Finding a Motif in DNA

Orr Shomroni

29 September, 2022

Problem

Given two strings $\boldsymbol{s}$ and $\boldsymbol{t}$, $\boldsymbol{t}$ is a **substring** of $\boldsymbol{s}$ if $\boldsymbol{t}$ is contained as a contiguous collection of symbols in $\boldsymbol{s}$ (as a result, $\boldsymbol{t}$ must be no longer than $\boldsymbol{s}$).

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 ‘U’ in “AUGCUUCAGAAAGGUCUUACG” are 2, 5, 6, 15, 17, and 18). The symbol at position $\boldsymbol{i}$ of $\boldsymbol{s}$ is denoted by $\boldsymbol{s}[\boldsymbol{i}]$.

A substring of $\boldsymbol{s}$ can be represented as $s[j:k]$, where $\boldsymbol{j}$ and $\boldsymbol{k}$ represent the starting and ending positions of the substring in $\boldsymbol{s}$; for example, if $\boldsymbol{s}$ = “AUGCUUCAGAAAGGUCUUACG”, then $s[2:5]$ = “UGCU”.

The **location** of a substring $s[j:k]$ is its beginning position $\boldsymbol{j}$; note that $\boldsymbol{t}$ will have multiple locations in $\boldsymbol{s}$ if it occurs more than once as a substring of $\boldsymbol{s}$ (see the Sample below).

Given: Two DNA strings $\boldsymbol{s}$ and $\boldsymbol{t}$ (each of length at most 1 kbp).

Return: All locations of $\boldsymbol{t}$ as a substring of $\boldsymbol{s}$.

Sample Dataset

```
GATATATGCATATACTT

ATAT
```

Sample Output

```
2 4 10
```

```
f=open("rosalind_subs.txt",'r')
s=f.readlines()[0].replace("\n","")
f=open("rosalind_subs.txt",'r')
t=f.readlines()[1].replace("\n","")
if len(s)<len(t):
  print("Error: length of substring ",t," is longer than ",s)
  exit()

motifLoci=[]
for i in range(0,len(s)-len(t)):
  tmpText=s[i:i+len(t)]
  if tmpText==t:
    motifLoci.append(i+1)

print("String: "+s)
```

```
## String: CATCTTGGCATCTTGGTTATGACTCATCTTGCATCTTGTGACATCATCTTGCCATCTTGGCATCTTGCATCTTGACATCTTGACATCTTGAG
```

```
print("Substring: "+t)
```

```
## Substring: CATCTTGCA
```

```
print("Locations of motifs: "+str(motifLoci))
```

```
## Locations of motifs: [25, 61, 127, 159, 184, 191, 198, 205, 305, 312, 374, 381, 507, 514, 672, 687,
```