BioSeC: Bioinformatic Sequence Classification

Majid Soheili

September 18, 2022

1 Why

High-throughput sequencing technology and powerful bioinformatics approaches are boosting genomic and metagenomic analysis. This combination has led to an exponential increase in sequence data. In 2022, more than 13 TeraByte sequence data was stored in the SRA database daily [1]. The accuracy of the categorization of sequence data uploaded in the SRA is reliant on the submitters. The SRA curators aim to collect correct metadata on the sequences submitted; nevertheless, annotations are not standardized, different methods are used to classify sequences submitted to databases. The BioSeC project is an effort to check quality of the sequence file, and it going to detect the type of the sequence file according to the context of the file instead of depending on the metadata configured by submitters. The main objective of this project can be listed as:

1. Classify the sequence file into four categories:
   (a) Amplicon Sequence:
   (b) Whole Genome Sequence (WGS), Meta-Genome
   (c) Isolated Genome
   (d) Single Amplified Genome (SAG)
2. The learning model should be able to cope with large-scale data training.
3. The scalability and distributed approaches are welcomed in dealing with voluminous sequence files submitted to the SRA database formerly.

2 How

In this section, four major steps for achieving the goals are listed. Each step is detailed by a deadline and a measurement. Some steps required meeting for expertise marked with †.

[1] https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

1. Definition: Getting familiar with the problem 25, September, 2022
   (a) Measurement: Should be confirmed by expert.
2. Preparing the training Dataset: 17, October, 2022
   (a) Measurement: The method should be confirmed by expert.
   (b) The label of sequence files should be same in both datasets JGI[2] and SRA[3].
3. Extracting features: Different types of features will be extracted. 31, October, 2022
   (a) Measurement: Features selection methods for evaluating the features should be used.
4. Developing a classification model: 14, November, 2022
   (a) Measurement: Total Accuracy and Geometric-Mean by using 5-fold cross validation.
5. Submitting Manuscript: 14, December, 2022
   (a) Measurement: Confirmed by supervisor.
6. Launch of online system to check data quality. 31, January, 2022
   (a) Measurement: Getting feedback from experts.

3 What

1. Definition: Getting familiar with the main objective 01, October, 2022
   - Preparing a proper introduction of the issue and the idea of solution .
   - Preparing the simple explanations for each target label, and the biological differentiation among them.
2. Review some published paper and methods.
   - Feature Extraction papers.
   - Some papers like these should be considered and reviewed [**?**,].
3. Accessing to sequence files
   - Objective: The main idea is that we need to prepare a small subsample of the original sequence file instead of downloading and uncompressing the whole one.
   - Subsampling method: We need a reliable method such that we can extract some spots randomly from the whole sequence file (1000 - 3000 spots of each sequence file).

[2]https://jgi.doe.gov/
[3]https://www.ncbi.nlm.nih.gov/sra

- Scalable subsampling: The proper approach should be multiprocessing or multithreading.

4. Preparing the training Dataset:
   - We should prepare a reliable dataset, so we are going to get the sequence files from JGI and SRA.
   - We believe that for making a good learning model, the number of sequence samples in the training dataset should be around 5 thousand for each type of sequence.
5. Cleaning the Dataset
   - Invalidity: Removing the sequence file with insufficient numbers of spots (less than 3000).
   - Outlier detection: For preparing a more reliable training dataset removing outliers should be necessary.
6. Extracting features: Different types of features will be extracted.
   - Different types of features introduced in the literature should be used. For example, Numerical mapping, genomic signal processing (GSP), Chaos game representation, Entropy, and Graphs [**?**].
   - Various features with different natures should be extract. Owing to subsampling, some feature extraction methods can be impossible.
   - After feature extraction, the feature selection seems to be necessary because some features will be redundant and irrelevant.
7. Developing a classification model.
   - The type of the classification model will relay on the type and number of features, but we believe that the using Ensemble method should be useful.
   - Using the 5-fold cross validation approach for automatic evaluation should be reasonable.
   - Measures: Accuracy and Geometry-Mean
8. Implementing the web interface to use for others.
9. Beta Test and evaluating machine learning model for detecting the sequence file type.
   -
10. Launch of online system to check data quality. 31, January, 2022
11. Preparing the draft version of the manuscript
12. Review and Submitting the manuscript.

4 Mapping to TA2 Meetings

In this section, for each TA2 meeting in the NFDI project, an milestone will be allocated. We assume that on the first week of each month, the TA2 meetings will happen regularly. In each meeting will talk about the result obtained and the next milestone. Table **??** the illustrates some information of the meetings.

Table 1: TA2 meeting topics

#	Topic	Date
1	Issues and main Idea	03.10.2022
2	Preparing training dataset and Feature extraction	02.11.2022
3	Classification Model	02.12.2022
4	Final Result and Manuscript	02.01.2023
5	Online System	01.02.2023

References

[1] P. J. Torres, R. A. Edwards, and K. A. McNair, ‘‘PARTIE: a partition engine to separate metagenomic and amplicon projects in the sequence read archive,’’ *Bioinformatics*, vol. 33, pp. 2389--2391, Aug. 2017.

[2] R. P. Bonidia, D. S. Domingues, D. S. Sanches, and A. C. P. L. F. de Carvalho, ‘‘MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors,’’ *Briefings in Bioinformatics*, vol. 23, 11 2021. bbab434.