How to find identical sequences between two FASTA files

Lou LaMartina & Aurash Mohaimani
Updated June 17, 2020

Purpose

You have two FASTA files, and you are interested in what sequences they share. One is your "query" file, which usually has sequences from your samples of interest. You also have the "database" file, which is usually larger and has all possible matches. In this example, the database file contains all 16S V4-V5 sequences from a wastewater metastudy. The query file has all sequences belonging to *Flavobacteriales* from that study. After matching these, we will know which reads in the database file are from *Flavobacteriales*.

1. Put all necessary files in the same directory, and move to that directory in the terminal.

The folder **findExactFA** has two fasta files, **TimeSeries_V4V5_dada2.fasta** and **TimeSeries_Flavobacteriales_all.fasta**, and the script for the alignment tool, **findExactFA.pl**. Go to that directory. If the folder is on your Desktop, run `cd Desktop/findExactFA` in terminal. Make sure your computer has perl installed. You can find out by running `perl --version` in terminal. A bunch of stuff will come up if you have it; if you don't, it will say it does not recognize that command.

2. Determine your reference file.

In this case, it is the **TimeSeries_V4V5_dada2.fasta file**. This has all sequences from the wastewater metastudy that was processed in dada2.

3. Determine your query file.

TimeSeries_Flavobacteriales_all.fasta has all sequences from wastewater that were classified as *Flavobacteriales*.

4. Make the findExactFA tool executable.

At this point, the script is just a text file. You need to turn it into an actual tool, so your computer knows to run it when you ask it to. You do this with `chmod` ("change mode" but pronounced like "shmod"), which is just used to modify file types and control accessibility. The `u` flag means user (you), `+` means we are allowing you to use it, and `x` means we are making it an executable. So you are giving yourself permission to execute it. Run this in terminal:

```
chmod u+x findExactFA.pl
```

5. Run the tool!

The usage goes like this: `./findExactFA.pl ref_FASTA query_FASTA`, separated by spaces. At the end, add a file name and a >, which tells it to output the results into that file.

```
./findExactFA.pl TimeSeries_V4V5_dada2.fasta
TimeSeries_Flavobacteriales_all.fasta > test_align.txt
```

6. Analyze the results.

The first few lines of the output file look like this:

```
Sewage_Weeksellaceae_Cloacibacterium_ASV8 => ASV8;size=573170;
Sewage_Flavobacteriaceae_Flavobacterium_ASV11 => ASV11;size=486902;
Sewage_Weeksellaceae_Cloacibacterium_ASV32 => ASV32;size=146486;
Sewage_Flavobacteriaceae_Flavobacterium_ASV42 => ASV42;size=114606;
```

Left of the arrow has sequences from the query file; to the right has reference file sequences. These are exact matches. From here you can do whatever you like. For example, I copy & pasted this into excel, removed the arrow, and converted it to a CSV, so in R I had a table with two columns to easily work with.

Good luck!