

# BioE131: Perl exercises

I.Holmes, A.Chau  
Dept Bioengineering, UCB

Fall 2005

## First week

1. Reverse complement a FASTA file.
  - (a) Write a Perl program to do the following:
    - Open a file, whose filename is specified by the user as a single command-line argument.  
That is, if the name of your Perl script is `programname` and the name of the file is `filename`, then the script should be run by typing the following at the Unix commandline  

```
perl programname filename
```

or, alternatively (if you have made `programname` an executable file whose first line is `#!/usr/bin/perl`), by typing  

```
programname filename
```
    - Read the contents of this file, assuming it is a FASTA file of DNA sequences;
    - Print the reverse-complement of every sequence on the standard output, in FASTA format.
  - (b) Make a comprehensive list of errors that could occur at runtime, when this program is used.
  - (c) For at least two of these errors, modify the code to guard against the errors and issue informative warnings if they are encountered.
2. Write a Perl program to test if two FASTA files contain exactly the same set of sequences (and sequence names). If the files are equivalent,

the program should return without error; if not, it should return with an error.

The program should not care if the sequences are in a different order in each file, and it should not be case-sensitive with respect to the sequences; e.g. if one file contains sequences in upper-case, and the other contains sequences in lower-case, the program should not complain.

The program should be invoked by typing something like the following  
programname filename1 filename2

## Second week

1. Write a Perl program that, when invoked using a syntax like this  
programname filename  
reads in a FASTA file, named “filename”, of DNA sequences and reports the
  - (a) nucleotide composition;
  - (b) dinucleotide composition.
2. Write a Perl program that, when invoked using a syntax like this  
programname filename  
reads in a FASTA file of DNA sequences named “filename” then finds and prints the locations of:
  - (a) all the stop codons;
  - (b) all the single-nucleotide repeats (that is, sequences of the form “AAAAA...” where “A”, or any other nucleotide, appears five or more times consecutively);
  - (c) **(harder)** all the dinucleotide repeats (that is, sequences of the form “ATATAT...” where “AT”, or any other two-nucleotide sequence, appears three or more times consecutively);
  - (d) **(harder)** all the tandem repeats (that is, sequences like “ACGTACGT” where a given subsequence, in this case “ACGT”, appears at least twice consecutively).