Science - 2 Assignment 4

Genius
Page:
Date: / /

Naimeesh Narayan Tiwari
2020101074

Q1 Applications

1. powerful tool for many fields such as phylogenetic reconstruction, illumination of functionality important regions, and prediction of higher order structures of proteins and RNAs

2. useful in bioinformatics for identifying sequence similarity, ~~producing phylogenetic trees, developing homology models of~~

3. producing phylogenetic trees

4. developing homology models of protein structures

MSA carries more info than Pairwise alignment

- it is because they show conserved regions with a protein family which are of structural & functional importance. MSA is basically an extension of pairwise alignment to incorporate more than two sequences at a time.

Genius
Page:
Date: / /

Q2 Sum-of-pairs score

Multiple sequence similarity suggests a common underlying structure of the protein, a common function, or a common evolutionary source.

For measurement of the quality of alignment, we analyze the scoring function which is used.

After getting aligned sequences, we sum up the score in all columns.

Sum-of-pairs is one such scoring system. It is defined on columns & is the sum of all pairwise scores of the symbols in the column.

DRAWBACKS & ALTERNATIVE

→ takes all pairwise info into account which becomes easily biased if the sequences from the same family are over-represented in the input & thus does not provide theoretical justification for the score.

→ Alternative is to use weighted SOP.

Genius
Page:
Date: / /

(Q3) steps involved in Progressive alignment approach

Step 1: Pairwise alignments — calculate distance matrix

Step 2: Construct a phylogenetic tree, cluster closely related sequences.
(rooted neighbor joining tree (guide tree))

Step 3: Progressive alignment: Align following the guide tree

Align closely related sequences first,

Page:
Date: / /

DRAWBACKS

D(1) ~~greedy~~ depends on the very first closely related sequences used for constructing the multiple alignment.

D(2) More distantly related the sequences are more errors will get propogated through the alignment.

D(3) greedy approach results in efficiency of the algorithm at the cost of accuracy.

Fixes to these Drawbacks.

for D1 → if the sequences align well there will be few errors

D2 → using Bayesian methods such as Hidden Markov models (HMMs) may be useful.

Other solutions: Stochastic or Iterative methods

(Q9)

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

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

-10

-3445945

Q10

a) $\{T_2, T_3\}$

b) $\{T_2, T_3, T_5\}$

c) $\{T_1, T_6\}, \{T_2, T_3, T_5\}$

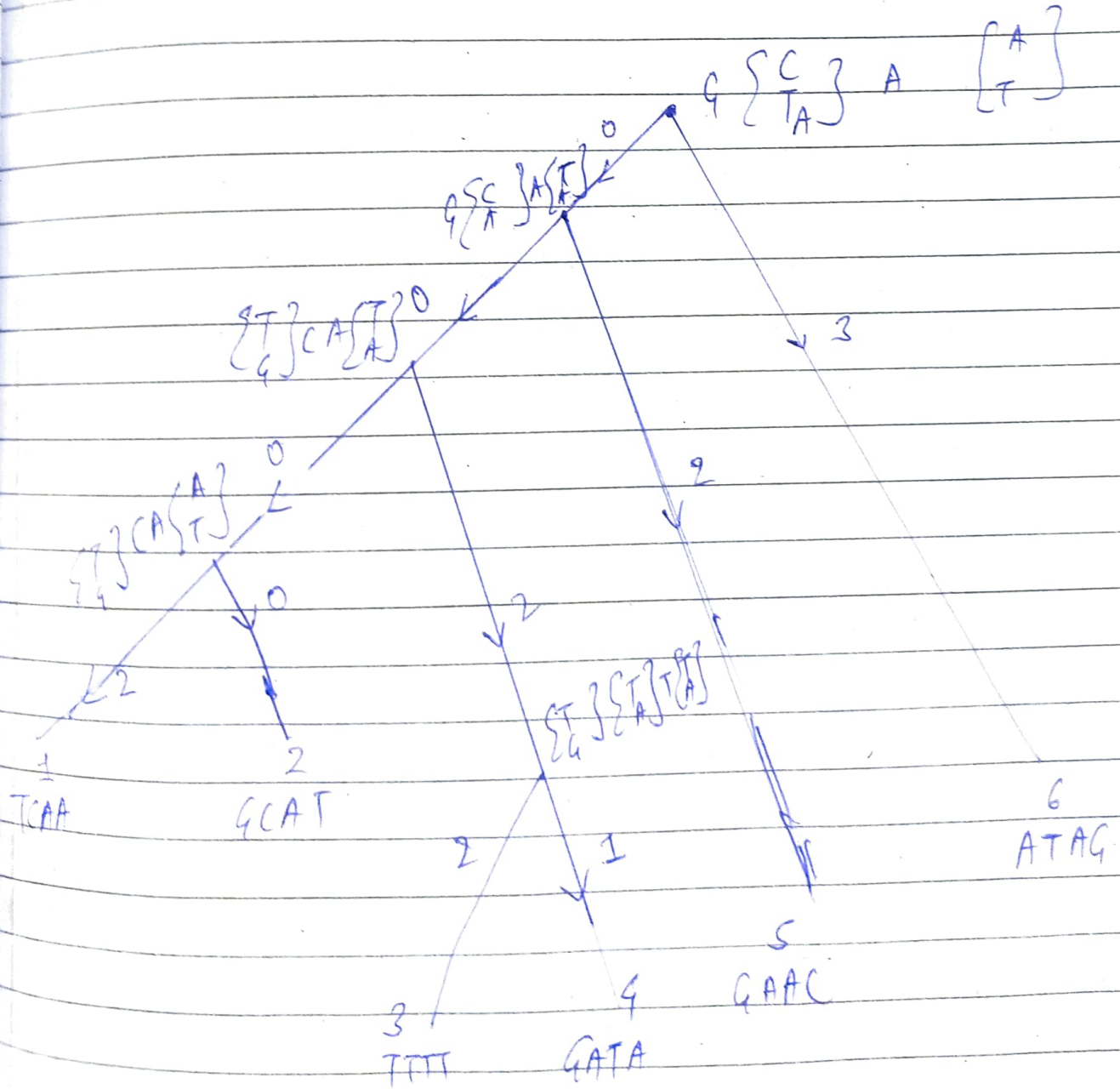
d) $\{T_1, T_2, T_3, T_4, T_5, T_6\}$

e) T_4, T_6

(Q7) MSA of 6 species

site :

Species	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



Genius
Page:
Date: / /

(Q4) We have to align N sequences of length 50, and knowing that a single pairwise comparison takes 1 second & the alignment of 4 of those sequences takes 10^4 seconds, & we have 5 billion years.

5×10^9 years = 15768×10^{13} seconds

$\therefore (2L)^{N-2} = 10^{2N-4} = 15768 \times 10^{13}$

take log,

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

$\therefore N = \left[\frac{\log(15768 \times 10^{13}) + 4}{2}\right]$

$= [10.5988] = 10$ sequences

$\therefore$ In 5 billion yrs, we can align 10 sequences.