Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 219 (2023) 1501–1508

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2022

Classifying and discovering genomic sequences in metagenomic repositories

Jorge Miguel Silva[a], João Rafael Almeida[a,b], José Luís Oliveira[a,]

[a]*DETI/IEETA, University of Aveiro, Portugal*
[b]*Department of Computation, University of A Coruña, Spain*

Abstract

The taxonomic and functional composition of microbial communities from environmental, agricultural, and therapeutic settings is increasingly being studied using metagenomic methodologies in large-scale genomic applications. This has led to exponential growth in the field and has impacted on healthcare, pharmacology and biotechnology. However, with the current methodologies, it is sometimes difficult to obtain conclusive identification of an organism. In addition, the growth of the metagenomic field has led to the creation of large amounts of data held by different hosts, which characterize data differently and make analysis difficult. Therefore, correct data aggregation and classification improve and facilitate the discovery of repositories of interest. This paper tackles these issues by proposing a methodology for organism identification, data aggregation and content characterization, visualization and selection. We propose a three-step pipeline for organism identification that uses compression-based metrics, an aggregation mechanism for content characterization, and a web database catalogue for data exposition and visualization.

Keywords: Taxonomic Classification; Organism Identification; Compression; Web Portal; Data Aggregation; Genomic Catalogue

1. Introduction

Metagenomics is a vast field that enables the discovery of taxonomic and physiological information of species collected from their natural environment. Acquiring this information is especially important given the role that microorganisms play in terms of the structural and functional balance of the ecosystem, medicine and patient health status, as well as industrial and economic activity [1]. Due to its importance, this field has become instrumental in medicine, forensics, and exobiology [2].

E-mail address: {jorge.miguel.ferreira.silva, joao.rafael.almeida}@ua.pt

1877-0509
10.1016/j.procs.2023.01.441

The rapid growth of metagenomics has led to the generation of large amounts of genomic data. Several institutions host distinct repositories of these data that can be used to help researchers conduct genomic analysis. Studying these sequences is time-consuming, and correctly identifying the repositories is essential to support researchers when selecting the repositories of interest. This characterization can be done by identifying the organisms in the metagenomic samples. However, identification of the organism is not always conclusive [3]. The leading cause of this is that most classification pipelines rely on reference-based comparison approaches to perform organism identification [4], where the reconstructed sequence is compared to a collection of references stored in a database [5]. Problems occur when performing reference-based identification since this method is ineffective when there is a variation between the sequences of known organisms in the database, when irregularities are introduced during the reconstruction process of the organism being identified, or when a new organism is being sequenced [6, 7].

Other approaches have recently been used to identify organisms without reference-based methods. Karlicki *et al.* [8] developed a deep learning-based categorization system. The system first classifies nuclear and organellar eukaryotic fractions. Afterwards, it separates organellar sequences into plastidial and mitochondrial categories. Zhang *et al.* [9] used k-mers as genomic features for viral genome identification. Silva *et al.*used feature-based classification to perform in-depth taxonomic classification of archaea [5] and viruses [10]. This model used compression-based measures (specifically Normalized Compression) with other simple features to achieve state-of-the-art results. A later study conducted by Silva and Almeida [11] showed that compression efficiency does not correlate with classification accuracy and that using multiple different compressors significantly increases the accuracy of organism identification. Another vital factor to consider is the need to accommodate mechanisms for sequence reconstruction to correct irregularities introduced during the reconstruction process and a reference-based identification mechanism to provide an initial identification. These two changes would considerably improve the identification tool's accuracy and reliability, respectively.

The information obtained using these approaches for metagenomic characterization provides valuable insights into the processed repositories. This information can then be used to help researchers identify repositories of interest. Currently, some platforms aggregate metadata about genomic samples, such as the NCBI database[1] and GISAID[2]. However, the information on these platforms is hard to filter and visualize. Another issue with these platforms is the lack of information due to some metadata being filled in manually, which often causes a lack of details to characterize a dataset.

A strategy to simplify data discovery can involve database catalogues [12]. These platforms aim to expose metadata that characterize relational databases, semi-structured datasets, and data repositories, among others [13]. Almeida *et al.* [14] proposed adopting a web-based platform to expose this information, aiming to optimize the analysis of distributed and private biobanks. However, the datasets in the platform were characterized manually by the data owners.

In this work, we propose a methodology to characterize unknown metagenomic sequences stored in genomic repositories and facilitate the discovery of such repositories using a web-based genomic catalogue. This work is mainly divided into two main components. The first is focused on a pipeline for accurate, fast identification of organisms from genomic samples. The latter describes an aggregation tool that collects information and sends it to a web catalogue platform where the genomic samples and their matching classification summaries can be exposed and quickly accessed by researchers.

2. Materials and Methods

In this section, we describe the methodology overview, followed by a more in-depth description of each part - specifically the three stages: i) genomic classification pipeline, ii) data characterization, and iii) genomics web catalogue.

[1] https://www.ncbi.nlm.nih.gov/
[2] https://www.gisaid.org/

2.1. Overview

A complete overview of this methodology is illustrated in Figure 1. The classification pipeline comprises a three-step process: i) reference-free reconstruction, ii) reference-based classification, and iii) features-based classification. Consequentially, this pipeline was designed to accommodate the following capabilities in three stages:

1. Performing reference-free reconstruction of a genomic sequence from a metagenomic sample;
2. Accommodating state-of-the-art reference-based classification;
3. Classifying sequences using compression-based features and machine learning classifiers.

The output from this pipeline is classification of the genomic data. This information includes the determined identification and taxonomic description. Finally, the output is uploaded to a centralized platform, aggregating and exposing the information in a web-database catalogue. All the process is described in more detail in the following sections.

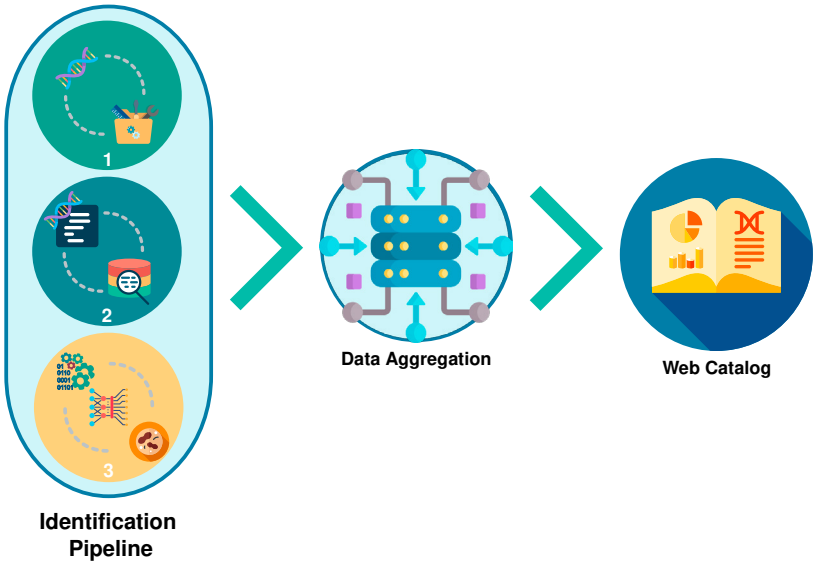

Fig. 1. Project description. First the 3 stages of organism identification (1-3), followed by data aggregation, and finally upload to the web catalogue.

2.2. Genomics classification

2.2.1. Stage 1 - Reference-free Reconstruction:

The first step is the reference-free reconstruction, which is characterized by assembling the genomes from FASTQ reads. Then, the reads go through a quality control process, entering the process of genome reconstruction by building scaffolds from overlapping reads assembly.

High-throughput sequencing reads are highly prone to error. In Illumina sequencing, the error is distributed non-randomly over the length of the read [15]. These facts create an additional layer of complexity for genomic analysis. Due to this extra layer of complexity, the reads must be trimmed and cleaned before being analyzed, removing eventual sequencing errors, and reads with low-quality scores have to be filtered [16].

For this purpose, our pipeline uses FASTP [17] to perform quality control, adaptive trimming, quality filtering, and per-read quality pruning, among other operations that provide clean data for downstream analysis [17]. In addition, this tool supports multi-threading, single-end and paired-end reads. Furthermore, as the read trimming stage is a fundamental process throughout this analysis, the tool is available for the user to trim high-throughput sequences. An advantage of FASTP is that it is a fast tool, allowing a performance 2 to 5 times faster than other FASTQ preprocessing tools [17].

Following trimming, *de-novo* reference-free assembly takes place. This step in the pipeline aims to reconstruct the genomes, starting from many reads without prior knowledge about the correct sequence, order, abundancy, or composition. The metaSPAdes tool is used in the *de-novo* assembly to assemble datasets with non-uniform coverage [18]. It was created primarily as a tool for metagenomic assembly rather than target-based assembly. It functions by activating a sequence of flags to improve the output data and reduce mismatches and short indels. Additionally, metaSPAdes supports data consisting of single-reads and paired-reads, both of which are supported in our pipeline. Depending on the genomic fragments used, paired reads can become much more beneficial than single reads, either in resolving structural rearrangements or in indicating the size of repetitive regions and how far contigs are from each other [19]. In the final part of the assembly, metaSPAdes creates scaffolds in a FASTA format with the reconstructed genomes or fragments of genomes. Several scaffolds are generated in cases where the sequencing depth is low.

2.2.2. Stage 2 - Reference-based Classification:

The second step is reference-based classification and occurs after the reconstruction process. Despite our pipeline focusing on compression-based classification, we also need to provide a reference-based classification tool to compare results with the compression-based classification. To this end, we adopted Kraken2, a state-of-the-art reference-based classification tool that creates or downloads a database to identify the organism's taxonomy [20].

Kraken2 uses a k-mer-based approach that provides a relatively fast-taxonomic classification of metagenomic sequence data. In addition, Kraken2 improves upon Kraken by reducing memory usage, allowing more significant amounts of genomic reference data to be used and increasing its computational speed. Furthermore, Kraken2 introduces a translated search mode, providing increased sensitivity in viral metagenomic analyses [20].

When using this tool, after sequence reconstruction, the sequence is fed to Kraken2 and using a k-mer-based approach, it is compared with the other sequences stored in a database. The results are taxonomic descriptions of the organism, which serve as a comparison baseline for our feature-based classification results. For instance, when processing a FASTA file containing bacteria, we obtain the taxonomic tree from root until Streptomyces sp. The data present in these trees were used in the characterization stage, to have refined information about the repository.

2.2.3. Stage 3 - Feature-based Classification:

The final stage of this classification pipeline is focused on performing reference-free classification. This was achieved by creating a database of features, computed from the existent sequences in the NCBI database. This process retrieves the genome and taxonomic information to construct the database. The former was compressed to obtain the various compression features, and the latter served as labelling information.

The features were obtained adopting the methodology that, according to Silva and Almeida *et al.* [11] provided the best results. The XGBoost Classifier was used with the conjugated features provided by the bzip2[3], Jarvis [21], MFCompress [22], NUHT [23] and zstd[4] compressors.

For each sequence obtained from the NCBI database, the compressors' NC value is determined and stored in the feature database. The NC is computed by dividing the size of the lossless compression file and the initial file size. The same features are extracted when a new sequence enters the pipeline, and the XGBoost Classifier determines its domain using the feature database as a training set.

2.3. Data characterization

After organism identification, the extracted classification of the dataset is detailed into files and sequences (reads). The information was then aggregated based on a set of characteristics, such as the number of files, reads, and domains of the entire dataset, among others. This aggregation was done using a tool created for this purpose (data aggregation stage represented in Figure 1).

The output is uploaded in MONTRA Framework, a flexible web-based database catalogue [13]. The Montra Framework is already used in other initiatives (such as EMIF-Catalogue[5] and EHDEN Portal [6]) and was designed as a general

[3] http://www.bzip.org/
[4] https://github.com/facebook/zstd/
[5] https://emif-catalogue.eu/
[6] https://portal.ehden.eu/

engine to build metadata catalogues of biomedical databases [24]. The system has a flexible data skeleton used to characterize data entities, making it easy to customize. In addition, MONTRA has a RESTfull API that allows creating new entries in the catalogues and updating specific details of each entry over time. This API was used to upload the aggregated information automatically to the catalogue.

In this case, we adapted the platform for easier sharing and exposing genomic data while preventing access to the data. The system also ensures access control and data privacy, following rule-based access policies crucial to control which data are accessible for the different access levels [25]. The platform does not expose sensitive information since the available data only characterize the repositories' content. Additionally, the platform has study management features where users can select relevant repositories and create research questions [26].

An entry in the catalogue contains the repository characteristics divided into groups, namely:

- repository general information: provides general information about the repository, for example, name, institution, data held and primary responsibilities;
- repository characteristics: holds the information extracted in the described pipeline (this group represents the repository directly and is the only group updated automatically);
- contact details: information about the people responsible for managing the repository;
- data access and ethical issues: description of the data access agreements and the ethical and governance policies;
- key publications: contains the publications associated with the repository.

3. Results

The results of the proposed pipeline are: i) classification of unknown genomic sequences and ii) characterization of these sequences' repositories. The former was validated using a public dataset, and the latter followed an empirical validation by exposing the data to the database catalogue.

3.1. Research application

To measure the accuracy of the first three stages of the pipeline, we used an already classified dataset [11]. This is described in Table 1 and is composed of a balanced pool of 450 random FASTA files obtained from NCBI, where each file has a varying number of reads. Moreover, the dataset sequences were from six domains: viral, bacteria, archaea, fungi, and protozoa. The feature dataset was compiled by applying the five compressors (zip2, Jarvis, MFCompress, NUHT and zstd) for each sequence in the dataset and obtaining their NC results. Afterwards, the validation classification was performed using a random 80-20 train-test split on the dataset. Furthermore, instead of performing k-fold cross-validation, we employed the random train-test split ten times and retrieved the average of the F1-score results. The F1-score results are also shown in Table 1.

Table 1. Dataset description and compression-based classification results obtained using multiple compressors. F1-score obtained per taxonomic group and average accuracy and F1-score.

Domain	**Viral**	**Bacteria**	**Archaea**	**Fungi**	**Protozoa**	**Total**
N. Samples	90	90	90	90	90	450
F1-score	84.85%	66.67%	69.39%	82.35%	76.47%	72.44%

3.2. Characterization metrics

In this use case, we already had the same information since we used a labeled dataset to validate the complete pipeline. The metrics presented here were extracted with the support of ORFm tool [27] and GTO toolkit [28]. Therefore, the metrics extracted from these 450 FASTA files used to validate the first part were:

- number of files;

- number of reads;
- domains (all domains, *e.g.*, viral, bacteria, archaea, fungi, and protozoa);
- number of entries by domain (in this case, already represented in Table 1);
- number of complete genomes (total and by domain);
- number of partial genomes (total and by domain).
- number of scaffold (total and by domain);
- percentage of each base (ACGTs);
- percentage of GC-content (by domain);
- sequence length (by domain);
- proteome size (by domain);
- percentage of each aminoacid (by domain);
- compression level of genome (by domain, for different compressors).

3.3. Genomics Catalogue

The final outcome of this work is the genomic catalogue. In Figure 2, we illustrate this catalogue to characterize the genomic repositories. The figure on the left shows a list of characterized repositories. In this example, we used labelled repositories, namely the one described previously. On the right, we have a partial view of the repository characteristics represented in this system.

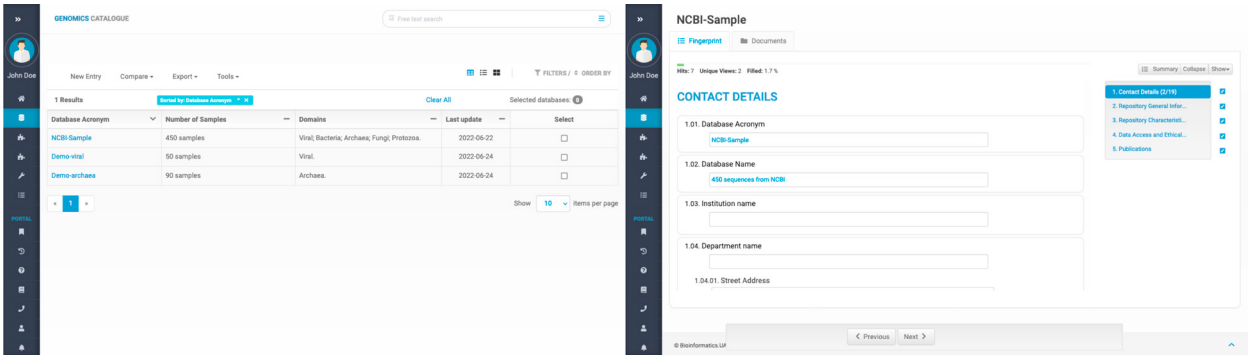

Fig. 2. The Genomics Catalogue showing demo repositories (on the left) and part of the characterization of the repository used to validated the pipeline (on the right).

Although we did not use real data, we obtained a complete pipeline that can process unknown sequences from new repositories, classify them and expose a set of characteristics in a centralized platform. Besides this, the data can be better associated with other resources or easily shared within the community.

4. Discussion

Discovering the proper genomic repositories is a time-consuming task that is necessary when conducting genomic studies. One of the challenges when doing this task is understanding the composition of the repositories currently available (*e.g.* on NCBI). With the creation of Next-Generation Sequencing (NGS) technology, an unprecedented amount of genomic data was produced. This method 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 [29]. Among other applications, NGS is employed in gene expression measurement, genotyping, genome reconstruction, identification of human vial communities, and uncovering genomic rearrangements [30, 31].

This unprecedented generation of genomic data enables researchers to conduct metagenomic studies at a new scale and use it, for instance, to analyze microorganisms that can be associated with infectious diseases. Metagenomics plays an important role when studying these organisms' genomics in a culture-independent way by analyzing their DNA obtained directly from environmental samples [32]. It not only allows the identification of the organism species present

in the sample but also provides insight into the functional roles and metabolic activities of the microorganisms [33]. Therefore, combining this information with function-based activity makes it possible to discover new functional genes from uncultured microbes.

As proposed in this work, characterizing genomic repositories and exposing their metadata to a centralized platform can optimize the discovery task. The concept of profiling datasets was already adopted in other health use cases. For instance, the European Medical Information Framework project (EMIF)[7] had the goal of facilitating the discovery of Electronic Health Records (EHR) databases available for study. In this project, a web-catalogue was created, using MONTRA Framework in its core [34].

One of the challenges in characterizing the genomic repositories in a catalogue is the correct identification of characteristics that would be useful for researchers. In our pipeline, we proposed a component to extract a set of characteristics that provide insights about each repository. This procedure aims to aggregate critical information and automatically upload it into the proposed genomic catalogue. When the sequences are already labelled, this procedure is simplified. However, taxonomic identification of each sequence is necessary when dealing with newly sequenced data.

To solve the organism identification challenge, we created a three-step pipeline that uses strategies validated in other use cases. Adopting metagenomics to provide a fast and accurate way to identify sample organisms is vital for academic research and industrial applications. However, as previously stated, the methodologies currently in use, despite being helpful, can be faulty at times, specifically in the cases of variation between the sequences of known organisms in the database, the presence of irregularities introduced during the reconstruction process of the organism being identified, and when a new organism is being sequenced [6, 7]. This pipeline solves these problems by accommodating genome reconstruction and possesses reference-based methods conjugated with feature-based classification to achieve a higher degree of accuracy in genome identification. On the one hand, genomic reconstruction solves possible irregularities introduced during genome sequencing, making the identification more accurate. On the other hand, a state-of-the-art reference-based solution assures researchers by providing well-known baseline classification results. Finally, the feature-based solution clears any doubt regarding inconclusive results by providing the flexibility required to identify species with variation between the sequences, as well as identifying unknown organisms.

5. Conclusions

In healthcare, metagenomics is currently applied to identify an unknown pathogen in disease outbreaks and to discover and detect pathogens in clinical samples. However, despite its importance in this field, many barriers make studying these samples difficult. One of the most significant ones is correctly identifying the best genomic repositories for conducting a study. This process is complex and time-consuming since the information is dispersed and often lacks a good description. Therefore, this work is designed to optimize this process.

The data gathered using the proposed methodology provide insightful knowledge about the processed repositories. Researchers can then use this information to find relevant repositories. Unfortunately, although some platforms aim to expose metagenomic information, they are hard to use to filter and select data for subsequent analysis. Moreover, in these platforms, some metadata is filled in manually, causing errors and a lack of details characterizing the dataset. In our work, we develop a method employing database catalogues to streamline the data finding process by aggregating and homogenizing metadata that can be present in silos with structured, semi-structured and unstructured formats. Additionally, we disclose these data via a web-based platform in an effort to streamline the study of scattered and private biobanks.

Acknowledgements

This work has received funding from the EC under grant agreement 101081813, Genomic Data Infrastructure. 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.

[7] https://www.emif.eu

References

[1] Choudhari J, Choubey J, Verma M, Chatterjee T, Sahariah B. Metagenomics: the boon for microbial world knowledge and current challenges. In: Bioinformatics. Elsevier; 2022. p. 159-75.

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

[3] 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.

[4] Chen S, He C, Li Y, Li Z, Charles EMI. A Computational Toolset for Rapid Identification of SARS-CoV-2, other Viruses, and Microorganisms from Sequencing Data. Briefings in Bioinformatics. 2021;22(2):924-35.

[5] Silva JM, Pratas D, Caetano T, Matos S. Feature-Based Classification of Archaeal Sequences Using Compression-Based Methods. In: Iberian Conference on Pattern Recognition and Image Analysis. Springer; 2022. p. 309-20.

[6] Abnizova I, Leonard S, Skelly T, Brown A, Jackson D, Gourtovaia M, et al. Analysis of context-dependent errors for illumina sequencing. J Bioinform Comput Biol. 2012;10(2).

[7] Boekhorst RT, Naumenko FM, Orlova NG, Galieva ER, Spitsina AM, Chadaeva I, et al. Computational problems of analysis of short next generation sequencing reads. Vavilov Journal of Genetics and Breeding. 2016;20(6):746-55.

[8] Karlicki M, Antonowicz S, Karnkowska A. Tiara: deep learning-based classification system for eukaryotic sequences. Bioinformatics. 2022;38(2):344-50.

[9] Zhang Q, Jun SR, Leuze M, Ussery D, Nookaew I. Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. Scientific reports. 2017;7(1):1-13.

[10] Silva JM, Pratas D, Caetano T, Matos S. The complexity landscape of viral genomes. GigaScience. 2022;11.

[11] Silva JM, Almeida JR. The value of compression for taxonomic identification. In: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2022. p. 276-81.

[12] Silva JM, Almeida JR. Characterizing genomics repositories using feature-based classification. Procedia Computer Science. 2022.

[13] 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.

[14] Almeida JR, Pratas D, Oliveira JL. A semi-automatic methodology for analysing distributed and private biobanks. Computers in Biology and Medicine. 2021;130:104180.

[15] MacManes MD. On the optimal trimming of high-throughput mRNA sequence data. Frontiers in Genetics. 2014;5:13.

[16] Kircher M. Analysis of High-Throughput Ancient DNA Sequencing Data. PubMed. 2012;840:197-228.

[17] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-90.

[18] Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5):824-34.

[19] Baker M. De novo genome assembly: what every biologist should know. Nature Methods. 2012;9:333-7.

[20] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome biology. 2019;20(1):1-13.

[21] Pratas D, Hosseini M, Silva JM, Pinho AJ. A reference-free lossless compression algorithm for DNA sequences using a competitive prediction of two classes of weighted models. Entropy. 2019;21(11):1074.

[22] Pinho AJ, Pratas D. MFCompress: a compression tool for FASTA and multi-FASTA data. Bioinformatics. 2014;30(1):117-8.

[23] Alyami S, Huang CH. Nongreedy unbalanced Huffman tree compressor for single and multifasta files. Journal of Computational Biology. 2020;27(6):868-76.

[24] Almeida JR, Monteiro E, Silva LB, Sierra AP, Oliveira JL. A recommender system to help discovering cohorts in rare diseases. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2020. p. 25-8.

[25] Almeida JR, Barraca JP, Oliveira JL. A secure architecture for exploring patient-level databases from distributed institutions. In: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2022. p. 447-52.

[26] Almeida JR, Gini R, Roberto G, Rijnbeek P, Oliveira JL. TASKA: a modular task management system to support health research studies. BMC medical informatics and decision making. 2019;19(1):1-9.

[27] Woodcroft BJ, Boyd JA, Tyson GW. OrfM: a fast open reading frame predictor for metagenomic data. Bioinformatics. 2016;32(17):2702-3.

[28] Almeida JR, Pinho AJ, Oliveira JL, Fajarda O, Pratas D. GTO: a toolkit to unify pipelines in genomic and proteomic research. SoftwareX. 2020;12:100535.

[29] Mardis ER. DNA sequencing technologies: 2006-2016. Nat Protoc. 2017;12(2):213-8.

[30] Fabbro CD, Scalabrin S, Morgante M, Giorgi FM. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoS ONE. 2013;8(12).

[31] Toppinen M, Sajantila A, Pratas D, Hedman K, Perdomo MF. The Human Bone Marrow Is Host to the DNAs of Several Viruses. Frontiers in Cellular and Infection Microbiology. 2021;11:329.

[32] George IF, Bouhajja E, Agathos SN. 6.11 - Metagenomics for Bioremediation. In: Moo-Young M, editor. Comprehensive Biotechnology (Third Edition). third edition ed. Oxford: Pergamon; 2011. p. 132-42.

[33] Matallana-Surget S, Jagtap PD, Griffin TJ, Beraud M, Wattiez R. Chapter 17 - Comparative Metaproteomics to Study Environmental Changes. In: Nagarajan M, editor. Metagenomics. Academic Press; 2018. p. 327-63.

[34] Oliveira JL, Trifan A, Silva LAB. EMIF Catalogue: a collaborative platform for sharing and reusing biomedical data. International journal of medical informatics. 2019;126:35-45.