Word composition X95544_en

A nucleic acid or amino acid sequence can be seen as composed of a number of possibly overlapping k-mers or words of length k, for a certain $k \geq 1$. The k-mer composition of a sequence is given by the frequency with which each possible k-mer occurs within the sequence. The 1-mer composition is related to the GC content of a DNA sequence, and the 2-mer, 3-mer, and 4-mer compositions are also known as the di-nucleotide, tri-nucleotide, and tetra-nucleotide compositions of a DNA sequence. For example, the di-nucleotide composition of TATAAT is given by one occurrence of AA, two ocurrences of AT, and two ocurrences of TA.

Write pseudocode, Python code, and C++ code for the word composition problem. The program must implement and use the word composition function in the pseudocode, which must be iterative and is not allowed to perform input/output operations. Make two submissions, including the pseudocode as a comment to both the Python and the C++ code.

Input

The input is a string s (a genomic sequence) over the alphabet $\Sigma = \{A, C, G, T\}$ and an integer k with $1 \leq k \leq \|s\|$.

Output

The output is a sorted list of all the k-mers of s and their frequencies.

Sample input 1

```
TATAAT
1
```

Sample output 1

```
A 3
T 3
```

Sample input 2

```
TATAAT
2
```

Sample output 2

```
AA 1
AT 2
TA 2
```

Sample input 3

```
TATAAT
3
```

Sample output 3

```
AAT 1
ATA 1
TAA 1
TAT 1
```

Sample input 4

```
TATAAT
4
```

Sample output 4

```
ATAA 1
TAAT 1
TATA 1
```

Sample input 5

```
TATAAT
5
```

Sample output 5

```
ATAAT 1
TATAA 1
```

Sample input 6

```
TATAAT
6
```

Sample output 6

```
TATAAT 1
```

Problem information

Author : Gabriel Valiente
Generation : 2021-10-24 10:30:06

© *Jutge.org*, 2006–2021.
https://jutge.org