Lab 03:

Introduction to GenBank, BLAST, and FASTA files;
sequence analysis and alignment

Updated by Ixchel González-Ramírez

1 Before you begin

1.1 Software needed

Please download and install the following software:

1. MAFFT: `http://mafft.cbrc.jp/alignment/software/`
2. AliView: `http://www.ormbunkar.se/aliview/`

Note for mac users: You will probably need an extra step for MAFFT software to run. Try to open it and you will receive a **”cannot be opened because it is from an unidentified developer”**. Clic the **?** sign and follow the instructions. You will be required to go to the General pan and type your password.

2 Introduction

Today we will examine tools that are useful for obtaining and preparing molecular sequence data for phylogenetic analysis. **GenBank** is the NIH (national Institutes of Health) sequence database. It contains sequence data for over 100,000 species, including over 388 billion nucleotide bases in more than 215 million sequences (December 2019 data release). **BLAST** is one of the most useful tools for working with molecular data; it allows a user to compare a query sequence against a database of sequences. Using BLAST, we will download sequences from GenBank in both **FASTA** and GenBank formats and align the sequences using two different alignment algorithms.

3 NCBI Databases

The National Center for Biotechnology Information (NCBI) is the branch of the NIH that houses GenBank. We’ll take a quick look at two of NCBI’s databases: the **Taxonomy** database and GenBank’s **Nucleotide** database. Note that there are many, many other resources!

Go to `http://www.ncbi.nlm.nih.gov/taxonomy`. Look up your favorite taxon. Also skim over the NCBI Taxonomy Handbook `http://www.ncbi.nlm.nih.gov/books/NBK21100/`.

Question 1:
What is the *Taxonomy ID* of your taxon? How many *Nucleotide* records are there for the taxon (see the box on the right side of the screen)? Explain the disclaimer at the bottom of the page. How is this taxonomy built? What kind of classification system is used by NCBI Taxonomy?

Back on the NCBI Taxonomy page of your favorite taxon, click on the link to the **Nucleotide** records. You'll now see a list of all the nucleotide sequences for your taxon. Each sequence is listed by its accession number, and information about the taxon, gene, etc. is also provided.

Follow the link for one of the sequences you've found. A new page with various information about the authors of the sequence, the taxon, gene, where it was published, etc. will appear. At the bottom of the page you will find the sequence itself. Near the top of the screen, you can see that there are several options for displaying and saving the sequence. Check out some of the display options (choose them from the pull-down menu and then push apply), but don't bother saving anything for now. If you're looking for sequences by a particular author or a particular gene, you can also type in those or any combination of them and do a search. Pick a sequence that you think would be a good one to use in a phylogenetic analysis of your group (e.g., a sequence that looks like it has been sequenced in many of the relevant species, that is conserved, named, etc.). Figure out how to download the sequence as both a FASTA file and a GenBank formatted file on your computer. Open the two files in a plain text editor.

Question 2:
When might you want to use the full GenBank format instead of a FASTA file? Think about what extra information is stored in the GenBank file compared to the FASTA file.

4 BLAST

Now we'll try a BLAST search on the sequence you just found. BLAST searches are useful for finding sequences similar to one you have generated or found. The BLAST algorithm is a less accurate but much faster approximation of the **Smith-Waterman algorithm**: `https://en.wikipedia.org/wiki/Smith%E2%80%93Waterman_algorithm`. Open the BLAST homepage in a new window `http://blast.ncbi.nlm.nih.gov/Blast.cgi`, and then click on the *nucleotide blast*. This is the option for searching for nucleotide sequences with a nucleotide sequence, but other options (such as searching for translated sequences, searching within the human genome, or searching for really close matches quickly) are available.

Now copy the sequence you found in GenBank, go back to the BLAST site and paste it into the search box. Pick an appropriate database to search.

Question 3:
What does BLAST stand for? What decision are you making by searching for sequences with the BLAST algorithm instead of some other algorithm. (hint: consider what the *LA* means, and see `https://en.wikipedia.org/wiki/Sequence_alignment#Global_and_local_alignments`). What's the default database? What database did you decide was appropriate to search?

When you've done that push the BLAST button. The search may take a couple of minutes, so be patient. Once the search is done, you can check out which sequences were found that generated significant alignments with your query sequence by scrolling down the page. You can also see the alignments with these sequences that the BLAST algorithm generated as well. There is a graphical representation (near the top of the results page) that shows where the various hits could be aligned with the query sequence and how good that alignment is. How many hits did you get? Did the taxa that 'should' have been the closest phylogenetic relatives, based on taxonomy, all come up as the closest matches to your sequence? If not, what are some possible reasons why not?

Question 4:

(a) What does *e-value* stand for? (look it up online if necessary)

(b) What does that value mean?

(c) What is a good e-value, and what is a bad e-value?

5 Sequence Alignment

In this section we will align a group of sequences. Remember that the process of alignment for molecular data is similar to the process of defining character states, we are establishing homology hypotheses. Go back to GenBank and search in *Nucleotide* for a taxonomic group that interests you. Select a manageable number of sequences (say, 5 to 20) of a group whose phylogeny phylogeny you want to infer. Make sure you only download data for the same gene region (eg. ITS, rbcL, 18S, COI, etc.). Again, keep the number of taxa reasonable (5 to 20). Pick FASTA from the display menu and then *file* from the *Send* menu. Save the file to your computer and rename it. (eg. my_sequences.fasta) Now that we have our sequences, we can do some aligning. We will practice making a manual aligning in AliView and using an algoritm in the program **MAFFT**.

5.1 AliView – Manual aligning

There are many different alignment viewer and editor tools, but I like **AliView** because it is fast and not bloated with too many extra functions (like Mesquite or MEGA). Using AliView open up your FASTA file. Use the `-` and the `+` buttons to zoom in and out of the sequences. Take a look at the options under the *Edit* menu such as *Reverse Complement Selected Sequences*. Click on the *Edit* menu on the top of the page and select the edit mode. Now you are able to insert spaces or *gaps* to match the nucleotides in vertical lines. Spend some time aligning your sequences.

5.2 Alignment with MAFFT (Multiple Alignment using Fast Fourier Transform)

MAFFT uses a technique called progressive alignment construction. Read about this here: `https://en.wikipedia.org/wiki/Multiple_sequence_alignment#Progressive_alignment_construction`.

- If you have a Windows machine you can use the MAFFT online alignment resource: `http://mafft.cbrc.jp/alignment/server/`.
- If you have Mac or linux open up a terminal window and navigate to the directory that you saved your FASTA file to. Enter the following command (changing the file names as necessary):

```
mafft --auto my_sequences.fasta > mafft_alignment.fasta
```

5.3 Alignment with MUSCLE

MUSCLE uses a sets of techniques called iterative alignment construction. Read about this here: `https://en.wikipedia.org/wiki/Multiple_sequence_alignment#Iterative_methods`. Now perform the MUSCLE alignment in this webpage: `https://www.ebi.ac.uk/Tools/msa/muscle/`

Finally, compare the alignments in AliView. To highlight differences, under the *View* menu, click *Highlight Non-consensus characters*.

Question 5:
Do you notice differences between the alignments? Do you think that a manual alignment is useful? Is it feasible? Should we visualize our data in a software like AliView? Why?

Please email me the following:

1. The answers to questions 1-5.
2. A copy of your FASTA alignments (MUSCLE, MAFFT and manual). It is Ok if the manual alignment is not finished