Bioinformatics Assignment - 4

1. - **Global Alignment (Needleman-Wusch Algorithm)**

Here, $F(0,0) = 0$, $F(i,0) = -id$, $F(0,j) = -jd$,

$$F(i,j) = \max \begin{cases} F(i-1,j-1) - s_{ij} \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases} \qquad [\text{where } d = 3]$$

Alignment:

S_1 : GGCTGCAACTAGCTC

S_2 : GGGTA–AGCTTG––C

Score: 23

- **Local Alignment (Smith-Waterman Algorithm)**

Here, $F(0,0) = F(i,0) = F(0,j) = 0$

$$F(i,j) = \max \begin{cases} 0 \\ F(i-1,j-1) - s_{ij} \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases} \qquad [\text{where } d = 3]$$

Alignment:

S_1 : GGCTGCAACTAGCTC

S_2 : GGGTA–AGCTTGC––

Score: 29

2. Taking match to be 1 and mismatch to be -1, we get repeat sequence CACACTCACACCACACAGACA, with a score of 15.

3. - During genomic sequences, there may occur cases such that:

   a.) one of the sequences is contained within the other

   or b.) both the sequence have a subsequence in common

   - Such a case is considered as an overlap between the sequences
   - Dynamic programming may be employed for both identification and quantification of the overlap

-

Overlap Case	Global Case
$F(0,0)=0$	$F(0,0)=0$
$F(i,0)=0$	$F(i,0)=-id$
$F(0,j)=0$	$F(0,j)=-jd$
$F(i,j) = \max \begin{cases} F(i-1,j-1) - S_{ij} \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$	$F(i,j) = \max \begin{cases} F(i-1,j-1) - S_{ij} \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$

Recurrence relationship of DP for overlaps is identical to that for global alignment, while both differ in the boundary conditions.

4.

- Presence of continuous gaps are often caused by a single mutation, which led to continuous insertions or deletions.
- Affine gap scores are more sensitive to this, in comparison to other scoring techniques.
- Extensions are penalised much lower than gap-opening penalty, resulting in a much more realistic scoring system
- Further, the complexity of this method, $O(nm)$, is much lower than the general method, whose complexity is $O(n^3)$
- This makes it highly suitable for comparing long DNA sequences.

5.

- For global alignment algorithm, backtracking is alway performed from the bottom-right (n, m) cell of the matrix and terminates in the top-left $(0,0)$ cell.
- Thus, global alignment algorithm always results in global alignment.
- For local alignment algorithm, a case such that backtracking is done from the bottom-right (n, m) cell of the matrix and such that it terminates in the top-left $(0,0)$ cell is possible.
- Thus, local alignment algorithm may results in global alignment.

6.

- Both space & time complexity of alignment using DP is $O(mn)$.
- High time complexity would result in unpractical run-times. Considering sequences of order 10^6 (1 million) will result in run-times in the orders of several hours.
- Similarly, high space complexity would result in unpractical storage

space requirements. For example, genomes of the order of few MBs will require several TBs of storage space.

7. Query - length = 10^3 bases

Computation time = 10^7 cells/sec

Size of UniProt Database = 75271144009 bases

Size of GenBank Database = 940513260726 bases

$\Rightarrow$ Time taken for UniProt Database $= 75271144009 \times 10^3 \times 10^{-7}$

$= 7527114.4009$ seconds

≈ 87 days

Time taken for GenBank Database $= 940513260726 \times 10^3 \times 10^{-7}$

$= 94051326.0726$ seconds

≈ 1089 days