Phylodynamics tutorial

In this tutorial, we will use genomics to investigate the dynamics of various phases of the SARS-CoV-2 pandemic. A set of 3 alignments are provided:

* sarscov2\_global.fasta, sequences sampled globally over the entire course of the pandemic
* sarscov2\_delta.fasta, sequences of the Delta variant sampled from the UK during the period of exponential growth of that variant
* sarscov2\_intro.fasta, sequences of the Delta variant from both the UK and India

You will need to have the following software packages available on your computer

* BEAST 1 (<https://beast.community/>) and its components BEAUTi and TreeAnnotator (part of the same package)
* TempEst (<http://tree.bio.ed.ac.uk/software/tempest/>)
* IQ-TREE (<http://www.iqtree.org/>)
* FigTree (<https://github.com/rambaut/figtree/releases>)
* Tracer (<https://github.com/beast-dev/tracer/releases>)

Using TempEst for checking a molecular clock

In this tutorial, we will explore the use of the interactive graphical program [TempEst](https://beast.community/tempest) to examine SARS-CoV-2 sequence data that has been sampled through time to look for problematic sequences and to explore the degree and pattern of temporal signal. This can be a useful way of examining the data for potential issues before committing significant time to running [BEAST](https://beast.community/beast).

Building a non-molecular clock tree

To examine the relationship between genetic divergence and time (temporal signal), we require a phylogenetic tree constructed without assuming a molecular clock. There is a wide range of suitable software packages (i.e., PhyML, RAxML, GARLI) but for this tutorial we are going to use [IQ-Tree](http://www.iqtree.org/) which uses a fast and effective stochastic algorithm to infer phylogenetic trees by maximum likelihood.

Install **IQ-Tree** using [the instructions on the website](http://www.iqtree.org/) and open a command-line prompt, navigating to the directory containing the data file sarscov2\_global.fasta.

To build a maximum likelihood phylogenetic tree using the GTR+gamma model type:

iqtree -s sarscov2\_global.fasta -m GTR+G

This will create a set of files in the directory containing various outputs and results. For our purposes we only need the maximum likelihood tree file sarscov2\_global.fasta.treefile. You can delete the other files if you like.

Running TempEst and loading the tree

A picture containing device

Description automatically generated

Run [TempEst](https://beast.community/tempest) by double clicking on its icon. TempEst is an interactive graphical application for examining the temporal signal in a tree of time-stamped sequences by plotting the divergence of each tip from the root against the date of sampling (a root-to-tip plot).

Once running, TempEst will look similar irrespective of which computer system it is running on. For this tutorial, the Mac OS X version will be shown but the Linux & Windows versions will have exactly the same layout and functionality.

When started, TempEst will immediately display a file selection dialog box in which you can select the tree that you made in the previous section.

Table

Description automatically generated

Select sarscov2\_global.fasta.treefile and click Open.

Parsing dates of sampling

Once the tree is loaded the main window will appear and look like this:

Graphical user interface, text, application

Description automatically generated

Ignore the panel on the left for the moment. The first thing that needs doing is to give the date of sampling to each of the sequences.

The date of sampling is given at the end of the name of each taxon. To specify the dates of the sequences in BEAUti we will use the Parse Dates button at the top of the panel. Clicking this will make a dialog box appear:

Graphical user interface, application

Description automatically generated

The defaults will in fact work fine here, so click OK.

The table will now have the year of sampling for each virus in the Dates column. Click on the Dates column header to sort the dates and check that they are all correct.

Table

Description automatically generated

The temporal signal and rooting

We can now explore the data using the tabs at the top of the window - Tree, Root-to-tip & Residuals. If you click on the Tree tab you will see the tree as loaded from the tree file. Because we constructed this tree using a non-molecular-clock model, it will be arbitrarily rooted. If you look at the date of each virus in the tree you will see that there is no correlation with the horizontal position:

Diagram

Description automatically generated

Now switch to the Root-to-tip panel. This shows a plot of the divergence from the root of the tree against time of sampling (a so-called ‘Root to tip plot’):

Chart, scatter chart

Description automatically generated

You can see that there is only a limited amount of correlation in this plot (the line is the best-fit regression). In the table on the left you can see the Correlation Coefficient is 0.39. This lack of correlation is expected as the root is arbitrarily set by the phylogeny reconstruction software and thus divergence from root is meaningless. **TempEst** can try to find the root of the tree that optimizes the temporal signal. It does this by trying all possible roots and picks the one that produces the optimal value of a range of statistics. The function it uses is selected in the menu at the top left. The options are to minimize the mean of the squares of the residuals (residual-mean-squared), or to maximize the correlation coefficient (correlation) or R2 (R squared). These are all *ad hoc* procedures and no particular one is best but residual-mean-squared may be most consistent with the investigations here.

Click Best-fitting root to root the tree at the place that minimizes the mean of the squares of the residuals.

Chart, scatter chart

Description automatically generated

Now there is a better correlation between the dates of the tips and the divergence from the root (the correlation coefficient has more than doubled). Return to the tree to look where the root was placed:

Graphical user interface, diagram

Description automatically generated

The high correlation coefficient suggests that there is sufficient temporal signal in this data for us to continue, and employ BEAST for a phylodynamic reconstruction.

Running BEAST for the first time

This tutorial will guide you through running BEAST and some of its accessory programs to do a simple phylogenetic analysis. If you haven’t already, [download and install BEAST following these instructions](https://beast.community/installing).

Running BEAUti

A picture containing text

Description automatically generated

Run BEAUti by double clicking on its icon. [BEAUti](https://beast.community/beauti) is an interactive graphical application for designing your analysis and generating the control file (a BEAST XML file) which BEAST will use to run the analysis.

Once running, BEAUti will look similar irrespective of which computer system it is running on. For this tutorial, the Mac OS X version will be shown but the Linux & Windows versions will have exactly the same layout and functionality.

Loading the NEXUS file

When running, you will see a window like this:

Graphical user interface, text, application

Description automatically generated

The first thing you need to do is to load some data, in this case sarscov2\_global.fasta.

To load a NEXUS format alignment, simply select the Import Data... option from the File menu. You can also click the + button at the bottom left of the window or just drag-and-drop the file into the main window.

Once loaded, the alignment will be displayed in the main window in a table:

Graphical user interface, text, application

Description automatically generated



Each of these has settings and options and in general you should work from left to right (although not all the tabs will be relevant to all analyses). After the Partitions tab where you imported data, skip the Taxa tab for now and move straight on to the Tips tab.

Setting the dates of the taxa

As in TempEst, you need to click Parse Dates in order to assign each sequence a sampling date. Also as in TempEst, the default settings will work.

Skip the Traits tab and move on to the Sites tab.

Setting the evolutionary model

In the Sites tab you can set the model of molecular evolution for the sequence data you have loaded. Exactly which options appear depend on whether the data are nucleotides or amino acids (or other forms of data). The settings that will appear after loading the sarscov2\_global.fasta data set will be as follows:

Graphical user interface, text

Description automatically generated

The default is a simple HKY ([Hasegawa, Kishino & Yano (1985) *J Mol Evol* **22**: 160-174](https://www.ncbi.nlm.nih.gov/pubmed/3934395)) model of nucleotide evolution. We are going to use this model.

Setting the molecular clock model

In the next tab Clock we set the model of molecular clock we will use. Unlike many other phylogenetic software BEAST exclusively uses molecular clock models so that trees have a timescale (an inferred root and direction in time). The simplest model is the default ‘Strict clock’. This assumes that all branches on the tree have the same rate of evolution. The other molecular clock models relax this assumptions in various ways which later tutorials will discuss.

Graphical user interface, application

Description automatically generated

The default setting here is a Strict clock, which assumes that mutations happen at a constant rate over the entire tree. This leads to inaccurate estimates of the TMCA of the pandemic, however, so change this to Uncorrelated relaxed clock with a lognormal distribution. This assumes that each branch in the phylogeny can have a separate rate drawn from a parameterised lognormal distribution.

Graphical user interface, text, application, email

Description automatically generated

Setting the tree prior model

In this panel you can set the phylodynamic model that generates on the tree.

Graphical user interface, text, application

Description automatically generated

The Tree Prior option has many choices divided generally into ‘Coalescent’ and ‘Speciation’ models, the latter of which which includes phylodynamic birth-death models. We would like to investigate the past population dynamics of the epidemic, which we will do using the Coalescent: Bayesian SkyGrid  model. Set the Number of Parameters to 20 and the Time at last transition point to 2.5. The former is reduced to help the model coverge quickly, while the latter is an estimate of the time frame of the entire tree. Since the latest sequences in the alignment were collected in January, 2.5 years is sufficient.

Skip the Priors and Operators tabs and move on to the last tab, MCMC.

Setting the MCMC options

The last tab, MCMC, provides settings to control the actual running of BEAST. BEAST estimation uses a statistical sampling routine called Markov Chain Monte Carlo (MCMC). This works by exploring the posterior probability distribution of trees and parameters, one set at a time.

Graphical user interface, text, application, email

Description automatically generated

Firstly we have the Length of chain. This is the number of steps the MCMC will make in the chain before finishing. How long this should be depends on the size of the dataset, the complexity of the model and the quality of answer required. The default value of 10,000,000 is entirely arbitrary and should be adjusted according to the size of your dataset. For the sake of this exercise, I suggest setting it to 15,000,000.

The next options specify how often the current parameter values should be displayed on the screen and recorded in the log file. The screen output is simply for monitoring the programs progress so can be set to any value (although if set too small, the sheer quantity of information being displayed on the screen may actually slow the program down). For the log file, the value should be set relative to the total length of the chain. Sampling too often will result in very large files with little extra benefit in terms of the precision of the analysis. Sample too infrequently and the log file will not contain much information about the distributions of the parameters. You probably want to aim to store about 10,000 samples so this should be set to the chain length / 10,000, so 1500 in our case.

Generating the BEAST XML file

We are now ready to create the BEAST XML file. Select Generate XML... from the File menu (or the button at the bottom of the windo) and save the file with an appropriate name — it will offer the name you gave it in the MCMC panel and we usually end the filename with ‘.xml’ (although see the note, above, about extensions on Windows machines – you may want to give the file the extension ‘.xml.txt’).

We are now ready to run the file through BEAST.

Running BEAST

A picture containing chain

Description automatically generated

Run BEAST by double clicking on the BEAST icon in the package you downloaded.

The following dialog box will appear:

Graphical user interface, application

Description automatically generated

All you need to do is to click the Choose File... button, select the XML file you created in BEAUti, above, and press Run. For information about the other options see the page on the [BEAST program](https://beast.community/beast).

When you press Run BEAST will load the XML file, setup the analysis and then run it with no further interaction. In the output window you will see lots of information appearing. It prints up various pieces of information that is useful for keeping track of what is happening. The first column is the ‘state’ number — in this case it is incrementing by 1000 so between each of these lines it has made 1000 operations. The screen log shows only a few of the metrics and parameters but it is also recording a log file to disk with all of the results in it (along with a ‘.trees’ file containing the sampled trees for these states).

Wait until the 10,000,000 states have passed. Congratulations! You have finished your first BEAST run.

# Analysing the results

## Summarizing parameter estimates

Loading the .log files, containing parameter samples from all the parameters in our analysis with exception of the trees and their corresponding branch lengths, into Tracer yields the following screen:

Graphical user interface

Description automatically generated

The ESS values in this display describe how well the MCMC procedure worked. If the values of these for any numerical parameters are less than 200, and most definitely if they are less than 100, then that indicates that the MCMC run was not long enough and the analysis should be re-run with a longer length of chain.

Note also the Burn-in entry. The first steps in any MCMC run should be discarded as the algorithm has not yet reached the desired posterior distribution. You can observe its progress by examining the joint trace with different burn-ins:

Graphical user interface, application

Description automatically generated

The run is “burnt in” one the line stops ascending and starts to oscillate around a given value. You should adjust the burn-in setting such that the ascent is entirely excluded (greyed-out). Doing this can help with ESS values.

We can use the Tracer outputs to examine the posterior distribution of parameters of interest and other features of the reconstruction. For example, the age(root) entry is the age of the tree root, i.e the date of the MRCA of the entire sample. You can see its distribution by clicking its name and then on Marginal Prob Distribution. What is the means of this distribution. What does this imply about the pandemic? What

Graphical user interface, application

Description automatically generated

More detailed Tracer tutorials are available, focusing on [analysing BEAST output](https://beast.community/analysing_beast_output) and [assessing convergence](https://beast.community/tracer_convergence).

## Reconstructing the skygrid plot

## Tracer can also be used to reconstruct past population dynamics from a coalescent model such as the Bayesian skygrid used here. To do this, choose SkyGrid Reconstruction from the Analysis menu.

## Graphical user interface, application Description automatically generated

## The age of the youngest tip (the last sampling date in the dataset) is the 9th January 2022, which is 2022.02 as a decimal date. Enter that in the last box and click OK*. Chart, line chart Description automatically generated*

The middle line here is the estimated posterior median effective population size of the epidemic as it progresses over time. The exact numerical values of this parameter are hard to interpret, but when do they peaks and troughs appear? What was happening in the pandemic at those times?

## Building an MCC tree

It’s simply not feasible to inspect every tree that was visited during the BEAST analysis, hence we will create a consensus tree summarizing the posterior tree distribution. Within the BEAST package, this is done by constructing a maximum clade credibilty (MCC) tree using the program [TreeAnnotator](https://beast.community/treeannotator).

Upon running [TreeAnnotator](https://beast.community/treeannotator), you will be presented with the following window:

Graphical user interface, application

Description automatically generated

You can set the burnin (in states) to the value usied in Tracer For the Input Tree File, the sarscov2\_global.trees file that generated during the BEAST run needs to be selected. For the Output File, enter any file name you want, with the .nexus extension.

After all the required settings have been entered, you can start constructing the MCC tree by clicking Run.

## Visualising the MCC tree in FigTree

Once the MCC tree has been generated, the final step is to load it [FigTree](https://beast.community/figtree), which allows to visualise the tree and accompanying summary information produced by [TreeAnnotator](https://beast.community/treeannotator). After starting [FigTree](https://beast.community/figtree), simply go to File and Open... the apes.mcc.tree file. Typically, nodes from the tree are drawn in increasing node order (go to Tree and select Increasing Node Order to do so).

Diagram, schematic

Description automatically generated

Sometimes (as in this screenshot) you may see backwards branches in the MCC tree. This is a consequence of how the posterior tree set is summarised, and is not particularly serious.

# Investigating introductions of the Delta variant to the UK

Now let’s set up an analysis with ancestral state reconstruction in order to look at geographical spread of the SARS-CoV-2 virus. Construct a BEAST analysis as before, using the sarscov2\_intro.fasta alignment this time. Also, this time we want to add location as a discrete state in the **Traits** tab in BEAUTi

# Graphical user interface, text, application Description automatically generated

Add a new discrete trait and give it a suitable name (e.g. “Country”). You can use the **Guess trait values**button to get this information from the tip labels; the country is the second-to-last block with underscore as a delimiter. Also select **Asymmetric substitution model**in the Sites tab under your new trait – this relaxes the assumption that migration rates are the same in both directions.

Graphical user interface, text, application

Description automatically generated

You can set up and run the rest of BEAST as before. Note that this is a bigger alignment and it will take longer to run.

When it is finished, generate the MCC tree and view it in FigTree. You can display ancestral state countries by selecting your trait under **Node Labels**. What does this suggest about the sources of UK lineages of this variant?

# Investigating introductions of the Delta variant to the UK

The file sarscov2\_delta.fasta contains just sequences sampled in the first, exponential, phase of the Delta outbreak in the UK. Repeat the BEAST analysis using an exponential growth tree prior rather than the skygrid. What is the estimated exponential growth rate of the virus during this time?