BIO-321 ABIN Assignment 1 [50 Marks]

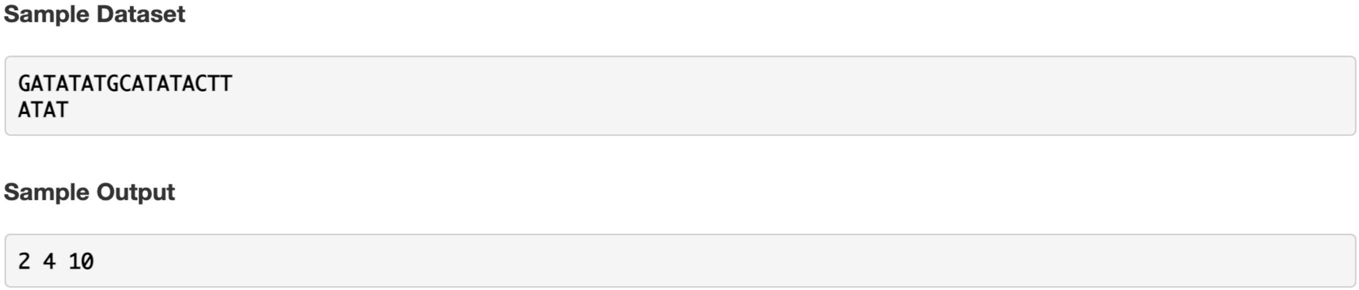
Plagiarism: All submitted codes are expected to be the result of your individual effort. You should never misrepresent someone else's work asyour own. In case of any plagiarism institute policy will be followed.

Instructions:
1) Allowed programming language is python.
2) Write clean code with proper comments at appropriate places as it will be checked.
3) You should submit a ZIP file consisting of the program and a pdf file listing the program and sample screenshots of its working. Name the ZIP file as: ABIN-A1-<Name>-<RollNo>. Store each problem with rollno_problemX.ipnyb, where X is the problem number.

Ques 1) Given two strings s and t, t is a substring of s if t is contained as a contiguous collection of symbols in s. The position of a symbol in a string is the total number of symbols found to its left, including itself (e.g., the positions of all occurrences of 'CU' in "AUGCUUCAGAAAGGUCUUACG" are 4, 16). The symbol at position i of s is denoted by s[i]. (10 Marks)

Given: Two DNA strings s and t.

Return: All locations of t as a substring of s



Sample Dataset

```
GATATATGCATATACTT
ATAT
```

Sample Output

```
2 4 10
```

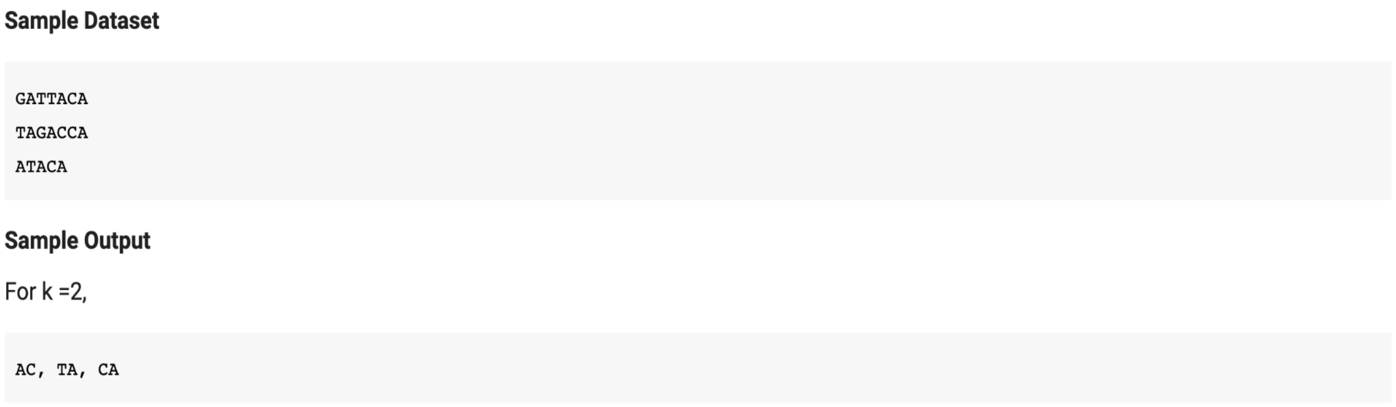
Ques 2) (10 Marks)

a) There are two sequences A and B of L nucleotides each. What is the probability to observe a common substring of at least k nucleotides between those two strings? Explain with the help of an example. (No need to provide the code for this question)

b) A common substring of a collection of strings is a substring of every member of the collection. We say that a common substring is a longest common substring if there does not exist a longer common substring. For example, "CG" is a common substring of "ACGTACGT" and "AACCGTATA", but it is not as long as possible; in this case, "CGTA" is a longest common substring of "ACGTACGT" and "AACCGTATA".

Given: A collection of DNA strings.

Return: A longest common substring of the collection of a given size.



Sample Dataset

```
GATTACA
TAGACCA
ATACA
```

Sample Output

For k =2,

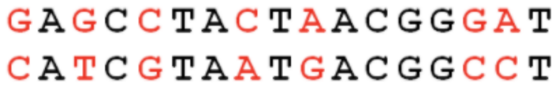
```
AC, TA, CA
```

Ques 3) Implement the Pattern Branching algorithm for motif finding. (10 Marks)

Ques 4) Given two strings s and t of equal length,the Hamming distance between s and t denoted by D(s,t) is the number of corresponding symbols that differ in s and t.

(10 Marks)

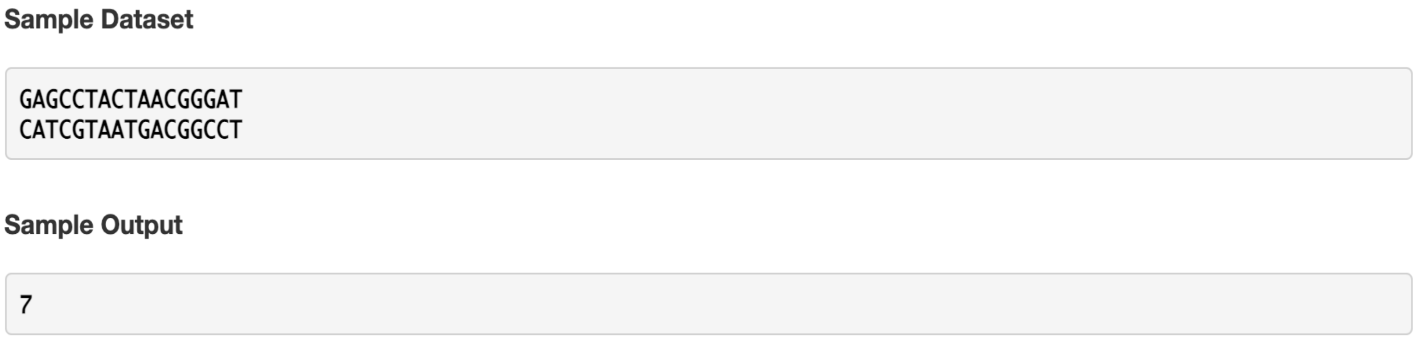
See the below figure for a more clear understanding of the problem.



```
GAGCCTACTAACGGGAT
CATCGTAATGACGGCCT
```

In the figure,the Hamming distance between the two strings is 7. Mismatched symbols are coloured red.

Given: Two DNA strings s and t of equal length.
Return: The Hamming distance D(s,t)



Sample Dataset

```
GAGCCTACTAACGGGAT
CATCGTAATGACGGCCT
```

Sample Output

```
7
```

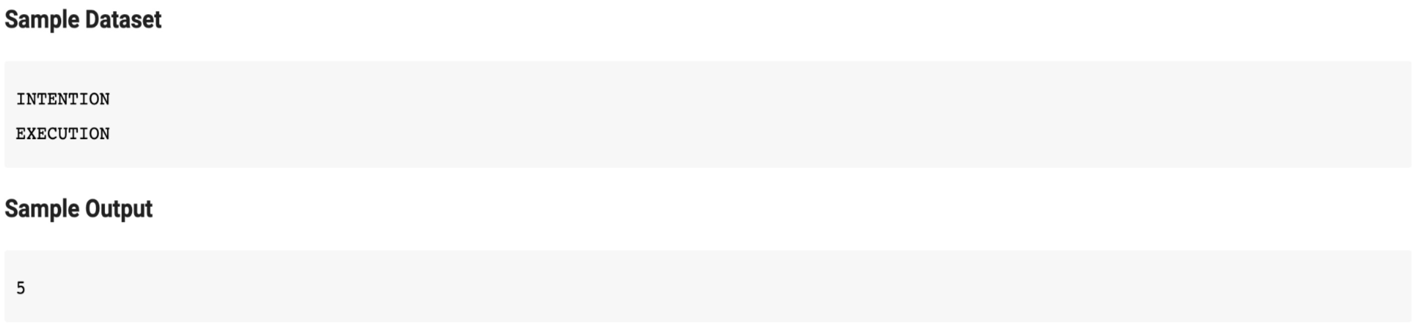
Ques 5) Given two strings s and t (of possibly different lengths), the edit distance d(s,t) is the minimum number of edit operations needed to transform s into t where an edit operation is defined as the substitution, insertion, or deletion of a single symbol. All of the operations are of equal cost.

Example:
Input: str1 = "sunday", str2 = "saturday"
Output: 3
Last three and first characters are the same. We basically need to convert "un" to "atur". This can be done using three operations: Replace 'n' with 'r', insert t, insert a

(10 Marks)

Given: Two protein strings s and t.

Return: The edit distance d(s,t).



Sample Dataset

```
INTENTION
EXECUTION
```

Sample Output

```
5
```