

## CS196-1 HW3

*Due: Thursday, Feb 28th 2:30pm*

### Problem Reading

- Chapter 3 of How To Lie with Statistics
- Dynamic Programming: pp. 167-183 in the textbook
- Could Your iPod Be Holding the Greatest Mystery in Modern Science?  
<http://www.cs.princeton.edu/%7Echazelle/iPod>
- Expandable DNA repeats and human disease  
<http://www.nature.com/nature/journal/v447/n7147/full/nature05977.html>

### Problem Mathematica (25)

We have compiled all of the DNA and amino acid sequences of the BRCA1 homologs from Homework 1. You can now find these sequences on the course website:

<http://www.cs.brown.edu/courses/cs196-1/hw3.html>

You will also find the code for two Mathematica programs on the website. One program is for DNA global alignment and the other is for protein global alignment. You can run Mathematica on CS department computers with the command `mathematica` or download it onto your personal computer from Brown's software download page:

<http://software.brown.edu>

After installing Mathematica, open up these two files. Anything between (\* and \*) is a comment. You will notice that one of the first comments says to enter two sequences. You must copy and paste DNA/protein sequences from the files provided on the website (genes-AA and genes-DNA) between the quotation marks in order for the program to align them. Note that the sequences you enter cannot have any spaces or line breaks.

If you scan through the code, you will notice various sections for finding the max, creating tables with zeros, defining the scoring matrix, converting letters to numbers, running the DP algorithm, and performing the traceback. The scoring matrix is of particular importance. Note that in the protein code, you can set the value of indels and in the DNA code, you can set the value of matches, mismatches, and indels. You might try playing around with various scoring matrices and seeing the difference in the result.

To complete this problem you must fill out the Excel file located on the website with a comparison of one gene's DNA and amino acid sequence to that of the other homologs. You can choose whichever gene you like to begin. You will need to run the Mathematica code with the appropriate sequences and record the resulting score, number of matches, number of mismatches, and number of indels.

Since the sequences we are working with are quite long and the DP algorithm is quadratic in both time and space, truncate all sequences to 500 nucleotides/amino acids.

In order to run the Mathematica code, your cursor must be in the code region, not the results region. You must then type `*shift* + *enter*`.

## Problem Statistics (25)

For this problem, you will be collecting statistics about the BRCA1 homolog DNA sequences available at: <http://www.cs.brown.edu/courses/cs196-1/hw3.html>

1. Determine each of the following for the entire length of one of the BRCA1 homolog sequences linked above:
  - What percentage of the nucleotides are As? Cs? Gs? Ts?
  - What percentage of the gene sequence consists of two consecutive As? Cs? Gs? Ts? (Note that such substrings may overlap, so the sequence “CAAAG” has two substrings of “AA.”)
  - What percentage of the gene sequence consists of three consecutive As? Cs? Gs? Ts?
  - What is the longest inexact repeat, allowing for at most one mismatch? (You may find it useful to reuse code you wrote for Problem 1 of Homework 2.)
  - What is the longest inexact repeat, allowing for at most two mismatches?

## Problem Contamination (25)

For this problem, you will need to use two resources available at:  
<http://www.cs.brown.edu/courses/cs196-1/hw3.html>

A *contaminated sequence* is one that does not faithfully represent the genetic information from the biological source organism because it contains one or more sequence segments of foreign origin. A common source of contamination is when a sequence of interest is inserted within another sequence, called a *cloning vector*, which allows biologists to easily clone, propagate, and manipulate it. Failure to remove the vector sequence is often the source of contamination.

Using the sequences and vector library provided at the above link, determine which of the sequences is contaminated with which of the vectors. You may assume that a sequence of 10 bases from the library that is found in a sequence means that the sequence has been contaminated.

## Problem The Car and Goat Revisited (25)

Consider that the car and goat problem discussed in class may have different probabilistic properties if the quantities of doors, cars, and goats are varied. In each of the following scenarios, determine whether you should switch doors, and explain why:

1. 5 doors, 3 goats, 2 cars (10 pts)
2.  $(m + n)$  doors,  $m$  goats,  $n$  cars (15 pts)