

5 Exercise (16 points)

Write a Python program that takes DNA sequences in 5'–3' direction from an input file in multi-FASTA format and computes the corresponding amino acid sequences they may encode (for simplicity, do not take splicing into account). To do so, perform the following steps for each DNA sequence:

1. Check if the string is a valid DNA sequence.
2. Perform transcription into an mRNA sequence (consider the sequences from the input file as template strands and bear in mind in which direction transcription takes place).
3. Perform translation into amino acid sequences (1-letter notation) by applying the genetic code to all sub-sequences delimited by a start and a stop codon (in 5'–3' direction).

Group the above steps in functions with meaningful interfaces. The program should read from an input file and write the *DNA*, *mRNA* and *amino acid sequences* to an output file (according to the supplied `sample_output.txt` file). Both files should be specified as command line arguments (please adapt the file name accordingly):

```
1 $ ./ex05_0123456.py <input_file> <output_file>
```

Apply your program to the set of sequences in `test_sequences.fa`.

6 Exercise (14 points)

Consider the following hypothetical multiple alignment over the alphabet of nucleotides {A, C, G, T}:

```
ATGGAATCCC
ATGGAATCCA
ATGGATTCCA
ATTGACACCA
GATGAATCCT
GAGTTATCTT
```

Compute (using pen and paper) and compare the resulting *BLOSUM100* and *BLOSUM75* scoring matrices. What do you observe? Interpret and discuss your results.

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/>