BioSeC: Bioinformatic Sequence Classification

Majid Soheili

September 16, 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.

There are two orthogonal approaches commonly used to explore the microbial universe: (i) Amplicon, where a part of a single gene (usually the 16S gene) is amplified and sequenced. (ii) Shotgun Metagenomics(random), where all the DNA is extracted and sequenced. The main objective of this project is proposing a new method to classify the sequence file into four subcategories.

1. Amplicon Sequence:
2. Whole Genome Sequence (WGS), Meta-Genome
3. Isolated Genome
4. Single Amplified Genome(SAG)

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. Definition: Getting familiar with the problem 01, October, 2022
2. Preparing the training Dataset:

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

3. Extracting features: Different types of features will be extracted, and evaluating and pruning some of them seems to be necessary.
4. Developing a classification model:
5. Test and evaluating machine learning model for detecting the sequence file.
6. Launch of online system to check data quality. 31, January, 2022

1. Definition: Getting familiar with the problem 01, October, 2022
    - Preparing the proper introduction of the problem.
    - Preparing the simple explanations for each target label, and the biological differentiation among them.
2. Review some published paper and methods.
    - Some papers like these should be considered and reviewed [?,].
3. Preparing the training Dataset:
    (a) We should prepare a reliable datasets, so we are going to get the sequences files from, JGI and SRA.
    (b) According to our estimation, the number of sequences that would be enough will be around 5 thousand in each type of sequence .
    (c) Removing the noise it will be necessary to the illumination of the outliers files.
4. Cleaning the Dataset
    (a) Invalidity: Remove the sequence file includes small spots (less that 3000)
    (b) Outlier detection: For preparing more reliable training dataset removing some outlier should be necessary.
5. Extracting features: Different types of features will be extracted
    (a) Different type of features are introduced before like, Numerical mapping, genomic signal processing (GSP), Chaos game representation, Entropy, and Graphs [?].
    (b) We should produce some features with different nature if they be possible.
    (c) After feature extraction, the feature selection seems to be necessary.
6. Accessing to Sequence files
    (a) Aim: 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.

   (b) Sampling 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).

   (c) Scalable subsampling: The proper approach should be multiprocessing or multithreading.

7. Developing a classification model

   (a) 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.

8. Test and evaluating machine learning model for detecting the sequence file.

9. Implementing the web interface to use for others.

10. Launch of online system to check data quality. 31, January, 2022