Lab 03: GenBank and Sequence Alignment

Introduction

Today we will examine three tools that are useful for obtaining and preparing molecular sequence data for phylogenetic analysis. GenBank is the NIH sequence database. It is can be accessed, searched, etc. on the Internet and contains sequence data for over 100,000 species. Jalview is a multiple alignment editor written in Java, so it will work on PCs or Macs. It was developed by Andrew Waterhouse, Jim Procter, David Martin, and Geoff Barton and is freely available on the web (http://www.jalview.org/). A number of other alignment programs are available both for free and at a cost, and for all major computer platforms. And the third tool is your good old friend Mesquite.

GenBank

First, we'll try out some of GenBank's functions. Open your web browser on your computer and go to the site http://www.ncbi.nlm.nih.gov/. This is the National Center for Biotechnology Information website, where GenBank (among many other things) resides.

It is important for you to be aware of the diversity of resources available at NCBI. Under "Popular Resources," you will see the following:

Popular Resources

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

Question 1: **Click on each of these, figure out what they basically do, and write a (short!) description of each.**

Note that there are many, many other resources!

Nucleotide database

Select 'Nucleotide', GenBank's DNA/RNA sequence database. Try typing in the name of your favorite taxon in the search bar to see if there are sequences for it. Once you've done this, a list of sequences will appear (if there are sequences for your organism). You can also search for by taxon using the 'Taxonomy' database. Type the name of e.g. a family. To see all of the nucleotide sequences available for your taxon, check the "Nucleotide" box, then "Display."

Question 2: **What does NCBI say about its taxonomy data at the bottom of your list of found taxa?**

Click on the number next to your taxon of interest. This displays the matching sequences. 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. Feel free to try this if you like.

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 save it to a FASTA file on your hard drive.

Also, just cut-and-paste the sequence from the webpage to a text file.

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. 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. (Hint: the numbers get ignored, just the letters are read.)

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 what other options there might be among search algorithms.**

Question 4: **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.

Question 5: **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 6:

(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?
(d) Does an e-value represent Manhattan distance or Euclidean distance between two sequences?

Finally, click on the 'Taxonomy Reports' link towards the top of the page if you want to learn more about the organisms that matched the query sequence. Obviously, you can do a lot more on GenBank. Feel free to explore the site further if time allows.

Jalview

Now we'll search for and download the sequences that we'll use in Jalview.

1. Use a manageable number of sequences from your own group (say, 5 to 100)

OR

2. 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 to menu. Save the file to your desktop and rename it. (eg. Yourname_beetles.fasta)

Now that we have our sequences, we can do some aligning. The techniques we will be using in Jalview are relatively simple. The program has numerous other functions that we will not use today, but that are useful for exploring various properties of molecular data. If you are planning on including molecular data in your project, you may wish to explore these options further by using the extensive Help information included with the program.

Open Jalview. If you have downloaded Jalview for Desktop, right click the .jnlp file and select **Open With >Java Web Start**. After the program starts, three windows will appear. Close them all. Now go to the file menu and select **Input Alignment > from File**. A dialogue box will appear. Change the Format to **All Files**. Select your saved sequences (Yourname_beetles.fasta) and click **Open**. Once you've imported all of your sequences, they will appear in the alignment window and a consensus sequence will appear along the bottom. Each sequence will be identified by its accession number.

Select **Colour** > **Percent Identity**. This indicates what percentage of the residues in a column match the consensus sequence. Columns that are shaded dark purple are more than 80% conserved, columns that are purple are more than 60% conserved, columns that are light purple are more than 40% conserved, and columns that are white are less than 40% conserved.

Now Select **Colour > Nucleotide**. This colors the sequence according to the nucleotide identity. Many of the other shading options have to do with what types of Amino Acids the sequence would code for in a protein sequence alignment. You can translate a sequence using this program, but we won't get into that now.

One thing we might do is use ClustalW through Jalview to align all of our sequences automatically. Go to **Web Service > Alignment > Clustal O > with defaults**. This will align all the sequences using Clustal Omega online, which we'll deal with more later. Did this alignment change anything? [Unless you have a super conserved gene, then yes].

Question 7: **Take a screen shot showing a specific region of your sequence (eg. nucleotide positions 400–450) that shows a difference between the original import of sequences into Jalview and after the alignment using Clustal O.**

You may have noticed under the Web Service tab that there are other alignment programs that you can use through Jalview such as MAFFT, Muscle, and TCoffee. I recommend you check each of these programs to see how they affect your alignment. You should also be aware of what the default parameters are. You worked hard to get your sequence data and your alignment directly affects your phylogenetic analysis – so you should give this plenty of thought! When you are finished with your alignments, you may wish to save you work to import it into other programs (*e.g.*, PAUP*). Make sure you are in the alignment window and select **file > Save as**. Choose the fasta format and name your file (eg. aligned_Traci_beetles.fasta)

Mesquite

Now let's see if we can improve our alignment further by aligning some additional areas by hand, using Mesquite. You already had some practice with this in a previous lab.

First, open Mesquite. Now, go to **File > Open File**. Choose your aligned FASTA file. Mesquite will ask you what file format you are importing. Choose **Fasta DNA/RNA**.

Mesquite will ask you to save a new .nex file. By default, it just adds .nex to your existing file name.

Ok, now you can see your alignment. Scroll through the sequence to see if you can find an area where there are several columns in a row where things match up less than ideally. Try to find a place where adding some gaps might improve the alignment. Now go to where you want to add a gap. There are two tools you can use to move the sequence around, a hand on a green block (**Push Sequence**) and a hand with a double-headed arrow (**Move Blocks**). The sequence pusher moves an entire row at a time, including gaps. The block mover just moves one block at a time, and slides it across gaps. When you hold the block mover over the line between two nucleotides, you will see it change to a double-headed arrow with two lines across it. This can split the sequence to create new gaps.

Play around with the sequence alignment tools to see if you can improve the sequence. When you are done, you could save changes, but you don't have to.

Question 8: **Email me a NEXUS file containing some aligned sequences.**