

Problem 1

Obtain a *.fasta file of a DNA sequence from Genbank or another source. For ease of manipulation you should select a relatively short DNA sequence, for example from a single protein. You will manipulate this DNA sequence in the exercises of this problem.

- a) Create subroutines to perform each of the following tasks. You should have one subroutine for each task.
 - i. Prompt a user for the name of a fasta file to open. Open the file, read the data, extract the sequence information and assign sequence to a scalar variable.
 - ii. Calculate the percentage of each base type
 - iii. Prompt user for a motif and determine if the sequence contains it
 - iv. Generate the reverse complement of the DNA sequence
 - v. Transcribe the DNA sequence to RNA
 - vi. Translate the DNA sequence to a protein
 - vii. Output the protein sequence to a file
- b) Compile your subroutines into a library
- c) Write a perl program that 1) loads the library of perl subroutines that you created in part b), and 2) prints out results of the subroutines sufficient to show that the subroutines are functioning properly, e.g., print out a portion of the original DNA sequence, the portion's reverse complement, the transcribed sequence, and translated portion.

Problem 2

From the class website, obtain the *.fasta files of the DNA sequence for the large subunit of the protein ribulose 1, 5-diphosphate carboxylase (Rubisco). This enzyme catalyzes the addition of CO₂ to ribulose 1,5 diphosphate and subsequent creation of two molecules of 3-phosphoglycerate, which constitutes the first step of the Calvin Cycle in plants. The gene for Rubisco is located in chloroplasts in higher plants. The two *.fasta files for the Rubisco gene are from the Korean pine (*Pinus koraiensis*) and an algal species (*Chlorella vulgaris*).

Write a perl program to perform the following tasks. Consider using subroutines where appropriate.

- a) Calculate the percentage of base pairs that match in the two sequences
- b) Generate a random sequence with the same nucleotide percentages as the Korean pine sequence
- c) Calculate the percentage of base pairs that match between the Korean pine sequence and the random sequence. Comment on how closely the Korean pine sequence matches the random sequence and the *C. vulgaris* sequence.