Bioinformatics Assignment - 3

1. **What is Sum-of-Pairs score?**

- Sum-of-pairs (SP) scoring is the standard scoring method for MSA.
- It is essentially a cost-weight function, calculated column-by-column manner, such that each column is assumed to be independent.
- The score is calculated by adding up the scores of all possible pairs in a column
- Scoring for this method is done using a substitution matrix such as BLOSUM or PAM scoring matrices.
- The alignment is fixed by maximising the score over all columns.

Drawbacks of Sum-of-Pairs scoring

- Sum-of-pairs (SP) scoring is largely heuristic and there is no theoretical justification.
- The score detoriates very quickly, when the number of differences are low (even as low as 1 disagreement).
- The rapid detorication might lead to incorrect overall alignments.

Alternative scoring system

- An entropy-based-score, based on scaling and by taking the natural log, can be used instead.
- This could solve the issue of rapid score detoriation and hence give more accurate alignments.
- Scores can be calculated as:

$$-\sum_{i} (C_i/C) \log (C_i/C)$$, where: C_i = no. of occurence of a amino acid in ith column

& C = no. of different symbols in given column.

- Here minimisation will be done, instead of maximisation.

2. Length of residues in each sequence $(L) = 50$

Time taken for aligning N sequences $= (2L)^{N-2} = 10^{2N-4}$ seconds

Now, no. of seconds in 5 billion years $= 5 \times 10^9 \times 365.25 \times 24 \times 60 \times 60$

$= 1.57788 \times 10^{17}$ seconds

$\Rightarrow$ No. of sequences that can be aligned $= \left\lfloor \frac{\log_{10}(1.57788 \times 10^{17}) + 4}{2} \right\rfloor$

$= \left\lfloor \frac{\log_{10}(1.57788) + 21}{2} \right\rfloor$

$= \left\lfloor \frac{0.198073971 + 21}{2} \right\rfloor$

$= 10$ sequences

3. * Initial condition: $\alpha_{i_1, i_2, 0}$, where it is the score of best alignment of the 2 most similar alignments.

* Boundary conditions: ○ Global alignment: $\alpha_{0,0,0} = 0$;

$\alpha_{i_1, 0, 0} = -i_1 d$;

$\alpha_{0, i_2, 0} = -i_2 d$;

$\alpha_{0, 0, i_3} = -i_3 d$;

○ local alignment: $\alpha_{i_1, 0, 0} = \alpha_{0, i_2, 0} = \alpha_{0, 0, i_3} = \alpha_{0, 0, 0} = 0$

* Recursion relation: $\alpha_{i_1, i_2, i_3} = \max_{\Delta_1 + \Delta_2 + \Delta_3 > 0} \left\{ \alpha_{i_1 - \Delta_1, i_2 - \Delta_2, i_3 - \Delta_3} + S(\Delta_1 \cdot x^1_{i_1}, \Delta_2 \cdot x^2_{i_2}, \Delta_3 \cdot x^3_{i_3}) \right\}$

4. ∘ Iteration 1 :

S_2 & S_4 are the most similar, hence we start by aligning them

S_2 : G T C T G A

S_4 : G T C A G C

∘ Iteration 2 :

Now, S_1 is most similar to the alignment obtained

S_2 : G - T C T - G - A

S_4 : G - T C - A G C -

S_1 : G A T - T - - C A

∘ Iteration 3 :

Now we align the remaining sequence S_3

S_2 : G - T C - T G - A

S_4 : G - T C A - G C -

S_1 : G A T - - T - C A

S_3 : G A T - A T T - -

Score: $C_1 = 6$

$C_2 = -2$

$C_3 = 6$

$C_4 = -2$

$C_5 = -2$

$C_6 = 0$

$C_7 = -2$

$C_8 = -2$

$C_9 = -2$

Total score = 0

5.