CSE 5290: Algorithms for Bioinformatics
Assignment 2 (Released Oct 18, 2011)
Submission deadline: Nov 2, 2011

1. The assignment can be handwritten or typed. It MUST be legible.
2. Please submit the R code using the submit command.

Question 1

Use seqinr as required.

1. Create a pseudogenomic sequence of length 1 million base pairs. This sequence should have randomly distributed exons of total length about 100,000. Make the sizes range from 500 to 10,000. You can choose the lengths of introns. For introns, use $p(A) = p(C) = p(T) = p(G) = 0.25$ in each position. For exons, use the strategy from the last homework - i.e., 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.
2. Go to the URL `http://www.ncbi.nlm.nih.gov/nuccore/NC_011748` and download the annotations as a GenBank (.gb) file and the sequence as a FASTA (.fa) file.

Question 2

Simple exon prediction: we will use the $N/3$ coefficient of the Fourier spectrum to predict exons. We saw in Assignment 1 that this coefficient has a low value in introns and a high value in exons.

1. For the synthetic sequence, run the following algorithm. Set window size $w = 351$. Slide a window of size w by 3 nucleotides across the sequence. For each position of the window, compute the $w/3$ Fourier coefficient (magnitude only) from the G indicator sequence, and plot it against the window position. You need to decide on a threshold to determine the value of the $N/3$ coefficient above which you will classify it as an exon. Determine a good value of the threshold from the synthetic data.
2. For the E.Coli sequence you downloaded, run the algorithm and get exon predictions. Define a nucleotide to be correctly predicted if it is in an exon and predicted to be in an exon, or if it is not in an exon and not predicted to be in an exon. Otherwise classify it as a false positive (algorithm predicts an exon when there is none) or a false negative (the algorithm did not predict it but it is part of an exon). Count the number of false positives and false negatives.
3. Repeat the experiment on a G+C indicator sequence. That is the replace G's and C's by 1 and A's and T's by 0. Is the accuracy better in this case?

Question 3

You will try to improve the accuracy of the previous algorithm as follows.

1. First, you will modify the algorithm by first applying a triangular window on the indicator sequence and then computing the FFT as before. Do you notice anything different in the N/3 coefficient vs position graph? Is the accuracy any better?

Note: a window function is a function that starts at $x = 0$, goes up (to 1) linearly to $x = 176$ and then decreases linearly (to 0) to $x = 351$. Thus the function (when plotted) makes an isoceles triangle whose base is the window size and height is 1. Instead of taking the FFT of the sequence $X[i]...X[i + 351 - 1]$ for window i, you will take the FFT of the sequence $X[i]w[1], X[i + 1]w[2], ...X[i + 351 - 1]w[351]$, where $w()$ is the window function.

2. In fig 6.24 (page 196) of the text, it says that exons are typically flanked with AG and GT. Try to extend your exons to have these terminals. You have to resolve ambiguities in some simple heuristic manner, e.g by using the nearest AG or GT). Is the accuracy any better?