Rubric for ABIN Mid-Sem

Q1.

a. What are the algorithmic strategies (Greedy, Divide & Conquer, DP etc.)? Briefly explain the reason for each. (Each carry 1 point with explanation)

Ans –

Gibbs' sampling for motif discovery- Greedy

an iterative procedure that discards one l-mer after each iteration and replaces it with a new one. Gibbs Sampling proceeds more slowly and chooses new l-mers at random, increasing the odds that it will converge to the correct solution.

Pattern branching- Greedy

Sequentially evaluate the best neighbours.
For any l-mer- 1st find best 1-hop neighbour N-1
Explore 1-hop neighbourhood of N-1 and find the best neighbour N-2
We need to define a sorting scheme for evaluating every candidate solution.

Smith Waterman Algorithm-

It follows DP approach -Top down or Recursive

b. Analyze the time complexity of a brute force motif search algorithm (2 points)

Ans –

The brute force algorithm determines the consensus string by determining which set of starting positions produces the best DNA score. It is an exhaustive search because it checks every possible set of starting positions

Varying $(n - l + 1)$ positions in each of t sequences, we 're looking at $(n - l + 1)^t$ sets of starting positions,

Given t sequences of length n each, l the length of the motif, using a Brute force approach the complexity is $l\,(n - l + 1)^t = O\,(l\,n^t)$.

Q2.

a. Map the following use-cases to local/global alignment. [2 points]

- **Short read alignment**: Local
- **Phylogeny analysis**: Local/Global
- **Identifying functional domain within protein sequences**: Local
- **Comparing proteins sequences from different species**: Global

b. Write down the generic steps of a Gibbs' sampler (not in the context of motif search).

Gibbs sampling

Suppose $\theta_1, \theta_2 \sim p(\theta_1, \theta_2)$ and we can sample from $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$. Beginning with an initial value $\left(\theta_1^{(0)}, \theta_2^{(0)}\right)$, the Gibbs sampler is

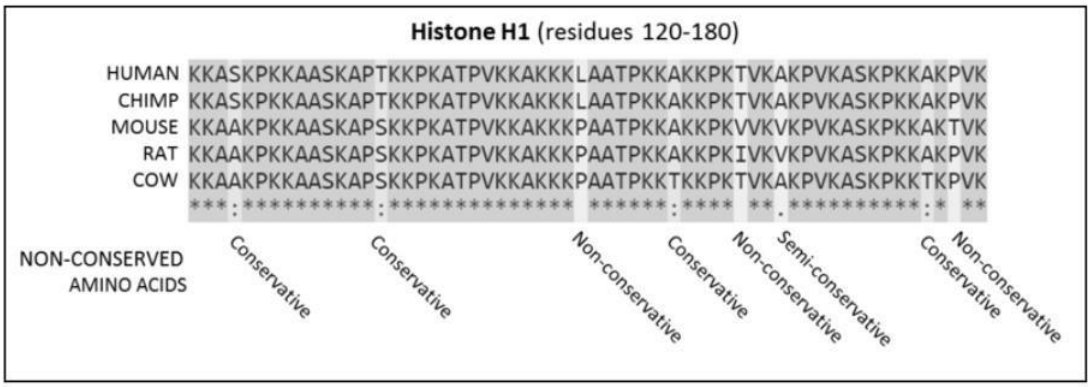
1. sample $\theta_1^{(j)} \sim p\left(\theta_1|\theta_2^{(j-1)}\right)$ and then
2. sample $\theta_2^{(j)} \sim p\left(\theta_2|\theta_1^{(j)}\right)$.

Notes:

- $\theta^{(j)} \xrightarrow{d} \theta$ where $\theta = (\theta_1, \theta_2)' \sim p(\theta_1, \theta_2)$.

Q3. Answer the following –

a. What do you understand from the below figure? (2.5points)



Ans. This is a figure of Multiple sequence alignment of mammalian histone proteins. Sequences are the amino acids for residues 120-180 of the proteins.

As a whole, the protein is totally conserved sequences across species, though it does have some non-conserved and semi-conserved amino acid sequences.

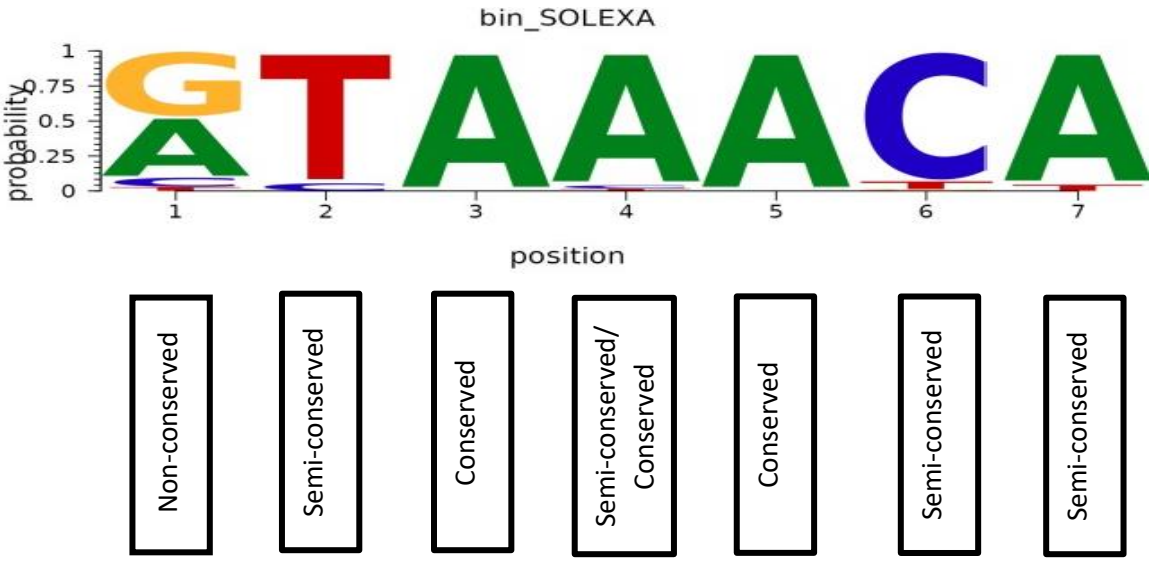
Below the protein sequences are symbols denoting the degree of conservation observed in each column –

- An * (asterisk) indicates positions which have a single, fully conserved residue (Conserved Sequence or identical).
- A colon (:) indicates that some of the sequences have different amino acids at that position, but that the chemical properties of the different amino acids are pretty similar. When an amino acid is substituted by a very similar amino acid at a position in an alignment, we call it a conservative substitution.
- A period (.) indicates a semi-conservative substitution, which is somewhere between a conservative and nonconservative substitution. This means that the chemical properties of the amino acids at that position are similar, but not that similar.
- A( ) indicates non-conservation indicating that the amino acids are very different at that position (Non-conservative mutation).

b. How do you interpret the following logo? (2.5points)

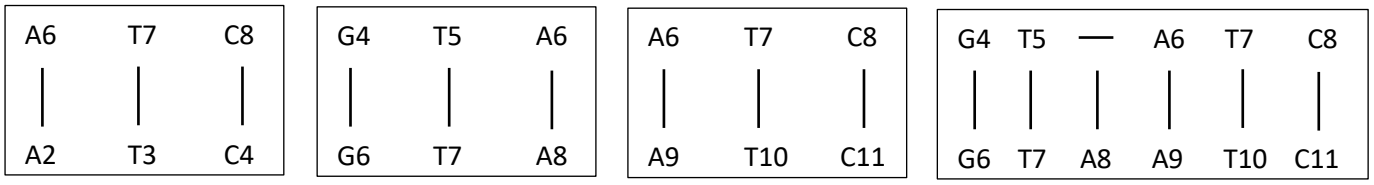
Ans.

Here, at position,



Q4. Report ALL alternative local alignments (following Smith Waterman algorithm) from the following alignment table considering the following – Match 1, Mismatch -1 and Gap – 2. No need to write the algorithm. (5 points)

Ans.



Or

A T C G T A A T C G T A - T C
A T C G T A A T C G T A A T C

Q5. State the pattern branching algorithm for motif discovery and analyze its time complexity.

(3 +2 points)

Ans-

Input: – n sequences of length m each.

Output: – Motif M, of length l – Variants of interest have a hamming distance of d from M

It is a local searching algorithm.

There are a total of $n(m - l + 1)$ l-mers in the input. Each of these l-mer has ${}^{l}C_{d}\, 3^{d}$ neighbours. For each such neighbor, a score can be computed. Having computed the scores of all of these $n(m - l + 1)({}^{l}C_{d}\, 3^{d})$ neighbours, the best scoring neighbor is output.

Pseudocode:

Algorithm PatternBranching(S, l, d)

```
Let M be an arbitrary l-mer;
for each l-mer u in S do
    for j := 0 to d do
        if d(u_j, S) < d(M, S) then M := u_j;
        u_{j+1} :=BestNeighbor(u_j);
output M;
```

Time complexity:

Assume there are t sequences, each sequence of a fixed length N

l-mer with l=10

from any of the t sequences of length N, we can generate (N-l+1) l-mers.

Therefore, from t sequences we can generate a total of (N-l+1) *t l-mers.

For finding neighbhors:

O(t(N-l+1) *(l d) *3^d)

If d=2

Then **t(N-l+1) *[$lC_1$3^0+ lC_2*3^1+ lC_3*3^2]**

(OR)

t(N-l+1) ((3l* (N-l+1)t*k)

O (t^2 N^2 l k)

Q6. For T = aataatct, provide the steps for Burrows–Wheeler transform i.e., BWT(T) and reversal of the same. [5 points]

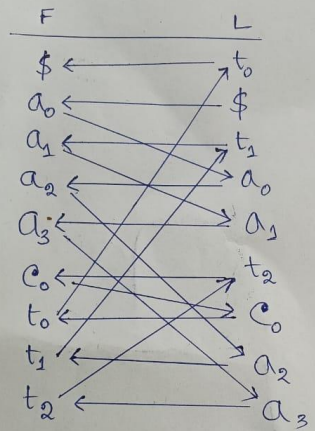
Ans -
str = **aataatct**

	rotation	sorted rotation
0	aataatct$	$aataatct
1	$aataatct	aataatct$
2	t$aataatc	aatct$aat
3	ct$aataat	ataatct$a
4	tct$aataa	atct$aata
5	atct$aata	ct$aataat
6	aatct$aat	t$aataatc
7	taatct$aa	taatct$aa
8	ataatct$a	tct$aataa

BWT(str) = **t$taatcaa**

Inverse BWT:

Inverse BWT:



Inverse BWT (str) = $a_0 a_2 t_1 a_1 a_3 t_2 c_0 t_0 \$$

= aataatct. (Answer)