

Downloading genomes from the online database, NCBI

Week 10 Activity

Aim

For this activity, we are going to download a genome from the NCBI database, and calculate the ATGC content using our lab03.py scripts. We are going to be working on Trypsin, because reference genomes can be large and take up a lot of harddrive space.

Protocol

1. Log on to genome@trypsin.sfsu.edu (Reminder: password is DNA4rosalind!)
2. Navigate to your student directory
3. Download the reference genome to your student directory.
 - a. To download the reference genome, search for “Homo sapiens” and choose “Genome” dropdown menu.
 - b. The genome can be downloaded by clicking the “genome” link next to “Download sequences in FASTA format for”.
 - c. Right click the “genome” link and copy the link address.
 - d. On trypsin, paste the URL after the word wget
 - i. `wget [URL]`
 - e. The reference genome comes compressed, unzip the reference genome using the following command
 - i. `gunzip [name of reference file]`
4. Upload your lab03.py script to Trypsin.
 - a. Open another command line window
 - i. If you’re on a Mac, navigate to your lab03.py directory using the usual commands (`cd`, `ls`, `pwd`)
 - ii. If you’re on a PC, navigate to your lab03.py directory.
 1. `ls = dir`
 - b. Use the `scp` command to upload your lab03.py to Trypsin
 - i. Find the directory on Trypsin you want to upload your script to. This should be your student directory address. To find the absolute path of your student directory, type in `pwd` when in your student directory to find its address.
 - ii. Next, use the `scp` command where your lab03.py is:

```
scp [lab03.py name] genome@trypsin.sfsu.edu: [absolute path to your student directory]
```

5. Check to see if lab03.py and the reference genome is now in your student directory on Trypsin (use the command `ls` while in your student directory)
6. If both are there, run lab03.py on the reference genome to output the nucleotide composition of the human genome.

7. Check to see if the GC content your script computed is the same as the GC content listed in "<https://www.ncbi.nlm.nih.gov/genome/?term=homo+sapiens>" under "Summary" and subheading "Statistics".
8. Submit your ATCG composition output to iLearn Assignment "Week 10 - NCBI Activity". If your output contains the DNA codon composition, you can submit that too.