Tutorial-1

By 2018101075

1. **1)** SARS-COV2 is similar to SARS-COV.

   2) It is easier to identify similarities using protein sequences than dna as when comparing DNA sequences, we get significantly more random matches than we get with proteins.
   There are several reasons for that:
   1. DNA is composed of 4 characters: A,G,C,T.Hence, two unrelated DNA sequences are expected to have 25% similarity.
   2. In contrast, the protein sequence is composed of 20 characters (AA). The sensitivity of the comparison is improved. It is accepted that convergence of proteins is rare, meaning that high similarity between two proteins always means homology.

   3) Graphs are submitted in folder Q1.
   k-tuple size is 5 for all 4 images of Dottup.
   Window size is 10 for all 4 images of Dotmatcher
   Threshold is 23 for all 4 images of Dotmatcher

2. **a)** Pairwise Alignment of SARS-COV2 and SARS-COV
   1) Global Alignment:
   For Protein:
   Identity: 70.2%
   Similarity: 80.8%
   For DNA:
   Identity: 72.7%
   Similarity: 72.7%
   Local Alignment:
   For Protein:
   Identity: 74.9%

Similarity: 86.0%

For DNA:

Identity: 73.2%

Similarity: 73.2%

Identity of DNA is more than protein as length of DNA is more as 1 amino acid equals 3 nucleotides and as explained in Q1 part 2, pair matching is more in DNA.

Similarity in Protein is more than DNA because there is no concept of similarity in DNA, either 2 nucleotides are equal or not. It is equal to identity while two not identical AA can be similar in Proteins.

2) Sequence identity is the amount of characters which match exactly between two different sequences. Gaps are not counted and the measurement is relational to the shorter of the two sequences. This has the effect that sequence identity is not transitive, i.e. if sequence A=B and B=C then A is not necessarily equal C (in terms of the identity distance measure) :

A: AAGGCTT, B: AAGGC, C:AAGGCAT

Here identity(A,B)=100% (5 identical nucleotides /min(length(A),length(B))). Identity(B,C)=100%, but identity(A,C)=85% ((6 identical nucleotides / 7)). So 100% identity does not mean two sequences are the same.

Sequence similarity is the degree of resemblance between two sequences when they are compared. It shows the extent to which residues are aligned. Similar sequences have similar properties. It is a more or less common practice to define similarity as an optimal matching problem (for sequence alignments or unless defined otherwise). Hereby, the optimal matching algorithm finds the minimal number of edit operations (inserts, deletes, and substitutions) in order to transform the one sequence into an exact copy of the other sequence being aligned (edit distance). Using this, the percentage sequence similarity of the examples above are sim(A,B)=60%, sim(B,C)=60%, sim(A,C)=86% (semi-global, sim=1-(edit

distance/unaligned length of the shorter sequence)). But there are other ways to define similarity between two objects (e.g. using tertiary structure of proteins).

3) There is a difference in global and local alignment of protein and global alignment is longer than local alignment with more insertions(-).
But global and local alignment is the same for dna sequences.

4) Alignments are in the Q2 folder.
For Needle:
Output Format: pair
Matrix: BLOSUM62,
Gap Open: 10, Gap Extend: 0.5
End Gap Open: 10, End Gap Extend: 0.5
End Gap Penalty: false

For Water:
Output Format: pair
Matrix: BLOSUM62
Gap Open: 10
Gap Extend: 0.5

b) Pairwise Alignment of SARS-COV2 and MERS-COV
Global Alignment:
For Protein:
Identity: 28.2%
Similarity: 52.8%
For DNA:
Identity: 47.7%
Similarity: 47.7%
Local Alignment:
For Protein:
Identity: 29.7%

Similarity: 45.3%

For DNA:

Identity: 47.8%

Similarity: 47.8%

1) Based on the above results, two proteins are not homologs.
2) Based on the above data, % of identity is very less in protein so we can say this inference was made from alignment of protein.

3.

1. SPIKE_CVHSA(the Sars coronavirus)
2. **UNIPORT:**

   Protein:-

   E-value: 0.0
   Score: 4817
   Identity: 75.9%
   Positives: 86.9%
   Query Length: 1281
   Match Length: 1255

   DNA:-

   E-value: 0.0
   Score: 5204
   Identity: 76.0%
   Positives: 86.8%
   Query Length: 3822
   Match Length: 1255

3. Yes it's there in the result .

   The score for SARS-COV using water:

   Percentage Similarity: 86
   Percentage identity: 74.9

   This is because BLAST looks for local alignment of the given sequence and WATER also calculates local alignment so the results of both are close
4. Yes, the percentage identity is high. The details:

   Score: 1816.6

Percentage identity: 74.6
Percentage similarity: 84.4
Length of alignment: 1242
E-value: 0

4.

1. Size of **UniProt** is 59974041839 amino acids .
   Size of GenBank is 399376854872 bases
   Length of query sequences = 1000
   Total number of matrix cells = mn = size*1000

a) Uniplot: 59974041839*1000=59974041839000
b) Genbank: 399376854872*1000=399376854872000

Total time taken = mn/10^7

a) Uniplot: 59974041839000/10^7 = 5997404.1839
b) Genbank: 399376854872000/10^7 = 39937685.4872

2. Space Complexity = mn, where m = 1000

a) Human: 249 * 10^6 * 2 * 1000 = 498 * 10^9
b) Mouse: 195 * 10^6 * 2 * 1000 = 390 * 10^9