

Lab 3: 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 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 old friend Mesquite. You can also experiment with MacClade if you have time and want to.

GenBank

First, we'll try out some of GenBank's functions. Open Internet Explorer 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 other things) resides.

At the top of the page, you will find a blank where you can begin searching. Pull down the 'All Databases' menu and 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 'Taxonomy,' and see all the sequences available for a given genus, family, etc. 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 display), 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.

Now we'll try a BLAST search on the sequence you just found. BLAST (Basic Local Alignment Search Tool) searches are useful for finding sequences similar to one you have generated or found. Open the BLAST homepage **in a new window** (<http://www.ncbi.nlm.nih.gov/BLAST/>), and then click on the 'standard nucleotide-nucleotide BLAST (blastn).' 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. When you've done that push the blast button. Once you've done that, a new page will appear telling you your query has been submitted. Push the 'format' button on this page, and it will take you to the results page. This page will be blank at first, but updates automatically until the search is finished. 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. Finally, click on the 'taxonomy report' link to 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.

Now we'll search for and download the sequences that we'll use in Jalview. Go back to the main GenBank web page, and search in 'Nucleotide' for "*emydidae feldman*" this is the taxon (turtles) and the author (Feldman, who submitted the sequences.) When the results appear, select the cytochrome b gene for *Terrapene carolina* (the accession number should be AF258871), *Emydoidea blandingii*, *Chrysemys picta*, *Clemmys guttata*, and *Clemmys marmorata*, pick 'FASTA' from the display menu and then 'file' from the send to menu. Save the file to your desktop. Name it yourname_turtles.fasta.

Jalview

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.

-First, we have to download a new version of JalView. Go to www.jalview.org/download.html, and choose “Start with Java Web Start.” Save the file to your desktop.

-Open Jalview from the desktop. 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_turtle.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.

-GeneDoc can shade the nucleotides in several different ways, showing different properties of the sequences. Pull down the colour menu and select ‘Percent Identity’ which indicates what percentage of the residues in a column match the consensus sequence. Columns that are shaded dark blue are more than 80% conserved, columns that are blue are more than 60% conserved, columns that are light blue are more than 40% conserved, and columns that are white are less than 40% conserved. As you can see, even without doing any additional aligning, these sequences have large conserved regions, which is not surprising given that these turtles are relatively closely related.

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. The most commonly used coloring for nucleotides is ‘nucleotide’. This colors the sequence according to the nucleotide identity.

-Now switch to 'Percent Identity' and scroll down through the different blocks of sequences. As you can see, the sequences generally match up pretty well for most of their length, except at the end where the *Chrysemys picta* sequence is notably different than the others. Also note that this sequence is four base pairs longer than the others. We could simply leave the sequences as they are, but we might be able to do some additional aligning to get us closer to the true phylogenetic signal.

One thing we might do is use ClustalW through Jalview to align all of our sequences automatically. Go to the 'Web Service' menu and select 'ClustalW Multiple Sequence Alignment.' This will align all the sequences using ClustalW online, which we'll deal with more on Thurs. Once it is done, be sure to check out the area near the end of the sequence, which is where most of the changes took place. As you can see, the new alignment added a few gaps, but resulted in a much closer fit between this sequence and the others.

-When you are finished with your alignments, you may wish to save your work to import it into other programs (e.g., PAUP*). Go to the 'file' menu, and select the 'save as' option. You can see that several formats for saving are available. Choose the .fasta format and name your file aligned_turtles.fasta.

Mesquite

Now let's see if we can improve our alignment further by aligning some additional areas by hand, using Mesquite. You can also use MacClade to do this if you prefer.

-First, open Mesquite. Wait for it to start up...

-Now, go to File > open file...

-Choose your aligned_turtles.fasta file from the desktop.

-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: aligned_turtles.fasta.nex

-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 (the sequence pusher) and a hand with a double-headed arrow (the block mover.) 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. There is no assignment to turn in for this lab; instead your assignment is to download some sequences for your project. It will be due in two weeks, on Feb 28th.