**Assignment 2: Learning splice sites from sequence**

**Download sequence files.**

Download the relevant sequence files for your organism from UCSC genome browser.

<http://hgdownload.soe.ucsc.edu/downloads.html>

Select the correct organism and assembly, and click on “Full Data Set”:

Human, Feb. 2009 (hg19)

Insect, D. melanogaster, Aug. 2014 (dm6)

Nematode, C. elegans, May 2006 (ce6)

Deuterostome, C. intestinalis, Mar. 2005 (ci2)

Then click on:

chromFa.tar.gz (human)

dm6.fa.gz (fly)

chromFa.tar.gz (worm)

ScaffoldFa.zip (ciona)

Untar/ungzip/unzip these archives and note the path containing your files.

**Create a bed file defining locations of all 5' splice sites.**

See the bed file specification: <https://genome.ucsc.edu/FAQ/FAQformat.html#format1>

Use the table from Assignment 1 to generate this file. Include 3 bases of the exon, and 5 bases of the intron (highlighted).

…55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77…

-----XXXXXXXXXXX-------

If this exon were on chromosome 1 on the + strand, a row in your bed file describing the coordinates for this splice site, with 3 bases of exon and 5 bases of intron, would look like this:

chr1 67 75 name score +

Note that necessary fields in the bed file are chromosome, start, end, name, score, and strand. Make sure to separate columns with the tab character (‘\t’). The “name” and “score” fields are just placeholder text – they are not necessary for our task but we need those columns to be present in the bed file.

**Create a bed file defining locations of all 3' splice sites.**

Include 20 bases of the intron, and 3 bases of the exon (highlighted).

…55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83…

------------------------XXXXX

**Fetch sequence defined in your bed file from your genome files.**

Download this python script to help you fetch defined sequence from fasta files:

<https://www.dropbox.com/s/qhpfthvj4b7m6lf/fetchFromBed.py>

To use this script, you would type:

python [path to script]/fetchFromBed.py --fetch bed\_f genomeDir out\_f

where

bed\_f is the bed file

genomeDir is the path to your genome directory

out\_f is the output file

**Taking data for all 5' splice sites in aggregate, compute the base composition at each position.**

1. That is, compute the percentage of A, C, T, and G at each position.
2. Plot the data as a line graph, where each series represents each letter.
3. Plot the data as a stacked bar graph.
4. Plot the data as pie graphs, where each pie is a position. Check out the “subplot” function.
5. Find one additional way to represent the data in a format that is easy for someone else to understand.

**Create the same plots for 3' splice sites.**