Lab 03:
Introduction to GenBank, BLAST, and FASTA files;
sequence analysis and alignment

Updated by Will Freyman

1 Before you begin

1.1 Software needed

Please download and install the following software:

1. MUSCLE: `http://www.drive5.com/muscle/`
2. MAFFT: `http://mafft.cbrc.jp/alignment/software/`
3. AliView: `http://www.ormbunkar.se/aliview/`

2 Introduction

Today we will examine tools that are useful for obtaining and preparing molecular sequence data for phylogenetic analysis. **GenBank** is the NIH sequence database. It contains sequence data for over 100,000 species, including over 150 billion nucleotide bases in more than 162 million sequences. **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 scree)? 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:

1. What does *e-value* stand for? (look it up online if necessary)
2. What does that value mean?
3. What is a good e-value, and what is a bad e-value?
4. Does an e-value represent Manhattan distance or Euclidean distance between two sequences?

5 Sequence Alignment

Now use a manageable number of your own sequences (say, 5 to 100), OR go back to the main GenBank web page, and search in *Nucleotide* for a taxonomic group that interests you. Make sure you only download data for the same gene region (eg. 18S, COI, etc.). Again, keep the number of taxa reasonable (5 to 100). 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)

5.1 AliView

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 alignment. Take a look at the options under the *Edit* menu such as *Reverse Complement Selected Sequences*.

Now that we have our sequences, we can do some aligning. We will practice using two alignment programs, **MUSCLE** and **MAFFT**, using the command line (for Mac OSX and Linux machines). Knowing the basics of using the command line is essential when using various bioinformatic software tools or writing your own scripts. If you have a Windows machine you can look at the alignment programs websites' for help, or use various online alignment resources such as `http://www.ebi.ac.uk/Tools/msa/muscle/` and `http://mafft.cbrc.jp/alignment/server/`.

5.2 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`. 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 MUSCLE (multiple sequence alignment by log-expectation)

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 with the following command:

```
muscle -in my_sequences.fasta -out muscle_alignment.fasta
```

Compare the two alignments in AliView. To highlight differences, try clicking *Highlight Non-consensus characters.* under the *View* menu. If you want, you can set up AliView to use MUSCLE or MAFFT for aligning under the *Align* menu.

Question 5:
Do you see any differences between the two alignments? Both methods use distance-based *guide trees* to build alignments. Do you think using guide trees to build alignments, which are then used to build trees, is circular logic? Explain.

Please email me the following:

1. The answers to questions 1-5.
2. A copy of your two FASTA alignments (MUSCLE and MAFFT).