Algoritmos Bioinformática /Bioinformática 2021/2022

Assignment 4

Sequence Alignment

This assignment consists of two parts. In the first part, you will produce the global alignment of two nucleic sequences. In the sequence part, you will apply the developed code for finding the most similar sequences to the Spike proteins.

Task 1 – Global Sequence Alignment

Use paper and pen to solve the following exercise. Apply the Needleman-Wunsch to the following sequences and parameters:

S1: TTACT
S2: TATAT
g = -3; Match = 2; Mismatch = -1

Calculate matrices S and T, the optimal alignment and its score. Check if there are alternative optimal alignments. Discuss this.

In the end make a photo of the solved alignment and submit as **png or pdf** format file. Different formats will not be considered. If you prefer you can do a scheme/computer graphic with the above resolution but submit it as png or pdf.

Task 2 – Global Sequence Alignment

In the SarsCov2 genome the four structural proteins are crown-like spike (S), envelope (E), membrane (M) and nucleocapsid (N). The S glycoprotein plays essential roles in virus attachment, fusion and entry into the host cell and is the focus of study for the development of vaccines and therapies.

In this task, we will analyze the similarity of the glycoprotein sequence in SarsCov2 with that of other similar genomes.

In the provided sequence file, the proteins YP_009724390.1 and QHO60594.1 represent isolates from Wuhan Hu-1 (2020) and USA-WA-1 (2020).

Perform a global sequence alignment using the *blosum62* substitution matrix and a gap value of -8. Once the alignment is produced count the number of mismatches in the alignment and keep track of the alignment score.

1) Consider the sequence YP_009724390.1 as the reference. Determine the order of the most similar sequences in the provided file. Compare this sequence with all the

remaining sequence using a global sequence alignment and using the obtained score as a similarity measure, i.e. the higher the score the higher the similarity. Your program should output a list of the sequences (one per line) and corresponding score in sorted descending order.

2) Write a small comment (in the standard output), of 2 to 3 lines, on your interpretation of the results. Look at the corresponding species of each sequence and discuss if the results make sense.

3) Create two matrices between every two pairs of sequences: i) scores of the alignments; ii) number of mismatches. In the matrix use the row and column indices as the order of the sequences in the fasta file. Extra point: plot the ids of the sequences as rows and columns identifiers.

4) Discuss evident properties of these matrices. Look how the values are distributed. Write a comment on the output. Extra point: modify the plot of the matrix accordingly to your comment.

5) Repeat the steps 1) to 4) but now using a local sequence alignment. In the output create a separation line indicating the start of the procedure for local alignment. Extra point: develop the code to make it less repetitive.

Write a **script called similarity.py** that takes as input the file with sequences and produces the expected results. You can assume the file *blosum62.fas* is in the same directory you are running your script. Run as:

python similarity.py glycoproteinS.fas

Merge the file.png + similarity.py in a zip file and submit as similarity.zip.