**Computer Lab 3 – Sequence Alignment**

**Conservation Genetics (BIOL 4174 / 5174)**

**Part 1 – GenBank and BLAST**

GenBank provides a central database of all publicly available DNA sequences. It can be accessed at <http://www.ncbi.nlm.nih.gov/genbank/>. The search box at the top of the page allows you to search the database by any of a variety of terms. For example, various taxonomic names can be used (order, family, genus, etc.). You can also search for specific genes (e.g., “CytB”). Sequences are also stored with additional information, so other terms such as an autor or publication name can be searched. Each sequence is assigned a unique accession number, which can also be used for retrieval (e.g., GU599988).

GenBank can also be searched using BLAST. This allows you to enter one or more DNA sequences, which it will then attempt to match to the sequences stored in the database. You will now do this for one of the sequences you edited in the previous section.

**Disclaimer**: The layout of the GenBank website changes frequently. If you return to these instructions at a later date to use GenBank for your own research projects you may find that things have changed slightly.

BLAST Instructions:

1. Go to GenBank
2. Click the GenBank dropdown box at the top of the screen and select “BLAST”
3. Choose “Nucleotide Blast”
4. Copy and paste one of the 842bp sequences from Part I of Computer Lab 2 into the “Enter Query Sequence” box
   1. Note: A sequence is provided for you in: ConsGen2018/Lab3/part\_1/62S8W01.fasta
   2. Open this sequence in a text editor or use the “cat” command, and copy the nucleotide sequence only
5. Under “Choose Search Set” make sure the “Others (nr etc.)” and “Nucleotide Collection (nr/nt)” options are selected.
6. Under “Program Selection”, choose to search for highly similar sequences.
7. Click “BLAST”

After a few seconds, the browser will return your results. This provides a table with several numerical measures of similarity between your sequence and the results found in GenBank. These include:

* Alignment Score – a measure of similarity between sequences. Greater scores indicate greater similarity. This is divided into Maximum and Total Scores. Multiple parts of your query sequence may match a sequence stored in GenBank. The maximum score corresponds to the alignment of segments with the greatest similarity. The total score is the sum of all scores for aligned segments of your query sequence and a particular GenBank sequence.
* Query Coverage – the percentage of your submitted sequence that is covered by the Genbank sequence.
* E-Value – this provides a means of assessing the “significance” of a match. It represents the number of matches expected to occur by chance given the size of the database being searched.
* Identity – percentage of bases in the query coverage area that match between your submitted sequence and the GenBank sequence.

**At this point you should answer questions 1 through 4 in the homework document.**

**Dowloading Genbank Sequences**

You will now use GenBank to obtain a Python regius ATPase 6 sequence.

1. Go back to the GenBank home page. In the search box at the top of the screen enter “python regius atp6” (without quotes) and click “search.” The only result should be a complete mitochondrial genome for this species.
2. Search for the gene within the complete genome (you can do this by using the search function of your web browser, or Command+f, to find “atp6”). When you find it, click on the “gene” link to the left and this will highlight the gene you want to download.
3. At the bottom of your screen you should now click “FASTA.” Then on the next page click “Send to:” make sure “Complete Record” is selected, click “File,” make sure “FASTA” is selected in the drop down box, and click “Create File.”
4. Now save the file to your lab\_3/part\_iii folder and rename it “python.fasta.”
5. Repeat steps 1 through 4 to download an ATPase 6 sequence for *Boa constrictor*. Save this file as “boa.fasta.” You will use the boa and python files in part III of this exercise.

**Answer question 6 in the homework.**

**Part 2 – Clustal alignment**

ClustalX 2.1 can be downloaded from <http://www.clustal.org/download/current/>. It is available for Windows, Mac OS X, and Linux. A command line only version (ClustalW 2.1) is also available. Although this program is no longer being updated, it is still a popular choice among biologists for multiple sequence alignment (MSA).

In your “part\_2” folder for Lab 3, you will find a sequence file called “heloderma.fasta”. Open ClustalX by typing “clustalx” in the Terminal.

Open “heloderma.fasta” file in ClustalX 2.1 (File 🡪 Load Sequences). Before starting, make sure the program has been set to the default parameters by going to “Alignment 🡪 Set All Parameters to default.” Perform an alignment using the default settings (Alignment 🡪 Do Complete Alignment). This will bring up a window asking you to save a guide tree and an alignment file. These should be saved in the same directory where your heloderma.fasta file is located. Click OK to proceed.

Import your python.fasta file into your group of sequences (File 🡪 Append Sequences). Before proceeding, make sure that the program will reset your alignment by removing any gaps it may have inserted (Alignment 🡪 Alignment Parameters 🡪 Reset New Gaps before Alignment). This option will now remain checked until you uncheck it, reset the default parameters, or close the program. Now perform the alignment again. Inspect the resulting alignment and then repeat this process by adding your boa.fasta file as well.

**Answer questions 8 and 9 in the homework document**

Finally, try changing the default costs of inserting and extending gaps in ClustalX. Go to Alignment 🡪 Alignment Parameters 🡪 Multiple Alignment Parameters. Try at least two alternate settings. Record these settings and report if/how they changed the alignment. The settings you choose should differ from the default settings, which are Gap open: 15 and Gap extend: 6.66. Increasing the Gap Open or Gap Extend penalties will reduce the chances of inserting or extending gaps. Try with both high and low gap penalties. **Record your observations under question 10 in the homework document.**

**Part 3 – Clustal Omega**

Clustal Omega is the successor to the ClustalX and ClustalW programs. Until recently ClustalO had been capable of handling only amino acid data, but a recent update gave it capabilities for handling nucleotide data as well.

**Run Clustal Omega**

1. In yout Terminal, change directories into your Lab3/part\_3 folder
2. Run Clustal Omega using the following command:

**clustalo -i reptiles.fasta -t DNA -o reptiles.omega.fasta --outfmt=fa**

If you received no errors, the program should have worked properly and produced an output file in fasta format within the same directory where your reptiles.fasta file is located. **We will return to this file later in lab.**

Options for the program are used as follows:

**-i <file>** is used to specify an input file

**-t <type>** is used to specify data type (DNA, RNA, or Protein are acceptable inputs)

**-o <file>** is used to specify an output file

**--outfmt=fa** is used to specify fasta format as the output format. The program can output several other common file formats as well.

Clustal Omega has several additional options. If you would like to use this program for your own research, you can find these options in the online documentation located at <http://www.clustal.org/omega/README>.

**Part 4 – Muscle**

The last program we will use for MSA is Muscle v3.8.31. Curiously, this program was not written by a biologist. Rather, it was written by an independently wealthy former physicist whose software company was purchased by Intel in the 1990s. After the sale of his company he took an interest in the human genome project, became dissatisfied with the state of alignment programs, and decided to create something that was supposed to be more accurate and efficient.

**Run Muscle**

1. In your Terminal, change directories into the Lab3/part\_4 folder
2. Run Muscle on the reptiles.fasta dataset using the following command:

**muscle -in reptiles.fasta -out reptiles.muscle.fasta**

This command will save an output file in your current directory. **We will return to this file later in lab.**

**Part 5 – Manual Alignment**

**Viewing the amino acid translation in Mesquite**

You can open Mesquite by clicking the C:\Program Files (x86)\Mesquite_Folder\images\mesquiteIcon48.gif icon in Applications/Mesquite\_Folder

You will now view the translations of the alignments from Clustal Omega and Muscle. First, view the Muscle alignment by opening the “reptiles.muscle.fasta” file in Mesquite (File 🡪 Open file…). When prompted to choose an interpreter for the file, select “FASTA (DNA/RNA)” and click “Ok.” You will also be prompted to save the alignment as a Nexus (.nex) file. Accept whatever name the program chooses and click “Save.”

To view the sequences, click on “Show Matrix” in the next window. Before proceeding, we will set the reading frame of the sequences and the genetic code. Note: *Mesquite has many strange quirks, one being that the contents of the menu bar at the top of the screen will change often depending on the type of data you are viewing. For the “Matrix” menu to be present, you must be viewing your character matrix.*

Go to Matrix 🡪 Genetic Code. Click on the resulting “Characters” tab, then select all characters (“Apple Key/Command” + “A”). Set the codon position (Codon Position 🡪 Set Codon Position 🡪 123123…) and the genetic code (Genetic Code 🡪 Genetic Code 🡪 Vertebrate Mitochondrial). To view the amino acid translation of the sequences, switch back to the “Character Matrix” tab, and recolor the character matrix (Display 🡪 Color Matrix Cells 🡪 Color Nucleotide by Amino Acid).

Take a moment to scan through the alignment. After inspecting the alignment, repeat the process listed above to view your Clustal Omega alignment (i.e., use the “reptiles.omega.fasta”). **Answer the last question in the homework document, then leave your Clustal Omega alignment open for the final portion of today’s lab.**

**Conduct a manual alignment**

Alignment programs often will fail to provide a proper alignment, but will give a good starting point to work from. Use the amino acid translation to manually fix the Clustal Omega alignment. Use your knowledge of how protein-coding genes mutate to complete this task. You can use the tools in the sidebar to select and drag sequences. The C:\Users\Steve\Desktop\Dell Laptop\C drive\Program Files (x86)\Mesquite_Folder\mesquite\align\SidePusher\rightGrab.gif tool will allow you to select and drag an entire sequence. The  and C:\Users\Steve\Desktop\Dell Laptop\C drive\Program Files (x86)\Mesquite_Folder\mesquite\align\aAlignIntro\images\BlockMover.gif tools will allow you to split and drag part of a sequence.

Hints:

1. You should be able to align the sequences by slightly modifying the gap in the *Python* and *Boa* sequences that was inserted by Clustal Omega.
2. Manual alignments can be tricky, and there are often multiple answers that are “not wrong.” Don’t overthink this. If you have difficulties don’t be afraid to ask for help.
3. Note that some sequences may not be complete, meaning they won’t include the stop codon.

**When finished, save your alignment as a NEXUS file (File 🡪 Save File) and exit Mesquite.**