1. Concatenation is attaching something to the end of something else. A biologist is well aware that joining DNA sequences is a common task in the biology lab, for instance when a clone is inserted into a cell vector or when splicing exons together during the expression of a gene. Write a program to store two DNA fragments into two variables called $DNA1 and $DNA2. Print the DNA onto the screen and concatenate the DNA fragments into a third variable $DNA3 and print them.

2. Genetic information normally flows from DNA to RNA. There are certain instances where the flow is in opposite direction. For example, certain viruses such as HIV can allow RNA to serve as a template in assembling and making DNA strands for its replication. Write a program that reverse transcribe a given RNA sequence into DNA.

3. The close relationship between the two strands of DNA in a double helix is such that, given the sequence of one strand, you can easily predict the other. Such a calculation is an important part of many bioinformatics applications. For instance, when searching a database with some query DNA, it is common to automatically search for the reverse complement of the query as well, since you may have in hand the opposite strand of some known gene. Write a program to calculate the reverse complement of a strand of DNA.

4. Programs interact with files on a computer disk. Develop a program that reads the protein sequence data from the file and stores it into a variable. First, create a file on your computer (use your text editor) and add the below protein sequence into it. You will be using the following data (part of the human zinc finger protein NM_021964):
MNIDDKLEGLFLKCGGIDEMQSSRTMVVMGGVSGQSTVSGELQDSVLQDRSM
PHQEILAADEVLQESEMRQQDMISHDELMVHEETVKNDEEQMETHERLPQGL
QYALNVPISVKQEITFTDVSEQLMRDKKQIR

5. Since DNA sequence data can use both upper- and lowercase letters, write a program that prints DNA sequences (containing both upper- and lowercase letters) in lowercase (acgt); write another that prints the DNA in uppercase (ACGT). Use the tr operator.

6. Perl has several convenient ways to get information into a program. One of such ways is to ask the user for specific input. Write a program that prompts the user to enter a DNA sequence. Print the given sequence to the screen and calculate the length of the input sequence.

7. An array is a variable that can hold many scalar values. Each item or element in the array can be referenced by giving its position in the array. Initialize an array that holds the four bases A, C, G, and T and write a program to use POP, PUSH, SHIFT, and UNSHIFT functions in the array.

8. A palindromic sequence is a nucleic acid sequence in a double-stranded DNA or RNA molecule wherein reading in a certain direction on one strand matches the sequence reading in the opposite direction on the complementary strand. Write a script that uses conditional statements to check a given DNA sequence is palindromic or not.

9. The genomic DNA base composition is a highly variable trait and is a major factor that significantly affects genome functioning and variation among organism. Write a program to calculate the nucleotide composition (A, T, G, C %) for a given sequence.

10. One of the most common things we do in bioinformatics is to look for motifs, short segments of DNA or protein that are of particular interest. They may be regulatory elements of DNA or short stretches of protein that are known to be conserved across many species. Write a Perl script that reads in protein sequence data from a file and scans for motifs given by the user as standard input (Download the protein file with accession NP_001278826.1 for this exercise)

11. A codon is a trinucleotide sequence of DNA or RNA that corresponds to a specific amino acid. The genetic code describes the relationship between the sequence of DNA bases (A, C, G, and T) in a gene and the corresponding protein sequence that it encodes. Create a Perl program that stores a DNA sequence of any length (preferably in multiples of 3), extracts all codons in the sequence, and stores them in an array and print the resulting output.

12. In computational biology, gene prediction or gene finding refers to the process of identifying the regions of genomic DNA that encode genes. Let's assume that a gene is a stretch of DNA sequences that codes for a protein. Usually, a coding sequence begins by ATG and ends with either TAA, TGA or TAG codons. Write a script to search for coding sequences in the given genome. You can download the genomic fasta file of 'Escherichia coli str. K-12 substr. MG1655, complete genome' using NC_000913.3 accession number.

13. A loop allows you to repeatedly execute a block of statements enclosed within matching curly braces. Construct a while loop code to read a multiline protein sequence file while printing each line as it is read. Additionally, add a condition to print an error message and exit the program, if the program fails to open the file.

14. Regular expressions are ways of matching one or more strings using special wildcard-like operators. Write a program that can read a protein file and check for a specific motif that begins with an Asparagine residue and ends with two Threonine residues (Use the given file ZNF148.fasta for this exercise).

15. Subroutines are an important way to organize a program and are used in all major programming languages. A subroutine wraps up a bit of code, gives the code a name, and provides a way to pass in some values for its calculations and then report back the results. Construct a subroutine that can calculate the total length of a polypeptide encoded by a given sequence. Integrate the subroutine to your program which will print the length of the polypeptide when prompted with a DNA sequence.

16. Translation is a process by which the genetic code contained within a messenger RNA (mRNA) molecule is decoded to produce a specific sequence of amino acids in a polypeptide chain. Write a Perl program to simulate how the genetic code flows from DNA into protein. Print out the mRNA sequence and the amino acids peptides that would be formed by the DNA sequence "ATGGTTGCACCACAACCGGTTGTGCTCTGCAGTTGA".

17. Over the fairly short history of bioinformatics, several different biologists and programmers invented several ways to format sequence data in computer files. There are many such formats, perhaps as many as 20 in regular use for DNA alone. Because of its simplicity, the FASTA format is perhaps the most widely used of all formats. Write a subroutine that can handle FASTA-style data.

18. Write a program that picks one of the four nucleotides and then keeps prompting until the user correctly guesses the nucleotide it picked.

19. Randomization is a computer technique that crops up regularly in everyday programs. Using randomization, it's possible to simulate and investigate the mechanisms of mutations in DNA and their effect upon the biological activity of their associated proteins. Design a program to model the accumulation of mutations in DNA with random numbers. The output of the program should be able to print out 10 successive steps of random mutation and respective changes in the input sequence.