Assignment 3 answers :

Answer 1 :

a. SARS-COV2 is more similar to SARS - COV than with MERS-COV
b. It is easier to identify similarity using protein sequences than DNA sequences. This is because Dna sequences for these 3 are less similar even from an initial stage (or word sizes) and thus differences in the graphs are not as stark as those between the Proteins' graphs of the three.
c. The graphs are available at the following link :
    i. https://docs.google.com/document/d/1O0WH6CKlhgMA9VyJxc68gtSK5JR6uz1l3ce-CZhOO_0/edit?usp=sharing
    ii. A pdf of the same has been attached with the final submission folder.

Here, we observe that while at the DNA level, both the identity and similarity are the same, at the protein level, the similarity is much larger than the identity. This is because the two protein sequences have a common ancestry and hence have similar properties. However, the actual elements in the sequence might be different, resulting in lower identity

Answer 2 : PART A

A. The numbers are as follows :
    a. At protein level :
        i. Percentage Similarity = 80.8%
        ii. Percentage Identity = 70.2%
    b. At DNA level :
        i. Percentage Similarity = 73.4%
        ii. Percentage Identity = 73.4%
    c. Reasons : In these numbers , we see that while at the DNA level , both the identity and similarity turn out to be the same, at the protein level however, the similarity is decently larger than the identity. This is because the two proteins sequences have a

common ancestry and hence have similar properties. However, the actual sequences might still be different, resulting in the lower identity.

B. Difference between Identity and similarity :

  i. Similarity signifies likeness between two sequences in comparison as a whole while Identity is the number of characters that match exactly between two different sequences.

C. Difference between local and global Identity:

  1. Global alignment is when you take the entirety of both sequences into consideration when finding alignments, whereas in local , you may only take a small portion into account.
  2. At the DNA level, there does not seem to be present much difference between the local and global similarities and identities. However, at the protein level, there seems to be more local similarity (86%) than global similarity (80.6%).

D. The data from the text files have been included in full in the following file :

  ii. https://docs.google.com/document/d/1O0WH6CKlhgMA9VyJxc68gtSK5JR6uz1l3ce-CZhOO_0/edit?usp=sharing

  iii. The same file has also been included in the pdf format with the final submission folder.

ANSWER 2 : Part B :

a. A general rule of thumb says that two sequences are approximate homologues if sequence similarity is above around 40 percent.

  i. Based on this rule of thumb and the data obtained from the analysis, we say that, yes these two sequences are approximate homologues.

b. We get a better idea of the same from the similarity number in the DNA based analysis. They are higher than those of the protein based analysis.

- The assessment has been included in the links:
    - Protein based :
        - https://www.ebi.ac.uk/Tools/services/rest/emboss_needle/result/emboss_needle-I20200404-182321-0550-94690044-p2m/aln
    - Dna based:
        - https://www.ebi.ac.uk/Tools/services/rest/emboss_needle/result/emboss_needle-I20200404-182528-0595-99860947-p2m/aln
    - The following links contain the respective alignment files.

Answer 3: The solutions is as follows

1. Link for the protein based search:
    a. https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=fasta-I20200404-183206-0162-65266355-p2m
2. Link for the Dna based search:
    a. https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=fasta-I20200404-183343-0848-74276617-p2m

A. According to the list, the closest homologue seems to be the HUMAN SARS Coronavirus

    a. Spike glycoprotein OS=Human SARS coronavirus OX=694009 GN=S PE=1 SV=1

B. The details as asked are as follows (for DNA based, but can be easily got for the protein based search too) :

    a. Score = 1416.3
    b. percentage identity = 76.2 %
    c. Positives : 91.5
    d. length of the alignment = 1255
    e. The expect or e-value = 0.0

C. Yes, it is one of the hits (the second one). The results match with the alignment numbers obtained using ‘water’: 73.2%. We can see that they are similar to those obtained from the database. This indicates that both these viruses have a common ancestry.

D. Yes indeed, The BAT coronavirus is at the 3rd rank in the DNA analysis list. The similarity scores are as under :
   a. Score = 1137.8
   b. Percentage Identity = 76.2
   c. Positives = 89.5
   d. Length of Alignment = 1242
   e. Expect or e-value = 0.0

Answer 4 :

- The links that are being used are :
  - https://www.ebi.ac.uk/uniprot/TrEMBLstats
  - https://www.ncbi.nlm.nih.gov/genbank/statistics/
- The sizes are as follows :
  a. Uniprot Database :
     i. 177754527 Sequence
     ii. 59974041839 amino acids
  b. Size of GenBank Database :
     i. 216214215 sequences
     ii. 399376854872 bases

A. The calculations are as follows :
   a. Memory complexity required for checking the similarity of two sequences = m*n
      i. Where m and n are the number of bases in the two sequences
   b. Time complexity will also be the same = O(mn)
   c. Length of the query sequence = 1000

i. Total number of matrix cells (for dna) = (GenBank database size)*(len of query sequence) = 399376854872*1000

ii. If we assume that 10^7 iterations would be completed in 1 second, then, (which is the standard for most x86 processors active today) Total time taken = 5997404.1839

iii. Total number of matrix cells (for proteins) = (size of uniprot) * (len of query seq) = 59974041839 * 1000

iv. Number of iterations that would be made for comparing DNA sequence = total number of bases in GenBank * query sequence = 399376854872 * 1000

v. If we assume that 10^7 iterations per second (which is the standard number in any x86 processor running today) then, Total time taken = 39937685.4872 seconds

B. Solution is as under :

a. Human Chromosome 1

i. Calculations are as follows:
m = 1000
n = 249M bp = 249 * 10^6 * 2 = 498 * 10^6
Memory requirement = m * n = 498 * 10^9
Time requirement = m * n = (Same as memory Complexity as established above)

b. Mouse Chromosome 1

i. Calculations are as follows :
m = 1000
n = 195Mbp = 195 * 10^6 * 2 = 390 * 10^6
Memory requirement = m * n = 390 * 10^9
Time requirement = m * n = 390 * 10^9 (Again, same as Memory requirement as established above)