Assignment 2

Biological Sequence Analysis (UE)
WS 2025/26

3 Exercise (12 points)

Generally, a dot plot is a graphical method that can be used to compare two sequences. It allows to identify regions of close similarity, as exemplarily depicted below.

	B	I	O	I	N	F	O	R	M	A	T	I	C	S
B	●													
O			●				●							
I		●		●								●		
L														
I		●		●								●		
N					●									
G														
F						●								
O			●				●							
R								●						
M									●					
A										●				
N					●									
I		●		●								●		
C													●	
S														●

Tabelle 1: Dot plot comparing the two sequences `BIOINFORMATICS` and `BOILINGFORMANICS`, whereas both are of a non-biological kind (for demonstration purposes). Normally, especially in the setting of bioinformatics, the sequences to be compared are biological ones (e.g. DNA).

Write a Python program that takes two strings (not necessarily of the same length) specified by the command line and prints the resulting dot plot to the standard output. Your program should accept a *window size* as the third command line argument. The command line interface should be specified as follows (please adapt the file name accordingly):

```
$ ./ex03_0123456.py <sequence_one> <sequence_two> <window_size>
```

If all characters inside the window match between the two sequences, a *dot* (or a similar symbol of the default `ASCII` table) should be printed at the position of the character in the centre of the window. Therefore only *odd* integers should be accepted as the window size. Consequently, a window size of 1 corresponds to a simple dot plot as depicted in Tabelle 1. The example shown in Tabelle 2 illustrates a dot plot with a window size of 3 (command invocation below).

```
$ ./ex03_0123456.py bioinformatics boilingformanics 3
```

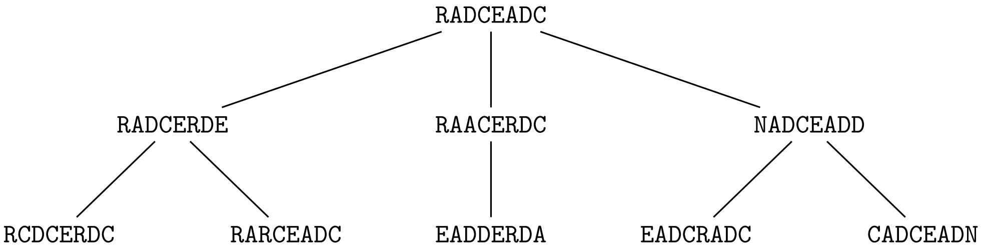
Document and test your program in detail and provide checks for the command line arguments. Is a window size larger than 1 even meaningful? How does the output change with the window size and which interpretations can we derive? Argument and discuss your answer.

	B	I	O	I	N	F	O	R	M	A	T	I	C	S
B														
O														
I														
L														
I														
N														
G														
F														
O							●							
R								●						
M									●					
A														
N														
I														
C													●	
S														

Tabelle 2: Dot plot comparing the two sequences `BIOINFORMATICS` and `BOILINGFORMANICS`, whereas a window size of 3 is applied.

4 Exercise (12 points)

In the following, an imaginary phylogenetic tree comprising nine amino acid sequences over the alphabet {`A, C, D, E, N, R`} is specified. Compute (using pen and paper) the *PAM1* matrix as described in the lecture notes. Additionally, derive the corresponding scoring matrix.



How would you obtain a *PAM250* matrix from this *PAM1* matrix? What is the rationale behind this procedure? Is it theoretically (from a biological point of view) sound?

Submission Instructions

Submit a PDF file that contains your results and explanations plus Python code to obtain the results! You are also allowed to submit your Code as Jupyter Notebook (PDF and ipynb file).

Submission: electronically via Moodle:

`https://moodle.jku.at/`