Homework 1 Part II - Implementation

Electronic submission Due: 11:59pm, Friday Sept 17

(40 points): Open reading frame finder & codon bias analysis

An Open Reading Frame (ORF) is a continuous stretch of codons (nucleotide triplets) that contain a start codon (i.e., ATG) at the beginning and a stop codon (i.e., TAA, TAG or TGA) at the end only, i.e., with no stop codon in the middle (https://en.wikipedia.org/wiki/Open reading frame). Note that there are three different ways (frames) that you can convert a DNA strand into triplets, each shifting one nucleotide from another, and a total of six different ways to find ORFs on a double stranded sequence.

Problem 1 (5 points): Reverse complementary strand.
Write a MATLAB function getReverseComp that takes a string as a DNA sequence and return its reverse complementary strand. Save as file getReverseComp.m.

Test it on the command line by typing in: getReverseComp('ACGTGCA') or run the corresponding cell in hw1script.m.

Problem 2 (2 points): Identify possible start codons.
Write a MATLAB function findStartCodon that takes a string as a DNA sequence, and returns the indices of all possible start codons. You may need the function strfind (type "help strfind" on matlab command line for usage.) Save as file findStartCodon.m.

Test it on the command line by typing in: findStartCodon('AATGTATGA') or run the corresponding cell in hw1script.m.

Problem 3 (3 points): Identify possible stop codons.
Write a MATLAB function findStopCodon that takes a string as a DNA sequence, and returns the indices of all possible stop codons. The indices should be sorted. Save as file findStopCodon.m.

Test it on the command line by typing in: findStartCodon('ATAAGTAGGA') or run the corresponding cell in hw1q2script.m.

Problem 4 (15 points) Identify the longest open reading frame (ORF)

Write a MATLAB function that takes as input a DNA sequence, and returns the longest ORF, which is given as two numbers (the start index of the start codon, and the END index of the stop codon). (The length = stop – start + 1 should be a multiple of 3.) Save this as file findLongestORF.m.

To test your function on the command line, type in:
findLongestORF('GGAGGCGTAAAATGCGTACTGGTAATGCAAACTAATGG') or run the corresponding cell in hw1script.m.

Problem 5 (15 points) Find longest ORF and analyze codon bias

Download sequence.fa, which contains a single DNA sequence related to the covid-19 virus. The file is in FASTA format (which is one of the most popular and simplest format), and can be read using the fastaread function in MATLAB Bioinformatics Toolbox. Use the code you developed above to find the longest ORF (considering both strands of the sequence). (The longest ORF is more than 10,000 bases, so depending on the efficiency of your algorithm, it may take a few minutes or longer. To debug, you can start testing your program on the first 1000 bases of the sequence.)

Save the subsequence corresponding to the longest ORF as a new variable, named longest_ORF.

Plot the codon distribution of longest_ORF using the codonbias function in MATLAB (with ‘pie’ plotting option set to “true”) (Save as Fig 1).

As a comparison, try to shift the ORF to the left or to the right by 1 base, and replot the codonbias (Display as Fig 2 and Fig 3).

Submission: upload your matlab .m files, and a short report with the following information

1. What is the start and end index of the longest ORF and which strand is it on?
2. What is the three most frequent amino acid encoded by the ORF? (Use MATLAB function nt2aa and aacount to find out.
3. Include the three codonbias plots with clear title.
4. In Fig 1, which codon is most frequently used for the amino acid Alanine and which codon is most frequently used for the amino acid Valine, and what are the percentage of usage?
5. In Fig 2 and Fig 3, what are the most frequent codons used for Alanine and Valine, respectively, and what are their percentages?