

## **Objectives**

Using user-friendly software learn how to perform different kind of phylogenetic-related tasks:

- Create tree images with annotations (iTOL, Microreact).
- Infer distance-based phylogenetic tree from a distance matrix (FastME)
- Infer ancestral characters (PASTML)
- Perform hierarchical clustering based on cgMLST profiles (BIGSdb)

## Exercise 1

In 2016, cases of *Bacillus cereus* fatal infection in two premature newborns were observed in the neonatal service of a French hospital<sup>1</sup>. Supplies of breastmilk being suspected as the source of the contamination, the genome sequences of the two involved isolates (here labelled Bxxx) were sequenced and assembled, as well as those of several strains gathered from the milk bank (here labelled Byyy).

To quickly assess whether the different strains are closely related or not we can use phylogenetic inferences. The pairwise  $p$ -distance was estimated between each pair of genomes with the program Mash<sup>2</sup>. This distance is the simplest one: it is the number of observed differences between two genomes divided by the genome length. So, it is a proportion, between 0 and 1.

The dataset was completed with 103 public genomes that are representative of different species of the *Bacillus cereus* group: *B. anthracis* (Bant), *B. cereus* (Bcer), *B. cytotoxicus* (Bcyt), *B. mycoides* (Bmyc), *B. thuringiensis* (Bthu), *B. toyonensis* (Btoy), *B. weihenstephanensis* (Bwei) and *B. wiedmannii* (Bwie).

First, we will infer a phylogenetic tree using the Minimum Evolution criterion.

- Download the file B.cereus.d that contains the square matrix of estimated  $p$ -distances.
- Use FastME to infer a phylogenetic tree. FastME can be used online using the web server available at: [www.atgc-montpellier.fr/fastme/](http://www.atgc-montpellier.fr/fastme/)
  - Upload the file B.cereus.d and set the data type to Distance matrix
  - Use both NNI\_BalME and SPR\_BalME and run the analysis.
  - When the analysis is done, download all the result files.
- Read the file b\_cereus\_d\_fastme-stats.txt. Are topological improvements (NNI and SPR) useful?
- One of the two types of topological moves induced a better tree. Which one? What could explain this difference between the two methods?

---

<sup>1</sup> <https://www.aphp.fr/contenu/point-detape-sur-la-suspension-par-mesure-de-precaution-de-la-delivrance-de-lait-provenant>

<sup>2</sup> <https://doi.org/10.1186/s13059-016-0997-x>

Now, we will display and edit the tree to make it easier to interpret.

- Use iTOL (<https://itol.embl.de/>) to open the tree. At the top of the home page, click on Upload. In the Tree file box click on the Browse button and select the previously downloaded tree file (b\_cereus\_d\_fastme-tree.nwk).
- Once the tree is displayed, reroot it by using the *B. cytotoxicus* taxon as an outgroup<sup>3</sup>:
  - Find the B.cyt leaf. You can use the search function (magnifying glass icon on the left).
  - Click on the branch connecting the leaf. In the popup menu, find the Tree structure item at the bottom, and click on Re-root the tree here.
- In the control panel on the right, open the Advanced tab. At the bottom, find the Other functions menu, and click on Label functions: Multi-style.
- In front of Label part 1, change the color to red. At the bottom, type Byyy (respect the case) in the Include only box. Then click Update labels and close the menu. The labels of the Byyy strains should appear in red.
- Repeat the process for the Bxxx strains, with a different color.
- Are the two strains Bxxx closely related? Are they closely related to the Byyy ones? Are the breastmilk supplies the source of contamination?

Finally, for the last question we will need to read easily the branch lengths. For this, we will switch to the rectangular tree mode.

- In the Basic tab of the control panel, select the Rectangular mode at the top.
- In the Advanced tab, in the Branch metadata display section, click on the Display button next to the Branch lengths item.
- Knowing that the genome of *Bacillus thuringiensis* is roughly 5.5Mb long, what is the number of SNP between the strains Bthu.HD12 and Bthu.HD-771? Justify your answer.

---

<sup>3</sup> <https://doi.org/10.1038/srep14082>

## Exercise 2

Using a phylogenetic tree, it is possible to describe groups of strains/species/etc. and thus perform a clustering. But there are other and more efficient ways to cluster taxa. A phylogenetic inference can be a costly procedure and might not always be necessary depending on the question asked.

In this exercise, we will compare the results of a phylogenetic analysis and of a hierarchical clustering on the same dataset of *Staphylococcus epidermidis* genomes to decide which tool is the best to assign an unknown strain to a clade.

We will use the *Staphylococcus epidermidis* database located at <https://bigsdbs.pasteur.fr/epidermidis>

First, we will perform a hierarchical clustering based on cgMLST profiles to try to define groups of isolates.

- Go to <https://bigsdbs.pasteur.fr/epidermidis> and click on Isolates & genomes database.
- Click on the + button next to the SEARCH menu and then click on Search database.
- Search for isolates whose id is smaller or equal to 20. Then at the bottom search for the ReporTree plugin and launch it.
- In the next page, in the Schemes section, select the cgMLST.
- In the Options section select HC (Hierarchical Clustering) and click on SUBMIT. The analysis should take a few seconds to run.
- When the job is finished, click on the single\_HC button. A text representing a tree will be displayed. Select this text and copy it.
- Go to iTOL (<https://itol.embl.de/>) and go to the upload page. Paste the text in the Tree text box and click on Upload. Keep this page opened.

We will then compare this tree to a proper phylogenetic tree. Note that different kind of BIGSdb plugins allow performing limited phylogenetic analyses (iTOL, GrapeTree, Microreact, etc.) but they are distance-based methods. To perform ML analyses, we need to use command-line interface software e.g., RAxML-NG<sup>4</sup>.

- Download the file STEP.raxml.nwk. This is a ML tree inferred from a supermatrix (concatenation of genes alignments) of all genes from the cgMLST scheme for the 20 isolates. Upload this tree in a new iTOL window.

---

<sup>4</sup> <https://github.com/amkozlov/raxml-ng>

- Download the file STEP.itol.txt. This file can be used to annotate the tree. Read it (this is a text file) to understand its format. Then drag and drop the file from the file explorer onto the tree.
- This dataset does not contain any legend information. In the Control panel on the right, go to the Datasets tab and at the bottom click on Legend. In the popup window, click on Automatic legend and then on Update dataset legend.
- Root the tree using Clade B as an outgroup.
- Repeat these steps (dropping the annotation file and rooting with Clade B) for the HC tree.
- Compare the ML tree and the HC tree. Remember that they have been obtained using completely different procedures (ML phylogenetic inference from concatenation of alignments of cgMLST alleles on one side, hierarchical clustering of the cgMLST profiles on the other side). Is the classification of the isolates different?
- If you were given the task to assign a clade to an unknown isolate, what method would you chose? Why?

### Exercise 3

A phylogenetic tree is rarely a result on its own. It is often used for further analyses, e.g., ancestral character reconstruction.

In 2021, Alexandra Moura and her colleagues of the French Listeria National Reference Center published an analysis of the most prevalent clonal group of *Listeria monocytogenes* associated with human listeriosis (CC1)<sup>5</sup>. One of their findings is that this group originated from North America.

For this exercise, we will collect the country of origin of almost 273 CC1 *L. monocytogenes* isolates and using phylogenetic tools we will try to infer the ancestral origin of this group.

- From <https://bigsdbs.pasteur.fr/listeria> different information from isolates can be exported. Here we gathered the BIGSdb id and country of origin of 273 isolates of the CC1 group. Download this file (LMO.id.country.xlsx).
- Download the tree LMO.raxml.nwk which have been inferred similarly to the *S. epidermidis* tree of the previous exercise.

Creating an iTOL dataset from the Excel file can be difficult for some people. An alternative is to use Microreact. Go to <https://microreact.org/>

- At the top of the home page, click on UPLOAD.
- On the next page, drop both the tree and the Excel file at the same time. Select id as the ID column and click CONTINUE (twice).
- You get a tree with colors representing countries, and you can export this in different formats. Note that if we had the ISO 3166 Codes for the countries (or the geographic coordinates) we could create a map.

Now we will infer the most probable country of origin of the ancestral isolate. For this, we will use the software PASTML<sup>6</sup>:

- First open the Excel file and save it as a txt file (tab separator).
- Go to <https://pastml.pasteur.fr/> and click on RUN PASTML.
- Upload the Tree and the txt file.
- Chose MAP as the prediction method and click on Reconstruct ancestral states.

---

<sup>5</sup> <https://www.science.org/doi/10.1126/sciadv.abj9805>

<sup>6</sup> <https://academic.oup.com/mbe/article/36/9/2069/5498561>

- When this is done you obtain an interactive view of all the different chronological series of event that happened. You can also display this as a traditional phylogenetic tree by clicking the [here](#) link above the image.
- What is the origin of the CC1 *L. monocytogenes* isolates according to this analysis? Is this coherent with the results of Moura *et al.*?