CSCI 291T Spring 2015 Programming Assignment 1 25 pts. Due: 2/3 (Tue)

Practice for Exon_masking and gene extraction from genome

From HG19 chr1.fa, extract following 5 RefSeq gene sequences and make a fasta file.
In each extracted gene sequence, mark exon regions in upper_case letters and intron regions in lower_case letters.

Given 5 gene_IDs:

NM_032291
NM_024066
NM_001199739
NM_003689
NM_001201547

Sequences in the fasta file should have the following format:

```
>gene_ID (+ or -)
ACGTACGT. . .
. . .
>gene_ID (+ or -)
ACGTACGT. . .
. . .
```

For each sequence, the first line starts with symbol '>' followed by gene_ID (from annotation) and strand (+ or -). Actual sequence starts from the 2nd line. You can limit each line length by 50, 70, or unlimited (i.e., one line per sequence).

Note: You should consider the strand when extracting a gene sequence from the genome, i.e., if strand is '-', extracted sequence should be reverse-complemented.

Input: hg19_chr1.fa //you downloaded this in Activity1
hg19_chr1_refseq_exon_annotation //placed in the Blackboard

Output: a fasta file having 5 extracted gene sequences in which exon regions are marked with upper_case letters and intron regions are marked with lower_case letters.

Submission:

1. hardcopy of all source codes used – please include documentation each.
2. a fasta file containing the five extracted genes.
   Since this file is too big to print, send it by email to: jpark@csufresno.edu