GMR

Evaluation of genome similarities using the non-decimated wavelet transform

L.M. Ferreira, T. Sáfadi and R.R. Lima

Departamento de Estatística, Universidade Federal de Lavras, Lavras, MG, Brasil

Corresponding author: T. Sáfadi
E-mail: safadi@des.ufla.br

Genet. Mol. Res. 16 (3): gmr16039758
Received June 23, 2017
Accepted August 24, 2017
Published September 21, 2017
DOI http://dx.doi.org/10.4238/gmr16039758

ABSTRACT. The wavelets have become increasingly popular in the field of bioinformatics due to their capacity in multiresolution analysis and space-frequency localization; the latter particularity is acquired due to a moving window that runs through the analyzed space. As a feature, they have a better ability to capture hidden components of biological data and an efficient link between biological systems and the mathematical objects used to describe them. The decomposition of signals/sequences at different levels of resolution allows obtaining distinct characteristics in each level. The energy (variance) obtained at each level provides a new set of information that can be used to search similarities between sequences. We show that the behavior of GC-content sequence can be succinctly described regarding the non-decimated wavelet transform, and we indicate how this characterization can be used to improve clustering of the similar strains of the genome of the *Mycobacterium tuberculosis*, having a very efficient level of detail. The clustering analysis using the energy obtained at each level of the analyzed sequences was essential to verify the dissimilarity of the sequences.

Key words: Non-decimated wavelet transform; Cluster analysis; Genome

INTRODUCTION

In the last decades, the analysis using technique of wavelets has been growing increasingly. One of the great advantages associated with this method corresponds to the computational gain, that is, the analyses are processed almost in real time. The applicability is in several areas of science, like Physics, Mathematics, Engineering, Genetics, among others.

The wavelet transform is a technique of seeing and represents a signal. Mathematically, it is represented by a function oscillating in time or space. As a characteristic, it has sliding windows that expand or compress to capture low- and high-frequency signals, respectively (Percival and Walden, 2000). Its origin occurred in the field of seismic study to describe the disturbances arising from a seismic impulse (Morlet et al., 1982).

Among the wavelet techniques, we have the discrete non-decimated wavelet transform (NDWT), whose main characteristic is that it can work with any size of signals/sequences. In this technique, the coefficients are translation invariants, that is, the choice of origin is irrelevant since all the observations are used in the analysis, a situation that does not occur in the discrete decimated wavelet transform (DWT).

Discrete wavelet transforms were used to identify gene locations in genomic sequences (Ning et al., 2003), identifying long-range correlations, locating periodicities in DNA sequences (Vannucci and Liò, 2001), and in the analysis of G+C patterns (Dodin et al., 2000).

The clustering analysis is often adopted to deal with DNA sequences efficiently. A wavelet-based feature vector model was proposed by Bao and Yuan (2015) for clustering of DNA sequences.

Human tuberculosis (TB) is caused by an intracellular pathogen, *Mycobacterium tuberculosis* (MTB) and it replicates rapidly in the lungs with high oxygen concentration. Global TB control measures are affected by the emergence of drug resistant (DR), multidrug resistant (MDR), and extensively drug resistant (XDR) strains. Resistance in these MTB strains to anti-TB drugs occurs due to chromosomal mutations (Saini and Dewan, 2016). Global control of tuberculosis is hampered by slow, insensitive diagnostic methods, particularly for the detection of DR forms and in patients with human immunodeficiency virus infection. Early detection is essential to reduce the death rate and interrupt transmission. Boehme et al. (2010) concerned with this situation and developed a more efficient method for the detection of DR and MDR strains. Perdigão et al. (2010) worked to characterize the genetic changes associated with the high number of XDR that threatens the global control of TB worldwide.

Apart from molecular methods based on whole genome sequences of MTB, signal processing of complete genomic sequences can help display and explore structural patterns capable of being interpreted and compared. Graphical representations obtained from signal processing methods can provide insight into the evolution, structure, and function of genomes (Anastassiou, 2000).

The genome of MTB is approximately 4.4 million base pairs long and is one of the largest known bacterial genomes. This bacterium is the cause of the TB disease that has killed thousands of people around the world. The GC-content is an important parameter of bacterial genomes used to scan the basic composition of the genome as well as to understand the evolution of the coded sequence.

Saini and Dewan (2016) highlights the potential of discrete wavelet transforms in the analysis and comparison of genomic sequences of MTB with different resistance characteristics. Based on the calculation of the energy of wavelet decomposition coefficients

of complete genomic sequences, they showed that the genomic sequences could be grouped into two broad categories wherein the DR and drug susceptible (DS) sequences formed one group while the MDR and XDR sequences formed the other group.

The main advantage of the NDWT method concerning DWT is the possibility to work with any size of genome sequences, that is, there is no loss of genome information. In the DWT method, the sequence must have a power of two, where loss of information of the genome inevitably occurs.

The NDWT method can be used in any genome type, increasing the speed of the analysis, because the analyses with this method are processed almost in real time.

In this study, the NDWT was applied to the GC-content sequences of the MTB genome strains and the energies obtained to the detail level coefficients were used to study their similarities. Stacked plots of the detail levels provide an effective means of exploring the relationships between genomic strains at different scales. The proposed methodology is applied to MTB sequences, being 4 DR, 4 DS, 1 MDR, and 1 XDR.

MATERIAL AND METHODS

The sequences analyzed correspond to the strains of the MTB genome. Ten sequences were analyzed, obtained from the National Center for Biotechnology Information (NCBI, 2017).

The GC-content of all the sequences was evaluated using a sliding window of 10,000 bases. The sequences were decomposed using a discrete NDWT. We considered the Daubechies wavelet (4 null moments) with 5 levels of decomposition. Statistical measures of energy for each of the decomposed sequences were evaluated. We used the measures of energy to know if the sequences were similar or not, and the clustering analysis was performed using the Mahalanobis distance in a hierarchical method with the average linkage.

The free software R (R Core Team, 2017) was used.

Wavelet

A wavelet function is the interpretation of a short-wave with rapid growth and decay. The theory is based on the representation of functions in different scales and resolutions (time-scale) that is considered one of its main characteristics (Daubechies, 1992).

In the analysis of wavelet, the window is oscillating and is called the mother wavelet. There are arbitrary translations and dilations. In this way, the mother wavelet generates other wavelets (Hernandez and Weiss, 1996).

By definition: a wavelet is a function $\psi(t) \in L^2(\mathbb{R})$, such that their family of functions is given by Equation 1:

$$\psi_{j,k} = 2^{-j/2}\psi\left(2^{-j}t - k\right) \quad \text{(Equation 1)}$$

where j and k are arbitrary integers on an orthonormal basis in Hilbert space $L^2(\mathbb{R})$ (Wojtaszczyk, 1997).

Pyramidal algorithm

The pyramidal algorithm, developed by Mallat (1989), uses the low-pass filter l_k

obtained using the scale function or father wavelet (ϕ) and the high-pass filter h_k obtained using the mother wavelet (ψ) with coefficients given by Equation 2:

$$l_k = \sqrt{2}\int_{-\infty}^{+\infty}\phi(t)\phi(2t-k)\,dt$$

$$h_k = \sqrt{2}\int_{-\infty}^{+\infty}\psi(t)\phi(2t-k)\,dt \qquad \text{(Equation 2)}$$

The approximation coefficients (coarse scales) are obtained by a low-pass filter, and the coefficients of details (finer scales) are obtained by a high-pass filter.

Discrete NDWTs

The characteristic of the discrete NDWT is to keep the same amount of data in even and odd decimations on each scale and continue to do the same on each subsequent scale, being D_0 the dyadic decimation, D_1 the odd decimation, H the high-pass filter, and L the low-pass filter. Consider, for example, an input vector $(y_1,..., y_n)$. Then, apply and keep both D_0H_y and D_1H_y, even and odd indexed of the observation-filtered wavelets. Each of these sequences is length $n/2$. Thus, in total, the number of wavelet coefficients in both decimals on the finer scale is 2 x $n/2 = n$ (Nason, 2008).

Figure 1 shows that in the finer scale the coefficients of wavelets are d_0 and d_1. The next finer scales are d_{00}, d_{01}, d_{10}, d_{11}.

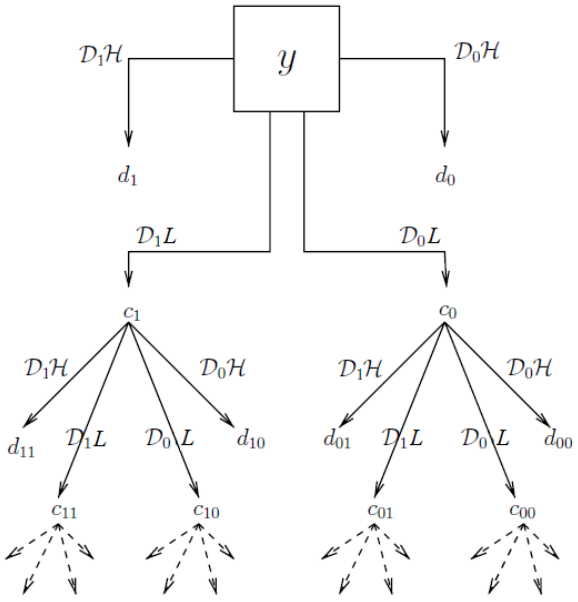

Figure 1. Diagram of the discrete non-decimated wavelet transform (NDWT). Source: Nason (2008).

In the discrete NDWT, the coefficients are translational invariants, that is, it means that the circular displacement of the data was reflected in the same direction of the coefficients. Another feature is the ability to handle data of an arbitrary size that does not require the sample size n to have a power of two, which is what occurs in the discrete DWT. The main advantage of this method is associated with zero pass filters, which means that it operates circularly to the data allowing functionalities at different scales to be aligned with the sequence of the original data (Vannucci and Liò, 2001).

Scalogram

The scalogram is a very effective tool for interpreting the sign of the wavelet. It is defined as a graph of the sum of squares of the wavelet coefficients at the different levels.

The energy *E* (j) for the coefficients, d_{jk}, of the wavelet on each level *j*, is given by:

$$E(j) = \sum_{k=0}^{n} d_{j,k}^2 \qquad j = 1, \ldots, J. \qquad \text{(Equation 3)}$$

Equation 3 corresponds to the calculation of the energy in the discrete NDWT on level *j* (Gençay et al., 2002).

Daubechies wavelets

The Daubechies wavelet is a family of orthogonal wavelets that define a discrete wavelet transformation and are characterized by a maximum number of null moments (degree of smoothing) for some given support. With each wavelet type of this class, there is a scaling function (called father wavelet), which generates an orthogonal multiresolution analysis.

According to Daubechies (1992), for each integer *r*, the orthonormal basis for $L^2(\mathbb{R})$ is defined as in Equation 4:

$$\phi_{r,j,k} = 2^{-j/2} \phi_r \left(2^{-j} x - k\right), \qquad j, k \in \mathbb{Z} \qquad \text{(Equation 4)}$$

where the function $\phi_r(x)$ in $L^2(\mathbb{R})$ has the property that $\phi_r(x-k) \mid k \in \mathbb{Z}$ is an orthonormal sequential basis in $L^2(\mathbb{R})$. Here *j* is the scale index, *k* is the translation index, and *r* is the filtering index.

Null moments

The waveforms have some null moments, that is, a function $\psi \in L^2(\mathbb{R})$ has *m* null moments if it satisfies Equation 5:

$$\int x^l \psi(x)\, dx = 0 \qquad \text{(Equation 5)}$$

where $l = 0, \ldots, m - 1$

If the wavelets have *m* null moments, then all the coefficients of the wavelet of any polynomial of degree *m* or less can be exactly zero (Nason, 2008).

Null moments and smoothness are mathematically related. The greater the number of null moments of a wavelet, the smoother it is. Moreover, the greater the smoothness of the wavelet, the greater is the probability of perfect reconstruction of the signal decomposed by the wavelet transform.

Clustering analysis

When two samples are close, they should also have similar values for the measured variables; therefore, the greater proximity between the measures related to the samples, the greater the similarity between them. The dendrogram (which has a tree structure) hierarchized this similarity so that one can have a two-dimensional view of the similarity of the whole set of samples used in the study, that is, the dendrogram organizes certain factors and variables (Abonyi and Feil, 2007).

The average linkage clustering was developed as an antidote to the extremes of both single and complete linkage. Although there are some variants of the method, each essentially computes an average of the similarity of a case under consideration with all cases in the existing cluster and, subsequently, joins the case to that cluster if a given level of similarity is achieved using this average value. The most commonly used variant of average linkage computes the arithmetic average of similarities among the cases. Other variants of average linkage are designed to calculate the similarity between the centroids of two clusters that might be merged. The average linkage has been used extensively in the biological sciences but has only recently begun to see much use in the social sciences (Aldenderfer and Blashfield, 1984).

The Mahalanobis distance increases with increasing distances between the two groups and with decreasing within-group variation. By also employing within-group correlations, the Mahalanobis distance takes account of the (possibly nonspherical) shape of the groups (Everitt et al., 2011).

The use of Mahalanobis D^2 according to Equation 6 implies that the investigator is willing to assume that the covariance matrices are at least approximately the same in the two groups. When this is not so, D^2 is an inappropriate inter-group measure, and for such cases, several alternatives have been proposed.

$$D^2 = \left(\overline{x}_A - \overline{x}_B\right)' W^{-1} \left(\overline{x}_A - \overline{x}_B\right) \quad \text{(Equation 6)}$$

where $\overline{x}_A{}'$ means vector of the group A, $\overline{x}_B{}'$ means vector of the group B, and W is the pooled within-group covariance matrix for the two groups.

GC-content

The GC-content is an important parameter of bacterial genomes used to scan the basic composition of the genome as well as to understand the evolution of the coded sequence. Historically, this rate is represented in a range of 25 to 75% in bacterial genomes (Mann and Chen, 2010). In the mammalian genome, approximately 50% of all genes are controlled by promoters with high GC-contents. Chang et al. (2015) examined a method for stable quantification of such GC-rich DNA sequences.

For each genome sequence, the GC-content is calculated as the ratio of the sum of bases G, C, under the sum of the bases A, G, C, and T, according to Equation 7:

$$GCcontent = \frac{nG + nC}{nA + nG + nC + nT} \quad \text{(Equation 7)}$$

where nA, nG, nC, and nT represent the number of nucleotide bases A, G, C, and T, respectively, in a sequence. The GC-content can also be calculated for a part of the sequence using the window technique, wherein the GC-content is calculated for a fixed length of a specific window of the sequence. The determination of GC-content ratio helps in identifying gene-rich regions of the genome (Saini and Dewan, 2016). These gene-rich regions bring significant biological information about the genome. Cheng et al. (2016) and Wei et al. (2016) worked with high GC-content, aiming the development of new molecular markers, highlighting the importance of working with gene-rich regions.

RESULTS AND DISCUSSION

Table 1 shows the information for each sequence obtained at the site of the NCBI. In the second column, the identification of each accession number is shown. The third column shows the characteristic of each strain according to Ilina et al. (2013), i.e., DS, DR, MDR, and XDR. In the fourth column, we have the total rate of GC-content, whose values are very close. The last column shows the total size of each sequence; note that all sequences have the length around 4 million.

Table 1. Description of 10 sequences of the *Mycobacterium tuberculosis* genome.

Sequence number	NCBI accession number	Resistance type	Total rate of GC-content	Total sequence length
Seq1	CP002992.1	DS	0.6560	4398525
Seq2	CP000717.1	DS	0.6562	4424435
Seq3	CP001641.1	DS	0.6561	4398812
Seq4	CP001642.1	DR	0.6559	4405981
Seq5	CP001664.1	DR	0.6563	4408224
Seq6	CP001658.1	MDR	0.6561	4398250
Seq7	CP001976.1	XDR	0.6561	4399120
Seq8	CP002884.1	DS	0.6561	4414325
Seq9	AL123456.3	DS	0.6561	4411532
Seq10	CP000611.1	DS	0.6561	4419977

DS = drug susceptible; DR = drug resistant; MDR = multidrug resistant; XDR = extensively drug resistant.

Initially, we show the dissimilarity between the rate of GC-content sequences considering the mean of all sequences and how much each sequence differs from the general average. The result is shown in Figure 2. The sequences that presented averages smaller than the general mean were: 1 (DS), 4 (DR), 9 (DS), and 10 (DS). Moreover, those with averages higher than the general average were: 2 (DS), 3 (DS), 5 (DR), 6 (MDR), 7 (XDR).

Figure 3 shows the decomposition on 5 levels of details of the sequence 1, using the NDWT. The wavelet used is Daubechies with 4 null moments.

The signal was decomposed up to the fifth level, because from the sixth level onwards the energy is quite low, as can be seen in Figure 4, that is, there is a decay of the energy along of the levels. It is also noted that energy concentrates more on the first two levels.

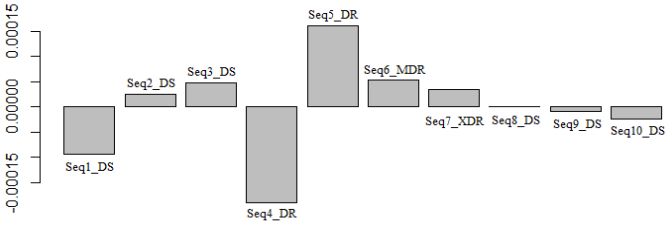

Figure 2. Dissimilarity of the sequence signals. DS = drug susceptible; DR = drug resistant; MDR = multidrug resistant; XDR = extensively drug resistant.

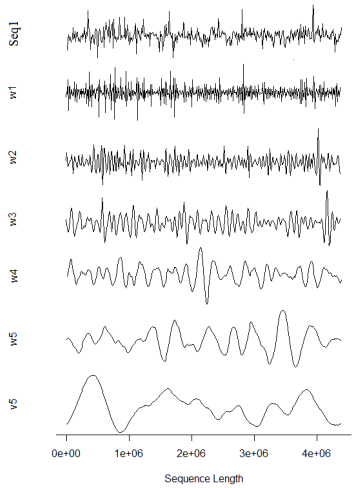

Figure 3. Decomposition of the sequence 1 in 5 levels using the discrete non-decimated wavelet transforms (NDWT). The first signal corresponds to the original signal, and in the sequence we have, w1 is the 1st level of decomposition, w2 the second level of decomposition, and so on. The last signal v5 represents the approximation coefficients of the smoother level.

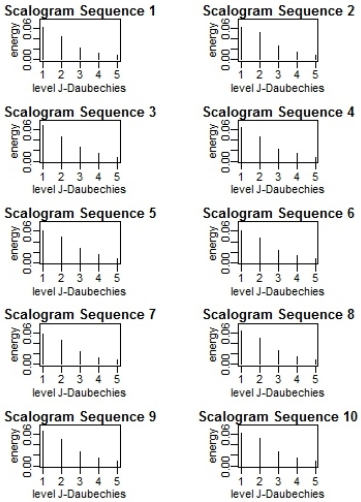

Figure 4. Scalogram for each GC-content sequence.

The approximation coefficients (v5) for all sequences are plotted (Figure 5), and it is possible to note that the regions of higher and lower peaks are coincident. For the windows of the 2 million nucleotides, the sequences 6 (MDR) and 7 (XDR) presented peaks in distinct regions from the other sequences. The highest peaks are regions rich in genes.

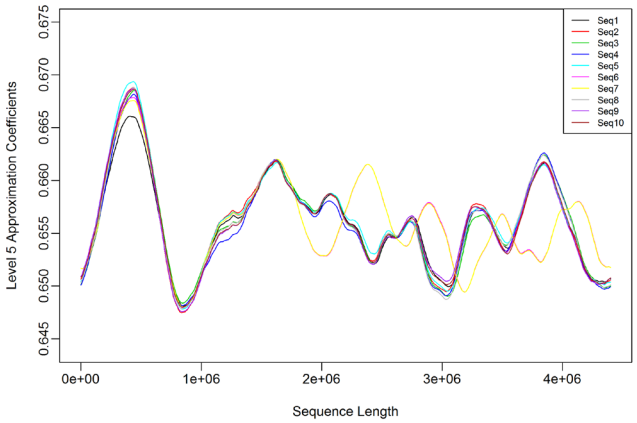

Figure 5. Approximation coefficients (v5) for all sequences.

The energies obtained in each level were considered to the clustering analysis. We use the Mahalanobis distance in a hierarchical method using the average linkage. The formation of 3 groups was verified (Figure 6).

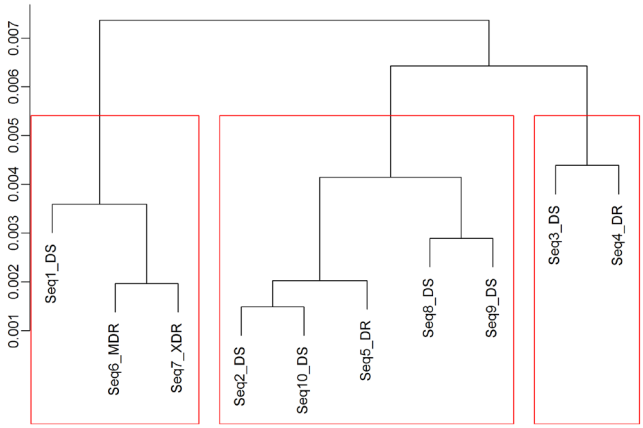

Figure 6. Clustering of sequences based on the energies of each level. DS - drug susceptible, DR - drug resistant, MDR - multidrug resistant, XDR - extensively drug resistant.

The first group (shown on the left side of Figure 6) has the sequences: 1 (DS), 6 (MDR), and 7 (XDR). The second group has the sequences: 2 (DS), 10 (DS), 5 (DR), 8 (DS), and 9 (DS), and the third group has the sequences: 3 (DS) and 4 (DR).

Figures 7 to 9 show each group with their features. The first group (Figure 7) show that the sequences 6 (MDR) and 7 (XDR) almost overlap completely. These strains correspond to a single patient in KwaZulu-Natal, South Africa. However, the sequence 1 (DS) presents different peaks. As a characteristic, this 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.

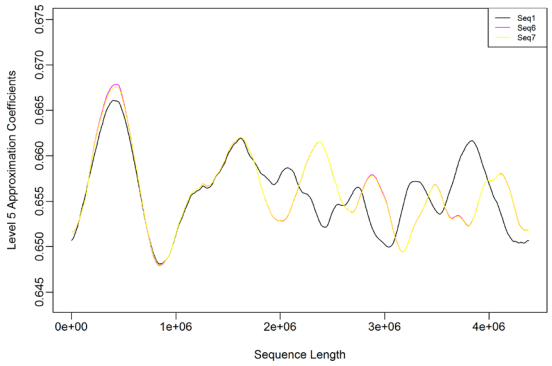

Figure 7. Clustering of sequences 1 (DS), 6 (MDR), and 7 (XDR).

Figure 8 shows the second group, with the sequences 2 (DS), 5 (DR), 8 (DS), 9 (DS), and 10 (DS) that present very similar behavior. The sequence 2 (DS) is a susceptible strain representing the largest portion of patients' tuberculosis isolates recovered during an epidemic in the Western Cape of South Africa. The sequence 5 (DR) corresponds to a drug-resistant strain, having an accelerated rate of transmission between humans under agglomeration conditions. The sequence 8 (DS) is a susceptible strain used for comparative genomic studies. The sequence 9 is a 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. Moreover, the sequence 10 (DS) is an avirulent 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.

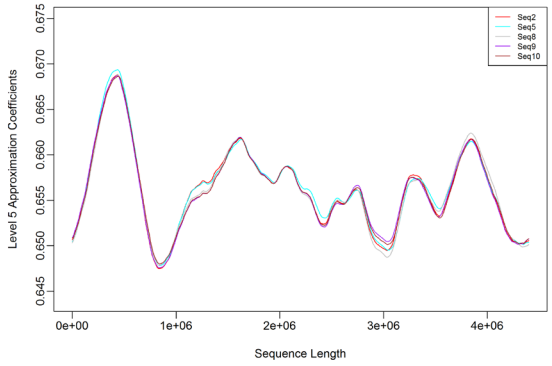

Figure 8. Clustering of sequences 2 (DS), 5 (DR), 8 (DS), 9 (DS), and 10 (DS).

Figure 9 shows the third group with the sequences 3 (DS) and 4 (DR). The sequence 3 (DS) is a susceptible strain belonging to the Beijing family, sequenced for comparative genomic studies. The sequence 4 (DR) is a resistant strain isolated in 2004, referring to a patient with secondary pulmonary tuberculosis, sequenced for comparative genomic studies.

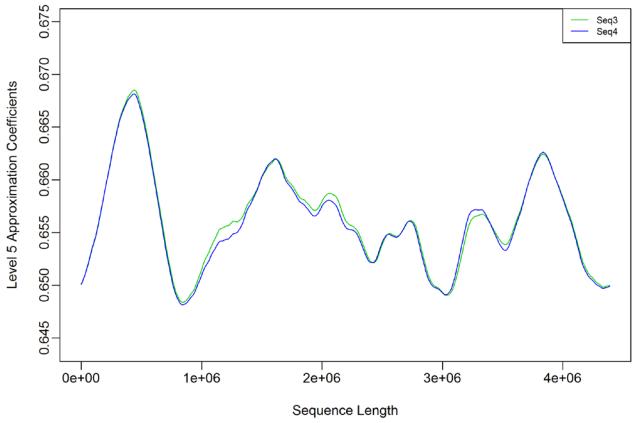

Figure 9. Clustering of sequences 3 (DS) and 4 (DR).

It is interesting to note that although the graphs 8 and 9 are very similar, the proposed methodology detected differences between the analyzed sequences.

Saini and Dewan (2016) based on the calculation of the energy of wavelet decomposition coefficients of complete genomic sequences showed that the genomic sequences of MTB could be grouped only into two groups. The first group with DS and DR sequences (lower energy) and the second group with MDR and XDR sequences (highest energy).

Our method, considering the energy at each level of detail, was able to identify more than two groups. It was possible to detect particularities of sequences 1 (DS), 3 (DS), and 4 (DR) with the proposed methodology.

CONCLUSIONS

The use of the discrete NDWT in the analysis of MTB genome strains allowed us to consider the entire genome sequence without the need to have a power of 2.

Similarities between genome sequences are best detected if one considers the energy at each level of detail of the wavelet decomposition.

ACKNOWLEDGMENTS

L.M. Ferreira thanks CAPES (Aperfeiçoamento de Pessoal de Nível Superior) for the financial support.

REFERENCES

Abonyi J and Feil B (2007). Cluster analysis for data mining and system identification. Birkhäuser, Basel, Boston, Berlin.

Anastassiou D (2000). Frequency-domain analysis of biomolecular sequences. *Bioinformatics* 16: 1073-1081. https://doi.org/10.1093/bioinformatics/16.12.1073

Aldenderfer MS and Blashfield RK (1984). Cluster analysis sage university papers series. Quantitative applications in the social sciences. Sage Publications, Inc., Beverly Hills.

Bao JP and Yuan RY (2015). A wavelet-based feature vector model for DNA clustering. *Genet. Mol. Res.* 14: 19163-19172. https://doi.org/10.4238/2015.December.29.26

Boehme CC, Nabeta P, Hillemann D, Nicol MP, et al. (2010). Rapid molecular detection of tuberculosis and rifampin resistance. *N. Engl. J. Med.* 363: 1005-1015. https://doi.org/10.1056/NEJMoa0907847

Chang GJ, Seyfert HM and Shen XZ (2015). Adaption of SYBR Green-based reagent kit for real-time PCR quantitation of GC-rich DNA. *Genet. Mol. Res.* 14: 8509-8515. https://doi.org/10.4238/2015.July.28.20

Cheng JL, Li J, Qiu YM, Wei CL, et al. (2016). Development of novel SCAR markers for genetic characterization of *Lonicera japonica* from high GC-RAMP-PCR and DNA cloning. *Genet. Mol. Res.* 15: gmr7737. https://doi.org/10.4238/gmr.15027737

Everitt BS, Landau S, Leese M and Stahs D (2011). Cluster analysis. John Wiley & Sons, King's College London, UK.

Daubechies I (1992). Ten Lectures on Wavelets. Society for industrial and applied mathematics, Philadelphia.

Dodin G, Vandergheynst P, Levoir P, Cordier C, et al. (2000). Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. *J. Theor. Biol.* 206: 323-326. https://doi.org/10.1006/jtbi.2000.2127

Gençay R, Selçuk F and Whicher B (2002). An introduction to wavelets and other filtering methods in finance and economics. Academic Press, San Diego, California.

Hernandez E and Weiss G (1996). A first course on wavelets. CRC Press LLC, USA.

Ilina EN, Shitikov EA, Ikryannikova LN, Alekseev DG, et al. (2013). Comparative genomic analysis of *Mycobacterium tuberculosis* drug resistant strains from Russia. *PLoS One* 8: e56577. https://doi.org/10.1371/journal.pone.0056577

Mallat SG (1989). Multiresolution approximations and wavelet orthonormal bases of L2(R). *Trans. Am. Math. Soc.* 315: 69-87.

Mann S and Chen YPP (2010). Bacterial genomic G+C composition-eliciting environmental adaptation. *Genomics* 95: 7-15. https://doi.org/10.1016/j.ygeno.2009.09.002

Morlet J, Arens G, Fourgeau E and Giard D (1982). Wave propagation and sampling theory- Part I: Complex signal and scattering in multilayered media. *Geophysics* 47: 203-221. https://doi.org/10.1190/1.1441328

Nason GP (2008). Wavelet methods in statistics with R. Springer, New York.

National Center for Biotechnology Information (2017). *Mycobacterium tuberculosis*. Genoma. Available at [https://www.ncbi.nlm.nih.gov/assembly/GCF_000224435.1/]. Accessed April 2, 2017.

Ning J, Moore CN and Nelson JC (2003). Preliminary wavelet analysis of genomic sequences. In: Proceedings of the IEEE computer society conference on bioinformatics CSB '03, Stanford, California, 509-510.

Percival DB and Walden AT (2000). Wavelet methods for time series analysis. Cambridge University Press.

Perdigão J, Macedo R, Malaquias A, Ferreira A, et al. (2010). Genetic analysis of extensively drug-resistant *Mycobacterium tuberculosis* strains in Lisbon, Portugal. *J. Antimicrob. Chemother.* 65: 224-227. https://doi.org/10.1093/jac/dkp452

R Core Team (2017). A Language and environment for statistical computing. Vienna, Austria. Available at [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* 5: 64. https://doi.org/10.1186/s40064-016-1668-9

Vannucci M and Liò P (2001). Non-decimated wavelet analysis of biological sequences: applications to protein structure and genomics. *The Indian J. Statistics* 63: 218-233.

Wei CL, Cheng JL, Khan MA, Yang LQ, et al. (2016). An improved DNA marker technique for genetic characterization using RAMP-PCR with high-GC primers. *Genet. Mol. Res.* 15: gmr8721. https://doi.org/10.4238/gmr.15038721

Wojtaszczyk P (1997). A mathematical introduction to wavelets. Cambridge University Press, New York.