Sequence Bioinformatics - Assignment 1

Tobias Fehrenbach and Ignacio García Ribelles

November 2, 2021

1 Setup Python

The chosen IDE was PyCharm.

2 FastA input and output

A program was implemented with two functions: read() and write(). The latter took a FastA-file as input and outputted a list of paired tuples consisting of header and sequence called "data". The write() function took

In the first task we had to implement a Python file which contained two methods for FastA-file processing. The first method read(file_path), that takes the path to a FastA file as input and returns its contents as a list of header and sequence tuples as follows [(header[0], seq[0]), ...].

The second method write(data, file_path=none) takes the data as a list of tuples (as described above) and prints the content of data to the console if no file path is specified as second argument. If a viable file path is specified write() writes the data to a FastA-file.

3 FastA echo

The read() and write() functions from exercise 2 were joined together, asking the user only for an input FastA-file path and outputting its contents to the console.

4 DNA translation

The third exercise asked to read in a FastA-file and translate the DNA sequences to amino acid sequences. To read and write the files we used the functions implemented in exercise two. For translation we implemented a function called translate() which takes the sequences as a parameter and compares them directly to a dictionary containing the corresponding amino acids. This dictionary corresponds to the standard genetic code in its DNA version according to M.Varus(1).

The program takes an input file path as first argument and an optional second command line argument as output file path. If no second argument is given, the program prints the amino acids sequences to the console.

5 Edit score

For the edit distance calculation we wrote a function called calc_edit_dist() which takes headers and sequences as lists to calculate an edit distance matrix. We defined the edit distance as the number of positions at which each pair of sequences s and t show different symbols. The output is printed out as a pandas matrix for better visualization, with headers as both columns and rows. The larger the edit distance the more dissimilar the two sequences are, thus the diagonal is filled with zeros.

6 References

Michael Yarus, Evolution of the Standard Genetic Code, J Mol Evol . 2021 Feb;89(1-2):19-44. doi: 10.1007/s00239-020-09983-9 Add to Citavi project by DOI. Epub 2021 Jan 24.