Master thesis proposal

Development of the R package for the bioinformatics tool

Krzysztof Bartoszek and Sebastian Sakowski

October 10, 2022

1. Background

The number of different pieces of information contained in biomedical databases has increased in recent years [1]. Of particular interest are the data containing the results of the next-generation sequencing (NGS), which are stored as FASTA files and contain nucleotide sequences. The GenBank database is one of the most important places to collect information about FASTA files. There are many tools that allow the analysis of FASTA files, the result of which is information that takes into account the various needs of the researcher [2-5].

The main goal of this Master thesis is to develop an appropriate interface in the R language (CRAN R package) to the data resulting from the operation of the selected bioinformatics tool. The interface developed in R will have the ability to load, visualize, and analyze data. Another result will be the development of a certain scheme that will describe the regularities in the genetic phenomenon. A bioinformatics tool will be improved for the statistical analysis of FASTA files, as well as the further interpretation of data analysis. An additional effect of the thesis is to select appropriate statistical methods for data received from a bioinformatics tool.

2. Research questions in bullet form

- How should the user statistical interface in the R language be developed to help the user application selected statistical methods?
- Which universal statistical method shall be used and what performance will it have on genetic data?
- Which visualization method should be implemented to help the user understand the genetic data?

3. Data Description

- The data cover GenBank - FASTA genetic sequence database files.

4. References:

1. S. Sakowski, J. Waldmajer, I. Majsterek, T. Popławski: *DNA Computing: Concepts for Medical Applications*. **Applied Sciences**, 2022, 12, 6928.
2. F. Escudié, C. Van Goethem, D. Grand, J. Vendrell, A. Vigier, P. Brousset, S. M, Evrard, J. Solassol, J. Selves: *MIAmS: microsatellite instability detection on NGS amplicons data*. **Bioinformatics** 2020, 36(6), 1915–1916.
3. S. B. Mudunuri, H. A. Nagarajaram: *IMEx: Imperfect Microsatellite Extractor*. **Bioinformatics** 2007, 23, 1181–1187.
4. S. B. Mudunuri, S. Patnana, H. A. Nagarajaram: *MICdb3.0: a comprehensive resource of microsatellite repeats from prokar-551 yotic genomes*. **Database** 2014, bau005–bau005.
5. G. Benson: *Tandem repeats finder: a program to analyze DNA sequences*. **Nucleic Acids Res.** 1999, 27, 573–580.