CMPSC 300
Introduction to Bioinformatics
Fall 2017

Lab 4: Genomic Regions Associated with Parkinson's Disease
Save this lab assignment to: `labs/lab4`

Objectives

- To learn how to use a Web-based genomic databases and tools.
- To understand the types of information stores in genomic databases.
- To learn how to use different interfaces to find and retrieve genomic information.
- To Learn how to write a program in python to load files using built-in control statements.

Reading Assignment

Chapter 1 in the *Exploring Bioinformatics* textbook.

PART 1: Retrieving Sequences

A *single-nucleotide polymorphism* (**SNP**, pronounced snip) is a DNA sequence variation occurring when a single nucleotide adenine (A), thymine (T), cytosine (C), or guanine (G) in the genome (or other shared sequence) differs between members of a species or paired chromosomes in an individual. As mentioned in class, the appearance of these SNPs may be used to separate DNA samples from each other based on the presence or absence of a particular type of nucleotide.

For instance, The research group, Do *et al.* genotyped 3,426 PD patients and 29,624 healthy control individuals for 522,782 known simple nucleotide polymorphisms (SNPs). They identified 11 SNP sites where one allele was correlated with PD with a statistically significant frequency. Known (SNPs) in the human genome are recorded in the primary genomic database dbSNP, where each SNP has a unique accession number that identifies it. In this laboratory assignment you will investigate the genomic neighborhood of one of these SNPs with the accession number *rs11868035*, which was identified by Do *et al.*

Research Steps and Questions

1. From the chapter reading, it is known that the SNP, *rs11868035*, is located within the gene, *SREBF1*. We would like to find the string of nucleotides of this gene to analyze it further.
2. Perform a search for this SNP at using the online web database provided by the National Center for Biotechnology Information (NCBI) at `https://www.ncbi.nlm.nih.gov/`.
    - Q1: According to NCBI, which particular database are you using to find this gene?

3. Determine your results from the Entrez database from this query.

   - Q2: What general observations can you make regarding the usefulness of your results (i.e., How easy is it to locate the gene of interest?
   - Q3: The data of several different organisms is mixed together in the result of your query. What are four (4) different organisms in which this same gene (*SREBF1*) may be found?

4. Now, reduce the number of results by limiting the search to only genes found in the genetics of humans (*homo sapiens*). Make sure your search results eliminate results that are not actually for SREBF1 but for some nearby gene.

   - Q4: What is the exact search query that you used to give you results which concerned only the human versions of the gene?

5. Find the gene whose LOCUS number is *NG_029029*.

   - Q5: What is the full title of this gene?

6. You are currently looking at the "Genbank" format of the gene. Now, click on the "FASTA" link on the top of the page. to change the view of the gene's information.

   - Q6: What is the major difference between the "Genbank" and "FASTA" formats?

7. Now, save the gene in a file using the FASTA format by clicking on the "Send to" button. Place this FASTA format file in your lab4 directory. You may have to create that directory if you don't have one in your course repository (cs300f2017-bbill/labs/). Return to GenBank and navigate through the features list. You can click on the links associated with features to alter the sequence display to show only the desired feature. You can also choose Highlight Sequence Features from the *Analyze this Sequence* list of links on the top right side of the page to visualize the locations of the features within the sequence. The list on the right also provides links to other additional information about this gene. Explore the various types of information available on this page.

   - Q7: Through your exploration, what (exactly) are the highlighted regions?

PART 2: Python Exercise for FASTA File

Use the included source code written in Python (*pythonCode/fastaReadWrite.py*). Use this code to load your FASTA file which you previously saved from Part1.

- Show a screen shot of your running the program using your saved FASTA file.
- Please provide a detailed idea about how the code works.
- Now, make a copy of this Python program to edit the source code so that it is able to count the numbers of bases in the sequence. Please use the sequence loading features (involving the biopython libraries) from the given code (*pythonCode/fastaReadWrite.py*), and borrow code from lab2's submitted code, *baseCounter.py* to count the number of bases in the sequence.

- Your output should include the name of the file, and the counts of the bases, {A,T,C,G}.
- Your source code file will be called: *lab4/fastaReadWriteBaseCounter.py*

Required Deliverables

All of the deliverables specified below should be placed into a new folder named 'lab03' in your Bitbucket repository (`cs300f2017-bbill`) and shared with the instructor by correctly using appropriate Git commands, such as `git add`, `git commit -m ''your message''` and `git push` to send your documents to the Bitbucket's server. When you have finished, please ensure that you have sent your files correctly to the Bitbucket Web site by checking the `source` files. This will show you your recently pushed files on their web site. Please ask questions, if necessary.

- An electronic version of a report in which you have answered the seven (7) comprehension questions from above (shown in red text). Please use Libre Office for this because you will be adding the screenshot and the output of your program from Part 2.
- A python source code file (called: *lab4/fastaReadWriteBaseCounter.py*). Your program is to be completed, properly commented and formatted Python program from part 2 conforming to the mentioned output specifications. Please make sure that your program has a comment header with the Honor code, your name, date and the description of the program (as shown in the class example programs).
- At the bottom of your submitted document containing the answers to the above questions, please include a copy of the output produced by running your program over the FASTA file that you downloaded in Part 1. Include a screen shot of the output in your report.

You should see the instructor if you have questions about assignment submission.