Assignment 5: perl

1. Write a program in perl that counts the number of occurrences of some pattern. Make the program executable from the shell and use the following syntax:

   myCnt.pl [<filename>] [<pattern>] [<overlap>]

Further details and elaborations:
   a. Program takes up to 3 arguments (all optional):
      i. The first argument, <filename>, is the file name to be searched
      ii. The second argument, <pattern>, is either a regular expression or text
         1. When <pattern> is a regular expression it should be delimited by the character ‘/’: /A-C/
         2. Under any other circumstances <pattern> is presumed to be text.
      iii. The third argument determines whether or not overlaps occur.
         1. When <pattern> is a regular expression the value of overlap is ignored.
   b. If an argument is omitted, then the user should be asked to provide the argument, if the argument is needed. This means that if <pattern> is a regular expression then <overlap> is ignored if it occurs as the third argument, and that the user is not prompted for its value if it <overlap> is omitted.

2. Use your program to calculate the number of pyridinines and the number of purines that occur in the genome for the Japanese Marten (NC_009685). How long (roughly) does the calculation take?

3. Use perl to produce a 4 by 4 table:
   a. Vertical entries are labeled by A, C, G, and T (in that order).
   b. Horizontal entries are labeled by A, C, G and T (also in that order).
   c. The (i,j)th entry in the table (row i, col j) should be the probability that nucleotide i is followed by followed by nucleotide j in the Japanese Marten mitochondrial DNA (forward strand).
   d. So, for example, the entry in the ‘C’ row and the ‘T’ column is the probability that a ‘C’ is followed by a ‘T’ in the Japanese Marten mitochondria. In this specific example the desired probability is calculated as the number of matches to /CT/ divided by the number of matches to /C[ACGT]/. **NOTE:** It is important to find ALL matches. This is my not-so-subtle way to remind you that overlaps need to be considered—and that you’ll need a slightly more sophisticated regular expression to do your counting. Just to drive the point home... consider the sequence CCCT. There is 1 occurrence of /CT/, but 3 occurrences of /C[ACGT]/.