Prof. Dr. Daniel Huson
Algorithmen der Bioinformatik
Fachbereich Informatik
Mathematisch-Naturwissenschaftliche Fakultät

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

**Sequence Bioinformatics** **WS 2021/22**

**Assignment 1** **Due: Nov-3, 10 am**

## 1 Setup Python

Download and install `Python 3.8` (or later). Find and install your favorite IDE (we recommend PyCharm) or advanced text editor (such as Sublime). A list of all of IDEs can be found on Wikipedia (search for Integrated Development Environments).

The aim of the following tasks is to get you working with Python and to have you implement some functions that we will need in some of the following assignments. Please do not use any bioinformatics-specific libraries to solve these tasks. The task here is to implement the methods "from first principles".

## 2 FastA input and output (3 points)

Implement a Python file called `fasta_YOUR_NAME.py` that provides two methods, called `read()` and `write()`. The `read()` method takes one parameter called `file_name`. Assuming that `file_name` is the name of a file containing sequences in FastA format, the method reads in the data and returns a list of pairs of strings, containing the header line and sequence of each entry.

The `write` methods takes two parameters, `fasta_pairs` and `file_name`, and writes all FastA pairs to the named file. If `file_name` is not specified, write the sequences to the console instead.

(Implement the two methods in such a way that `write` can be applied to the output of `read`.)

## 3 FastA echo (1 point)

Using your class `fasta_YOUR_NAME.py`, write a program `fasta_echo_YOUR_NAME.py` that reads in a specified FastA file and writes it to the console. The file is to be specified as a "command-line option".

## 4 DNA translation (4 points)

Write a Python program `translate_YOUR_NAME.py` that reads in DNA sequences from a file in FastA format, translates all sequences in to amino acid sequences using the standard genetic code (for bacteria), and writes the resulting sequences to an output file, or to the console, in FastA format. You only need to translate the first frame, *not* all six possible reading-frames. (The input file is specified as the first command-line option. If an output file is desired, then it is specified as the second command-line option).

## 5 Edit score (2 points)

Write a Python program `edit_distance_YOUR_NAME.py` that reads in sequences from a file in FastA format. The program first checks whether all sequences have the same length. (What is the correct way to deal with this, if the condition is not fulfilled? Please implement it.)

If all sequences has the same length, then the program reports the matrix of edit distances between each pair of sequences $s$ and $t$, where the edit distance $d(s,t)$ is defined as the number of positions at which $s$ and $t$ show different symbols.

(Note that this is a straight-forward comparison between each pair of sequences and does not involve dynamic programming or any alignment algorithm!)

## Handing in

You have two weeks to solve and hand in the first assignment. However, there will be a new assignment handed out next week (on Oct 27). If you have any problems, please discuss them in the first tutorial next week. Also, please use the course forum on Ilias to ask any questions of general interest that you have.

**For all assignments, make sure that your names always appears in the file name (as discussed above) and inside the file. Uploaded files that do not fulfill both requirements will be ignored.** Upload your solutions to the upload directory inside your tutorial directory.

## Tutorial sessions

The first session of group A will be on Monday, Oct 25. The first session of group B will be on Wednesday, Oct 27. The first session of group C will be on Thursday, Oct 28.

The first session is aimed at helping you to install and run Python on your computer and to help you with programming issues. **Attendance of the first session is optional, but attendance of all following tutorials is mandatory.**