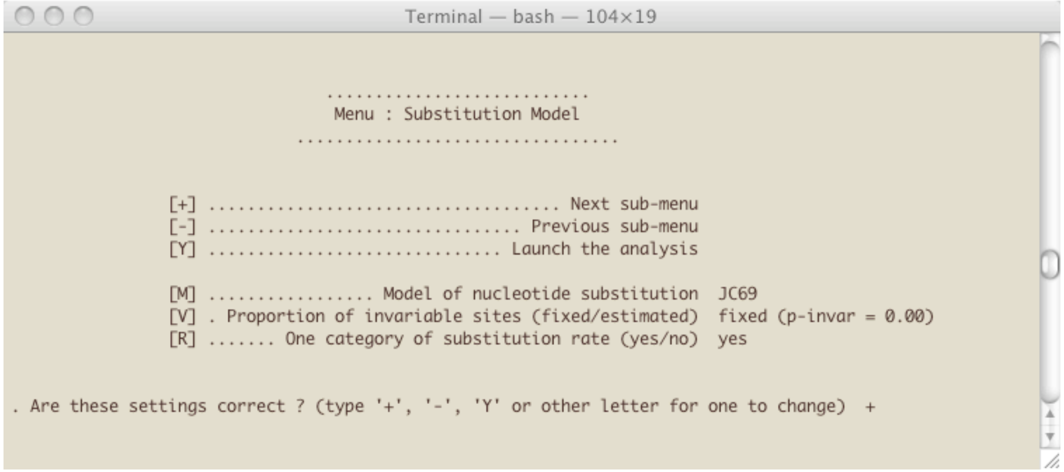
```
                               ..........................
                                Menu : Substitution Model
                            ................................


                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [M] ................. Model of nucleotide substitution  HKY85
                [F] ................. Optimise equilibrium frequencies  no
                [T] .................... Ts/tv ratio (fixed/estimated)  estimated
                [V] . Proportion of invariable sites (fixed/estimated)  fixed (p-invar = 0.00)
                [R] ....... One category of substitution rate (yes/no)  no
                [C] ........... Number of substitution rate categories  4
                [A] ... Gamma distribution parameter (fixed/estimated)  estimated
                [G] .........'Middle' of each rate class (mean/median)  mean


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  m
```

Figure 8: This menu allows you specify the substitution model. By default, PhyML uses the HKY85 model for nucleotide data. Instead, let's use the simpler JC69 model. Enter 'm' to cycle through the available model options. Stop cycling when you see the model named 'JC69'.



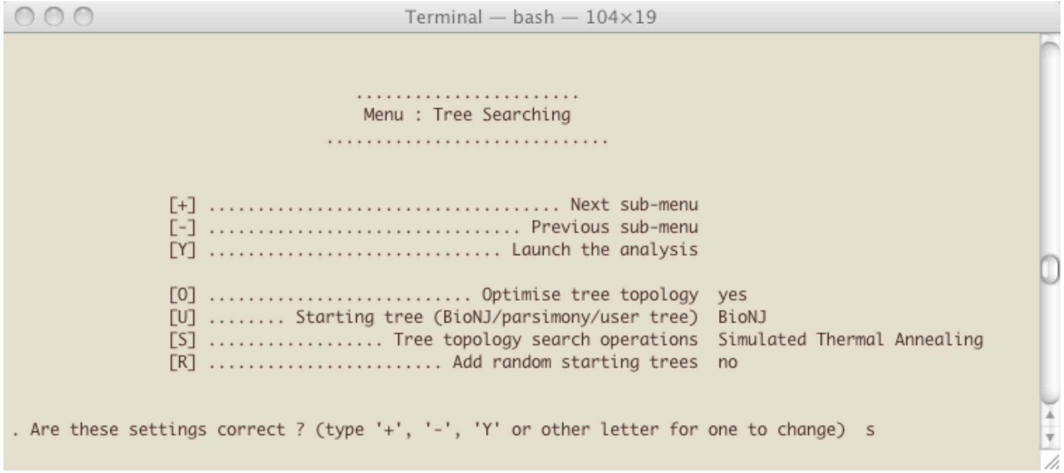
```
                              ...........................
                               Menu : Substitution Model
                           .................................


               [+] .................................... Next sub-menu
               [-] ................................ Previous sub-menu
               [Y] .............................. Launch the analysis

               [M] ................. Model of nucleotide substitution  JC69
               [V] . Proportion of invariable sites (fixed/estimated)  fixed (p-invar = 0.00)
               [R] ....... One category of substitution rate (yes/no)  no
               [C] ........... Number of substitution rate categories  4
               [A] ... Gamma distribution parameter (fixed/estimated)  estimated
               [G] .........'Middle' of each rate class (mean/median)  mean


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  m
```

Figure 9: By default, PhyML uses four gamma-distributed evolutionary rates. For this example, let’s use a simpler model with only one evolutionary rate. Enter ‘r’ to disable the gamma-distributed model.



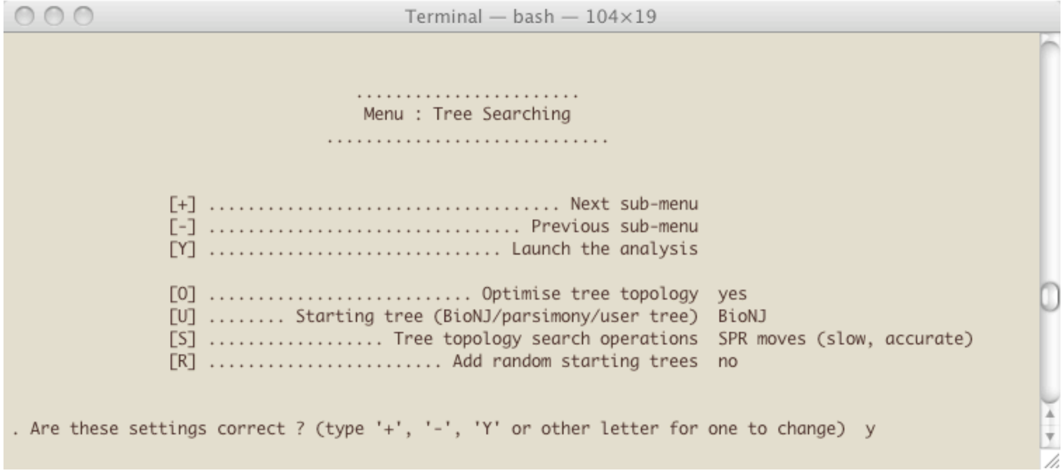
```
                                 ...........................
                                  Menu : Substitution Model
                              .................................


                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [M] ................. Model of nucleotide substitution  JC69
                [V] . Proportion of invariable sites (fixed/estimated)  fixed (p-invar = 0.00)
                [R] ....... One category of substitution rate (yes/no)  yes


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  +
```

Figure 10: After disabling the gamma-distributed model, the menu should look like this. Enter '+' to advance to the next menu.



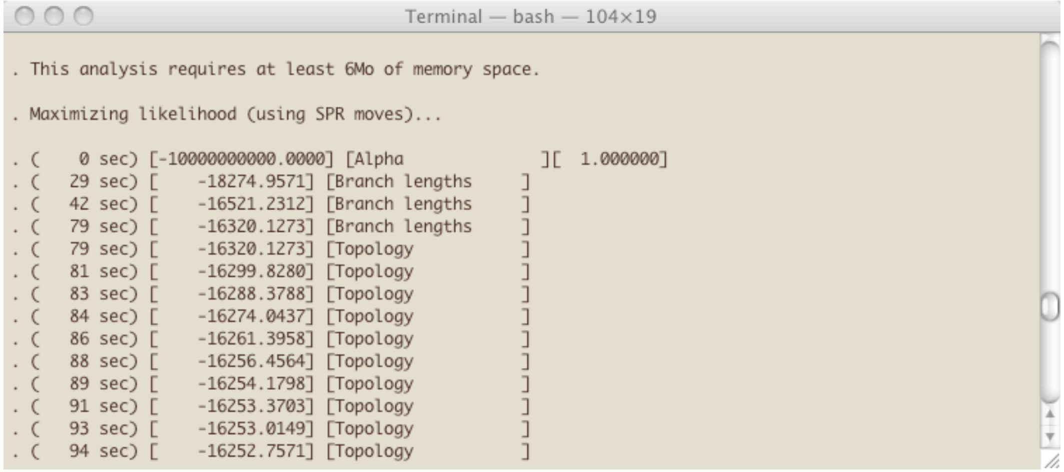
```
                                   .......................
                                    Menu : Tree Searching
                                 .............................


                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [O] ........................... Optimise tree topology  yes
                [U] ........ Starting tree (BioNJ/parsimony/user tree)  BioNJ
                [S] .................. Tree topology search operations  Simulated Thermal Annealing
                [R] ........................ Add random starting trees  no


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  s
```

Figure 11: When using multiple branch lengths, PhyML+M3L enables simulated thermal annealing by default. For this example, let's instead use hill-climbing with subtree pruning and regrafting. Enter ‘s’ to cycle through the search options. Stop cycling when you see the option named ‘SPR moves (slow, accurate)’.



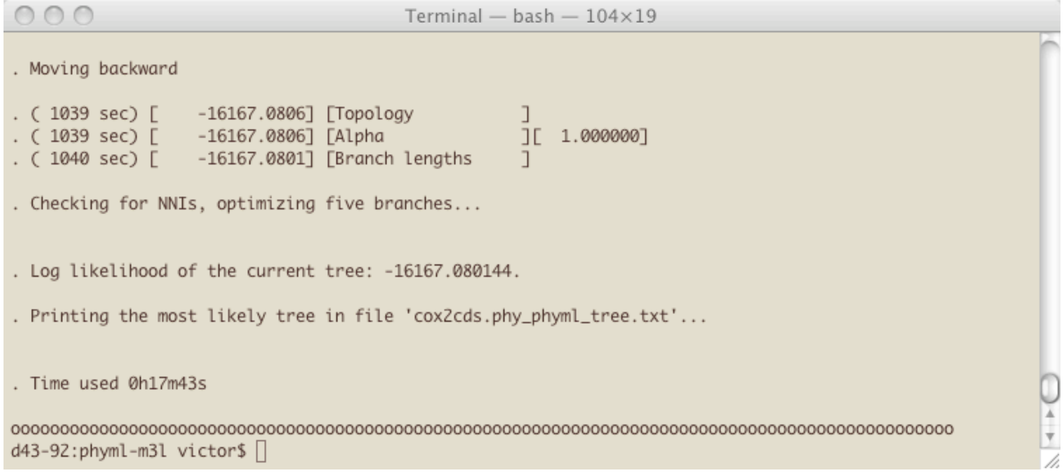
```
                                   .......................
                                    Menu : Tree Searching
                                 .............................


                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [O] ........................... Optimise tree topology  yes
                [U] ........ Starting tree (BioNJ/parsimony/user tree)  BioNJ
                [S] .................. Tree topology search operations  SPR moves (slow, accurate)
                [R] ........................ Add random starting trees  no


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  y
```

Figure 12: After selecting SPR, the menu should like this. Although there exist other menu options we could consider, let's start the analysis. Enter 'y' to begin.



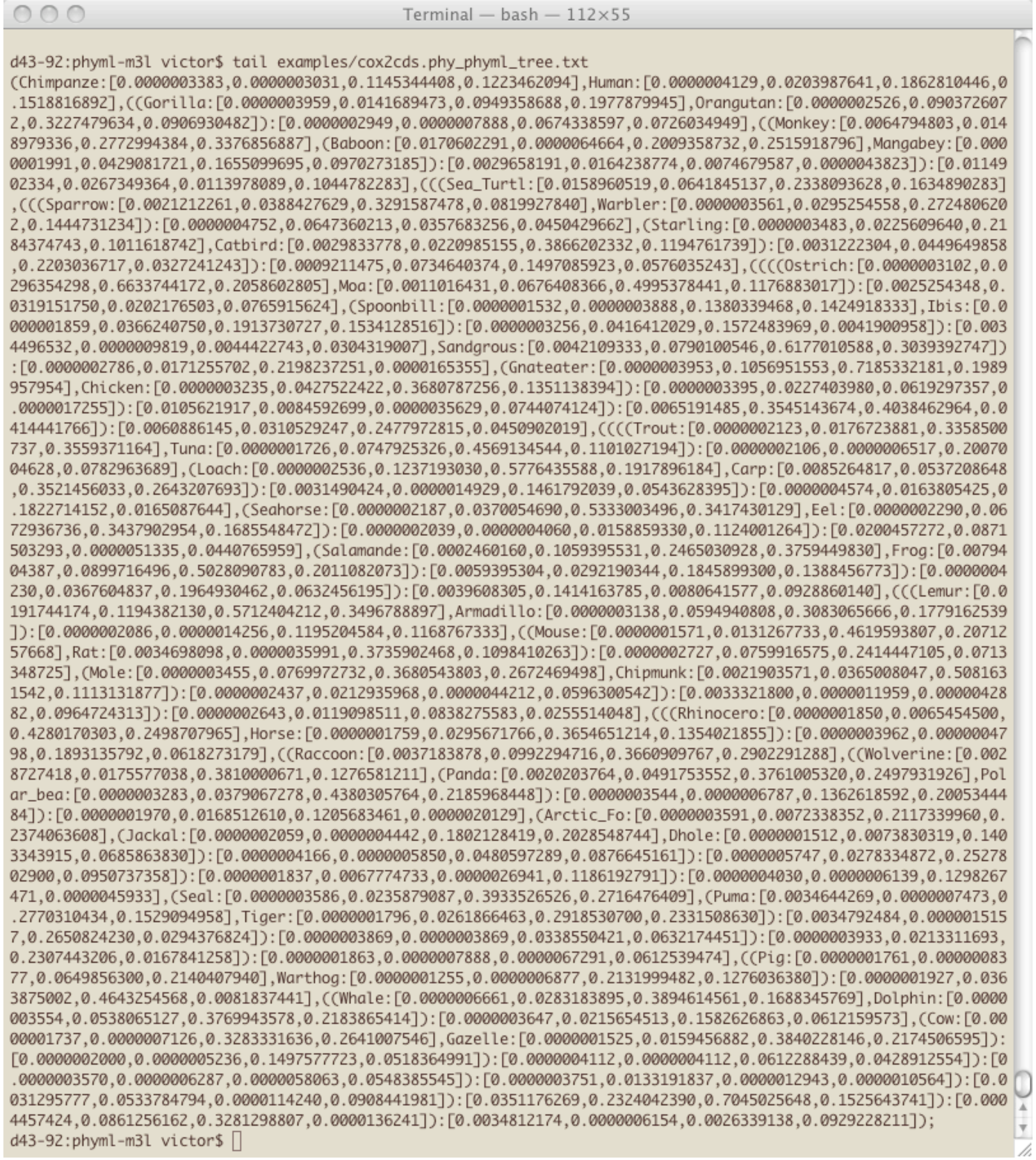
```
. This analysis requires at least 6Mo of memory space.

. Maximizing likelihood (using SPR moves)...

. (    0 sec) [-10000000000.0000] [Alpha             ][  1.000000]
. (   29 sec) [    -18274.9571] [Branch lengths     ]
. (   42 sec) [    -16521.2312] [Branch lengths     ]
. (   79 sec) [    -16320.1273] [Branch lengths     ]
. (   79 sec) [    -16320.1273] [Topology           ]
. (   81 sec) [    -16299.8280] [Topology           ]
. (   83 sec) [    -16288.3788] [Topology           ]
. (   84 sec) [    -16274.0437] [Topology           ]
. (   86 sec) [    -16261.3958] [Topology           ]
. (   88 sec) [    -16256.4564] [Topology           ]
. (   89 sec) [    -16254.1798] [Topology           ]
. (   91 sec) [    -16253.3703] [Topology           ]
. (   93 sec) [    -16253.0149] [Topology           ]
. (   94 sec) [    -16252.7571] [Topology           ]
```

Figure 13: After entering ‘y’, you should see output looking similar to this. The actual numbers will likely be different. Each row of this output displays (1) the total time used, (2) the log likelihood of the current phylogeny, and (3) the parameter that was modified to find this tree. This output will be grow until the hill-climbing algorithm converges on a likelihood maxima. Be patient, the analysis can take a while.



```
. Moving backward

. ( 1039 sec) [    -16167.0806] [Topology           ]
. ( 1039 sec) [    -16167.0806] [Alpha              ][  1.000000]
. ( 1040 sec) [    -16167.0801] [Branch lengths     ]

. Checking for NNIs, optimizing five branches...


. Log likelihood of the current tree: -16167.080144.

. Printing the most likely tree in file 'cox2cds.phy_phyml_tree.txt'...


. Time used 0h17m43s

oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
d43-92:phyml-m3l victor$ 
```

Figure 14: Eventually the analysis will finish, and you should see output that looks similar to this. In this example, the best-found phylogeny has a log-likelihood of -16167.080144. On a dual-core Intel MacBook, the analysis required 17 minutes and 43 seconds. On your computer, perhaps it will take less time.

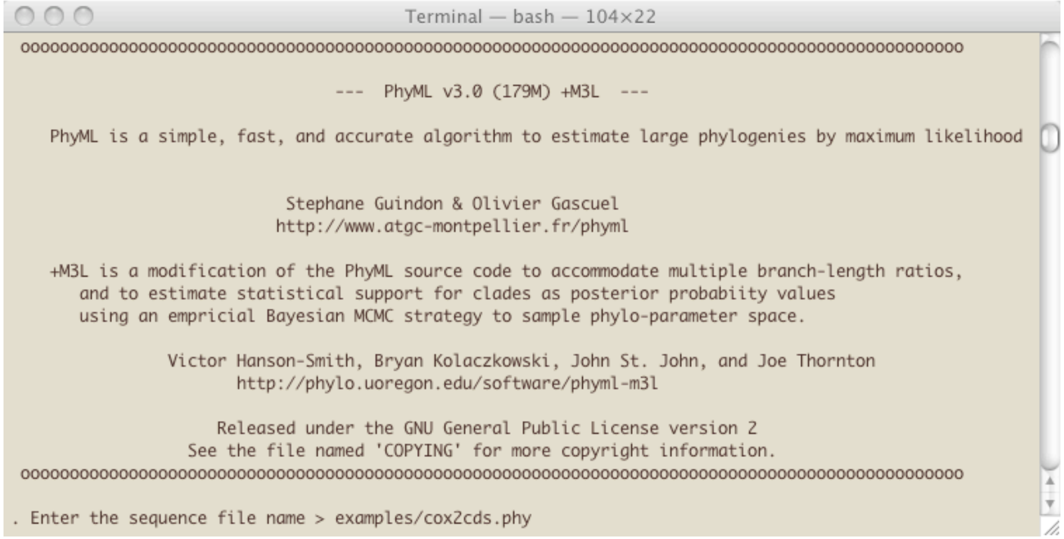


```
d43-92:phyml-m3l victor$ tail examples/cox2cds.phy_phyml_tree.txt
(Chimpanze:[0.0000003383,0.0000003031,0.1145344408,0.1223462094],Human:[0.0000004129,0.0203987641,0.1862810446,0
.1518816892],((Gorilla:[0.0000003959,0.0141689473,0.0949358688,0.1977879945],Orangutan:[0.0000002526,0.090372607
2,0.3227479634,0.0906930482]):[0.0000002949,0.0000007888,0.0674338597,0.0726034949],((Monkey:[0.0064794803,0.014
8979336,0.2772994384,0.3376856887],(Baboon:[0.0170602291,0.0000064664,0.2009358732,0.2515918796],Mangabey:[0.000
0001991,0.0429081721,0.1655099695,0.0970273185]):[0.0029658191,0.0164238774,0.0074679587,0.0000043823]):[0.01149
02334,0.0267349364,0.0113978089,0.1044782283],(((Sea_Turtl:[0.0158960519,0.0641845137,0.2338093628,0.1634890283]
,(((Sparrow:[0.0021212261,0.0388427629,0.3291587478,0.0819927840],Warbler:[0.0000003561,0.0295254558,0.272480620
2,0.1444731234]):[0.0000004752,0.0647360213,0.0357683256,0.0450429662],(Starling:[0.0000003483,0.0225609640,0.21
84374743,0.1011618742],Catbird:[0.0029833778,0.0220985155,0.3866202332,0.1194761739]):[0.0031222304,0.0449649858
,0.2203036717,0.0327241243]):[0.0009211475,0.0734640374,0.1497085923,0.0576035243],((((Ostrich:[0.0000003102,0.0
296354298,0.6633744172,0.2058602805],Moa:[0.0011016431,0.0676408366,0.4995378441,0.1176883017]):[0.0025254348,0.
0319151750,0.0202176503,0.0765915624],(Spoonbill:[0.0000001532,0.0000003888,0.1380339468,0.1424918333],Ibis:[0.0
000001859,0.0366240750,0.1913730727,0.1534128516]):[0.0000003256,0.0416412029,0.1572483969,0.0041900958]):[0.003
4496532,0.0000009819,0.0044422743,0.0304319007],Sandgrous:[0.0042109333,0.0790100546,0.6177010588,0.3039392747])
:[0.0000002786,0.0171255702,0.2198237251,0.0000165355],(Gnateater:[0.0000003953,0.1056951553,0.7185332181,0.1989
957954],Chicken:[0.0000003235,0.0427522422,0.3680787256,0.1351138394]):[0.0000003395,0.0227403980,0.0619297357,0
.0000017255]):[0.0105621917,0.0084592699,0.0000035629,0.0744074124]):[0.0065191485,0.3545143674,0.4038462964,0.0
414441766]):[0.0060886145,0.0310529247,0.2477972815,0.0450902019],((((Trout:[0.0000002123,0.0176723881,0.3358500
737,0.3559371164],Tuna:[0.0000001726,0.0747925326,0.4569134544,0.1101027194]):[0.0000002106,0.0000006517,0.20070
04628,0.0782963689],(Loach:[0.0000002536,0.1237193030,0.5776435588,0.1917896184],Carp:[0.0085264817,0.0537208648
,0.3521456033,0.2643207693]):[0.0031490424,0.0000014929,0.1461792039,0.0543628395]):[0.0000004574,0.0163805425,0
.1822714152,0.0165087644],(Seahorse:[0.0000002187,0.0370054690,0.5333003496,0.3417430129],Eel:[0.0000002290,0.06
72936736,0.3437902954,0.1685548472]):[0.0000002039,0.0000004060,0.0158859330,0.1124001264]):[0.0200457272,0.0871
503293,0.0000051335,0.0440765959],(Salamande:[0.0002460160,0.1059395531,0.2465030928,0.3759449830],Frog:[0.00794
04387,0.0899716496,0.5028090783,0.2011082073]):[0.0059395304,0.0292190344,0.1845899300,0.1388456773]):[0.0000004
230,0.0367604837,0.1964930462,0.0632456195]):[0.0039608305,0.1414163785,0.0080641577,0.0928860140],(((Lemur:[0.0
191744174,0.1194382130,0.5712404212,0.3496788897],Armadillo:[0.0000003138,0.0594940808,0.3083065666,0.1779162539
]):[0.0000002086,0.0000014256,0.1195204584,0.1168767333],((Mouse:[0.0000001571,0.0131267733,0.4619593807,0.20712
57668],Rat:[0.0034698098,0.0000035991,0.3735902468,0.1098410263]):[0.0000002727,0.0759916575,0.2414447105,0.0713
348725],(Mole:[0.0000003455,0.0769972732,0.3680543803,0.2672469498],Chipmunk:[0.0021903571,0.0365008047,0.508163
1542,0.1113131877]):[0.0000002437,0.0212935968,0.0000044212,0.0596300542]):[0.0033321800,0.0000011959,0.00000428
82,0.0964724313]):[0.0000002643,0.0119098511,0.0838275583,0.0255514048],(((Rhinocero:[0.0000001850,0.0065454500,
0.4280170303,0.2498707965],Horse:[0.0000001759,0.0295671766,0.3654651214,0.1354021855]):[0.0000003962,0.00000047
98,0.1893135792,0.0618273179],((Raccoon:[0.0037183878,0.0992294716,0.3660909767,0.2902291288],((Wolverine:[0.002
8727418,0.0175577038,0.3810000671,0.1276581211],(Panda:[0.0020203764,0.0491753552,0.3761005320,0.2497931926],Pol
ar_bea:[0.0000003283,0.0379067278,0.4380305764,0.2185968448]):[0.0000003544,0.0000006787,0.1362618592,0.20053444
84]):[0.0000001970,0.0168512610,0.1205683461,0.0000020129],(Arctic_Fo:[0.0000003591,0.0072338352,0.2117339960,0.
2374063608],(Jackal:[0.0000002059,0.0000004442,0.1802128419,0.2028548744],Dhole:[0.0000001512,0.0073830319,0.140
3343915,0.0685863830]):[0.0000004166,0.0000005850,0.0480597289,0.0876645161]):[0.0000005747,0.0278334872,0.25278
02900,0.0950737358]):[0.0000001837,0.0067774733,0.0000026941,0.1186192791]):[0.0000004030,0.0000006139,0.1298267
471,0.0000045933],(Seal:[0.0000003586,0.0235879087,0.3933526526,0.2716476409],(Puma:[0.0034644269,0.0000007473,0
.2770310434,0.1529094958],Tiger:[0.0000001796,0.0261866463,0.2918530700,0.2331508630]):[0.0034792484,0.000001515
7,0.2650824230,0.0294376824]):[0.0000003869,0.0000003869,0.0338550421,0.0632174451]):[0.0000003933,0.0213311693,
0.2307443206,0.0167841258]):[0.0000001863,0.0000007888,0.0000067291,0.0612539474],((Pig:[0.0000001761,0.00000083
77,0.0649856300,0.2140407940],Warthog:[0.0000001255,0.0000006877,0.2131999482,0.1276036380]):[0.0000001927,0.036
3875002,0.4643254568,0.0081837441],((Whale:[0.0000006661,0.0283183895,0.3894614561,0.1688345769],Dolphin:[0.0000
003554,0.0538065127,0.3769943578,0.2183865414]):[0.0000003647,0.0215654513,0.1582626863,0.0612159573],(Cow:[0.00
00001737,0.0000007126,0.3283331636,0.2641007546],Gazelle:[0.0000001525,0.0159456882,0.3840228146,0.2174506595]):
[0.0000002000,0.0000005236,0.1497577723,0.0518364991]):[0.0000004112,0.0000004112,0.0612288439,0.0428912554]):[0
.0000003570,0.0000006287,0.0000058063,0.0548385545]):[0.0000003751,0.0133191837,0.0000012943,0.0000010564]):[0.0
031295777,0.0533784794,0.0000114240,0.0908441981]):[0.0351176269,0.2324042390,0.7045025648,0.1525643741]):[0.000
4457424,0.0861256162,0.3281298807,0.0000136241]):[0.0034812174,0.0000006154,0.0026339138,0.0929228211]);
d43-92:phyml-m3l victor$ 
```

Figure 15: The resulting phylogeny is printed as a Newick-formatted string in a text file. You can view that text file using your favorite text editor, or using unix commands such as *tail*, *less*, or *cat*. Notice that each branch has four length values wrapped in brackets '[' and ']'.

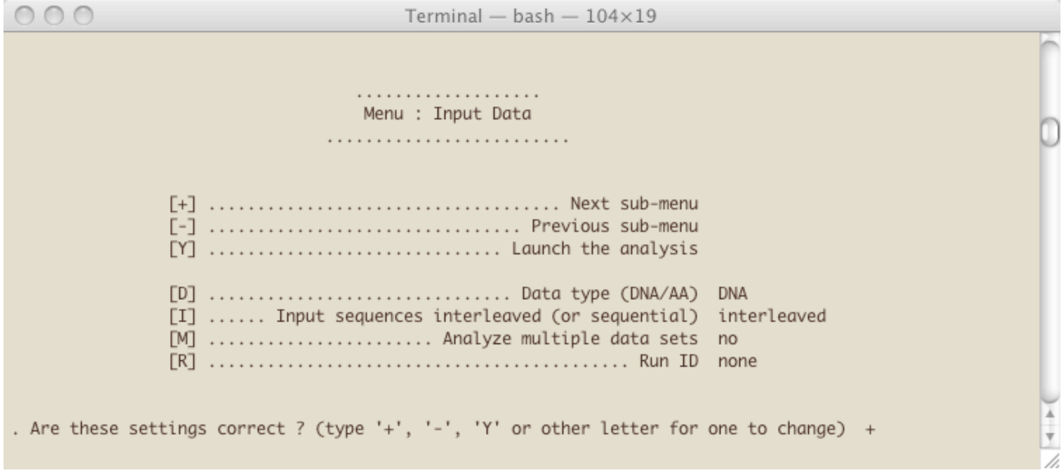
5.2 Using empirical Bayes MCMC to estimate clade support

Begin by navigating to the folder that contains the PhyML+M3L program. This folder should contain a subfolder named *examples*, in which the *cox2cds.phy* alignment is located.



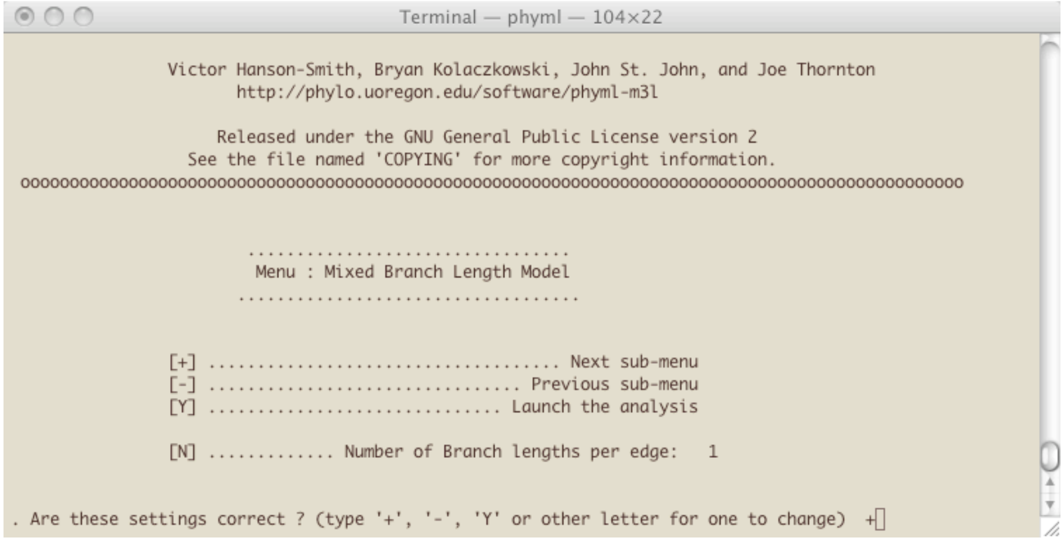
```
 ooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo

                                 ---  PhyML v3.0 (179M) +M3L  ---

    PhyML is a simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood


                            Stephane Guindon & Olivier Gascuel
                           http://www.atgc-montpellier.fr/phyml

    +M3L is a modification of the PhyML source code to accommodate multiple branch-length ratios,
       and to estimate statistical support for clades as posterior probabiity values
       using an empricial Bayesian MCMC strategy to sample phylo-parameter space.

                Victor Hanson-Smith, Bryan Kolaczkowski, John St. John, and Joe Thornton
                       http://phylo.uoregon.edu/software/phyml-m3l

                     Released under the GNU General Public License version 2
                  See the file named 'COPYING' for more copyright information.
 ooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo

. Enter the sequence file name > examples/cox2cds.phy
```

Figure 16: After PhyML+M3L loads, enter the filepath of the *cox2cds.phy* alignment.



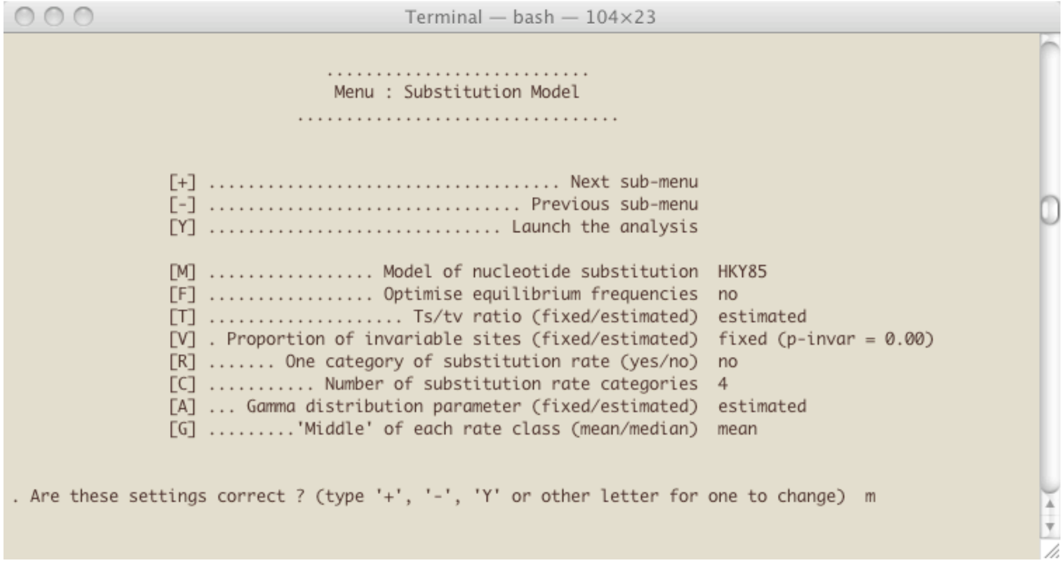
```
                                   ...................
                                    Menu : Input Data
                                ......................... 


                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [D] ............................... Data type (DNA/AA)  DNA
                [I] ...... Input sequences interleaved (or sequential)  interleaved
                [M] ....................... Analyze multiple data sets  no
                [R] ........................................... Run ID  none


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  +
```

Figure 17: The Cox-2 data is DNA (nucleotide) data, so no changes need to be made on this menu page. Enter '+' to advance to the next menu page.



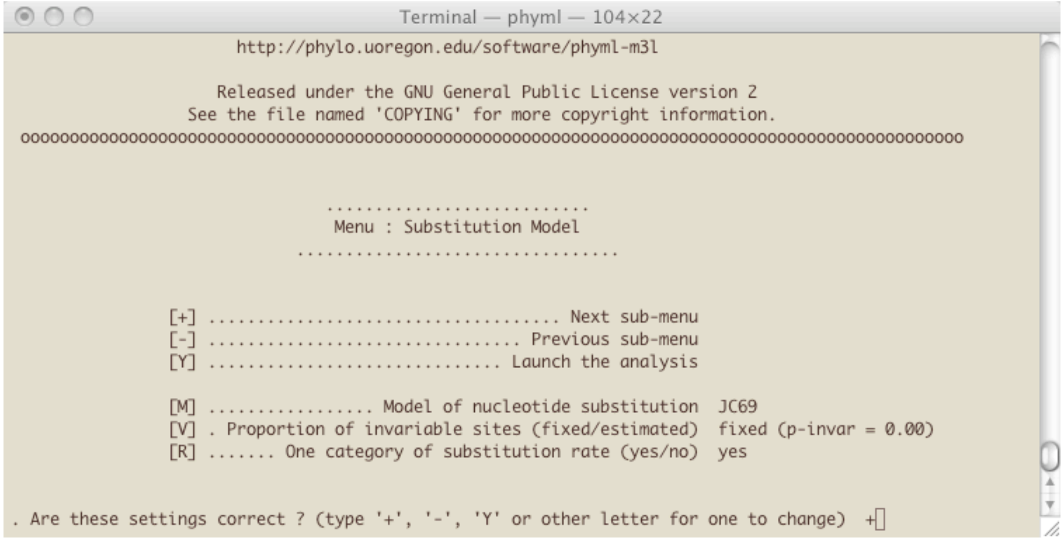
```
                Victor Hanson-Smith, Bryan Kolaczkowski, John St. John, and Joe Thornton
                       http://phylo.uoregon.edu/software/phyml-m3l

                     Released under the GNU General Public License version 2
                  See the file named 'COPYING' for more copyright information.
 ooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo


                        .................................
                         Menu : Mixed Branch Length Model
                       ...................................


                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [N] ............. Number of Branch lengths per edge:   1


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  +
```

Figure 18: For this example, we'll use only one branch length set. Leave this menu alone and enter '+' to advance to the next menu.



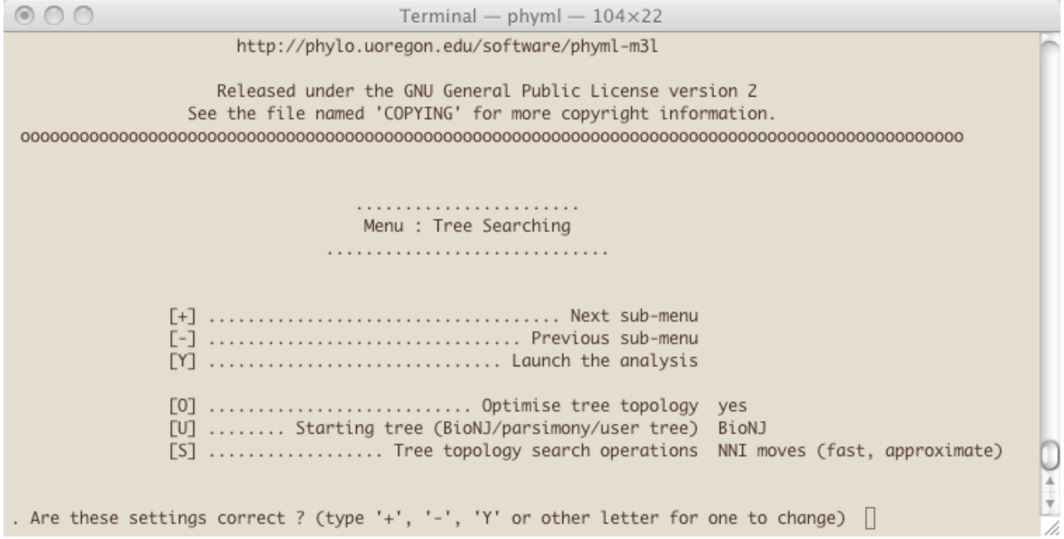
```
                          ..........................
                           Menu : Substitution Model
                       ................................


             [+] .................................... Next sub-menu
             [-] ................................ Previous sub-menu
             [Y] .............................. Launch the analysis

             [M] ................. Model of nucleotide substitution  HKY85
             [F] ................. Optimise equilibrium frequencies  no
             [T] .................... Ts/tv ratio (fixed/estimated)  estimated
             [V] . Proportion of invariable sites (fixed/estimated)  fixed (p-invar = 0.00)
             [R] ....... One category of substitution rate (yes/no)  no
             [C] ........... Number of substitution rate categories  4
             [A] ... Gamma distribution parameter (fixed/estimated)  estimated
             [G] .........'Middle' of each rate class (mean/median)  mean


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  m
```

Figure 19: This menu allows you specify the substitution model. By default, PhyML uses the HKY85 model for nucleotide data. Instead, let’s use the simpler JC69 model. Enter ‘m’ to cycle through the available model options. Stop cycling when you see the model named ‘JC69’. Also, enter ‘r’ to disable the gamma-distributed rates model.



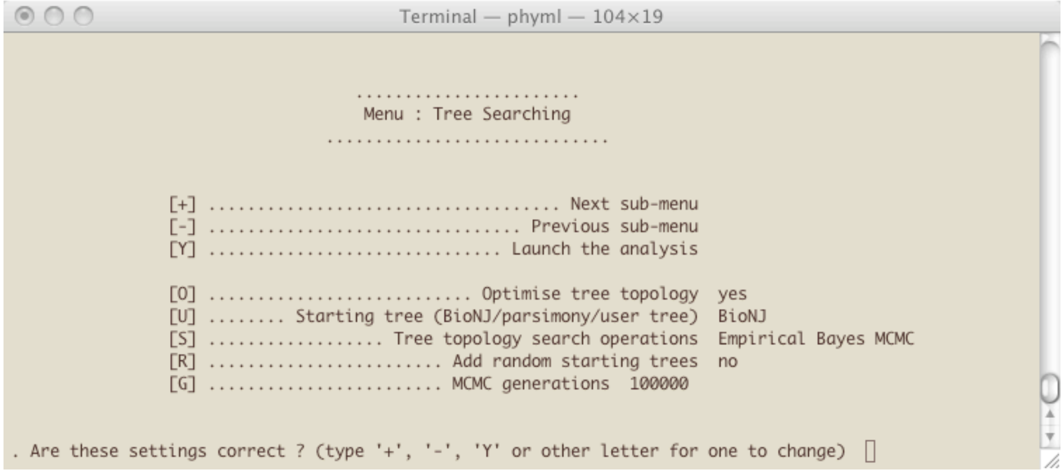
```
                       http://phylo.uoregon.edu/software/phyml-m3l

                     Released under the GNU General Public License version 2
                  See the file named 'COPYING' for more copyright information.
 oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo



                                ...........................
                                 Menu : Substitution Model
                             .................................



                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [M] ................. Model of nucleotide substitution  JC69
                [V] . Proportion of invariable sites (fixed/estimated)  fixed (p-invar = 0.00)
                [R] ....... One category of substitution rate (yes/no)  yes


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)  +
```

Figure 20: At this point, the menu should look like this. Enter '+' to advance to the next menu.



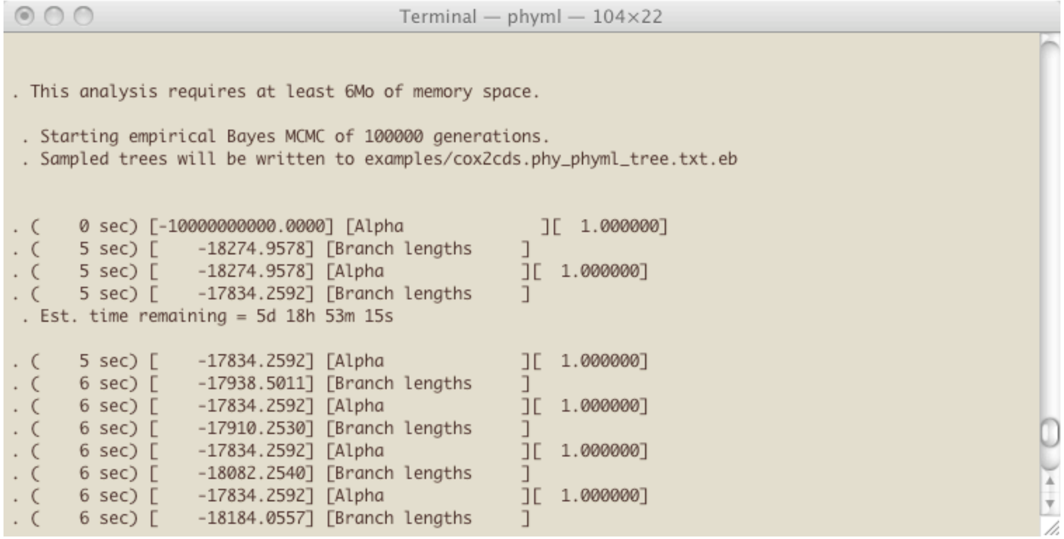
```
                        http://phylo.uoregon.edu/software/phyml-m3l

                      Released under the GNU General Public License version 2
                   See the file named 'COPYING' for more copyright information.
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo



                                  .......................
                                   Menu : Tree Searching
                               .............................



                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [O] ........................... Optimise tree topology  yes
                [U] ........ Starting tree (BioNJ/parsimony/user tree)  BioNJ
                [S] .................. Tree topology search operations  NNI moves (fast, approximate)


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)
```

Figure 21: By default, PhyML uses a tree searching strategy based on nearest neighbor interchange (NNI) or subtree pruning and regrafting (SPR). To enable empirical Bayes MCMC, enter ‘s’ to cycle through the options. Stop cycling when you see ‘Empirical Bayes MCMC’



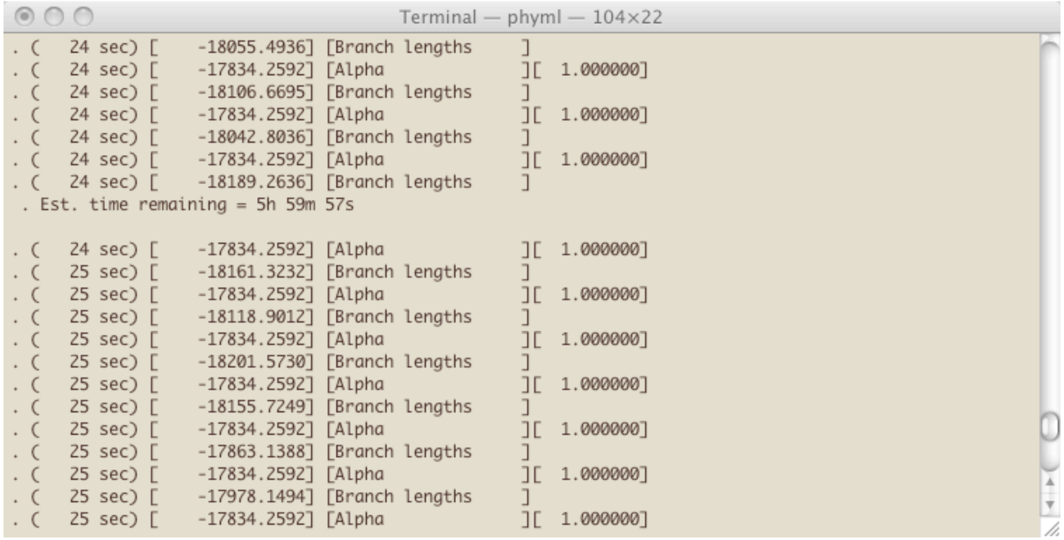
```
                                 .......................
                                  Menu : Tree Searching
                              .............................


                [+] .................................... Next sub-menu
                [-] ................................ Previous sub-menu
                [Y] .............................. Launch the analysis

                [O] ........................... Optimise tree topology  yes
                [U] ........ Starting tree (BioNJ/parsimony/user tree)  BioNJ
                [S] .................. Tree topology search operations  Empirical Bayes MCMC
                [R] ........................ Add random starting trees  no
                [G] ........................ MCMC generations  100000


. Are these settings correct ? (type '+', '-', 'Y' or other letter for one to change)
```

Figure 22: Notice that new options appear when you select empirical Bayes. The option named 'MCMC generations' specifies how many generations to run the MCMC analysis. More generations will lead to a more accurate estimate of posterior probability values. However, more generations will take more time. By default, PhyML+M3L uses 100,000 generations. For most practical analysis, we consider 100,000 to be a minimum number of generations to yield an accurate sample. Enter 'y' to start the analysis.



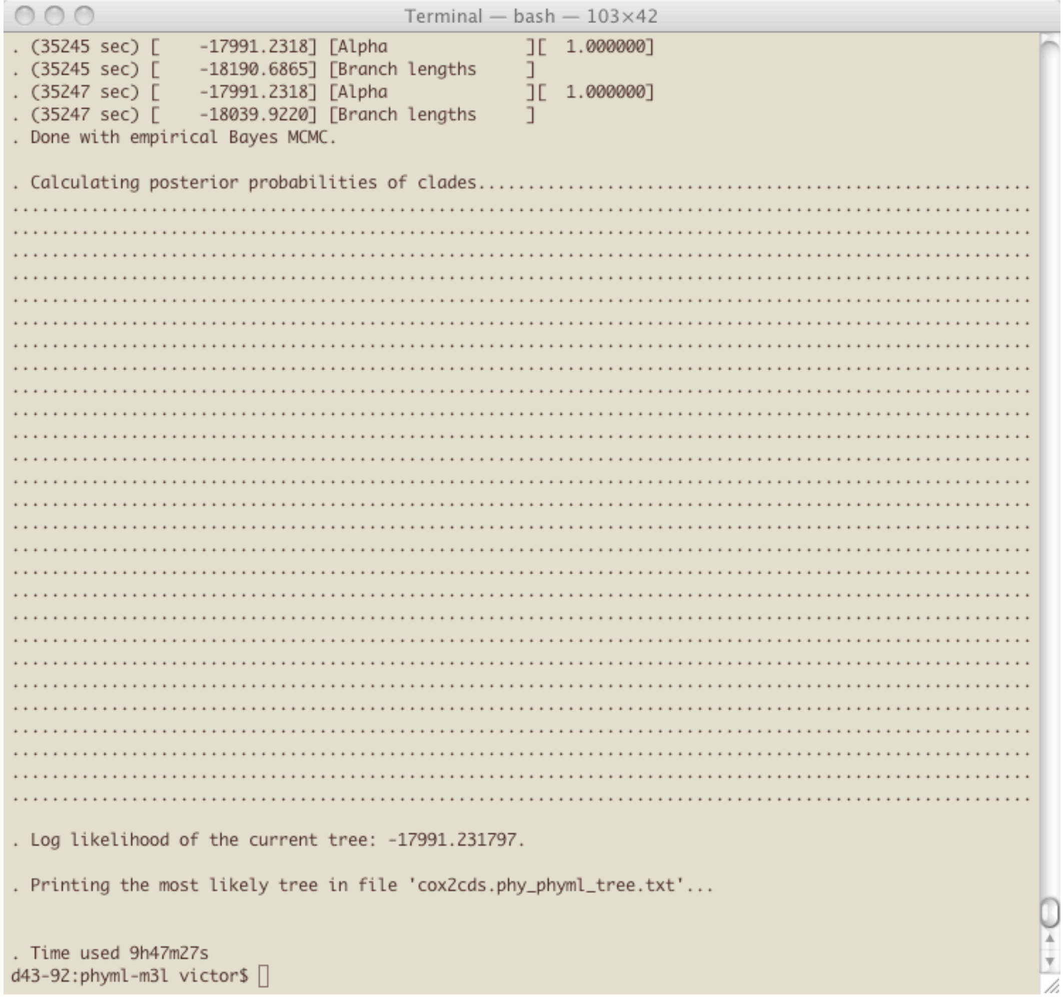
```
. This analysis requires at least 6Mo of memory space.

 . Starting empirical Bayes MCMC of 100000 generations.
 . Sampled trees will be written to examples/cox2cds.phy_phyml_tree.txt.eb


. (    0 sec) [-10000000000.0000] [Alpha             ][  1.000000]
. (    5 sec) [    -18274.9578] [Branch lengths     ]
. (    5 sec) [    -18274.9578] [Alpha              ][  1.000000]
. (    5 sec) [    -17834.2592] [Branch lengths     ]
 . Est. time remaining = 5d 18h 53m 15s

. (    5 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (    6 sec) [    -17938.5011] [Branch lengths     ]
. (    6 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (    6 sec) [    -17910.2530] [Branch lengths     ]
. (    6 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (    6 sec) [    -18082.2540] [Branch lengths     ]
. (    6 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (    6 sec) [    -18184.0557] [Branch lengths     ]
```

Figure 23: Once the analysis begins, you will see output that looks like this. MCMC analysis can be very time consuming, so be patient. If the estimated time remaining is large, we strongly suggest you run the analysis on a computer that remains powered-on with few computational distractions. If you are use PhyML+M3L on a remote cluster, we encourage you to learn about the unix command named *screen* (it allows you to continue running the job after you logout).



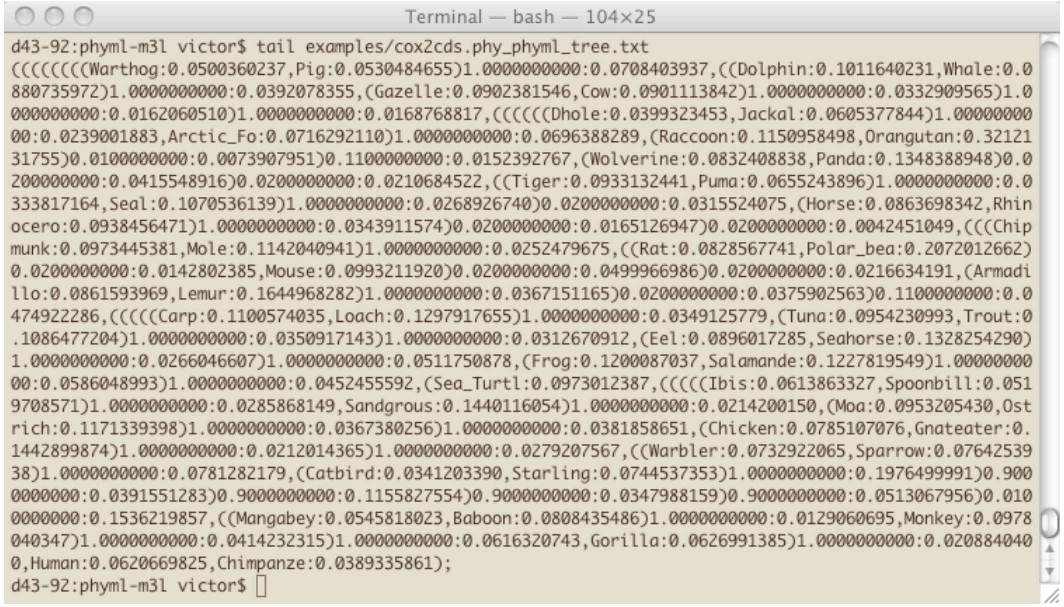
```
. (   24 sec) [    -18055.4936] [Branch lengths     ]
. (   24 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (   24 sec) [    -18106.6695] [Branch lengths     ]
. (   24 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (   24 sec) [    -18042.8036] [Branch lengths     ]
. (   24 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (   24 sec) [    -18189.2636] [Branch lengths     ]
 . Est. time remaining = 5h 59m 57s

. (   24 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (   25 sec) [    -18161.3232] [Branch lengths     ]
. (   25 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (   25 sec) [    -18118.9012] [Branch lengths     ]
. (   25 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (   25 sec) [    -18201.5730] [Branch lengths     ]
. (   25 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (   25 sec) [    -18155.7249] [Branch lengths     ]
. (   25 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (   25 sec) [    -17863.1388] [Branch lengths     ]
. (   25 sec) [    -17834.2592] [Alpha              ][  1.000000]
. (   25 sec) [    -17978.1494] [Branch lengths     ]
. (   25 sec) [    -17834.2592] [Alpha              ][  1.000000]
```

Figure 24: As the analysis continues, the estimate time remaining will decrease.



```
. (35245 sec) [    -17991.2318] [Alpha               ][  1.000000]
. (35245 sec) [    -18190.6865] [Branch lengths      ]
. (35247 sec) [    -17991.2318] [Alpha               ][  1.000000]
. (35247 sec) [    -18039.9220] [Branch lengths      ]
. Done with empirical Bayes MCMC.

. Calculating posterior probabilities of clades........................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................
.......................................................................................................

. Log likelihood of the current tree: -17991.231797.

. Printing the most likely tree in file 'cox2cds.phy_phyml_tree.txt'...


. Time used 9h47m27s
d43-92:phyml-m3l victor$ 
```

Figure 25: Eventually the analysis will finish and posterior probabilities of clades will be calculated. The posterior probability of a clade equals the proportion of the MCMC samples in which that clade appears.



```
d43-92:phyml-m3l victor$ tail examples/cox2cds.phy_phyml_tree.txt
(((((((Warthog:0.0500360237,Pig:0.0530484655)1.0000000000:0.0708403937,((Dolphin:0.1011640231,Whale:0.0
880735972)1.0000000000:0.0392078355,(Gazelle:0.0902381546,Cow:0.0901113842)1.0000000000:0.0332909565)1.0
000000000:0.0162060510)1.0000000000:0.0168768817,((((((Dhole:0.0399323453,Jackal:0.0605377844)1.00000000
00:0.0239001883,Arctic_Fo:0.0716292110)1.0000000000:0.0696388289,(Raccoon:0.1150958498,Orangutan:0.32121
31755)0.0100000000:0.0073907951)0.1100000000:0.0152392767,(Wolverine:0.0832408838,Panda:0.1348388948)0.0
200000000:0.0415548916)0.0200000000:0.0210684522,((Tiger:0.0933132441,Puma:0.0655243896)1.0000000000:0.0
333817164,Seal:0.1070536139)1.0000000000:0.0268926740)0.0200000000:0.0315524075,(Horse:0.0863698342,Rhin
ocero:0.0938456471)1.0000000000:0.0343911574)0.0200000000:0.0165126947)0.0200000000:0.0042451049,(((Chip
munk:0.0973445381,Mole:0.1142040941)1.0000000000:0.0252479675,((Rat:0.0828567741,Polar_bea:0.2072012662)
0.0200000000:0.0142802385,Mouse:0.0993211920)0.0200000000:0.0499966986)0.0200000000:0.0216634191,(Armadi
llo:0.0861593969,Lemur:0.1644968282)1.0000000000:0.0367151165)0.0200000000:0.0375902563)0.1100000000:0.0
474922286,(((((Carp:0.1100574035,Loach:0.1297917655)1.0000000000:0.0349125779,(Tuna:0.0954230993,Trout:0
.1086477204)1.0000000000:0.0350917143)1.0000000000:0.0312670912,(Eel:0.0896017285,Seahorse:0.1328254290)
1.0000000000:0.0266046607)1.0000000000:0.0511750878,(Frog:0.1200087037,Salamande:0.1227819549)1.00000000
00:0.0586048993)1.0000000000:0.0452455592,(Sea_Turtl:0.0973012387,(((((Ibis:0.0613863327,Spoonbill:0.051
9708571)1.0000000000:0.0285868149,Sandgrous:0.1440116054)1.0000000000:0.0214200150,(Moa:0.0953205430,Ost
rich:0.1171339398)1.0000000000:0.0367380256)1.0000000000:0.0381858651,(Chicken:0.0785107076,Gnateater:0.
1442899874)1.0000000000:0.0212014365)1.0000000000:0.0279207567,((Warbler:0.0732922065,Sparrow:0.07642539
38)1.0000000000:0.0781282179,(Catbird:0.0341203390,Starling:0.0744537353)1.0000000000:0.1976499991)0.900
0000000:0.0391551283)0.9000000000:0.1155827554)0.9000000000:0.0347988159)0.9000000000:0.0513067956)0.010
0000000:0.1536219857,((Mangabey:0.0545818023,Baboon:0.0808435486)1.0000000000:0.0129060695,Monkey:0.0978
040347)1.0000000000:0.0414232315)1.0000000000:0.0616320743,Gorilla:0.0626991385)1.0000000000:0.020884040
0,Human:0.0620669825,Chimpanze:0.0389335861);
d43-92:phyml-m3l victor$ 
```

Figure 26: The clade posterior probabilities are printed with Newick-formatted phylogeny in the text file. You can view that text file using your favorite text editor, or using unix commands such as *tail*, *less*, or *cat*.

5.3 Using simulated thermal annealing

Simulated thermal annealing (STA) provides an alternative heuristic for optimizing the topology, branch lengths, and model parameters during phylogenetic inference. For more information about STA in general, see Kirkpatrick et al. (1983) and Kirkpatrick (1984). For information about STA for optimizing phylogenies, see Kolaczkowski and Thornton (2008).

Please be aware that STA is computationally demanding and can require several hours (or days!) to complete.

The STA algorithm is configured using a large number of parameters, defining the length of computation and the shape of several probability distributions. We have provided default values for all these parameters, but you can also specify custom values from the Terminal interface or the command-line. Here we describe these parameters in detail:

Number of annealing stages (`--num_anneal_stages`)

This integer value specifies how many intermediate temperatures will be evaluated as the temperature incrementally decreases from the starting temperature to the final temperature. From some temperature t_i, the next temperature to be evaluated will equal:

$$t_{i+1} = e^{\frac{log(\frac{t_f}{t_s})}{n}}$$

where t_f is the final temperature, t_s is the starting temperature, and n is the number of annealing stages.

Starting temperature (`--temp_start`)

The annealing algorithm will start the system with temperature equal to this value.

Final temperature (`--temp_end`)

The annealing algorithm will terminate when the temperature eventually equals this value.

Iterations per stage (`--iters_per_stage`)

This value specifies how long the algorithm will dwell at each annealing stage. In other words, at each temperature stage, the algorithm will propose (and potentially accept) a number of new parameter combinations equal to this value.

Set-back interval size `( --set_back )`

The algorithm will "set back" the iteration counter for the current stage when the algorithm discovers a combination of parameters that yield a likelihood value higher than any previously observed likelihood. By setting-back the counter, the algorithm will spend more time dwelling at this stage. The rationale for this strategy is that we should not prematurely cool the system if the current temperature is still yielding optimal solutions.

Acceptance ratio `( --acc_ratio )`

This value specifies the acceptance ratio for new proposals of parameter values. Larger values are more permissive to changes that result in worse likelihoods. The probability of accepting a new proposal equals:

$P = e^{\frac{L_{i+1} - L_i}{r - t}}$

where L_{i+1} is the log likelihood of the proposed state, L_i is the log likelihood of the current state, r is the acceptance ratio, and t is the current temperature of the system.

Maximum alpha value `( --max_alpha )`

This value specifies the maximum α value for a Dirichlet distribution. The annealing algorithm draws random arrays of values from a Dirichlet distribution for the purposes of proposing a new set of state frequencies, a new set of relative substitution rates (for nucleotide data using HKY85 or GTR models), and a new set of branch lengths.

Do not confuse this α value with the α parameter in the gamma-distributed model of among-site rate variation.

Branch length sigma value `( --brlen_sigma )`

This value specifies the σ value for describing a Gaussian distribution from which random branch lengths will be drawn to generate a new annealing proposal.

P-invar sigma value `( --pinvar_sigma )`

This value specifies the σ value for describing a Gaussian distribution from which a random proportion of invariant sites will be drawn to generate a new annealing proposal. This value is only valid if the invariant sites model is enabled.

Gamma sigma value (`--gamma_sigma`)

This value specifies the σ value for describing a Gaussian distribution from which a random α value will be drawn to generate a gamma-distributed set of evolutionary rates. This value is only valid if the gamma-distributed model is enabled.

Probability of stepping . . .

At each iteration, the annealing algorithm generates a new proposed combination of parameter values, where each parameter value is probabilistically perturbed from the current set of parameter values. The probability of perturbing, or "stepping," each parameter is defined by a unique value; these values are described below.

`--prob_topology`

Probability of proposing a new topology.

`--prob_spr`

Probability of proposing a new tree using subtree pruning and regrafting (SPR).

`--prob_brlen`

Probability of proposing a new set of branch lengths.

`--prob_kappa`

Probability of proposing a new κ value (for nucleotide data), where κ is the ratio of transitions to transversions.

`--prob_lambda`

Probability of proposing a new λ value (for nucleotide data), where λ defines the ratio of transitions to transversions in the F84 and TN93 models.

`--prob_gamma`

Probability of proposing a new α value for describing the shape of the gamma-distribution for among site rate heterogeneity. This value is only value if the gamma-distributed model is enabled.

`--prob_rr`

Probability of proposing a new set of relative substitution rates (for nucleotide data).

`--prob_rate_proportions`

Probability of proposing a new set of proportions for the gamma-distributed model of among site rate variation.

`--prob_pinvar`

Probability of proposing a new proportion of invariant sites. This value is only valid if the invariant-sites model is enabled.

`--prob_pi`

Probability of proposing a new set of relative state frequencies. This value is only valid if the state frequencies are not fixed.

6 Frequently Asked Questions

6.1 Why is this software named PhyML+M3L?

The name "M3L" was originally M^{3}L (or M "cubed" L), standing for *m*ixed *m*odel *m*aximum *l*ikelihood. M3L's project scope later grew to include the empirical Bayes MCMC sampler and multiprocessor parallelization. At that point, we decided to keep the M3L moniker, despite the fact that M3L does not capture the entire project scope.

6.2 How do I visualize a phylogeny with mixed branch lengths?

Unfortunately, we do not know of any software for visualizing phylogenies with mixed branch lengths. If your tree has only one branch length category, we recommend using the software FigTree:

```
http://tree.bio.ed.ac.uk/software/figtree/
```

6.3 What about Windows?

We have not tested PhyML or PhyML+M3L on Windows PCs. In theory, you should be able to download the C source code and build the application yourself. If you find success running our software on Windows, send us an email; we would love to know it works!

6.4 I cannot compile the source code! I get an error relating to the 'gsl', 'libgsl', or the GNU Scientific Library.

The +M3L extensions heavily employ methods from the GNU Scientific Library (GSL). Among other features, the GSL provides useful C code for sampling random numbers from distributions. On some -ix operating systems, including some distributions of Linux, the GSL BLAS library is installed in unique locations. If your C compiler (such as **gcc** or **g++**) is complaining about missing references from GSL, try opening the file named 'Makefile.am' and add the following tag to the 'AM_LDFLAGS' line:

```
-lgslcblas
```

In other words, transform this original line:

```
AM_LDFLAGS=-lm -lgsl
```

. . . into this new line:

```
AM_LDFLAGS=-lm -lgsl -lgslcblas
```

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281.

Brent, R. P. (1972). *Algorithms for Minimisation without derivatives (automatic computation)*. Prentice Hall.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704.

Huelsenbeck, J. P. and Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecological Systems*, 28:437–466.

Huelsenbeck, J. P. and Rannala, B. (1997). Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science*, 276(5310):227–32.

Huelsenbeck, J. P. and Ronquist, F. (2001). Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34(5):975–986.

Kirkpatrick, S., Jr., C. G., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kolaczkowski, B. and Thornton, J. W. (2007). Effects of branch length uncertainty on bayesian posterior probabilities for phylogenetic hypotheses. *Molecular Biology and Evolution*, 24(9):2108–2118.

Kolaczkowski, B. and Thornton, J. W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution*, 25(6):1054–1066.

Kolaczkowski, B. and Thornton, J. W. (2009). Long-branch attraction bias and inconsistency in bayesian phylogenetics. *PLoS ONE*, 4(12):e7891.

Penny, D., McComish, B. J., Charleston, M. A., and Hendy, M. D. (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J Mol Evol*, 53(6):711–23.

Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *J Mol Evol*, 39(3):306–14.