Aluno: Felipe Colombelli          Cartão: 00287698

Nota inicial: foi utilizado Python 3 - com auxílio da biblioteca Bio (para abertura do arquivo .fasta) - para a resolução da lista de exercícios.

## Questão 1-a:

### Pseudocódigo
S: sequência do cromosso 7

s': sequência mutada dada no enunciado do problema

D: dicionário contendo todas as sequências possíveis mudando apenas um nucleotídeo de s'

```
1. Abrir arquivo com S no formato string
2. Para cada nucleotideo de s':
3.     Para cada nucleotideo existente:
4.         Gerar uma sequência nova
5.         Adicionar como chave de D associando o valor [indice
   do nucleotideo, 0]
6. Para cada chave em D:
7.     Contar as ocorrências (oc) da chave
8.     Trocar o valor de [indice, 0] para [indice, oc]
9. Para cada chave em D:
10.     Retirar todas as chaves cujo oc seja diferente de 1
11.     Verificar com indice qual foi o nucleotideo alterado
```

Considerações: foram encontradas mais de uma subsequência que, mudando apenas um nucleotídeo da sequência mutada dada no enunciado, fosse única na sequência orginal do cromossomo 7.

### Resultados:

Original  given subsequence: CAATTGAATAATTG

----------------------------------------------------------------

Unique sequence: CAATAGAATAATTG

ith position that changed: 5

Starting index position to find the subsequence in the original chromosome 7 sequence: 20843949

Original (mutated) given nucleotide: T

Changed (theoretical unique and non-mutated) nucleotide: A

----------------------------------------------------------------

Unique sequence: CAATTAAATAATTG

ith position that changed: 6
Starting index position to find the subsequence in the original chromosome 7 sequence: 35769315
Original (mutated) given nucleotide: G
Changed (theoretical unique and non-mutated) nucleotide: A
----------------------------------------------------------------

Unique sequence: CAATTCAATAATTG
ith position that changed: 6
Starting index position to find the subsequence in the original chromosome 7 sequence: 48205636
Original (mutated) given nucleotide: G
Changed (theoretical unique and non-mutated) nucleotide: C
----------------------------------------------------------------

Unique sequence: CAATTTAATAATTG
ith position that changed: 6
Starting index position to find the subsequence in the original chromosome 7 sequence: 47003091
Original (mutated) given nucleotide: G
Changed (theoretical unique and non-mutated) nucleotide: T
----------------------------------------------------------------

Unique sequence: CAATTGAATGATTG
ith position that changed: 10
Starting index position to find the subsequence in the original chromosome 7 sequence: 3630527
Original (mutated) given nucleotide: A
Changed (theoretical unique and non-mutated) nucleotide: G
----------------------------------------------------------------

Unique sequence: CAATTGAATATTTG
ith position that changed: 11
Starting index position to find the subsequence in the original chromosome 7 sequence: 58013515
Original (mutated) given nucleotide: A
Changed (theoretical unique and non-mutated) nucleotide: T
----------------------------------------------------------------

Unique sequence: CAATTGAATAACTG
ith position that changed: 12
Starting index position to find the subsequence in the original chromosome 7 sequence: 36409919
Original (mutated) given nucleotide: T
Changed (theoretical unique and non-mutated) nucleotide: C
----------------------------------------------------------------

Unique sequence: CAATTGAATAATTT
ith position that changed: 14
Starting index position to find the subsequence in the original chromosome 7 sequence: 8010831
Original (mutated) given nucleotide: G
Changed (theoretical unique and non-mutated) nucleotide: T

## Questão 1-b:

### Pseudocódigo
S: sequência do cromosso 7

```
1. Abrir arquivo com S no formato string
2. Funcao foo_conta(tam_pal):
3.     Numero de palindromos n encontrados recebe 0
4.     Indice icial i recebe 0
5.     Indice final j recebe tam_pal
6.     Enquanto j <= tamanho de S:
7.         Sequencia atual s' recebe S de i a j
8.         Palindromo atual p' recebe s' ao contrario
9.         Se s' = p' aumenta 1 em n
10.    Retorna n

1. foo_conta(11)
2. foo_conta(9)
```

Considerações: Existem muitos palíndromos de tamanho 11 e 9 que englobam o nucleotideo inválido N, presente no arquivo da sequência disponibilizado, esse fato foi desconsiderado e as contagens finais permaneceram considerando os palíndromos com N (assim como foi previamente debatido com o senhor no final de uma aula).

### Resultados:

Palindromes of size 11: 363615
Palindromes of size 9: 720286

**Questão 1-c:**

**Pseudocódigo**
S: sequência do cromosso 7
D: dicionário com as subsequências de tamanho 37 e sua respectiva contagem.

```
1. Abrir arquivo com S no formato string
2. Indice inicial i recebe 0
3. Indice final j recebe 37
4. Equanto j <= tamanho de S
5.     Se a string s' - S no intervalo [i,j) - está em D:
6.         Aumenta em 1 o valor associado à chave s' de D
7.     Senão:
8.         Adiciona s' como chave e associa o valor 1
9.     i recebe i+1
10.    j recebe j+1
```

**Resultados:**

56024904 different subsequences of size 37 were found.

The first 100 sequences with their respective count:
Sequence: CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC Count: 11
Sequence: TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT Count: 10
Sequence: AACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA Count: 11
Sequence: ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA Count: 11
Sequence: CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC Count: 11
Sequence: CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACC Count: 10
Sequence: CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACA Count: 1
Sequence: CTAACCCTAACCCTAACCCTAACCCTAACCCTAACAA Count: 1
Sequence: TAACCCTAACCCTAACCCTAACCCTAACCCTAACAAC Count: 1
Sequence: AACCCTAACCCTAACCCTAACCCTAACCCTAACAACC Count: 2
Sequence: ACCCTAACCCTAACCCTAACCCTAACCCTAACAACCC Count: 2
Sequence: CCCTAACCCTAACCCTAACCCTAACCCTAACAACCCT Count: 2
Sequence: CCTAACCCTAACCCTAACCCTAACCCTAACAACCCTA Count: 2
Sequence: CTAACCCTAACCCTAACCCTAACCCTAACAACCCTAA Count: 2
Sequence: TAACCCTAACCCTAACCCTAACCCTAACAACCCTAAC Count: 2
Sequence: AACCCTAACCCTAACCCTAACCCTAACAACCCTAACC Count: 2
Sequence: ACCCTAACCCTAACCCTAACCCTAACAACCCTAACCC Count: 2
Sequence: CCCTAACCCTAACCCTAACCCTAACAACCCTAACCCT Count: 2
Sequence: CCTAACCCTAACCCTAACCCTAACAACCCTAACCCTA Count: 3
Sequence: CTAACCCTAACCCTAACCCTAACAACCCTAACCCTAA Count: 3
Sequence: TAACCCTAACCCTAACCCTAACAACCCTAACCCTAAC Count: 3
Sequence: AACCCTAACCCTAACCCTAACAACCCTAACCCTAACC Count: 2
Sequence: ACCCTAACCCTAACCCTAACAACCCTAACCCTAACCC Count: 2

Sequence: CCCTAACCCTAACCCTAACAACCCTAACCCTAACCCT Count: 2
Sequence: CCTAACCCTAACCCTAACAACCCTAACCCTAACCCTA Count: 2
Sequence: CTAACCCTAACCCTAACAACCCTAACCCTAACCCTAA Count: 2
Sequence: TAACCCTAACCCTAACAACCCTAACCCTAACCCTAAC Count: 2
Sequence: AACCCTAACCCTAACAACCCTAACCCTAACCCTAACC Count: 3
Sequence: ACCCTAACCCTAACAACCCTAACCCTAACCCTAACCC Count: 3
Sequence: CCCTAACCCTAACAACCCTAACCCTAACCCTAACCCT Count: 3
Sequence: CCTAACCCTAACAACCCTAACCCTAACCCTAACCCTA Count: 3
Sequence: CTAACCCTAACAACCCTAACCCTAACCCTAACCCTAA Count: 3
Sequence: TAACCCTAACAACCCTAACCCTAACCCTAACCCTAAC Count: 3
Sequence: AACCCTAACAACCCTAACCCTAACCCTAACCCTAACC Count: 3
Sequence: ACCCTAACAACCCTAACCCTAACCCTAACCCTAACCC Count: 3
Sequence: CCCTAACAACCCTAACCCTAACCCTAACCCTAACCCT Count: 3
Sequence: CCTAACAACCCTAACCCTAACCCTAACCCTAACCCTA Count: 3
Sequence: CTAACAACCCTAACCCTAACCCTAACCCTAACCCTAA Count: 3
Sequence: TAACAACCCTAACCCTAACCCTAACCCTAACCCTAAA Count: 1
Sequence: AACAACCCTAACCCTAACCCTAACCCTAACCCTAAAC Count: 1
Sequence: ACAACCCTAACCCTAACCCTAACCCTAACCCTAAACC Count: 1
Sequence: CAACCCTAACCCTAACCCTAACCCTAACCCTAAACCT Count: 1
Sequence: AACCCTAACCCTAACCCTAACCCTAACCCTAAACCTA Count: 1
Sequence: ACCCTAACCCTAACCCTAACCCTAACCCTAAACCTAA Count: 1
Sequence: CCCTAACCCTAACCCTAACCCTAACCCTAAACCTAAC Count: 1
Sequence: CCTAACCCTAACCCTAACCCTAACCCTAAACCTAACC Count: 1
Sequence: CTAACCCTAACCCTAACCCTAACCCTAAACCTAACCC Count: 1
Sequence: TAACCCTAACCCTAACCCTAACCCTAAACCTAACCCT Count: 1
Sequence: AACCCTAACCCTAACCCTAACCCTAAACCTAACCCTA Count: 1
Sequence: ACCCTAACCCTAACCCTAACCCTAAACCTAACCCTAA Count: 1
Sequence: CCCTAACCCTAACCCTAACCCTAAACCTAACCCTAAC Count: 1
Sequence: CCTAACCCTAACCCTAACCCTAAACCTAACCCTAACC Count: 1
Sequence: CTAACCCTAACCCTAACCCTAAACCTAACCCTAACCC Count: 1
Sequence: TAACCCTAACCCTAACCCTAAACCTAACCCTAACCCT Count: 1
Sequence: AACCCTAACCCTAACCCTAAACCTAACCCTAACCCTA Count: 1
Sequence: ACCCTAACCCTAACCCTAAACCTAACCCTAACCCTAA Count: 1
Sequence: CCCTAACCCTAACCCTAAACCTAACCCTAACCCTAAC Count: 1
Sequence: CCTAACCCTAACCCTAAACCTAACCCTAACCCTAACC Count: 1
Sequence: CTAACCCTAACCCTAAACCTAACCCTAACCCTAACCC Count: 1
Sequence: TAACCCTAACCCTAAACCTAACCCTAACCCTAACCCT Count: 1
Sequence: AACCCTAACCCTAAACCTAACCCTAACCCTAACCCTA Count: 1
Sequence: ACCCTAACCCTAAACCTAACCCTAACCCTAACCCTAA Count: 1
Sequence: CCCTAACCCTAAACCTAACCCTAACCCTAACCCTAAC Count: 1
Sequence: CCTAACCCTAAACCTAACCCTAACCCTAACCCTAACA Count: 1
Sequence: CTAACCCTAAACCTAACCCTAACCCTAACCCTAACAA Count: 1
Sequence: TAACCCTAAACCTAACCCTAACCCTAACCCTAACAAC Count: 1
Sequence: AACCCTAAACCTAACCCTAACCCTAACCCTAACAACC Count: 1
Sequence: ACCCTAAACCTAACCCTAACCCTAACCCTAACAACCC Count: 1
Sequence: CCCTAAACCTAACCCTAACCCTAACCCTAACAACCCT Count: 1

Sequence: CCTAAACCTAACCCTAACCCTAACCCTAACAACCCTA Count: 1
Sequence: CTAAACCTAACCCTAACCCTAACCCTAACAACCCTAA Count: 1
Sequence: TAAACCTAACCCTAACCCTAACCCTAACAACCCTAAC Count: 1
Sequence: AAACCTAACCCTAACCCTAACCCTAACAACCCTAACC Count: 1
Sequence: AACCTAACCCTAACCCTAACCCTAACAACCCTAACCC Count: 1
Sequence: ACCTAACCCTAACCCTAACCCTAACAACCCTAACCCT Count: 1
Sequence: TAACAACCCTAACCCTAACCCTAACCCTAACCCTAAC Count: 2
Sequence: AACAACCCTAACCCTAACCCTAACCCTAACCCTAACA Count: 1
Sequence: ACAACCCTAACCCTAACCCTAACCCTAACCCTAACAA Count: 1
Sequence: CAACCCTAACCCTAACCCTAACCCTAACCCTAACAAC Count: 1
Sequence: AACCCTAACCCTAACAACCCTAACCCTAACA Count: 1
Sequence: ACCCTAACCCTAACAACCCTAACCCTAACAA Count: 1
Sequence: CCCTAACCCTAACAACCCTAACCCTAACAAC Count: 1
Sequence: CCTAACCCTAACAACCCTAACCCTAACAACC Count: 1
Sequence: CTAACCCTAACAACCCTAACCCTAACAACCC Count: 1
Sequence: TAACCCTAACAACCCTAACCCTAACAACCCT Count: 1
Sequence: AACCCTAACAACCCTAACCCTAACAACCCTA Count: 1
Sequence: ACCCTAACAACCCTAACCCTAACAACCCTAA Count: 1
Sequence: CCCTAACAACCCTAACCCTAACAACCCTAAC Count: 1
Sequence: CCTAACCCTAACAACCCTAACCCTAACAACCCTAACC Count: 1
Sequence: CTAACCCTAACAACCCTAACCCTAACAACCCTAACCC Count: 1
Sequence: TAACCCTAACAACCCTAACCCTAACAACCCTAACCCT Count: 1
Sequence: AACCCTAACAACCCTAACCCTAACAACCCTAACCCTA Count: 1
Sequence: ACCCTAACAACCCTAACCCTAACAACCCTAACCCTAA Count: 1
Sequence: CCCTAACAACCCTAACCCTAACAACCCTAACCCTAAC Count: 1
Sequence: CCTAACAACCCTAACCCTAACAACCCTAACCCTAACC Count: 1
Sequence: CTAACAACCCTAACCCTAACAACCCTAACCCTAACCC Count: 1
Sequence: TAACAACCCTAACCCTAACAACCCTAACCCTAACCCT Count: 1
Sequence: AACAACCCTAACCCTAACAACCCTAACCCTAACCCTA Count: 1
Sequence: ACAACCCTAACCCTAACAACCCTAACCCTAACCCTAA Count: 1
Sequence: CAACCCTAACCCTAACAACCCTAACCCTAACCCTAAC Count: 1

## Questão 1-d:

### Pseudocódigo
S: sequência do cromosso 7

D: dicionário onde a chave é o nucleotídeo ou caracter "intruso" e o valor associado é o número de vezes que aparece na sequência.

```
1. Abrir arquivo com S no formato string
2. Para cada caractere c em S:
3.     Se existe em D:
4.         Aumenta 1 na contagem
5.     Se não existe:
6.         Adiciona c como chave em D associando o valor 1
```

**Resultados:**

C - amount found: 11918693
T - amount found: 17132531
A - amount found: 17146584
G - amount found: 11909449
Non-nucleotide found: N - amount found: 2396


## Questão 1-e:

**Pseudocódigo**
S: sequência do cromosso 7
s': string inicialmente fazia correspondente ao cDNA do cromossomo 7

```
1. Para cada caractere c em S:
2.     Se c é A:
3.         Concatena s' com T
4.     Se c é T:
5.         Concatena s' com A
6.     Se c é C:
7.         Concatena s' com G
8.     Se c é G:
9.         Concatena s' com C
10.    Se c não é nenhum dos acima (else):
11.         Concatena s' com c
```

Considerações: Os caracteres encontrados que não correspondem a nenhum nucleotideo conhecido, foram adicionados inalterados à fita de DNA complementar final.


**Resultados:**

The first 100 given nucleotides:

CTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACAACCCTAACCCTAACCC

The first 100 nucleotides of the cDNA:

GATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATT
GGGATTGGGATTGGGATTGGGATTGTTGGGATTGGGATTGGG