LAB 03
Jennifer Lin

Question 1:

What is the Taxonomy ID of your taxon?

Taxonomy ID: **9986** (*Oryctolagus cuniculus*)

How many Nucleotide records are there for the taxon (see the box on the right side of the screen)?

Subtree links: 2,297,144 or Direct links: 2,296,539

Explain the disclaimer at the bottom of the page.

Disclaimer: The NCBI taxonomy database is not an authoritative source for nomenclature or classification - please consult the relevant scientific literature for the most reliable information.

How is this taxonomy built? What kind of classification system is used by NCBI Taxonomy?

The taxonomy database is **manually curated** by a small group of scientists at the NCBI who **use the current taxonomic literature** to maintain a phylogenetic taxonomy for the source organisms represented in the sequence databases (Federhen, 2012)

Question 2:

When might you want to use the full GenBank format instead of a FASTA file? Think about what extra information is stored in the GenBank file compared to the FASTA file.

In the FASTA file, there is only the geneID, name, species, and the DNA sequence of the gene. The information in the FASTA file is enough for us to do the analysis on the sequence data. Yet, we have no further information about it. Nevertheless, in the full GenBank format file, we have a lot of information about that sequence. For instance, we have the date where it was published, which may be very helpful for us to determine if the data is reliable. We also have more information about the organisms. We can clearly see several other organisms have this gene as well, which may help us to compare the gene afterwards. Other than these, there are all the informations we might need to use in the future. Thus, it would be better if we download the full GenBank format file while collecting the data.

Question 3:

What does BLAST stand for?

The Basic Local Alignment Search Tool (BLAST)

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 see https://en.wikipedia.org/wiki/Sequence_alignment#Global_ and_local_alignments).

LA: Local Alignment

Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context (Polyanovsky et al., 2011). In the other hand, global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size (Polyanovsky et al., 2011).

What's the default database?

Nucleotide collection (nr/nt)

What database did you decide was appropriate to search?

I used **Reference RNA sequences (refseq_rna)** instead because the sequence I collected was a mRNA sequence.

Question 4:

(a) What does e-value stand for? (look it up online if necessary)

Expect (E) Value

(b) What does that value mean?

The BLAST E-value is the number of expected hits of similar quality (score) that could be found just by chance (http://www.metagenomics.wiki/tools/blast/evalue).

(c) What is a good e-value, and what is a bad e-value?

According to the definition, a lower e-value would be good e-value. For instance, e-value of 10 means 10 of the sequences we found might just be found by chance, not by similarity. Yet, if we had another blast search and got the e-value of 1, that means only 1 of the sequences we found might be found just by chance. Comparing these two situations, we can see that lower e-values would be better.

Question 5:

Do you notice differences between the alignments?

The conserved region, which is in the middle, of the alignments of MAFFT and MUSCLE look very similar to me. I think when aligning the conserved region, different algorithms can still get the similar results. Nonetheless, since the length of the sequences are quite different, MAFFT and MUSCLE chose to add gaps or accepted a substitution in different place. Thus, the region at the beginning or at the end look totally different given the two alignment methods.

As for manual alignment, I didn't finished all of it and thus most of the alignment looks totally different from the other two alignment tools. I only managed to adjust the beginning part of the first sequence to make it align with other sequences.

Do you think that a manual alignment is useful? Is it feasible? Should we visualize our data in a software like AliView? Why?

I think manually alignment is very helpful because I can take a look of what the data looks like and have a sense of what results should I expect. Otherwise, I think most of the time people just click those bottoms without understanding the concepts. However, manually aligning sequences is a bit difficult for me since the lengths of the sequences are quite different. Yet, I think if I start from the data that has already been aligned by other tools, then manually check the problems, and fix it, it would be much easier, efficient, and feasible for me.