Classification of DNA Sequences based on Data Analysis

Bacem Saada, *Member, IAENG*, Jing Zhang

***Abstract*— The technological progress of recent years has led biologists to extract, examine and store huge quantities of genetic information. Thus, the databases become very large and contain a huge amount of redundant or poorly analyzed information. Through this article, we propose a comparative study based on the percentages of similarity between the optimal alignment algorithm and the taxonomic classification of eight genera from Bacillales order to detect if they are similar or not.**

***Index Terms*— Bioinformatics, DNA sequence alignment, Data analysis, High Throughput Sequencing;**

I. INTRODUCTION

Nowadays, large amounts of existing information on databases are used, shared, analyzed and stored. Powerful computers are used to properly analyze this content and bioinformatics as a new field of research that analyzes cells of living beings has been introduced. A DNA sequence is a biomolecule present in all living being cells. it contains the genetic information required for the functioning and development of all living beings. Each nucleotide is composed of a nitrogenous base; adenine (A), cytosine (C), guanine (G) or thymine (T). Online databases (GENBANK, EMBL, etc.) store gigabytes of DNA sequences. This quantity continues to grow to reach the double in only 18 months. Every day, new strains are sequenced and identified as belonging to a specie. Complex algorithms are used to properly analyze this data. In this paper, we will discuss the relationship between species. We studied 33 species of 8 genera from the order Bacillales. By calculating the matrix of similarity between them, we will present some errors of taxonomic classification between those species.

Manuscript received December 14, 2016; revised January 05, 2017.

Bacem Saada, Ph.D. Student with Harbin Engineering University, College of Computer Science and Technology, Harbin, China, (email:basssoum@gmail.com).

Jing Zhang, Professor with Harbin Engineering University, College of Computer Science and Technology, Harbin, China, (email: zhangjing@hrbeu.edu.cn).

II. STATE OF THE ART

Some researchers have attempted to reduce the complexity of the dynamic programming algorithms. Some of them have tried to reduce the total execution time of the algorithm by exploiting multicore architecture graphics processors NVidia Cuda technology and optimizing the use of these graphics processors [1,2].

From another point of view, researchers have endeavored to compress DNA sequences. The DNA compression is based on text compression algorithms. However, researchers proved that conventional text compression algorithms are insufficient for DNA sequences compression.

Therefore, they proposed specific compression algorithms. For example, Biocompress [3] seeks repetitions in a sequence. Biocompress-2[4] uses a Markov model to compress non-repetitive regions of a DNA sequence. By applying these algorithms to the standard benchmark data [5], the compression ratio is 1.85 BpB for Biocompress and 1.78 BPB for biocompress-2.

Other researchers presented an algorithm for dividing a genome to several parts and detecting similar regions to reduce the alignment operations between nucleotides [6].

Recent algorithms used new techniques to convert the DNA sequence to hexadecimal representation and detect regions of similarities between a set of sequences [7].

Saada, B. and Zhang, J. introduced an algorithm that detects the longest common chain for a set of sequences belonging to the same genus and uses it as representative of the whole set [8].

III. CLASSIFICATION OF SPECIES BASED ON DATA ANALYSIS APPROACHES

The main uses of DNA sequence alignment algorithms are the examination of the percentage of similarity between sequences, their classification in order of similarity and the checking of the conformity of these algorithms to the classification methods used by biologists.

A large part of genetic data should be analyzed using mathematical and computational tools to properly identify the observations and conclusions that describe the data representation and the relationships between them.

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Some of the most used methods for biological data analysis are:

- The Factorial Analysis of Correspondences,
- Ascending Hierarchical Classification.

1.Similarity Percentage for Smith and Waterman algorithm

In this part we will present a study of the Smith and Waterman algorithm based on the multidimensional scaling and hierarchical classification method approaches. We will present the species used in our study and analyze the results. Finally, we will present a comparative study of the classification of species resulting from these two approaches and the taxonomic classification of biological species.

1.1 Evaluation metrics

In order to analyze the hierarchical classification of species made by taxonomists, we decided to use DNA sequences from several genera but which are all from the same order. In this way, it is possible to analyze the:

- **Genus discrimination:** detect if the species of the same genus are grouped together.
- **Tolerance to the similarity errors:** detect if the alignment between an unidentified sequence and a sequence whose species is known allows to clearly identify the exact species of this new sequence.
- **Relationships between species of the same order:** Check if the choice of species of the same order will influence the classification of these species or not.

The table below includes all the species used in our analysis:

Table 1: number of species of every genus

Genus	Alicyclobacillus	Anoxybacillus	Bacillus
Species	2	3	12
Genus	Geobacillus	Lactobacillus	LysiniBacillus
Species	3	3	2
Genus	Paenibacillus	Sporosarcina	
Species	7	1	

To conduct this study, we used the software XLSTAT which is a data analysis tool integrated to Excel.
The computer used for this analysis has the following features:

- Processor Intel i3 2200Mhz,
- Ram 4 Go,
- Windows 7 Version Ultimate 64 bits.

1.2 Multidimensional positioning

In this part we will present the analysis made using the multidimensional scaling approach by presenting the type of the data presented and the interpretation of results.

Data Preparation

We have developed an algorithm which is based on the Smith and Waterman algorithm that allows to draw a matrix of similarity containing the percentages of similarity between the different DNA sequences.

Experimental results

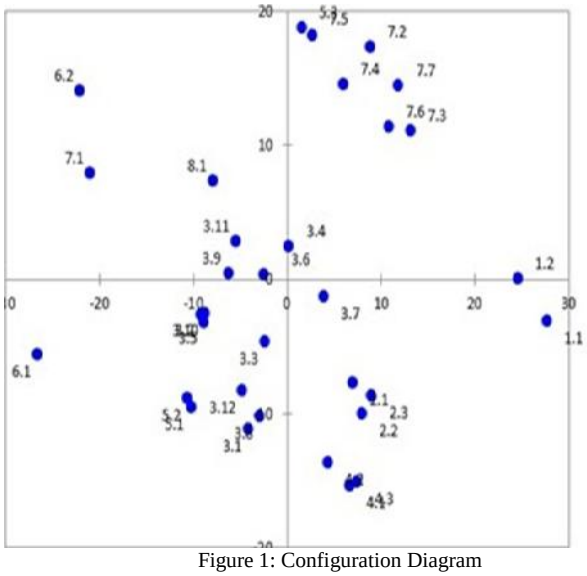

Figure 1: Configuration Diagram

Through this figure (fig. 1) We noticed that some species of the same genus are scattered and not grouped together. Thus, they have low similarity percentages. This brings into question the similarity between these species and leads to even say that a review of the classification made by taxonomists is highly recommended. We also noticed that some sequences of different genera are closely gathered. Hence, they have high similarity percentages.

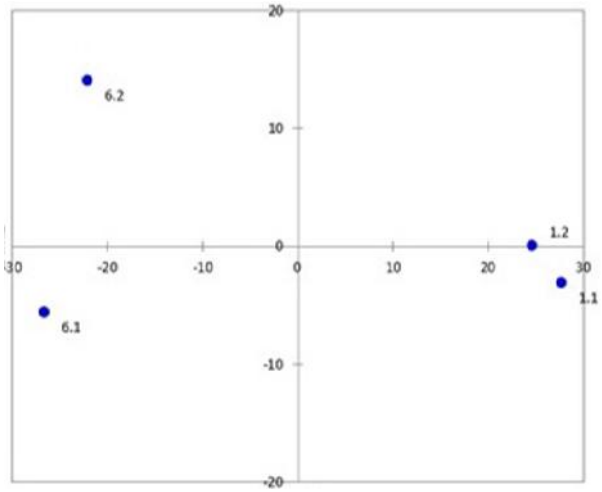

Figure 2: Species 1.1, 1.2, 6.1 and 6.2

By analyzing the above figure (Fig.2), we can conclude that some species of same genus are closely grouped (such as

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

species of the genus Alicyclobacillus [1.1 and 1.2]). Paradoxically, others, from the same genus, are far situated from each other (such as Lysinibacillus [6.1 and 6.2]).

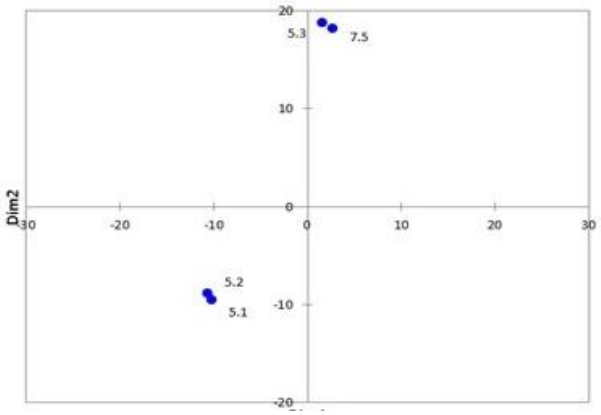

Figure 3: Species 5.1, 5.2, 5.3 et 7.5

The second example (Fig. 3) shows that the species Lactobacillus Paraplatarum [5.3] is closer to the species Paenibacillus polymyxa [7.5] than to the other sequences of the genus Lactobacillus [5.1 and 5.2].

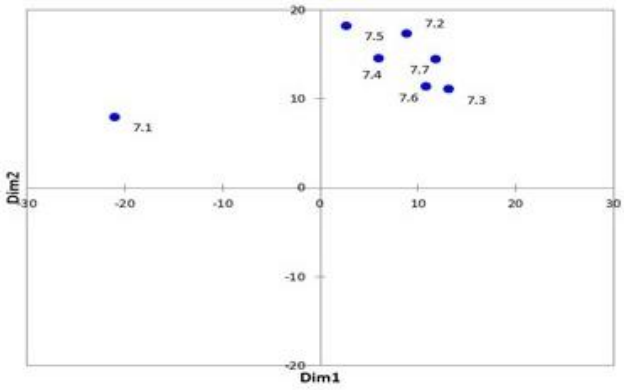

Figure 4: Species of the genus Paenibacillus

We also noted that for the genus Paenibacillus, which is represented by seven different species (fig.4), six are closely grouped while the seventh species (Paenibacillus azotofixans [7.1]) is situated quite far from them.

The big problem here is that if we want to associate an unidentified DNA sequence to a species, the result may be wrong.

To determine if the error rate is important, we will consider the following two factors:

- Probability of error: It represents the probability that the association with a case would be wrong.
- Frequency of occurrence in the data banks: in databases, each species is represented by a definite number of sequences. If the probability of error is large and the frequency of occurrence is high, the percentage of identification error will be higher.

2. Classification Errors

In this part we will present three cases showing the influence of the probability of error in the results of species classification. The appearance of DNA sequences rate will be calculated from The Ribosomal Database Project [9].

A. Lactobacillus Paraplatarum (5.3) and Paenibacillus Polymyxa (7.5)

According to Figure 3, we see that these two species (A. Lactobacillus Paraplatarum (5.3) and Paenibacillus Polymyxa (7.5)) are nearly coincident. The most probable hypothesis is that it was a wrong identification of one of these species which was identified as a new one.
Moreover, by consulting the database, we noted that the Lactobacillus Paraplatarum is represented by only two strains while Paenibacillus Polymyxa is represented by 214 strains.

B. Paenibacillus azotofixans (7.1) and Lysinibacillus sphaericus (6.2)

According to the scatter plot shown in Figure 1, these two species (Paenibacillus azotofixans (7.1) and Lysinibacillus sphaericus (6.2)) are quite close to each other. Paradoxically, the species Lysinibacillus sphaericus (6.2) is closer to the specie Paenibacillus azotofixans (7.1) than to the other species of the genus Lysinibacillus (6.1).
With regard to the frequency of occurrence in the database, the Lysinibacillus sphaericus appears 202 times in the database while the Paenibacillus azotofixans appears only 21 times.

C. Paenibacillus azotofixans (7.1) and other species of the genus Paenibacillus

If an identification operation is done with the different species of the Paenibacillus genus, the possibility of error might occur too.

All Paenibacillus sequences represent 2548 strains. Among them, 21 strains represent the species Paenibacillus azotofixans and since this species is closer to the Lysinibacillus sphaericus species, analysis can deduce that this strain belongs to the genus Lysinibacillus not to the genus Paenibacillus.

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

IV. Conclusion and Future Work

As a conclusion, we can say that the hierarchical classification made by taxonomists is not conform to the statistical approaches classification and does not respect the similarities percentages of the execution of optimal algorithms of DNA sequences alignment. This classification method might lead to errors of identification of a not yet identified strain.

In our future work, we will attempt to use DNA sequences compression methods for detecting the common data between a set of DNA sequences and verify if this degree of similarity is conforming to the taxonomic similarity or not.

ACKNOWLEDGMENT

This paper is funded by the International Exchange Program of Harbin Engineering University for Innovation-oriented Talents Cultivation.

REFERENCES

[1] Yongchao Liu, Douglas L. Maskell, Bertil Schmidt: CUDASW++: optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units". BMC Research Notes,2009, 2:7

[2] Yongchao Liu, Bertil Schmidt, Douglas L. Maskell: CUDASW++2.0: enhanced Smith-Waterman protein database search on CUDA-enabled GPUs based on SIMT and virtualized SIMD abstractions". BMC Research Notes, 2010, 3:93

[3] Matsumoto, T., Sadakane, K., Imai, H., et al., 2000, Can General-Purpose Compression Schemes Really Compress DNA Sequences?, Computational Molecular Biology, Universal Academy Press, 76–77.

[4] Grumbach S. and Tahi F.: Compression of DNA Sequences. In Data compression conference, pp 340-350. IEEE Computer Society Press, 1993.

[5] Grumbach, S., Tahi, F.: A new Challenge for compression algorithms: genetic sequences. Journal of Information Processing and Management 30, 866–875 (1994).

[6] Granger G. Sutton, Owen White, Mark D. Adams, and Anthony R. Kerlavage. Genome Science and Technology. 1995, 1(1): 9-19. doi:10.1089/gst.1995.1.9.

[7] Saada, B., & Zhang, J. (2015, November). DNA sequences compression algorithms based on the two bits codation method. In Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on (pp. 1684-1686). IEEE.

[8] Saada, B., & Zhang, J. (2015). Vertical DNA Sequences Compression Algorithm Based on Hexadecimal Representation. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 2).

[9] Bonnie L. Maidak, Gary J. Olsen, Niels Larsen1, Ross Overbeek2, Michael J. McCaughey and Carl R. Woese.The RDP (Ribosomal Database Project) .Department of Microbiology, University of Illinois.

ISBN: 978-988-14048-3-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)