

Bioinformatics Lab Assignment 2

ON

Longest Common Subsequence

Submitted by

Amit Kumar

BT18CSE073

Code in Python

```
from difflib import SequenceMatcher

str__a = open("C:/Users/ak781/Desktop/BIO Info/lab2/str1.txt")
str__b = open("C:/Users/ak781/Desktop/BIO Info/lab2/str2.txt")
str_a = str__a.read()
str_b = str__b.read()
# print(str_a)
# print(str_b)
# str_a = "aaccttgg"
# str_b = "acactgtga"

def lcs(s1, s2):
    matrix = [["" for x in range(len(s2))] for x in range(len(s1))]
    for i in range(len(s1)):
        for j in range(len(s2)):
            if s1[i] == s2[j]:
                if i == 0 or j == 0:
                    matrix[i][j] = s1[i]
                else:
                    matrix[i][j] = matrix[i-1][j-1] + s1[i]
            else:
                matrix[i][j] = max(matrix[i-1][j], matrix[i][j-1],
key=len)
        cs = matrix[-1][-1]
        return len(cs), cs

print(lcs(str_a, str_b))
```

Advantages and disadvantages

- Developed problem solving skills
- Learned about dynamic programming
- Coming with the exact approach is time taken
- Coding the approach is also tough
- Learning the approach and dynamic programming is time consuming

Code in R

```
library(qualV,stringi,readr)
# str__a = "aaccttg"
# str__b = "acactgtga"
str__a = read_file("C:/Users/ak781/Desktop/BIO Info/lab2/str1.txt")
str__b = read_file("C:/Users/ak781/Desktop/BIO Info/lab2/str2.txt")
# str3 = read_file("C:/Users/ak781/Desktop/BIO Info/lab2/str3.txt")
print(nchar(str__a))
print(nchar(str__b))
str_a = strsplit(str__a, split=NULL)[[1]]
str_b = strsplit(str__b, split=NULL)[[1]]
result = (LCS(str_a,str_b))
str = result[4]
typeof(str)
str <- stri_join_list(str,sep="")
print(str)
print(nchar(str))
sink("C:/Users/ak781/Desktop/BIO Info/lab2/answer.txt")
cat(str)
sink()
# print(nchar(str3))
file.show("answer.txt")
```

Advantages and disadvantages

- Large Variety of Libraries
- Learned about File I/O operations in R
- Due to inbuilt Libraries we don't need to write the whole algo
- Learned about the functions and types of parameter to include in the LCS
- Finding the exact packages takes is time consuming in the R documentation
- Limited resources available online