**Practical 1c: Maximum likelihood analysis in PhyML**

**Data set**

* Sequence alignment in phylip format of samples collected before the 2013-2016 epidemic:

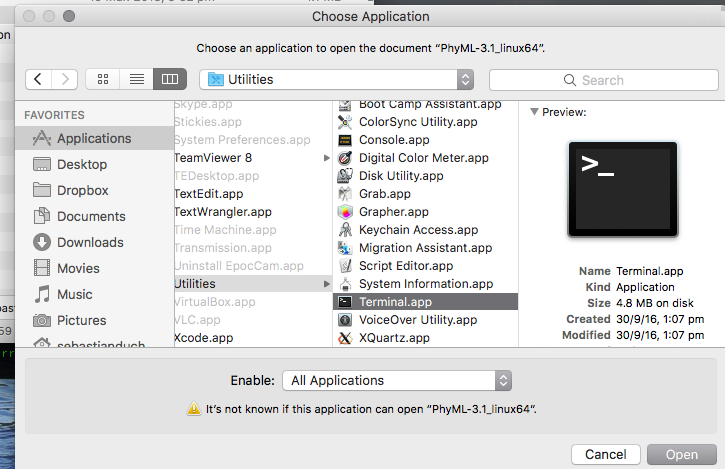
EBOV\_N2\_aligned\_early\_samples.phy

**Software**

* PhyML
* FigTree

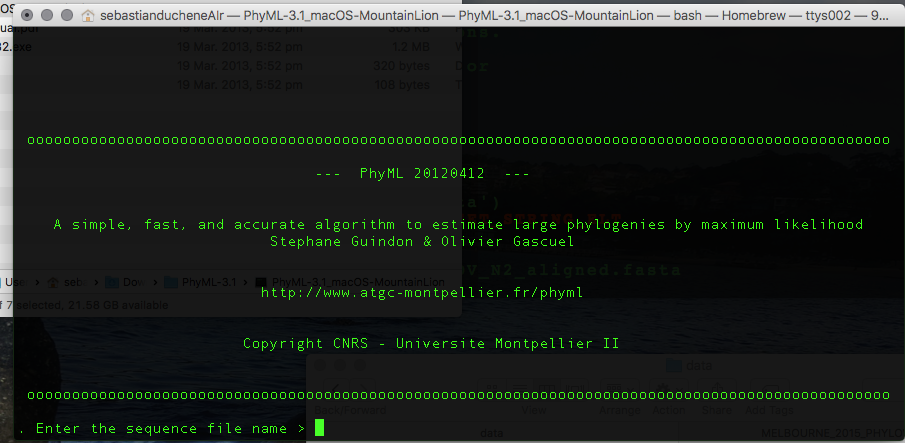
PhyML is typically used through the command-line. We will learn how to use it interactively. Please refer to the user manual, or ask me later about running PhyML in batch mode.

Open the PhyML folder (PhyML-3.1). It should contain executable versions for different operating systems. In windows, double click the .txt file (PhyML-3.1\_win32.exe). In OSX and Linux machines the system might prompt you to choose a program to open this file. If this happens, click on *Choose Application...* and find a folder called *Utilities*. Make sure that at the bottom of the window you select *Enable: All Applications* (Fig 1):



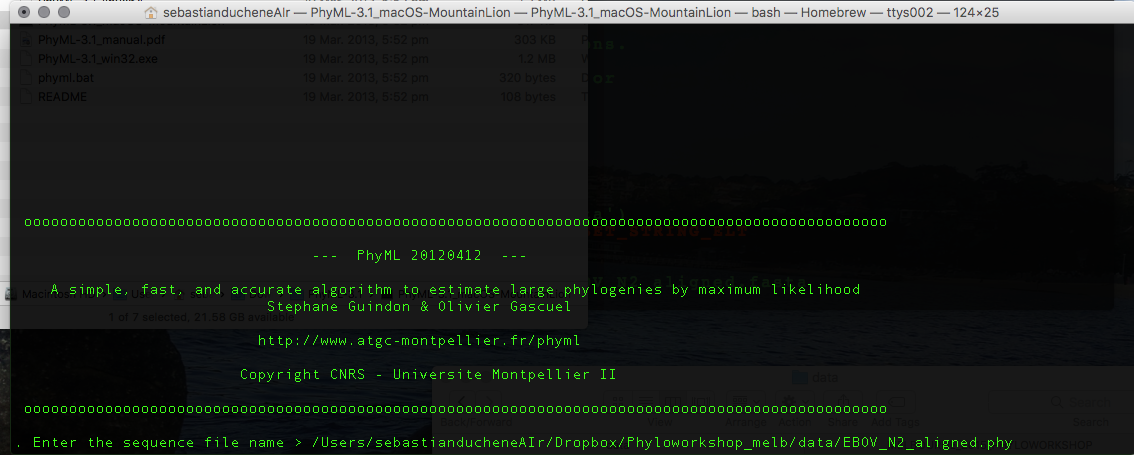
**Fig 1.** Choosing Terminal to open PhyML.

In all operating systems, a command-line window will appear (Fig 2).



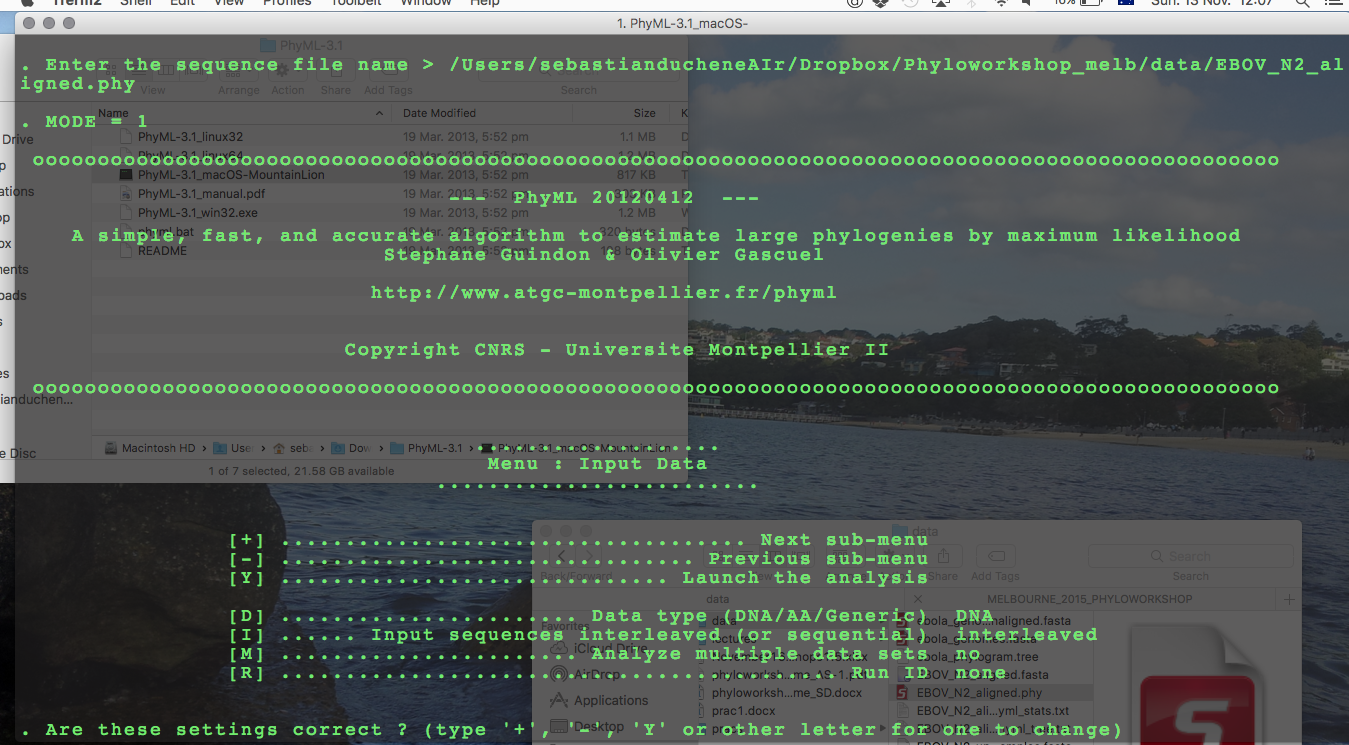
**Fig 2.** Command-line interface for PhyML.

It will ask for the sequence alignment file. Drag the EBOV\_N2\_aligned\_early\_samples.phy. The full path to the file will appear in the window. **Sometimes doing this adds a space at the end of the file path.** Delete it and hit Enter. If you run into an error, it might be that some of the folder names have spaces. In this case, change the location of the alignment, or delete all spaces from file and folder names.



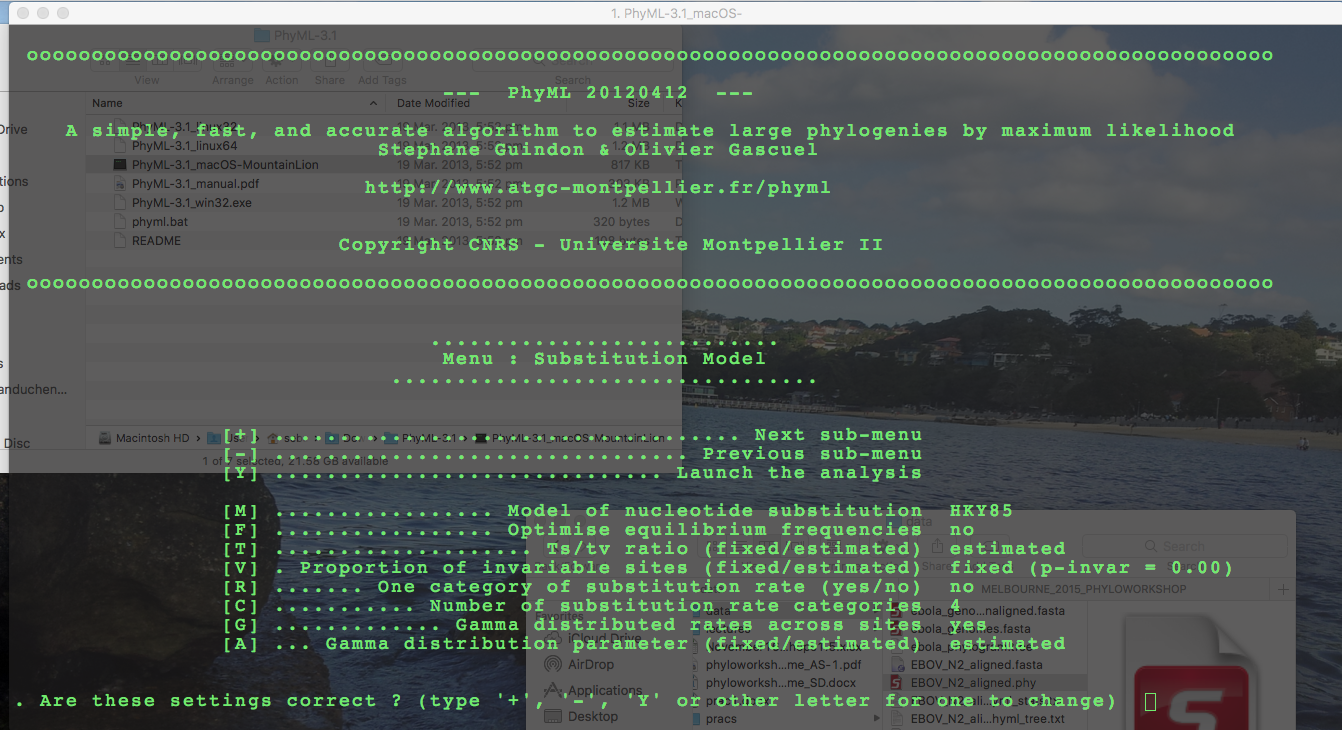
**Fig 3.** File path in PhyML. Note that there should be no spaces at the end of the file path.

You should now see a menu with some options (Fig 4).

****

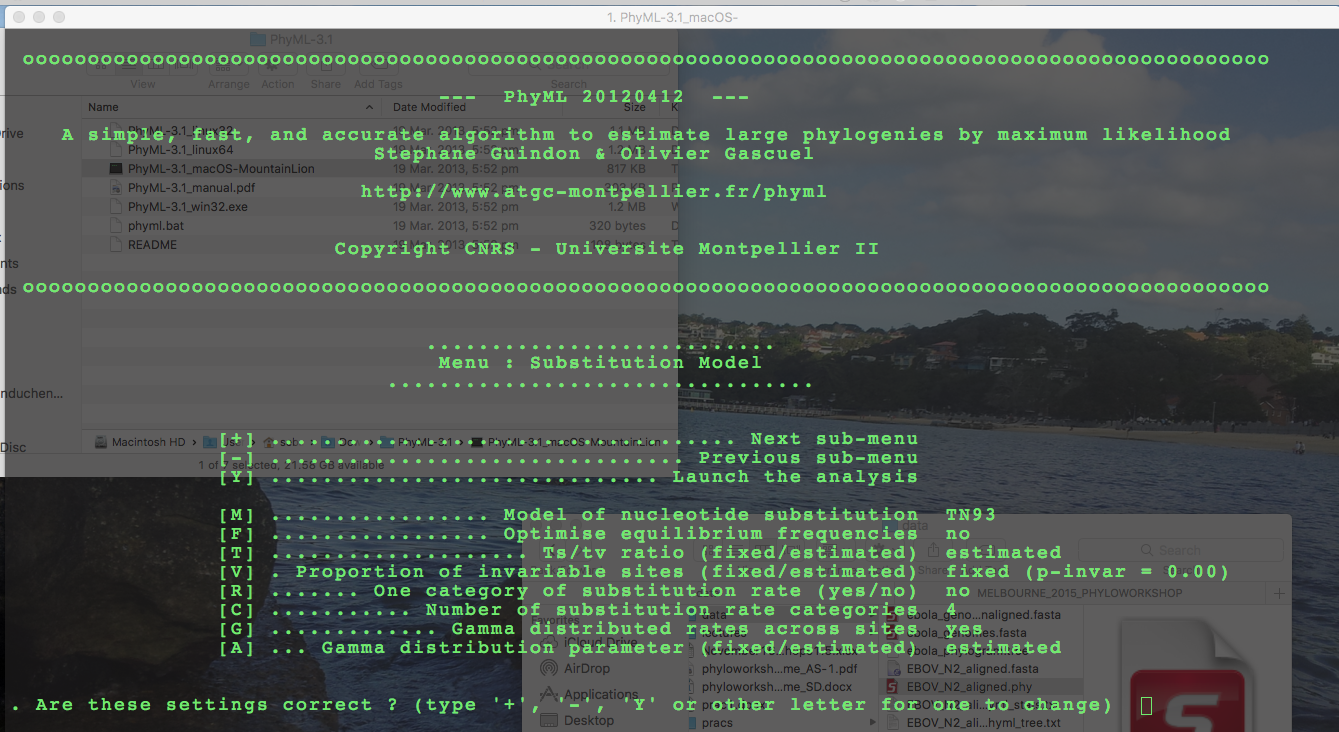
**Fig 4.** Options in PhyML.

These settings are OK. Type + to go to the next menu. This will present some options for the substitution model (Fig 5).



**Fig 5.** Substitution model options in PhyML.

Type M and hit enter a few times until you see the TN93 model, which was selected in MEGA. The resulting set up is in Fig 6.

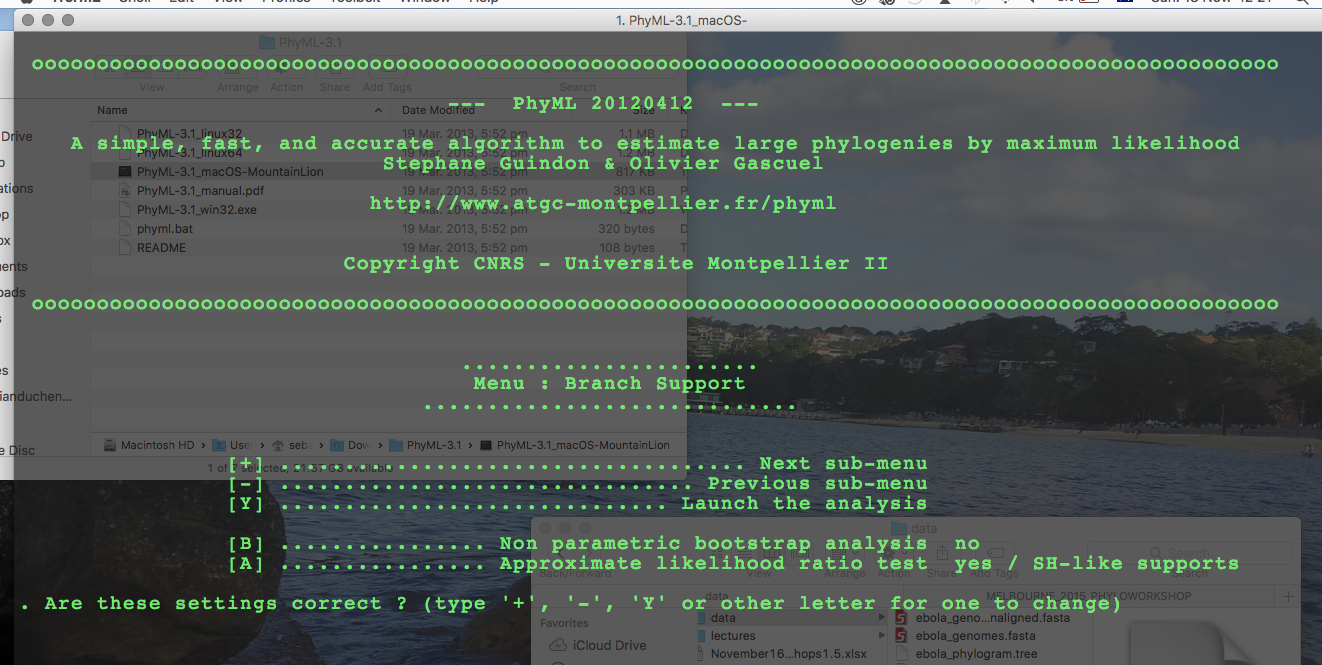


**Fig 6.** Substitution model set up in PhyML.

The substitution model should match the optimal model from MEGA. In this case, we are using a TN93 substitution matrix, which allows for transition-to-transversion bias, a gamma distribution with four categories, and different frequencies for nucleotides. We are assuming no invariable sites.

To go to the next submenu, type + and hit enter. We can set up other options about optimising the tree topology. These are fine for this analysis, but to get more accurate results, change the Tree topology search operations from NNI to SPR (this might not make a difference in this analysis because the data are informative, and are not many sequences). Type + and hit enter.

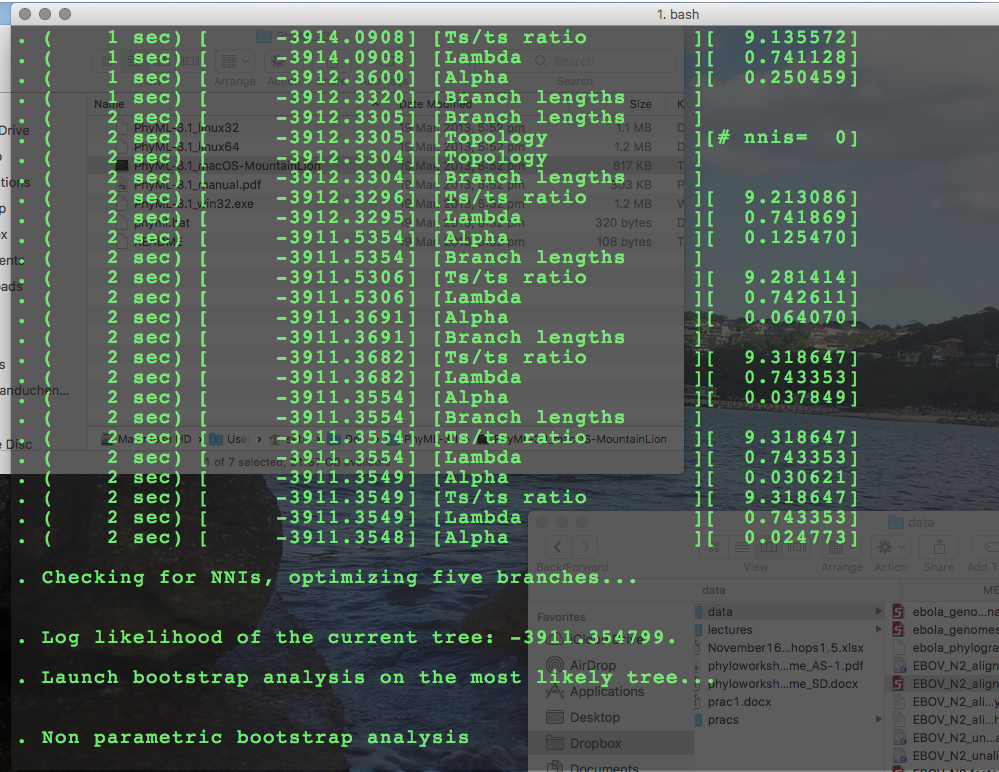
The next submenu is for options of assessing branch support (Fig 7).



**Fig 7.** Options for branch support in PhyML.

PhyML can perform a local topology test to assess branch support, which is very fast and performs similarly to the non-parametric bootstrap. For comparison, however, we will use the non-parametric bootstrap here.

Type A and hit enter to disable the Approximate likelihood ratio test (you might need to do this twice), and then type B and hit enter. The program will ask you how many bootstrap replicates it should perform. Type 100 and hit enter. It will then ask you to whether it should print bootstrap trees and statistics, which we do not require here. Select N and click enter. Out analysis in PhyML is now ready to run. Type Y and hit enter. The analysis will start. You should see the screen in Fig 8.



**Fig 8.** Progress of maximum likelihood analysis in PhyML.

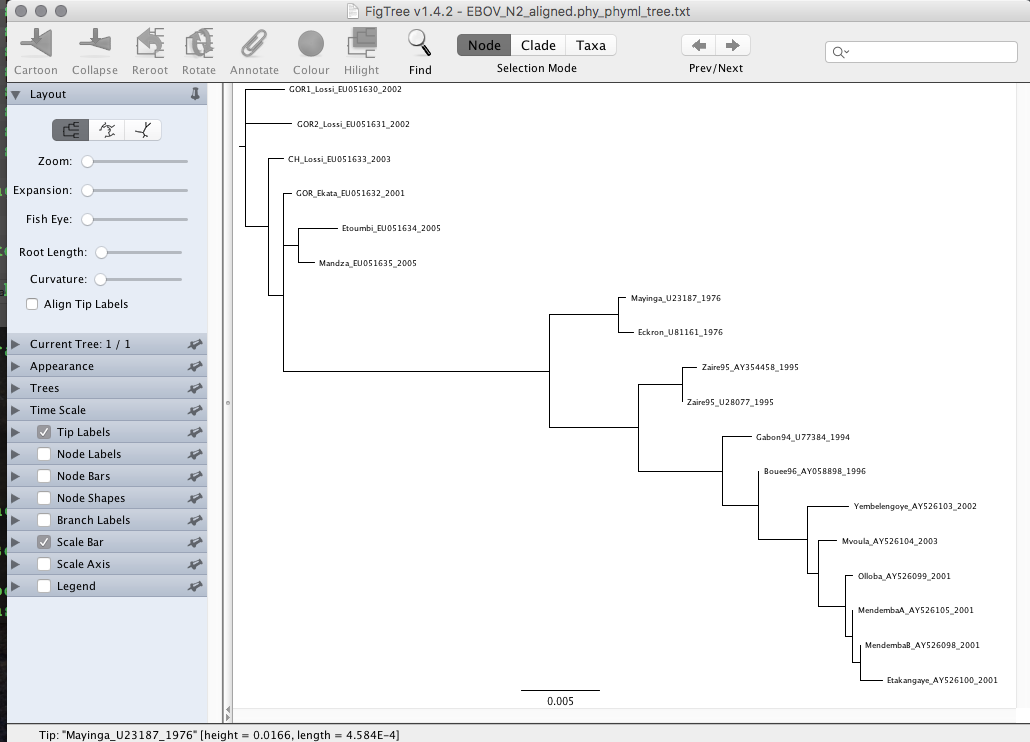
The output from PhyML is a tree file (EBOV\_N2\_aligned\_early\_samples.phy\_phyml\_tree.txt) and a file with the maximum likelihood parameter estimates (EBOV\_N2\_aligned\_early\_samples.phy\_phyml\_stats.txt). Open both in a text editor.

**Question 3.1:** Is there a strong transition-to-transversion bias in these data? Does it appear to differ between purines and pirimidines?

**Question 3.2:** Do you think that it is reasonable to assume that all sites evolve at the same rate in these data?

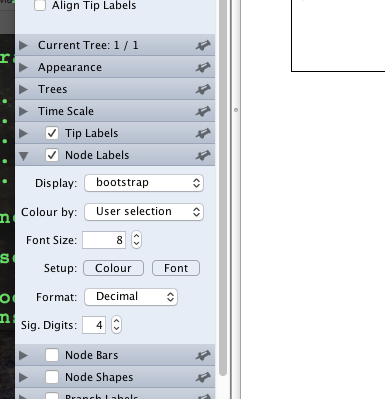
**Question 3.3:** Why do we expect the bootstrap analysis to take longer than an ordinary analysis without branch support?

To visualise the tree we will use FigTree. Click on the program icon . The to the *File* menu and select *Open*. Find the tree file from PhyML (EBOV\_N2\_aligned\_early\_samples.phy\_phyml\_stats.txt). The program will ask us about some branch labels in the tree, which we know to be the bootstrap values. Type ‘bootstrap’ in lieu of ‘label’, and click OK. A window with the tree will appear (Fig 9).



**Fig 9.** Phylogenetic tree displayed in FigTree.

Select *Node Labels* in the menu at the left. Then click on the arrow at the right and select *Display: bootstrap* (Fig 10).



**Fig 10.** Displaying bootstrap values in nodes in FigTree.

The values that appear at the nodes represent the number of bootstrap replicates that contained a particular clade. Use the options in the window to display the tree in different ways. Try selecting different branches and rerooting it.

**Question 3.4:** Sequences starting with CH were obtained from chimps, those starting with GOR were sampled from gorillas, and the rest are from humans. Do human samples all form a monophyletic group? Does this support the notion that Ebola virus has an animal reservoir? If so, can we establish what this reservoir would be?

**Question 3.5:** What does the scale bar at the bottom mean?

**Optional Exercise:** Conduct the same analysis in PhyML, but instead of conducting non-parametric bootstrap, use the SH-like branch support. These values range between 0 and 1. Do they produce similar support as the non-parametric bootstrap?

**Optional Exercise**: Open the Virus Pathogen Resource data base (<https://www.viprbrc.org>) in a web browser and check the location of the strains in this analyses, which are shown for each sequence. For example, Mayinga\_U23187\_1976 means that the sample is from the Mayinga strain, it has the GenBank accession number U23187, and it was sampled in 1976. Is there evidence for geographic clustering in these data?

**Running PhyML in batch mode**

PhyML can also be run using a single command. To do this open you command line interface (Terminal in mac), drag and drop PhyML and use the –h flag before hitting Enter (Fig 11):



**Fig 11.** Command line to get help in PhyML.

A set of options will appear on screen, which we can use to set up an analysis in a single line. Set your working directory in the command line window as we did above so that it is in the same folder as the Ebola sequence alignment. Drag and drop PhyML and use the following command (edit the PhyML path for your machine):

phyml -i EBOV\_N2\_aligned\_early\_samples.phy -a e -s BEST -b 100 --run\_id commandLineTesting --n\_rand\_starts 10

**Question 3.6:** What do each of these options do (-i, -a, -s, –b –run\_id, and –n\_rand\_starts)?

**Question 3.7:** Open the file called EBOV\_N2\_aligned\_early\_samples.phy\_phyml\_tree\_commandLineTesting.txt in a text editor. Is there any variation in the likelihood for the analyses with different seeds? What does this imply about our confidence in finding the trees with the maximum likelihood?