CMPSC 300
Bioinformatics
Spring 2021

Lab 4 Assignment:
What Is This Sequence?
Submit deliverables through your assignment GitHub repository and complete the Google Form.

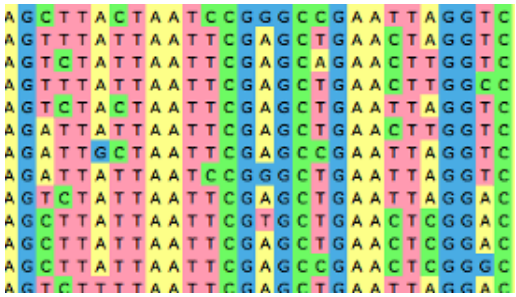

Figure 1: Multiple Sequence Alignment: When you compare each sequence in a group with each column lined-up, you can see commonalities across sequences. These commonalities inform researchers of the origins of the sequence.

Objectives

To gain some experience working with multiple sequence alignment outputs similar to that of Figure 1 to determine commonalities between sequences. These commonalities will be used to infer relatedness between unknown and known sequences. This lab also offers the student to become familiar using Clustal Omega, a tool to align sequences which is hosted by EMBL-EBI.

Clone Your Assignment Repository

In this section, we will be using `Git` commands. It is suggested that the reader refer to online searches for help. For example, GitHub provides good documentation at the following link; `https://git.github.io/htmldocs/git.html`.

In many cases, you will be given a new repository containing assignment materials and you will save your files in this assignment repository as you continue to work on them. Copy and paste the assignment repository cloning command into your terminal to create your assignment repositories. Be sure to place your assignment repositories in a directory such as `cs300/` to keep your class materials organized by class.

Today's assignment repository can be found at the below link to a GitHub Classroom repository. Here you will work on your assignment and then push your work to the cloud where the instructor will be able to view your work for grading. Often, there will be files in your assignment repositories which you are to edit before you submit them by using the below commands for `git`.

https://classroom.github.com/a/aC4fW6Py

To use this link, please follow the steps below.

- Click on the link and accept the assignment
- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab,
- Clone this repository (bearing your name) and work locally
- As you are working on your lab, you are to commit and push regularly. The commands are the following.
  - `git add -A`
  - `git commit -m ‘‘Your notes about commit here’’`
  - `git push`

Check Your Submission

After you have pushed your work to your repository, please visit the repository the GitHub website (you may have to log-in) to verify that your files were correctly sent. Importantly, please check that GitHub Actions has checked your submission. For this, look for an orange dot that will turn into a red check mark to indicate errors, or a green check on the top line of your repository to indicate that all checks have passed.

Alignment

In bioinformatics, much effort is spent comparing genetic sequences (i.e., sequences of DNA, RNA or protein) to determine information about the origins of the sequences. For example, when a patient has become sick due to a bacterial disorder, it is important to determine which bacterial organism is responsible for the ailment so that proper therapy can be administered. Unfortunately, since the specific sequence of DNA of the bacterial organism causing distress may not be known, researchers will still compare this new DNA sequence to known sequences to find resemblances from which a likely identification can be concluded.

In this lab, you are given five (*known*) organismal sequences and five (*unknown*) mutated sequences. Your task is to determine which unknown sequence is a relation of what organism, according to a multi-sequence alignment task. To determine relatedness, you will use the Clustal Omega tool from EMBL-EBI which can be found at the below link.

`https://www.ebi.ac.uk/Tools/msa/clustalo/`

As you study your results, please determine which results would be necessary to determine close relationships between the sequences. You will be taking screenshots of this irrefutable proof to place into your deliverable for the lab.

What To Do

Locate the `data/` directory in your GitHub repository. You will note that there are five DNA sequences which were derived from five organisms (i.e., coronavirus, dog, orchid, rabbit and rat). There are also five unknown sequences which come from organisms which are related to these organisms in some way. Using the Clustal Omega tool, you will be able to compare these sequences in order to determine which unknown sequence is related to what organism.

After your experiment, you are to offer provide irrefutable proof of the relatedness of each unknown sequence to a known one. In addition, you are to offer clear and meaningful language to explain what the results are and how the matching between known and unknown sequences can be understood to be *irrefutable*.

Fasta and GenBank Formatting

When you obtain genetic sequences from online databases such as NCBI (National Center for Biotechnology Information) at `https://www.ncbi.nlm.nih.gov/` as shown in Figure 2, sequences may be downloaded different formats. Depending on the requirement of the project, more information may be necessary, in addition to the genetic sequence itself. The two most commonly used formats are GenBank and FASTA. Below, we spend a moment to understand the main strengths of each format.

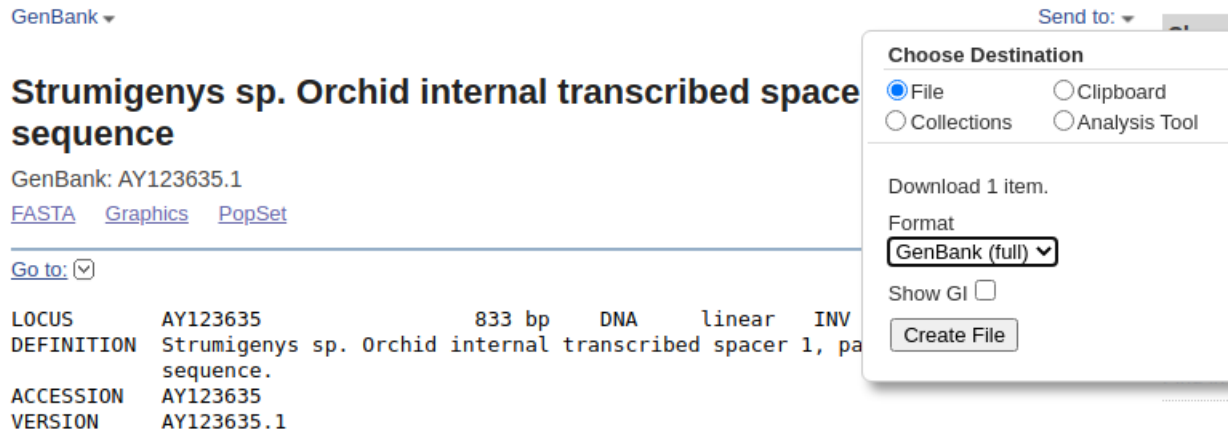

Figure 2: You can download you data for your genetic sequences in many different formats from NCBI. Here, the download file is requested in the GenBank formatting which has much additional information, besides the genetic sequence.

GenBank Formatting

In the GenBank format, for example, a record shown in Figure 2 of a sequence contains much extra information concerning the research behind the sequence. Shown below, we expand on the

information that is shown in a typical record to note that this particular format may be entirely unnecessary, if only the sequence is required.

```
LOCUS       AY123635                 833 bp    DNA     linear   INV 05-AUG-2002
DEFINITION  Strumigenys sp. Orchid internal transcribed spacer 1, partial
            sequence.
ACCESSION   AY123635
VERSION     AY123635.1
KEYWORDS    .
SOURCE      Strumigenys sp. Orchid
  ORGANISM  Strumigenys sp. Orchid
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Holometabola; Hymenoptera; Apocrita; Aculeata;
            Formicoidea; Formicidae; Myrmicinae; Strumigenys; unclassified
            Strumigenys.
REFERENCE   1  (bases 1 to 833)
  AUTHORS   Hung,Y.-T., Wu,W.-J., Lin,C.-C., Chen,C.A. and Shih,C.-J.
  TITLE     Strumigenys ITS1
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 833)
  AUTHORS   Hung,Y.-T., Wu,W.-J., Lin,C.-C., Chen,C.A. and Shih,C.-J.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-JUN-2002) Entomology, National Taiwan University, No.
            1, Sec. 4, Roosevelt Road, Taipei, Taiwan 106, Republic of China
FEATURES             Location/Qualifiers
     source          1..833
                     /organism="Strumigenys sp. Orchid"
                     /mol_type="genomic DNA"
                     /isolate="Orchid"
                     /db_xref="taxon:202929"
     misc_RNA        <1..>833
                     /product="internal transcribed spacer 1"
ORIGIN
        1 acgttccgga ggtcctgctc tgtggcctgt gttccgtccc ttggacgaac gcgcgcgcag
       ...
      781 gtagaaagga tacaccgtga gcgttctatg gaacgcgaac gaaacgatta ccc
//
```

Note: This GenBank formatting is may be found for the organism at link: `https://www.ncbi.nlm.nih.gov/nuccore/AY123635.1`.

FASTA Formatting

In the FASTA format, there is much less information present in comparison to that offered by GenBank formatting. Known for its simplicity, the FASTA format has only two lines for a genetic

sequence; a name and its corresponding sequence. The name of the organism will be the same as the one given in the GenBank record, however all other information, except for the sequence, has been omitted. Below is an example of the genetic sequence of the *Strumigenys sp. Orchid* which has been prepared in a FASTA format.

```
>AY123635.1 Strumigenys sp. Orchid internal transcribed spacer 1, partial sequence
ACGTT...ACCC
```

Note: This GenBank formatting is may be found for the organism at link: `https://www.ncbi.nlm.nih.gov/nuccore/AY123635.1?report=fasta`.

Several FASTA formatted files may be concatenated into one file (containing the names and sequence-details of an arbitrary number of organisms) which can be uploaded to many different tools for sequence analysis. For your work, try working with this concept to enter your sequence data into Clustal Omega for your multi sequence alignment task. Your completed file, containing the known and unknown sequences will resemble the general form shown below.

```
>seq_x1
AT...G
>seq_x2
GC..T
>seq_x3
TA...C
...
>AY123635.1 Strumigenys sp. Orchid internal transcribed spacer 1, partial sequence
ACGTT...ACCC
```

Ethics

Read the March 18, 2021 article from *The Atlantic*, by Zeynep Tufekci, *3 Ways the Pandemic Has Made the World Better* at link: `https://www.theatlantic.com/health/archive/2021/03/three-ways-pandemic-has-bettered-world/618320/?utm_source=pocket-newtab`.

In the article, it mentions that vaccines (such as the one for SARS-CoV-2) may now easily be manufactured in labs by working only with the spike proteins that initiate infections because, *"we instruct our cells to make only the spike portion to give our immune system practice with something that cannot infect us—the rest of the virus isn't there"*. In addition, the article mentions that, *Pfizer and Moderna are both already working on boosters that better target the new variants we've seen so far, and the FDA has said it can approve these tweaks quickly.* Both of these quotes suggest that there is research and governmental-supported safety in using vaccines.

Three questions to consider:

- **1. We Now Know How to Code for Our Vaccines**: If safe vaccines can be created to combat viruses, then, in your opinion, would providing (free) technical education about the design and creation of vaccines (i.e., vaccine transparency) be enough to help change the minds of those who oppose vaccines? Why or why not? What else might be necessary to make

an argument for vaccine safety? Note: if you are not personally convinced that vaccines are safe, please take the position that you would like to find a way to convince others of vaccine safety.

- **2. We Actually Learned How to Use Our Digital Infrastructure**: The second part of the article mentions that the pandemic has taught people how to use online tools such as Zoom to connect people online for the attendance of meetings, work, classes and other events. Tele-medicine, or using online tools to connect patients to their doctors for appointments, was mentioned as a favorable example for maintaining health in the community.

  In your opinion, will tele-medicine become the normal way of allowing patients and doctors to hold appointments after the pandemic? In your opinion, how will the in-person patient-doctor relationship be impacted as a result of online tele-medicine visits after the pandemic? For instance, could there be trust issues if the patient believes that there is a technical boundary between him or herself, and the doctor? Why or why not?

- **3. We've Unleashed the True Spirit of Peer Review and Open Science**: In the article, it is mentioned that, *Like many others, we didn't wait for formal peer review to end before sharing our findings* in scientific articles.

  In the interests of avoiding misinformation, do you think it is beneficial to circumvent the formal peer reviewing before releasing free information to the research community? In your response, discuss how misinformation (published incorrectly stated facts or misunderstood results) could be avoided in such a system where there is no peer-reviewing before publishing.

Required Deliverables

- **report.md**: Edit and complete the file, *writing/report.md* in markdown. In this file, you will offer clear and meaningful language to explain your results and argue which unknown sequence a relation of what organism. All your results are to stem from your analysis while working with EMBL-EBI's Clustal Omega. You are to include screenshots of particular results that **clearly show** your irrefutable proof to match the unknown sequences with the known ones.

- **reflections.md**: Edit and complete the file, *writing/reflections.md* in markdown. In this file, you will address the above questions-in-blue.

Grading

The grade that you receive for this lab assignment will be based on the following:

- 60% Your report document where you pair the lab's known and unknown sequences, in addition to providing your justification of this pairing using screenshots and clear and meaningful language.

- 30% Your reflection document where you address the above questions-in-blue.

- 10% Complete GitHub Actions CI build-pass corresponding to all the GatorGrader checks passing.

Please see the Technical Leaders or the instructor if you have questions about the assignment submission.