IQB Assignment - 1

Group - 25

Akshala Bhatnagar	2018012
Pragya Sethi	2018067
Arundhati Bhattacharya	2018225

q1.py

To run the python script type the following on the terminal:

```
python q1.py -i input_file -o1 output_rna_file -o2 output_protein_file
```

Input file : DNA.fa

Output file : q1_rna_output.fa, q1_protein_output.fa

The DNA sequence is extracted from the input fasta file. All the Ts are converted to Us and the corresponding RNA sequence is made. The RNA sequence is written in the output RNA fasta file. Then the genetic coding table is used in the first reading frame to translate to the corresponding protein sequence. The protein sequence is written in the output protein fasta file.

q2.py

To run the python script type the following on the terminal:

```
python q1.py -i input_file -o output_file
```

Input file : 3mgo.pdb

Output file : q2_output.txt

Title, header and resolution are extracted from the input pdb file and written in the output txt file.

q3.py

To run the python script type the following on the terminal:

```
python q3.py -i input_file -o output_file
```

Input file : protein.fa
Output file : q3_output.fa

Protein Sequence Alignment using Needleman-Wunsch algorithm (Dynamic Programming) and Identity scoring scheme

To align two sequences of lengths n and m each, the program does the following:

1. Similarity Matrix: An nxm similarity matrix is created in which 1 is put in the cell where the correspnding Proteins in sequence 1 and sequence 2 are identical, i.e. mat[i][j] = 1 if seq1[i] = seq2[j]

2. Sum Matrix (using Needleman Wunsch algorithm): A sum matrix is generated from the Similarity matrix starting from the lower and rightmost cell and moving from right to left till we reach the upper and leftmost cell. For each cell mat[i][j], the input is computed by the following: mat[i][j] = mat[i][j] + max(mat[i+1][j+1], mat[i+1][k] with k in range(j+2,m), mat[k][j+1] with k in range(i+2,n))

3. Traceback to find alignment: Starting from the first cell, the maximum value in the corresponding row and column of the cell is found; then we move to the cell diagonally down from this cell and proceed accordingly. For every Protein in one of the sequences whose match isn't found, a gap is produced in the other sequence. Proceeding likewise till we reach the lowermost cell.