Wavelet analysis of human DNA

J.A. Tenreiro Machado, António C. Costa, Maria Dulce Quelhas

ABSTRACT

This paper studies the human DNA in the perspective of signal processing. Six wavelets are tested for analyzing the information content of the human DNA. By adopting real Shannon wavelet several fundamental properties of the code are revealed. A quantitative comparison of the chromosomes and visualization through multidimensional and dendograms is developed.

Keywords

DNA, Chromosome, Wavelets, Signal processing, Multidimensional scaling, Dendograms

1. Introduction

With the progresses of genome sequencing considerable information is available for computational processing, leading to intense research in the information structure of DNA [1-10]. This paper addresses the deoxyribonucleic acid (DNA) code of the human being in the perspective of signal processing [11-13]. These ideas motivated the adoption of the wavelets for the study of the DNA information embedded in the human twenty four distinct chromosomes. Several types of wavelets are tested and, based on the emerging patterns, the real Shannon wavelet is considered as the best one for the analysis. The wavelet charts depict complex patterns and, due to the large number of cases, a comparison index is developed. Based on the similarity measure two visualization tools, namely multidimensional scaling (MDS) [14-18] and dendograms, are adopted [19]. The results reveal important properties, demonstrating the goodness of the proposed scientific tools.

Having these ideas in mind, this paper is organized as follows. Section 2 presents the fundamental biological concepts, the mathematical tools, and formulates the DNA sequence decoding algorithm. Section 3 analyzes the information content of human chromosomes in the perspective of wavelet analysis. Finally, Section 4 outlines the main conclusions.

2. Preliminaries: DNA, decoding and wavelets

Deoxyribonucleic acid (DNA) is a double helix constituted by two polymers connected by hydrogen atoms. Their structural and vibration regimes were studied in [20-24]. The polymers contain three types of nucleotides, namely deoxyribose (a five carbon sugar), a phosphate group, and a nitrogenous base. There are four distinct nitrogenous bases: thymine, cytosine, adenine, and guanine, denoted by the symbols {T, C, A, G}. Each type of base on one strand connects with only one type of base on the other strand, forming the "base pairing", with A and C bonding to T and G, respectively. The four bases are the basis of the genetic code, instructing cells how to synthesize enzymes and proteins. In a human being each cell holds twenty four chromosomes, each containing, on average, 160 million nucleotide pairs. This information is being collected during the last years and is available for scientific research.

Besides the four symbols {T, C, A, G}, the available chromosome data includes a fifth symbol "N" which has no practical meaning for the DNA decoding. Therefore, the translation of the symbols into numerical values must, on one hand, avoid improper effects that will deform the information processing, and, on the other hand, reflect the base pairing restriction and the fifth symbol existence. Bearing these ideas in mind, it is considered the symbol translation:

$$A = 1 + i0,\ C = -1 + i0, T = 0 + i1, G = 0 - i1, N = 0 + i0 \quad (1)$$

where $i = \sqrt{-1}$.

The definition of a translation scheme leads to a sequence of values along the DNA strand and, therefore, to a "signal" $x(t)$, where t may be interpreted as the "time" progress along each chromosome.

The conversion of the DNA message is not restricted to a symbol by symbol scheme and, in fact, can follow many distinct logical and mathematical algorithms. The authors developed experiments with translation of sequences of different lengths by means of a Gray code and a trigonometric function [11], a histogram counting [12] and entropy measure [13]. Additional experiments involved also the use of a four-dimensional Fourier transform, that is, with a separate dimension for each symbol. With some surprise it was observed that the results did not change significantly with the adopted scheme, which seems to support the idea that we are capturing global characteristics that emerge as long as the conversion scheme follows logical steps that do not deform the overall information content of the DNA message.

The signal $x(t)$ is complex and difficult to analyze in the "time domain". Therefore, the characteristics of $x(t)$ only emerge by applying signal processing tools. In this study is adopted the continuous wavelet transform [25-27] defined as:

$$\left[W_\psi x(t)\right](a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt, \; a>0 \tag{2}$$

where the symbol * denotes the complex conjugate, the parameters (a, b) represent the dyadic dilation and the dyadic position, respectively, and ψ is a function called the mother wavelet. The mother wavelet is the source for generating daughter wavelets, which are simply the translated and scaled versions of the mother wavelet. Often the parameter a is interpreted qualitatively as the inverse of the frequency of Fourier analysis.

In short, the wavelets process data at different scales or resolutions. If we look at a signal with a large window we notice gross features, while if we look at a signal with a small window we notice small details. The parameters a and b provide dilations and translations of the mother wavelet ψ and the temporal analysis is performed with a contracted version of the prototype wavelet, while frequency analysis is performed with a dilated version of the wavelet. Therefore, wavelets establish a balance between precision in the frequency and the time domains, according with an analogy the Heisenberg's uncertainty principle.

The choice of an appropriate mother wavelet plays an important role in the analysis and several methods were proposed. Nevertheless, often it is required some a priori knowledge of the signal characteristics, which is clearly not the present case. Therefore, in the study of the DNA message, the test of several functions and, in a second phase, the analysis of the results for choosing the best one, is considered to be a more robust approach to follow.

We investigate three real and three complex valued wavelets, namely the Haar, Ricker (also called Mexican hat), Shannon, Hermitian hat, Shannon complex and Morlet wavelets, denoted as HW, RW, SW, HHW, SCW, and MW, defined by the expressions:

$$Haar: \psi(t) = \begin{cases} 1 & 0 \le t < 1/2 \\ -1 & 1/2 \le t < 1 \\ 0 & \text{otherwise} \end{cases} \tag{3a}$$

$$\text{Ricker}: \psi(t) = \frac{2}{\sqrt{3\sigma}\,\pi^{\frac{1}{4}}}\left[1-\left(\frac{t}{\sigma}\right)^2\right]e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}, \; \sigma>0 \tag{3b}$$

$$\text{Shannon real}: \psi(t) = \frac{1}{\pi t}\left[\sin(2\pi t) - \sin(\pi t)\right] \tag{3c}$$

$$\text{Hermitian hat}: \psi(t) = \frac{2}{\sqrt{5}\pi^{\frac{1}{4}}}\left(1-t^2+it\right)e^{-\frac{t^2}{2}} \tag{3d}$$

$$\text{Shannon complex}: \psi(t) = \frac{\sin(\pi t)}{\pi t}e^{-i2\pi t} \tag{3e}$$

$$\text{Morlet}: \psi(t) = c_\sigma \pi^{-\frac{1}{4}}\left(e^{-\frac{1}{2}(t-\beta)^2}e^{i\alpha t} - e^{-\frac{1}{2}\sigma^2}\right); \; c_\sigma = \left(1+e^{-\sigma^2}-2e^{-\frac{3}{4}\sigma^2}\right)^{-\frac{1}{2}}; \; \sigma \in \mathbb{R}^+ \tag{3f}$$

Clearly we can get a considerable volume of information. Therefore, for condensing the results of the wavelet charts a similarity measure r between two plots is developed in the next section. This index allows the construction of a symmetrical correlation matrix R that compares all cases. Based on the matrix it is then possible to use visualization tools for establishing graphical locus of the twenty four chromosomes. In this paper are considered the MDS and the dendogram techniques. MDS assigns a point to each item in a multi-dimensional space and arranges them in a low-dimensional space in order to reproduce the observed similarities. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.

In synthesis, in the next section for the DNA human analysis is adopted (*i*) the set of twenty four human chromosomes {Hu1, Hu2, ..., HuX, HuY}, (*ii*) the translation of the DNA code by scheme (Eq. (1)), (*iii*) the continuous wavelet transform for the signal analysis, (*iv*) the six wavelets {HW, RW, SW, HHW, SCW, MW} defined in Eqs. (3a)-(3f), (*v*) the measure of similarity between wavelet charts using an appropriate index, and (*vi*) the adoption of two visualization techniques for obtaining a graphical output.

3. Wavelet analysis of human DNA

For each of the twenty four distinct human chromosomes the corresponding complex valued signal $x(t)$ is obtained and the wavelet transform $[W_\psi x(t)](a, b)$ is calculated. Nevertheless, the results of the wavelet analysis depend on the mother function ψ to be adopted. Therefore, before comparing all chromosomes, a preliminary evaluation is developed in order to evaluate the characteristics of each function ψ for the observation of the DNA signal. In this line of thought Fig. 1 depicts the absolute value of the wavelet for the Hu1 chromosome and the six functions listed in Eqs. (3a)-(3f), where it is adopted $\sigma = 1$ and $\sigma = 5$ for the RW and the MW, respectively. Furthermore, it is established a "time step" $dt = 1$ for the sequence base increment along the DNA, and the parameters (a, b) are considered to vary from zero up to the maximum length of the chromosome. In the charts we must take care with the results at the limits of the intervals of variation of the parameters (a, b) due to truncation effects.

We verify immediately that we get very different charts for each function ψ, that is, we conclude that the observation lens provided by ψ hide or reveal different signal characteristics and that some may be better adapted to this than others. The HW seems to be the "worst" probably because it is more adapted to digital signals while in the present case we have a different type of time evolution. We observe that the RW has similarities to HHW, and, identically, the SCW to the MW. In these four cases we observe a pattern for a in the middle of the interval and for low values of b. Particularly the RW seems to present a slight higher level of detail with the presence of three objects. The "best" wavelet is clearly the SW, with a clear emergence of three objects roughly for $a \approx \{0.7, 1.4, 2.1\} \times 10^8$, $b \approx 10^8$.

Other chromosomes were tested leading also to the conclusion that the SW produces charts with more clear patterns. Therefore, the SW is adopted in the analysis of the twenty four chromosomes. Fig. 2 depicts the absolute value of the SW of the human chromosomes where, for the sake of completeness, Hu1 is repeated.

We verify that the charts are distinct from each other but reveal similarities namely, the emergence of several objects for low values of b. While for most cases we have three objects (e.g., chromosomes Hu1 and Hu2) some charts present a larger number (e.g., chromosomes Hu3 and Hu18), or reveal several levels according with b following a

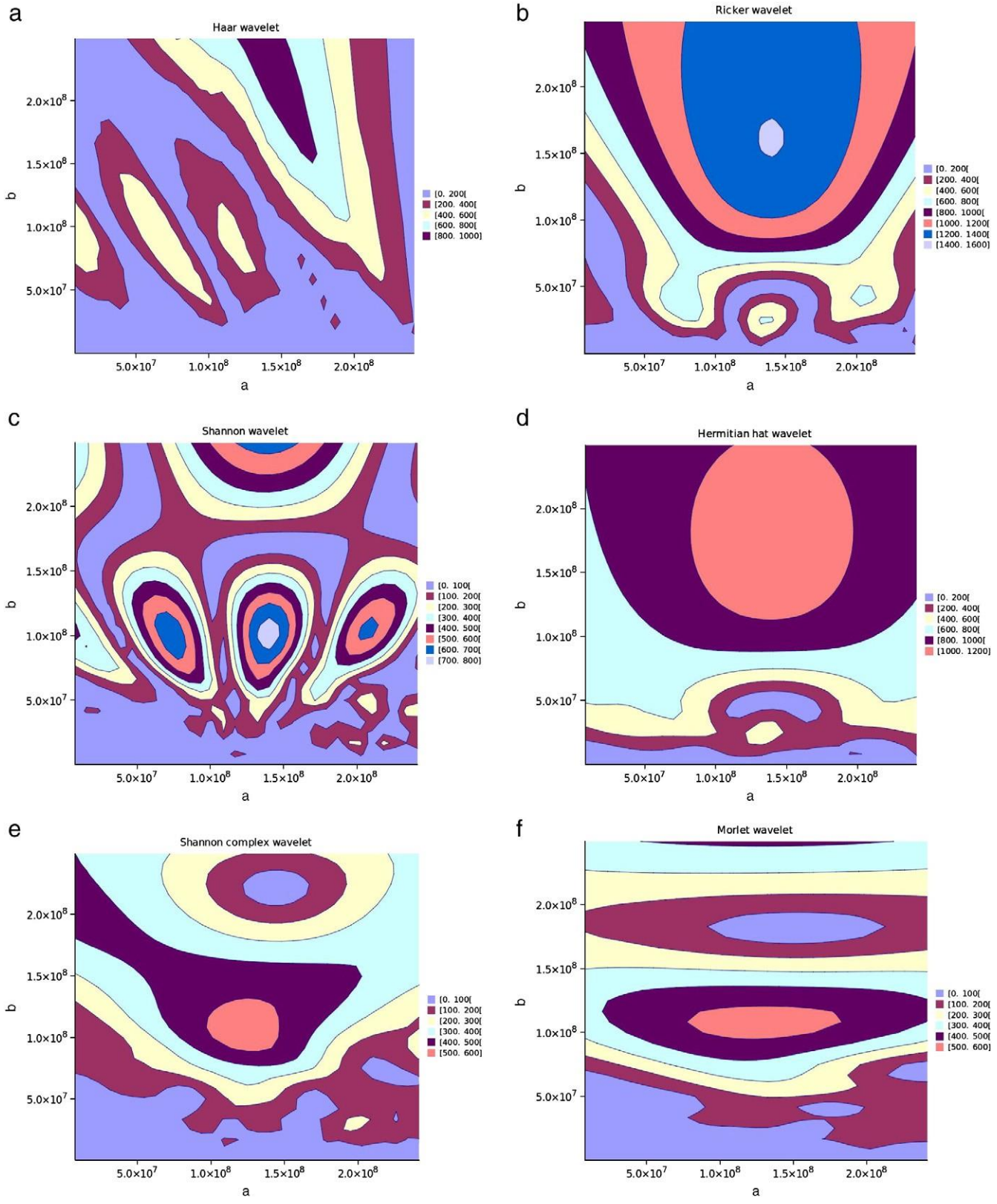

Fig. 1. Absolute values of the HW, RW, SW, HHW, SCW, and MW versus (a, b) for the human chromosome Hu1, with dt =1.

logic of different scales of resolution (e.g., chromosomes Hu17 and Hu22).

We now analyze the absolute value of the SW and the values of b that lead to the independent peaks, visible in the lower half of the chart. We choose only the most relevant peaks, leading to a collection of data representing their coordinates (a,b). Each chromosome has a different length and, consequently, distinct maximum values of a and b denoted by max(a) and max(b). Therefore, having in mind a global comparison of the results for the 24 chromosomes, the absolute coordinates (a,b) of the peaks are

converted into relative values, namely $a/\max(a)$ and $b/\max(b)$, and the histogram of relative frequency is calculated. Fig. 3 represents the relative frequency of peaks of the absolute value of the SW versus the corresponding relative coordinates $a/\max(a)$ and $b/\max(b)$ showing that usually the peaks are in most cases three, but that can go up to six, that they are almost equally spaced along the a axis, and that they occur in the range $0.3 \leq b/\max(b) \leq 0.5$.

We conclude that it is possible to compare visually the plots and to establish a qualitative grouping by similarities. Nevertheless, it is preferable to define a quantitative measure avoiding subjective

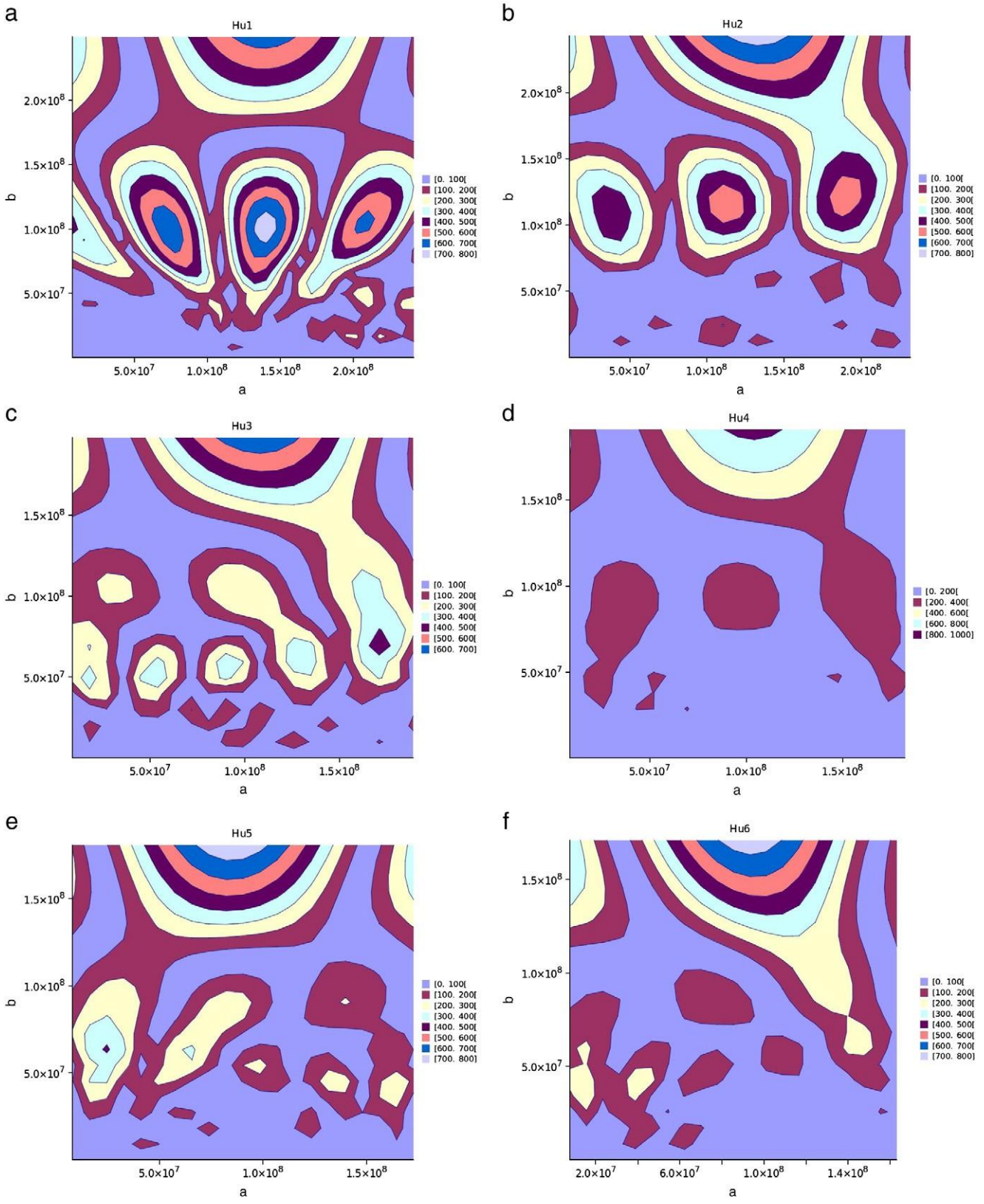

Fig. 2. Absolute value of the SW versus (a, b) for the human twenty four chromosomes {Hu1, Hu2, ..., HuX, HuY}, with dt =1.

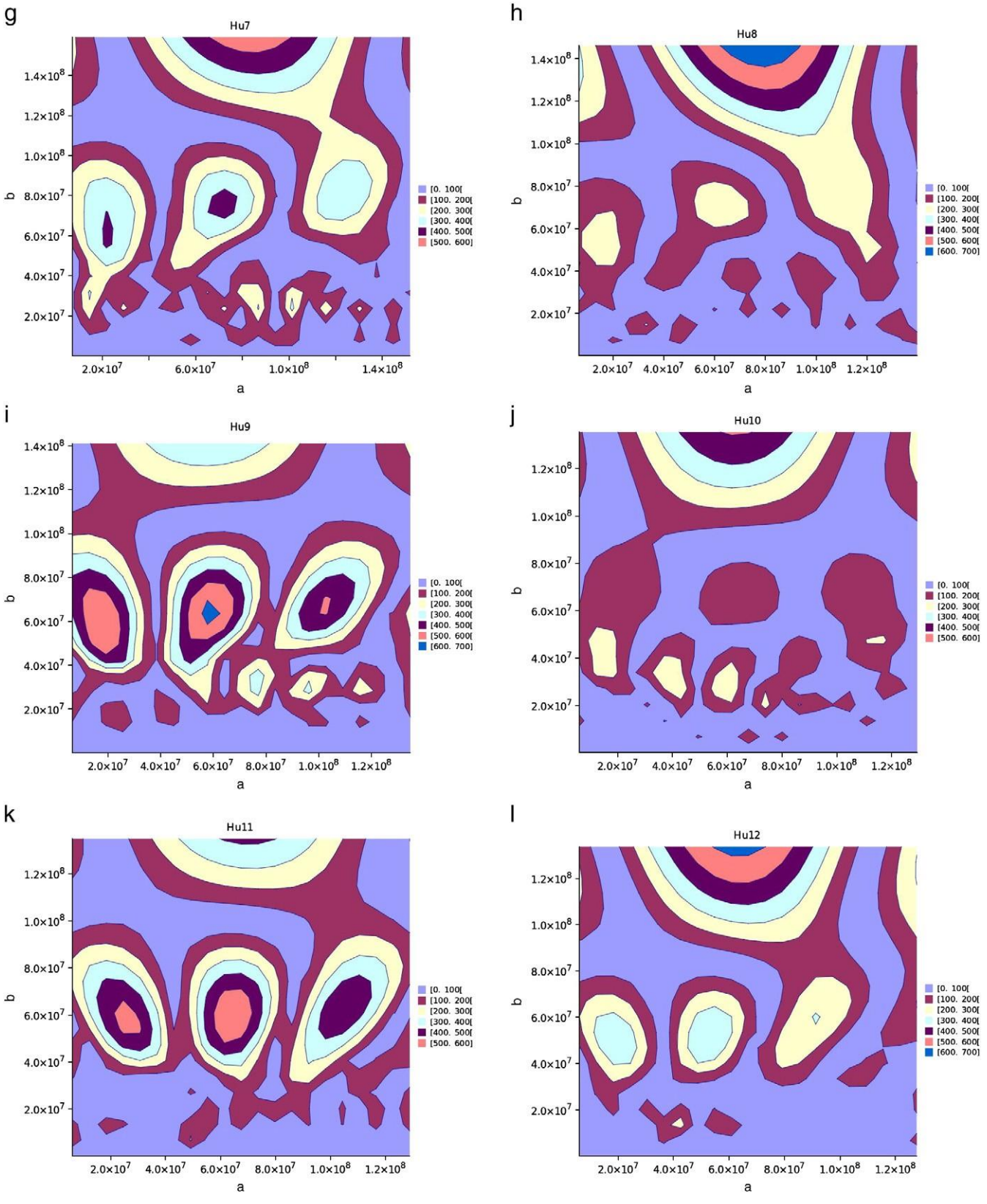

Fig. 2 (*continued*).

issues. Before continuing it should be noted that wavelet results are complex valued and that Figs. 1 and 2 represent only the absolute value. Nevertheless, this approach is frequently adopted because it is simple and produces good results. In order to overcome the effects of the different chromosome sizes, and to emphasize the shape of the wavelet plots, it was decided to normalize the charts, by

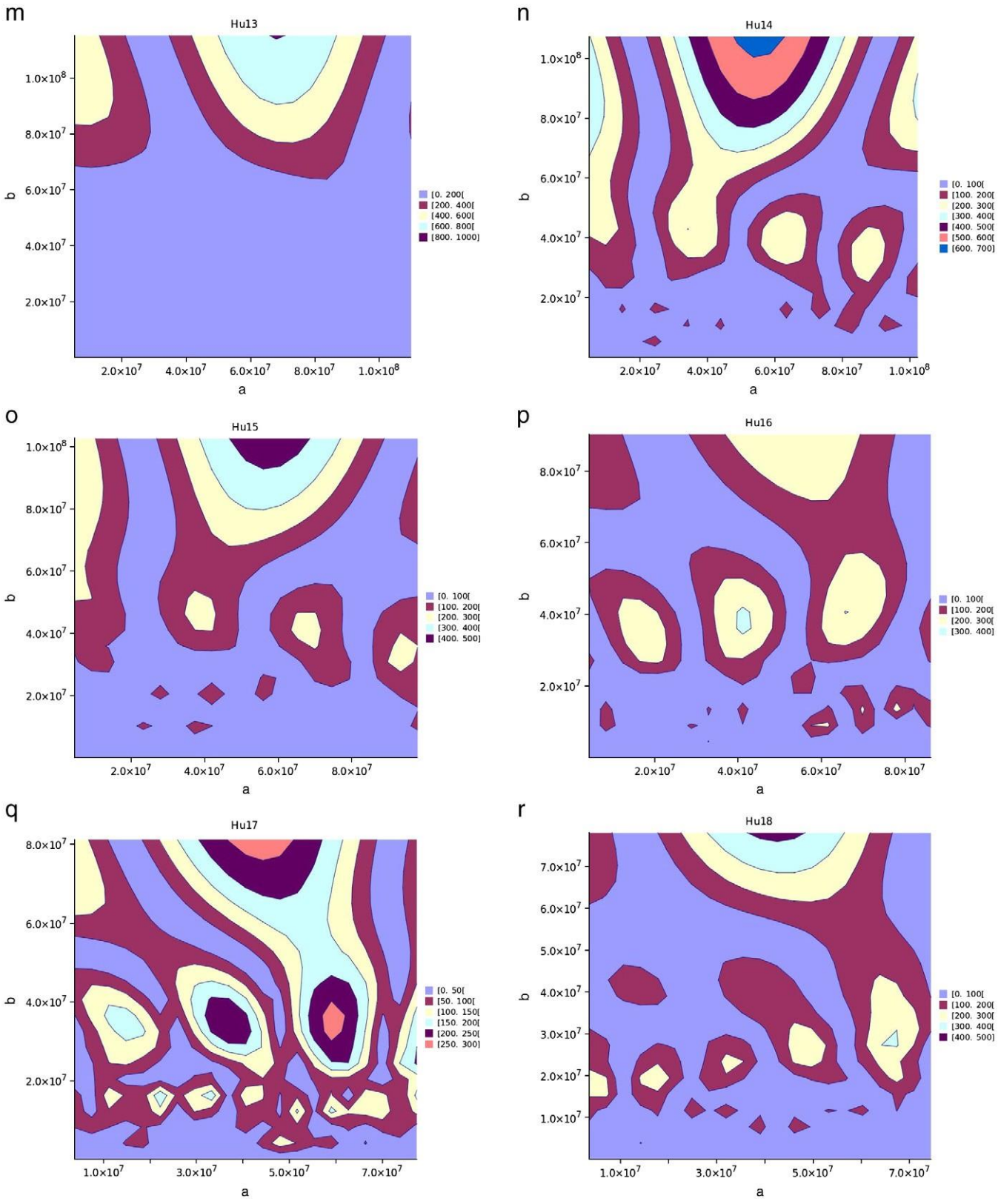

Fig. 2 (*continued*).

converting the a and b axis into the interval [0, 1] and by re-scaling the wavelet absolute values so that the total volume becomes one. In other words, for each plot is considered for the x and y scale axis the values $a/\max(a)$ and $b/\max(b)$, and for the z axis the values $\left\|[W_\psi x(t)](a,b)\right\| / \iint_{(a,b)} \left\|[W_\psi x(t)](a,b)\right\| da\, db$. Each plot can now be interpreted as a probability density function and for comparing the normalized plots it is adopted the measure:

$$r_{ij} = \sqrt{\left(\mu_{a,i}-\mu_{a,j}\right)^2 + \left(\sigma_{a,i}-\sigma_{a,j}\right)^2 + \left(\mu_{b,i}-\mu_{b,j}\right)^2 + \left(\sigma_{b,i}-\sigma_{b,j}\right)^2},\ i,j = 1,\dots,24 \quad (4)$$

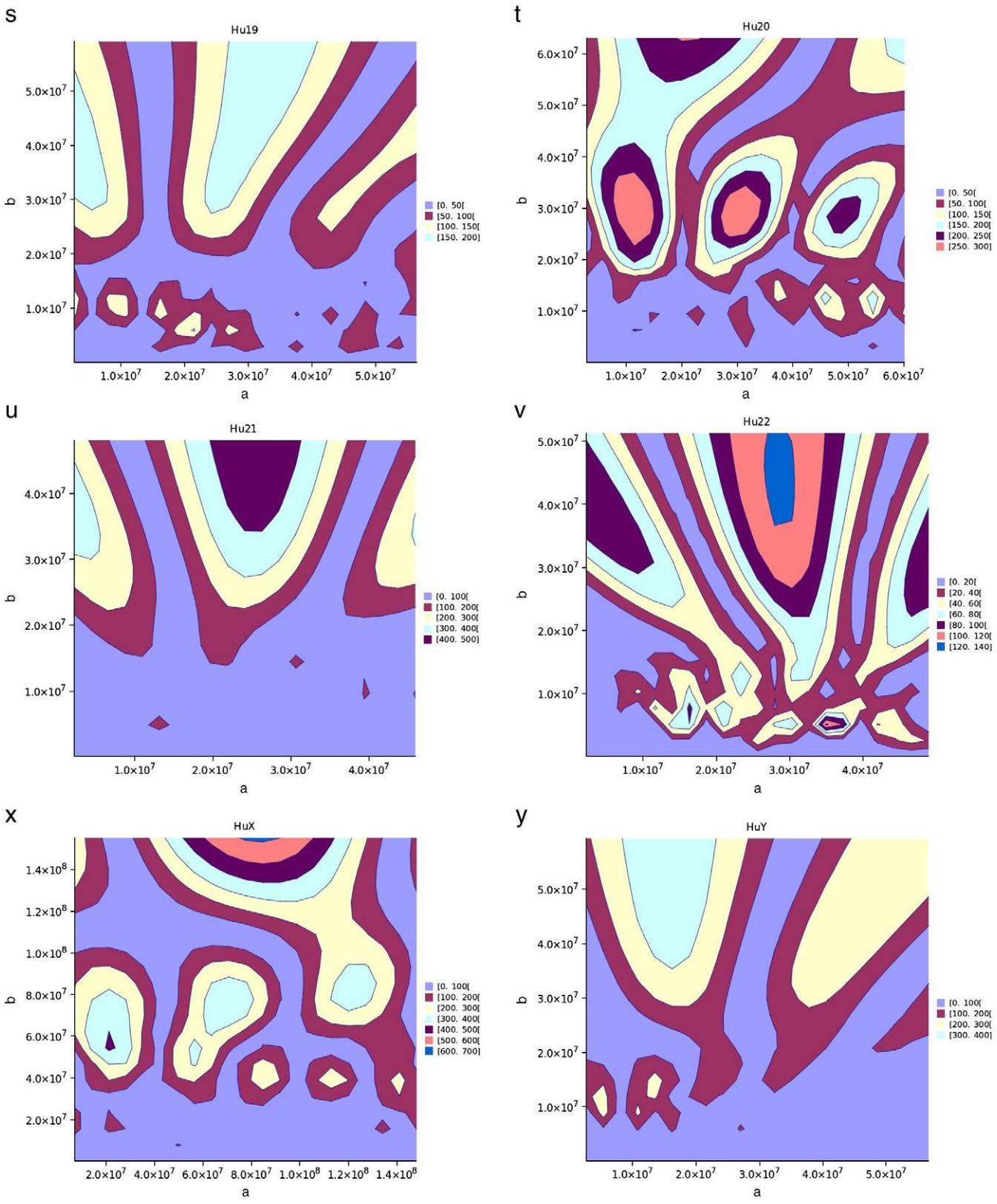

Fig. 2 (*continued*).

where the symbols μ and σ represent the arithmetic average and standard deviation, and the indices *i*,*j* list the set of chromosomes.

Based on the r_{ij} index it is now possible to calculate a $R_{24 \times 24}$ symmetric matrix of distances in the sense of Eq. (4) and to use a visualization tool for mapping the chromosome characteristics.

Fig. 4 depicts the two-dimensional map generated by the MDS tool. Usually MDS output quality is evaluated by the Sheppard and stress diagrams. The first plots the distances versus the original dissimilarities and the second plots a measure of the mapping difficulty (called stress) versus the number of dimensions in the

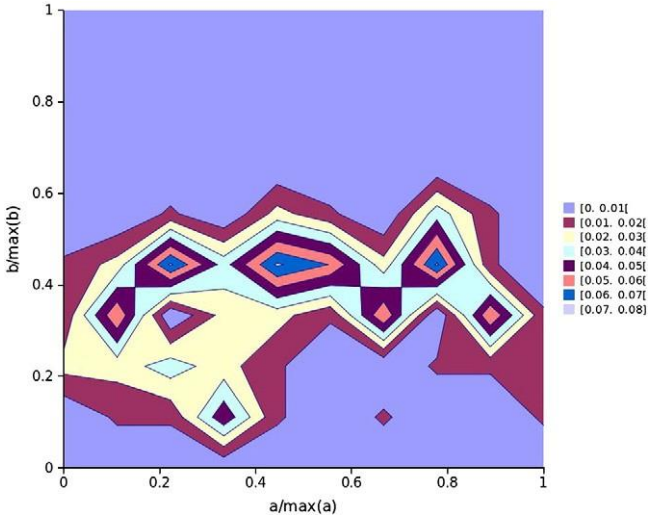

Fig. 3. Relative frequency of peaks of the absolute value of the SW for the 24 chromosomes versus (a/max(a), b/max(b)).

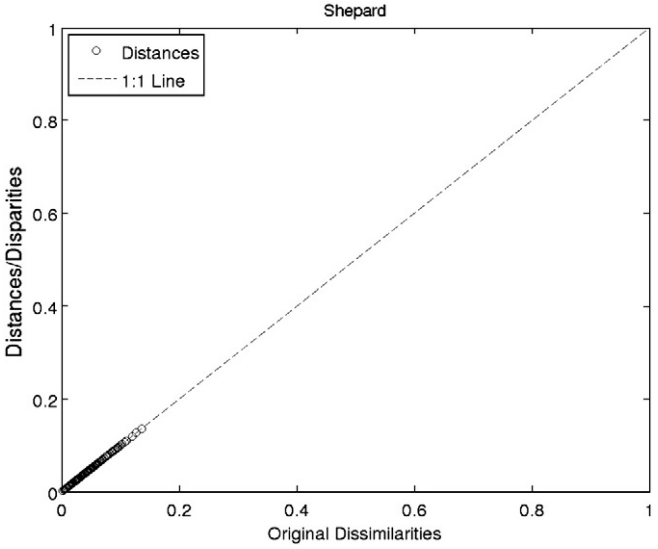

Fig. 5. Sheppard diagram of the two dimensional MDS map.

MDS representation. Obviously the closer to the 45° line, in the first case, or the lower the values, in the second case, the better is the MDS map. In this perspective, Figs. 5 and 6 show the corresponding Sheppard and stress diagrams, respectively, demonstrating a good fit of the two dimensional MDS map.

As mentioned in Section 2, another visualization tool is the dendogram. Fig. 7 depicts the dendogram plot based on the unweighted average method. Since the R matrix feeds both the MDS and the dendogram it is of no surprise that the grouping has strong resemblances.

At a different level we can discuss several aspects namely the restrictions posed by decisions such as (*i*) the conversion scheme (Eq. (1)) for obtaining a numerical signal, (*ii*) the adoption of Shannon real wavelet defined in Eq. (3c), (*iii*) the use of absolute value of the wavelet and (*iv*) the comparison of the plots by the normalization scheme followed by the r_{ij} measure defined in Eq. (4). Therefore, in spite of the comprehensible results obtained, the approach developed in this study does not preclude other strategies and motivates further research by exploring other routes.

4. Conclusions

Chromosomes have a code based on a four symbol alphabet. This information can be analyzed with tools usually adopted in signal processing. In this paper it was developed a conversion scheme for establishing a numerical signal and the resulting sequence of values was studied by means of continuous wavelets. The application to the human DNA of six different wavelets revealed that the Shannon real wavelet is the more promising one, leading to the emergence of patterns capable of being interpreted and compared. For condensing and visualizing the results a comparison index was developed and MDS and dendogram techniques were used successfully. Besides the aforementioned aspects, another merit of the overall processing is to open new research directions in pursuit of DNA decoding.

Acknowledgments

We thank the following organizations for allowing access to genome data of the Human-Genome Reference Consortium, http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/.

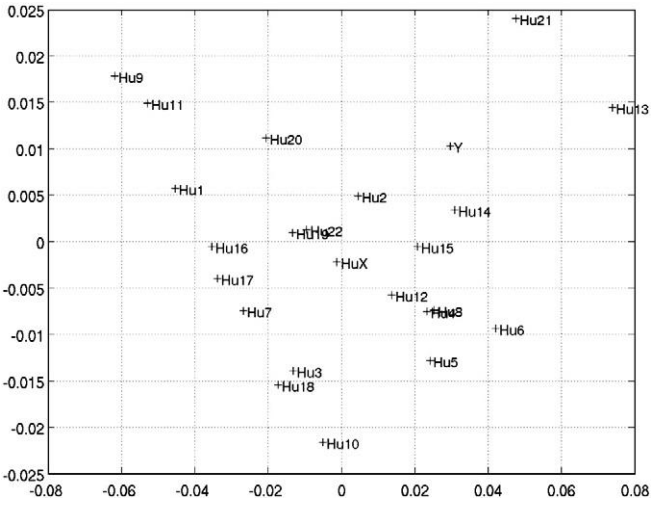

Fig. 4. Two dimensional MDS map of the human twenty four chromosomes {Hu1, Hu2, ..., HuX, HuY} in the perspective of the SW and the r_{ij} index.

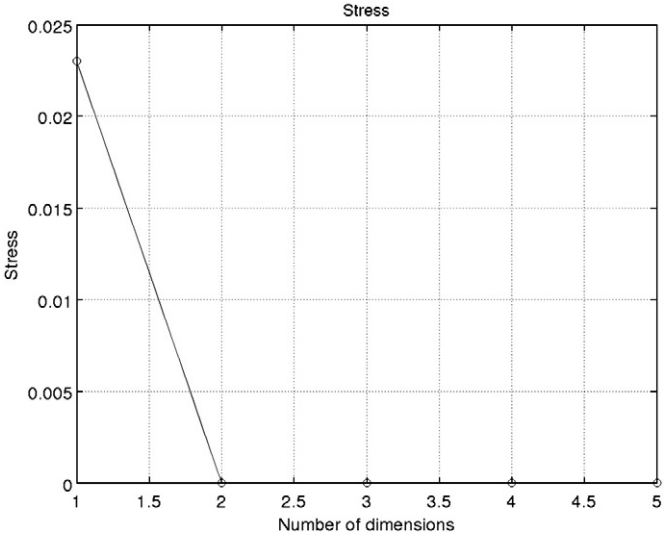

Fig. 6. Stress versus number of plotting dimensions of the MDS map.

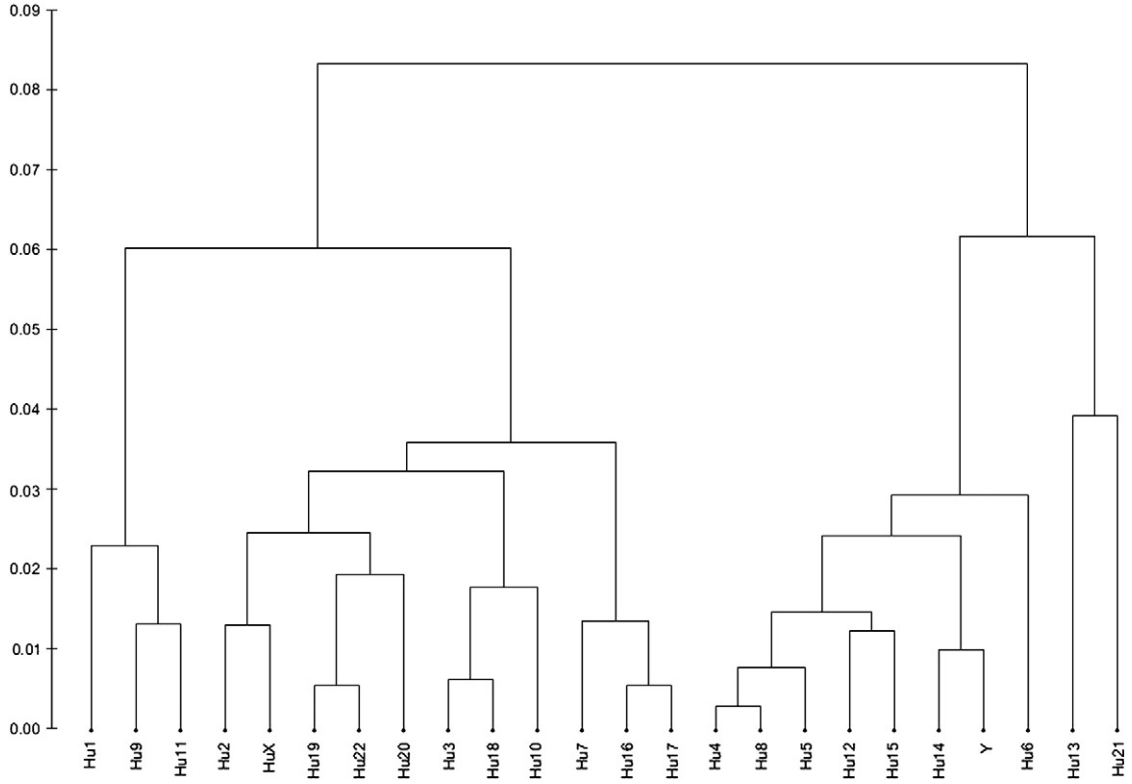

Fig. 7. Dendogram of the human twenty four chromosomes {Hu1, Hu2, ..., HuX, HuY} in the perspective of the SW and the r_{ij} index.

References

[1] R.T. Schuh, A.V.Z. Brower, Biological Systematics: Principles and Applications, 2nd edition Cornell University Press, 2009.

[2] Harald Seitz (Ed.), Analytics of Protein-DNA Interactions, Advances in Biochemical Engineering Biotechnology, Springer, 2007.

[3] H. Pearson, Genetics: what is a gene? Nature 441 (7092) (2006) 398-401.

[4] UCSC Genome Bioinformatics, http://hgdownload.cse.ucsc.edu/downloads.html.

[5] Gregory E. Sims, Se-Ran Jun, Guohong A. Wu, Sung-Hou Kim, Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions, Proceedings of the National Academy of Sciences of the United States of America 106 (8) (2009) 2677-2682.

[6] William J. Murphy, Thomas H. Pringle, Tess A. Crider, Mark S. Springer, Webb Miller, Using genomic data to unravel the root of the placental mammal phylogeny, Genome Research 17 (2007) 413-421.

[7] Hao Zhao, Guillaume Bourque, Recovering genome rearrangements in the mammalian phylogeny, Genome Research 19 (2009) 934-942.

[8] Arjun B. Prasad, Marc W. Allard, Confirming the phylogeny of mammals by use of large comparative sequence data sets, Molecular Biology and Evolution 25 (9) (2008) 1795-1808.

[9] Ingo Ebersberger, Petra Galgoczy, Stefan Taudien, Simone Taenzer, Matthias Platzer, Arndt von Haeseler, Mapping human genetic ancestry, Molecular Biology and Evolution 24 (10) (2007) 2266-2276.

[10] Casey W. Dunn, et al., Broad phylogenomic sampling improves resolution of the animal tree of life, Nature 452 (2008) 745-750.

[11] J.A. Tenreiro Machado, António C. Costa, Maria Dulce Quelhas, Fractional dynamics in DNA, Communications in Nonlinear Science and Numerical Simulations 16 (8) (2011) 2963-2969.

[12] António C. Costa, J.A. Tenreiro Machado, Maria Dulce Quelhas, Histogram-based DNA Analysis for the Visualization of Chromosome, Genome and Species Information, Bioinformatics, 27(9), Oxford University Press, 2011, pp. 1207-1214.

[13] J.A. Tenreiro Machado, António C. Costa, Maria Dulce Quelhas, Entropy Analysis of DNA Code Dynamics in Human Chromosomes, Computers and Mathematics with Applications, Elsevier (Available online 25 March 2011). doi:10.1016/j.camwa.2011.03.005.

[14] W.S. Torgerson, Theory and Methods of Scaling, Wiley, New York, 1958.

[15] R.N. Shepard, The analysis of proximities: multidimensional scaling with an unknown distance function, Psychometrika 27 (I and II) (1962) 219-246 and 219-246.

[16] J. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1) (1964) 1-27.

[17] J.B. Kruskal, M. Wish, Multidimensional Scaling, Sage Publications, Newbury Park, 1978.

[18] W.L. Martinez, A.R. Martinez, Exploratory Data Analysis with MATLAB, Chapman & Hall/CRC, Boca Raton, 2005.

[19] S. Gómez Fernández, Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms, Journal of Classification 25 (1) (2008) 43-65, doi:10.1007/s00357-008-9004-x.

[20] Javier Arsuaga, Robert K.-Z. Tan, Vazquez Mariel, De Witt Sumners, Stephen C. Harvey, Investigation of viral DNA packaging using molecular mechanics models, Biophysical Chemistry 101-102 (2002) 475-484.

[21] N. Kovaleva, L. Manevich, Localized nonlinear oscillation of DNA molecule, 8th Conference on Dinamical Systems Theory and Applications, December 12-15, 2005, Lodz, Poland, 2005.

[22] N. Kovaleva, L. Manevich, V. Smirnov, Analitical study of coarse-grained model of DNA, 9th Conference on Dinamical Systems Theory and Applications, December 17-20, 2007, Lodz, Poland, 2007.

[23] Katica R. (Stevanovic) Hedrih, Andjelka N. Hedrih, Eigen modes of the double DNA chain helix vibrations, Journal of Theoretical and Applied Mechanics (JTAM-Poland) 48 (1) (2010) 219-231.

[24] C. Frontali, E. Dore, A. Ferrauto, E. Gratton, A. Bettini, M.R. Pozzan, E. Valdevit, An absolute method for the determination of the persistence length of native DNA from electron micrographs, Biopolymers 18 (1979) 1353-1357.

[25] Gilbert G. Walter, Wavelets and Other Orthogonal Systems, Second Edition CRC, 2000.

[26] Stéphane Jaffard, Yves Meyer, Robert D. Ryan, Wavelets: Tools for Science & Technology, SIAM, 2001.

[27] Hans-Georg Stark, Wavelets and Signal Processing: An Application-Based Introduction, Springer, 2005.