Egyetemi doktori (PhD) értekezés tézisei

Computational Methods and Applications for Real-Time Identification of Species and Pathogens from Raw Read Sequencing Data.

Ramin Karimi

Témavezető: Dr. András Hajdu

DEBRECENI EGYETEM
Informatikai Tudományok Doktori Iskola
Debrecen
2017

Table of Contents

1. Introduction

The number of bacterial cells (10^{14}) inhabiting in an average healthy adult human body is estimated tenfold more than human cells (10^{13}) [1][2] and the number of existing microbial species in the world is estimated at 10^7 to 10^9 [3]; Therefore, they play a pivotal role not only in Human life but also the whole life on Earth. Over 99% of microorganisms in our planet are not cultivable in vitro or require a long and difficult cultivation period [4]; therefore, how can we discover these unseen occupiers of our body and our planet? What is the appropriate solution for the rapid identification of these best Friends and the worst enemies?

Polymerase Chain Reaction (PCR) is a quick, easy, and inexpensive laboratory technique to amplify a single copy or a few copies of a small fragment of target DNA (the template) to unlimited copies. PCR can synthesize any particular piece of DNA.

The invention of the PCR technology provided a new era of studying uncultivable microorganisms, but there was still a long way ahead. PCR has some limitations.

1- Amplification inhibitors such as detergents, antibiotics, enzymes, polysaccharides, fats, proteins, and salts can reduce the amplification efficiency [5]

2- PCR is a highly sensitive; any form of contamination of the sample can produce false positives or false negative results [6].

3- Another limitation of PCR is the length of the fragment that can be amplified. PCR works well over short stretches of DNA up to about 2 kbp.

4- PCR can only be used to identify the presence or absence of a known pathogen or gene [7].

5- PCR cannot be used to identify species in complex communities.

Later with emerging the microarray approaches, identifying uncultivable microorganisms entered into a new stage. The microarray is the combination of a very large set of distinct probes attached to a solid structure (Glass slides). Probes are small fragments of sequences

which are complementary to a pathogen-specific gene sequence [8]. According to the reports, microarray techniques bring the possibility of species identification or detecting and diagnose of various bacterial samples at the same time with main advantages of high throughput, parallelism, miniaturization, speed, and automation.

Microarray has been limited to a small set of functional genes such as 16S rRNA genes and it is not a suitable approach to investigate the uncultivable majority of the species in the environment [8].

With respect to remarkable abilities of PCR and microarray approaches, they are not enough powerful for studying complex environmental and clinical samples which contain hundreds or thousands of different species.

With the advancement of Next-generation sequencing (NGS) which is a combination of massively parallel sequencing technologies besides PCR and microarray techniques, considerable progress has occurred not only in the phylogenetic and functional analysis of microbial communities but also in their affiliated science, significantly. It is a culture-free method that enables analysis of the entire microbial community within a sample. It has the ability to combine many samples in a sequencing run.

16S ribosomal RNA (16S rRNA) gene with about 1500 bp length is part of DNA and generally contain nine "hypervariable regions" (V1 – V9) that represent considerable sequence diversity among different bacterial species and can be used for species identification [9].

16S rRNA is the most common standard culture-free approach which is currently used for taxonomic assignments, bacterial identification, and studying bacterial diversity in ecology and clinical microbiology [10]. It is also used to design the primers for Polymerase Chain Reaction (PCR) and probes for microarray studies.

There are a considerable number of publications about limitations of 16S rRNA gene:

The major limitation is that the copy numbers of 16S rRNA per genome vary from 1 up to 15 or more copies [10][11]. Therefore, the amount of 16S rRNA variants is estimated to

be 2.5-fold greater than the number of bacterial species [10][12]. Moreover, 16S sequences of the same species or even the same genome are often different [10].

The ambiguous and incorrect identity of species and also the artificial classification of an organism into more than one species can be led by divergent evolution of rRNA genes [13]. This problem can be solved in cultivable species by cloning rRNA genes from the pure culture of that species to identify the degree of variation [13]. As mentioned above, more than 99% of species are culture-independent, so in a complex and mixed community of microorganisms, sequence heterogeneity of 16S rRNA within a single genome can lead overestimation of microbial diversity [12].

Short reads as the output of the Next-generation high-throughput sequencing technology are very noisy and partial, with too many missing parts. Moreover, most of the reads from recent NGS platforms are too short in length [12], thus, de novo assembly is required in order to make longer sequences. It may represent an extra limitation to use 16S rRNA fragments for taxonomic assignments. It can be argued that in the case of reads with the short length, 16S rRNA is more efficient to identify a higher level of taxonomic assignments such as phyla, classes, orders, families, and genera than species or strains [14].

As an example, *Escherichia coli* is a bacterial species with several strains. Some of these strains have nearly similar 16S rRNA-encoding genes but have very dissimilar functional capabilities [15].

The application of next-generation sequencing technologies has provided a set of technical innovation called “Metagenomics” as a culture-free method to study the genetic content of all organisms in a community obtained directly from their natural environment

Debility of 16S rRNA to identify and especially rapid identification of microorganisms in the metagenomics reads is more visible.

This doctoral research is intended to answer the following questions:

1. What is the proper alternative alignment-free method for rapidly identifying the species and strains from raw read sequencing data?
2. What is the proper method of finding DNA signatures (alternative agents) from genome databases?
3. What is the proper method of short reads classification?
4. What is the proper method to use less computational resources?

The key contribution of this research is the development of an alternative fast and cost-effective method to allow identifying species and strains from raw read metagenome sequencing data, regardless of aforementioned limitations and without further processes such as assembling and alignment.

This method involves two pipelines with multiple stages and complex computerizes applications such as Hadoop and Hive in a cluster of low-cost nodes, using parallel and distributed computing. The second contribution of this research has tended to develop the required applications with the automated process in order to facilitate the method to be applicable to the entire research community.

The motivation behind this research is developing alignment-free approaches, not only to shortcut identification into a quick and accurate process using parallel and distributed computing on commodity hardware, but also for other purposes in bioinformatics and metagenomics studies such as the accurate estimation of microbial community composition based on metagenomics sequencing data, the alignment and assembly of short reads, and other Next-generation sequencing analysis.

2. Proposed Methods

2.1. Proposed Method of DNA Signature Discovery

There are two common types of sequence-based identification methods, the alignment-based and the alignment-free methods. The inability of alignment-based approaches for

rapid identification purposes caused a necessity for shifting into alignment-free approaches as an alternative method.

Among the alignment-free approaches, the most popular option is to use marker genes such as the 16S rRNA gene. We discussed the limitations of using this gene for the analysis of complex metagenomics sequencing data.

In this research, we proposed using DNA signature as an alternative fast and cost-effective method to allow a rapid identifying the species and the strains from raw read metagenomics sequencing data, regardless of aforementioned limitations.

DNA signature is a short k-mer oligonucleotide fragment with an arbitrary length k, which is unique or specific to a particular group of species selected from a target genome database. There are two categories of unique and common signatures according to the purpose of usage. The presence of a unique DNA signature in any volume of sequences and genetic materials represents the existence of the corresponding species. Therefore, signature discovery is the action of finding specific fragments of the genome in a database [16]. Any pipeline, application or algorithm which is designed for DNA signature discovery, has to detect an entire database or multiple databases recursively. The procedure varies according to the purpose of using DNA signatures. It has already been used, to design primers and probes for PCR and microarray assays.

A major advantage of the use of DNA signatures is the gene-independent and alignment-free nature of this approach [17]. DNA signatures are well suited to the analysis of the sequences with lacking a robust estimate of phylogenetic relationships and to the analysis of complex sequences such as metagenomics sequences that alignment-based methods often perform poorly [18]. Sequence analysis of high-throughput technologies with DNA signature is easier than the other methods. The number of DNA signatures and their specificity increase with adding the length of signature. It causes a wide flexibility of using the method.

Regarding a large number of DNA signatures in different species and the possibility to choose arbitrary lengths of them for identification, this approach is suitable not only for

PCR and microarray-based assays but also has great potential for next-generation sequencing analysis. The flexibility to choose targeted and non-targeted databases and an arbitrary length of signatures are other advantages that allow reducing the cost of sequencing by lower-coverage sequencing. Since the length of signatures is short, the size of the reads does not a serious matter.

Due to the large size of databases, most existing methods of DNA signature discovery require significant computational resources; they are not applicable to the entire research community. A large amount of RAM and CPU capacity and long execution times are major limitations of most of the methods that are based on pattern comparison and pairwise alignment of the genomes. The determination of the mismatch tolerance level as a discovery condition also influences the results.

In some cases, it is necessary to load the whole dataset into the main memory for searching for unique or common signatures. When the size of the data exceeds the available memory, the execution will fail. For the sequential algorithms, increasing the number of CPU cores does not increase the discovery efficiency of the algorithm. Another limitation for most of the existing methods is the lack of efficiency to find both unique and common signatures simultaneously. Most of them are capable of finding only DNA signatures of a single genome. Another limitation of some of these methods is the lack of the possibility to select an arbitrary length (k) for the signatures.

The additional challenge as another major limitation for DNA signature discovery methods is the lack of option in the choice of targeted and non-targeted genome databases.

To overcome the aforementioned challenges of DNA signature discovery, we proposed our method as a powerful pipeline. The pipeline **HTSFinder** (High-throughput signature finder) [16] has been designed in order to enhance the usability of DNA signatures for massively high-throughput sequencing analysis.

The pipeline HTSFinder has significant advantages compared with the existing DNA signature discovery pipelines and algorithms.

- First, HTSFinder is capable of detecting all unique, common, and maximal group coverage signatures of the entire database or multiple databases simultaneously.
- Second, it becomes possible to select target and non-target genome databases, based on user requirements. For instance, we have the ability to use both forward and reverse-complement genome sequences of a database for detecting DNA signatures.
- Third, the pipeline can consider either a cluster of low-cost computer nodes that are commonly available in research facilities or a high-performance computer (HPC).
- Finally, the flexibility of the different phases of the pipeline makes it suitable for other Bioinformatics and metagenomics studies such as Next-Generation Sequencing (NGS) analysis.

HTSFinder consists of three computational phases as shown in Figure 1. This pipeline generates all the possibilities of k-mers for every genome individually and then determines their frequency in the entire database. Finally, DNA signatures of every species or strain are obtained in the database or multiple databases that have been involved in the pipeline. HTSFinder implements the parallel and distributed computational tool Hadoop for the second and third phases.

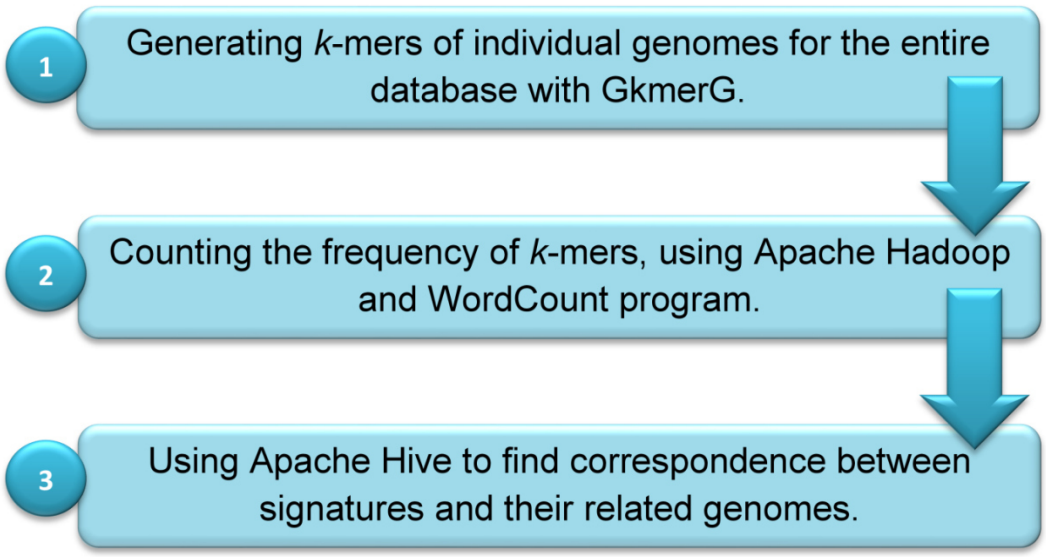

Figure 1. The three main phases of HTSFinder for detecting DNA signatures. We can repeat the second phase with the obtained results if required

The first phase of the pipeline is carried out by GkmerG (Genome k-mer Generator) that is a software designed to obtain all the possibilities of k-mers of genome sequences with FASTA format (*.fna or *.fa).

Figure 2 illustrates the split of the genome by GkmerG. Concatenating the files, sorting k-mers and removing all duplicates except one are the last steps of GkmerG. For the species with multiple chromosomes and some bacterial genomes that are comprised of multiple chromosomes and plasmids, GkmerG concatenates them into a single file before sorting at the end of the first phase.

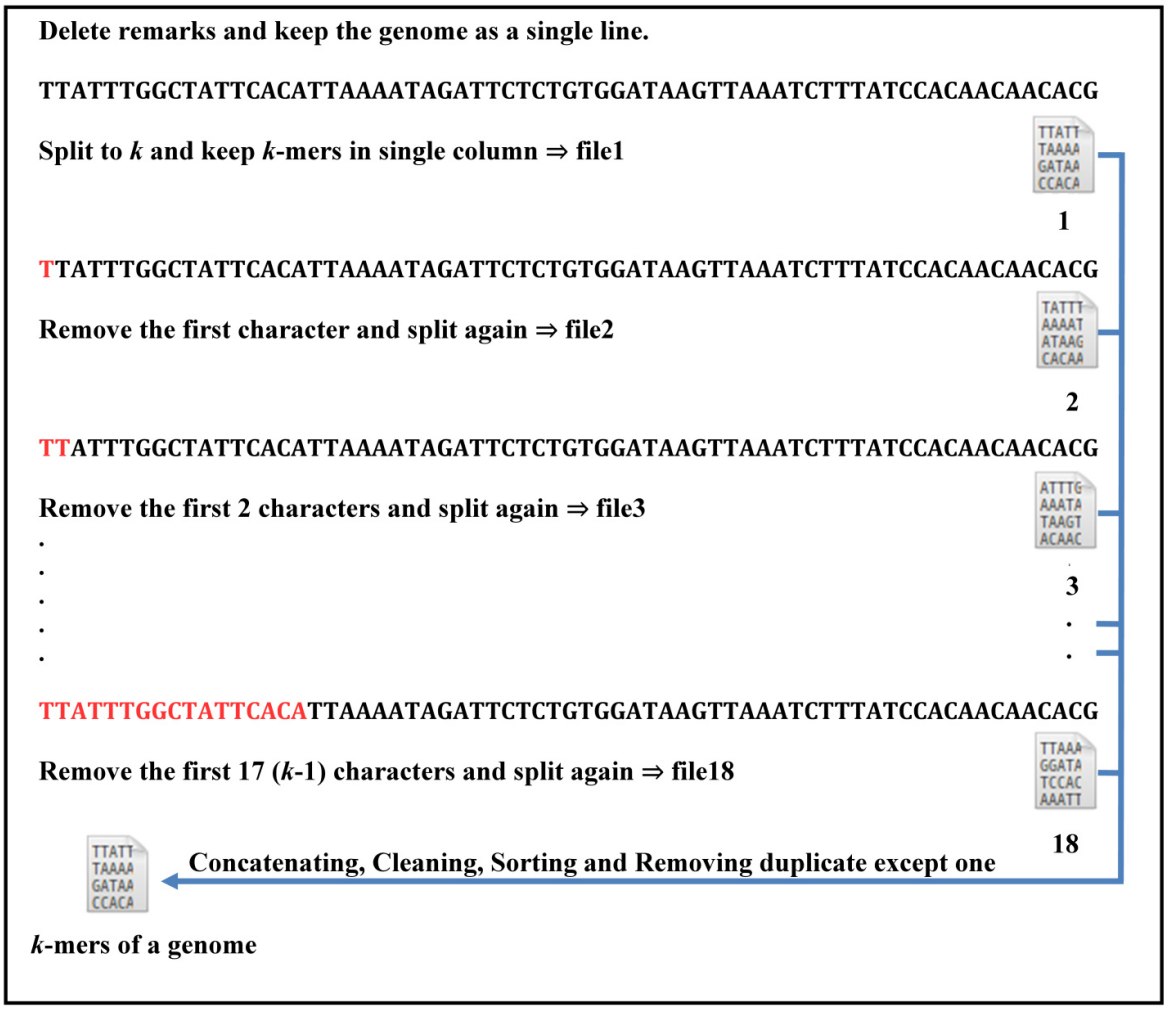

Figure 2. Splitting of the genome by GkmerG for k = 18 to get all the possibilities of 18-mers. Generating k-mers for a single genome with GkmerG includes purgation, splitting, concatenation, cleaning, sorting, and removing duplicate except one. The output of GkmerG is a file containing k-mers of a genome in a single column. The labels above the file numbers in this figure represent the beginning of four k-mers in the head of files.

The output of GkmerG is the input for the second stage of the pipeline. For designing this software, the method of splitting genome sequence to all the possibilities of k-mers is inspired from De Bruijn graph. It is an efficient way to represent all the possible k-mers (subsequences of length k) of a sequence. These types of graphs are important because of their usefulness in the reconstruction of genomic sequences and are used in most of the applications for de novo assembly of short reads. The n-dimensional Bruijn graph of "m" characters is a directed graph representing overlaps between sequences of characters. The graph has m^n vertices consisting of all possible k-mers of the given sequence of characters present in the reads sequence. For example, given the alphabet comprising A, T, G and C, there are 4^3 = 64 nucleotides of length k=3 [19]. In a string with length S, The amount of all possible k-mers is {S – k + 1}. As an example in Figure 3 the length of the sequence is S = 17, therefore the amount of all the possibilities of k=7 are: S – k + 1 = 17 – 7 + 1 = 11

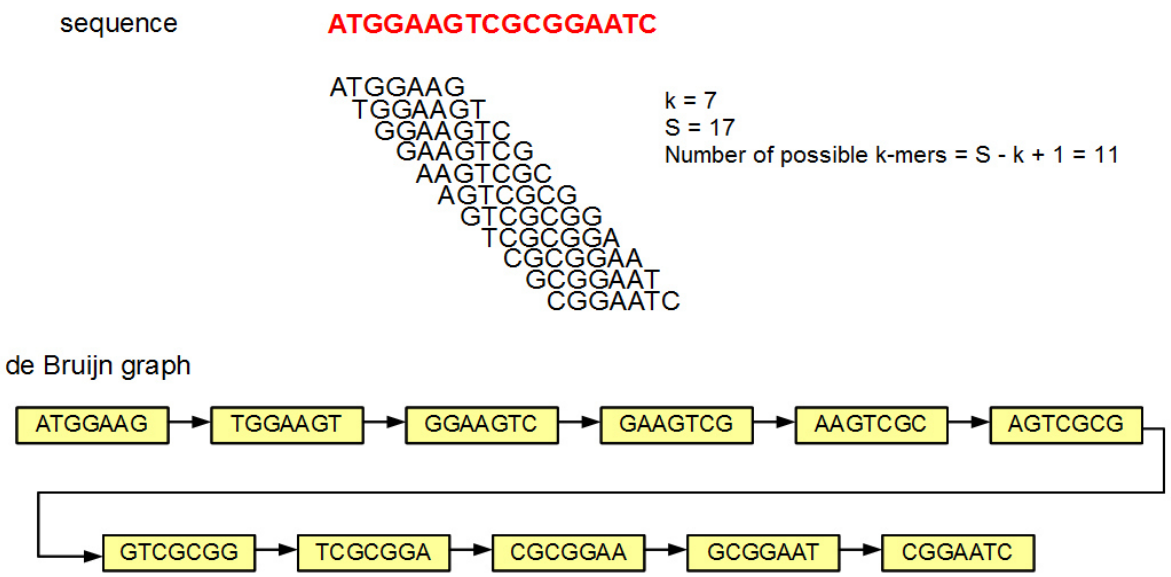

Figure 3. An example of de Bruijn graph.

In the second stage of the pipeline, we used the Hadoop framework and WordCount program to count the frequency of k-mers in very large genome datasets. The result of this stage is a large file containing sorted and a non-duplicate list of k-mers obtained from all the genomes generated in the first step. In the second stage, we can extract all unique

signatures or group-specific signatures due to the frequency, but we cannot determine the owner of the signatures. For instance, the 18-mer with frequency=8 in the first row of Table 1 means that this 18-mer occurs in 8 genomes among the 2,773 bacterial ones, while the 18-mer in the fourth row is a unique signature in the database.

Table 1. An example of Hadoop and WordCount results in the second stage.

Signatures or 18-mers	Frequency in the database
AAAAAAAAAAAAAAAGAG	8
AAAAAAAAAAAAAAAGAT	25
AAAAAAAAAAAAAAAGCA	20
AAAAAAAAAAAAAAAGCC	1
AAAAAAAAAAAAAAAGCG	5
AAAAAAAAAAAAAAAGCT	6
AAAAAAAAAAAAAAAGGA	9
AAAAAAAAAAAAAAAGGC	3
AAAAAAAAAAAAAAAGGG	6
AAAAAAAAAAAAAAAGGT	38

The input for the third phase of the pipeline is the output of the first and second phases. In the third phase of the pipeline we use Apache Hive and HiveQL queries in order to extract all the unique signatures of a specific species in the database or extract group-specific signatures which are common between 2, 3, 4, etc. Hive lets the user process large datasets with relatively little effort and in a reasonably short time. This research proves the efficiency of Hive to handle querying on billions of rows in a table or multiple tables. Due to the flexibility of querying in Hive, there are various ways to create the tables and design the queries in the third stage. Optimization techniques are very effective for reducing the time-consumption and computational resource usage in a Hadoop cluster. Running time, CPU usage, memory usage, and speed of disk reading of the nodes are the subjects of optimization in a Hadoop cluster.

The final output of this pipeline is a table containing the signatures with the desired length and the reference number that indicate the owner of the signature. The pipeline can handle very large data sets (e.g. 287.85 GB data in a single run).

2.2. Proposed Methods of Short Reads Classification

In this doctoral research, we have proposed two different methods of short reads classification for matching the DNA signatures, reads, and their related species. Unlike the other existing methods that have focused on the speeding up the process using High-performance computers or large cluster of nodes, our goal is to process the huge analysis using ordinary desktop computers that are available everywhere with the aid of parallel and distributed computing.

2.2.1. The First Proposed Reads Classifier Based on Bitmap Indexes and NoSQL

In this method, we use optimization techniques borrowed from database technology, namely bitmap indexes. They are used to speed up searching and matching of billions of DNA signatures in the short reads of thousands of different microorganisms, using commodity High-performance computing, such as Hadoop MapReduce and Hive [20].

Bitmap Index is an efficient way to speed up the queries and improve performance in data warehouse environments, which contain tables with low cardinality columns. As the example given in Table 2, we index the values of the column Grade having low cardinality. In this case, our index has the same number of rows and the number of columns is equal to the number of distinct values in column Grade. In Table 2, the cardinality of the column Grade is 4 because we have 4 different values in it.

Table 2. An example of a bitmap index defined on Grade column.

RID	Name	Nationality	Grade
1	John	FRANCE	B
2	Sara	USA	D
3	Piter	RUSSIA	C
4	David	ENGLAND	A
5	Tania	GERMANY	B
6	Daniel	POLAND	A
7	Tom	CANADA	C
8	Robert	ITALY	C
9	Jain	FRANCE	D

RID	A	B	C	D
1	0	1	0	0
2	0	0	0	1
3	0	0	1	0
4	1	0	0	0
5	0	1	0	0
6	1	0	0	0
7	0	0	1	0
8	0	0	1	0
9	0	0	0	1

Bitmap index techniques are used to create the index table by searching the existence of signatures in short reads. '1' represents the existence of the signature in the short reads and '0' represents non-existence. This process is done with Java programming.

Table 3 is an example of the created index table with keeping every bacterium as a column. We store '1' if any signature of the bacteria exists in a short read, '0' if not.

Table 3. An example of input table for Hive. Each column of "0" and "1" is owned by a specific species. The first 2 columns represent the reads and their identification numbers.

1	R1	0	0	0	1	0
2	R2	1	0	0	0	0
3	R3	0	0	0	1	0
4	R4	0	0	0	0	0
5	R5	0	0	1	0	0
6	R6	0	0	0	0	1
7	R7	1	0	0	0	0
8	R8	0	0	0	0	0
9	R9	0	0	0	0	0
10	R10	1	0	0	0	0

Although in this method, the running time of the queries is very short; constructing the index files for each bacterium is time-consuming. This problem becomes more visible when a large number of bacterial species are considered (such as the real metagenomics samples). Therefore, we were motivated by the need to make an extra effort to come up with better and more creative solutions to address this problem.

2.2.2. The Second Proposed Reads Classifier Based on SRIdent Pipeline

This pipeline is based on generating k-mers from the short reads and searching the existence of DNA signatures in the Reads k-mers, by using Apache Hive data warehousing. RkmerG (Read k-mers Generator) is a software presented in this study, for producing k-mers of the short reads.

The **SRIdent** pipeline [21] consists of two computational stages. Data preparation is the first stage which is done by RkmerG. DNA signatures with specific length (k) of every individual known species and short reads generated by sequencing technologies are two types of data that are used in this pipeline. Figure 4 described the algorithm of RkmerG.

```
for r in file
        do
                remove information line of each r
                make each r as a single line
                copy file to reference-file
                add line number to reference-file
                copy file to tmp-file
        done
while read line {1,...,n} in tmp-file
        do
                for i in {1,...,k}
                        do
                                copy line for k times
                                delete i-1 first character of lines
                                split lines to k-mers
                                delete mers<k
                                add reference number to k-mers
                                move k-mers to output
                if end of the tmp-file
                        then
                                exit 0
                fi
        done
done
```

Figure 4. RkmerG algorithm. (r=Short read, n=Total number of reads, k=Length of mers, i=each of the lines (reads) that is copied from a single short read for k times, illustrated in figure 15.)

In the final output of this pipeline, we can find the signature, the owner of the signature, and the short read containing the signature. In fact, we can classify the reads according to the appearance of the signatures in reads. In the other word, we can find the species presented in the raw reads sequencing samples.

2.3. Computational Resources of the Proposed Methods

Although a dramatic reduction in the cost and the time-consumption of sequencing technologies have led the applicability of them as a routine tool for diagnostic and public health microbiology, the computational analysis is still a barrier. Using less computational resources was a primary goal in this research; therefore, in the proposed methods we have applied parallel and distributed computing on commodity hardware to achieve this goal. Table 4 contains the computational resources used by existing methods that are using the same data format with HTSFinder and SRIdent [16][21]. As it is shown in this table we could process the analysis with ordinary cheap computers.

Table 4. Comparison of the computational resources of our two pipelines with the existing methods.

Name	Data Format	adopted Platform according to the publication
TOFI	FASTA	64 x 1.5 GHz Itanium 2 processors running on Linux with 64 GB of shared memory
TOPSI	FASTA	98-cores Linux cluster
Insignia	FASTA	192-node Linux cluster
Kaiju	FASTA	HP Apollo 6000 System ProLiant XL230a Gen9 Server, with two 64-bit Intel Xeon E5-2683 2 GHz CPUs (14 cores each), 128 GB DDR4 memory
Kraken	FASTA	48 AMD Opteron 6172 2.1 GHz CPUs and 252 GB of RAM, running Red Hat Enterprise Linux 5.
CLARK	FASTA	Dell PowerEdge T710 server, dual Intel Xeon X5660 2.8 GHz, 12 cores, 192 GB of RAM
HTSFinder Multi-nodes	FASTA	The master node: Intel Core2 Quad CPU Q6600 at 2.40 GHz and 8 GB of RAM and 6 slaves Intel Core i3-2100 CPU at 3.10 GHz with 4 GB of RAM for each.
SRIdent Multi-nodes	FASTA	The Master node and 4 Slave nodes. All with 4 GB of RAM, Intel Core i3- 2100 CPU at 3.10GHz

3. Suggestions and Future Works

The overriding purpose of this research was to overcome the limitations in the computational analysis of rapid diagnostic, identification, and characterization of species and infectious pathogens from raw reads sequencing data, in particular, complex metagenomics data. This research has focused on 2 main primary goals:

1- Proposing a fast, flexible, and independent alignment-free method for real-time identification of microorganisms from High-throughput sequencing reads.
2- Reducing the cost and time-consumption of the computational resources by the use of parallel and distributed computing and related optimization techniques, in order to utilize ordinary desktop computers for Big Data analysis.

We believe that the ability to use sequencing technologies and their related analysis as a routine process in the medical and biological laboratories will revolutionize the medical, biological and environmental science in the near future.

Developing automated and powerful software and applications regarding the goals of this research are strongly suggested.

References

[1] Savage, Dwayne C. "Microbial ecology of the gastrointestinal tract." Annual Reviews in Microbiology 31, no. 1 (1977): 107-133.

[2] Gerritsen, Jacoline, Hauke Smidt, Ger T. Rijkers, and Willem M. Vos. "Intestinal microbiota in human health and disease: the impact of probiotics." Genes & nutrition 6, no. 3 (2011): 209.

[3] Curtis, Thomas P., William T. Sloan, and Jack W. Scannell. "Estimating prokaryotic diversity and its limits." Proceedings of the National Academy of Sciences 99, no. 16 (2002): 10494-10499.

[4] Wagner, Michael, Rudolf Amann, Hilde Lemmer, and Karl-Heinz Schleifer. "Probing activated sludge with oligonucleotides specific for proteobacteria: inadequacy of culture-dependent methods for describing microbial community structure." Applied and environmental microbiology 59, no. 5 (1993): 1520-1525.

[5] Al-Soud, Waleed Abu, and Peter Rådström. "Capacity of nine thermostable DNA polymerases to mediate DNA amplification in the presence of PCR-inhibiting samples." Applied and environmental microbiology 64, no. 10 (1998): 3748-3753.

[6] Bolognia, Jean L., Dennis L. Cooper, and Earl J. Glusac. "Toxic erythema of chemotherapy: a useful clinical term." Journal of the American Academy of Dermatology 59, no. 3 (2008): 524-529.

[7] Garibyan, Lilit, and Nidhi Avashia. "Polymerase chain reaction." Journal of Investigative Dermatology 133, no. 3 (2013): 1-4.

[8] Schrenzel, Jacques, Tanja Kostic, Levente Bodrossy, and Patrice Francois. "Introduction to Microarray-Based Detection Methods." Detection of Highly Dangerous Pathogens: Microarray Methods for BSL 3 and BSL 4 Agents (2009): 1-34.

[9] Van de Peer, Yves, Sabine Chapelle, and Rupert De Wachter. "A quantitative map of nucleotide substitution rates in bacterial rRNA." Nucleic acids research 24, no. 17 (1996): 3381-3391.

[10] Klindworth, Anna, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. "Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies." Nucleic acids research 41, no. 1 (2013): e1-e1.

[11] Klappenbach, Joel A., Paul R. Saxman, James R. Cole, and Thomas M. Schmidt. "rrndb: the ribosomal RNA operon copy number database." Nucleic acids research 29, no. 1 (2001): 181-184.

[12] Acinas, Silvia G., Luisa A. Marcelino, Vanja Klepac-Ceraj, and Martin F. Polz. "Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons." Journal of bacteriology 186, no. 9 (2004): 2629-2635.

[13] Pei, Anna Y., William E. Oberdorf, Carlos W. Nossa, Ankush Agarwal, Pooja Chokshi, Erika A. Gerz, Zhida Jin et al. "Diversity of 16S rRNA genes within individual prokaryotic genomes." Applied and environmental microbiology 76, no. 12 (2010): 3886-3897.

[14] Siqueira Jr, Jose F., Ashraf F. Fouad, and Isabela N. Roˆcas. "Pyrosequencing as a tool for better understanding of human microbiomes." Journal of oral microbiology 4 (2012).

[15] Welch, R. A., V. Burland, G. Plunkett, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. T. Mobley, M. S. Donnenberg, and F. R. Blattner. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 99:17020-17024.

[16] Karimi, Ramin, and Andras Hajdu. "HTSFinder: Powerful Pipeline of DNA Signature Discovery by Parallel and Distributed Computing." Evolutionary bioinformatics online 12 (2016): 73.

[17] Ogilvie, Lesley A., Lucas D. Bowler, Jonathan Caplin, Cinzia Dedi, David Diston, Elizabeth Cheek, Huw Taylor, James E. Ebdon, and Brian V. Jones. "Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences." Nature communications 4 (2013).

[18] Pride, D. T., Wassenaar, T. M., Ghose, C. & Blaser, M. J. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 7, 8 (2006).

[19] Rødland, Einar Andreas. "Compact representation of k-mer de Bruijn graphs for genome read assembly." BMC bioinformatics 14, no. 1 (2013): 313.

[20] Karimi, Ramin, Ladjel Bellatreche, Patrick Girard, Ahcene Boukorca, and Andras Hajdu. "BINOS4DNA: Bitmap Indexes and NoSQL for Identifying Species with DNA Signatures through Metagenomics Samples." In International Conference on

Information Technology in Bio-and Medical Informatics, pp. 1-14. Springer International Publishing, 2014.

[21] Karimi, R., and Hajdu, A., "SRIdent: A novel pipeline for real-time identification of species from high-throughput sequencing reads in Metagenomics and clinical diagnostic assays." In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, pp. 6481-6484. IEEE, 2015.

List of Submitted Publications

1- Karimi, R., Hajdu, A. “How Can Bring the Big Data Analysis of High-Throughput Sequencing Technologies into the Routine Clinical Diagnostic Assays?” Journal of Annales Mathematicae et Informaticae (submitted) 2017.

2- Tótha, A., Karimi, R. “Optimization of Hadoop Cluster for Analyzing Large-scale Sequence Data in Bioinformatics.” Journal of Annales Mathematicae et Informaticae (submitted) 2017.

Registry number: DEENK/194/2017.PL
Subject: PhD Publikációs Lista

Candidate: Ramin Karimi
Neptun ID: UDTVQ8
Doctoral School: Doctoral School of Informatics

List of publications related to the dissertation

Foreign language international book chapters (2)

1. **Karimi, R.**, Hajdu, A.: SRIdent: A novel pipeline for real-time identification of species from high-throughput sequencing reads in metagenomics and clinical diagnostic assays. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, 2015. Proceedings, IEEE, [Piscataway], 6481-6484, 2015.
2. **Karimi, R.**, Bellatreche, L., Girard, P., Boukorca, A., Hajdu, A.: BINOS4DNA: Bitmap indexes and NoSQL for identifying species with DNA signatures through metagenomics samples. In: Information Technology in Bio- and Medical Informatics : 5th International Conference, ITBAM 2014, Munich, Germany, September 2, 2014. Proceedings. Ed.: Miroslav Bursa, Sami Khuri, M. Elena Renda, Springer International Publishing, Cham, 1-14, 2014, (Lecture Notes in Computer Science, ISSN 0302-9743 ; 8649) ISBN: 9783319102658

Foreign language scientific articles in international journals (1)

3. **Karimi, R.**, Hajdu, A.: HTSFinder: Powerful pipeline of DNA signature discovery by parallel and distributed computing.
   Evol. Bioinform. 12, 73-85, 2016. EISSN: 1176-9343.
   DOI: http://dx.doi.org/10.4137/EBO.S35545
   IF: 1.5

Total IF of journals (all publications): 1,5
Total IF of journals (publications related to the dissertation): 1,5

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of Web of Science, Scopus and Journal Citation Report (Impact Factor) databases.

27 June, 2017