Brazilian Journal of Biology®

ISSN 1519-6984 (Print)
ISSN 1678-4375 (Online)

Original Article

K-mer applied in *Mycobacterium tuberculosis* genome cluster analysis

K-mer aplicado na análise de agrupamento de genomas de *Mycobacterium tuberculosis*

Leila Maria Ferreira[a]*, Thelma Sáfadi[a] and Juliano Lino Ferreira[b]

[a]Universidade Federal de Lavras, Departamento de Estatística, Lavras, MG, Brasil
[b]Empresa Brasileira de Pesquisa Agropecuária, Embrapa Pecuária Sul, Bagé, RS, Brasil

Abstract

According to studies carried out, approximately 10 million people developed tuberculosis in 2018. Of this total, 1.5 million people died from the disease. To study the behavior of the genome sequences of *Mycobacterium tuberculosis* (MTB), the bacterium responsible for the development of tuberculosis (TB), an analysis was performed using k-mers (DNA word frequency). The k values ranged from 1 to 10, because the analysis was performed on the full length of the sequences, where each sequence is composed of approximately 4 million base pairs, k values above 10, the analysis is interrupted, as consequence of the program's capacity. The aim of this work was to verify the formation of the phylogenetic tree in each k-mer analyzed. The results showed the formation of distinct groups in some k-mers analyzed, taking into account the threshold line. However, in all groups, the multidrug-resistant (MDR) and extensively drug-resistant (XDR) strains remained together and separated from the other strains.

Keywords: DNA word frequency, genome, similar sequences.

RESUMO

De acordo com estudos realizados, cerca de 10 milhões de pessoas desenvolveram tuberculose em 2018. Desse total, 1,5 milhão de pessoas morreram devido à doença. Procurando estudar o comportamento das sequências do genoma da *Mycobacteruim tuberculosis* (MTB), bactéria responsável por desenvolver a Tuberculose (TB), foi realizada uma análise aplicando o *k-mer* (frequência de palavras do DNA). Os valores de *k* variaram de 1 a 10, pois devido a análise ter sido feita no comprimento total das sequencias, onde cada sequencia é composta por aproximadamente 4 milhões de pares de bases, valores de *k* acima de 10, a análise é interrompida, como consequência da capacidade do programa. O intuito do trabalho foi de verificar a formação da árvore filogenética em cada *k-mer* analisado. Os resultados obtidos evidenciaram a formação de grupos distintos em alguns *k-mers* analisados, levando-se em consideração a linha de corte. Entretanto, em todos os grupos formados as cepas multidroga resistente (MDR) e extensivamente resistente à droga (XDR) permaneceram juntas e separadas das demais cepas.

Palavras-chave: frequência de palavras do DNA, genoma, sequências similares.

1. Introduction

Mycobacterium tuberculosis (MTB), also known as Koch's bacillus, is highly contagious, airborne, slow growing, Gram stain positive, rich in GC content, aerobic, and rod shaped and has an unusual layer of wax on its cell surface due to mycolic acid. Its cell wall has a high lipid content, which allows the bacteria to survive inside macrophages, providing the organism with a barrier resistant to many common medicines. Humans are MTB's main host (Subhasree et al., 2017). Infection is transmitted by airborne spread of aerosolized bacteria containing 1-to-5 µm diameter droplet nuclei that transport MTB from an individual with Tuberculose (TB) infection to an uninfected individual. Infectious droplet nuclei are inhaled and localized to the distal airway alveoli. MTB is then absorbed by alveolar macrophages, initiating a cascade of events that result in the successful containment of infection or progression to active TB disease. The risk of developing active disease varies with time since infection, age and host immunity. Koch's bacillus has the ability to remain dormant for many years. Once active, TB attacks the respiratory system and other organs, destroying body tissues (Koch et al., 2018).

TB persists as a global health concern because it infects individuals who generally do not adhere to treatment for 6 months or more (Gagneux et al., 2006). This circumstance is particularly true in the developing world, where over 95% of infections occur (Niemann et al., 2009; Delogu et al.,

*e-mail: leilamaria2003@yahoo.com.br
Received: November 15, 2021 – Accepted: May 26, 2022

2013). Noncompliance with treatment has contributed to the current TB pandemic, increasing the probability of transmission and sustaining the development of resistant strains (Bertholet et al., 2010). Thus, more advanced studies of these resistant strains are needed to develop vaccines that effectively control the disease, which often require high research costs (Tacconelli et al., 2018; Walker et al., 2018).

The current state of affairs can make a parallel approach with Covid 19 pandemic. A disease of the coronavirus presents manifestations clinics similar to others in infections also transferred via the airways, such as pulmonary tuberculosis (TB). Visca et al., 2021 affirm that TB is still a lethal and neglected disease in the Covid 19 era. They substantiate this claim by saying little progress has been made with the prevention, diagnosis, and treatment up to date, despite the Who declaring this disease a global emergency in 1993, while the few investments in research funds in Brazil were negligible over the years. If in tuberculosis, the incorrect treatment causes resistant strains. The novel coronavirus causes severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, also called COVID-19 and 2019-nCoV). In light of this, the virus continues to surround the people who do not take the vaccines correctly, and more sinister, it mutates (Sharifipour et al., 2020). In that regard, the present SARS-CoV2, like TB, poses an increased risk to global health due to its high speed of spread, which induces mutations and consequently causes severe forms of respiratory illness (Campos et al., 2020). Genetic mutations of viruses are a common phenomenon. Currently, 4,000 mutations in the spike protein of SARS-CoV-2 have already been identified, and most of them do not affect the virus regarding its ability to spread or cause disease (Hossain et al., 2021). The SARS-CoV-2 already exceeded the Ebola virus outbreak, continuously escalating, with over 14 million cases and more than half-million deaths confirmed in more than 180 countries, mainly in the USA, Brazil, India, Russia, and Peru (Pratas and Silva, 2020). Maia et al. 2022, in a letter to the editor, address the impact of the COVID-19 pandemic on tuberculosis in Brazil. The data extracted from the Brazilian public database - the Ministry of Health, reveal that the average number of TB consultations jumped from 2017 to 2019 from 48,688 to 108,269 in 2020 (Covid-19 pandemic): an increase of 122.4%. However, a reduction in confirmed cases of pulmonary TB in Brazil occurred, with a decrease of 6,501 reported cases from the period of 2017 to 2019 (a drop of 7,9%). Consider that TB is well-known to be a seasonal disease. According to the findings of that letter, Brazil experienced different levels of interruption of the health system, which may result in a reduction in the total notifications of pulmonary TB. Furthermore, during the pandemic period, essential services for TB were restricted due to the decrease in resources and supplies, prioritizing the mitigation of COVID-19; a similar fact occurred in other countries, according to the authors. That analyzed data also appoint an increase in treatment dropouts and deaths from TB during the Covid-19 pandemic.

There are two types of MTB drug resistance: genetic resistance and phenotypic resistance (Zhang and Yew, 2015). Genetic drug resistance is due to mutations in chromosomal genes in growing bacteria, while phenotypic resistance or drug tolerance is due to epigenetic changes in gene and protein expression, which causes drug tolerance in nongrowing bacteria (Lange et al., 2018). These two types of resistance are responsible for several problems in the effective control of TB, with genetic resistance, as present in multidrug-resistant (MDR) or extensively drug-resistant (XDR) strains, causing problems worldwide. The subtler phenomenon of phenotypic drug resistance or tolerance is classified as persistent, implying prolonged treatment and risk of relapse after treatment (Campaniço et al., 2018).

Over recent years, there has been an increase in the incidence of TB in developing and industrialized countries due to the widespread emergence of drug-resistant strains and synergy with human immunodeficiency virus (HIV) infection, as appointed by the studies published to date by Dalcolmo (2000), Kwan and Ernst (2011), Pawlowski et al. (2012), Maiti and Maiti (2021), Sharma et al. (2021) and WHO (2021). In 2000, official estimates from the WHO's Global Health Estimates, HIV infected some 33 million people. Even though the data validating this notion may be questioned, a third of the world's population was also infected with tuberculosis bacillus in an active or latent form (Dalcolmo, 2000). Kwan and Ernst (2011) make evident the syndemic relationship between HIV and TB and captured a positive linear Pearson correlation (r) of 0,799 from the data of 132 countries in 2008. They also conclude that the HIV-TB syndemic has disproportionately impacted people in Africa. Later, another study said that about 14 million people were dually infected in 2010. In that scenery, TB accounts for roughly 26% of AIDS-related deaths, the most significant single cause of death in the context of AIDS (Mazurek et al., 2012). According to Maiti and Maiti (2021), TB is the most common opportunistic infection among humans infected with HIV. Thus a synergistic interaction occurs between HIV and *Mycobacterium*; each emphasizes the progression of the other. MDR-TB remains a public health crisis and a health security threat. In 2020, only about one in three people with drug-resistant TB accessed treatment (WHO, 2021). TB occurs in every part of the world. The most significant number of new TB cases occurred in the South-East Asian Region, with 43% new cases, followed by the African Region, with 25% new cases, and the Western Pacific with 18%. Approximately 86% of new TB cases occurred in the 30 high TB burden countries. Eight countries accounted for two-thirds of the new TB cases: India, China, Indonesia, the Philippines, Pakistan, Nigeria, Bangladesh, and South Africa (WHO, 2021).

Radical measures are needed to prevent dire WHO predictions from coming true. The combination of genomics and bioinformatics has the potential to generate the information and knowledge that will enable the design and development of new therapies and interventions needed to treat this airborne disease and to elucidate the unusual biology of its etiological agent, MTB (Cole et al., 1998)

Given the large amount of genomic data, alignment-free sequence comparison methods are required due to their low computational complexity. Most methods without alignment are based on word frequencies. For a word length, a k is fixed, thus calculating a frequency of (relative) words, which will compose a vector for each of the input sequences (Leimeister et al., 2014). By sequencing

additional complete genomes, it becomes possible to move from theoretical to empirical studies and examine the properties of DNA words and how their distributions vary among different species or elements of genomes. The most basic empirical question that has been investigated is k-mers (Chor et al., 2009). K-mer-based methods can improve comparison accuracy by extracting an effective feature from genome sequences (Han and Cho, 2019). Several works have been published using this method, such as (Singh et al., 2017; Cheng et al., 2013; Li et al., 2010; Ondov et al., 2016; Wang, 2013; Yin and Yau, 2015)

The purpose of this paper is the use of k-mer analysis to verify the position of the words (nitrogenous bases of the genome) of each sequence related to the MTB genome in order to identify the phylogenetic tree. From this analysis, it will be possible to visualize sequences that are similar to each other.

2. Material and Methods

The sequences (MTB strains) used in this work were extracted from the National Center for Biotechnology Information (NCBI) website and are described in Table 1 (NCBI, 2018). These same strains were also used in (Saini and Dewan, 2016; Ferreira et al., 2017, 2018, 2020).

2.1. K-mer analysis

K-mers are substrings of length k contained within a biological sequence. Used mainly in the context of computational genomics and sequence analysis, k-mers are composed of nucleotides (i.e., A, T, G and C), to assemble DNA sequences, improve gene expression, identify species in metagenomic samples, create attenuated vaccines, etc. Generally, the term k-mer refers to all subsequences in length of a sequence k, such that the sequence given by AGAT would have four monomers (A, G, A, and T), three {2-mers} (AG, GA, and AT), two {3-mers} (AGA and GAT) and one {4-mer} (AGAT). More generally, a sequence of length L will have L-k + 1 k-mers, with n^k being a possible total of k-mers, where n is the number of possible monomers, i.e., four in the case of DNA (Huang, 2016).

Table 2 shows the possible k-mers of a DNA sequence.

According to Allman et al. (2017), given A is a sequence in the letter alphabet L, $L = \{1,2,\ldots,L\}$; to a natural number k, given Y is the vector of the k-mer counts extracted from A. Consequently, for each $B = b_1 b_2 \ldots b_k \in L^k$, the coordinate Y^B records the number of times B occurs as a contiguous substring in A. A standard k-mer method calculates a distance between two sequences A_1 and A_2 of length n_1 and n_2 by first calculating their respective k-mer vectors Y_1 and Y_2 and then computing the Euclidean squared distance, according to Equation 1.

$$\left|Y_1 - Y_2\right|_2^2 = \sum_{B \in L^k} \left(Y_1^B - Y_2^B\right)^2 \quad (1)$$

Analyzing two sequences descending from a common ancestor, they go through a base substitution process described by standard assumptions of phylogenetic modeling. We can assume that one of the sequences, A_1, is ancestral to the other, A_2, and their locations are assigned states in L according to an i.i.d (independently

Table 1. Description of each of the sequences analyzed.

Sequences	Description of strains	Total length of sequences
Seq1_DS	Strain was isolated in Russia belonging to the AI family (according to RFLP genotyping) and it is sensitive to all common drugs used in the treatment of tuberculosis.	4,398,525
Seq2_DS	Susceptible strain representing the largest portion of patients' tuberculosis isolates recovered during an epidemic in the Western Cape of South Africa.	4,424,435
Seq3_DS	Susceptible strain belonging to the Beijing family, sequenced for comparative genomic studies.	4,398,812
Seq4_DR	Resistant strain isolated in 2004, referring to a patient with secondary pulmonary tuberculosis, sequenced for comparative genomic studies.	4,405,981
Seq5_DR	Drug resistant strain, having an accelerated rate of transmission between humans under agglomeration conditions.	4,408,224
Seq6_MDR	Strain correspond to a single patient in KwaZulu-Natal, South Africa.	4,398,250
Seq7_XDR	Strain correspond to a single patient in KwaZulu-Natal, South Africa.	4,399,120
Seq8_DS	Susceptible strain used for comparative genomic studies.	4,414,325
Seq9_DS	Susceptible strain derived from the original human lung H37, isolated in 1934. It has been widely used all over the world in biomedical research. Unlike some clinical isolates, it retains total virulence in animals with tuberculosis and is susceptible to drugs and receptive to genetic manipulation.	4,411,532
Seq10_DS	A virulent susceptible strain derived from its virulent parent strain H37 (isolated from a 19 year old male patient with chronic pulmonary tuberculosis, named Edward R. Baldwin in 1905). This strain was obtained through an aging and dissociation process of an in vitro culture in the year 1935.	4,419,977

and identically distributed) process with state probability vector $\pi = \left(\pi^b\right)_{b \in L}$.

Moreover, π is the stationary distribution of a transition probability matrix R, dimension $L \times L$, describing the process of state change from a single location from sequence A_1 to sequence A_2. In the context of continuous time models with a D rate matrix and time (or branch length) t, $R = \exp(Dt)$. Thus, the probability of a k-mer $B = b_1 b_2 \ldots b_k \in L^k$ in any consecutive k of a single sequence is $\pi^B = \prod_{j=1}^{k} \pi^{b_j}$. The k-mer vectors Y_1 and Y_2 are random variables that summarize A_1 and A_2.

The method corresponds to a frequency analysis of words (nitrogenous bases of the genome) called k-mers. The package used was {Kmer}, where we worked with the kdistance function (Wilkinson, 2019). The k values tested ranged from 1 to 10. The analysis was performed by spanning the entire genome, where each sequence is over 4 million base pairs.

To perform the analysis, the free program R was used (R Core Team, 2019).

Table 2. K-mers for GTAGAGCTGT.

k	k-mers
1	G, T, A, G, A, G, C, T, G, T
2	GT, TA, AG, GA, AG, GC, CT, TG, GT
3	GTA, TAG, AGA, GAG, AGC, GCT, CTG, TGT
4	GTAG, TAGA, AGAG, GAGC, AGCT, GCTG, CTGT
5	GTAGA, TAGAG, AGAGC, GAGCT, AGCTG, GCTGT
6	GTAGAG, TAGAGC, AGAGCT, GAGCTG, AGCTGT
7	GTAGAGC, TAGAGCT, AGAGCTG, GAGCTGT
8	GTAGAGCT, TAGAGCTG, AGAGCTGT
9	GTAGAGCTG, TAGAGCTGT
10	GTAGAGCTGT

3. Results

In Figure 1A, the formation of two distinct groups is shown, obtained by the analysis with the value of k = 1; that is, the genome was traversed verifying each letter, which, in this case, are the nitrogenous bases.

The first group formed contains multidrug-resistant (MDR) and extensively drug resistant (XDR) strains, which correspond to the MTB genomes of most resistant strains. The second group contains drug-resistant (DR) strains along with drug-susceptible (DS) strains.

As shown in Figure 1A, with respect to the largest group formed, the Seq3_DS and Seq10_DS sequences stay together, as well as the Seq2_DS and Seq9_DS sequences. The sequences Seq1_DS, Seq5_DR, Seq4_DR and Seq8_DS appear isolated.

In Figure 1B, the results with k = 2 are shown. We verified the formation of three groups according to the cut line. The first group contains the sequences Seq6_MDR and Seq7_XDR. The second group contains the sequence Seq1_DS alone. The third group contains the sequence Seq5_DR alone. The fourth group contains the sequences Seq4_DR, Seq8_DS, Seq2_DS, Seq10_DS, Seq3_DS and Seq9_DS.

With respect to Figure 2A, using k = 3, the groups formed looking at the threshold line were as follows: the first group contained the sequences Seq6_MDR and Seq7_XDR; the second group contained the sequence Seq1_DS; the third group contained the sequence Seq5_DR; the fourth group contained the sequence Seq4_DR; and the fifth group contained the sequences Seq2_DS, Seq8_DS, Seq3_DS, Seq9_DS and Seq10_DS.

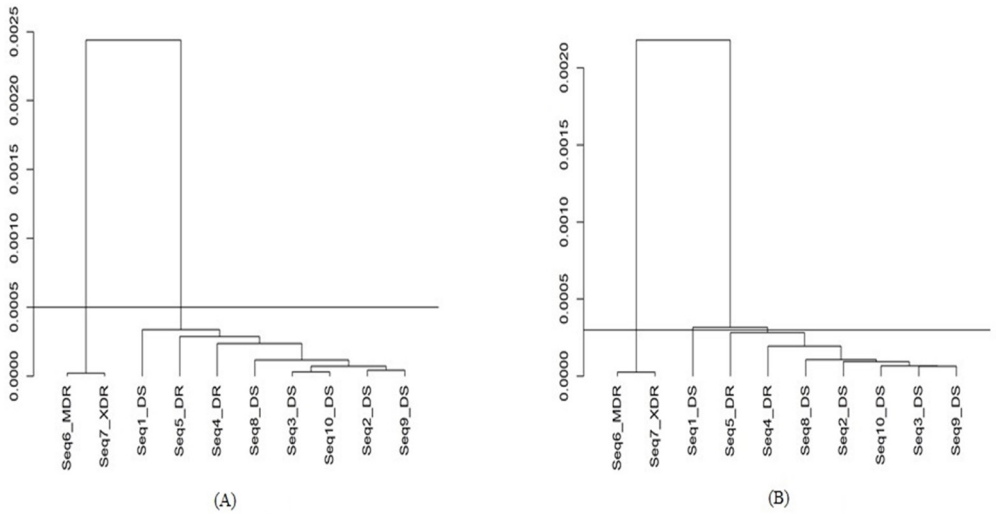

Figure 1. Formation of groups using k-mer analysis, (A) with k = 1, (B) with k = 2.

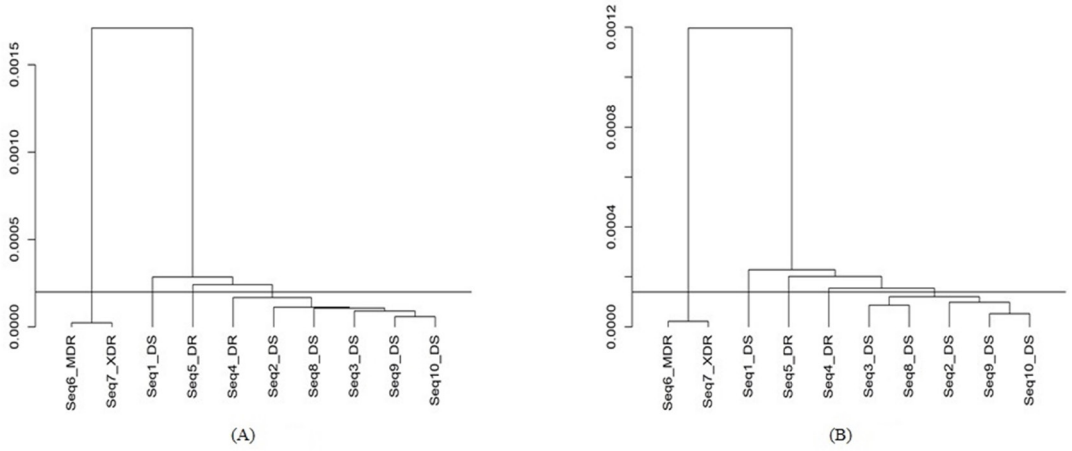

Figure 2. Formation of groups using k-mer analysis, (A) with k = 3, (B) with k = 4, 5, 6 and 7.

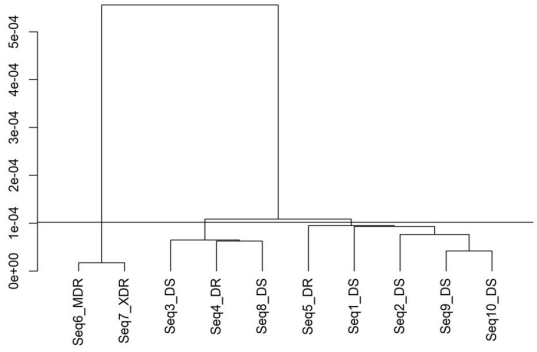

Figure 3. Formation of groups using k-mer analysis, with k = 8, 9 and 10.

Verifying the results shown in Figure 2B, with k = 4, 5, 6 and 7, the results were the same for these values of k. Taking into account the threshold line, the following groups were identified: the first group contained the sequences Seq6_MDR and Seq7_XDR; the second group contained the sequence Seq1_DS; the third group contained the sequence Seq5_DR; the fourth group contained the sequence Seq4_DR; the fifth group contained the sequences Seq3_DS and Seq8_DS; and the sixth group contained the sequences Seq2_DS, Seq9_DS and Seq10_DS.

As shown in Figure 3, with k = 8, 9 and 10, the same results were obtained for these k values. Analyzing the threshold line, we verified the formation of three groups. The first group contained the sequences Seq6_MDR and Seq7_XDR. The second group contained the sequences Seq3_DS, Seq4_DR and Seq8_DS. The third group contained the sequences Seq5_DR, Seq1_DS, Seq2_DS, Seq9_DS and Seq10_DS.

4. Discussion

The result shown in Figure 1 was also found in the work of (Saini and Dewan, 2016). The authors used was the discrete decimated wavelet transform (DWT) methodology, employing the Haar wavelet at five levels of decomposition. The extraction of MTB genome characteristics was performed using GC content with a 10,000 base pair sliding window. To verify the formation of the groups, they used the total energy from the DWT.

Already at work of Ferreira et al. (2017), it was used discrete nondecimated wavelet transform (NDWT) methodology, using the Daubechies wavelet with 4 null moments at five levels of decomposition. The extraction of MTB genome characteristics was also performed using GC content with a 10,000 base pair sliding window. Clustering was verified using the energy obtained at each level of decomposition.

The work of Ferreira et al. (2018), who also arrived at the same result, employed the same methodology of Ferreira et al. (2017) in the first part of the analysis. To visualize the formation of groups, elastic net methodology was used, whose advantage is the ability to see the formation of groups at each level of decomposition. As the softer level corresponds to the approximation of the last level of decomposition, that is, the level that brings provides details, it was possible to see the formation of two groups.

In the most recent work of Ferreira et al. (2020), who also employed the same methodology of Ferreira et al. (2017) in the first part of the analysis. The verification of the formation of the groups was made using the technique of the exponent of Hurst obtained through five different methods.

The effectiveness of k-mer in the analysis, proven through the results obtained in the four articles mentioned above, reinforces the computational gain in genome analysis, both in the analysis processing and in the reliability of the results, due to covering reliable regions of the genomes. In the work of Jaillard et al. (2020) they also reinforce the advantages of the analyzes using the k-mer, because first they covering conserved genomic regions are redundant, and while they can be easily detected and filtered, they define groups of equivalent k-mers, which are not always straightforward to interpret as genomic

determinants. Second, k-mers may not be specific of a given genomic region and hence may be hard to annotate. Now for Bussi et al. (2021) the distributions of DNA words (i.e. oligomers of length k–also known as k-mers, n-tuples, n-grams) within long fragments of DNA has been shown to be highly characteristic of an organism. Thus, by using extracted short k-mers, sufficiently long DNA sequences could be taxonomically classified to different genomes efficiently, a common task in processing metagenomic data. Currently, tetranucleotide frequencies are used in the most popular tools for this process of binning metagenomic sequences, however, longer lengths of k have been shown to improve the resolution of taxonomic classification. In the works of Humphrey et al. (2020) and Kafri et al. (2021) reinforce the efficiency of analyzes working with k-mer, both with bacterial genomes and with other species.

Regarding the k-mers used in Figure 3, we found that with k = 8, 9 and 10, more similar sequences were found, where no sequence is isolated; that is, there is a similarity detection capability that is distinct from that with the other k-mers.

The following are the descriptions of the sequences that were grouped in Figure 3.

The Seq6_MDR and Seq7_XDR sequences that make up the first group are strains corresponding to two patients in KwaZulu-Natal, South Africa. Analyzing the similar sequences in the second group, it is interesting to highlight that the sequences Seq3_DS, Seq4_DR and Seq8_DS presented the same descriptions, since they are sequenced strains for comparative genomic studies.

In the third group, the Seq5_DR sequence is a strain that has an accelerated rate of transmission between humans under crowded conditions. The sequence Seq1_DS is a strain sensitive to all common medicines used to treat tuberculosis. The Seq2_DS sequence is a strain representing the majority of tuberculosis isolates from patients that recovered during an epidemic in the Western Cape of South Africa. The Seq9_DS and Seq10_DS sequences are strains derived from the virulent progenitor strain H37.

A cluster-randomized trials study is critical when working with antimicrobial resistance as a significant public health concern. In this view, a comparative approach was made by Yang et al. (2021) to understand the genetic variability and antibiotic resistance of P. aeruginosa isolated from patients with LTRIs admitted to the intensive care unit. P. aeruginosa infection leads to a deterioration of pulmonary function comparable to TB. Furthermore, it causes lower respiratory tract infections (LTRIs), the most common infection among hospitalized patients, associated with increased levels of morbidity and mortality. In addition, this pathogen, like TB, can develop antibiotic resistance through several mechanisms. In that investigation, it was used three types of DNA fingerprint markers were: Restriction Fragment Length Polymorphism (RFLP), Random Amplified Polymorphic DNA (RAPD), and Repetitive Extrapalindromic PCR (REP-PCR). The difference in similarity observed between those markers indicates high variability between strains; highlighted by the different number of clusters detected in the phylogenetic tree of each method considering 100% intra-strain similarity: RAPD - 8, RFLP - 13, and REP-PCR - 9. Along the same line, it emphasizes the importance of our innovative technique applied to MTB sequences to detect clusters with different degrees of drug resistance to understand their phylogenetic pattern.

5. Conclusions

The analysis using k-mers, taking into account the distance of the composition of each monomer (DNA word) in the MTB genome sequences, was able to capture the similarities of the strains well.

Of the 10 k-mers analyzed, 5 distinct group formations were identified, according to the threshold line. However, in all groups, the multidrug-resistant (MDR) and extensively drug-resistant (XDR) strains remained together and separated from the other strains.

References

ALLMAN, E.S., RHODES, J.A. and SULLIVANT, S., 2017. Statistically consistent k-mer methods for phylogenetic tree reconstruction. *Journal of Computational Biology*, vol. 24, no. 2, pp. 153-171. http://dx.doi.org/10.1089/cmb.2015.0216. PMid:27387364.

BERTHOLET, S., IRETON, G.C., ORDWAY, D.J., WINDISH, H.P., PINE, S.O., KAHN, M., PHAN, T., ORME, I.M., VEDVICK, T.S., BALDWIN, S.L., COLER, R.N., and REED, S.G., 2010. A defined tuberculosis vaccine candidate boosts BCG and protects against multidrug-resistant Mycobacterium tuberculosis. *Science Translational Medicine*, vol. 2, no. 53, pp. 53ra74. https://doi.org/10.1126/scitranslmed.3001094.

BUSSI, Y., KAPON, R. and REICH, Z., 2021. Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLoS One*, vol. 16, no. 10, pp. e0258693. http://dx.doi.org/10.1371/journal.pone.0258693. PMid:34648558.

CAMPANIÇO, A., MOREIRA, R. and LOPES, F., 2018. Drug discovery in tuberculosis. New drug targets and antimycobacterial agents. *European Journal of Medicinal Chemistry*, vol. 150, pp. 525-545. http://dx.doi.org/10.1016/j.ejmech.2018.03.020. PMid:29549838.

CAMPOS, D.M.O., OLIVEIRA, C.B.S., ANDRADE, J.M.A. and OLIVEIRA, J.I.N., 2020. Fighting COVID-19. *Brazilian Journal of Biology = Revista Brasileira de Biologia*, vol. 80, no. 3, pp. 698-701. http://dx.doi.org/10.1590/1519-6984.238155. PMid:32555974.

CHENG, J., CAO, F. and LIU, Z., 2013. AGP: a multimethods web server for alignment-free genome phylogeny. *Molecular Biology and Evolution*, vol. 30, no. 5, pp. 1032-1037. http://dx.doi.org/10.1093/molbev/mst021. PMid:23389766.

CHOR, B., HORN, D., GOLDMAN, N., LEVY, Y. and MASSINGHAM, T., 2009. Genomic DNA k-mer spectra: models and modalities. *Genome Biology*, vol. 10, no. 10, pp. R108. http://dx.doi.org/10.1186/gb-2009-10-10-r108. PMid:19814784.

COLE, S.T., BROSCH, R., PARKHILL, J., GARNIER, T., CHURCHER, C., HARRIS, D., GORDON, S.V., EIGLMEIER, K., GAS, S., BARRY 3RD, C.E., TEKAIA, F., BADCOCK, K., BASHAM, D., BROWN, D., CHILLINGWORTH, T., CONNOR, R., DAVIES, R., DEVLIN, K., FELTWELL, T., GENTLES, S., HAMLIN, N., HOLROYD, S., HORNSBY, T., JAGELS, K., KROGH, A., MCLEAN, J., MOULE, S., MURPHY, L., OLIVER, K., OSBORNE, J., QUAIL, M.A., RAJANDREAM, M.A., ROGERS, J., RUTTER, S., SEEGER, K., SKELTON, J., SQUARES, R., SQUARES, S., SULSTON, J.E., TAYLOR, K., WHITEHEAD, S. and BARRELL, B.G.. 1998. Deciphering the biology of Mycobacterium

tuberculosis from the complete genome sequence. *Nature*, vol. 393, no. 6685, pp. 537-544. http://dx.doi.org/10.1038/31159. PMid:9634230.

DALCOLMO, M.P., 2000. AIDS e tuberculose: novo problema, velho problema. *Jornal de Pneumologia*, vol. 26, no. 2, pp. 1-4. http://dx.doi.org/10.1590/S0102-35862000000200001.

DELOGU, G., SALI, M. and FADDA, G., 2013. The biology of mycobacterium tuberculosis infection. *Mediterranean Journal of Hematology and Infectious Diseases*, vol. 5, no. 1, pp. e2013070. http://dx.doi.org/10.4084/mjhid.2013.070. PMid:24363885.

FERREIRA, L.M., SÁFADI, T. and LIMA, R.R., 2017. Evaluation of genome similarities using the non-decimated wavelet transform. *Genetics and Molecular Research*, vol. 16, no. 3, pp. 1-12. http://dx.doi.org/10.4238/gmr16039758. PMid:28973739.

FERREIRA, L.M., SÁFADI, T. and FERREIRA, J.L., 2018. Wavelet-domain elastic net for clustering on genomes strains. *Genetics and Molecular Biology*, vol. 41, no. 4, pp. 884-892. http://dx.doi.org/10.1590/1678-4685-gmb-2018-0035. PMid:30508009.

FERREIRA, L.M., SÁFADI, T. and FERREIRA, J.L., 2020. Evaluation of genome similarities using a wavelet-domain approach. *Revista da Sociedade Brasileira de Medicina Tropical*, vol. 53, pp. e20190470. http://dx.doi.org/10.1590/0037-8682-0470-2019. PMid:32428175.

GAGNEUX, S., LONG, C.D., SMALL, P.M., VAN, T., SCHOOLNIK, G.K. and BOHANNAN, B.J.M., 2006. The competitive cost of antibiotic resistance in Mycobacterium tuberculosis. *Science*, vol. 312, no. 5782, pp. 1944-1946. http://dx.doi.org/10.1126/science.1124410. PMid:16809538.

HAN, G.B. and CHO, D.H., 2019. Genome classification improvements based on k-mer intervals in sequences. *Genomics*, vol. 111, no. 6, pp. 1574-1582. http://dx.doi.org/10.1016/j.ygeno.2018.11.001. PMid:30439480.

HOSSAIN, M.K., HASSANZADEGANROUDSARI, M. and APOSTOLOPOULOS, V., 2021. The emergence of new strains of SARS-CoV-2. What does it mean for COVID-19 vaccines? *Expert Review of Vaccines*, vol. 20, no. 6, pp. 635-638. http://dx.doi.org/10.1080/14760584.2021.1915140. PMid:33896316.

HUANG, H.H., 2016. An ensemble distance measure of k-mer and Natural Vector for the phylogenetic analysis of multiple-segmented viruses. *Journal of Theoretical Biology*, vol. 398, pp. 136-144. http://dx.doi.org/10.1016/j.jtbi.2016.03.004. PMid:26972479.

HUMPHREY, S., KERR, A., RATTRAY, M., DIVE, C. and MILLER, C.J., 2020. A model of k-mer surprisal to quantify local sequence information content surrounding splice regions. *PeerJ*, vol. 8, pp. e10063. http://dx.doi.org/10.7717/peerj.10063. PMid:33194378.

JAILLARD, M., PALMIERI, M., BELKUM, A.V. and MAHÉ, P., 2020. Interpreting k-mer–based signatures for antibiotic resistance prediction. *GigaScience*, vol. 9, no. 10, pp. giaa110. https://doi.org/10.1093/gigascience/giaa110.

KOCH, A., COX, H. and MIZRAHI, V., 2018. Drug-resistant tuberculosis: challenges and opportunities for diagnosis and treatment. *Current Opinion in Pharmacology*, vol. 42, pp. 7-15. http://dx.doi.org/10.1016/j.coph.2018.05.013. PMid:29885623.

KAFRI, A., CHOR, B. and HORN, D., 2021. Inter-chromosomal k-mer distances. *BMC Genomics*, vol. 22, pp. 644. http://dx.doi.org/10.1186/s12864-021-07952-0.

KWAN, C.K. and ERNST, J.D., 2011. HIV and tuberculosis: a deadly human syndemic. *Clinical Microbiology Reviews*, vol. 24, no. 2, pp. 351-376. http://dx.doi.org/10.1128/CMR.00042-10. PMid:21482729.

LANGE, C., CHESOV, D., HEYCKENDORF, J., LEUNG, C.C., UDWADIA, Z. and DHEDA, K., 2018. Drug-resistant tuberculosis: an update on disease burden, diagnosis and treatment. *Respirology (Carlton, Vic.)*, vol. 23, no. 7, pp. 656-673. http://dx.doi.org/10.1111/resp.13304. PMid:29641838.

LEIMEISTER, C.A., BODEN, M., HORWEGE, S., LINDNER, S. and MORGENSTERN, B., 2014. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics (Oxford, England)*, vol. 30, no. 14, pp. 1991-1999. http://dx.doi.org/10.1093/bioinformatics/btu177. PMid:24700317.

LI, R., ZHU, H., RUAN, J., QIAN, W., FANG, X., SHI, Z., LI, Y., LI, S., SHAN, G., KRISTIANSEN, K., LI, S., YANG, H., WANG, J. and WANG, J., 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, vol. 20, no. 2, pp. 265-272. http://dx.doi.org/10.1101/gr.097261.109. PMid:20019144.

MAITI, S. and MAITI, K.B., 2021. Gastrointestinal tuberculosis and HIV association in tropics. *Indian Journal of Surgery*, vol. 83, no. S4, pp. 1-6. http://dx.doi.org/10.1007/s12262-021-02844-9.

MAZUREK, J., IGNATOWICZ, L., KÄLLENIUS, G., JANSSON, M. and PAWLOWSKI, A., 2012. Mycobacteria-infected bystander macrophages trigger maturation of dendritic cells and enhance their ability to mediate HIV transinfection. *European journal of immunology*, vol. 42, no. 5, pp. 1192-1202. https://doi.org/10.1002/eji.201142049.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION – NCBI, 2018 [viewed 10 February 2020]. *Mycobacterium tuberculosis* [online]. Rockville: National Library of Medicine. Available from: https://www.ncbi.nlm.nih.gov/genome/166

NIEMANN, S., KÖSER, C.U., GAGNEUX, S., PLINKE, C., HOMOLKA, S., BIGNELL, H., CARTER, R.J., CHEETHAM, R.K., COX, A., GORMLEY, N.A., KOKKO-GONZALES, P., MURRAY, L.J., RIGATTI, R., SMITH, V.P., ARENDS, F.P., COX, H.S., SMITH, G. and ARCHER, J.A., 2009. Genomic diversity among drug sensitive and multidrug resistant isolates of Mycobacterium tuberculosis with identical DNA fingerprints. *PLoS One*, vol. 4, no. 10, pp. e7407. http://dx.doi.org/10.1371/journal.pone.0007407. PMid:19823582.

ONDOV, B.D., TREANGEN, T.J., MELSTED, P., MALLONEE, A.B., BERGMAN, N.H., KOREN, S. and PHILLIPPY, A.M., 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, vol. 17, no. 1, pp. 132. http://dx.doi.org/10.1186/s13059-016-0997-x. PMid:27323842.

PAWLOWSKI, A., JANSSON, M., SKÖLD, M., ROTTENBERG, M.E. and KÄLLENIUS, G., 2012. Tuberculosis and HIV Co-Infection. *PLoS pathogens*, vol. 8, no. 2, pp. e1002464. https://doi.org/10.1371/journal.ppat.1002464.

PRATAS, D. and SILVA, J.M., 2020. Persistent minimal sequences of SARS-CoV-2. *Bioinformatics (Oxford, England)*, vol. 36, no. 21, pp. 5129-5132. http://dx.doi.org/10.1093/bioinformatics/btaa686. PMid:32730589.

R CORE TEAM, 2019 [viewed 11 February 2020]. *A Language and environment for statistical computing* [online]. Vienna, Austria. Available from: https://www.R-project.org/

SAINI, S. and DEWAN, L., 2016. Application of discrete wavelet transform for analysis of genomic sequences of Mycobacterium tuberculosis. *SpringerPlus*, vol. 5, pp. 64. http://dx.doi.org/10.1186/s40064-016-1668-9. PMid:26839757.

SHARIFIPOUR, E., SHAMS, S., ESMKHANI, M., KHODADADI, J., FOTOUHI-ARDAKANI, R., KOOHPAEI, A., DOOSTI, Z. and EJ GOLZARI, S., 2020. Evaluation of bacterial co-infections of the respiratory tract in COVID-19 patients admitted to ICU. *BMC Infectious Diseases*, vol. 20, pp. 1646. http://dx.doi.org/10.1186/s12879-020-05374-z. PMid:32873235.

SHARMA, A., DE ROSA, M., SINGLA, N., SINGH, G., BARNWAL, R.P. and PANDEY, A., 2021. Tuberculosis: an overview of the immunogenic

response, disease progression, and medicinal chemistry efforts in the last decade toward the development of potential drugs for extensively drug-resistant tuberculosis strains. *Journal of Medicinal Chemistry*, vol. 64, no. 8, pp. 4359-4395. http://dx.doi.org/10.1021/acs.jmedchem.0c01833. PMid:33826327.

SINGH, R., SEKHON, A., KOWSARI, K., LANCHANTIN, J., WANG, B. and QI, Y., 2017. Gakco: a fast gapped *k-mer* string kernel using counting. In: M. CECI, J. HOLLMÉN, L. TODOROVSKI, C. VENS and S. DŽEROSKI, eds. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2017*. Cham: Springer, Lecture Notes in Computer Science, vol. 10534, pp. 356-373. http://dx.doi.org/10.1007/978-3-319-71249-9_22

SUBHASREE, C.R., PRIYA, R.S.K., DIWAKAR, M., SUBRAMANIAM, S. and SHYAMA, S., 2017. Review on comparative genomics for mycobacterium tuberculosis strains. *International Journal of Pharmaceutical Sciences and Research*, vol. 8, no. 12, pp. 5022-5042. http://dx.doi.org/10.13040/IJPSR.0975-8232.8(12).5022-42.

TACCONELLI, E., CARRARA, E., SAVOLDI, A., HARBARTH, S., MENDELSON, M., MONNET, D.L., PULCINI, C., KAHLMETER, G., KLUYTMANS, J., CARMELI, Y., OUELLETTE, M., OUTTERSON, K., PATEL, J., CAVALERI, M., COX, E.M., HOUCHENS, C.R., GRAYSON, M.L., HANSEN, P., SINGH, N., THEURETZBACHER, U. and MAGRINI, N., and WHO PATHOGENS PRIORITY LIST WORKING GROUP, 2018. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *The Lancet. Infectious Diseases*, vol. 18, no. 3, pp. 318-327. http://dx.doi.org/10.1016/S1473-3099(17)30753-3. PMid:29276051.

VISCA, D., ONG, C.W.M., TIBERI, S., CENTIS, R., D'AMBROSIO, L., CHEN, B., MUELLER, J., MUELLER, P., DUARTE, R., DALCOLMO, M., SOTGIU, G., MIGLIORI, G.B. and GOLETTI, D., 2021. Tuberculosis and COVID-19 interaction: a review of biological, clinical and public health effects. *Pulmonology*, vol. 27, no. 2, pp. 151-165. http://dx.doi.org/10.1016/j.pulmoe.2020.12.012. PMid:33547029.

WALKER, T.M., MERKER, M., KNOBLAUCH, A.M., HELBLING, P., SCHOCH, O.D., VAN DER WERF, M.J., KRANZER, K., FIEBIG, L., KRÖGER, S., HAAS, W., HOFFMANN, H., INDRA, A., EGLI, A., CIRILLO, D.M., ROBERT, J., ROGERS, T.R., GROENHEIT, R., MENGSHOEL, A.T., MATHYS, V., HAANPERÄ, M., SOOLINGEN, D.V., NIEMANN, S., BÖTTGER, E.C. and KELLER, P.M., and MDR-TB CLUSTER CONSORTIUM, 2018. A cluster of multidrug-resistant Mycobacterium tuberculosis among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. *The Lancet. Infectious Diseases*, vol. 18, no. 4, pp. 431-440. http://dx.doi.org/10.1016/S1473-3099(18)30004-5. PMid:29326013.

WANG, J.D., 2013. Comparing virus classification using genomic materials according to different taxonomic levels. *Journal of Bioinformatics and Computational Biology*, vol. 11, no. 6, pp. 1343003. http://dx.doi.org/10.1142/S0219720013430038. PMid:24372032.

WILKINSON, S.P., 2019 [viewed 12 February 2020]. *Kmer: an R package for fast alignment-free clustering of biological sequences. R package version 1.0.0* [online]. Available from: https://cran.r-project.org/package=kmer

WORLD HEALTH ORGANIZATION – WHO, 2021 [viewed 14 May 2022]. *Tuberculosis* [online]. Available from https://www.who.int/news-room/fact-sheets/detail/tuberculosis

YANG, X., LAI, Y., LI, C., YANG, J., JIA, M. and SHENG, J., 2021. Molecular epidemiology of Pseudomonas aeruginosa isolated from lower respiratory tract of ICU patients. *Brazilian Journal of Biology = Revista Brasileira de Biologia*, vol. 81, no. 2, pp. 351-360. http://dx.doi.org/10.1590/1519-6984.226309. PMid:32491054.

YIN, C. and YAU, S.S.T., 2015. An improved model for whole genome phylogenetic analysis by Fourier transform. *Journal of Theoretical Biology*, vol. 382, pp. 99-110. http://dx.doi.org/10.1016/j.jtbi.2015.06.033. PMid:26151589.

ZHANG, Y. and YEW, W.W., 2015. Mechanisms of drug resistance in Mycobacterium tuberculosis: update 2015. *The International Journal of Tuberculosis and Lung Disease*, vol. 19, no. 11, pp. 1276-1289. http://dx.doi.org/10.5588/ijtld.15.0389. PMid:26467578.