Using Sequences from GenBank to Make Your Own Alignment

Your Mission: Download sequences from your taxa of interest (or outgroups or related taxa) from GenBank and align them, either by hand or using Clustal. Print out or email the resulting alignment by February 28.

Before you start, you need the following programs:

- Java <http://www.java.com/en/download/index.jsp>
- Mesquite <http://mesquiteproject.org/mesquite/download/download.html>
Or MacClade

1. Go to GenBank, and search the nucleotide (or protein → just change everything in this document to protein format) database for the taxon and gene of interest. It is easiest and most sensible to download one gene at a time.
2. Select the sequences you would like to include by checking the little box on the left of each blue underlined number. If you do not select any sequences, GenBank will include everything on the page.
3. From the drop down menu next to "Display," choose the fasta format.
4. From the drop-down menu labeled "Send to," choose file. Save the file somewhere convenient.
5. The next step is to align your sequences. You can do this by hand, or you can use a program to do it, which is quicker. You usually still need to take a good hard look at the alignment afterward though. The program that is most widely used is ClustalW (you can find it online at <http://www.ebi.ac.uk/clustalw/>.)
6. Clustal will allow you to upload your sequences, then download a result after a few minutes. You don't need to enter your email, but you can change the defaults if you have any reason to think your gene is particularly gappy or not. Scroll to the bottom of the page, next to "Upload a file." Click browse to find the file you saves in step 4.
7. Wait patiently while Clustal aligns your sequences.
8. When "Results of search" pops up, download the file that ends in .aln
9. Now, you will need to use an alignment program to view your alignment. You can look at it in a text editor, but this is cumbersome. MacClade is expensive, but nice and stable. It is also a bit old school, and it is only for Macs. Mesquite does all the same stuff, plus more, is free, and crashes with moderate frequency. You will probably need it anyway to translate between all the different file formats you'll be using. It is available at <http://mesquiteproject.org/mesquite/download/download.html>. Mesquite requires the Java virtual machine to run (this is actually the source of a lot of its slowness and instability, but also the reason it can easily be run on any system, Mac, Windows, Linux, etc.) The Java virtual machine is available online at <http://www.java.com/en/download/index.jsp>, and it's also free.
10. You can open this file in Mesquite; just let the program know that you are using the Clustal(DNA/RNA) format. MacClade offers the option to use this format, but it doesn't seem to work very well. Mesquite will automatically ask you to save a new Nexus file (.nex). Clustal often adds extra, uninformative blank spaces; since they are uninformative it actually

doesn't matter whether you delete these or not. Don't worry about the ! characters; these will disappear once you close and open the file.

11. You can review your alignment in Mesquite, or save it and open it in MacClade. If it won't open in MacClade, try exporting it as an "Old School Nexus File:" File > Export… > Simplified NEXUS > OK
12. One thing you will probably want to take care of sooner rather than later is to fix up the taxon names. Clustal saves the first 30 characters – generally just numbers in GenBank. The names are not in the same order as the FASTA file. It is also possible to open the .fasta file in either MacClade or Mesquite before you run it through Clustal, fix the names, then use Mesquite to export it as a .fasta file again.
13. Once you have looked at the alignment and are satisfied with it, save it as a .nex file.
14. That's it!