Assignment 1

Biological Sequence Analysis (UE)
WS 2025/26

1 Exercise (10 points)

A palindrome is a word, phrase, number or other sequence of units that can be read the same way in either direction. E.g. the word `level`, the number `1234321`, the phrase `Step on no pets`.

Write a Python program, that reads a text file and searches for all palindromes in this file. The program should write all found palindromes (except phrases), together with their multiplicity to an output file. Handle all strings case insensitive, i.e. the word `Level` is also a palindrome. Both input and output files should be specified as command line arguments. Copy some arbitrary text (e.g. from the internet) and apply your program to it.

2 Exercise (12 points)

Write a Python program, that finds restriction sites in a DNA sequence. Restriction sites are positions where restriction enzymes cut the DNA. They are usually recognised by a short, specific sequence motif.

The respective sequences for the restriction enzymes *PpuMI*, *MspA1I* and *MslI* are defined as:

PpuMI	`RG^GWCCY`
MspA1I	`CMG^CKG`
MslI	`CAYNN^NNRTG`

Note: `K` stands for `{G, T}`, `M` for `{A, C}`, `N` for `{A, C, G, T}`, `R` for `{A, G}`, `W` for `{A, T}` and `Y` is short for `{C, T}`. The caret (^) indicates the cut site.

Given a file with DNA sequences apply regular expressions to look for all restriction sites of these three enzymes and print the position after the cutting site to an output file (e.g. the position of the `G` for *PpuMI*). Make sure that the names of the input and output files can be specified as command line arguments and exactly two command line arguments have been specified.

Use the UCSC Genome Browser (`https://genome.ucsc.edu/`) in order to download the DNA sequence from the human reference genome version hg38 of chromosome 22 band q13.1 (`chr22 bp 37200001-40600000`). The resulting file serves as the input of your program, whereas the header line should be skipped programmatically. Usual file extension is .fasta (`https://en.wikipedia.org/wiki/FASTA_format`)

Many restriction enzymes exhibit so called palindromic recognition sequences. Read up what *palindromic* means in the context of DNA sequences. Which of the three enzymes listed above exhibits such a palindromic recognition sequence (argument your choice)? What is the advantage of a palindromic recognition sequence?

Hint: The beginning of the downloaded file should look similar to this.

```
>hg38_dna range=chr22:37200001-40600000 5'pad=0 3'....
CTCTGGGGCCTGTTGAGCCAGCAGTTCCCCTGAGCAAATATTGACACATT
TGCTGGCCTTTAAAGCGGACAGGAGGGTGGAGAGGCCACATCCCAGCTCT
.....
```

Hint: For biologists the first base of a sequence is at position/index 1, not 0.

Submission Instructions

Submit a PDF file that contains your results and explanations plus Python code to obtain the results!

Submission: electronically via Moodle:

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