Hasan and Skyler

BNFO 420: Applications in Bioinformatics

Python Review: The Serpent Strikes Back

Due: 1/26/2017

**Deliverables:** Each group will work together as a team to solve all of the following problems below (1-4). This is more of exercise to get you into the motions of programming as a team. Search the deep recesses of your brain and awaken your inner pythonista. when in doubt **“import this”** to find your zen…

After class is over, you will responsible for completing the remaining questions (questions 1-5), and you will hand in this completed assignment as your first grade.

**Background:** Antivirals are a class of medication used to treat viral infections. They can also be used to prevent an infection when given shortly before or after exposure to a pathogen. They interfere with the life cycle of the a virus and prevent it from multiplying in a deleterious manner (thus preventing or lessening the infection). A key difference **between a vaccine and antiviral drug is that the antiviral drug will prevent an infection only when administered within a certain time frame-- either before or after exposure-- and is only effective during that timeframe**. A vaccine can be given long before exposure to a virus and can thus provide protection over a long period of time.

1. Neuraminidase is a surface protein that is commonly targeted by antivirals. The file “NA\_swine.txt” contains several neuraminidase DNA segments that were isolated from swine. Please parse the fasta file, and create a function named “revcomp()” to find the reverse complement of each sequence. Please write these sequences to a file named “reverse.txt”.

*Output:*

>revcomp\_header1

AATAGATAGATAGATCGCTAGCTTCGCTCG

>revcomp\_header2

TTCGCTGATCGTCGAGCTCGCTAGCTCTT

..

...

….

>revcomp\_header999

CAGCACGTACAGACACTTTGATAGTAAAA

1. Please parse the fasta file of transcripts (see “translateMeFasta.txt”), and create a function named “translation()” that returns translated sequences. Write those translated sequences to a file named “translated.txt” (this functionality can be embedded within your translation function or else where).

This translation function will require a dictionary. To save you some time we have given you the dictionary in “codon\_table.txt” (copy-pasta).

*Output:*

>aa\_header1

RYLGYNSNYFPWWESYLGYNSNYFPWWES

>aa\_header2

GYCARYLNSLDDYWLGCARYLNSLDDYWLG

..

...

….

>aa\_header999

FNYSSGCKIGYYYWGSANYSSGCKIGYYYW

1. “flu\_results.tsv” includes a list of commonly used antivirals for influenza A and B. Write a python script to parse through the file, and using any data structure of your choice (use a dictionary) save some valuable information of your choosing. For example, maybe the name of a drug and its mechanism of action (MOA).

*Output:*

{“Drug Name” : [“Drugbank\_ID”, “Description”, “Indication\_Targeted\_Virus(es)”]}

1. We want to compare the average GC-content of the hemagglutinin segment between type A influenza and type B influenza. Like neuraminidase, hemagglutinin is also a surface protein. This protein is is critical for the production of vaccines. Please create a function that takes a fasta formatted file as a parameter and returns the gc-content. You are not to use the “.count()” method. Please write this function in pure python. Calculate average GC content for each file:
   1. HA\_NorthAmerica\_A.fasta
   2. HA\_NorthAmerica\_B.fasta

Your output can look like whatever you want. Obviously the more information the better.

Example output:

Hemagglutinin from Type A

GC content: 56%

Hemagglutinin from Type B

GC content: 51%

1. Please find the dinucleotide frequencies of the two hemagglutinin segments (type A influenza and type B influenza). Please create a function that takes a fasta formatted file as a parameter and returns a dictionary of dinucleotide frequencies. Please write this function in pure python. Calculate dinucleotides frequencies for each file:
   1. HA\_NorthAmerica\_A.fasta
   2. HA\_NorthAmerica\_B.fasta

Your output can look like whatever you want. Obviously the more information the better.

Example of output:

|  |  |  |
| --- | --- | --- |
| Dinucleotide | HA\_NorthAmerica\_A.fasta | HA\_NorthAmerica\_B.fasta |
| AA | 0.80 | 0.82 |
| AC | 0.72 | 0.68 |
| AG | 0.56 | 0.52 |
| AT | 0.67 | 0.62 |
| CA | 0.32 | 0.29 |
| CG | 0.22 | 0.26 |
| CT | 0.89 | 0.87 |
| CC | 0.91 | 0.89 |
| GA | 0.66 | 0.65 |
| GC | 0.11 | 0.08 |
| . | \*This should add up to 1.0\* |  |
| .. |  |  |
| ... |  |  |