Created by Paul Yenerall, 2010

How to access GenBank

The purpose of this tutorial is to teach students how to access GenBank and explain to students some of the features of GenBank. After completing this tutorial students will have a greater understanding of what GenBank is, how to access it, and what GenBank displays.

For questions, corrections, or help with this tutorial, please e-mail paulyenerall@gmail.comor lzhou1@pitt.edu

**Accessing GenBank**

GenBank is the National Institute of Health’s (NIH) genetic sequence database and contains a collection of annotated DNA sequences. GenBank contains over 100 million DNA sequence records. If you’re ever looking for any type of nucleotide sequence and information about the sequence, GenBank would be a good place to look. For proteins, however, you need to access GENPEPT.

**1)** Now, let’s access a nucleotide record in GenBank. Open up an internet browser, preferably Mozilla Firefox, and go to <http://www.ncbi.nlm.nih.gov/>. The following page should be displayed:



Now, in the search box displayed at the top of the page, type in the accession number AB004506 (we will use this just as an example), as displayed below:



Then, click the dropdown box labeled “All Databases”, scroll down the list a little, and then change the database to “Nucleotide”, as illustrated below:



Then hit the search button. The following page should be displayed (next page):



Note that this record retrieved is a GenBank record. You can tell this by the “Format” section listed at the top of the page, which is highlighted in the previous picture.

**2)** Let’s take a little more in-depth look at some of the features of this page. Note the highlighted regions in the next picture (next page):



1 – This is the Accession number. You can put this into a NCBI search to retrieve this record.
2 – This is a brief description of the record we’ve retrieved. It will generally have the amount of base pairs the record is (bp), the type of sequence it is (in our case, RNA, in other cases it may be mRNA or Gene), and any other pertinent information
3 – This is the GenBank ID that can be used to link directly to this record (note: you can always use the common name or keyword to find records as well)
4 – This part is the organism the sequence originates from, as well as the lineage of the organism.
5 – This part gives us any references to work done on this sequence. This can be useful to find research articles on the function of this nucleotide sequence or what research is being conducted on this nucleotide sequence.
6 – This part is any distinguishable features of the sequence as well as the length of the sequence. For instance in our example, we see that this is the 5’ untranslated region (5’ UTR) of the sequence. Other times this part will show you exons, introns, whole genes, or sequence tagged sites (STS) just to name a few.

Scroll down the page a little more the “Origin” section of the page. This part will contain the nucleotide sequence and/or protein sequence (CDS – Coding Sequence). In our example, we only have the nucleotide sequence. If you want the nucleotide sequence in FASTA format, click the FASTA link at the top of the page, as illustrated below:



**3)** To download the record, click the download link (and then select the type of file you wish to download) as illustrated below:

