**BIO/CMPSC 300 Introduction to Bioinformatics   
Spring 2016  
Kristen Webb and Janyl Jumadinova***http://cs.allegheny.edu/sites/jjumadinova/300*

Lab Week 4 – Genomic Regions Associated with Parkinson's DiseaseGiven: Friday February 12, 2016  
 Due: Friday February 19 by 2:30pm

**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.**
* **Write Python program with file input by correctly employing control statements.**

**Reading Assignment:**

**Chapter 1 in Exploring Bioinformatics textbook.**

**Required Deliverables (submitted through your Bitbucket repository):**

1. **An electronic version of the report containing the answers to the lab comprehension questions in red and the first computational question in the last part of the lab.**
2. **A completed, properly commented and formatted Python program. 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).**
3. **An output produced by running your program.**

**General Guidelines for Labs**

**Work on the Alden Hall computers.** If you want to work on a different machine, be sure to transfer your programs to the Alden machines and re-run them before submitting.

**Keep all of your files!** Don't delete your programs and reports after you hand them in---you might need them again later.

**Back up your files regularly.** Use a flash drive or Google Drive or whatever your favorite backup method is.

**Review the Honor Code policy on the syllabus.** Remember that you may discuss experiments and programs with others, but copying answers or programs is a violation of the Honor Code

**Part 1: Genome Browsing to Identify Possible Disease-Associated Genes**

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.

1. Go to UCSC genome browser and using *rs11868035* as the search term, search for the SNP identified by this accession number. In your results you should see various data sets that include this SNP.
2. We are interested in the listing for the NHGRI Catalog of Published Genome-Wide Association Studies, because this means that the SNP has been identified as a point of interest in at least one GWAS. Click on the NHGRI link to find the SNP in the genome.   
   Examine the resulting browser view.   
   a) How many nucleotides are displayed in this view?
3. You should be zoomed in your browser view to visualize a very small region of the genome (note from the scale bar). Below the scale bar is the Ref Seq track followed by a track labeled NHGRI Catalog of Published Genome-Wide Association Studies. This track is turned on since you selected that particular link and the genome view is centered on the *rs11868035*. The point where the SNP occurs is shown by a bar and its accession number is highlighted. Above this display you should see an ideogram (schematic representation of the chromosome where this SNP occurs). Bands along the p (short) and q (long) arms of the chromosome are numbered, allowing the chromosome position to be displayed using notations such as 6q25.1.   
   b) What is the chromosome position for the *rs11868035* SNP?
4. Toward the bottom of the track display, you should see a track labeled Simple Nucleotide Polymorphisms and see bars for other SNPs in this region. Right-click on the track title and choose *full* to see their labels. Click on *rs11868035* SNP accession number in this track. You should see a page containing information from the dbSNP database, which specifies what nucleotides have been observed in human alleles at this position. In this case, the SNP is a base substitution, and either A or G can occur at this position, giving two known alleles. Return to the track display (previous page) and now click the SNP's accession number in the NHGRI Catalog track. In NHGRI Catalog of Published Genome-Wide Association Studies (rs11868035) page that comes up, note it has been identified in the Do et al. Study.   
   c) Has it also been identified by other GWASs?
5. Return to the track display and zoom out by clicking *10x zoom out* button. You should see more SNPs and start seeing the boundaries of genes. Right-click on the SNP track to change the display to *dense* to keep it from getting too long. At the top of the track display you should see three tracks showing different views of genes in this chromosome region: GENCODE V22 Comprehensive Transcript Set, RefSeq Genes, and Human mRNAs.   
   d) Does the *rs11868035* SNP occur within a gene? If so, what is the name of the gene?
6. Zoom in to read the actual nucleotide sequence by clicking on the base button. The display should be focused on your highlighted SNP, allowing you to see the position of the SNP within the gene. We note that the SNP is within an intron, not within the coding sequence itself, as can be seen from the Human mRNA track. Mutations within introns can affect the expression of a gene by impeding the splicing process, however it could also mean that the SNP identified in the GWAS occurs within a chromosome region important in PD but not within the actual gene correlated with the disease. We need evidence for possible disease involvement. One line of evidence is conservation – maintenance of the region in the genome over evolutionary time. A track labeled Multiz Alignments should be visible by default, showing an alignment between the human genome region you are browsing and the genomes of other sequenced vertebrates.   
   e) Is the gene in which *rs11868035* SNP is found conserved? Is the specific sequence where this SNP occurs conserved?
7. Identified *rs11868035* SNP was just a marker in the study by Do et al. And any genes in its neighborhood could be responsible for the correlation between this region and PD. We will turn on some tracks that are not displayed by default to analyze additional evidence. Zoom out until you get a view showing approximately 1,000 kilobases (1,000,000 bases) of the genome. This should bring the entire length of several genes into view, while centering on your original SNP. In the drop-down boxes below the track display, you will notice several options available for turning on specified tracks: dense will display the track with all features collapsed into a single line, full displays the track with each annotation feature on a separate line, squish displays the track with each annotation feature shown separately but at 50% height of the full mode, and pack displays the track with each annotation feature shown separately and labeled but not necessarily on a separate line. Turn on the track labeled OMIM Genes, showing genes listed in the OMIM database associating human genes with phenotypes. Under *Expression*, turn on the track labeled GNF Atlas 2, which summarizes experiments testing expression of the genes in various tissues. The OMIM Genes track uses a color scale to show how clearly a gene has been associated with a disorder or a phenotype; you can click on one of the green or gray bars for a description of this scale, then click on the OMIM entry number link to get a summary of information and click either the gene link or the disorder link to get the entries from OMIM. In the GNF Expression Atlas track, red bars indicate genes that were strongly expressed: the brighter red, the more expression. Continue to explore the various types of information related to this SNP and the associated genome region using the Genome Browser by adding and studying tracks that are not displayed from the drop-down options.   
   f) Using information from the various pieces, what evidence can you find to support the identification of one or more of the genes in this region as a candidate for a PD-associated gene?   
     
   g) Do you agree that the genes in this region most likely to be involved in PD were SREBF1 and RAI1, as was concluded by Do et al.? Why?

**Part 2: Retrieving Sequences**

You can retrieve the DNA or a protein sequence of a gene from the genome browser, but for the purpose of learning to use the NCBI Entrez interface, in this lab we will retrieve sequences by searching GenBank directly.

1. As discussed in the first part of this assignment, *rs11868035* SNP was located within the gene SREBF1. To download the sequence of this gene and/or the protein it encodes, go to the NCBI homepage.
2. Using Entrez interface do a simple search by typing SREBF1 into the search box and choose Nucleotide as the database to search. Look through your search results.   
     
   h) What general observations can you make regarding the usefulness of your results?
3. Now, conduct a narrower search by limiting the search to human genes. Make sure your search results eliminate results that are not actually for SREBF1 but for some nearby gene.   
     
   i) What search terms did you use?
4. Find an entry that includes “RefSeqGene” in its title, it should have accession number NG\_029029.1. Click on this sequence and observe the GenBank record for this gene. Click on the FASTA link on the top of the page and save it in a FASTA format in the *lab4* directory (create that directory if you don’t have one) in your own *300s2016-name repository*. 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 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.   
     
   j) Through your exploration, what did you learn about the function of this gene?

**Part 3: Computational Exercise for Gene Examination**

Gene finding in an organism, especially prokaryotes, starts from searching for an open reading frames (ORF). They are used in initial identification of candidate protein coding regions and can assist in gene prediction. An open reading frame starts with an ATG (Met) in most species and ends with a stop codon (TAA, TAG or TGA). In this part of the lab, you will take a step toward examining the SREBF1 sequence that you obtained in part 2 of this lab using computation. Please note that this portion of the lab contains the required portion and an optional portion. Optional portion consists of incremental steps for implementation of a program that would find all possible ORFs.

**Required**:

1. Write an algorithm in English (in your report text file) for finding all possible ORFs in a sequence. Think about the sequence of steps you would need to go through. You may assume that you start with a sequence as an input.

2. Write a Python program that reads a DNA sequence from a FASTA file and counts the number of “ATG”s. Your program should print the sequence name (not the sequence) and then print the number of ATGs found in the sequence. Your program should also print the position/index of the first occurring ATG. Make sure that your program has a comment header with the Honor code, your name, the lab number, the date and the description of the program. Take a **snapshot** of your program's output after you have completed and tested your program. To test your program's correctness, make a small text file with a few “ATG”s, and manually verify the results, before running your program on SREBF1.

**Optional**:

Extend your program to (save each one of these as separate programs):

1. Count the number of “TAA”s, the number of “TAG”s and the number of “TGA”. Take a snapshot of your program's output after you have completed and tested your program.

2. Find and print the first possible open reading frame, that is the first sequence of codons that starts with “ATG” and ends with “TAA”, “TAG” or “TGA”, print its length and its starting position/index and ending position/index. Take a snapshot of your program's output after you have completed and tested your program.

3. Find and print all possible open reading frames. You can explore the additional material on regular expressions to accomplish this task or learn the functionality of BioPython for finding ORFs. For each possible ORF, your program should output the possible ORF, its length, and it should also say the starting and ending point of each ORF. Take a snapshot of your program's output after you have completed and tested your program.