CSE 5290: Algorithms for Bioinformatics
Assignment 2 (Released Oct 1, 2009)
Submission deadline: 1 pm, Oct 15, 2009

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

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 predicted exon to be correct if it overlaps a real exon, otherwise classify it as a false positive. Count the number of false positives (algorithm predicts an exon when there is none) and false negatives (the algorithm misses exons).
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

We will try to improve the accuracy of the previous algorithm as follows. We will modify the algorithm to first apply a triangular window on the indicator sequence and then compute the FFT as before. Do you notice anything different in the N/3 coefficient vs position graph? Is the accuracy any better?