We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,300
Open access books available

116,000
International authors and editors

125M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter

The Bioinformatics Tools for Discovery of Genetic Diversity by Means of Elastic Net and Hurst Exponent

Leila Maria Ferreira, Thelma Sáfadi, Tesfahun Alemu Setotaw and Juliano Lino Ferreira

Abstract

The genome era allowed us to evaluate different aspects on genetic variation, with a precise manner followed by a valuable tip to guide the improvement of knowledge and direct to upgrade to human life. In order to scrutinize these treasured resources, some bioinformatics tools permit us a deep exploration of these data. Among them, we show the importance of the discrete non-decimated wavelet transform (NDWT). The wavelets 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 analysis using technique of wavelets has been growing increasingly in the study of genomes. One of the great advantages associated to 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, such as physics, mathematics, engineering, and genetics, among others. In this context, we believe that using R software and applied NDWT coupled with elastic net domains and Hurst exponent will be of valuable guideline to researchers of genetics in the investigation of the genetic variability.

Keywords: wavelet, genome, NDWT, elastic net, Hurst exponent

1. Introduction

The genome era allowed us to evaluate different aspects on genetic variation, with a precise manner followed with a valuable tip to guide the improvement of knowledge and direct to upgrade to human life. In order to scrutinize these treasured resources, some bioinformatics tools permit us a deep exploration of these data. Among them, we display the significance of the discrete non-decimated wavelet transform (NDWT). The wavelets they possess improved capability to identify occult constituents of biological data and do a well-organized connection amid biological systems and the mathematical items used to designate them. The decomposition of signals/sequences at diverse stages of resolution allows

obtaining different characteristics in each level. The analysis using technique of wavelets has been growing increasingly in the study of genomes. One of the great advantages associated to this method corresponds to the computational gain, that is, the analyses are processed almost in real time. The applicability is in numerous themes of science, as physics, mathematics, engineering, genetics, meteorology, and oceanography, among others. The wavelet transform comprehends a technique of see and represents a signal. This signal is decomposed in resolution intensities, where each level brings a detailing. Mathematically, it is embodied by a function oscillating in time or space. As characteristic, it has sliding windows that expand or compress to capture low- and high-frequency signals. Its starting point arose in the field of seismic training to designate the instabilities ascending from a seismic impulse. Among the wavelets 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 procedure, the inductance is paraphrase invariants, to be exact; the selection of origin is irrelevant, provided all the observations are used in the analysis, a condition that does not happen in the discrete decimated wavelet transform (DWT). The technique of discrete wavelet transforms is being used to find gene locations in genomic sequences, detecting long-range correlations, discovering periodicities in sequences of DNA and analysis of G + C patterns. The NDWT technique may be applied in any genome type, increasing the promptness of the analysis, because the analyses with this method are processed almost in real time. The wavelets have demonstrated to be an efficient method in the analysis of DNA sequences. This tool is imperative to be applied to elastic net. The main feature of the elastic net technique is the grouping of correlated variables where the quantity of predictors is greater than the quantity of remarks. Furthermore, the Hurst exponent allows the evaluation of genome similarities. In the same way, the NDWT is crucial to evaluate the Hurst exponent. Strictly speaking, the bioinformatics tool NDWT is a fundamental step to allow the examination of genomic variation through the other subsequent bioinformatics tools, like elastic net and Hurst exponent, which allow us to understand, interpret, and identify the genome variation in a certain species.

2. Wavelet

Wavelet analysis, nowadays, is used widely in subjects such as signal processing, engineering, physics, genetics, mathematics, medical sciences, economics, astronomy, etc. The genetic approach of this tool appears to be a valuable and interesting possibility in science.

Wavelet is miniature wave. Whatsoever their form has a distinct number of oscillations and lasts through a definite period of time or space. Wavelets hold countless appropriate properties. Wavelets possess gender categories: there are father wavelets φ and mother wavelets ψ. The father wavelet fits to 1, and the mother wavelet fits to 0. Wavelets also arise in different shapes. There are the discrete ones, the symmetric, the nearly symmetric, and the asymmetric. The key aspect of wavelet investigation is that it allows the researcher to separate out a variable or signal into its essential multiresolution components [1].

In the last 21 years, more than 2000 articles were published with wavelet technique in wide-ranging subjects.

Wavelet theory delivers an integrated background for number methods which had been established autonomously for several signal processing applications [2]. Wavelet concept is established on Fourier analysis [3], in which all function may be denoted as the sum of sine and cosine functions.

Non-decimal wavelet transform (NDWT) possesses ample spectra of application, including mammographic imaginings, geology, genomes, applied mathematics, applied physics, atmospheric sciences, and economics, among other applications. In our specific case, we will approach the genomic approach.

When working with the complete genome, which is all the heritable information of an organism that is set in DNA or, in some viruses, in RNA, this includes both the genes and the noncoding sequences of a specific species; the main feature we find is the large volume of data. To elucidate this problem, the technique called wavelets has emerged as an efficient alternative in data compression, owning one of the main advantages that this technique offers. However, wavelet functions are also commanding apparatuses in signal processing, noise elimination, separation of components in the signal, identification of singularities, and detection of self-similarity, among others.

The goals of this examination address to show how wavelets possibly will be used in the analysis of genome clustering using the energy and interaction of wavelet functions with data grouping techniques (elastic net and Hurst exponent).

Structure of the analysis: first it is required to acquire the signal of the genome that will be analyzed; for this purpose, it is used to the tool called GC content. The signal if is required to apply a wavelet transform, in this case the NDWT will be used, working with the Daubechies wavelet with a certain number of null moments. The amount of decomposition levels will depend on the size of the genome. The scalogram is calculated using the detail coefficients obtained through the decomposition levels. The clustering analysis is done using the dendrogram with medium binding and applying the Mahalanobis distance.

In order to apply the elastic net technique in wavelet transform (NDWT), all levels of decomposition are used; as a characteristic of this interaction, it is possible to see the groupings at each of the decomposition levels.

Applying the Hurst exponent technique on the levels of signal decomposition, each level brings information regarding the amount degree of Hurst exponent index. All values found for the Hurst exponent are used in the dendrogram with the mean binding and the distance of Mahalanobis. There are several methods of estimation of Hurst exponent; the most commonly used is the R/S method.

3. Wavelet transform

Wavelet analysis has arisen as a possible device for spectral investigation owing to the interval incidence localization which makes it appropriate for multifaceted and motionless signals. The wavelet transform has added meaningfully in the training of many processes/signals in virtually all areas of earth science [4].

Wavelet is mathematical function. To be considered a wavelet, it must have the total area on the function curve equals to zero. The energy of the behavior must be limited (regularity and well located). Another need in the art is the speed and ease of calculating the wavelet transform and the inverse transform.

Among various application areas of wavelets are computer vision, data compression, fingerprint compression at the FBI, data recovery affected by noise, similar behavior detection, musical tones, astronomy, meteorology, numerical image processing, and many others.

The wavelet transform rots a function demarcated in the period domain into another function, well-defined in time domain and frequency domain. It is defined as

$$W(a,b) = \int_{\infty}^{\infty} f(t) \frac{1}{\sqrt{|a|}} \psi^{*}\left(\frac{t-b}{a}\right) dt, \tag{1}$$

which is a behavior function of two real parameters, a and b. If we define $\psi_{a,b}(t)$ as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi^*\left(\frac{t-b}{a}\right), \tag{2}$$

we may put another way the transform as the inner output of the functions $f(t)$ and $\psi_{a,b}(t)$:

$$W(a,b) = \left\langle f(t), \psi_{a,b}(t)\right\rangle = \int_{-\infty}^{\infty} f(t)\,\psi_{a,b}^*(t)dt. \tag{3}$$

The function $\psi(t)$ which equals $\psi_{1,0}(t)$ is entitled the mother wavelet, while the other functions $\psi_{a,b}(t)$ stay called daughter wavelets. The parameter b designates that the function $\psi(t)$ has been translated on the t axis of a distance equivalent to b, being then a translation parameter. The parameter causes a change of scale, increasing (if $a > 1$) or decreasing (if $a < 1$) the wavelet formed by the function. Consequently, the parameter "a" remains known as the scaling parameter.

4. Wavelet analysis

There are abundant types of wavelet transform. Rely on the procedure one can be desired that others. The wavelet analysis is prepared by the successive procedure of wavelet transform with several values for the criterion a and b, representing the decomposition of the signal components located in period and the agreeing to these parameters. Each wavelet has a better or worse location in the domains of frequency and of the time, so the analysis can be done with wavelets according to the desired result. Wavelet analysis brings with it an analysis of where the resolution level is set by the index a.

Discrete wavelets: among them are the Daubechies wavelet, wavelet of Cohen-Daubechies-Feauveau (occasionally mentioned as CDF N/P or Daubechies biorthogonal wavelets), Beylkin [5], BNC wavelets, Coiflet, Mathieu wavelet, Haar wavelet, binomial-QMF, Villasenor wavelet, Legendre wavelet, and symlet.

Continuous wavelets: (1) the real-valued wavelets are Mexican hat wavelet, Hermitian wavelet, beta wavelet, Hermitian hat wavelet, and Shannon wavelet, and the (2) complex-valued wavelets are Shannon wavelet, Morlet wavelet, complex Mexican hat wavelet, and modified Morlet wavelet.

In the latest decades, the investigation using method of wavelets has been rising progressively. One of the great rewards related with this method links to the computational improvement, that is, the analyses are treated virtually in real time. The applicability is in numerous areas of science, like physics, mathematics, engineering, and genetics, among others.

The wavelet transform is a method of sighted and characterizes a signal. Mathematically, it is characterized by a function wavering in time or space. As a characteristic, it has sliding windows that increase or bandage to capture low- and high-frequency signals, respectively [2]. Its origin arose in the field of seismic study to define the instabilities ascending from a seismic impulse [6].

Among the wavelet techniques, we have the discrete non-decimated wavelet transform (NDWT), whose main characteristic is that it may work with any extent of signals/sequences.

In this procedure, the coefficients are translation invariants, that is, the selection of source is unrelated since all the annotations are done in the investigation, a condition that does not happen in the discrete decimated wavelet transform (DWT).

In recent period, the discrete wavelet transforms were worn to find gene sites in sequences of the genome [7], finding long-range correlations, finding periodicities in sequences of the DNA molecule [8], and also in the scrutiny of G + C patterns [9].

The clustering analysis is often assumed to deal with DNA sequences proficiently. A wavelet-based element vector model was anticipated for grouping of DNA sequences [10].

The distinction of the discrete NDWT is to retain the similar extent of data in even and odd decimations on each measure and remain to do the identical on each subsequent scale, being D0 the dyadic decimation, D1 the odd decimation, H the high-pass filter, and L the low-pass filter. Consider, for example, an input path $(y_1,\ldots,y_n)$. Then, put on and preserve both D_0H_y and D_1H_y, even and odd indexed of the observation-filtered wavelets. Each of these sequences is length $n/2$. Consequently, in whole, the amount of wavelet coefficients in both decimals on the better scale is $2 \times n/2 = n$ [11].

5. GC content

An important parameter in genetics is the GC content. They are referred as the percentage of each bases of nitrogen composition of the molecule of DNA or RNA. We own the adenine, cytosine, guanine, thymine, and uracil. They are called by the acronyms A, C, G, T, and U, respectively. The last one belongs to RNA molecule replacing thymine. They are applied to the complete genome or determined fragment. This concept may be applied in coding or noncoding molecule segment. The adenine has the same quantity of thymine (DNA) or uracil (RNA). The cytosine has the same sum of guanine in either RNA or DNA. The amount of GC is related to high-stability one which value is less than AT or AU. In the opposite is low stability when this quantity is relatively small compared with AT or AU. This detail is because GC has three hydrogen bonds, although AU or AT has two of them.

The GC proportion inside a genome is established to be evidently variable. The DNA coding section is straight proportional to stand-up G + G.

In varied organisms, GC content is found to be too variable, which donate the dissimilarities in recombination pattern, including association with DNA repair, selection, and in the alteration of mutational bias patterns. Due to the essence of the genetic coding, it is nearly incredible for an organism to have a genome with a GC content pending either 0 or 100%. An organism species with an exceptionally low GC content is *Plasmodium falciparum* having about 20% of GC amount, published at NCBI—available at https://www.ncbi.nlm.nih.gov/bioproject?cmd=Retrieve&dopt=Overview&list_uids=148.

The GC percentage is the largely used systematic approaches in many prokaryotic organisms mainly in bacteria species. Actinobacteria are one example of uppermost GC bacterial content. Another example is *Streptomyces coelicolor* being 72% of G + G amount.

Interestingly, the software apparatuses GCSpeciesSorter [12] and TopSort [13] are used for categorizing species centered on their GC contents.

6. Daubechies wavelet

The Daubechies wavelets, established on the study done by Ingrid Daubechies, comprise of a series of orthogonal wavelets determining a discrete wavelet transform and categorized by a greatest amount of disappearing moments for certain given provision. With every wavelet assembly of this category lies in a scaling function (entitled the father wavelet) that produces an orthogonal multiresolution investigation.

Ingrid Daubechies is a Belgian physicist and mathematician. Daubechies was the first female to be chair of the International Mathematical Union (2011–2014). She is very well acknowledged for her study using wavelets in image compression.

Daubechies earned the Louis Empain Prize for Physics in 1984, conferred once every 5 years to a Belgian scientist on the basis of a study done before the age of 29. In the middle of 1992 and 1997, she stood a partner of the MacArthur Foundation, in addition in 1993, she was designated to the American Academy of Arts and Sciences. In 1994, she earned the American Mathematical Society Steele Prize for explanation for her book *Ten Lectures on Wavelets* and was requested to provide an entire talk in Zurich at the International Congress of Mathematicians. In 1997, she stood granted the AMS Ruth Lyttle Satter Prize available at http://www.ams.org/profession/prizes-awards/pabrowse#year=1997. In 1998, she was selected to the United States National Academy of Sciences, which can be visualized at http://nas.nasonline.org/site/Dir/1753239219?pg=vprof&mbr=1001102&returl=http%3A%2F%2Fwww.nasonline.org%2Fsite%2FDir%2F1753239219%3Fpg%3Dsrch%26view%3Dbasic&retmk=search_again_link and acquired the Golden Jubilee Award for Technological Innovation from the IEEE Information Theory Society (https://www.itsoc.org/honors/golden-jubilee-awards-for-technological-innovation). She turns into an overseas fellow of the Royal Netherlands Academy of Arts and Sciences in 1999 accessible at https://www.knaw.nl/en/members/foreign-members/4013.

In 2000, Daubechies turns out to be the pioneer lady to obtain the National Academy of Sciences Award in Mathematics, stated every 4 years for excellence in published mathematical investigation. The prize honored her for important findings on wavelets and wavelet growths and designed for her accomplishment in building wavelet methods a constructive elementary apparatus of applied mathematics. This achievement is presented on https://www.knaw.nl/en/members/foreign-members/4013. She was also conferred the Basic Research Award, German Eduard Rhein Foundation, which could be visualized on https://web.archive.org/web/20110718233021/http://www.eduard-rhein-stiftung.de/html/Preistraeger_e.html and https://web.archive.org/web/20110718234059/http://www.eduard-rhein-stiftung.de/html/2000/G00_e.html and the NAS Prize in Mathematics https://web.archive.org/web/20101229195210/http://www.nasonline.org/site/PageServer?pagename=AWARDS_mathematics.

Generally, the Daubechies wavelet properties stay preferred to have the maximum sum A of vanishing moments (this does not make sure of indicating the preeminent levelness) on behalf of assumed provision measurement 2A-1 [3]. It is present in two designation patterns in routine, DN via the extent or total of blows and dbA stating to the quantity of vanishing moments. Thus db2 and D4 stand the equivalent wavelet transform.

Among the 2A-1 thinkable resolution of the arithmetical calculations for the moment and orthogonal circumstances, the one is elected whose scaling filter has extreme phase. Wavelet transform remains too easy to place hooked on training through the debauched wavelet transform. Daubechies wavelets are broadly used in answering wide-ranging problems, for example, self-homology assets of sign or fractal difficulties and sign cutoffs, among others.

Daubechies wavelets remain not demarcated in footings of the subsequent scaling and wavelet functions; actually, they are not probable to inscribe down in locked procedure.

In the production of a wavelet scaling arrangement, low-pass filter and the wavelet sequence band-pass filter will standardized to ensure entirety unenliven 2 and summation of squares unenliven 2. In particular requests, they are standardized to require sum$\sqrt{2}$; thus one and other arrangements and entirely changes of them by an even sum of coefficients are orthonormal to each other.

The employment of Daubechies wavelets though software such as Mathematica rope straight mode is available at https://reference.wolfram.com/language/ref/DaubechiesWavelet.html, a basic execution is humble in MATLAB. This application routines periodization to grip the problematic of limited measurement signals. Other, further refined devices are accessible, but habitually it is not required to use these as it merely touches the many split ends of the converted signal. The periodization is fulfilled in the onward transform straight in MATLAB vector system and the inverse transform by means of the circshift() function.

7. Non-decimal wavelet transform

Non-decimal wavelet transform (NDWT) has the benefits of period invariance and redundancy, paralleled to the typical orthogonal wavelet transformations. NDWT owns properties beneficial in various wavelet applications. Furthermore, NDWT matrix is capable to powerfully map a signal arising from an acquirement field to the wavelet sphere with humble matrix multiplication and deprived of the prerequisite of the whole quantity of the signal [14].

A widespread version of wavelet transform is a NDWT, which can overwhelm sensitivity to translations in time and change found in typical [15] orthogonal wavelet transform. Initially in the 1990s, NDWT arose in scientific literature using several names for a figure of applications and purposes [16].

A process that approaches nonstop wavelet transform with an iterative algorithm, which evicted to be corresponding to a shift-invariant representation, was put forward by [17]. Furthermore, a resourceful algorithm was defined with $O(n \log 2\ (n))$ complexity for scheming wavelet coefficients that stand shift-invariant, to be exact, humble repetitious wavelet coefficients at wholly N circulant shift for an input signal size of N [5, 18]. In addition, a wavelet packet decomposition for time invariance and applied it to estimation and detection problems was proposed by Pesquet and collaborators [19] and lengthy finished in the study [20], uses an over ample wavelet decomposition, which is stated to as discrete wavelet frame, for arrangement of texture. After that, two other studies [21, 22] tested translation-invariant transform to verge for noise reduction. Then, the study of stationary wavelet transform with example applications for local spectra estimation was published [23]. Finally, an examination of applied translation-invariant wavelet algorithm for data compression was done [24].

The time-invariance property of NDWT generates a reduced mean square error and also reduces the Gibbs phenomenon in d-noising applications [21]. Conversely, the defilement of variance maintenance in NDWT embarrasses the signal restoration [16].

Major benefits of a NDWT matrix are squeezability, calculation promptness, and tractability in magnitude of an input signal. We previously deliberated the superior compressibility when NDWT matrices are well-worn for 2-D scale-mixing transforms.

NDWT possess ample spectra of application, including mammographic imaginings, geology, genomes, physics, atmospheric sciences, and economics, among other applications.

8. Scalogram

Spectrogram is an ample prevalent tool in signal analysis because it provides a scattering of signal energy in time-frequency plane. The wavelet spectrogram

is broadly known like scalogram [25]. Comprehend a distribution of energy in timescale plane. The scalogram yields a more or less simply intelligible visual in two-dimensional representations of signals [26].

The scalogram is a valuable device for the understanding of the wavelet signal represented. It is like a graph of the square sum of the wavelet coefficients in different levels. In the occurrence of discrete transformation, it embodies a decomposition of function energy without timescale. One of its features is the aptitude to detect periodic components of the signal; either apparatuses will result in peaks in the scalogram. These apparatuses may be mined from the signal by dividing the ripple coefficients into different sets, where each of these sets is at the same peak. High- and low-frequency apparatuses of a signal might be restored by applying a reverse loop transformation to separate sets [27].

The energy $E\ (j)$ aimed at the wavelet d coefficients in each level j, which corresponds to the scalogram, is represented by

$$E(j) = \sum_{k=0}^{n} d_{j,k}^2 \quad para \quad j = 1, \ldots, J \tag{4}$$

9. Cluster analysis

Cluster analysis also known as unsupervised classification is a grouping of items into diverse groups, each of which requisite be assembled rendering to the rules of programming. This assembly must be handled computationally, without user intervention.

The term clustering analysis, early termed by [15], actually contains an assortment of different grouping algorithms, all of which address an important issue in several areas of research: how to organize observed data into structures that make sense or how to develop taxonomies capable of classifying data observed in different classes. Important is to even consider that these assembly must be classes that occur naturally in the dataset.

Clustering analysis is the designation given to the group of computational techniques whose purpose is to separate objects into groups, based on the characteristics that these objects have. The basic idea is to put objects in the same group that are similar in some predetermined criteria. The criterion is usually based on a dissimilarity function, which function receives two objects and returns the distance between them. The groups determined by a quality metric should have high internal and high homogeneity separation (external heterogeneity). This implies that the elements of a given set should be mutually similar and, preferably, have a high amount of differences from the elements of other sets [28].

Biologists, for instance, have to organize data observed in structures that make sense, that is, develop taxonomies. Microbiologists confronted with a variety range of species of a certain type, for example, must be capable to classify the observed specimens into clusters before it has been possible to describe these microorganisms in detail in ways to detach in detail the differences between species and subspecies.

Grouping procedures have been practiced in a huge range of areas. Ref. [29] already provides a broad overview of several published studies on the use of grouping analysis techniques. In the medical field, for example, grouping of diseases by symptom or cures can lead to very useful taxonomies. In areas of psychiatry, for example, clustering of syndrome, for instance, paranoia, schizophrenia, and others, is considered essential for proper therapy. In archeology, conversely, one has also tried to group civilizations or times of civilizations based on tools of stone, funerary

objects, etc. In general, whenever a "mountain" of unknown data is required to be classified into manageable cells, grouping methods are used.

10. Elastic net

In statistics and specifically in the suitable of linear or logistic regression models, the elastic net is a standardized regression method that linearly couples the L1 and L2 punishments of the lasso and ridge approaches. **Figure 1** shows the elastic net typical design.

Lasso is a regression method broadly worn in domains with huge datasets, such as genomic data, where proficient and agile algorithms are vital [30]. Ridge regression is a procedure for investigating manifold regression data that arise out of multicollinearity. When multicollinearity arises, least squares estimates are unbiased, but their variances are huge so they might be outlying from the accurate value. In 1970, the investigation of [31] published an article about ridge regression, approaching the tendentious appraisal for nonorthogonal issues. In 2009, [32] study examined the ridge regression and their extensions applied to genome-wide selection into *Zea mays* L.

R software, available at https://www.r-project.org/, has the packing necessary to do a wavelet and elastic net based on genome sequence. Furthermore, the elastic net may be also used with microsatellite (SSR) data. This tool could be used in any genetic data of all organisms.

The most relevant article about elastic net was published in 2005 [33]. They say that elastic net is of pronounced interest especially when the predictors' number is considerably higher than the sum of observations. This might be useful in real or in simulation data.

The recent evolution of science brought a fast deeper understanding of the genome. In this sense, through several methods with varying levels of complexity added to the computational efficiency at the present days, we may easily compare organisms based on their genetic dissimilarity. Along these lines, we used accurate

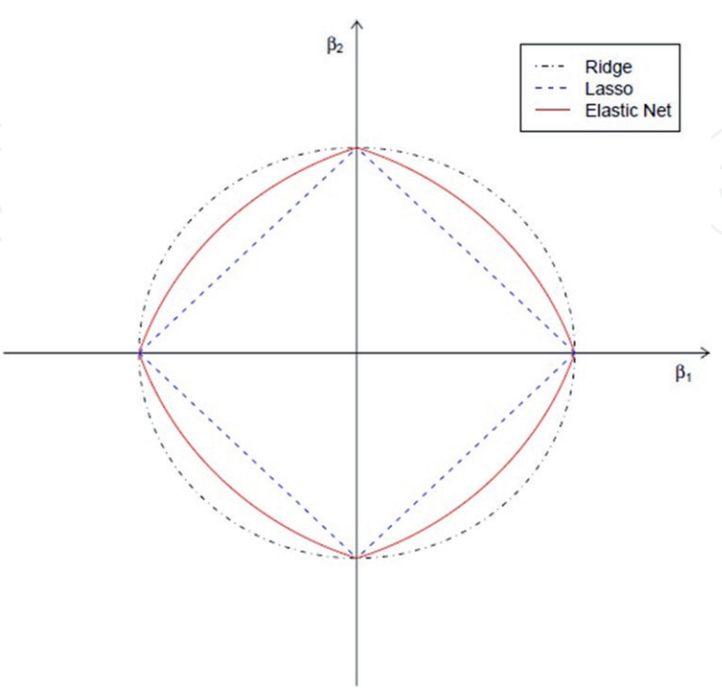

Figure 1.
Elastic net standard scheme.

genomic selection methods dropping the penalties of each approaches like in elastic net, enabling the fitting of a certain statistical model. Therefore, an outstanding methodology to analyze genome is elastic net domain used in several study, like [33–36].

Recently, the tuberculosis strain's differences were evaluated using the elastic net domain [34]. In that examination, 10 genome sequences of *Mycobacterium tuberculosis* with a window size of 10,000 bp were assessed combining the NDWT and elastic net domain. This study encompasses 10 strains: 2 from drug resistant, 6 from drug susceptible, 1 from multidrug resistant, and finally 1 from extensively drug resistant. The clustering detected on that analysis indicated to be real adequate.

11. Hurst exponent

Hurst exponent is applied as a degree of long-standing memory of time series. It associates to the autocorrelations of time series and the degree at which these decline as the lag between pairs of values intensifications. This coefficient has started to be established in hydrology, used to understand the variation level of dam size at Nile River over a long cycle of time. Harold Edwin Hurst was a British engineer that worked with hydrology; for this reason the coefficient has his surname. Later, this exponent was used in several areas, including fractal geometry, storage process, trends in financial market analyzing economic time series, mechanics, physics, mathematics, computation, and finally to the long-ranging dependency in DNA. **Figure 2** displays the values of Hurst exponent and their interpretation in a long-standing.

Using the genetic data, the Hurst exponent approach is able to build genetic cluster based on genome sequence. There are a lot of estimation methods of Hurst exponent: the original and best-known is the alleged rescaled range (R/S) analysis promoted by [37, 38] and based on previous hydrological findings [39]. Alternatives include DFA, periodogram regression [40] aggregated variances [41], local Whittle's estimator [42], and wavelet analysis [43, 44] both in the time domain and frequency domain.

In our case, we performed a Hurst exponent in the bacterial strains used in article [34]. We did many methods of Hurst exponent. Interestingly, the R/S methodology was the most similar to the cluster obtained on elastic net domain approach. This data is not shown due to being in a review process to an International journal currently. Our data agree with the majority of scientific papers published approaching the Hurst exponent, which so far applying the R/S method.

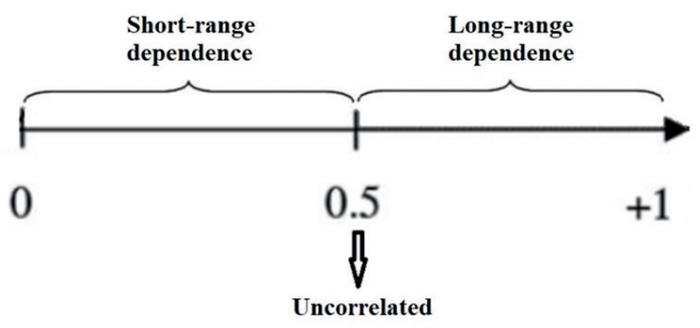

Figure 2.
Hurst exponent pattern interpretation of the index value.

DOI: http://dx.doi.org/10.5772/intechopen.82755

12. Conclusion

We strongly believe that exploring the genetic variability of any organism using wavelet coupled with elastic net domain and/or Hurst exponent will be a valuable and interesting tool. It is not difficult and the free R software could solve easily the approach. In this way, it gives reliability and robustness in your results. Therefore, these bioinformatics apparatuses provide more possibility to scrutinize the genetic divergence of living organisms.

Conflict of interest

The authors do not have conflict of interests.

Author details

Leila Maria Ferreira[1], Thelma Sáfadi[1], Tesfahun Alemu Setotaw[2] and Juliano Lino Ferreira[3*]

1 Universidade Federal de Lavras, Lavras, Minas Gerais, Brazil

2 Ethiopian Institute of Agricultural Research, Addis Ababa, Ethiopia

3 Empresa Brasileira de Pesquisa Agropecuária, Bagé, Rio Grande do Sul, Brazil

*Address all correspondence to: juliano.ferreira@embrapa.br

IntechOpen

© 2018 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. CC BY

References

[1] Crowley PM. A guide to wavelets for economists. Journal of Economic Surveys. 2007;**21**:207-267. DOI: 10.1111/j.1467-6419.2006.00502.x

[2] Percival DB, Walden AT. Wavelet Methods for Time Series. 1st ed. Cambridge: Analysis Cambridge University Press; 2000. 594 p. DOI: 10.1017/CBO9780511841040

[3] Dodin G, Vandergheynst P, Levoir P, et al. Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. Journal of Theoretical Biology. 2000;**206**:323-326

[4] Chamoli A. Wavelet analysis of geophysical time series. e-Journal Earth Science India. 2009;**2**:258-275

[5] Beylkin G. On the representation of operators in bases of compactly supported wavelets. SIAM Journal on Numerical Analysis. 1992;**29**:1716-1740

[6] Morlet J, Arens G, Fourgeau E, Giard D. Wave propagation and sampling theory–Part II: Sampling theory and complex waves. Geophysics. 1982;**47**:222-236. DOI: 10.1190/1.1441329

[7] Ning J, Moore CN, Nelson JC. Preliminary wavelet analysis of genomic sequences. In: Proceedings of the IEEE Computer Society Conference on Bioinformatics CSB '03. Stanford, California: IEEE; 2003. pp. 509-510

[8] Vannucci M, Liò P. Non-decimated wavelet analysis of biological sequences: Applications to protein structure and genomics. Sankhyā: The Indian Journal of Statistics, Series B. 2001;**63**:218-233. DOI: 10.2307/25053172

[9] Daubechies I. Ten Lectures on Wavelets. 1st ed. Berlin: Springer-Verlag; 1992. 344 p

[10] Bao J, Yuan RY. A wavelet-based feature vector model for DNA clustering. Genetics and Molecular Research. 2015;**14**:19163-19172. DOI: 10.4238/2015.December.29.26

[11] Nason G. Wavelet Methods in Statistics with R. 1st ed. New York: Springer-Verlag; 2008. 259 p. DOI: 10.1007/978-0-387-75961-6

[12] Karimi K, Wuitchik D, Oldach M, Vize P. Distinguishing species using GC contents in mixed DNA or RNA sequences. Evolutionary Bioinformatics Online. 2018;**14**:1-4. DOI: 10.1177/1176934318788866

[13] Lehnert E, Mouchka M, Burriesci M, Gallo N, Schwarz J, Pringle J. Extensive differences in gene expression between symbiotic and aposymbiotic cnidarians G3: Genes, genomes. Genetics. 2014;**4**: 277-295. DOI: 10.1534/g3.113.009084

[14] Zhou H, Narayanan RM. Microwave imaging of non-sparse object using dual-mesh method and iterative method with adaptive thresholding. IEEE Transactions on Antennas and Propagation. 2018; early access. DOI: 10.1109/TAP.2018.2876164

[15] Tryon RC. Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. 1st ed. Ann Arbor: Edwards Brothers; 1939. 122 p

[16] Kang M. Non-decimated wavelet transform in statistical assessment of scaling: Theory and applications [thesis]. Atlanta: Georgia Institute of Technology; 2016

[17] Mallat S. Zero-crossings of a wavelet transform. IEEE Transactions on Information Theory. 1991;**37**:1019-1033. DOI: 10.1109/18.86995

[18] Shensa MJ. The discrete wavelet transform: Wedding the a trous and mallat algorithms. IEEE Transactions on Signal Processing. 1992;**40**:2464-2482. DOI: 10.1109/78.157290

[19] Pesquet JC, Krim H, Carfantan H. Time-invariant orthonormal wavelet representations. IEEE Transactions on Signal Processing. 1996;**44**:1964-1970

[20] Unser M. Texture classification and segmentation using wavelet frames. IEEE Transactions on Image Processing. 1995;**4**:1549-1560

[21] Coifman RR, Donoho DL. Translation-invariant de-noising. In: Antoniadis A, Oppenheim G, editors. Wavelets and Statistics. Lecture Notes in Statistics. Vol. 103. New York: Springer; 1995. pp. 1-26. DOI: 10.1007/978-1-4612-2544-7_9

[22] Lang M, Guo H, Odegard JE, Burrus CS, Wells RO Jr. Nonlinear processing of a shift-invariant discrete wavelet transform (dwt) for noise reduction. In: Szu HH, editor. Wavelet Applications. 2nd ed. Orlando: Proc. SPIE 2491; 1995. pp. 640-651. DOI: 10.1.1.24.4098

[23] Nason GP, Silverman BW. The stationary wavelet transform and some statistical applications. In: Antoniadis A, Oppenheim G, editors. Wavelets and Statistics. 1st ed. New York: Springer; 1995. pp. 281-299. DOI: 10.1007/978-1-4612-2544-7_17

[24] Liang J, Parks TW. A translation-invariant wavelet representation algorithm with applications. IEEE Transactions on Signal Processing. 1995;**44**:225-232

[25] Rioul O, Vetterli M. Wavelets and signal processing. IEEE Signal Processing Magazine. 1991;**8**:14-38. DOI: 10.1109/79.91217

[26] Grossmann A, Kronland-Martinet R, Morlet J. Reading and understanding continuous wavelet transforms. In: Combes JM, Grossmann A, Tchamitchian P, editors. Wavelets. Inverse Problems and Theoretical Imaging. Berlin, Springer; 1990. pp. 2-20. DOI: 10.1007/978-3-642-75988-8_1

[27] Liò P, Vannucci M. Finding pathogenicity islands and gene transfer events in genoma data. Bioinformatics. 2000;**16**:932-940. DOI: 10.1093/bioinformatics/16.10.932

[28] Linden R. Técnicas de agrupamento. Revista de Sistemas de Informação da FSMA. 2009;**4**:18-36

[29] Hartigan JA. Clustering Algorithms. 99th ed. New York: John Wiley & Sons; 1975. 369 p

[30] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010;**33**:1-22

[31] Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;**12**:55-67

[32] Piepho HP. Ridge regression and extensions for genome wide selection in maize. Crop Science. 2009;**49**:1165-1176. DOI: 10.2135/cropsci2008.10.0595

[33] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B: Statistical Methodology. 2005;**67**:301-320. DOI: 10.1111/j.1467-9868.2005.00503.x

[34] Ferreira LM, Sáfadi T, Ferreira JL. Wavelet-domain elastic net for clustering on genomes strains. Genetics and Molecular Biology. 2018;**4**:884-892. DOI: 10.1590/1678-4685-GMB-2018-0035

[35] Ogutu JO, Schulz-Streeck T, Piepho HP. Genomic selection using regularized

linear regression models: Ridge regression, lasso, elastic net and their extensions. BMC Proceedings. 2012;**6**:1-6. DOI: 10.1186/1753-6561-6-S2-S10

[36] Waldmann P, Mészáros G, Gredler B, Fuerst C, Sölkner J. Evaluation of the lasso and the elastic net in genome-wide association studies. Frontiers in Genetics. 2013;**4**:270. DOI: 10.3389/fgene.2013.00270

[37] Mandelbrot BB, Wallis JR. Noah, Joseph, and operational hydrology. Water Resources Research. 1968;**4**:909-918

[38] Mandelbrot BB, Wallis JR. Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. Water Resources Research. 1969;**5**:967-988

[39] Hurst HE. Long-term storage capacity of reservoirs. Transactions of the American Society of Civil Engineers. 1951;**116**:770-799

[40] Geweke J, Porter-Hudak S. The estimation and application of long memory time series models. Journal of Time Series Analysis. 1983;**4**:221-238. DOI: https://doi.org/10.1111/j.1467-9892.1983.tb00371.x

[41] Beran J, Terrin N. Estimation of the long-memory parameter, based on a multivariate central limit theorem. Journal of Time Series Analysis. 1994;**15**:269-278. DOI: 10.1111/j.1467-9892.1994.tb00192.x

[42] Robinson PM. Gaussian semiparametric estimation of long-range dependence. The Annals of Statistics. 1995;**23**:1630-1661

[43] Riedi RH. Multifractal processes. In: Doukhan P, Oppenheim G, Taqqu MS, editors. Theory and Applications of Long-Range Dependence. 1st ed. Boston: Birkhäuser Boston; 2003. pp. 625-716

[44] Simonsen I, Hansen A, Nes OM. Determination of the Hurst exponent by use of wavelet transforms. Physical Review E. 1998;**58**:2779-2787. DOI: 10.1103/PhysRevE.58.2779