Homework #2
Perl for bioinformatics (140.636)

If you can do this assignment, you have mastered most of the material covered in chapters 1-6 of "Learning Perl."

The assignment is due on Wednesday Sep 13 at midnight.

Problem #1 Calculate the length and GC content of a DNA sequence

Grab the file **`pf_chrom01.txt`** from

`/users/sph140636/shared/homework2`

It contains the published DNA sequence of chromosome 1 of *Plasmodium falciparum*. The file contains row upon row of sequence data.

Write a stand-alone script named `problem1.pl` that performs the following tasks:

(1) Read the file line-by-line from STDIN and append the lines into single variable that contain the sequence.
(2) Count the number of lines in the file.
(3) Count the number of nucleotides in the sequence.
(4) Count the number A's, T's, C's and G's.
(5) Finally, calculate the GC content of the sequence (count of C+G over the total count of A+T+C+G, in percent). You might need: **`<STDIN>`**, **`while`**, **`chomp`**(), **`tr///`**, **`int()`** , and the concatenation operator "**`.`**"

The script should be runnable with the following command line

```
./problem1.pl
```

Problem #2 Calculate the annealing temperature along the sequence

In this problem you will write a standalone script named `problem2.pl`that calculates the annealing temperature of 20nt long oligomers along a DNA sequence.

Background: In the *polymerase chain reaction* (PCR), double stranded DNA is first separated into two strands by heating to 95°C -- typically for around 60 seconds. The temperature is then reduced to around 55°C for about 30 seconds to allow complementary *primer* sequences to anneal to the template DNA strands. After annealing, polymerization is achieved by increasing the temperature to 72°C for 60-90 seconds. The cycle is then repeated. The yield of the PCR reaction depends, among other factors, on the choice of annealing temperature. This depends on the composition of the primer. For short oligonucleotides (< 25 nt), a commonly used approximate expression for the optimal annealing temperature (in °C) is $T = 2*(n_A + n_T) + 4*(n_G + n_C)$, where n_i represents the number of times the i-th nucleotide occurs in the primer.

Here is how I recomment solving this problem. Grab the sequence data in **pf_chrom01.txt** from

```
/users/sph140636/shared/homework2
```

First you want to step through the sequence, in one nucleotide increments, and print 20 nucleotides at a time starting at the current position. Thus for example, if the input sequence is

```
ATCCCGCTAATGCATCCCGCTAATGCCCCGCTA
```

The first four lines of output would be

```
ATCCCGCTAATGCATCCCGC
TCCCGCTAATGCATCCCGCT
CCCGCTAATGCATCCCGCTA
CCGCTAATGCATCCCGCTAA
```

To verify that the script is working correctly (and so you aren't flooded by subsequences on your terminal window), redirect the results to a file and view the first 10 and the last 10 subsequences in the editor of your choice. Compare these with the original sequence to verify if your script is working correctly. Note that all subsequences (even the last one) must be exactly 20nt long. Manually check that your code is producing the correct output.

Next you want to calculate the annealing temperature of each oligo. Use the formula given above to calculate the annealing temperature. Of oligo before it is printed. Modify your print statement so it prints the oligo sequence followed by the annealing temperature. Format the value of the temperature so that there is one significant digit after the decimal point, e.g. `21.1`

.

Hint: In problem 1 you used the **`tr`** operator. This won't work for problem 2.

Use R to plot the sequence of annealing temperatures. Save the plot as a jpeg file named `plot.jpg` and display the plot in your web page along with a descriptoin of the problem.

Note: you get a much more interesting plot if you smooth the data before plotting it. Can you explain, what's happening at the ends of the sequence?

Homework submission

Submit your homework in your homework directory for homework2:

```
/users/sph140636/homework/<your_userid>/homework2
```

The directory should contain at least the following 4 files:

```
problem1.pl
problem2.pl
plot.jpg
index.html
```

The two scripts must be standalone scripts, e.g. the script from problem 1 can be run as follows:

```
./problem1.pl
```

If you have to run the script by invoking perl from the command line, i.e.

```
perl problem2.pl
```

then it is not a stand-alone script.