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 detoriation 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 alligning N sequences $= (2L)^{N-2} = 10^{2N-4}$ seconds

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

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

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

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

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

$$= 10 \text{ sequences}$$

3. * <u>Initial condition</u>: $\alpha_{i_1, i_2, 0}$, where it is the score of best allignment of the 2 most similar allignments.

* <u>Boundary conditions</u>: ○ <u>Global allignment</u>: $\alpha_{0,0,0} = 0$;

$$\alpha_{i_1, 0, 0} = -i_1 d\ ;$$
$$\alpha_{0, i_2, 0} = -i_2 d\ ;$$
$$\alpha_{0, 0, i_3} = -i_3 d\ ;$$

○ <u>Local allignment</u>: $\alpha_{i_1, 0, 0} = \alpha_{0, i_2, 0} = \alpha_{0, 0, i_3} = \alpha_{0,0,0} = 0$

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

4. **★ Obtaining the guide tree**

- Iteration 1:

	S_1	S_2	S_3	S_4
S_1		3	3	1
S_2			1	(4)
S_3				2
S_4				

$\left(\begin{array}{c} \text{Matrix} \\ \text{of Matches} \end{array} \right)$
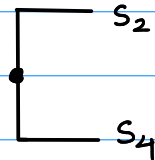
Tree:

- Iteration 2:

	S_1	S_2/S_4	S_3
S_1		2	3
S_2/S_4			1.5
S_3			

$\left(\begin{array}{c} \text{Matrix} \\ \text{of Matches} \end{array} \right)$

Tree:
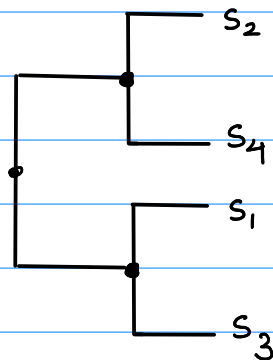
★ Obtaining the alignment

○ **Alignment between S_2 & S_4**

S_2: G T C A G C

S_4: G T C T G A

		G	T	C	A	G	C
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
T	-2	0	2	1	0	-1	-2
C	-3	-1	1	3	2	1	0
T	-4	-2	0	2	2	1	0
G	-5	-3	-1	1	1	3	2
A	-6	-4	-2	0	2	2	2

○ **Alignment between S_1 & S_3**

S_1: G A T – T C A

S_3: G A T A T – T

		G	A	T	A	T	T
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
A	-2	0	2	1	0	-1	-2
T	-3	-1	1	3	2	1	0
T	-4	-2	0	2	2	3	2
C	-5	-3	-1	1	1	2	2
A	-6	-4	-2	0	2	1	1

- **Alignment between S_1 & S_3 with S_2 & S_4:**

$S_1: G A T - T C A$

$S_2: G - T C A G C$

$S_3: G A T A T - T$

$S_4: G - T C T G A$

		GG	AA	TT	_A	TT	C_	AT
	0	-4	-8	-12	-16	-20	-24	-28
GG	-4	4	0	-4	-8	-12	-16	-20
TT	-8	0	0	4	0	-4	-8	-12
CC	-12	-4	-4	0	0	-4	-4	-8
AT	-16	-8	-4	0	-1	0	-4	0
GG	-20	-12	-8	-4	-4	-4	-4	-4
CA	-24	-16	-12	-8	-5	-8	-5	-5