CSE221 **Algorithms**

Assignment: Greedy and DP approach

There are two problems in this assignment.

---

1. DNAs contain genetic and hereditary information of almost every living organism. DNA Sequence Matching is a procedure to compare two DNA sequences and find similarity between them. Given below is the recurrence relation of a dynamic programming algorithm to perform sequence matching of two DNAs.

$$seq(i,j) = \max \begin{cases} seq(i, j-1) - 1 \ , \\ seq(i-1, j) + p(i, j) \ , \\ seq(i-1, j) - 1 \end{cases}$$

where,

$$seq(i,j) = 0, \quad if\ i = 0\ or\ j = 0$$
$$p(i,j) = 1, \quad if\ DNA1[i] = DNA2[j]$$
$$p(i,j) = -2, \quad otherwise$$

As you can see, the formula takes two indices, $i\ and\ j$ of two DNA sequences given as input. Then it produces a similarity score as output. We want to apply this formula to compare the following DNA sequences.

index	1	2	3	4	5
DNA1	G	C	G	T	A
DNA2	C	T	G	A	G

Now answer the following questions:

a. Put appropriate values to fill the gaps in the line below: **[marks: 1]**
$seq(\_\_\ ,\ \_\_)\ represents\ the\ final\ similarity\ score\ of\ the\ given\ DNAs.$

b. Apply the formula to calculate a similarity score between the DNA sequences given above. Show your work with either a recursion tree or a memory table. **[marks: 5]**

c. State the time complexity of this algorithm with proper reasoning.

You can assume that, the lengths of the DNA sequences are $N\ and\ M$ respectively. **[marks: 2]**

d. Mriaslun, a talented algorithmist, claims that he can implement the algorithm in an optimized way so that it only takes $O(min(N, M))$ space. Can you do it too? Explain your idea. **[marks: 1]**

2. Following are the codes generated from a text for a Huffman tree construction.

H - 1000	*u* - 000	*e* - 011
o - 1001	*d* - 001	*l* - 110
<*space*> - 1010	*n* - 010	*t* - 111
S - 1011		

You are also given the following information:

- The frequency of each leaf node except *e*, *l*, and *t* is 1.
- The left and right child nodes of the root have frequencies 5 and 8 respectively.

Now answer the following questions.

a) Suppose in a Huffman tree, the distances from the root to the pair of leaves denoting the letters *k* and *b* are 5 and 2 respectively, which letter between them is more frequent in the original text? Just **mention** the letter. **[marks: 1]**

b) **Draw** the Huffman tree from the given coding table above. **[marks: 3]**

c) Continuing on Q(b), what are the frequencies of *l*, and *t* in the original text? Just **mention** the frequencies. **[marks: 2]**

END!