Lab 8: GenBank and Sequence Alignment

Introduction

Today we will examine two 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 Bartonand 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.

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). 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 and the author. 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 marmorat*, pick 'FASTA' from the display menu and then 'file' from the send to menu. Save the file to your desktop.

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.

-Open Jalview in the IB200A folder. After the program starts, a blank window will appear. This is the alignment window (or view). 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 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.

-Now let's see if we can improve our alignment further by aligning some additional areas by hand. First, 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. The 'nucleotide' colour option may be helpful for this. Now go to where you want to add a gap. Hold down the mouse and drag it across all the nucleotides that you want to move so they are highlighted. To add or remove gaps click the nucleotide you want to move, while holding down *Ctrl* to move one sequence or *Shift* to move all the selected sequences, and drag the sequences in the appropriate direction. Did this improve your percent identity? Is it realistic? How do we know weather or not we should add a gap?

-Finally, when you are finished with your alignments, you may wish to save you 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. Don't bother saving anything for now, but once you have exported the file, you can import it into MacClade or PAUP* or other programs to build a data matrix that you can then analyze.