EVALUATION OF GENOME SIMILARITIES USING INDEPENDENT COMPONENTS

Thelma SÁFADI[1]
Leila Maria FERREIRA[1]

■ ABSTRACT: We propose the use of independent component analysis to find similarities of genomes. Considering different numbers of independent components, the complete linkage method was used to identify groups based on the estimated coefficients of the mixing matrix. The sequences analyzed correspond to the strains of the *Mycobacterium tuberculosis* genome, ten sequences were analyzed, obtained from the National Center for Biotechnology Information (NCBI, 2017). The GC-content of each sequence was evaluated using a sliding window of 10,000 bases. The clustering analysis using the independent components of the analyzed sequences was essential to verify the dissimilarity of the sequences.

■ KEYWORDS: GC-content; *Mycobacterium tuberculosis* genomes; cluster analysis.

1 Introduction

Independent component analysis (ICA) is a dimension reduction technique that uses the existence of independent factors in multivariate data and decomposes an input data set into statistically independent components. ICA can reduce the effects of noise and is ideal for separating mixed signals. The use of independent component analysis not only enables the clustering of time series as also provides us with information about the characteristics common to groups from the analysis of the components.

ICA has been used successfully in various areas, for example, sea levels, electroencephalographic (EEG), seed X-ray images and functional magnetic resonance imaging (fMRI) data.

[1] Universidade Federal de Lavras - UFLA, Departamento de Estatística, Caixa Postal: 3037, CEP: 37200-900, Lavras, MG, Brasil. E-mail: *safadi@ufla.br; leilamaria2003@gmail.com*

Global control of tuberculosis is hampered by slow, insensitive diagnostic methods, particularly for the detection of drug resistant (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 multidrug resistant (MDR) strains. Perdigão et al. (2010) worked to characterize the genetic changes associated with the high number of extensively drug resistant (XDR) that threatens the global control of human tuberculosis (TB) worldwide.

Apart from molecular methods based on whole genome sequences of *Mycobacterium tuberculosis* (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 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.

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).

Ferreira et al (2017) showed that the behavior of GC-content sequence can be succinctly described regarding the non-decimated wavelet transform, and indicated 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.

Ferreira et al (2018) proposed to evaluate the similarity of genomes by combining discrete non-decimated wavelet transform and elastic net. The main feature of the elastic net is the grouping of correlated variables where the number of predictors is greater than the number of observations. The combination of these two methodologies applied in the analysis of clustering of the strains of genome of the *Mycobacterium tuberculosis* proved to be quite effective, being able to identify the clusters at each level of decomposition.

This paper proposes the use of independent component analysis technique to find similarities of *Mycobacterium tuberculosis* genomes. A hierarchical clustering is applied to the parameter estimates of the mixing matrix. Ten sequences of the MTB genome are analyzed, obtained from the National Center for Biotechnology Information (NCBI, 2017) by using the R software (R CORE TEAM, 2019).

2 Methodology

We describe briefly the Independent Component Analysis and the GC-content method.

2.1 Independent component analysis

Blind source separation (BSS), a well-known problem, aims at recovering the sources from a set of observations. Applications include separating individual voices in cocktail party. BSS is a difficult task because we do not have any information about the sources and the mixing process. Independent component analysis, ICA, is a method tackling this problem by assuming that the sources are independent to each other (HYVÄRINEN; OJA, 2000), and finds the demixing matrix and corresponding independent signals from the observations with some criteria making the separated signals as independent as possible.

ICA can also be contrasted with principal component analysis (PCA). Both ICA and PCA linearly transform the observed signals into components. The key difference however, is in the type of the components obtained. The goal of PCA is to obtain principal components which are uncorrelated. Moreover, PCA gives projections of the data in the direction of the maximum variance. The principal components are ordered in terms of their variances. In ICA however, we seek to obtain statistically independent components. PCA algorithms use only second order information. On the other hand, ICA algorithms may use higher order statistical information for separating the signals (see for example COMON,1994) . For this reason non-Gaussian signals (or at most, one Gaussian signal) are normally required for ICA algorithms based on higher order statistics (CARDOSO, 1999; HYVÄRINEN; OJA, 2000).

Various ICA algorithms have been proposed. Karhunen, Pajunen and Oja (1998) studies allowed to interpret the ICA as a nonlinear extension of principal components analysis. This approach played a key role in understanding the ICA as a relevant issue in multivariate data analysis. Hyvärinen contributed to the development of criteria based on maximization of non-gaussianity, which is based on the algorithm FastICA (Fast Independent Component Analysis) (HYVÄRINEN; OJA, 2000).

Given a microarray data set $Y = (\mathbf{y_{ij}})_{m \times N} = (y_1, y_2, ..., y_m)^T$ (T means transpose) whose m rows are N-dimensional time series, each element y_{ij} in the matrix Y corresponds to the value at time j for the i-th series. We consider that the matrix Y is generated by mixing m mutually independent components expressed by

$$\mathbf{Y}_{m \times N} = \mathbf{A}_{m \times m}.\mathbf{S}_{m \times N} \tag{1}$$

where $\mathbf{A}$ is the matrix of coefficients (a_{ij}) of the linear combination, named mixing matrix, while $\mathbf{S}$ is the matrix of independent components $\mathbf{s_j}$.

Aimed at size reduction, a number $k < m$ of independent components (IC) can be selected by using principal component analysis (PCA) as pre-processing for

ICA, so that

$$\mathbf{Y}_{m\times N} \approx \mathbf{A}_{m\times k}.\mathbf{S}_{k\times N}. \tag{2}$$

Each series y_i is decomposed into a linear combination of ICs (basis) given by

$$y_i = a_{i1}\mathbf{s}_1 + a_{i2}\mathbf{s}_2 + \ldots + a_{ik}\mathbf{s}_k, \tag{3}$$

for every $i = 1, 2, \ldots, m$, so that each series is represented by the coefficients of each independent component of the mixture.

Based on the estimates of rows $\mathbf{a_i}$ of the mixing matrix $\mathbf{A}$, a complete linkage method is performed to identify groups based on the independent components.

2.2 Mycobacterium tuberculosis strains and GC-Content

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; DEWAN, 2016).

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; 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 :

$$GCcontent = \frac{(nG + nC)}{(nA + nG + nC + nT)}, \tag{4}$$

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; DEWAN, 2016). Theses gene-rich regions bring significant biological information about the genome. Cheng et al. (2016) worked with high GC content, aiming the development of new molecular markers, highlighting the importance of working with gene-rich regions.

The sequences analyzed correspond to the strains of the MTB genome. Ten sequences were analyzed, being 2 Drug Resistant (DR), 6 Drug Susceptible (DS), 1 Multi Drug Resistant (MDR) and 1 Extensively Drug Resistant (XDR). Table 1 shows the description of each strain of the Mycobacterium tuberculosis genome, obtained from the National Center for Biotechnology Information (NCBI, 2017).

Table 1 - Descriptions of the *Mycobacterium tuberculosis* strains

Sequences	Descriptions of the strains
$Seq1\_DS$	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.
$Seq2\_DS$	Susceptible strain representing the largest portion of patients' tuberculosis isolates recovered during an epidemic in the Western Cape of South Africa.
$Seq3\_DS$	Susceptible strain belonging to the Beijing family, sequenced for comparative genomic studies.
$Seq4\_DR$	Resistant strain isolated in 2004, referring to a patient with secondary pulmonary tuberculosis, sequenced for comparative genomic studies.
$Seq5\_DR$	Drug resistant strain, having an accelerated rate of transmission between humans under agglomeration conditions.
$Seq6\_MDR$	Strain correspond to a single patient in KwaZulu-Natal, South Africa.
$Seq7\_XDR$	Strain correspond to a single patient in KwaZulu-Natal, South Africa.
$Seq8\_DS$	Susceptible strain used for comparative genomic studies.
$Seq9\_DS$	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.
$Seq10\_DS$	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 This strain was obtained through an aging

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

The GC-content of all the sequences was evaluated using a sliding window of 10,000 bases.

3 Results

Figure 1 shows the size and signal behavior visualization of each GC content sequences. Note that the sequences show practically the same behavior. In the x-axis we have the amount of nucleotides of each sequence.

The FastICA algorithm was applied with 2, 3, 4 and 5 ICs into each GC-content sequences. A hierarchical clustering was applied considering the estimated values of the mixing matrix. Figure 2 shows the dendrogram in each case.

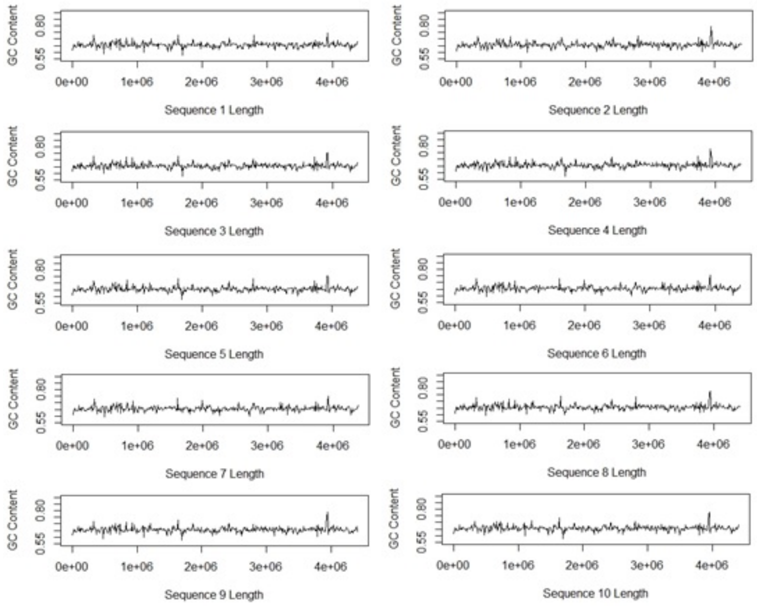
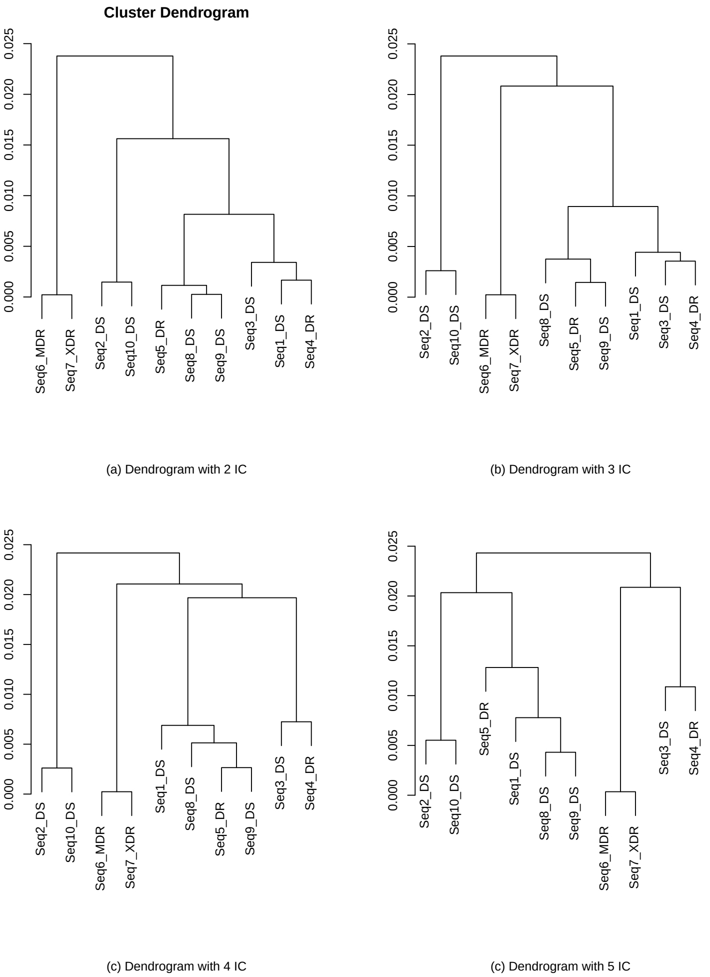


Figure 1 - GC content sequence sign (10,000 bp window) of MTB strains.



(a) Dendrogram with 2 IC

(b) Dendrogram with 3 IC

(c) Dendrogram with 4 IC

(c) Dendrogram with 5 IC

Figure 2 - Genome clusters (a) 2 Independent Components, (b) 3 Independent Components, (c) 4 Independent Components, (d) 5 Independent Components.

We observed that for 2 and 3 ICs we have the formation of the same groups. The same occurs when 4 and 5 CIs were considered. Regardless of the number of components considered, the sequences Seq6-MDR and Seq7-XDR; Seq2-DS and Seq10-DS are always grouped. The only difference concerns the Seq1-DS sequence which, with 4 and 5 CIs, is no longer grouped with Seq3-DS and Seq4-DR.

By increasing the number of independent components, we get more information about the common characteristics of the groups, this is observed in the classification of the sequence Seq1-DS in distinct group of the sequences Seq3-DS and Seq4-DR. .Interestingly, Seq3-DS and Seq4-DR are sequenced for comparative genomic studies, while Seq1-DS is sensitive to all common medicines used to treat tuberculosis.

Saini and Devan(2016) based on the calculation of the energy of wavelet decomposition coefficients of complete genomic sequences showed that the genomic sequences of MTB can be grouped only into two groups. The first group with the DS and DR sequences (lower energy) and the second group with MDR and XDR sequences (highest energy). Ferreira, Sáfadi and Lima(2017) considering the energy in each level of detail, were able to identify more than two groups. It was possible to detect particularities of sequences Seq1-DS, Seq3-DS and Seq4-DR with the proposed methodology.

Conclusions

The methodology applied in the analysis of clustering of the strains of genome of the Mycobacterium tuberculosis proved to be quite effective.

Acknowledgments

The authors thank editors and two anonymous referees for their helpful comments and suggestions.

SÁFADI, T.; FERREIRA, L. M. Avaliação de similaridades em genomas utilizando componentes principais. *Rev. Bras. Biom.,* Lavras, v.38, n.1, p.92-101, 2020.

- *RESUMO: Propomos o uso da análise de componentes independentes para encontrar semelhanças entre genomas. Considerando diferentes números de componentes independentes, o método de ligação completo foi utilizado para identificar grupos com base nos coeficientes estimados da matriz de mistura. As sequências analisadas correspondem às cepas do genoma do Mycobacterium tuberculosis, sendo analisadas dez sequências, obtidas no Centro Nacional de Informações em Biotecnologia (NCBI, 2017). As sequências de GC-content foram obtidas usando uma janela deslizante de 10.000 bases. A análise de agrupamento utilizando os componentes independentes das seqüências analisadas foi essencial para verificar a dissimilaridade das seqüências.*

- *PALAVRAS-CHAVE: GC-content; genomas do Mycobacterium tuberculosis; análise de agrupamento.*

References

ANASTASSIOU, D. Frequency-domain analysis of biomolecular sequences. *Bioinformatics*, v.16, p.1073-1081, 2000.

BAO, J.P. ; YUAN, R.Y. A wavelet-based feature vector model for DNA clustering. *Genetics and Molecular Research*, v.14, p.19163-19172, 2015.

BOEHME, C. C.; NABETA, P.; HILLEMANN, D.; NICOL, M. P., SHENAI, S.; KRAPP, F.; ALLEN, J.; TAHIRLI, R.; BLAKEMORE, R.; RUSTOMJEE, R.; MILOVIC, A.; JONES, M.; O'BRIEN. S. M.; PERSING, D. H.; GERDES, S. R.; GOTUZZO, E.; RODRIGUES, C.; ALLAND, D.; PERKINS, M. D Rapid molecular detection of tuberculosis and rifampin resistance. *The New England Journal of Medicine*, v.363, p. 1005-1015, 2010.

CARDOSO, J. F. High-order contrast for independent component analysis. *Neural Computation*, v.11, n.1, p.157–192, 1999.

CHANG, G. J. ; SEYFERTY, H. M.; SEN, X. Z. Adaption of SYBR green-based reagent kit for real-time PCR quantitation of GC-rich DNA. *Genetics and Molecular Research*, v.14, p.8509-8515, 2015.

CHENG, J. L.; QIU, Y. M.; WEI, C. L.; YANG, L. QL ; FU, J. J. Development of novel SCAR markers for genetic characterization of Lonicera japonica from high GC-RAMP-PCR and DNA cloning. *Genetics and Molecular Research* 15:gmr7737, 2016.

COMON, P. Independent component analysis, a new concept? *Signal Processing*, v.36, p.287–314, 1994.

FERREIRA, L. M. ; SÁFADI, T. ; FERREIRA, J. L. Wavelet-domain elastic net for clustering on genomes strains. *Genetics and Molecular Biology*, v.41, p.1-9, 2018.

FERREIRA, L. M. ; SÁFADI, T. ; LIMA, R. R. de . Evaluation of genome similarities using the non-decimated wavelet transform. *Genetics and Molecular Research*, v.16, n.3, gmr16039758, 2017.

HYVÄRINEN, A.; OJA, E. Independent component analysis: algorithms and applications. *Neural Networks*, v.13, n.4, p.411–430, 2000.

KARHUNEN, J.; PAJUNEN, P.; OJA, E., The nonlinear PCA criterion in blind source separation: relations with other approaches, *Proc. Neurocomput*, v.22, n.1, p.5-20, 1998.

MANN, S. ; CHEN, Y. P. P. Bacterial genomic G+C composition-eliciting environmental adaptation. *Genomics*, v.95, p.7–15, 2000.

NCBI, National Center for Biotechnology Information. Mycobacterium tuberculosis. Genoma, https://www.ncbi.nlm.nih.gov/assembly/GCF_000224435.1/ (July 2, 2017).

PERDIGÃO, J., MACEDO, R; MALAQUIAS, A.; FERREIRA, A; BRUM, L.; PORTUGAL, I. Genetic analysis of extensively drug-resistant Mycobacterium tuberculosis strains in Lisbon, Portugal. *J Antimicrob Chemother*, p.65224-65227, 2010

R CORE TEAM. *R: a language and environment for statistical computing.* Vienna, Austria. https://www.R-project.org/, 2019

SAINI, S ; DEWAN L . Application of discrete wavelet transform for analysis of genomic sequences of Mycobacterium tuberculosis. *SpringerPlus*, v.5, n.64, 2016.

Received on 12.06.2019.

Approved after revised on 30.09.2019.