A picture containing light, blur

Description automatically generatedSherlock Biosciences Inc

Bioinformatics Internship Program

March 23, 2022

Upload final files to github by: March 30, 2022 End of Day

Please shoot any questions/troubleshooting via email to [alex@sherlock.bio](mailto:alex@sherlock.bio)

Happy to help clarify or guide!

*The purpose of this assignment is to assess your ability to efficiently and independently gather information from public sources.*

Question 1:

First, you should build a simple web scraper to access the FDA’s page on available genetic tests (<https://www.fda.gov/medical-devices/in-vitro-diagnostics/nucleic-acid-based-tests>), and pull the following information for human tests:

* Disease/Use
* Trade Name
* Submission
* URL of C
* Classification Product Code

The output should be stored in a structured file format of your choice. We prefer Python but you may use a different programming language. Your code should be legible and contain comments where necessary.

We will judge your work based on functionality, interpretability of code and edge case handling (i.e., broken submission link, etc.)

Your final submission should consist of the output data file and your code. Both should be uploaded to a github repo.

Question 2:

Purpose: Interact with NCBI API wrappers to access the ***nuccore*** database to pull genome records associated with *Human adenovirus A.*

Use the Query - **Human adenovirus A AND Human adenovirus A[Organism]** when querying the nuccore entrez API.

Write a script to pull genome records from the ncbi-nuccore database. There is a command line tool called esearch that is best for accessing records or you can use one of the a few other methods available and shown here: <https://www.ncbi.nlm.nih.gov/books/NBK25500/>

Plot a histogram of the sequence lengths pulled and attempt to identify any outliers.

Hint (Biopython or Bioperl have some of these built in functions.)