PHYG – TMEG

2019-2020

Yasser MOHSENI
yasser.mohseni_behbahani@upmc.fr

12 December 2019

General rules

- Reports must be sent by e-mail, by the 19th of December, using the subject "`[PHYG] TME6`", mentioning the names of the students who worked on it.
- Multiple files should be grouped in a compressed archive (`.tar.gz` or `.zip`)
- Your report *must be* in PDF format and named `student1_student2_TME6.pdf`. It should be simple, clear and well organized. Answers should be given in an exhaustive manner.
- All required materials can be found in the repository `https://github.com/yassermb/PHYG2019.git`

In this TME you will find the following tools:

- the archive `mummer-4.0.0beta2.tar.gz` contains the MUMmer package. We will use it in order to compare complete genomes. In order to build the executables, run `configure` script and then `make` command from the extracted `mummer-4.0.0beta2` folder. In particular, we will consider the commands `nucmer` and `mummerplot`.
- the Artemis Comparison Tool (ACT) is available through the `ACT.jar` package and can be run using the command `java -jar ACT.jar`. It is useful for comparing sequences/genomes.
- the archive `clustalx-2.1.tar.gz` contains the command `clustalx` which is a tool for manipulating multiple sequence alignments. In particular, it allows you to visualize, build and edit them. It also allows to build a N-J tree using bootstrapping. We will perform all these operations on a small group of sequences.

Exercise 1 – Whole Genome Alignments

Consider the `genomes` folder. In this exercise you will compare different strains of the following two organisms: *Helicobacter pylori* (26695 and J99) and *Porphyromonas gingivalis* (ATCC33277, HG66, and 381).

Use the `nucmer` command in order to compare two sequences (use `--prefix` to give a name to the output). It will generate a `.delta` file which should be passed as input for the `mummerplot` command in order to draw a dotplot of the whole genome alignment. You can use the `--postscript --prefix <output>.ps` to generate the output in PostScript format (or use `--png` instead of `--postscript` to generate a PNG image).

1. Compare the two strains of *Helicobacter pylori* and include the dotplot to your report. What kind of rearrangements do you notice? Describe two of them. Then, obtain the total number of relocations, insertions and deletion in each sequence (with respect to the other) using the `dnadiff` command.

2. Consider *Porphyromonas gingivalis* and use the sequence of strain 33277 as reference. Compare the strains HG66 and 381 against it. For each of the two comparisons answers the same questions of the previous point.

Exercise 2 – Comparative Genomics

The Artemis Comparison Tool accepts in input two sequence files and a comparison file (*e.g.*, a formatted whole genome alignment). It allows you to easily visualize an alignment between two sequences/genomes. The visualization can be further enriched with an annotation file. Here there are some quick tips related to the usage of the tool:

- First, right-click on the top and bottom sequence in order to hide *Stop codons* (this will make the program run faster).

- **Aligned blocks.** Red blocks depict same-strand aligned regions, while blue blocks are, instead, different-strand alignments.

- **Sliders.** The top and bottom sliders allow you to move to different regions with respect to one of the sequence (right-click in the central part of the window and unlock the sequences to do that). A double-click on a particular block centers the view with respect to it. Top and bottom sliders on the right-hand side allow you to zoom, while the middle one filters blocks with respect to their lengths. You can also apply a percentage identity threshold.

- **Annotations.** You can add annotated regions to a specific sequence from *File* → *SequenceName* → *Read Entry.*

In this exercise you will compare more in detail some related genomes (or parts of them). You will be looking at the degree of conservation of the gene order (*i.e.*, synteny) and possibly identify genes in one of the two species considered. All needed files are found in the `comparative` folder.

1. Take a couple of minutes to explore the user interface (see for example the operations presented above) using the files

   - `Styphi.dna.gz`: sequence of *Salmonella Typhi*
   - `Ecoli.dna.gz`: sequence of *Escherichia coli*, strain K12
   - `Styphi_Ecoli.comp.gz`: comparison file

   If you take a global look at the two genomes, do you notice something interesting? (Hint: play a bit with the thresholds of the matches).

2. You will compare a genomic DNA fragment of *Plasmodium knowlesi* and the annotated chromosome 13 of *Plasmodium falciparum.* For this point you will need to load the following files:

   - `Pfal_chr13.embl.gz`: *P. falciparum* annotation file (with sequence);
   - `Pknowlesi_contig.seq.gz`: *P. knowlesi* sequence file (without annotation);
   - `Pknowlesi_contig.embl.gz`: *P. knowlesi* annotation file (without sequence);
   - `Plasmodium.comp.gz`: comparison file between *P. falciparum* and *P. knowlesi.*

3. First, use the slider to obtain a global view of the sequence comparison and remove short similarity hits. What effects does this have? Can you see conserved gene order between the two species? If yes, include export a picture which shows it.

4. Can you see any region in *P. falciparum* where similarity is broken up? Zoom in and retrieve the name of the genes involved in this discrepancy and include a picture of the region.

5. Can you identify genes in conserved regions that have not been annotated in the *P. knowlesi* contig which, on the contrary, are present in *P. falciparum*'s sequence?

6. **(Optional)** Find the *P. falciparum* gene `PFM1010w` which corresponds to the predicted gene named `Phat4` in *P. knowlesi* (*Goto* → *Pfal_chr13.embl.gz* → *Navigator*). Compare the two gene models and identify the conserved exon(s) between the two species (include a figure which shows it). Then, from the *Graph* menu, enable the visualization of the GC-content in both sequences. Can you relate the GC-content with the exons of the annotated genes?

7. You will now compare the parasites *Trypanosoma brucei* and *Leishmania major*. You will need to load the following files:

   - `Leish.embl.gz`: *L. major* annotation file (with sequence);
   - `Tbrucei.embl.gz`: *T. brucei* annotation file (with sequence);
   - `Leish_Tbrucei.comp.gz`: comparison file between *L. major* and *T. brucei*.

   Can you see conserved gene order between the two species? Can you see any region where similarity is broken up? Zoom in and look at some of the genes encoded within these regions. Can you identify any genes in one organism that don't appear to be predicted in the other?

Exercise 3 – Bootstrapping

In 1990 a dentist with HIV was accused of infecting some of his patients during some dental procedures. A phylogenetic analysis was used in the trial as supporting evidence. The sequences considered came from three main sources: the dentist, infected patients and a control group of sequences. Your task is to re-analyze a couple of them.

The file `bootstrap/HIV.fasta` contains 31 of the aforementioned sequences. More in detail, sequence names beginning with `HIVFLD` and `HIVFLQ` are associated to the dentist (`D`) and to the control group (`Q`), respectively. Those having prefix `HIVFLP`, instead, refer to different infected patients (`P`) and the next letter defines a *specific* patient. For instance, sequences whose names begins with `HIVFLPB` belongs to the patient `B`.

- Run `clustalx` and align all the sequences. As you will notice, sequence `HIVFLPED` is quite incomplete and, for this reason, we will remove it from the analysis (select its name, then *Edit* → *Cut Sequences*).
- Now build a N-J tree using 1000 bootstrapping trials and excluding gap positions. You can do this from the *Tree* menu.
- Draw the tree (*e.g.*, using `http://itol.embl.de/upload.cgi`) and, in order to see it more clearly, re-root it using sequence `HIVFQ77` (left-click, then *Tree structure* → *Re-root the tree here*).
- Finally, display bootstrap values (from the *Advanced* menu) and color sequences coming from the three sources (`HIVFLD`, `HIVFLQ`, and `HIVFLP`) differently.

Now answer to the following questions:

1. Include the phylogenetic tree with bootstrapping values to your report.
2. Do you think the dentist was guilty? Did he infect all the patients?
3. How confident are you? Motivate your answer.