

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 deteriorates very quickly, when the number of differences are low (even as low as 1 disagreement).
- The rapid deterioration 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 deterioration and hence give more accurate alignments.
- Scores can be calculated as:

$$-\sum_i (C_i/C) \log(C_i/C), \text{ where: } C_i = \text{no. of occurrence of a amino acid in } i\text{th column}$$

& $C = \text{no. of different symbols in given column.}$

- 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 $= \left\lfloor \frac{\log_{10}(1.57788 \times 10^{17}) + 4}{2} \right\rfloor$

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

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

$$= 10 \text{ sequences}$$

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: \circ Global alignment: $\alpha_{0,0,0} = 0$;

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

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

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

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

* Recursion relation: $\alpha_{i_1, i_2, i_3} = \max_{A_1 + A_2 + A_3 > 0} \left\{ \alpha_{i_1 - A_1, i_2 - A_2, i_3 - A_3} + S(A_1 \cdot x_{i_1}^1, A_2 \cdot x_{i_2}^2, A_3 \cdot x_{i_3}^3) \right\}$

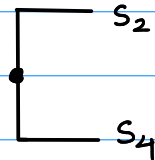
4. * Obtaining the guide tree

◦ Iteration 1:

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

(Matrix of Matches)

Tree:

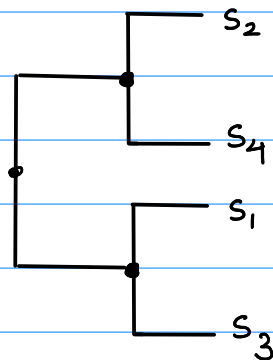


◦ Iteration 2:

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

(Matrix of Matches)

Tree:



* Obtaining the alignment

◦ Alignment between S_2 & S_4

S_2 : GTCAGC

S_4 : GTCCTGA

		G	T	C	A	G	C
	O	-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 : GAT - TCA

S_3 : GATA T - T

		G	A	T	A	T	T
	O	-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 :

		GG	AA	TT	-A	TT	C-	AT
S_1 : GAT-TCA		0	-4	-8	-12	-16	-20	-24
S_2 : G-TCAGC	GG	-4	4	0	-4	-8	-12	-16
S_3 : GATA T-T	TT	-8	0	0	4	0	-4	-8
S_4 : G-TCTGA	CC	-12	-4	-4	0	0	-4	-8
	AT	-16	-8	-4	0	-1	0	-4
	GG	-20	-12	-8	-4	-4	-4	-4
	CA	-24	-16	-12	-8	-5	-8	-5

* Obtaining the final score [taking $s(-, -) = -1$]

Column 1 = 6

Column 2 = -4

Column 3 = 6

Column 4 = -4

Column 5 = 0

Column 6 = -4

Column 7 = -4

⇒ Total score = -4

5. a.) log-odds score for a base 'b' at the i^{th} column is given as:

$$\text{score}(b, i) = \log_2 \left(\frac{f_{b,i}}{P_b} \right)$$

where, $f_{b,i}$ = frequency of base 'b' at column 'i'

P_b = background frequency / probability of 'b'

Here, we have: Background frequency = 0.25

$$\Rightarrow P_A = P_T = P_G = P_C = 0.25$$

Hence, we get PSSM as:

	Column 1	Column 2	Column 3	Column 4
A	1.26	-1.32	-0.32	-1.32
G	-1.32	1.49	-1.32	-1.32
C	-1.32	-1.32	1.26	-1.32
T	-0.32	-1.32	-1.32	1.49

b.) For each sliding window sequence, the score is given as:

$$\text{score} = \sum_{i=1}^4 \text{PSSM}(x_i, i)$$

[where x_i is the i^{th} base in the sliding window]

⇒ Scores :

- TGAG : -0.48
- AGCT : 5.50
- GAGC : -5.29
- GCTA : -5.29
- CTAA : -4.29

c.) Based on the scores calculated above, we can calculate ODDS scores for each of the window sequences.

⇒ ODDS scores :

- TGAG : $\text{antilog}(-0.48) = 0.716$
- AGCT : $\text{antilog}(5.50) = 45.131$
- GAGC : $\text{antilog}(-5.29) = 0.025$
- GCTA : $\text{antilog}(-5.29) = 0.025$
- CTAA : $\text{antilog}(-4.29) = 0.051$

The ODDS score of AGCT is maximum & hence this sequence is most probable.

We calculate probability of AGCT sequence as below:

$$\begin{aligned} P(ATGC) &= \frac{\text{ODDS}(\text{AGCT})}{\sum \text{ODDS}(\text{seq}_j)} \\ &= \frac{45.131}{0.716 + 0.025 + 45.131 + 0.025} \\ &= 0.98 \end{aligned}$$