EECS 4425: Introductory Computational Bioinformatics
Assignment 3 (Released Nov 19, 2018)
Submission deadline: December 1, 2018

1. The assignment can be handwritten or typed. It MUST be legible.
2. You must do this assignment individually.

Introduction

In this assignment you will use the phase of the Fourier Transform to get information about DNA sequences. In particular, you will see that the the phase of the $N/3$ coefficient can be used to detect deletions of 1 or 2 nucleotides. The algorithm is given below. You will implement the algorithm in R and display the results on a synthetic and a real sequence.

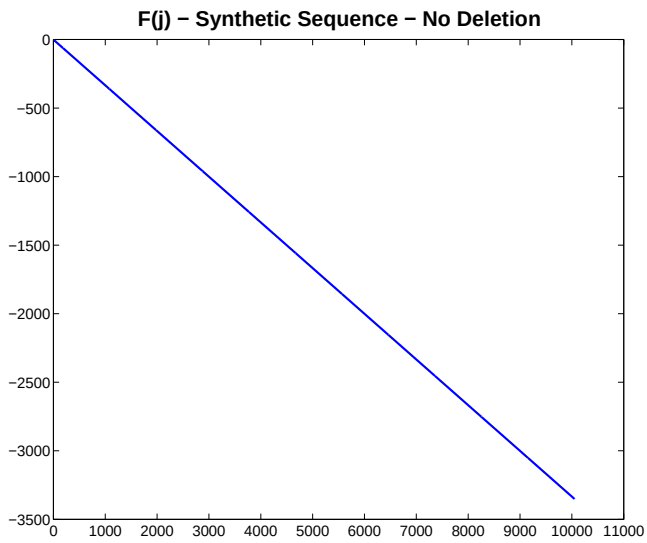
Question 1

In this question you will verify a property reported in the paper *D. Kotlar and Y. Lavner*, "Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions", Genome research, vol. 13, no. 8, pp. 19301937, 2003.
This paper observed that the phase of the Fourier spectral component at $N/3$ in each window from within a protein-coding region is tightly clustered around a single value, whereas the same quantity is uniformly distributed between 0 and 2π in non-coding regions.
Verify this by using sequences from the last assignment (you can choose one exon and one non-coding region of size about 5000. Use a window size of 351 and plot for each window position the phase of $N/3$ Fourier coeffficient.

Question 2

Next, implement the following algorithm for window size $W = 351$.

1. Compute the binary indicator function for nucleotide G the first W nucleotides of the query DNA sequence.
2. Calculate the DFTs of the binary indicator sequence obtained in step 1, and record the phase of $G[W/3]$.
3. The window is moved forward by 3 nucleotides, and repeat steps 1 and 2 until the entire sequence is read.
4. Compute a histogram from the phases of $G[W/3]$ that you recorded. The largest peak in these histograms is labeled as ϕ.
5. The phase in each window is classified into 3 classes $C_{\phi-2\pi/3}, C_{\pi}, C_{\phi+2\pi/3}$, and labeled -1,0,1. This classification is done using the nearest-neighbor heuristic, and the class assigned to a window is henceforth referred to as its phase label. The phase label thus defines a function $f : [1, n_W] \rightarrow 1, 0, +1$, where n_W is the number of windows used by the algorithm.



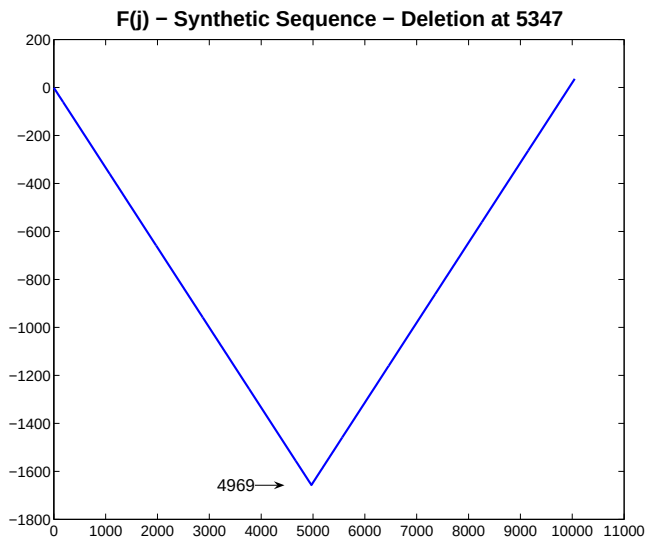


Figure 1: Sample graphs

6. We then define the cumulative function $F : [1, Nn + 1] \rightarrow [N, +N]$ as

$$F(j) = \sum_{k=1}^{j} f(k)$$

The function $F(j)$, when plotted against j, gives a visualization of the changes in phase. A sequence of windows with phases in the same class will yield a straight line. Any change in phase results in a change of slope (see figure 1.

To the above add a step to predict the location of a phase change. Test the algorithm on a coding region and a non-coding region by testing 1 or 2 nucleotides at 5 random locations for each and list the predictions from your algorithm.