Alignment

2024-02-21

Alignment

Task 1

Look at the following sequences:

`AGGACTAGCTA AGCACCTAGCTA AGCTA AGCACCTACCTA`

1. By eye, how would you align these 4 sequences?
2. What has happened to the third sequence? Can you think of a reason we might see this in actual sequencing?
3. Can you identify an indel (insertion/deletion)?
4. Can you identify a substitution?

Task 2

We'll be using SeaView to look at our sequences. This is pre-installed on the SSD but can also be downloaded from https://doua.prabi.fr/software/seaview

Locate the file `sequences.fasta` from the folder `RAGE-workshop-2024/4-command_line_and_phylogenetics/L-tutorial-a`

Open SeaView. Click File->Open FASTA and select the fasta file you just downloaded.

Scroll along to view the sequences.

1. Are they aligned? How can you tell?

Look at position 4946 to 4956.

2. What do you think is happening here, and how would you fix it in the alignment?
3. What effect does this have on sequence Z00861872 downstream?

Task 3

Create a new folder and copy the `sequences.fasta` from `RAGE-workshop-2024/4-command_line_and_phylogenetics/L-tut` to the new folder.

Using `cd` navigate to the new folder in the terminal. Use `ls` to check the fasta file is in the folder.

Enter the MADDOG_backup conda environment by typing `conda activate MADDOG_backup`. Note: If you are not working on a RAGE-on-SSD you'll first have to download the MADDOG repo and create the conda environment. Instructions on how to do this are found at: https://github.com/KathrynCampbell/MADDOG

Type `mafft -h`

This opens up the manual for mafft, the alignment tool we'll be using. We'll be using the basic high speed option today (input > output) but it's good to read through the manuals of these tools to check for useful options!

Write `mafft sequences.fasta > sequences_aligned.fasta`

This will take the input file (`sequences.fasta`) and align the sequences in it, producing an output file (`sequences_aligned.fasta`) with the sequences aligned.

It's a good idea to add _aligned in the output file name (and any other processes you may run!) to the file name like we've done here, so you don't overwrite the original file, and you can see how each file has been changed. This also makes it easier to go back a step if things don't work!

E.g. When processing file test.fasta, I may end up with test.fasta, test_outgroup.fasta and test_outgroup_aligned.fasta. Just from the file name, I can tell I added an outgroup and then aligned all the sequences in the file!

Task 4

When the process is complete, open the `sequences_aligned.fasta` file in SeaView.

Scroll along to look at the sequences again.

1. What has changed? How can you tell these are now aligned?
2. Look at position 4949 to 4959 now. What has changed here? Is this how you would have fixed the issue?
3. In sequence Z0828879 starting from position 2821 there is a sequence of Ns. How long is this section? What could be the cause of this?
4. Scroll along the rest of sequence Z0828879. Would you use this sequence for analysis? Why/why not?

It's really important to check through your sequences for any issues before starting more complex analysis! Any issues and large gappy regions can affect the analysis and give some unexpected and inaccurate results.