Digital Professionalism in Health and Care: Developing the Workforce, Building the Future
P. Scott et al. (Eds.)

© 2022 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI220932

Characterizing Genomics Repositories Using Feature-Based Classification

Jorge Miguel SILVA[a] and João Rafael ALMEIDA[a,b,1]
[a] *DETI/IEETA, University of Aveiro, Portugal*
[b] *Department of Computation, University of A Coruña, Spain*

Abstract. In the last decades, the field of metagenomics aided by NGS technologies has grown exponentially and is now a cornerstone tool in medicine. However, even with the current technologies, obtaining a conclusive identification of an organism can be challenging due to using reference-based methods. Consequently, when releasing a new repository of genomic data that contains de-novo sequences, it is problematic to characterize its content. In this paper, we propose a novel method for organism identification and the creation and characterization of genomic databases. For identification, we propose a three-step pipeline for reference-free reconstruction, reference-based classification and features-based classification. On the other hand, for content exposition and extraction, the sequences and their identification are aggregated into a web database catalogue.

Keywords. Metagenomics catalogue, Taxonomic identification, Genomic population characterization, Compression, Feature-based classification.

1. Introduction

Metagenomics studies the structure and function of entire nucleotide sequences isolated from all the organisms in genetic material recovered directly from environmental samples by applying massive sequencing technologies. This field has become increasingly critical and instrumental in the healthcare domain [1]. The most significant catalyst for the growth in metagenomics was Next-Generation Sequencing (NGS). NGS is a massively parallel sequencing technique with ultra-high throughput, adaptivity, and speed. Furthermore, NGS is a cost-effective way to screen large genomes and can be used to establish the order of nucleotides in entire genomes or specific portions of DNA or RNA. Due to its capabilities, NGS is currently used in genotyping, genome reconstruction, gene expression measurement, identification of human vial communities, and finding genomic rearrangements. Several steps must be performed when identifying an organism from metagenomic samples: genomic sequencing, trimming, filtering, assembly, and analysis. Although some steps, such as cloning and sequence enrichment, aim to improve the genome quality, identifying the organism is not always conclusive [2]. Consequently, when releasing a new repository of genomic data that contains de-novo sequences, it is problematic to characterize its content. This paper proposes a methodology for characterizing these datasets using compression-based approaches conjugated with simple features such as GC-content and sequence length.

[1] Corresponding Author, João Rafael Almeida; E-mail: joao.rafael.almeida@ua.pt.

2. Methods

The first part of our methodology was developing a full-fledged adaptable tool to accommodate various scenarios present in metagenomics. Therefore, we propose the creation of a complete pipeline for accurate and fast, in-depth taxonomic identification and classification of organisms in metagenomic samples. In our proposal, we devise a pipeline that uses other tools over distinct steps for reference-free reconstruction, reference-based classification and features-based classification.

After organism identification, the extracted classification of the dataset is detailed into files and sequences (reads). Then, we aggregated this information based on the number of files and the count of domains of the entire dataset. This information was then exposed in a web-database catalogue. This catalogue was built on top of MONTRA Framework [3] which was designed as a general engine to build metadata catalogues of biomedical databases. The system also ensures access control and data privacy, following rule-based access policies crucial to control which data is accessible.

3. Preliminary Results and Next Steps

To validate the accuracy of our newly proposed methods, we used open-source datasets, such as the NCBI dataset. With these datasets, we extracted features, namely: i) the normalized compression (using various compressors) from the genome and proteome; ii) GC-content; iii) sequence length. These features are labelled using taxonomic information retrieved from the NCBI database. Afterwards, we used the feature database to perform a 5-fold cross-validation on the dataset, applied to each domain. The average F1-score obtained for each domain were Archaea (69%), Bacteria (67%), Fungi (82%), Protozoa (76%) and Viral (84%). These results aim to evaluate which features and models outperform and generalize better for classification purposes. Then, we set up a genomic catalogue for uploading these summaries in a proof-of-concept environment. The summaries were extracted from the dataset used to validate the first component. The following steps aim to release this catalogue to the public, exposing metadata about genomic repositories in a user-friendly web catalogue.

Acknowledgments

This works is funded by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UIDB/00127/2020. J.M.S. and J.R.A are funded by the FCT - Foundation for Science and Technology (national funds) under the grants SFRH/BD/141851/2018 and SFRH/BD/147837/2019, respectively.

References

[1] Amorim A, Pereira F, Alves C, García O. Species assignment in forensics and the challenge of hybrids. Forensic Science International: Genetics. 2020;48:102333.

[2] Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, et al. Genome Detective: an automated system for virus identification from high-throughput sequencing data. Bioinformatics. 2019;35(5):871-3.

[3] Silva LB, Trifan A, Oliveira JL. Montra: An agile architecture for data publishing and discovery. Computer methods and programs in biomedicine. 2018;160:33-42.