Automated Genome Indexing for Metagenomic Analysis

Automatisierte Genom-Indexierung für metagenomische Analyse

Kristian Nikolov*, Boris Kirov*, †, Vassil Galabov*, †, ‡

* BioInfoTech, RDIC, Sofia, Bulgaria, bioinfotechlab@gmail.com
† Faculty of Automation, TU – Sofia, Sofia, Bulgaria
‡ FDIBA, TU – Sofia, Sofia, Bulgaria

Abstract — Metagenomics analysis is a novel technique with key importance for the development of modern ecology, biomedicine and food safety. This method is tightly connected to the advance of express genome sequencing, namely the "shotgun" techniques, and depends directly on the efficiency of automated data analysis methods. In the latter years some serious breakthroughs in the field have been made both in reduction of the size of the sample DNA needed for analysis (as low as 1 ng), as well as in the efficiency of the analytic software and its practically complete independence from the way reference libraries are built. In this paper we present the next step towards complete automation of the whole metagenomics analysis, namely the development of automated genome indexing software tool that greatly reduces the amount of labor required in the analysis process.

Zusammenfassung — Metagenomikanalyse ist eine neuartige Technik, die für die Entwicklung der modernen Ökologie, Biomedizin und Lebensmittelsicherheit von zentraler Bedeutung ist. Diese Methode ist eng mit dem Fortschritt der Express-Genom-Sequenzierung verbunden, nämlich der "Shotgun"-Techniken und hängt direkt von der Effizienz der automatisierten Datenanalyse ab. In den letzten Jahren wurden einige ernsthafte Durchbrüche auf diesem Gebiet sowohl in der Verringerung der Größe der für die Analyse benötigten DNA (so niedrig wie 1 ng) sowie in der Effizienz der analytischen Software und ihrer praktisch vollständigen Unabhängigkeit gemacht die Art und Weise, wie Referenzbibliotheken gebaut werden. In diesem Beitrag stellen wir den nächsten Schritt zur vollständigen Automatisierung der gesamten Metagenomik-Analyse vor, nämlich die Entwicklung eines automatisierten Genom-Indexing-Software-Tools, das den Arbeitsaufwand im Analyseprozess stark reduziert.

I. Introduction

Microorganisms are omnipresent in the natural human environment. There are numerous species everywhere, including air, water, soil, even our food. Unfortunately, many of those microbes are pathogenic. Therefore, it is imperative for us to be able to efficiently identify the microbial composition of different samples.

The complete species composition in a foodstuff or soil material, etc. samples is studied by Metagenomics and could be detected using a DNA-Based method, which involves (Fig. 1.):

- Extraction – procedure used to isolate DNA from the nucleus of cells. The purification uses a combination of physical and chemical methods;
- fragmentation – separation of DNA strands into pieces via a variety of methods involving mechanical breakage;
- sequencing – a process of determining the primary structure of DNA (the order of guanine, adenine, thymine and cytosine nucleotides within a DNA molecule).

Genomics is the science which studies the complete set of genes belonging to each organism. It uses a variety of methods, which are commonly referred to as "standard sequencing". The latter is very efficient for sequencing the full genomes of unknown organisms. However, for the detection of species in a sample, traces of standard known organisms are looked for, not their complete genomes. In order to achieve that, a more resource-saving and innovative approach is used [1]. Instead of the whole genome, small parts (i.e. "reads") of it are sequenced and then compared to a database of known reference genomes. Read classification puts the separated and yet read pieces of DNA strands back together using the data obtained through the application of random DNA shotgun sequencing method [2]. Thus, applying a software pipeline for quantitative measurement and taxonomic profiling, we are able to identify organisms found in the sample [3].

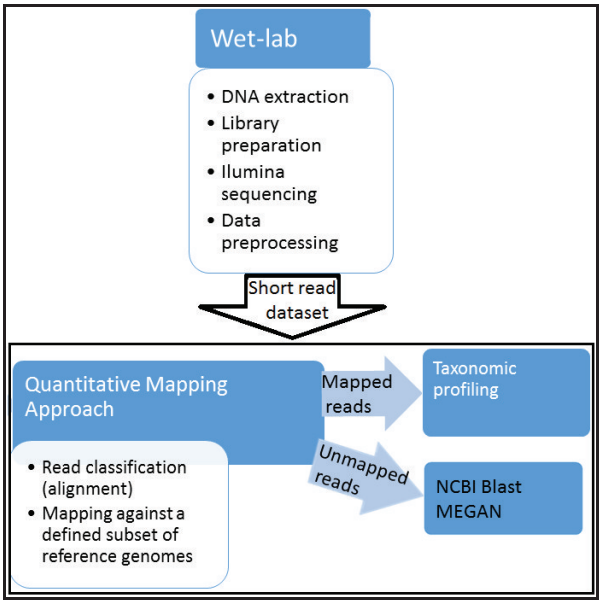

Fig. 1. Pipeline of the metagenomic analysis process.

The software, by itself, does the aligning and profiling of the reads comparing them to a database of full genomes. In order to detect a larger number of species – more genomes need to be added to the database. The latter is a time-consuming process, which could take a significant amount of time, attention, effort and energy, and also depends on internet connection speed, computer performance and size of the needed genome.

II. Genome Indexing

A. Pre-indexing

Full genomes are readily available in the NCBI (National Center for Biotechnology Information) database. They have to be downloaded in the FASTA file format (.fasta/.fna), which is a text-based format for representing nucleotide or peptide sequences using single-letter codes. Full genomes are commonly compressed and their size usually varies between 1MB and 1GB. The decompressed file size might expand up to 4GB.

B. Indexing

For the sake of time saving, every genome has to be indexed using a compressing algorithm, known as Burrows-Wheeler transform [4]. This algorithm scans for patterns, allowing mismatches and idles (which are consequence of random spontaneous mutations occurring naturally), which makes searching through genomes faster. A single indexing may require up to 50 minutes on a system with an Intel i5-4590 Quad-Core 3.30GHz CPU with 16GB of RAM and operating 64-bit Linux Mint 18.1. The number of iterations depends on the length of the indexed genome, often exceeding 100000.

C. Post-indexing

The sequential part is simple, but requires some patience and time, for the process is effort-consuming and labor-intensive. It involves many steps, all of which require one's supervision and attention. Sorted-pointer libraries are transferred to the analyzing software's database after a minor identifier and format conversion. The indexed genome's name needs to be included in a list for recognition. Usually, the process takes about 15 minutes to be completed, making a total of 25-65 minutes for a single organism. Given that, for proper analysis to be performed at least 20 indexed genomes are needed, the total amount of time required might exceed 10 hours.

III. Results

We have developed a script that (in the following order) (Fig. 2.):

- automates the process of full genome downloading – reaches the NCBI server and downloads all given paths freeing the user from any interactions;
- prepares the input data – decompresses the downloaded genome packs, using Gzip (GNU General Public License) and qualifies them for the next step;
- indexes the decompressed genome – runs the genome indexing software for every given genome;
- prepares the output data for the usage of the analyzing software – converts indexed genome files into the required from the metagenomic software format;
- loads formatted output data into the database – creates a directory for every organism and transfers the files into it, putting their names on the reference genome list;
- prevents future errors in case of failures;
- gives an exact count of the successfully indexed genomes and the total count of all attempted genomes;
- exerts control over the successful completion of every step and terminates the execution of the script for the failed genomes only.

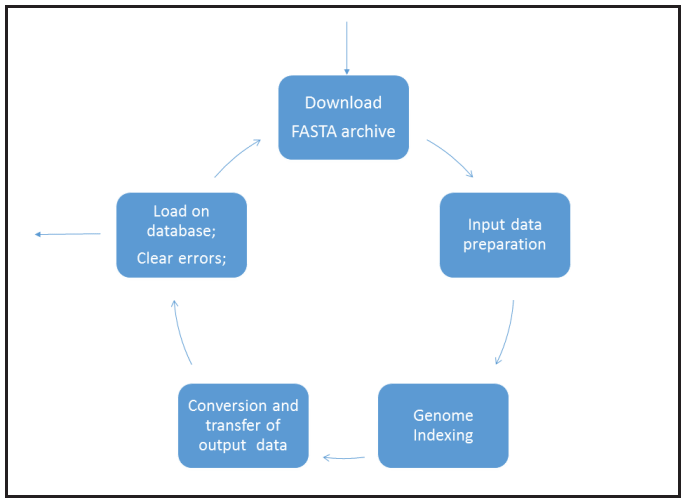

Fig. 2. Workflow of the developed script.

Thus, the user is left with the only responsibility to prepare a list of download URLs of the needed genome's FASTA files, to add their names next to each URL, and finally to run the script. In total, those actions would require only few minutes of labor overall for any number of organisms in comparison with the 15 minutes for each as described above. Therefore, we conclude that the proposed script is an efficient and useful addition to the existing software tools for automated metagenomic analysis, which could save a considerable amount of labor and time if used.

Acknowledgments

The work reported here has been conducted as part of the scientific program of BioInfoTech lab – RDIC, Sofia Tech Park. The authors thank everyone, who helped overcoming any difficulties and staying focused on the goal.

References

[1] Thomas et al., "Metagenomics - a guide from sampling to data analysis", *Microbial Informatics and Experimentation 2012*, 2:3

[2] Ripp et al., "All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing", *BMC Genomics 2014*, 15:639

[3] Liu et al., "AFS: identification and quantification of species composition by metagenomic sequencing", *Bioinformatics*, 2017, 1–3

[4] Burrows, Michael; Wheeler, David J., "A block sorting lossless data compression algorithm", 1994, *Technical Report 124*, Digital Equipment Corporation, Palo Alto, CA