Sci2 - Assignment - V

MSA & phylogeny

NAGA MANOJ KAR
2021101128

MSA

(1)

MSA is a tool used to align ≥3 sequences simultaneously.

Applications:

- used for phylogenetic analysis, functional annotation, structural analysis and identification of regulatory elements
- It provides more information than pairwise alignment as it identifies conserved regions & variations across multiple sequences
- This information can help identify functional domains, structural features and regular elements that may not be evident from pairwise comparisons.
- MSA provides a more accurate representation of the evolutionary relationships between sequences
- MSA can help identify convergent & divergent evolution across different lineages.

(2)

The sum-of-pairs (sp) score is a measure of the quality of a MSA, calculated by summing the scores of all pairs of aligned, residues in the MSA

However it has limitations, as it does not consider gaps and the variability of residues in columns.

Additionally it doesn't consider the importance of certain residues for protein functions.

It also tends to overweight the contributions of differences.

Other Techniques A free-form technique can be used to score. In this technique, the MSA software finds a phylogenetic tree showing the links among the sequences, the total lengths of the tree branches can be computed using the substitutions in the MSA column.

A simpler tree, with one of the sequences serving as the ancestor of all the others, can be used instead

(3) # Progressive Alignment Approach

Steps:

(1) All sequences are pairwise aligned to generate a matrix of pairwise similarity scores.

(2) A phylogenetic tree is constructed based on the pairwise similarity scores.

(3) A guide tree is constructed based on the phylogenetic tree

(4) Sequences are aligned one by one, starting from the most closely related sequences and gradually adding more distant sequences using the guide tree as a guide

(5) The alignment is refined to improve its quality

Drawbacks:

- NOT globally optimal
- Errors are propagated throughout the final result
- $O(n^3)$ time complexity makes it unsuitable for datasets with a large number of sequences.

The Relaxed Neighbour Joining Technique, which lowers the constraints for linking tree nodes can lessen the shortcomings of the progressive

alignment approach as a result, the time complexity is reduced to $O(n^2 \log n)$.

(4) $L = 50$

The alignment of N sequences takes $= (2L)^{N-2}$

$= 10^{2N-4}$

Time = 5 billion years

$= 5 \times 10^9$ years

$= 5 \times 10^9 \times 365 \times 86400$ sec

$= 15768 \times 10^{13}$ s

$10^{2N-4} = 15768 \times 10^{13}$

$\Rightarrow 2N-4 = 17.19$

$N = 10.595$

$\lfloor N \rfloor = 10$

∴ The computer can align sequences in 5 billion years

(6)

~~S1: GATTCA~~

match — +1

mismatch — −1

indel — −1

(S_1, S_2, S_3, S_4)

$^4C_2 = \frac{4 \times 3}{2} = 6$

↑

possible pairs, we need to do pairwise alignment.

(S_1, S_4), (S_1, S_2), (S_1, S_3), (S_2, S_3), (S_2, S_4), (S_3, S_4)

① S_1 and S_2

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

GAT_TCA
G-TCTGA

score = 1

mismatches = 3

② S_1 and S_3

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

GAT_TCA
GATAT-A

GAT_TCA
GATATT-

score = 1

mismatches = 3

(3)

S_1 and S_4

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

~~Score~~

GATTC--A
G--TCAGC

Score = -2

mismatch = 5

(4)

S_2 and S_3

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

G-TCTGA
GATAT-

G-TCTGA
GATATT-

Score = -1

mismatch = 4

(5) S_2 and S_4

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

GTCTGA
GTCTAGC

Score = 2

mismatch = 2

⑥ S_3 and S_4

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

GAT_ATT
G_TCAGC

score = -1

mismatch = 4

Score (pairwise) Matrix

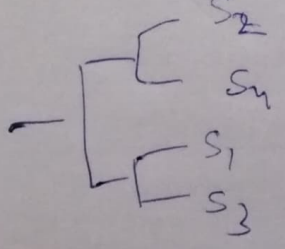
	S_1	S_2	S_3	S_4
S_1				
S_2	1			
S_3	1	-1		
S_4	-2	2	-1	

distance (pairwise)

	S_1	S_2	S_3	S_4
S_1				
S_2	3			
S_3	(3)	4		
S_4	5	(2)	4	

cluster division

(S_2 and S_4) (S_1 and S_3)



Final MSA Alignment:

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

SP Score:

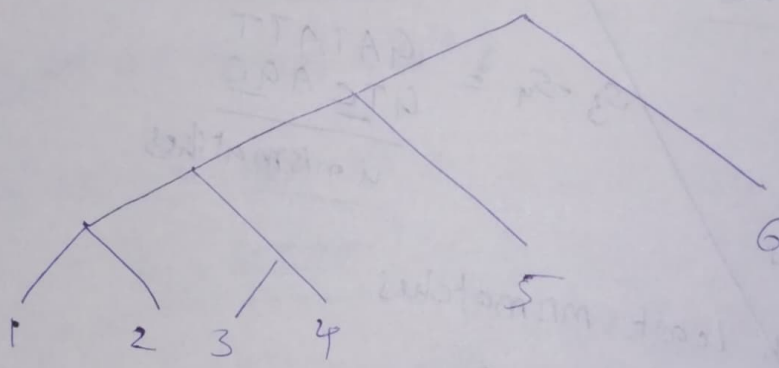
Col1 : 6
Col2 : −1 −1 −1 −1 −1 +1 = −4
Col3 : 6
Col4 : −1 −4
Col5 : 0
Col6 : −4
Col7 : −4

∴ Total score = −4

7) Given

Species	Site 1	2	3	4
1	T	C	A	A
2	G	C	A	T
3	T	T	T	T
4	G	A	T	A
5	G	A	A	C
6	A	T	A	G

the tree follows = $(((1,2),(3,4)),5),6)$



$(1,2) = \begin{Bmatrix} G \\ T \end{Bmatrix} C A \begin{Bmatrix} A \\ T \end{Bmatrix}$

$(3,4) = \begin{Bmatrix} T \\ G \end{Bmatrix} \begin{Bmatrix} A \\ T \end{Bmatrix} T \begin{Bmatrix} A \\ T \end{Bmatrix}$

$((1,2),(3,4)) = \begin{Bmatrix} T \\ G \end{Bmatrix} \begin{Bmatrix} A \\ C \\ T \end{Bmatrix} \begin{Bmatrix} A \\ T \end{Bmatrix} \begin{Bmatrix} A \\ T \end{Bmatrix}$

$(((1,2),(3,4)),5) = GAA \begin{Bmatrix} A \\ C \\ T \end{Bmatrix}$

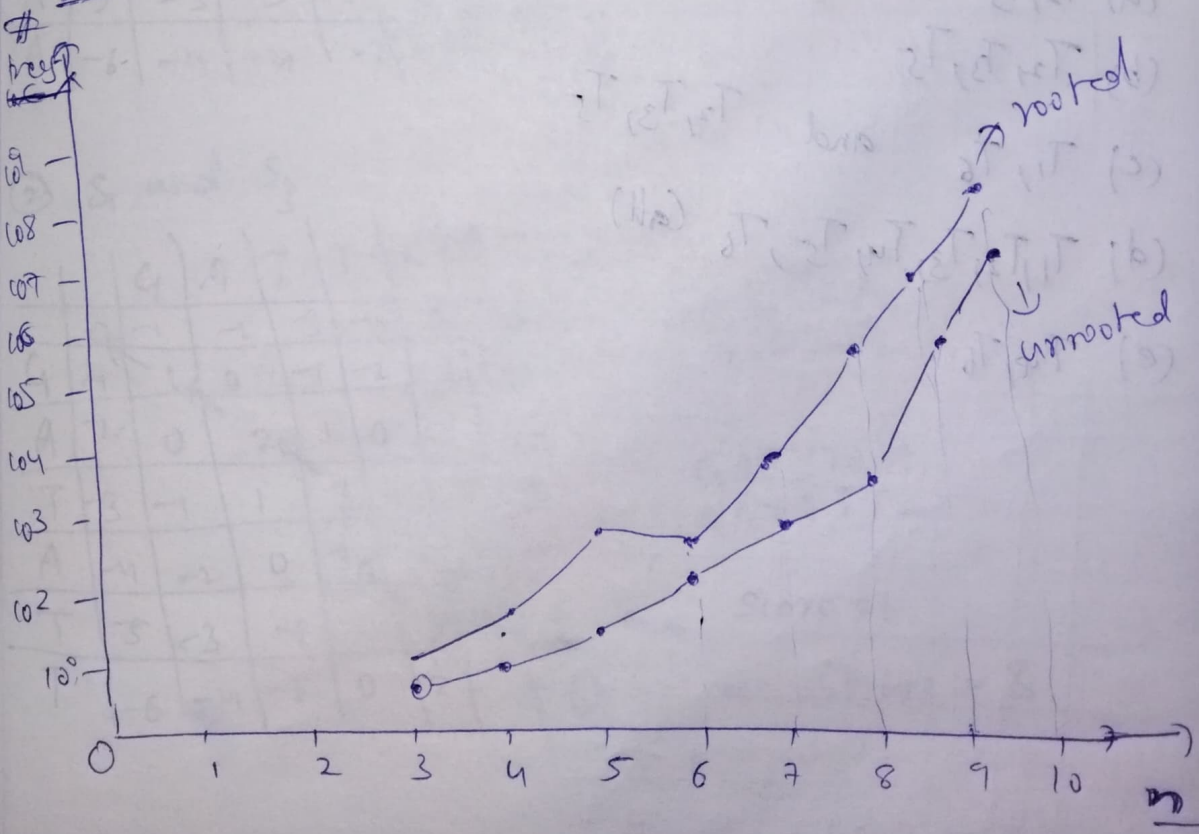
$((((1,2),(3,4)),5),6) = \begin{Bmatrix} A \\ G \end{Bmatrix} \begin{Bmatrix} A \\ T \end{Bmatrix} A \begin{Bmatrix} A \\ G \\ C \end{Bmatrix}$

Score = 2+3+2+1+3 = 11.

Table:

n	unrooted trees $\frac{(2n-5)!}{2^{n-3}\cdot(n-3)!}$	rooted trees $\frac{(2n-3)!}{2^{n-2}\cdot(n-2)!}$
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

Graph:



(9)

(a) T_2, T_3

(b) T_2, T_3, T_5

(c) T_1, T_6 and T_2, T_3, T_5

(d) $T_1, T_2, T_3, T_4, T_5, T_6$ (all)

(e) T_4, T_6