INFO-F-434 – Biological databases and analysis of macromolecular sequences
Basic programming exercises

Ex. 1 *A **string** is a ordered collection of symbols and its **length** is the number of symbols contained in the collection. DNA is basically a string composed of 4 types of character, called nucleotides : 'A', 'C', 'T' and 'G'.*

***Given** : A DNA string **s** of variable length*

Output** : Four integers separated by space counting the number of times 'A', 'C', 'T' and 'G' occurs in **s

Ex. 2 *As you are now able to count each type of nucleotide, here is a new problem. GC content varies accross DNA and influences DNA mechanism such as transcription, etc. and can be calculated as follows :*

$$GC = \frac{\#C + \#G}{total\ nt}$$

***Given** : A DNA string **s** of variable length*

***Output** : A decimal number equal to the GC content of s*

Ex. 3 *RNA is a string also composed of nucleotides. RNA is directly obtained from the DNA by replacing all occurences of 'T' by 'U'.*

***Given** : A DNA string **s** of variable length*

***Output** : The transcribed RNA string of s*

Ex. 4 *In DNA, "A" and "T" are complement of each other, just as "C" and "G". Reversing complementing a DNA strand is taking the reverse of the strand and replacing each nucleotide by its complement.*

***Example :** "AAAACCCGGT" $\rightarrow$ "ACCGGGTTTT"*

***Given** : A DNA string **s** of variable length*

***Output** : Its reverse complement DNA strand DNA_c*

Ex. 5 *The **Hamming distance** is defined as the number of mutations, corresponding symbols that differ, between two DNA sequences.*

***Given** : Two DNA string **s** and **t** of equal length*

***Output** : The Hamming distance between s and t*

Ex. 6 *In the genome, you can find some **motif** that repeats all across DNA. This represents the problem of finding a substring in a string, the substring being a contiguous collection of symbols in the string. The position of the substring is the total number of symbols found to its left, including itself. For example : 'U' in "AUGCUUCAGAAAGGUCUUACG" are at the positions 2, 5, 6, 15, 17 and 18.*

***Given** : Two DNA string **s** and **t** of variable length*

Output** : All locations of **t** as a substring of **s