

CS195L HW5

Due: Thursday, March 10th 2:30pm

0 Phase Transitions (20)

In this problem we will investigate the importance of selecting adequate alignment parameters to produce a biologically meaningful result. We've uploaded the Mathematica notebook `AlignmentStats.nb` to the homework page. Familiarize yourself with the code. Adjust the parameters for the probability of A, C, G, T using the sliders generated from the `Manipulate` Mathematica function and **describe how varying the probabilities of nucleotides affect the phase transition from a logarithmic-length local alignment to a linear-length local alignment**. Use as much or as little of the existing functionality as you'd like. Please provide support for your claims using Mathematica.

Also, for a particular set of parameters, investigate how the alignments look around the phase transition. **Please record at least one set of alignments representing the three phases (logarithmic local alignment, transitional phase, and linear local alignment). How do the parameters affect the alignments?** We recommend using the `DynamicStringAlignment.nb` Mathematica notebook included on the resources page of the website to visual alignments.

1 Probability and RNAseq (40)

The bacterial organism $\sum K\Omega 9$ has a 262,144bp length genome. (Only in this class would an organism have a power of 2 as genomic size). You've been given the task of identifying the coordinates of the genes of this organism. To infer the presence of a gene, we need to identify an mRNA transcript for the genomic sequence. However, in practice this is done using RNAseq. mRNA from cell extract are cut into many fragments (assume randomly), then one end of the fragments is sequenced up to a read length r . For this problem, $r = 8$.

Your task is this: we give you the genome of $\sum K\Omega 9$. Then we give you 2000 8-mer reads of mRNA transcripts. Assume that the genomic sequence is the template strand, thus the mRNA is in the same orientation as the genome, only with "U" substituted for "T". Output your inferred coordinates of the genes in the genome, given that this bacteria is intron-less. (Note: There is no "correct" answer as you might not have the read fragment for the ends. In fact, since the same end of the mRNA is sequenced each time, it's almost impossible to infer one of the termini!)

Document all of your methods, and provide code. For each gene you identify, give a statistical argument for that prediction. This may consist of one overarching paragraph, then an equation for each gene. Keep in mind that the mRNA reads will randomly match with other scattered 8-mers along the genome. Your solution must therefore first answer the following questions: What is the probability of a random 8-mer match by chance? (Assume equal probabilities of A,C,T,G.) What is the probability of k matches occurring in a "gene" region? Address any caveats in your

probabilistic model, or additional assumptions you make. You are graded on the strength of your model and the quality of argument.

Hint: This could help in a simple model: MATLAB function `binocdf(x, n, p)` gives the probability of $\leq x$ successes in n trials with p probability of success.

Extra-credit: Allow for sequencing error by implementing an inexact match of up to k mismatches. You will also need to add complexity to your probabilistic model in your argument.

2 De Bruijn Graphs (40)

Given an integer k , the De Bruijn for sequence assembly is constructed by creating 4^k nodes each representing a unique k -mer (piece of DNA of length k). Let a and b be two nodes with DNA sequences of a_1, a_2, \dots, a_k and b_1, b_2, \dots, b_k respectively. We place a directed edge from node a to node b if $a_2, a_3, \dots, a_k = b_1, b_2, \dots, b_{k-1}$, in other words, if the last $k - 1$ characters of a overlap with the first $k - 1$ characters of b .

Write a program that takes as input a FASTA sequence representing a genome and a k . Construct the De Bruijn graph for the input k and output a visual representation of the graph. Use the input genome to add the edges to the De Bruijn graph. You can use whatever external libraries to draw the graph but we'll also accept the graph as text output in DOT (<http://www.graphviz.org/doc/info/lang.html>) or Mathematica (look up GraphPlot function) format. The De Bruijn graph of the full $\sum K\Omega 9$ genome is likely too large to visualize so we've provided a reduced-sized $\sum K\Omega 9$ genome on the website. Feel free to try to use a large segment of the $\sum K\Omega 9$ genome or any other biological sequences as long as you can make sense of the output graph. Adjust the input k and describe how varying k perturbs the structure of the graph. The ultimate goal is to use De Bruijn graphs for de novo assembly; however, in this scenario, DNA sequencing errors greatly complicate the construction of the assembly. Can you think of how DNA sequencing errors might manifest themselves in the De Bruijn graph (an example would be great!) and suggest (*not code*) a method to correct them?

Extra-credit: Mathematica implementation *and/or* show the path through the graph representing the reconstruction of the input DNA sequence.