Foundations of Biology

1. The GC-content of a DNA string is given by the percentage of symbols in the string that is 'C' or 'G'. For example, the GC-content of "AGCTATAG" is 37.5%. Use the file names as “Input1.txt” for this question. This file contains 5 FASTA sequences along with their headers. You have to print the header of that sequence which has the highest GC content along with its GC content value followed by the next sequence header along with its value and so on.

2. Again use the file “Input1.txt” and convert each DNA sequence into its corresponding mRNA sequence. Print the header info along with the mRNA sequence.

3. Given two DNA strings s and t of equal length, the Hamming distance between s and t, denoted dH(s,t), is the number of corresponding symbols that differ in s and t. For eg

Seq1 (s)	A	T	T	G	C	T	A	C	T
Seq2 (t)	A	G	T	G	C	A	A	G	C

   The hamming distance between these two sequences is 4. Calculate the hamming distance of the sequence provided in “Input2.txt”

4. Calculate and print the dinucleotide frequency of the sequence given in “Input3.txt”

5. The file “Input4.txt” consists of an RNA sequence. Conver this RNA sequence into its corresponding protein sequence. Note multiple proteins can be formed from this sequence. Print only those protein sequences which are formed by proper start and stop codons. Once you obtain the protein sequence. Calculate and print its molecular weight (MW) in dalton using the table given below.

Full name	three-letter code	one letter code	MW(Da)
alanine	Ala	A	89
arginine	Arg	R	174
asparagine	Asn	N	132
aspartic	Asp	D	133
cysteine	Cys	C	121
glutamic	Gln	E	146
glutamine	Glu	Q	147
glycine	Gly	G	75
histidine	His	H	155
isoleucine	Ile	I	131
leucine	Leu	L	131
lysine	Lys	K	146
methionine	Met	M	149
phenylalanine	Phe	F	165
proline	Pro	P	115
serine	Ser	S	105
threonine	Thr	T	119
tryptophan	Trp	W	204
tyrosine	Tyr	Y	181
valine	Val	V	117