EECS 4425: Introductory Computational Bioinformatics
Assignment 1 (Released Oct 4, 2018)
Submission deadline: October 14, 2018

1. The assignment can be handwritten or typed. It MUST be legible.
2. You must do this assignment individually.
3. Installing R packages in the dept machines is a little trickier than doing it on your own computer, so use your own computer if possible.

Question 1

Generating pseudo-genomic sequences.

1. Create a sequence ACTGACTG.... of length 400.
2. Generate a random string of nucleotides of length 400 with equal probabilities.
3. Generate a random string of nucleotides of length 600 with equal probabilities of nucleotides in every position that is a not a multiple of 3, and with $p(a) = 0.5$, $p(c) = 0.25$, $p(t) = 0.15$ in every position that is a multiple of 3.

Question 2

The package seqinr.

1. Install the package in your computer (or directory, if you are using the departmental server).
2. Read the package documentation. Try the command lseqinr().
3. Get data files 1 and 2 from `https://www.ncbi.nlm.nih.gov/nuccore/257787102?report=fasta&log$=seqview` (choose save as fasta file) and `http://www.ncbi.nlm.nih.gov/sviewer/viewer.fcgi?tool=portal&db=nuccore&val=11497621&dopt=fasta&sendto=on&log$=seqview&extrafeat=0&maxplex=1`
4. Read the file 1 (fasta format). Output the following statistics of the files:
   (a) Percentages of a,c,t,g.
   (b) a table of the distribution of dimers (i.e. pairs of nucleotides). E.g., the segment acc has 1 ac and 1 cc.
   (c) a table of the distribution of trimers (also called codons), but non-overlapping; so acctcg has 1 acc and 1 tcg.

Question 3

Writing simple functions in R.

1. Write a R function that takes as input 2 indices and extracts the nucleotides between those indices (e.g. the inputs 10,15 should result in nucleotides 10 through 15 (inclusive) being extracted); then, the segment should be converted to an indicator sequence for the nucleotide g (it has a 1 in places where g occurs and 0 everywhere else). For this indicator sequence, plot the discrete fourier transform of the indicator sequence (plot the magnitude of the Fourier coefficients only).

2. Use your function on one long protein coding region and one long non-coding region from data file 2, as well as the sequence created in Q 1(c). There will be an annotation file for file 2 that you can use to identify these. Report any significant finding from these plots.