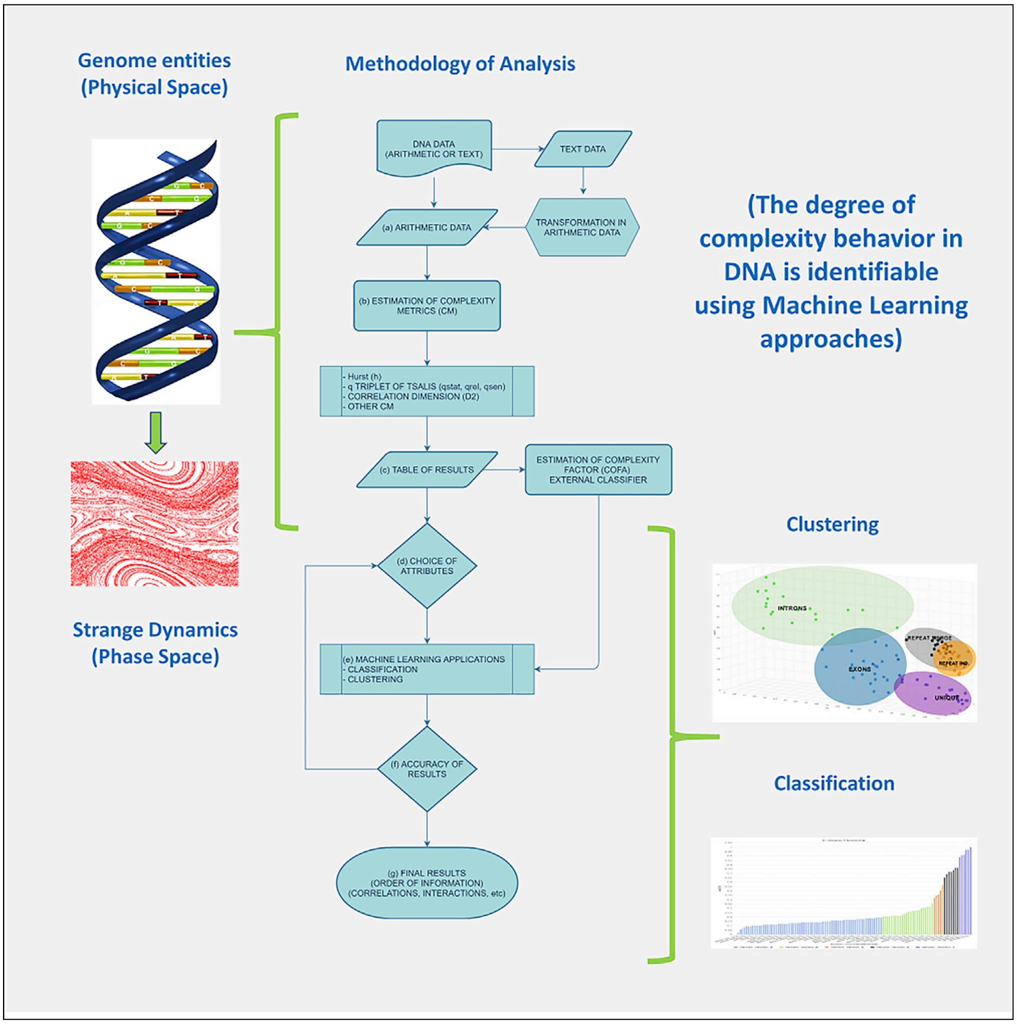
iScience

CellPress
OPEN ACCESS

Article

Spatial constrains and information content of sub-genomic regions of the human genome


Genome entities
(Physical Space)
Methodology of Analysis
DNA DATA (ARITHMETIC OR TEXT)
TEXT DATA
(a) ARITHMETIC DATA
TRANSFORMATION IN ARITHMETIC DATA
(b) ESTIMATION OF COMPLEXITY METRICS (CM)
- Hurst (h)
- q TRIPLET OF TSALIS (qstat, qrel, qsen)
- CORRELATION DIMENSION (D2)
- OTHER CM
(c) TABLE OF RESULTS
ESTIMATION OF COMPLEXITY FACTOR (COFA) EXTERNAL CLASSIFIER
(d) CHOICE OF ATTRIBUTES
(e) MACHINE LEARNING APPLICATIONS
- CLASSIFICATION
- CLUSTERING
(f) ACCURACY OF RESULTS
(g) FINAL RESULTS (ORDER OF INFORMATION) (CORRELATIONS, INTERACTIONS, etc)
Strange Dynamics
(Phase Space)
(The degree of complexity behavior in DNA is identifiable using Machine Learning approaches)
Clustering
INTRONS
EXONS
REPEAT MODE
REPEAT IND.
UNIQUE
Classification

Leonidas P. Karakatsanis, Evgenios G. Pavlos, George Tsoulouhas, ..., Jamie L. Duke, George P. Pavlos, Dimitri S. Monos

karaka@env.duth.gr (L.P.K.)
monosd@chop.edu (D.S.M.)

HIGHLIGHTS

The lengths of DNA subgenomic entities satisfied the Tsallis non-extensive statistics

The size distribution of the subgenomic entities within chromosomes follow specific patterns

A technical index COFA was introduced to characterize the degree of complexity

The degree of complexity behavior in DNA is identifiable using ML approaches

Karakatsanis et al., iScience *24*, 102048
February 19, 2021 © 2021 The Authors.
https://doi.org/10.1016/j.isci.2021.102048

iScience

CellPress
OPEN ACCESS

Article

Spatial constrains and information content of sub-genomic regions of the human genome

Leonidas P. Karakatsanis,[1,4,*] Evgenios G. Pavlos,[1,2] George Tsoulouhas,[1] Georgios L. Stamokostas,[1] Timothy Mosbruger,[3] Jamie L. Duke,[3] George P. Pavlos,[1] and Dimitri S. Monos[3,*]

SUMMARY

Complexity metrics and machine learning (ML) models have been utilized to analyze the lengths of segmental genomic entities of DNA sequences (exonic, intronic, intergenic, repeat, unique) with the purpose to ask questions regarding the segmental organization of the human genome within the size distribution of these sequences. For this we developed an integrated methodology that is based upon the reconstructed phase space theorem, the non-extensive statistical theory of Tsallis, ML techniques, and a technical index, integrating the generated information, which we introduce and named complexity factor (COFA). Our analysis revealed that the size distribution of the genomic regions within chromosomes are not random but follow patterns with characteristic features that have been seen through its complexity character, and it is part of the dynamics of the whole genome. Finally, this picture of dynamics in DNA is recognized using ML tools for clustering, classification, and prediction with high accuracy.

INTRODUCTION

The DNA structure in the human genome reflects the entire evolutionary process from simple to highly complex biological forms and organisms. Complexity theory indicates the existence of a strange and self-organizing dynamic process underlying the biological evolution process. As we have shown, in two previous studies (Pavlos et al., 2015; Karakatsanis et al., 2018) concerning the DNA sequence of the major histocompatibility complex (MHC), all DNA sequences of sub-genomic regions (exons, introns, intergenic) have structure and contain information. The DNA base sequence is constructed by nature as a long-range correlated self-organized system and emergent biological form through the co-evolution of biological and environmental subsystems. From mathematical point of view, nature realizes complex mathematical forms with spatiotemporal correlations. The non-linear and strange dynamics describes the evolution of complex systems such as biological systems as a non-linear complex process including critical states and critical points, where the system can develop ordered states and forms throughout the development of long-range spatiotemporal correlations. This mathematical behavior of nature is self-consistently described by the non-equilibrium thermodynamics and the non-extensive statistical theory of Tsallis. Nature works thermodynamically for the development of non-equilibrium stationary thermodynamic states where the entropy function is maximized (Prigogine, 1978; Nicolis and Prigogine, 1989; Nicolis, 1993; Tsallis, 2009). The development of the complexity theory (Prigogine, 1978, 1997; Nicolis and Prigogine, 1989; Nicolis, 1993; Tsallis, 2009), through the information theory can describe the redundancy of information in DNA. According to the classical biological description, only 1.5% of the human DNA is translated into proteins. The rest was traditionally of unknown significance and thought as non-essential (‘‘junk’’). To determine the role of the remaining part of the genome, many tries have been made, with the most notable one being the Encyclopedia of DNA elements project (Davis et al., 2017). To shed light on the problem from a different perspective, a significant increase in novel interdisciplinary approaches and methods were developed. More specifically for the last 30 years, or so, the complex character of biological systems, such as the order of information in genome, the origins of autoimmune diseases, etc, have been studied with the intent to shed light on DNA’s internal organization. Many researchers have developed computational methods to identify and characterize DNA motifs throughout the genome utilizing methods borrowed from the field of signal processing, information theory, non-linear dynamics, and the non-extensive statistics-based methods (Broomhead and King, 1986; Tsallis, 1988, 2002, 2004; Casdagli, 1989; Theiler, 1990; Grassberger

[1] Department of Environmental Engineering, Complexity Research Team (CRT), Democritus University of Thrace, 67100 Xanthi, Greece

[2] Department of Basic Sciences, School of Medicine, University of Crete, Heraklion, Crete 71003, Greece

[3] Department of Pathology and Laboratory Medicine, The Children’s Hospital of Philadelphia and Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[4] Lead Contact

*Correspondence: karaka@env.duth.gr (L.P.K.), monosd@chop.edu (D.S.M.)

https://doi.org/10.1016/j.isci.2021.102048

et al., 1991; Peng et al., 1992; Provenzale et al., 1992; Lorentz, 1993; Klimontovich, 1994; Provata and Beck, 2011; Kellis et al., 2014; Wu, 2014). These statistical metrics can be used to describe the dynamic characteristics and the structure and organization of the human genome. Many scientists (Voss, 1992; Li and Kaneko, 1992; Buldyrev et al., 1993, 1995; Grosberg et al., 1993; Ossadnik et al., 1994; Stanley et al., 1994) have introduced and studied the DNA random walk process as a basic physical process-model for the detection and understanding of the observed long-range correlations of nucleotides in DNA sequences. According to Voss (1992) there is no significant difference between long-range correlation properties of coding and non-coding sequences, and in contrast, other scientists (Peng et al., 1992; Li and Kaneko, 1992; Buldyrev et al., 1993, 1995; Grosberg et al., 1993) found that the non-coding sequences are characterized by long-range correlations, whereas coding sequences are not. This could mean that the dynamics that produced the spatial information of the DNA can be characterized by strange dynamics such as strange attractors, islands, and multifractal behavior in the reconstructed phase space. Some significant theories from the account of statistical physics, like self-organized criticality (SOC), strange dynamics, non-extensive statistical mechanics of Tsallis, fractional dynamics, etc., have been proposed to interpret the development of long-range correlations of the DNA sequences.

However more recently, thanks to advanced DNA sequencing (cumulatively named next-generation sequencing) technologies, a large-scale sequencing information has become available enabling and advancing the research that utilizes these computational and statistical methods to identify the principles that define DNA's internal structural characteristics (Oikonomou and Provata, 2006; Vinga and Almeida, 2007; Oikonomou et al., 2008; Kellis et al., 2014; Pavlos et al., 2015; Namazi and Kiminezhadmalaie, 2015; Woods et al., 2016; Karakatsanis et al., 2018).

More specifically, Melnik and Usatenko (2014), using an additive Markov chain approach, analyzed DNA molecules of different organisms, and they estimated the differential entropy for the biological classification of these organisms. Similarly, Papapetrou and Kugiumtzis (2014, 2020) studied DNA sequences, through the estimation of the Markov chain orders and Tsallis conditional mutual information. The results showed a different long memory structure in their DNA samples (coding and non-coding). In another study, Provata et al. (2014a, 2014b) analyzed the evolutionary tree of higher eukaryotes, amebae, unicellular eukaryotes, and bacteria with complexity tools to estimate the conditional probability, the fluxes, the block entropy, and the exit distance distributions. The study detected the changes in the statistical and complexity measures of the five organisms and proposed these measures as alternative methods for organism classification. Wu (2014) studied the *Synechocystis sp*. PCC6803 genome by using the recurrence plot method and the technique of phase space reconstruction. This analysis revealed periodic and non-periodic correlation structures in the DNA sequences. Costa et al. (2019), Machado (2019), and Silva et al. (2020) used tools from Kaniadakis statistics, power law distribution, and fractal and information theory to uncover the order information of the *Homo sapiens* DNA chromosomes.

There is additional literature related to the complexity metrics used in this study and other proposed complexity metrics that utilize these theoretical/statistical tools to address questions in the realm of genomics. Specifically, Corona-Ruiz et al. (2019) presented an analysis of the mitochondrial DNA of 32 species in the subphylum Vertebrata, divided in seven taxonomic classes, using stochastic parameters, like the Hurst and detrended fluctuation analysis exponents, Shannon entropy, and Chargaff ratio. Namazi et al. (2016) used fractal dimension to study the influence of changes in DNA (DNA mutation) on human characteristics and features. Liu et al. (2020) analyzed promoter sequences by calculating the information content of the sequences and the correlation between sequences in the subregion and other sequence features as supplements, such as the Hurst exponent, GC content, and sequence bending property. Li et al. (2019) analyzed exon and intron DNA sequences based on topological entropy calculation, genomic signal processing method, and singular value decomposition to explore the complexity of DNA sequences and its functional elements. Hsu et al. (2017) proposed a measure of complexity, called entropy of entropy analysis, useful for DNA sequences compared with Shannon entropy and application to the cardiac interbeat interval time series. Thanos et al. (2018) studied the local Shannon entropy in blocks as a complexity measure to study the information fluctuations along DNA sequences. Finally, Anitas (2020) analyzed DNA sequences based on Chaos game representation followed by a multifractal analysis studying the corresponding scaling properties.

The produced data from these large-scale DNA sequencing efforts have provided researchers the opportunity to develop additional approaches like machine learning (ML) techniques to analyze these sequences with models for clustering, pattern recognition, classification, and prediction with supervised and unsupervised learning (Manogaran et al., 2018; Apostolou et al., 2019; Washburn et al., 2019; Varma et al., 2019; Frey et al., 2019). The use of methods based on statistics and ML algorithms have many common and also separate routes with respective disadvantages and advantages. The limit between statistics and ML is often not visible (Bzdok et al., 2018; Xu and Jackson, 2019). For a review of the ML applications in genetics and genomics see Libbrecht and Noble (2015). The goal of all these methods of analysis is the deep understanding of DNA organization and therefore of the biological systems.

In this work we expand upon our previous studies (Pavlos et al., 2015; Karakatsanis et al., 2018), where we measured the dynamical and the non-extensive statistical characteristics in the DNA sequence of the whole MHC as a single unit and in the exonic, intronic, and intergenic sequences of the MHC as separate and independent entities. We seek to identify order information included in the whole genome and their possible interactive relationships focused in regions such as exons versus introns and repeat versus unique sequences. Our analyses are based upon the reconstructed phase space theorem (Takens, 1981; Theiler, 1990), the non-extensive statistical theory of Tsallis (Tsallis, 1988, 2004, 2009), and ML techniques, and we introduce a technical index, which we call complexity factor (COFA), as a more suitable one to our analysis.

The selected parameter of length for each of these genomic sub-regions was chosen because the overall intent of the study was to identify the potential relationships and order/information concealed within the spatial organization of each chromosome. Questions like possible relationships between and among the sizes of exonic/intronic, genic/intergenic, or repeat/unique regions, whereby occasionally overlapping sequences have different functions among the different chromosomes, are simple and fundamental questions that have never before been comprehensively addressed. Until recently the detailed and massive genomic data for the whole genome was not available and the computational and statistical tools were not fully developed. Their availability now enables our community to ask these questions and hopefully identify answers that at some future point can be confirmed experimentally; eventually a more thorough and comprehensive characterization of the human genome and its interactions will emerge.

Instructive literature to familiarize the readership of this article with basic concepts of complex systems would be the following books and articles: On Complexity—Self Organization (Nicolis and Prigogine, 1989; Bak, 2013), on Strange Attractors (Grebogi et al., 1987; Ben-Mizrachi et al., 1984; Grassberger and Procaccia, 1983), on Correlation Dimension (Grassberger and Procaccia, 2004; Argyris et al., 1998), on Multifractality (Stanley and Meakin, 1988), and on Non-extensive Statistical Theory (Tsallis, 2009).

RESULTS

The Genomic compartments we used in this study and the Gene definitions are taken from National Center for Biotechnology Information (NCBI). This database provides both, the gene and exon definitions. Based on these definitions we generated the intronic and intergenic region coordinates. For the repeat individual we used the Repeat Masker. We then merged the repeat individual to generate the repeat merge data. Coordinates for the non-repeat sequences were complementary to merged repeat sequences. Using both the curated and derived definitions we generated the data below.

Figures 1A–1C shows the set of data, which includes seven regions per chromosome (genic, intergenic, exonic, intronic, repeat individual, repeat merge, unique). The dataset equates 7 regions by 22 chromosomes = 154 raw data. Finally, we analyzed 5 regions (exonic, introns, repeat individual, repeat merge, unique) by 22 chromosomes = 110 raw data. The genic and intergenic regions do not have enough number of points to satisfy the statistics in the reconstruction state space, therefore they were not included in our analysis.

It is noteworthy to mention that the data in Figures 1A and 1C if combined reveal a proportional relationship between the number of exonic and genic regions, which is constant (see Table 1) in all chromosomes. Expectedly, the same is observed to couples exonic/intergenic, intronic/genic, and intronic/intergenic due to numeric relationships among these genomic fragments. The values of the fraction $\left(\frac{Number\ of\ exonic\ regions}{Number\ of\ genic\ regions}\right)$ per chromosome (Figure 1D) present a remarkable stability with average value 10.79 ± 1.01, where the linear fittings have very small deviations from chromosome to chromosome. It appears that this universal ratio is a deep structural symmetry reflecting the internal organization of the

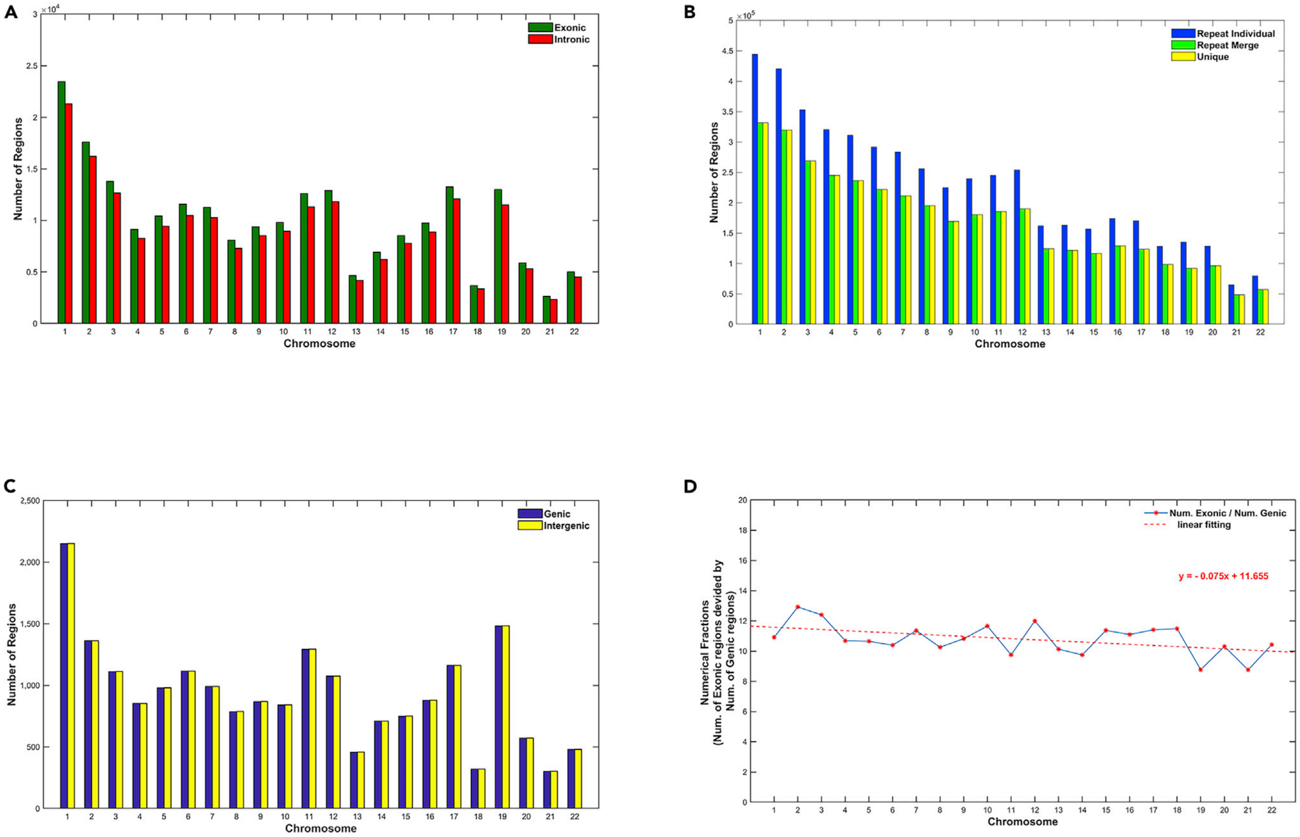


Figure 1. The set of data for the analysis in 22 Chromosomes and 7 regions

(A–C) (A) Exonic, Intronic; (B) Repeats, Unique; (C) Genic, Intergenic.

(D) Numerical fractions of regions in exonic/genic

genome, defining both spatial and functional relationships between the number of exonic and the number of genic regions.

In Figure 2, we present a sample of raw data (space series) for all regions from Chromosome 6. For example, each point on axis *x* corresponds to the i^{th} exon, intron, etc., whereas each point on axis *y* corresponds to the length of i^{th} exon, intron, etc.

In Figure 3 the general flowchart of the analysis method of DNA data (arithmetic or text) is shown. This general method can be an alternative view of the DNA entities in the entire genome based on the strange dynamics with the goal of revealing new symmetries and rules on this information.

Metrics of complexity theory

In this section the results from the estimation of complexity metrics in the distribution of the entities of genome are presented.

Hurst exponent

The Hurst exponent was estimated, for all genomic entities and for each chromosome. Figure 4 presents the estimated values of the Hurst exponent. The dashed line at 0.5 corresponds to a normal diffusion random walk process. As can be observed in Figure 4A, the Hurst exponent for intron data are much higher than 0.5 for all chromosomes and related with persistent (super-diffusion) random walk process. For the exon data, the Hurst exponent is higher than 0.5 and related with persistent (super-diffusion) random walk process for all chromosomes except Chromosome 13, which is lower than 0.5 and related with anti-persistent (sub-diffusion) random walk process. These findings reflect the degree of the multifractal character and the existence of different scaling along the distribution of the DNA entities. For Chromosomes 4,

Table 1. Numerical fractions

Chromosome	Exonic/genic	Intronic/genic	Exonic/intergenic	Intronic/intergenic
1	10.91	9.91	10.90	9.90
2	12.92	11.92	12.91	11.91
3	12.40	11.40	12.39	11.39
4	10.69	9.69	10.68	9.68
5	10.65	9.65	10.64	9.64
6	10.39	9.39	10.38	9.39
7	11.36	10.36	11.35	10.35
8	10.26	9.26	10.25	9.25
9	10.82	9.82	10.80	9.81
10	11.66	10.66	11.64	10.64
11	9.75	8.75	9.75	8.75
12	12.00	11.00	11.99	10.99
13	10.13	9.13	10.11	9.11
14	9.75	8.75	9.74	8.74
15	11.37	10.37	11.36	10.36
16	11.10	10.10	11.09	10.09
17	11.41	10.41	11.40	10.40
18	11.49	10.49	11.46	10.46
19	8.77	7.77	8.76	7.76
20	10.31	9.31	10.29	9.29
21	8.76	7.76	8.73	7.73
22	10.43	9.43	10.41	9.41
Average values	10.79 ± 1.01	9.79 ± 1.01	10.77 ± 1.01	9.78 ± 1.01

15, and 18, the Hurst exponent is almost equal to 0.5 and related with normal diffusion random walk process. This means that the profile is mono-fractal, and does not permit the existence of different scaling in the data. Similarly, in Figure 4B, for the Repeat Individual, Repeat Merge and Unique data, the Hurst exponent is much higher than 0.5 for all chromosomes and related with persistent (super-diffusion) random walk process.

q-triplet of Tsallis statistics

In Figures 5, 6, and 7, we present the estimation of Tsallis *q*-triplet for all genomic entities and for all chromosomes. Specifically, in Figure 5 we present the estimation of q_{stat} index; in Figure 6, the estimation of q_{rel} index; and finally in Figure 7, the estimation of q_{sen} index.

Concerning the q_{stat} index (Figure 5), as one can see, the value in all chromosomes in all genomic entities is higher than 1 and suggests the presence of long-range correlations, a distinctive property of open non-equilibrium systems, with underlying dynamics characterized by non-Gaussian (*q*-Gaussian) distributions. The variations of the Tsallis q_{stat} along the sizes of DNA entities is the quantitative manifestation of the biological evolution process throughout the constructive scenario of critical DNA turbulent phase transition processes. The development of long-range correlations means that the sizes of regions that are furthest between them are governed by fundamental rules on their size. Specifically, for the exonic genomic entity the index takes values mainly between 1 and 1.5, while for intronic one take values between 1.5 and 3. This means that the non-extensive character of the dynamics is much higher in introns than the exons and presents stronger long-range correlations in introns. Moreover, we observe a significant differentiation of the q_{stat} index between chromosomes in both exonic and intronic genomic entities, which means a significant differentiation of the non-extensive character of the dynamics between chromosomes in the same genomic entity (Figure 5A). Similarly in Figure 5B, the value of q_{stat} index is higher for the repeat genomic entities than the unique one, which means that in the repeat the non-extensivity is higher than the unique region. Furthermore, among chromosomes it is observed that the q stationary index of the unique genomic

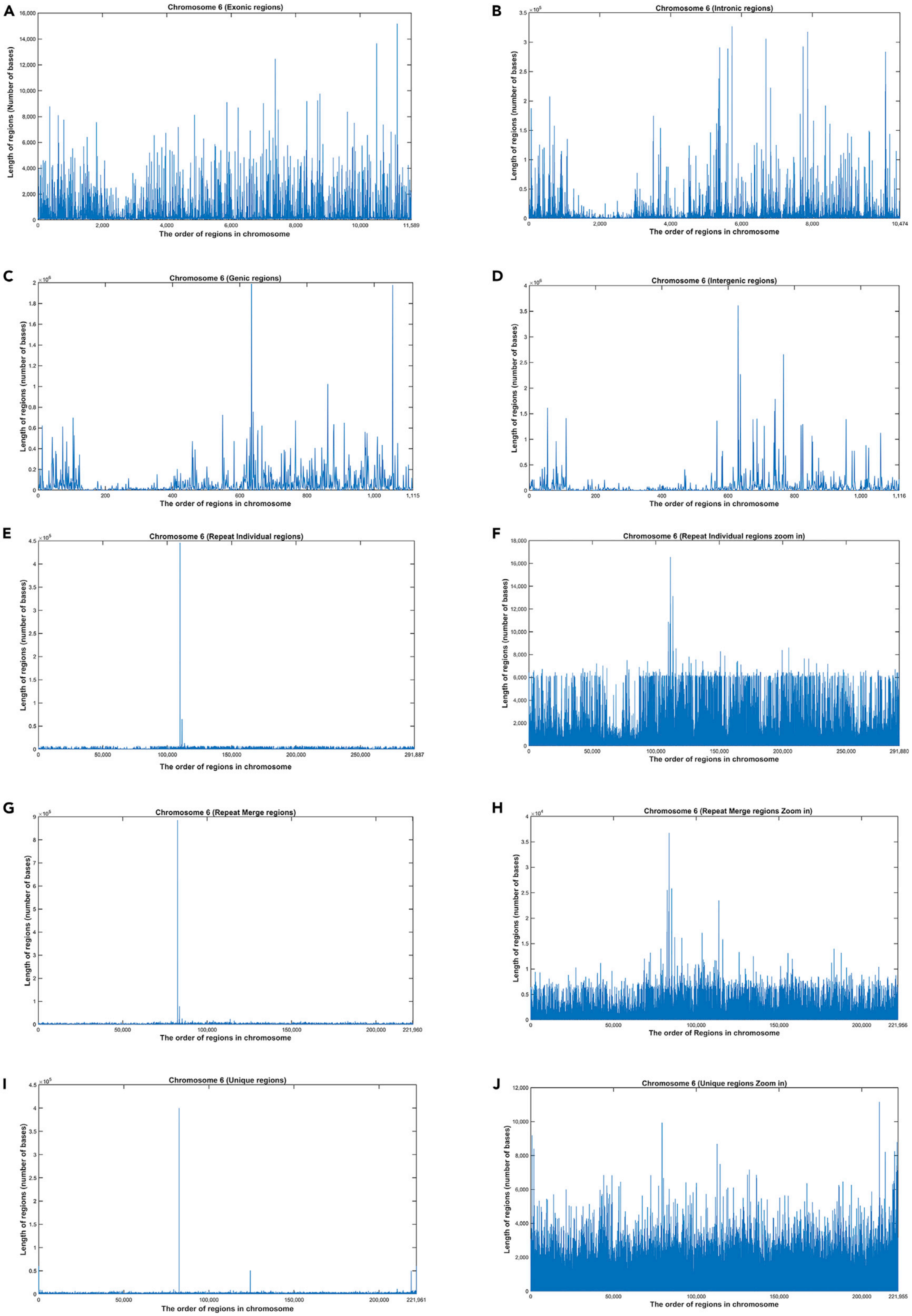


Figure 2. A sample of raw data to all regions from Chromosome 6

(A–J) (A) exonic; (B) intronic; (C) genic; (D) intergenic; (E) repeat individual; (F) repeat individual (zoom in); (G) repeat merge; (H) repeat merge (zoom in); (I) unique; (J) unique (zoom in). We see clearly here that the lengths of regions have a fractal shape, indicating a complex behavior.

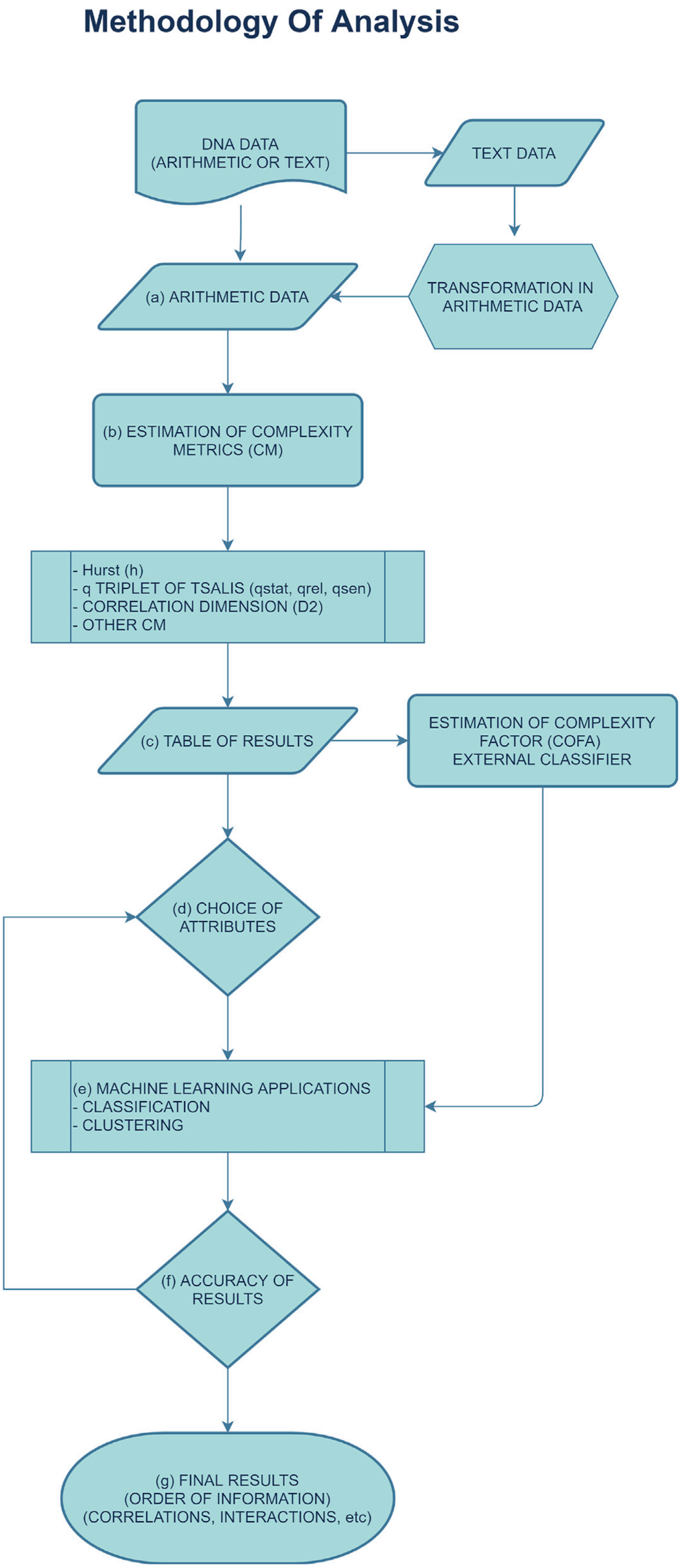


Figure 3. The flow chart diagram of the method of analysis

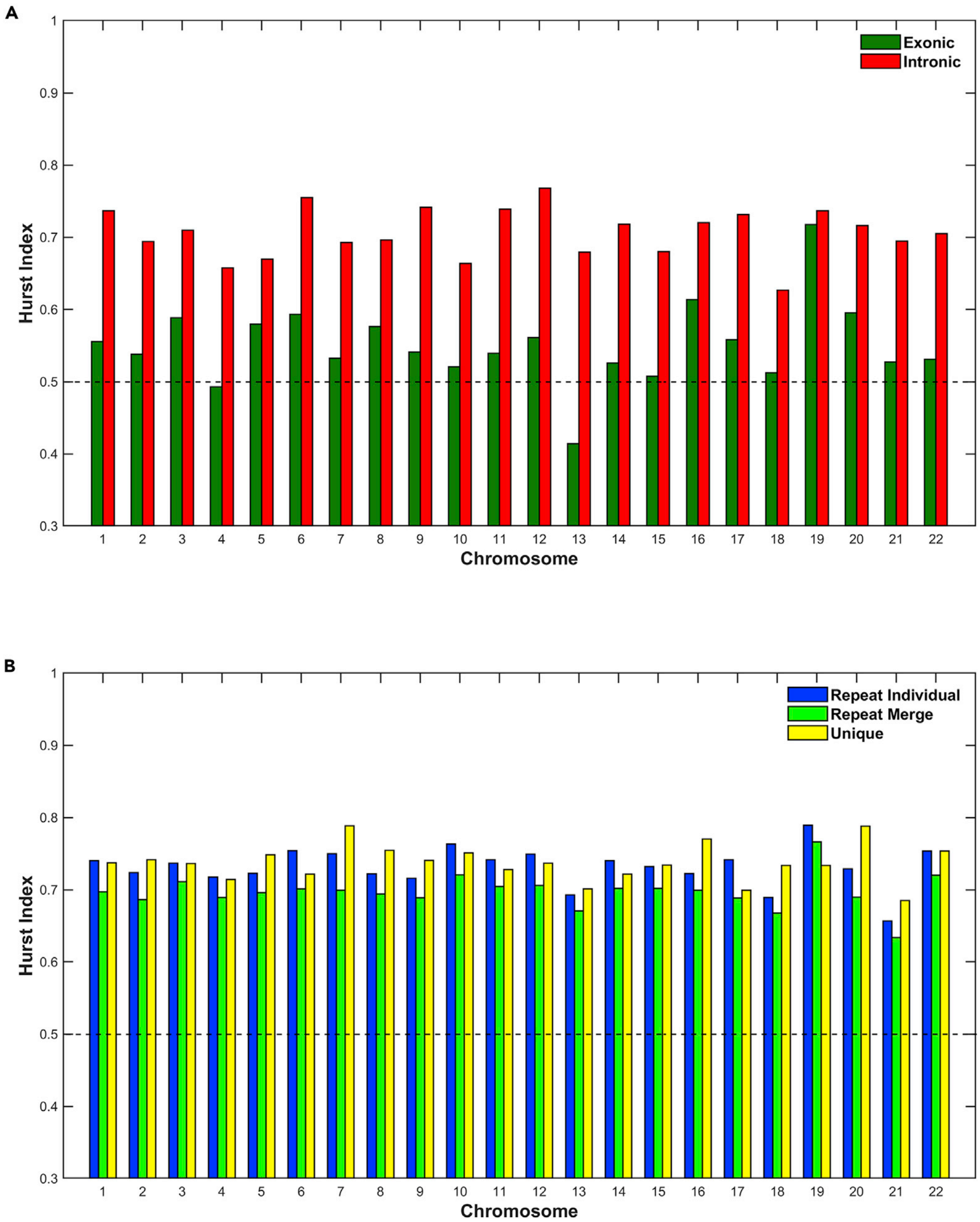


Figure 4. The estimation of Hurst exponent per chromosome and genomic entity
(A and B) (A) Exons, Introns; (B) Repeats, Unique.

sequences of the first larger Chromosomes 1–4 have an average value of 1.324 ± 0.003, whereas the last smaller in size four Chromosomes 19–22 have an average value of 1.599 ± 0.087. This difference is statistically significant ($p = 0.0008$) and denotes a differential character of q stationary even within the unique regions. The smaller chromosomes appear to have a higher index, suggesting a higher order of long-range correlations.

Concerning the q_{rel} index (Figure 6), there is a significant differentiation between exons and introns (Figure 6A). As we observe for all chromosomes, the q_{rel} index is higher for the intronic than the exonic regions. This reveals a non-Gaussian (q_{rel}>1) relaxation process of the system to its non-equilibrium steady states (NESS) for the data in intronic regions, whereas for the signals in exonic regions it reveals a near-Gaussian ($q_{rel}\approx 1$) or Gaussian ($q_{rel} = 1$) relaxation process of the system to its NESS. The results of the q_{rel} in these

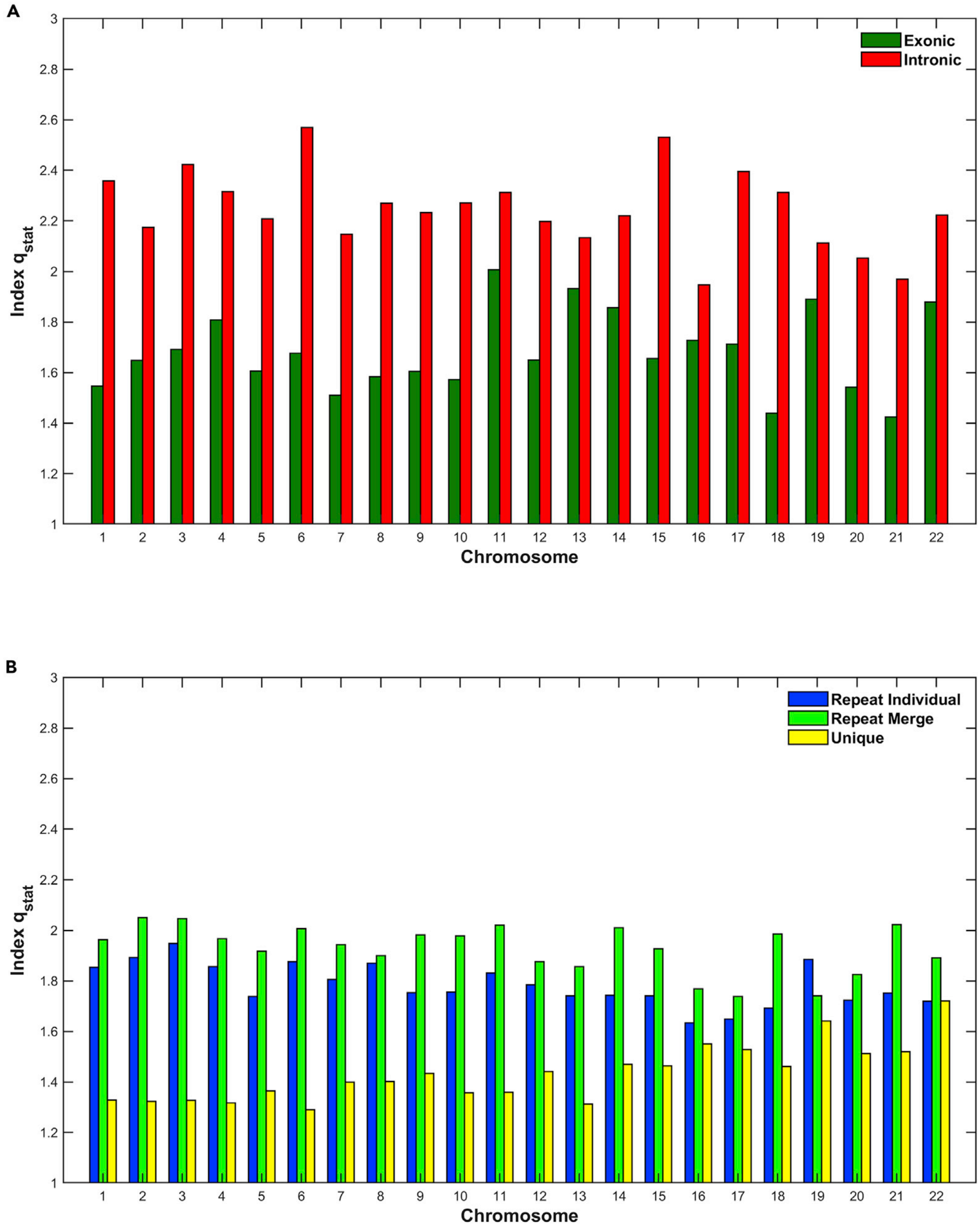


Figure 5. The estimation of q_{stat} index per chromosome and genomic entity

(A and B) (A) Exons, Introns; (B) Repeats, Unique.

regions suggest that the distribution of the sizes may reach a new metastable state with different time (space) profiles. Clearly, though, while all regions include information, they are of a complex character, such that there are differences in the degree of complexity, and therefore this complexity impacts the time (space) they take to transition to a new state of equilibrium upon being disturbed. In Figure 6B, we observe that in both repeat and unique regions the q_{rel} index is different than 1, and this reveals a non-Gaussian relaxation process of the system to its NESS. However, in certain chromosomes the q_{rel} index for one genomic entity is different from those of other(s), which means that in those cases the non-Gaussian relaxation process is stronger. Moreover, we observe differentiations of relaxation process between chromosomes within the same genomic entity. The dotted line in Figure 6B shows the limit of values in Figure 6A for visual comparison values of subfigures.

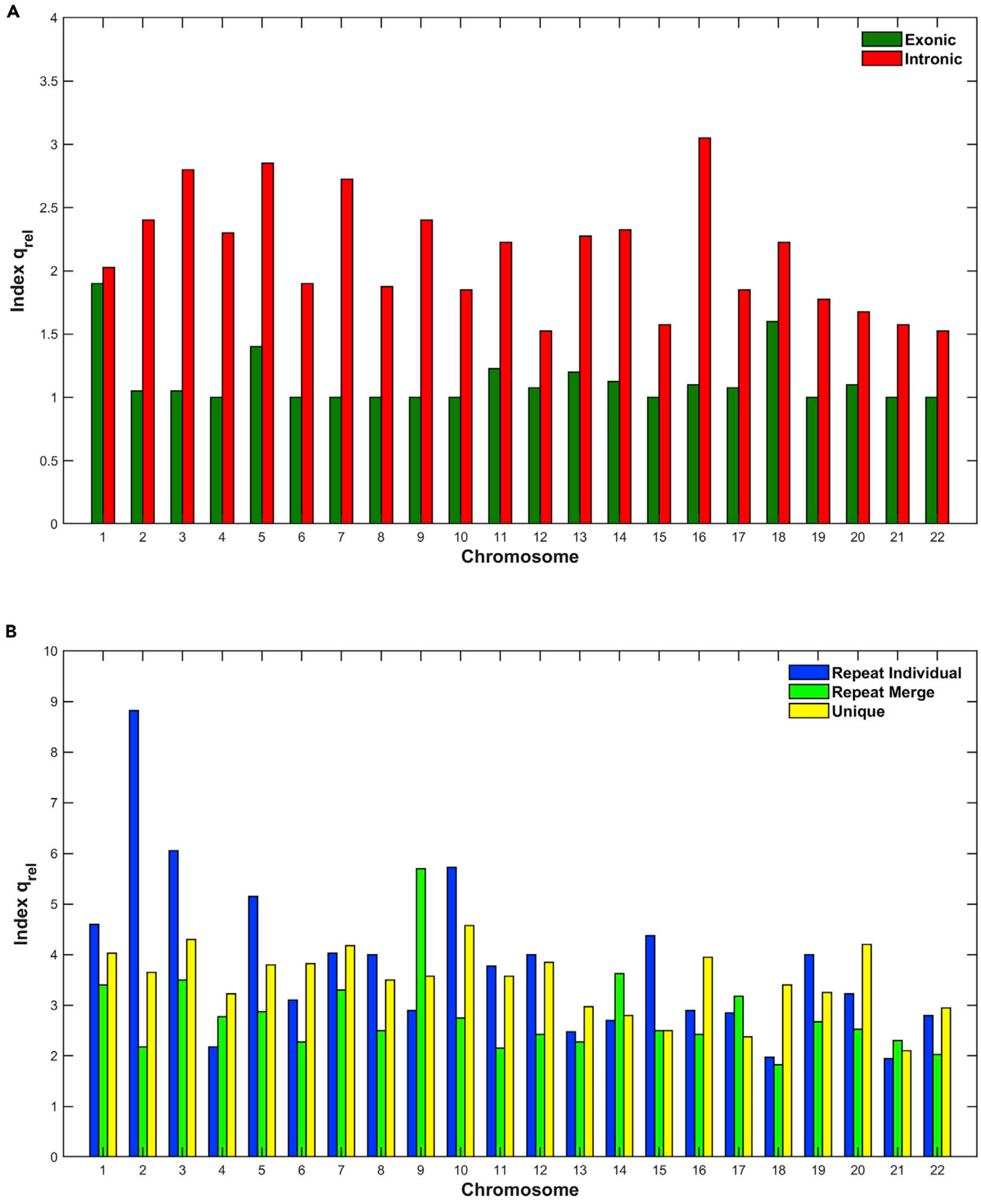


Figure 6. The estimation of q_{rel} index in per chromosome and genomic entity

(A and B) (A) Exons, Introns; (B) Repeats, Unique

Finally, concerning the q_{sen} index (Figure 7), there is a strong differentiation between chromosomes for exonic and intronic regions (Figure 7A). As one can observe, the q_{sen} index for intronic regions in all chromosomes takes much higher values than those in exonic regions, indicating that the multifractal character of the chromosomes is stronger within intronic regions. The multifractal profile verifies the presence of different scaling in physical space, which characterized the different order of information per region and per chromosome in the entire genome. Moreover, the multifractal character is different between chromosomes regarding intronic regions. Oppositely, in the exonic regions the multifractal character has almost the same behavior for most of chromosomes. For the repeat and unique genomic entities, we observe similar results as the exonic regions, but with smaller values of q_{sen} index (Figure 7B). The dotted line in Figure 7B shows the limit of values in Figure 7A for visual comparison values of subfigures. In certain chromosomes, there is no differentiation of multifractal character between different genomic entities.

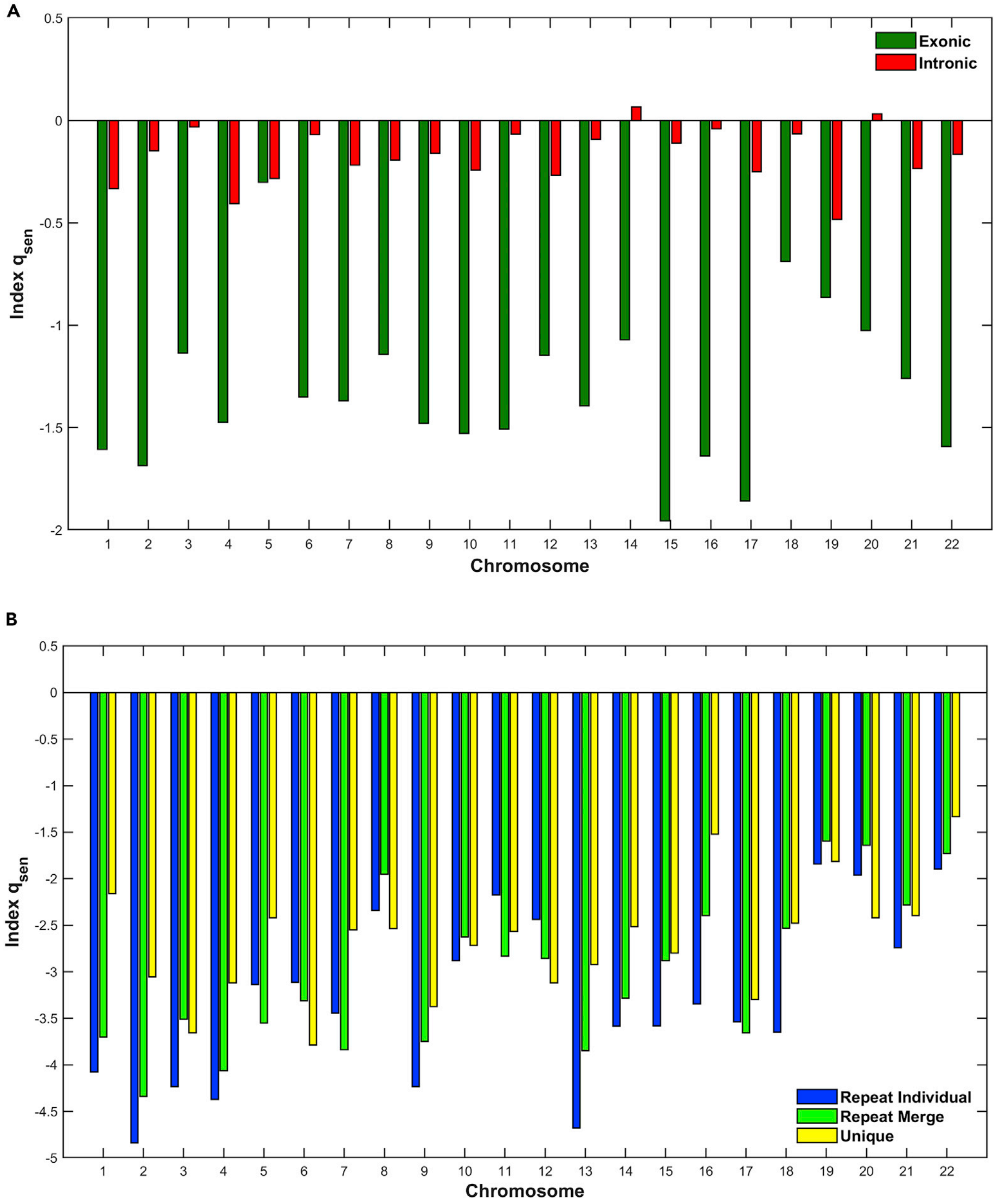


Figure 7. The estimation of q_{sen} index per chromosome genomic entities

(A and B) (A) Exons, Introns; (B) Repeats, Unique.

Moreover, there is a differentiation of multifractal character between chromosomes with higher values than the repeat and unique ones.

Correlation dimension

In Figure 8, the estimation of correlation dimension (D_2) is presented. For a random system the correlation dimension is approaching the embedding dimension. In contrast, a more deterministic self-organized system, the correlation dimension, remains at lower values from the embedding dimension. The estimation of the correlation dimension showed that the distribution of the sizes of the intronic regions reveals strong self-organization with strong variations per chromosome. The self-organized behavior means the existence of fundamental laws that produced the order of the sizes of intronic regions. Moreover, as we observe in Figure 8A, there is a differentiation between chromosomes, but the important thing here is the reduction of dimensionality of intronic and exonic genomic entities and even more the significant reduction of

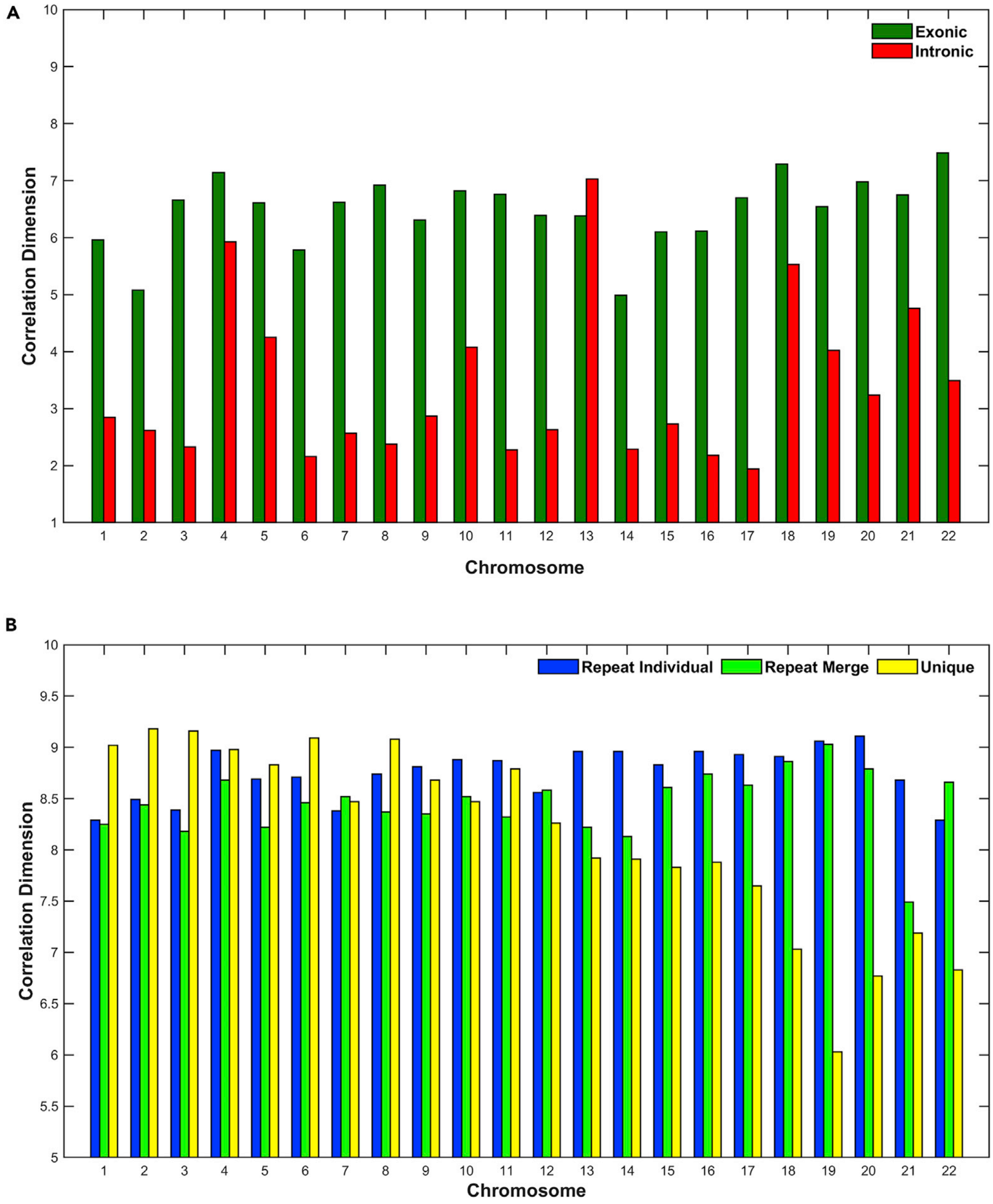


Figure 8. The estimation of Correlation Dimension (D_2) per chromosome and genomic entity

(A and B) (A) Exons, Introns; (B) Repeats, Unique.

dimensionality of intronic against exonic regions. As one can see, the correlation dimension for the intronic region is $D_2 \leq 5$ for almost all chromosomes (except Chromosomes 4, 13, and 18), whereas while the correlation dimension for the exonic region is $D \geq 5$ for all chromosomes. In Figure 8B, we observe the correlation dimension for Repeat Individual, Repeat Merge and Unique signals and does not seem to be any differentiation between chromosomes or genomic entities, except in cases of Chromosomes 18–22 where a significant reduction of dimensionality ($D \leq 7$) is observed in the Unique genomic entity and significant differentiation with the rest of the chromosomes.

Complexity factor (COFA)

A technical factor was introduced to characterize the degree of complexity in the phase space, taking into account the set of complexity metrics that we used in the analysis:

Table 2. COFA for known models

Models	Hurst	(D_2) (m = 10)	q_{stat}	q_{rel}	q_{sen}	EYKLIDEAN Dist. of q-triplet	Linear COFA
Gaussian (theoretical)	0.500	10	1.00	1.00	1.00	1.00	0.050
White noise	0.491	9.18	1.00	1.00	1.00	1.00	0.053
Henon map	0.415	1.26	1.75	1.30	1.00	2.40	0.790
Logistic map	0.466	0.54	1.65	2.25	0.24	2.80	2.416
xLorenz	0.621	2.12	0.93	1.15	−1.24	1.93	0.564

$$COFA = \left(\sqrt{(q_{stat})^2 + (q_{rel})^2 + (q_{sen})^2}\right) h \Big/ D_2$$

where q_{stat}, q_{rel},q_{sen} are the q-triplet indices from Tsallis non-extensive statistics, h is the Hurst exponent, and D_2 is the correlation dimension. The scale of the factor appears the degree of complexity in the phase space in the metric of the Euclidean space. For a pure Gaussian (random) signal the Euclidean distance of the q-triplet equals 1, h equals 0.5, and for embedding dimension $m = 10$ the estimation of D_2 gives a value $\cong 10$, so the COFA estimation is: $COFA = \frac{1x0.5}{10} = 0.05$. In the Table 2 we show the estimation of COFA for a various known models. The COFA creates a metric that characterizes the amount of the strange dynamics in a geometrical Euclidean space, and it can be used as an external classifier to the ML modeling. The COFA is a linear transformation of the complexity metrics that we used. In future studies a non-linear transformation of the factor will be presented as well.

In Figure 9, the estimation of the technical term COFA per chromosome and genomic entity is presented. The dotted line in Figure 9A shows the limit of values in Figure 9B for visual comparison values of subfigures. As we observe in Figure 9A, there is a significant variation of COFA between genomic entities and chromosomes. The Exonic are characterized by a low complexity ($COFA < 0.2$), whereas the Intronic by high complexity ($COFA > 0.6$). In Figure 9B, we observe Repeat Individual, Repeat Merge, and Unique genomic entities were characterized by low ($COFA < 0.2$) and medium ($0.2 < COFA < 0.6$) complexity.

Machine learning algorithms

In this section we used the estimation of complexity metrics as an input in ML algorithms for classification clustering and prediction with the thought to see if the variation of the metrics that correspond to each genomic entity for all chromosomes can be identified as a common dynamical feature that is characterizing these genomic entities. We analyzed these set of metrics first with a supervised classification based on Nave Bayes classifier, and second, with a k-means clustering.

Supervised classification (Naive Bayes classifier)

We used the supervised classification based on Naive Bayes classifier (see Supplementary Information for details). We prepare the model using a different set of complexity metrics every time we run the classification process. Table 3 shows the classification model's accuracy for each try, and the Figure 10 shows the block diagram of the model. These tables, also known as Confusion Matrices, reveal true versus predicted values. The diagonal of each matrix represents the correct predictions. The first set of variables (h,q_{stat}, q_{rel},D_2,ΔD_q) gives the highest accuracy ((Correct predictions)/(Number of Examples)) of 95.56%.

For the classifier's evaluation we used a 60/40 train/test set split. We split the dataset into a training dataset and a test dataset. Our model randomly selects 60% of the instances for training and uses the remaining 40% as a test dataset. On the test dataset the accuracy of our model is:

- 95.56% with attributes: h, q_{stat}, q_{rel}, D_2, ΔDq
- 92.59% with attributes: h, q_{stat}, q_{sen}, q_{rel}, D_2, ΔDq
- 75.56% with attributes: h, $(q_{stat})^2$+$(q_{rel})^2$+$(q_{sen})^2$, D_2, ΔDq
- 77.78% with attributes: $COFA$, ΔDq

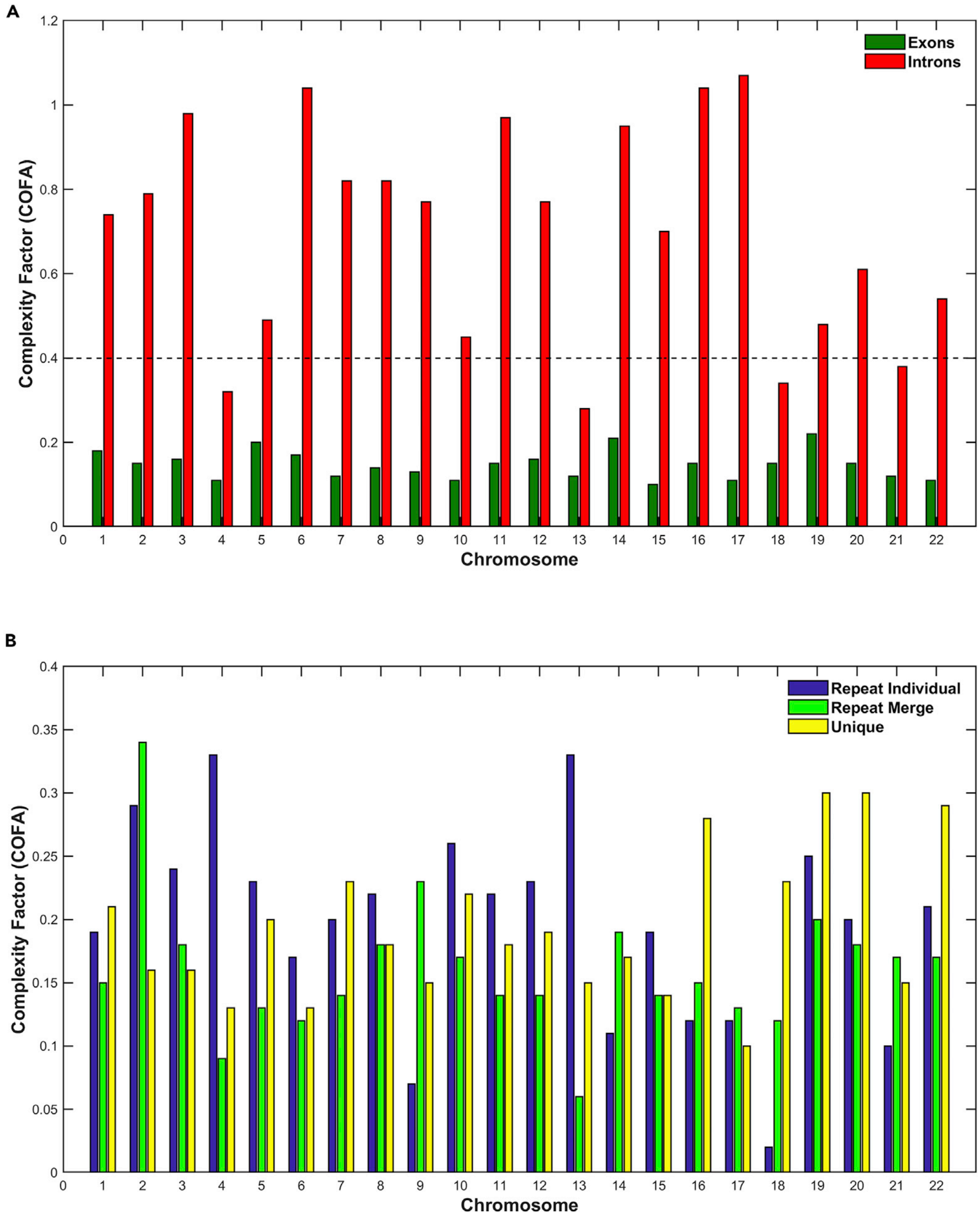


Figure 9. The estimation of the technical term Complexity Factor (COFA) per chromosome and genomic entiτυ

(A and B) (A) Exons, Introns; (B) Repeats, Unique. The dotted line in Figure 9A shows the limit of values in Figure 9B φo

K-means clustering (unsupervised)

Similar to the previous paragraph, we applied the unsupervised k-means clustering (see Supplementary Information for details). We prepared the model using a different set of complexity metrics every time we ran the clustering process. To evaluate each clustering process we used the Davies-Bouldin (DB) index (Davies and Bouldin, 1979). The DB index provides an internal evaluation schema (the score is based on the cluster itself and not on external knowledge such as labels) and is bounded from 0 to 1, where a lower score is better.

In Figure 11, the DB performance of each model with a different number of attributes for different values of k parameter is presented. The number, the type, and the combination of attributes characterized the

Table 3. Accuracy table/set of attributes

Accuracy: 95.56%	True exons	True introns	True Rep. Ind.	True Rep. merge	True unique	Class precision
Pred. Exons	9	0	0	0	0	100.00%
Pred. Introns	0	9	0	0	0	100.00%
Pred. Rep. Ind.	0	0	9	2	0	81.82%
Pred. Rep. Merge	0	0	0	7	0	100.00%
Pred. Unique	0	0	0	0	9	100.00%
Class recall	100.00%	100.00%	100.00%	77.78%	100.00%	

success of the model performance. The set of (h,qstat, qsen,qrel,D2,ΔDq) complexity metrics gave the best DB index performance (0.155), and specifically we had the lowest value for $k = 5$ parameter.

In Figure 12A we showed the number of regions per chromosome that are included in the cluster, and the block diagram of the model is shown in Figure 12B. It is clear that the model managed to separate the DNA regions in different clusters with very high accuracy for Exonic, Intronic, and Unique and high accuracy on Repeat Individual. The region Repeat Merge had the lowest accuracy.

In Figure 13A, the 3D scattered plot is presented with complexity metrics: h, q_{stat} and (D_2). In Figure 13B, the results of k-means model, with the same complexity metrics, with clusters $k = 5$ is shown. The variation of the complexity metrics is identified from the clustering model in high accuracy for regions Intronic, Exonic, and Unique and high-medium accuracy for the rest of the regions.

K-means clustering (unsupervised) based on COFA index

Similarly to the previous paragraph, we applied the unsupervised k-means clustering based on the COFA metric for different values of parameter k. The best results for the DB index versus the k parameter are presented in Figure 14. For the parameters k = 5 we had the lowest values of DB index performance. This means that the k-means model creates five clusters.

In Figure 15 the clusters for the best DB index performance are presented. Each cluster included a set of different genomic entities from different chromosomes with a common geometrical center of the variations of the COFA index. With this method of clustering based on the COFA index we discriminated sets of genomic entities per chromosomes, which appears to have similar dynamics or dynamics that live around a local center. These sets may contain specific flows of information that are produced from fundamental laws and symmetries. It would be promising to see these findings in the laboratory.

DISCUSSION

In this study, the size distribution of sub-genomic regions, were used to develop an insight of the degree of complexity behavior and internal organization of chromosomes, as reflected in the sizes of exonic, intronic, repeat individual, repeat merge, and unique regions of the genome. The analysis was based on complexity metrics to phase or physical space with the estimation of Hurst exponent, multifractal indices, q-triplet of Tsallis, correlation dimension, and COFA index and presented variations in the degree of complexity behavior per region and chromosomes. In particular, the low-dimensional deterministic non-linear chaotic dynamics (anomalous random walk-strange dynamics) and the non-extensive statistical character of the sizes of the sub-genomic regions were verified with strong multifractal characteristics and long-range correlations.

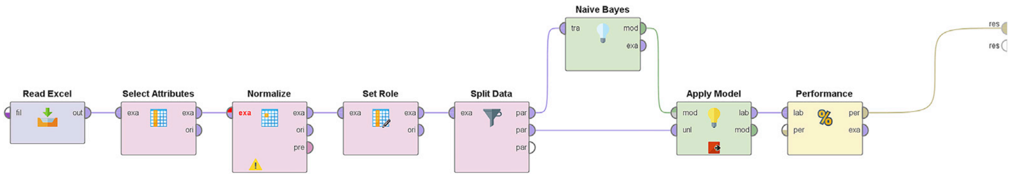


Figure 10. Block diagram of the model

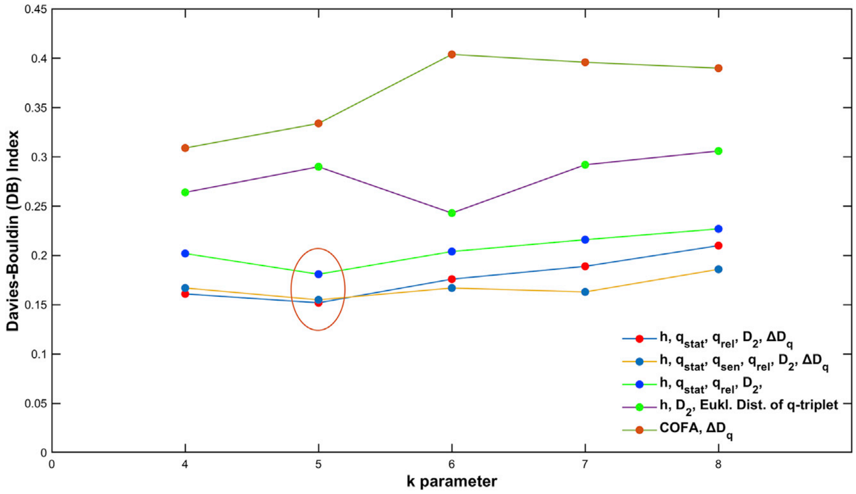


Figure 11. The DB index performance versus number of attributes for different k

The results of this study demonstrate that the DNA chromosomic system is a dynamic system working throughout an anomalous random walk and strange dynamic process underlying the biological temporal evolution and creating the DNA multifractal structure system. The multifractal DNA character reveals that the DNA system is a globally unified, multiscale self-correlated and information storage of a fractal system. The evolution of the chromosomic system includes consecutive critical points and self-organizing phase transition scenario included in the DNA dynamics. This process creates critical self-organized states with the DNA being a storage of information redundancy, according to the DNA entropy reduction and self-organization. This process corresponds to the maximization of Tsallis entropy function at different chromosomic regions. The DNA chromosomic system includes scales and fundamental laws everywhere as the DNA entities are built through the underlying DNA strange dynamics. Moreover, the findings of this study reveal the chromosomic DNA system as the storage of biological information of which only a small fraction has been decoded. In this direction, the complexity theory and the computational tools can lead to further decoding of the hidden information within the DNA. In addition, the Tsallis theory used in this study showed the existence of the non-Gaussian character everywhere in the DNA.

Notably the results of the Hurst exponent reveal that the distributions of sizes of all regions in the genome are characterized by memory character or persistent behavior in all chromosomes. Specifically, this memory character has a differential profile so much between exonic and intronic regions within a single chromosome and also among all chromosomes. Generally, it is observed that intronic regions maintain a higher Hurst exponent in all chromosomes suggesting that the size distribution of intronic regions possess an enriched multiplicity character with a high degree of organization, as opposed to exonic regions that maintain a lower degree of multiplicity and therefore a lower degree of organization. This, in biological terms, may suggest that intronic regions are engaged in multiple structural or functional roles, whereas exons are more restricted in terms of functionality and multiplicity of roles. Additionally, the distributions of sizes in repeat and unique regions are characterized by a similar memory/pattern behavior with small fluctuations, not only between themselves within each individual chromosome but also as a set of repeat and unique sequences among all chromosomes. We observed that these regions possess a high Hurst exponent in all chromosomes, similar to intronic regions, meaning that the size distribution of these regions appears to have a high degree of organization reflected at different levels (multi-scaling). This in turn suggests that the role of repeat/unique regions in the genome is of comparable complexity not only within each chromosome but also among all chromosomes.

The results of the *q* stationary reveals that the size distribution of the different genomic regions is characterized by long range correlations. This non-extensive behavior is stronger in intronic regions when compared with exonic regions with some degree of variations per chromosome. Similarly, the variations are also significant in the exonic regions, reflecting long-range correlations within chromosomes. Both intronic and exonic size distributions are independent of the chromosomal size. No particular trend was identified between the size of the

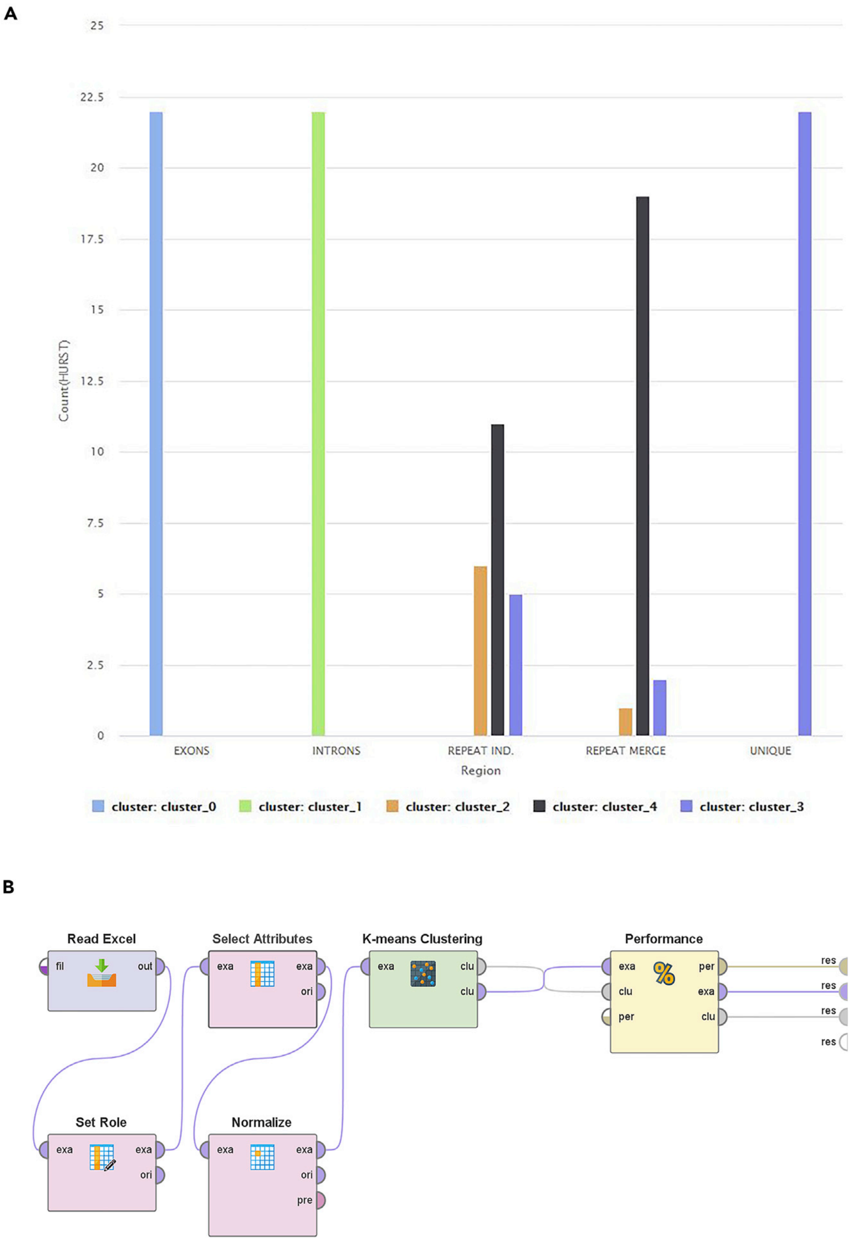


Figure 12. The model of the unsupervized k-means clustering

(A and B) (A) The clustering model's results for the best try; (B) block diagram of the model.

chromosome and the distribution of the size interactions of the two different sub-genomic regions. These results would suggest that the sizes of these two regions (exonic and intronic) in a particular location of the chromosome are coordinated with sizes located in other distant regions of the same kind (exon to exon or intron to intron) within a single chromosome and that all chromosomes have similar interactive structural relationships dictated by the same principles. These interactions and functionalities regarding the sizes of intronic regions, however, are more extensive than the exonic ones, as suggested by the Tsallis q stationary index. Regarding the unique versus repeat sequences our data suggest that the sizes of the repeat individual sequences are expectedly not very different from the sizes of the repeat merge and their long-range correlations are significantly more extensive than the size interactions of the unique sequences. Characteristically, the size interactions of the unique sequences are such that as the sizes of the chromosome decrease, the q stationary index increases, reflecting stronger interactions and interdependencies of the size of the unique regions within the four last shorter

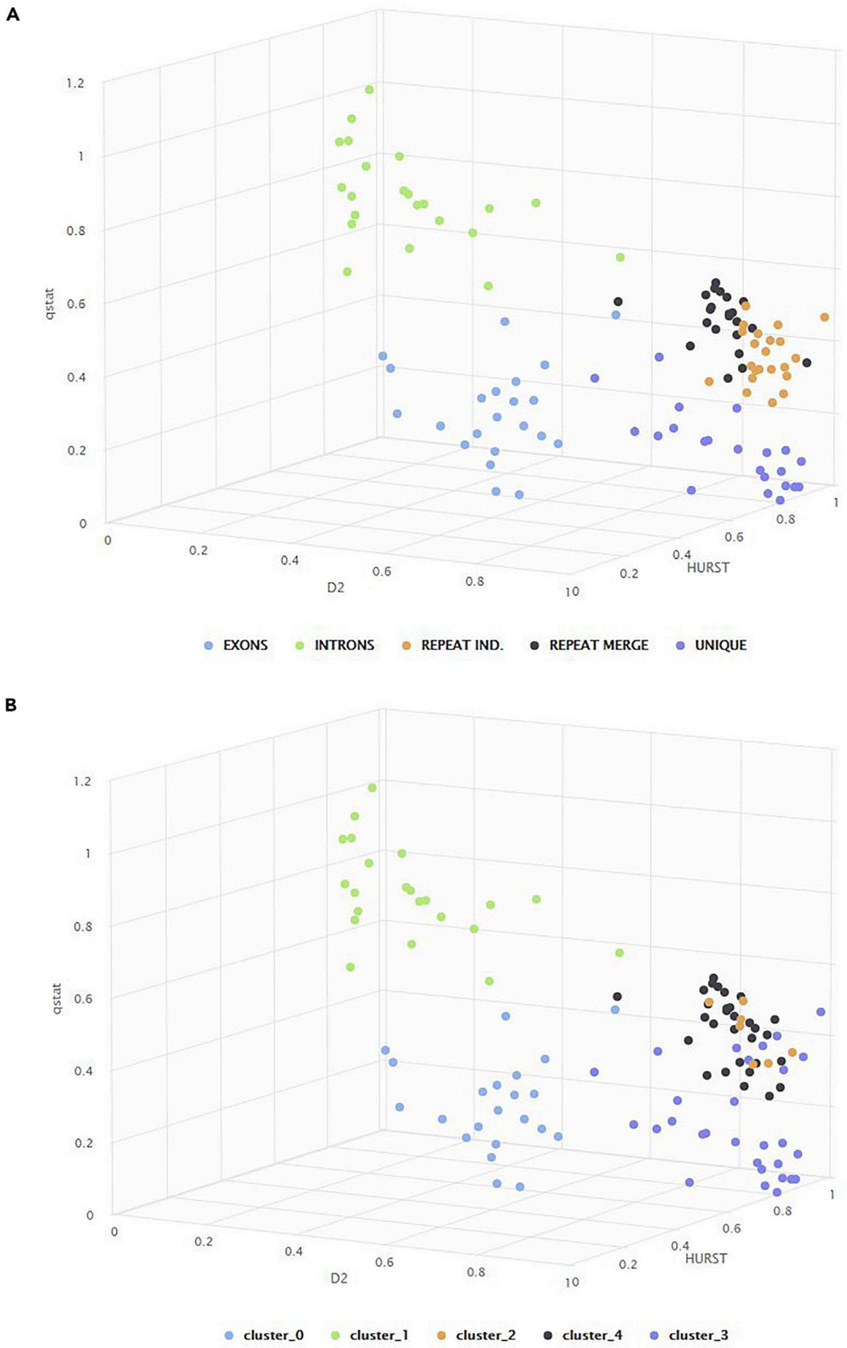


Figure 13. Visualization of the regions clusters in all genome

(A and B) (A) real 3D visualization; (B) 3D visualization after k-means clustering ($k = 5$).

chromosomes. The same does not apply for the sizes of the repeat sequences. The q stationary index clearly demonstrates that there is a coordination of the distributions of the sizes of the different sub-genomic regions (exonic, intronic, repeats, unique) within chromosomes characterized by specific profiles per genomic sub-region and chromosome.

The results of the *q* relaxation suggest that the distribution of the sizes of these regions, upon disturbing the particular order of sequences (different kind of genomic variations), may reach a new metastable state

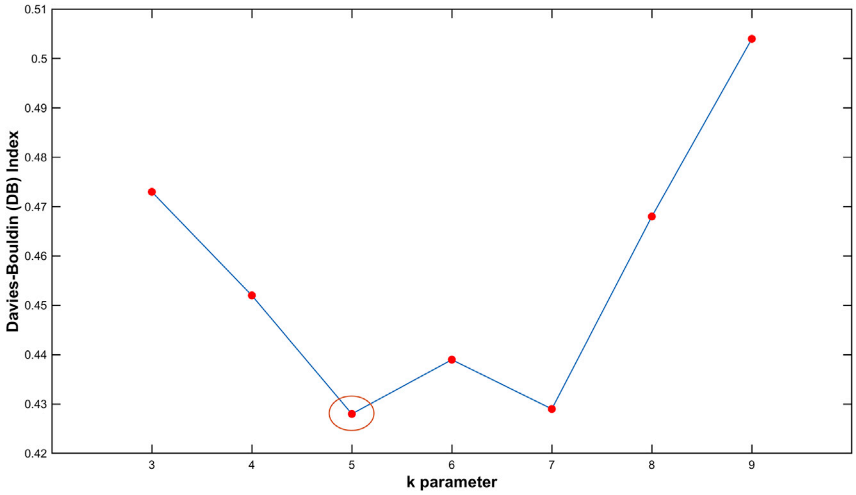


Figure 14. The DB index performance versus parameter k (where k is the number of clusters)

with different time profiles. Clearly though, while all regions include information, and they are all of a complex character, this complexity varies from region to region. Therefore, their degree of complexity impacts the time it takes to transition to a new state of equilibrium upon being disturbed. This character is reflected in the q relaxation index. Differences in the relaxation process per genomic entity, like between exonic and intronic (Figure 6A) indicate that the intronic regions with the higher q index would reach a new metastable equilibrium in a shorter period of time when compared with exonic regions with a lower q index that would take more time to reach a new metastable equilibrium. This observation is compatible with the results from the Hurst exponent, whereby the intronic regions are of a higher multifractal character when compared with exons, suggesting that intronic segments are of more complex nature, and that any disturbing event in these intronic regions needs to be addressed/restored in a shorter period of time. In more direct and simplified terms, the more complex the system the greater the need for its timely restoration. Intuitively, someone may think that higher complexity would dictate longer periods of restoration, but apparently for the proper balancing of the whole system, the degree of complexity may dictate degrees of priority in terms of functionality, and therefore, the more complex the system the higher the need for its immediate restoration. The enriched complex character of intronic regions, when compared with exonic, offers a multitude of alternative paths for restoration and therefore of a faster recovery time.

Moreover, the results of the q sensitivity reveal that the size distribution of regions have a multifractal profile in all chromosomes and also significant variations per chromosome. The multifractal profiles verify the presence of different scaling in phase space of different regions and at different chromosomes. Specifically, the multifractal profile is stronger in the distributions of sizes of intronic regions when compared with exonic regions. This result in biological terms may suggest that intronic regions operate at multiple structural or functional levels, whereas exonic regions reflect a different and less complex structural/functional mechanism. Additionally, the multifractal profile is weaker in the distributions of sizes in repeat and unique regions than exonic and intronic regions, and between them the multifractal profile has similar shape concluding similar number of subsets of structural or functional roles. These are reminiscent of our earlier observations from the Hurst exponent. Two different approaches reveal similar characters for the respective sub-genomic regions.

Correlation Dimension is another complexity metric reflecting the size of the strange attractor in the phase space. When the system is embedded in higher dimensions and the system shows strong self-organization then we have reduction of dimensionality in the phase space, and so the correlation dimension remains significantly in lower values from the embedding dimension (Argyris et al., 1998; Grassberger and Procaccia, 1983, 2004). Lower values of the Correlation Dimension index reflect higher self-organization. In the reconstructive phase space, the distribution of the intronic region reveals strong self-organization with significant variations per chromosome. The lower values of Correlation Dimension index of the sizes of intronic regions, when compared with exonic, demonstrate the stronger self-organization of the intronic segments. This, is turn, reflects the existence of fundamental laws, which produced the distribution of sizes in the

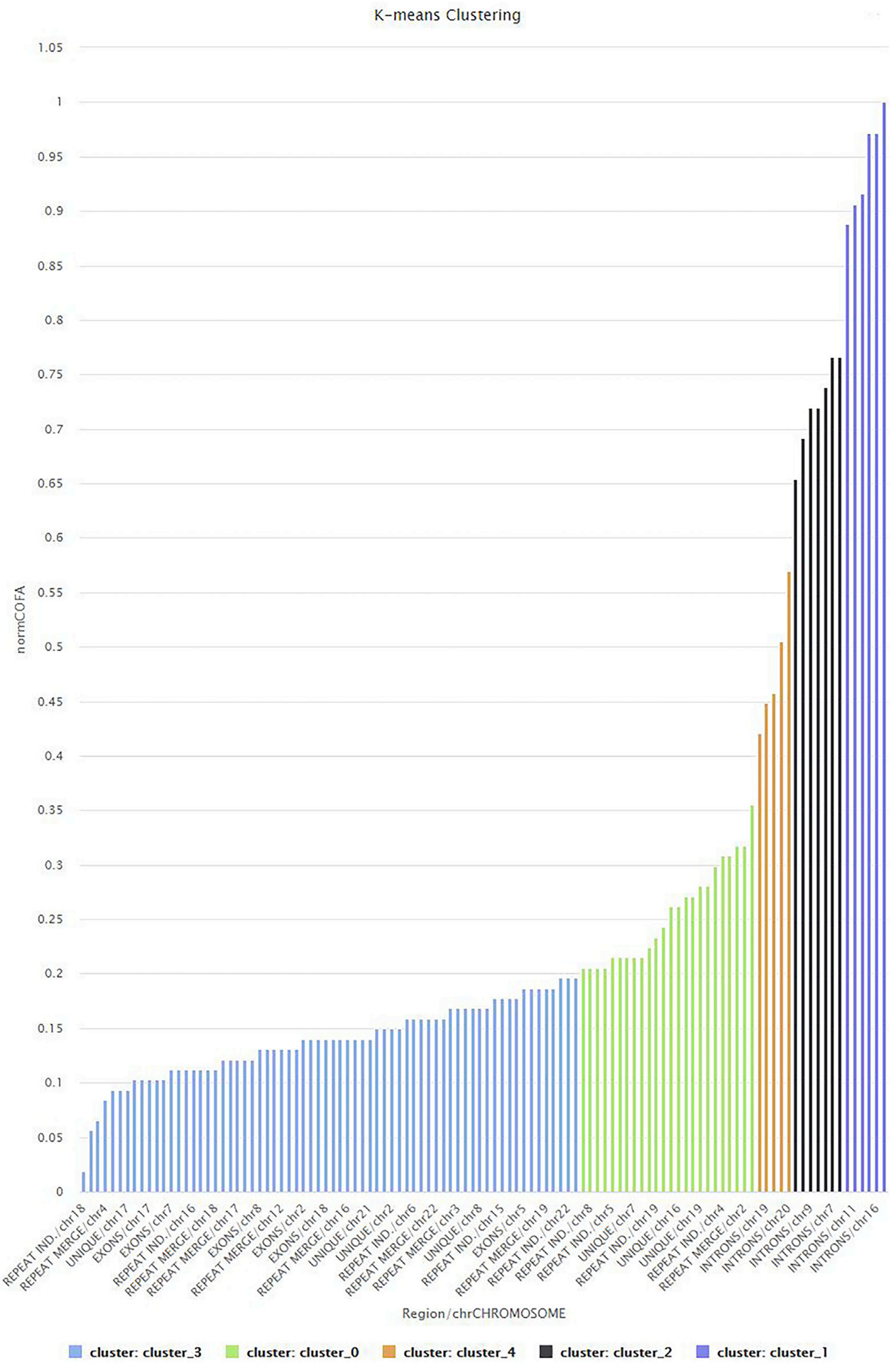


Figure 15. Visualization of the clusters in all genome after k-means clustering (k = 5) based on COFA index

We separate the results in five clusters. In the x axis are the genomic entities and chromosome reference; for example, the first one is Repeat individual genomic entity in Chromosome 18; the y axis is the COFA index.

aforementioned regions. This stronger self-organization would imply an enriched multilevel functional character for the intronic regions, quite different from that of the exonic regions. These findings are concordant with interpretations we have already provided using other complexity metrics like Hurst exponent and q sensitivity. Different complexity metrics reveal the same complexity character for the intronic/exonic regions and therefore strengthen the conclusions drawn regarding their complexity and therefore content information and multiple functionalities. Furthermore, the distributions of sizes of unique regions are such that the self-organization gradually increases, as reflected in the lower Correlation Dimension values, starting with Chromosome 12. The same does not apply for the repeat regions. However, the observation regarding the repeat regions using Correlation Dimension analysis is concordant with the earlier observations derived from the q stationary analysis demonstrating that the higher complexity character of unique regions is smaller chromosomes. Generally, the presence of self-organization regarding the sizes of the different sub-genomic entities follows the principle that the degree of multilevel functionality depends on the degree of self-organization.

The technical index COFA, which represents the geometrical measure of the complexity in a Euclidean space (Hurst exponent, the Euclidean distance of the *q*-triplet of Tsallis statistics, and the Correlation Dimension index) was successfully used as a technical term to describe cumulatively the degree of complexity. COFA index below 0.08 suggests that the system lacks structure and is low in complexity behavior. Values over 0.08 reflect higher complexity behavior and order of information content. In Figure 9A we observe that the size of intronic segments is permeated by a significantly higher complexity when compared with exonic regions. Basically, the data are a synthesis of all previous complexity indexes we have already discussed. Figure 9B respectively demonstrates the relative complexity of repeat and unique sequences, whereby each chromosome appears to have its own unique character. The COFA index for these regions is consistently below 0.35.

Overall, the existence of strange dynamics in the phase space with set of attractors with multifractal profile reflects the existence of symmetries and fundamental laws that finally produced the multi-dimensional structural-functional mechanism of the genome. Such symmetry is demonstrated and strongly suggested by the proportional relationship identified between the number of exonic and genic regions, which constant (see Table 1) in all chromosomes. It derives that all other ratios of exonic/intergenic, intronic/genic, and intronic/intergenic have the same ratio that is close to 10.

Additionally, the estimation of complexity metrics for a subset of chromosomes (60%) was used as an input in ML algorithms for classification clustering and prediction with the intent to assess whether the variation of the metrics that correspond to each genomic entity for all chromosomes can be identified as a common dynamical feature that characterizes these genomic entities. We used first a supervised classification based on Naive Bayes classifier and second with a k-means clustering. The models successfully re-create the size of all regions in different clusters with high accuracy for the rest of the 40% of the genome (chromosomes), basically confirming the validity of the overall approach. Furthermore, we used the COFA index along with ML models as a new external classifier. With COFA index and ML models we identify sets of genomic regions among all chromosomes that present similar dynamics (similar COFA index) or dynamics that lives around a local center (cluster center). These new sets may contain interactions of information among genomic entities and chromosomes based on internal laws and symmetries. This is a different way to assess either physical interactions or information flow among different genomic regions and chromosomes. The aforementioned approach can be subjected to modifications to further improve the accuracy and our results.

In conclusion, the results demonstrate that the underlying dynamical processes, which give rise to the organization of the genome, correspond to the extremization of *q*-entropy principle included in the non-extensive statistical mechanics of Tsallis (Broomhead and King, 1986; Klimontovich, 1994). The *q*-entropy principle of Tsallis applies the unification of the macroscopic to the microscopic level through the multi-scale interaction and the scale invariance principle included in the power laws of complex phenomena.

It is to be noted that in this work we used the Tsallis entropy. In general, different entropies are used for different reasons in different cases, depending on the particular application. For example, a system that moves toward thermodynamic equilibrium maximizes Gibbs entropy. Thus, when you are in this case, it is natural to use Gibbs entropy to analyze it. In case of far from equilibrium that we are at, as biological systems are, systems evolve toward maximizing Tsallis entropy (Tsallis, 2009). Specifically, for the DNA, it has been shown that it can be viewed as an out-of-equilibrium structure (Provata et al., 2014a, 2014b). Therefore, the Tsallis entropy analysis is appropriate and adequate. Furthermore, it has been verified from our previous studies that the probability distribution of the DNA structures follows the Tsallis entropy maximization probability distribution function (Pavlos et al., 2015). Other entropies may be relevant depending on the nature of the particular system under investigation. Most recent literature, as mentioned earlier in the Introduction, section, utilizes other entropies for different questions but not for the particular questions that we have addressed in this work. Considering that our previous work (Pavlos et al., 2015; Karakatsanis et al., 2018) has consistently utilized the Tsallis entropy, our current analysis was also performed in a similar fashion. As our approach to study the lengths of subgenomic regions of DNA has not studied by others, using other entropy approaches such comparisons are not presently possible. It should be mentioned, however, that although different methods have been used to analyze the DNA and its information content, they show some commonalities in their general findings. As a general trend, they distinguish between different structural regions of the genome, and differentiate between coding and non-coding regions of DNA (Karakatsanis et al., 2018, Thanos et al., 2018).

The projection of the dynamics to the statistics in the phase space develops a complete picture that integrated to the variations of the complexity metrics. The redundancy of information in DNA lies between randomness and order in a continuous evolutionary process of thousands of years (Beltrami, 1999) based on fluctuations and deterministic laws. This picture of dynamics can be identified from ML tools for clustering, classification, and prediction. The results of the ML tools successfully identified the different degree of complexity profile of the distribution of the regions with high accuracy, based on a given set of complexity metrics. In conclusion, the distribution of the size of the genome entities is characterized from different degree of complexity profiles, which is recognizable from the ML models. This integrated methodology (Figure 3) is a different approach for the identification of the symmetries and fundamental laws, which produces the order of information in all genome and generates strange dynamics that is observable and qualitatively measurable. Finally, the merging of interdisciplinary complexity theory and genomics can provide semantic results in the direction of a deeper understanding and promotion of the fundamental laws of biology with new motifs, patterns, and interactions of the complex biological information.

Limitations of the study

In our study, publicly available DNA sequences were used to analyze the lengths of the different genomic entities using complexity metrics and ML. Admittedly, however, these sequences are a compilation of many different sequencing projects resulting in a single reference sequence for the human genome. Optimally, a single source of DNA sequenced for the whole human genome would be a better source. As our tools for whole-genome sequencing improve and generate credible sequencing data, our computational analysis can be repeated and confirm our current findings.

Regarding limitations related to our computational approach further optimizations are possible and can be applied. For example, modification of the COFA index, by including additional metrics or modifying the form to reflect linear or non-linear relationships, may provide better identification of the dynamics in the phase space of the DNA system. Furthermore, to the well-known clustering and categorization algorithms used in our study, other specialized algorithms based on a self-organized neural network, like the self-organizing feature map, can be used to enhance the performance of the model.

Resource availability

Lead contact

karaka@env.duth.gr.

Materials availability

All data needed to evaluate the results and conclusions are presented in the main text. Scripts related to this paper are available from the corresponding authors.

Data and code availability

The Genomic compartments we used in this study and the Gene definitions are taken from National Center for Biotechnology Information (NCBI) RefSeq (RefSeq Annotation Release 108, https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/108/).

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102048.

ACKNOWLEDGMENTS

This research was partially funded by CHOP institutional funds to D.S.M.

AUTHOR CONTRIBUTIONS

L.P.K., G.P.P., and D.S.M., conceptualization; L.P.K., methodology; L.P.K., E.G.P., and G.T. software; L.P.K., E.G.P., and G.T. formal analysis; T.M. and J.L.D. investigation; L.P.K., E.G.P., and G.T. resources; T.M. and J.L.D. data curation; L.P.K., E.G.P., G.T., G.L.S., G.P.P., and D.S.M. writing – original draft preparation; L.P.K., E.G.P., G.T., G.L.S., T.M., G.P.P., and D.S.M. writing –review and editing; L.P.K., E.G.P., and G.T. visualization; L.P.K. and D.S.M. supervision; L.P.K. and D.S.M. project administration;

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 12, 2020
Revised: November 30, 2020
Accepted: January 6, 2021
Published: February 19, 2021

REFERENCES

Anitas, E.M. (2020). Small-angle scattering and multifractal analysis of DNA sequences. Int. J. Mol. Sci. *21*, 4651.

Apostolou, P., Iliopoulos, A.C., Parsonidis, P., and Papasotiriou, I. (2019). Gene expression profiling as a potential predictor between normal and cancer samples in gastrointestinal carcinoma. Oncotarget *10*, 3328–3338.

Argyris, J., Andreadis, I., Pavlos, G., and Athanasiou, M. (1998). The influence of noise on the correlation dimension of chaotic attractors. Chaos, Solitons & Fractals *9*, 343–361.

Bak. (2013). Per. How Nature Works: The Science of Self-Organized Criticality (Springer Science & Business Media).

Beltrami, E. (1999). What is Random?. Chance and Order in Mathematics and Life (Springer Nature).

Ben-Mizrachi, A., Procaccia, I., and Grassberger, P. (1984). Characterization of experimental (noisy) strange attractors. Phys. Rev. A *29*, 975.

Broomhead, D.S., and King, G.P. (1986). Extracting qualitative dynamics from experimental data. Physica D Nonlinear Phenomena *20*, 217–236.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M., Sciortino, F., and Stanley, H.E. (1993). Long-range fractal correlations in DNA. Phys. Rev. Lett. *71*, 1776.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.K., Simons, M., and Stanley, H.E. (1995). Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys. Rev. E *51*, 5084.

Bzdok, D., Altman, N., and Krzywinski, M. (2018). Points of significance: statistics versus machine learning. Nat. Methods *15*, 233–234.

Casdagli, M. (1989). Nonlinear prediction of chaotic time series. Physica D Nonlinear Phenomena *35*, 335–356.

Corona-Ruiz, M., Hernandez-Cabrera, F., Cantú-González, J.R., González-Amezcua, O., and Javier Almaguer, F. (2019). A stochastic phylogenetic algorithm for mitochondrial DNA analysis. Front. Genet. *10*, 66.

Costa, M.O., Silva, R., Anselmo, D.H.A.L., and Silva, J.R.P. (2019). Analysis of human DNA through power-law statistics. Phys. Rev. E *99*, 022112.

Davies, D.L., and Bouldin, D.W. (1979). A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intelligence PAMI- *1*, 224–227.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2017). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Frey, B.J., Delong, A.T., and Xiong, H.Y. (2019). U.S. Patent Application No. 16/179, p. 280. https://patents.google.com/patent/US20190073443A1/en.

Grassberger, P., and Procaccia, I. (1983). Characterization of strange attractors. Phys. Rev. Lett. *50*, 346.

Grassberger, P., and Procaccia, I. (2004). Measuring the strangeness of strange attractors. In The Theory of Chaotic Attractors (Springer), pp. 170–189.

Grassberger, P., Schreiber, T., and Schaffrath, C. (1991). Nonlinear time sequence analysis. Int. J. Bifurcation Chaos *1*, 521–547.

Grebogi, C., Ott, E., and Yorke, J.A. (1987). Chaos, strange attractors, and fractal basin boundaries in nonlinear dynamics. Science *238*, 632–638.

Grosberg, A., Rabin, Y., Havlin, S., and Neer, A. (1993). Crumpled globule model of the three-dimensional structure of DNA. Europhysics Lett. *23*, 373–378.

Hsu, C.F., Wei, S.Y., Huang, H.P., Hsu, L., Chi, S., and Peng, C.K. (2017). Entropy of entropy: measurement of dynamical complexity for biological systems. Entropy *19*, 550.

Karakatsanis, L.P., Pavlos, G.P., Iliopoulos, A.C., Pavlos, E.G., Clark, P.M., Duke, J.L., and Monos, D.S. (2018). Assessing information content and interactive relationships of subgenomic DNA sequences of the MHC using complexity theory approaches based on the non-extensive statistical mechanics. Physica A Stat. Mech. its Appl. *505*, 77–93.

Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., et al. (2014). Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. U S A *111*, 6131–6138.

Klimontovich, Y.L. (1994). Thermodynamics of chaotic systems: an introduction by C Beck, F Schlogel. Physics-Uspekhi *37*, 713–714.

Li, W., and Kaneko, K. (1992). Long-range correlation and partial 1/fα spectrum in a noncoding DNA sequence. Europhysics Lett. *17*, 655.

Li, J., Zhang, L., Li, H., Ping, Y., Xu, Q., Wang, R.,., and Wang, Y. (2019). Integrated entropy-based approach for analyzing exons and introns in DNA sequences. BMC Bioinformatics *20*, 1–7.

Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nat. Rev. Genet. *16*, 321–332.

Liu, X., Guo, Z., He, T., and Ren, M. (2020). Prediction and analysis of prokaryotic promoters based on sequence features. Biosystems *197*, 104218.

Lorentz, E. (1993). The Essence of Chaos (University of Washington Press).

Machado, J.T. (2019). Information analysis of the human DNA. Nonlinear Dyn. *98*, 3169–3186.

Manogaran, G., Vijayakumar, V., Varatharajan, R., Kumar, P.M., Sundarasekar, R., and Hsu, C.H. (2018). Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. Wireless Personal. Commun. *102*, 2099–2116.

Melnik, S.S., and Usatenko, O.V. (2014). Entropy and long-range correlations in DNA sequences. Comput. Biol. Chem. *53*, 26–31.

Namazi, H., and Kiminezhadmalaie, M. (2015). Diagnosis of lung cancer by fractal analysis of damaged DNA. Comput. Math. Methods Med. *2015*, 242695, https://doi.org/10.1155/2015/242695.

Namazi, H., Akrami, A., Hussaini, J., Silva, O.N., Wong, A., and Kulish, V.V. (2016). The fractal-based analysis of human face and DNA variations during aging. Bioscience Trends *10*, 477–481.

Nicolis, G. (1993). Physics of far-from-equilibrium systems and self-organization. In The New Physics, P. Davies, ed. (Cambridge University Press), pp. 316–347.

Nicolis, G., and Prigogine, I. (1989). Exploring Complexity: An Introduction (Freeman, W.H.).

Oikonomou, T., and Provata, A. (2006). Non-extensive trends in the size distribution of coding and non-coding DNA sequences in the human genome. Eur. Phys. J. B-Condensed Matter Complex Syst. *50*, 259–264.

Oikonomou, T., Provata, A., and Tirnakli, U. (2008). Nonextensive statistical approach to non-coding human DNA. Physica A: Stat. Mech. Its Appl. *387*, 2653–2659.

Ossadnik, S.M., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Peng, C.K., Simons, M., and Stanley, H.E. (1994). Correlation approach to identify coding regions in DNA sequences. Biophysical J. *67*, 64–70.

Papapetrou, M., and Kugiumtzis, D. (2014). Investigating long range correlation in DNA sequences using significance tests of conditional mutual information. Comput. Biol. Chem. *53*, 32–42.

Papapetrou, M., and Kugiumtzis, D. (2020). Tsallis conditional mutual information in investigating long range correlation in symbol sequences. Physica A: Stat. Mech. its Appl. *540*, 123016.

Pavlos, G.P., Karakatsanis, L.P., Iliopoulos, A.C., Pavlos, E.G., Xenakis, M.N., Clark, P., Duke, J., and Monos, D.S. (2015). Measuring complexity, nonextensivity and chaos in the DNA sequence of the Major Histocompatibility Complex. Physica A: Stat. Mech. Its Appl. *438*, 188–209.

Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H.E. (1992). Long-range correlations in nucleotide sequences. Nature *356*, 168–170.

Prigogine, I. (1978). Time, structure, and fluctuations. Science *201*, 777–785.

Prigogine, I. (1997). The End of Certainty: Time, Chaos, and the New Laws of Nature (The Free Press).

Provata, A., and Beck, C. (2011). Multifractal analysis of nonhyperbolic coupled map lattices: application to genomic sequences. Phys. Rev. E *83*, 066210.

Provata, A., Nicolis, C., and Nicolis, G. (2014a). Complexity measures for the evolutionary categorization of organisms. Comput. Biol. Chem. *53*, 5–14.

Provata, A., Nicolis, C., and Nicolis, G. (2014b). DNA viewed as an out-of-equilibrium structure. Phys. Rev. E *89*, 052105.

Provenzale, A., Smith, L.A., Vio, R., and Murante, G. (1992). Distinguishing between low-dimensional dynamics and randomness in measured time series. Physica D: Nonlinear Phenomena *58*, 31–49.

Silva, R., Silva, J.R.P., Anselmo, D.H.A.L., Alcaniz, J.S., da Silva, W.J.C., and Costa, M.O. (2020). An alternative description of power law correlations in DNA sequences. Physica A: Stat. Mech. its Appl. *545*, 123735.

Stanley, H.E., and Meakin, P. (1988). Multifractal phenomena in physics and chemistry. Nature *335*, 405–409.

Stanley, H.E., Buldyrev, S.V., Goldberger, A.L., Goldberger, Z.D., Havlin, S., Mantegna, R.N., Ossadnik, S.M., Peng, C.K., and Simons, M. (1994). Statistical mechanics in biology: how ubiquitous are long-range correlations? Physica A: Stat. Mech. Its Appl. *205*, 214–253.

Takens, F. (1981). Detecting strange attractors in turbulence. In Dynamical Systems and Turbulence, D. Rand and L.S. Young, eds. (Springer), pp. 366–381.

Thanos, D., Li, W., and Provata, A. (2018). Entropic fluctuations in DNA sequences. Physica A: Stat. Mech. its Appl. *493*, 444–457.

Theiler, J. (1990). Estimating fractal dimension. J. Opt. Soc. America A *7*, 1055–1073.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. J. Stat. Phys. *52*, 479–487.

Tsallis, C. (2002). Entropic nonextensivity: a possible measure of complexity. Chaos, Solitons and Fractals *13*, 371–391.

Tsallis, C. (2004). Dynamical scenario for nonextensive statistical mechanics. Physica A: Stat. Mech. its Appl. *340*, 1–10.

Tsallis, C. (2009). Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World (Springer).

Varma, M., Paskov, K.M., Jung, J.Y., Chrisman, B.S., Stockham, N.T., Washington, P.Y., and Wall, D.P. (2019). Outgroup machine learning approach identifies single nucleotide variants in noncoding DNA. Associated with autism spectrum disorder. Pac. Symp. Biocomputing *24*, 260–271.

Vinga, S., and Almeida, J.S. (2007). Local Renyi entropic profiles of DNA sequences. BMC Bioinformatics *8*, 393.

Voss, R.F. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett. *68*, 3805.

Washburn, J.D., Mejia-Guerra, M.K., Ramstein, G., Kremling, K.A., Valluru, R., Buckler, E.S., and Wang, H. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proc. Natl. Acad. Sci. U S A *116*, 5542–5549.

Woods, T., Preeprem, T., Lee, K., Chang, W., and Vidakovic, B. (2016). Characterizing exonic and intronic by regularity of nucleotide strings. Biol. Direct *11*, 6.

Wu, Z.B. (2014). Analysis of correlation structures in the Synechocystis PCC6803 genome. Comput. Biol. Chem. *53*, 49–58.

Xu, C., and Jackson, S.A. (2019). Machine learning and complex biological data. Genome Biol. *20*, 76.

iScience, Volume *24*

Supplemental Information

Spatial constrains and information content of sub-genomic regions of the human genome

Leonidas P. Karakatsanis, Evgenios G. Pavlos, George Tsoulouhas, Georgios L. Stamokostas, Timothy Mosbruger, Jamie L. Duke, George P. Pavlos, and Dimitri S. Monos

Supplemental Information

Spatial Constrains and Information Content of Sub-Genomic regions of the Human Genome

Leonidas Karakatsanis, Evgenios Pavlos, George Tsoulouhas, Georgios Stamokostas, Timothy Mosbruger, Jamie Duke, George Pavlos and Dimitri Monos

Theoretical highlights of complexity theory

The Complexity theory can give useful quantitative parameters for the description of DNA structure. Such quantities are the Tsallis q-triplet $(q_{sen}, q_{rel}, q_{stat})$, the Correlation dimensions, the Hurst exponent, the Lyapunov exponents, etc. According to Prigogine and Nicolis, far from equilibrium, nature can produce spatiotemporal self-organized forms through the extended probabilistic dynamics of correlations in agreement with the extended entropy principle (Prigogine, 1978; Nicolis and Prigogine, 1989; Nicolis, 1993; Prigogine, 1997; Davies, 2004). Far from equilibrium the entropy principle creates long-range correlations, as nature works to maximize the non-equilibrium entropy (Tsallis) function. The entropy principle for the far from equilibrium and open physical systems, as the biological systems are, leads to the creation of dissipative structures and self-organized multi-level and multi-scale long-range correlated physical forms. The maximization of Tsallis q-entropy can explain the formation of DNA structure as a non-equilibrium intermittent turbulence structure and a multiplicative self-organization process (Pavlos et al., 2015). From this point of view, the DNA structure is a constructed multifractal system of the four DNA bases (A, C, G, T) with high information redundancy. This in turn suggests that most, if not all of the DNA sequences are purposeful and relevant but not all of this information has been decoded.

The underlying intermittent DNA turbulence which constructs the DNA sequence and the chromosomic high ordered system is mirrored in the well-known q-triplet of non-extensive statistical theory of Tsallis including three characteristic parameters $(q_{sen}, q_{rel}, q_{stat})$. The q_{sen} parameter, describes the entropy production and the information redundancy, as the DNA sequence is constructed by the underlying DNA turbulence process, as multifractal DNA structure. The q_{rel} parameter describes the relaxation process of the DNA turbulence system to the meta-equilibrium stationary state of DNA structure, where the q-entropy (S_q) of Tsallis statistics is maximized. The meta-equilibrium state with maximized entropy function corresponds to the chromosomic DNA system. The q_{stat} parameter describes the statistical probability distribution function of the DNA complex or random structure at the DNA turbulent stationary state. The DNA turbulence system can be described dynamically as an anomalous random walk process creating the DNA bases series. This dynamic can include critical points where the DNA turbulence dynamics can change. This constructive biological evolutionary phase transition process can develop the entire self-organized multifractal dynamical system. The variations of the Tsallis q-triplet along the DNA sequence is the quantitative manifestation of the biological evolution process throughout the constructive scenario of critical DNA turbulent phase transition processes.

The multifractal character of this biological evolutionary process is mirrored at the evolution of the Hurst exponent along the DNA sequence. As the Hurst exponent changes along the DNA sequence it mirrors the degree of the multifractal character along the DNA structure.

The DNA structure can be explained as the dynamical evolution of the biological complex system in the underlined natural state space to the DNA turbulence dynamics. This state space can be reconstructed by numbering the DNA sequence, supposing that the constructed DNA sequence corresponds also to the temporal aspect of the DNA sequence. This means that the natural numbering of DNA bases corresponds to the temporal evolution of DNA structuring. This permit us, to use the embedding theory of Takens (Takens, 1981) for the multidimensional reconstruction of the state space underlying to the DNA dynamical process. This reconstructed state space describes the entire temporal biological evolution physical process. The DNA sequence is the one-dimensional time projection of the DNA Turbulence phase space in the form of the DNA "time series". The DNA reconstructed state space can explain the multifractal structuring of DNA sequence and mirrors the sequence of evolutionary phase transition biological process as the topological phase transition process of the biological dynamical state space topology. The DNA correlation dimension can be estimated in the reconstructed DNA state space, as well as, other useful geometrical and dynamical parameters of the biological evolution, can also be estimated (Pavlos et al., 2015; Karakatsanis et al., 2018). After all, the DNA structure mirrors also the multifractal topology of the underlying DNA turbulence state space, as well as the critical DNA phase transition process during the biological evolution of species related to the DNA construction through non-linear strange dynamics. Moreover, the self-consistently to the DNA strange dynamics maximization of Tsallis *q*-entropy, structures the DNA state space multifractal topology. According to these theoretical concepts, in this study we use the sizes of DNA regions for the reconstruction of DNA state space according to Takens embedding theory. All the estimated parameters are related with the fractal topology of DNA state space. The changes of the complexity metrics correspond to the evolutionary topological phase transition process of the DNA state space, as well as of the underlying intermittent turbulence process of the chromosomic dynamical self-organization.

Source of DNA Sequences

The Genomic compartments we used in this study and the Gene definitions are taken from National Center for Biotechnology Information (NCBI) (RefSeq Annotation Release 108, https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/108/). This data base provides both, the gene and exon definitions. Based on these definitions we generated the intronic and intergenic region coordinates. For the repeat individual we used the Repeat Masker. We then merged the repeat individual to generate the repeat merge data. Coordinates for the non-repeat sequences were the complementary to merged repeat sequences. Using both the curated and derived definitions we generated the data as shown in Figure 1.

Transparent Methods

Methodology of data analysis

The methodology of the analysis of data are supported from metrics in physical and phase space. In order to unravel the symmetries and the order of information on the distribution of the lengths of the regions in the entire genome, we used complexity theory tools for data analysis such as: a) *q*-triplet estimation, b) estimation of correlation dimension and c) estimation of Hurst exponent on these data. A new technical factor, which we name the complexity factor (COFA), and tools from machine learning (ML) algorithms are used to better describe the variation of the metrics between genomic entities, with the ultimate goal of improving our depth of understanding of the DNA system.

The dynamics of the DNA system in the phase space determines in the physical space the position of the fundamental four bases in the DNA chains in all genome entities. We understand that these positions included the necessary information for the following functions of DNA chains with an extended conclusion that the distribution of the genomic entities are not random, but it is a part of the dynamics. The information we get from a measured quantity from the physical space, is a part of the projection of the dynamics which produced this physical and measured quantity. The complexity metrics we used in the analysis reflected every time part of the dynamics in the physical or phase space. The statistics of the information and the dynamics are inextricably linked in a continuous interaction from physical space to phase space and vice versa. The dynamics produces in the phase space objects with strange geometry like strange attractors, islands, long range correlations, diffusion, multifractal behavior etc. Staying in that line of thinking, we supposed that the variations of the metrics for different entities of the whole genome corresponds to changes of the strange dynamics in the phase space, marking entities like regions, words, etc in the genome with the scope to input such information in supervised or unsupervised ML models and help us to uncover patterns and symmetries of information in the whole genome.

An integrated analysis method of DNA entities

We present in algorithmic steps the whole methodology:

a) We prepare the arithmetic or text data from DNA system. If the data are text (independence bases, words, etc) we apply specific routines: like the distance a base to the next similar base or other methods, to transform the text data in arithmetic data, else we go to the next algorithmic step. b) We apply on arithmetic data the complexity metrics like: Hurst exponent, q-triplet of Tsallis, correlation dimension, etc (other complexity metrics). c) We produce the table of results for the whole data set. The results are then used to estimate the COFA index, which will be used as an external classifier for the ML models. d) Next, we choose the attributes (Hurst, q-triplet, etc) that will be used for various ML models. e) We apply ML models for classification, clustering and prediction based on the external classifier COFA. f) We produce the table of accuracy from the previous step. If the accuracy is not acceptable, we return to step d) and we repeat the procedure until the accuracy is acceptable. g) Once the accuracy of the model is acceptable, we present the final results from ML models and extract the final symmetries and laws of information from the analysis of the whole data set.

Theoretical Framework

The DNA chromosomic system taking into account the nonlinear and strange dynamics can be described from the general equation:

$$\frac{d\vec{X}(\vec{r},t)}{dt} = F_{\lambda}(\vec{X}, W) \tag{S1}$$

where the vector $\vec{X}$ describes the state of the chromosomic chemical system, while the nonlinear function $F(\vec{X}, W)$ describes the temporal change $d\vec{X}/dt$ of the state vector. The state vector evolves temporarily in the state space of the biological evolution process. The control parameter λ describes the degree of physical connection of the DNA system with its biological and chemical environment while the quantity W corresponds to the temporal evolution of the system connection with its environment. The environment state function W can be high or low dimensional. The dimension of the DNA state vector $\vec{X}$ and the topology of the correspondent DNA state space can change according to the control parameter values. As the control parameter λ changes the profile of the dynamics of the system change also through phase transition self-organization of the entire system. Complexity theory is related with the nonlinear and strange dynamics included to the equation (S1) and the statistical character of the system evolution in the state space. The multifractal topology of the state space is created through the entropy maximization principle (Pavlos et al., 2015; Karakatsanis et al., 2018).

1. Non-extensive statistical mechanics

The non-extensive statistical theory is based mathematically on the nonlinear equation:

$$\frac{dy}{dx} = y^{q}, (y(0) = 1, q \in R \tag{S2}$$

with solution the q-exponential function such as: $e_q^x = [1 + (1 - q)x]^{\frac{1}{1-q}}$. For further characterizing the non-Gaussian character of the dynamics, we proceed to the estimation of Tsallis q-triplet based on Tsallis nonextensive statistical mechanics. Nonextensive statistical mechanics includes the q-analog (extensions) of the classical Central Limit Theorem (CLT) and α-stable distributions corresponding to dynamical statistics of globally correlated systems. The q-extension of CLT leads to the definition of statistical q-parameters of which the most significant is the q-triplet$(q_{sen}, q_{rel}, q_{stat})$, where the abbreviations sen, rel, and $stat$, stand for sensitivity (to the initial conditions), relaxation and stationary (state) in nonextensive statistics respectively (Tsallis, 2004; Umarov et al., 2008; Tsallis, 2011). These quantities characterize three physical processes: a) q-entropy production (q_{sen}), (b) relaxation process (q_{rel}), c) equilibrium fluctuations (q_{stat}). The q-triplet values characterize the attractor set of the dynamics in the phase space of the dynamics and they can change when the dynamics of the system is attracted to another attractor set of the phase space. Equation (S2) for $q = 1$ corresponds to the case of equilibrium Gaussian (Boltzmann-Gibbs (BG)) world (Tsallis, 2009). In this case, the q-triplet of Tsallis simplifies to $q_{sen} = 1, q_{stat} = 1, q_{rel} = 1$.

2. q-triplet of Tsallis theory $(\boldsymbol{q_{sen}, q_{rel}, q_{stat}})$

(a) q_{stat} *index*

A long-range-correlated meta-equilibrium non-extensive process can be described by the nonlinear differential equation (Tsallis, 2004; 2009):

$$\frac{\mathrm{d}(p_i \mathrm{Z}_{q_{stat}})}{\mathrm{dE_i}} = -\beta_{q_{stat}} \left(p_i \mathrm{Z}_{q_{stat}}\right)^{q_{stat}}, \tag{S3}$$

where *stat* stands for stationary state, and $\beta_{q_{stat}}$ is the adequate inverse temperature. The solution of this equation corresponds to the probability distribution:

$$p_i = \frac{e_{q_{stat}}^{-\beta_{q_{stat}} E_i}}{Z_{q_{stat}}} \tag{S4}$$

where $\beta_{q_{stat}} \equiv \frac{1}{KT_{stat}}$, and $Z_{q_{stat}} = \sum_j e_{q_{stat}}^{-\beta_{q_{stat}} E_j}$. Then the probability distribution is given:

$$\mathrm{p_i} \propto [1 - (1 - \mathrm{q})\beta_{q_{stat}} \mathrm{E_i}]^{1/q_{stat} - 1} \tag{S5}$$

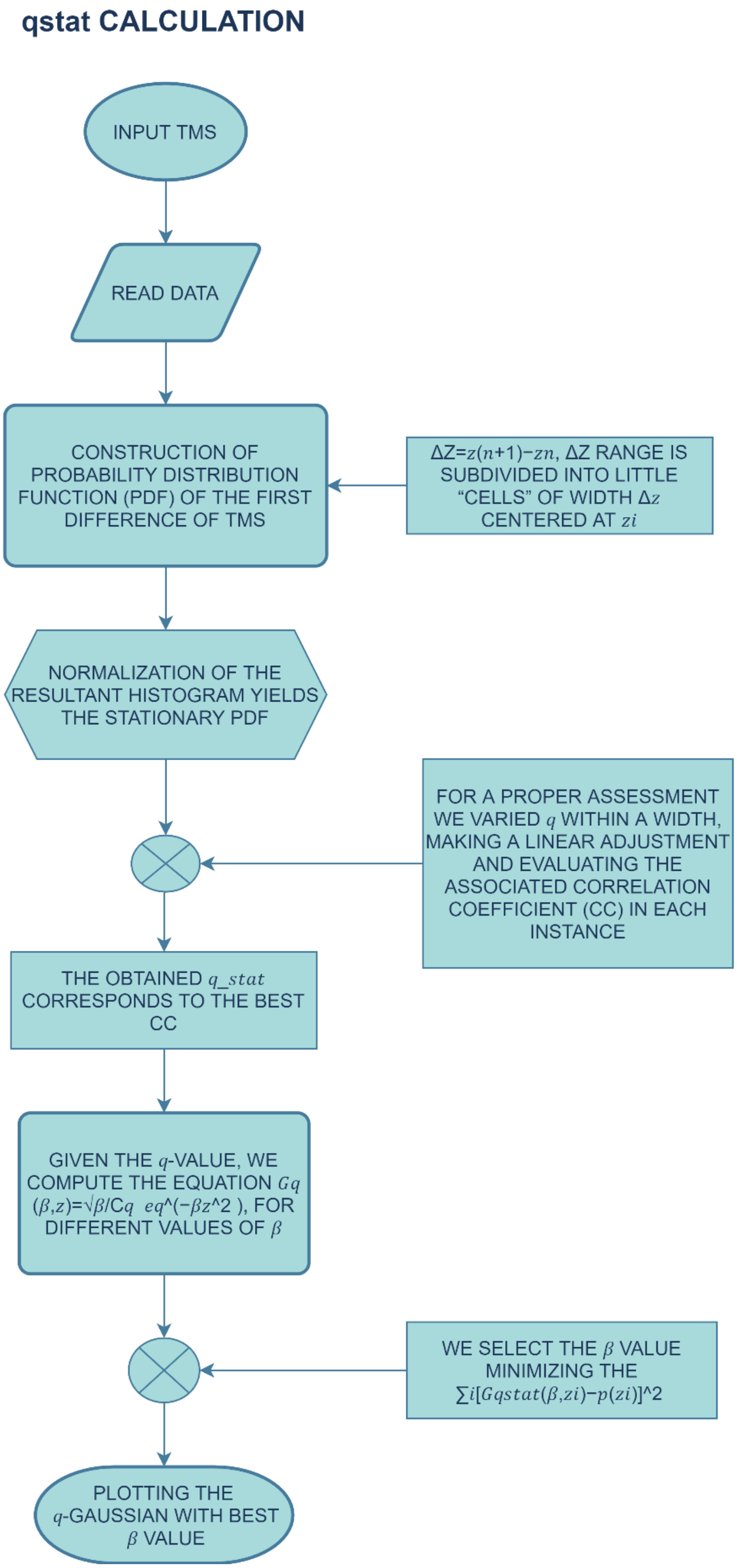
for discrete energy states $\{E_i\}$ and by

$$\mathrm{p(x)} \propto [1 - (1 - \mathrm{q})\beta_{q_{stat}} x^2]^{1/q_{stat} - 1} \tag{S6}$$

for continuous x states of $\{X\}$, where the values of the magnitude X correspond to the state points of the phase space. Distribution functions (S5) and (S6) correspond to the attracting stationary solution of the extended (anomalous) diffusion equation related to the nonlinear dynamics of the system. The stationary solutions $P(x)$ describe the probabilistic character of the dynamics on the attractor set of the phase space. The non-equilibrium dynamics can evolve on distinct attractor sets, depending upon the control parameters, while the q_{stat} exponent can change as the attractor set of the dynamics change. For the estimation of Tsallis q-Gaussian distributions we use the method described in Ferri (Ferri et al., 2010).

In the following we show the flow chart of the methodology of the q_{stat} index:

Figure s1: “The flow chart of the index q_{stat}, Related to Figure 5”


qstat CALCULATION
INPUT TMS
READ DATA
CONSTRUCTION OF PROBABILITY DISTRIBUTION FUNCTION (PDF) OF THE FIRST DIFFERENCE OF TMS
ΔZ=z(n+1)−zn, ΔZ RANGE IS SUBDIVIDED INTO LITTLE “CELLS” OF WIDTH Δz CENTERED AT zi
NORMALIZATION OF THE RESULTANT HISTOGRAM YIELDS THE STATIONARY PDF
FOR A PROPER ASSESSMENT WE VARIED q WITHIN A WIDTH, MAKING A LINEAR ADJUSTMENT AND EVALUATING THE ASSOCIATED CORRELATION COEFFICIENT (CC) IN EACH INSTANCE
THE OBTAINED q_stat CORRESPONDS TO THE BEST CC
GIVEN THE q-VALUE, WE COMPUTE THE EQUATION Gq (β,z)=√β/Cq eq^(−βz^2), FOR DIFFERENT VALUES OF β
WE SELECT THE β VALUE MINIMIZING THE ∑i[Gqstat(β,zi)−p(zi)]^2
PLOTTING THE q-GAUSSIAN WITH BEST β VALUE

(b) q_{sen} *index*

Entropy production is related to the general profile of the attractor set of the dynamics. The profile of the attractor can be described by its multi-fractality as well as by its sensitivity to initial conditions. The sensitivity to initial conditions can be expressed as:

$$\frac{d\xi}{dt} = \lambda_{q_{sen}} \xi^{q_{sen}} \tag{S7}$$

where ξ is the trajectory deviation in the phase space: $\xi \equiv \log_{\Delta x(0)\to 0} \Delta x(t)/\Delta x(0)$, where $\Delta x(t)$ is the distance between neighbouring trajectories (Tsallis, 2004). The solution of equation (S7) is given by:

$$\xi(t) = e_{q_{sen}}^{\lambda_{q_{sen}} t}, \tag{S8}$$

where *sen* stands for sensitivity.

The q_{sen}exponent is related to the multi-fractal profile of the attractor set according to

$$\frac{1}{1-q_{sen}} = \frac{1}{\alpha_{min}} - \frac{1}{\alpha_{max}}, \tag{S9}$$

where $\alpha_{min}, \alpha_{max}$ corresponds to zero points of the multi-fractal exponent spectrum $f(\alpha)$, that is $f(\alpha_{min}) = f(\alpha_{max}) = 0$. For the estimation of the multifractal spectrum we use the method described in Pavlos (Pavlos et al., 2014).

By using $D_{\bar{q}}$ spectrum we estimate the singularity spectrum $f(\alpha)$ using the Legendre transformation:

$$f(\alpha) = \bar{q}a - (\bar{q} - 1)D_{\bar{q}}, \tag{S10}$$

where $\alpha = \frac{d\tau(\bar{q})}{d\bar{q}}$. We note that the Tsallis q-entropy number is a special number corresponding to the extremization of Tsallis entropy of the system, while the $\bar{q}$ describe the range of real values of generalized dimension spectrum $D_{\bar{q}}$.

The degree of multifractality is given by:

$$\Delta a = a_{max} - a_{min} \tag{S11}$$

and the degree of asymmetry A can be estimated by the relation:

$$A = \frac{a_0 - a_{min}}{a_{max} - a_0} \tag{S12}$$

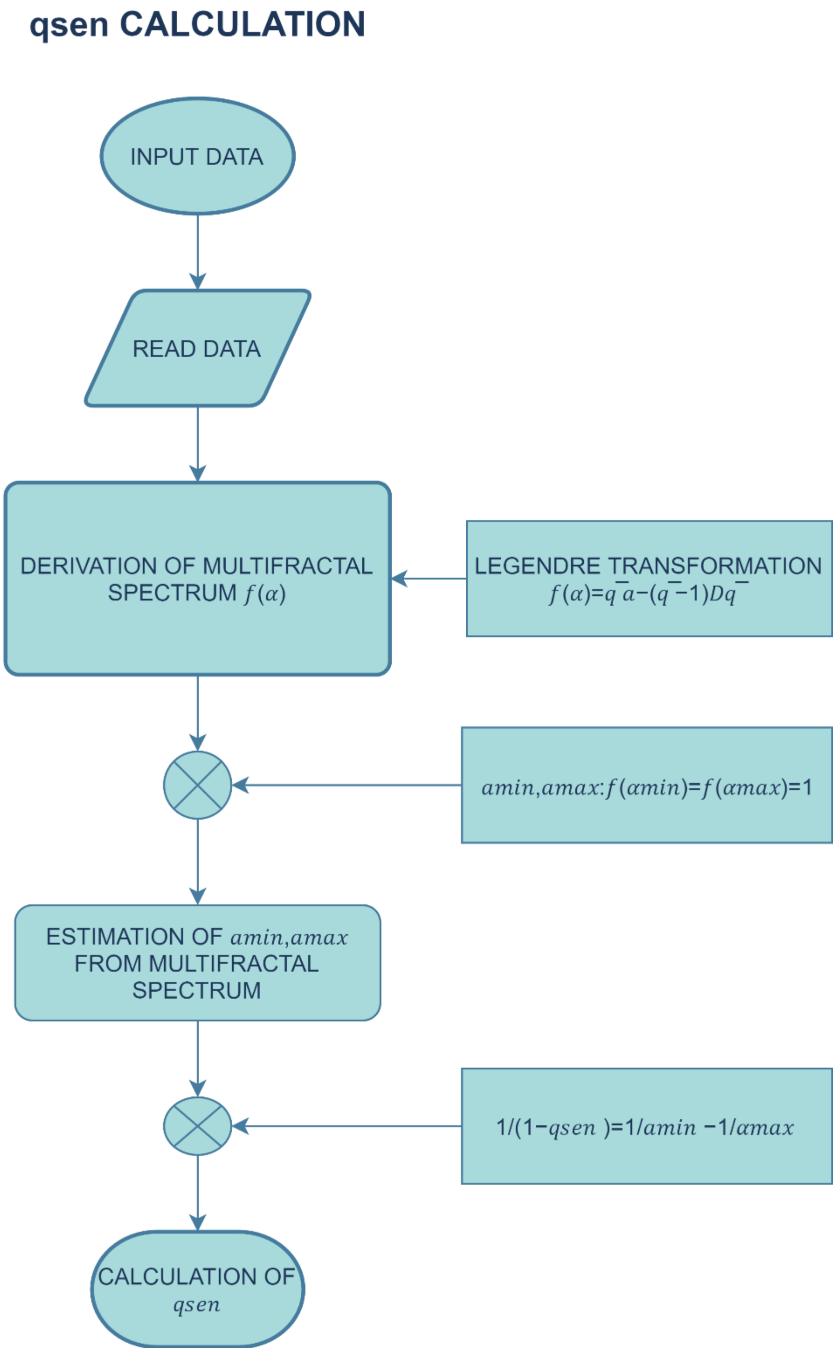
In particular, α_0 corresponds to the largest fractal dimension, which in this case is $f(\alpha) = 1$. It is important to note here that the singularity exponents α of the singularity spectrum $f(\alpha)$ corresponds to the Holder exponent and reveal the intensity of the topological singularity of the phase space as well as how irregular are the physical magnitudes defined in the phase space of the system. The value α_0, separates the values of α in two distinct intervals, $\alpha < \alpha_0$ and $\alpha > \alpha_0$ with different physical meaning. In particular, the left part of singularity spectrum $f(\alpha)$ is related with values α lower than the value α_0, and correspond to the low dimensional regions of the phase space, which is described by the right part of $D_{\bar{q}}$ spectrum. Similarly, the right part of the singularity spectrum $f(\alpha)$ is related with values α higher than α_0 and correspond to the high dimensional regions of the phase space, which is described by the left part of the curve $D_{\bar{q}}$ of the generalized dimension spectrum.

According to these characteristics of $f(\alpha)$ and $D_{\bar{q}}$ spectra, the high dimensional regions of phase space includes smoother fractal topology than the low dimensional regions, where the fractal character is stronger.

Low dimensional regions of phase space cause strong fractional acceleration and anomalous diffusion processes of the experimental TMS. The estimation of $\Delta D_{\bar{q}}$ between the low $(\bar{q} \rightarrow +\infty)$ and high $(\bar{q} \rightarrow -\infty)$ dimensional regions of the phase space reveals the multifractal behavior of the system. High (Low) values of $\Delta D_{\bar{q}}$ shows strong (weak) multifractality.

In the following we show the flow chart of the methodology of the q_{sen} index:

Figure s2: "The flow chart of the index q_{sen}, Related to Figure 7"


qsen CALCULATION
INPUT DATA
READ DATA
DERIVATION OF MULTIFRACTAL SPECTRUM $f(\alpha)$
LEGENDRE TRANSFORMATION $f(\alpha)=\bar{q}a-(\bar{q}-1)D\bar{q}$
$amin,amax$:$f(\alpha min)=f(\alpha max)=1$
ESTIMATION OF $amin,amax$ FROM MULTIFRACTAL SPECTRUM
$1/(1-qsen)=1/amin-1/\alpha max$
CALCULATION OF $qsen$

(c) q_{rel} index

Thermodynamic fluctuation-dissipation theory is based on the Einstein original diffusion theory (Brownian motion theory). Diffusion is a physical mechanism for extremization of entropy. The Einstein-Smoluchowski theory of Brownian motion was extended to the general Fokker Planck (FP) diffusion theory of non-equilibrium processes. The potential of FP equation may include many meta-equilibrium stationary states near or far away from thermodynamical equilibrium. Macroscopically, relaxation to the equilibrium stationary state of some dynamical observable $O(t)$ related to system evolution in the phase space can be described by the form of general form:

$$\frac{\mathrm{d}\Omega}{\mathrm{dt}} = -\frac{1}{\tau}\Omega, \tag{S13}$$

where $\Omega(t) \equiv [O(t) - O(\infty)]/[O(0) - O(\infty)]$ describes the relaxation of the macroscopic observable $O(t)$ towards its stationary state value and τ being the relaxation time (Tsallis, 2004). The non-extensive generalization of fluctuation-dissipation theory is related to the general correlated anomalous diffusion processes (Tsallis, 2009). The equilibrium relaxation process (S13) is transformed to the meta-equilibrium non-extensive relaxation process according to:

$$\frac{d\Omega}{dt} = -\frac{1}{\tau_{q_{rel}}}\Omega^{q_{rel}}, \tag{S14}$$

where *rel* stands for relaxation.

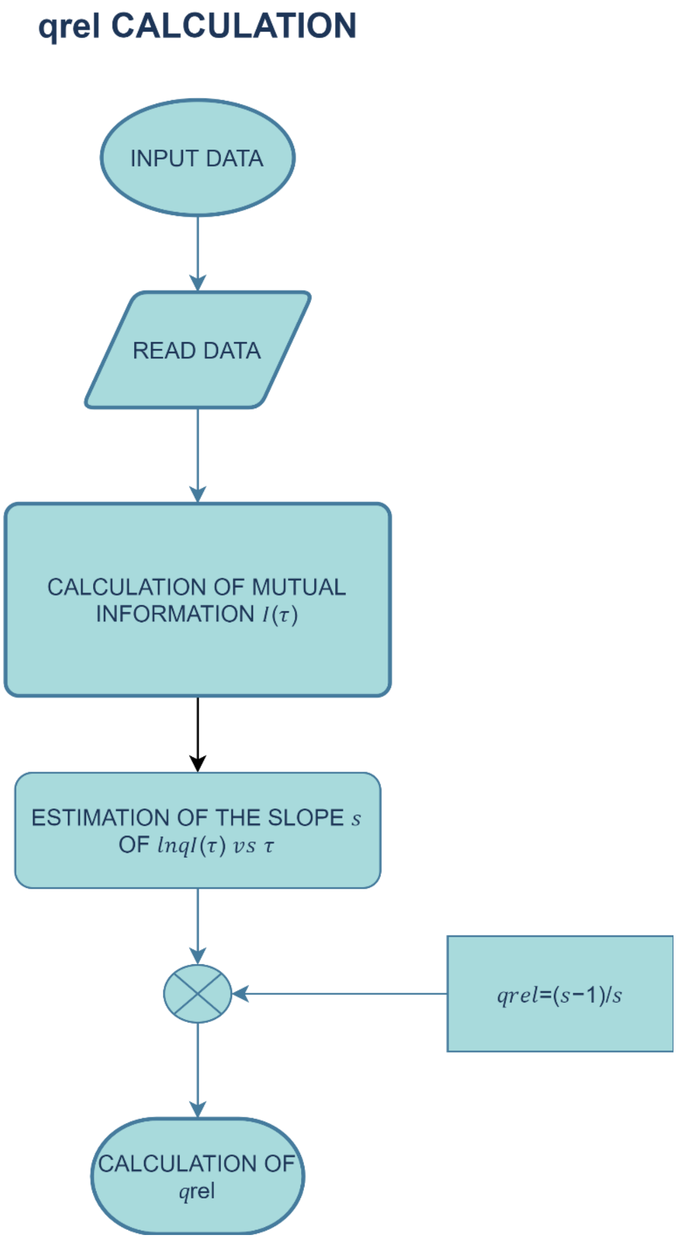
The solution of this equation is given by:

$$\Omega(\mathrm{t}) = \mathrm{e}_{\mathrm{q_{rel}}}^{-\mathrm{t}/\tau_{q_{rel}}} \tag{S15}$$

The autocorrelation function $C(t)$ or the mutual information $I(t)$ can be used as candidate observables $\Omega(t)$ for estimation of q_{rel}. However, in contrast to the linear profile of the correlation function, the mutual information includes the nonlinearity of the underlying dynamics and it is proposed as a more faithful index of the relaxation process and the estimation of the Tsallis exponent q_{rel}.

In the following we show the flow chart of the methodology of the q_{rel} index:

Figure s3: "The flow chart of the index q_{rel}, Related to Figure 6"


qrel CALCULATION
INPUT DATA
READ DATA
CALCULATION OF MUTUAL INFORMATION $I(\tau)$
ESTIMATION OF THE SLOPE s OF $lnqI(\tau)$ vs τ
$qrel$=(s−1)/s
CALCULATION OF $q$rel

3. Correlation Dimension $(\boldsymbol{D_2})$

In order to provide information for the dynamical degrees of freedom of the dynamics underlying the experimental time series we estimate the correlation dimension (D_2) defined as:

$$D_2 = \lim_{r \to 0} \frac{d[\ln C(r)]}{d[\ln(r)]} \quad \text{(S16)}$$

where C(r) is the so-called correlation integral for a radius r in the reconstructed phase space. When an attracting set exists then C(r) reveals a scaling profile:

$$C(r) \sim r^d \text{ for } r \rightarrow 0. \quad \text{(S17)}$$

The correlation integral depends on the embedding dimension m of the reconstructed phase space and is given by the following relation:

$$C(r,m) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1+1}^{N} \Theta(r - \|x(i) - x(j)\|) \quad \text{(S18)}$$

where Θ(a)=1 if a>0 and Θ(a)=0 if a ≤ 1, and N is the length of the time series. The low value saturation of the slopes of the correlation integrals is related to the number (d) of fundamental degrees of freedom of the internal dynamics. For the estimation of the correlation integral we used the method of Theiler (Theiler, 1991) in order to exclude time correlated states in the correlation integral estimation, thus discriminating between the dynamical character of the correlation integral scaling and the low value saturation of slopes characterizing self-affinity (or crinkliness) of trajectories in a Brownian process. When the dynamics possesses a finite (small) number of degrees of freedom, we can observe saturation to low values D_2 of the slopes D_m for a sufficiently large embedding m. The dimension of the attractor of the dynamics is then at least the smallest integer D_0 larger than D_2 or at most $2D_0+1$, according to Taken's theorem (Takens, 1981).

4. Hurst Exponent ($\boldsymbol{h}$)

The Hurst exponent (h) related to the fractal dimension (D). The relationship between the fractal dimension and the Hurst exponent is:

$$D = 2 - h \quad \text{(S19)}$$

The fractal dimension shows how rough a surface is. A small value of Hurst exponent shows a higher fractal dimension and a rougher surface. A larger Hurst exponent shows a smaller fractional dimension and a smoother surface. The values of the Hurst exponent range between 0 and 1. A value of 0.5 indicates a true random process (a Brownian time series). A Hurst exponent value $h, 0.5 < h < 1$ indicates "persistent behavior". Here an increase (decrease) probably followed by an increase (decrease). A Hurst exponent value $0 < h < 0.5$ indicates "anti-persistent behavior". Here an increase (decrease) probably followed by a decrease (increase). For the estimation of the Hurst exponent (h) in this study we use Rescaled Range Analysis (R/S) (Weron, 2002). The Hurst exponent (h), is defined in terms of the asymptotic behavior of the rescaled range (R/S) as a function of the time span of a time series as follows:

$$E\left[\frac{R(n)}{S(n)}\right] = Cn^h, n \rightarrow \infty, \quad \text{(S20)}$$

where, $R(n)$ is the range of the first n values and $S(n)$ is their standard deviation, $E(x)$ is the expected value, n is a number of data points in a time series and C is a constant.

5. Machine Learning Analysis

(a) Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

(b) k-means model

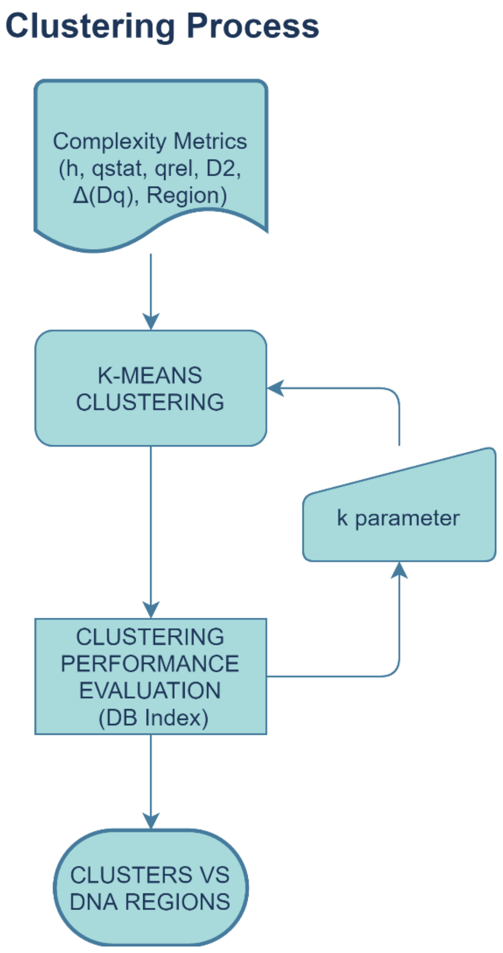
k-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Is a method of vector quantization, originally from signal processing. k-means clustering aims to partition n observations (Examples) into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Clustering can be used on unlabeled data.

The k-means algorithm determines a set of k clusters and assigns each Examples to exact one cluster. The clusters consist of similar Examples. The similarity between Examples is based on a distance measure between them. A cluster in the k-means algorithm is determined by the position of the center in the n-dimensional space of the n Attributes of the Example Set. This position is called centroid. It can, but do not have to be the position of an Example of the Example Sets. The k-means algorithm starts with k points which are treated as the centroid of k potential clusters. All Examples are assigned to their nearest cluster (nearest is defined by the measure type). Next the centroids of the clusters are recalculated by averaging over all Examples of one cluster. The previous steps are repeated for the new centroids until the centroids no longer move or max optimization steps is reached. The procedure is repeated max runs times with each time a different set of start points. The set of clusters is delivered which has the minimal sum of squared distances of all examples to their corresponding centroids. The objective function for the k-means clustering algorithm is the squared error function:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(\lVert x_i - u_j \rVert \right)^2 = 1, \qquad \text{(S21)}$$

where $\lVert x_i - u_j \rVert$ is the Euclidean distance between a point, x_i and a centroid, u_j , iterated over all k points in the i^{th} cluster, for all n clusters. In simpler terms, the objective function attempts to pick centroids that minimize the distance to all points belonging to its respective cluster so that the centroids are more symbolic of the surrounding cluster of data points. K-means clustering is a fast, robust, and simple algorithm that gives reliable results when data sets are distinct or well separated from each other in a linear fashion. It is important to keep in mind that k-means clustering may not perform well if it contains heavily overlapping data, if the Euclidean distance does not measure the underlying factors well, or if the data is noisy or full of outliers.
In the following we show the flow chart for the clustering method using in this study:

Figure s4: “The flow chart of the clustering process, Related to Figures 10-12”


Clustering Process
Complexity Metrics (h, qstat, qrel, D2, Δ(Dq), Region)
K-MEANS CLUSTERING
k parameter
CLUSTERING PERFORMANCE EVALUATION (DB Index)
CLUSTERS VS DNA REGIONS

(c) Supervised classification (Naive Bayes classifier)

For this work we used the Naive Bayes classifier for the classification process. Naive Bayes is a high-bias, low-variance classifier, and it can build a good model even with a small data set. It is simple to use and computationally inexpensive. Typical use cases involve text categorization, including spam detection, sentiment analysis, and recommender systems. The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naive Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical. Given a set of variables, $X = \{x_1, x_2, \dots, x_d\}$, we want to construct the posterior probability for the event Cj among a set of possible outcomes $C = \{c_1, c_2, \dots, c_d\}$. In a more familiar language, X is the predictors and C is the set of categorical levels present in the dependent variable. Using Bayes' rule:

$$p\left(C_j \vee x_1, x_2, \dots, x_d\right) \propto p\left(x_1, x_2, \dots, x_d \vee C_j\right) p\left(C_j\right), \tag{S22}$$

where $p\left(C_j \vee x_1, x_2, \dots, x_d\right)$ is the posterior probability of class membership, i.e., the probability that X belongs to C_j. Since Naive Bayes assumes that the conditional probabilities of the independent variables are statistically independent, we can decompose the likelihood to a product of terms:

$$p\left(X \vee C_j\right) \propto \prod_{k=1}^{d} p\left(x_k \vee C_j\right) \tag{S23}$$

and rewrite the posterior as:

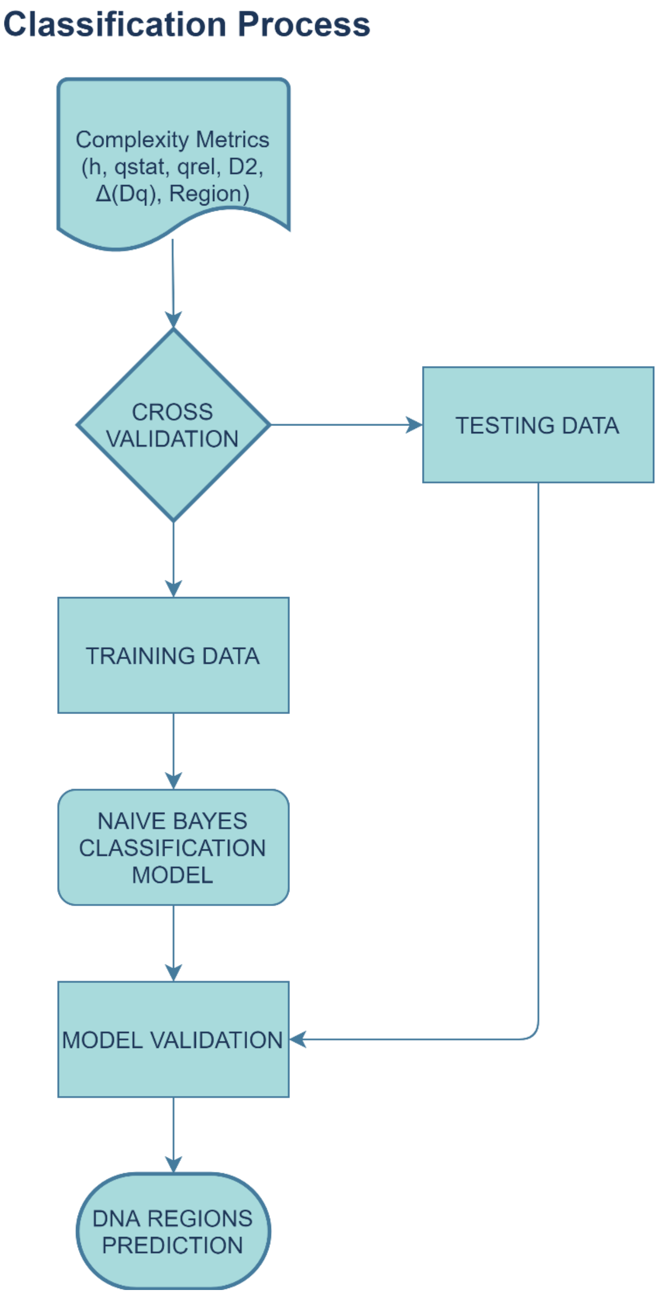
$$p\left(C_j \vee X\right) \propto p\left(C_j\right) \prod_{k=1}^{d} p\left(x_k \vee C_j\right) \tag{S24}$$

Using Bayes' rule above, we label a new case X with a class level C_j that achieves the highest posterior probability.

Although the assumption that the predictor (independent) variables are independent is not always accurate, it does simplify the classification task dramatically, since it allows the class conditional densities $\mathrm{p}\left(\mathrm{x}_\mathrm{k} \vee \mathrm{C}_\mathrm{j}\right)$ to be calculated separately for each variable, i.e., it reduces a multidimensional task to a number of one-dimensional ones. In effect, Naive Bayes reduces a high-dimensional density estimation task to a one-dimensional kernel density estimation. Furthermore, the assumption does not seem to greatly affect the posterior probabilities, especially in regions near decision boundaries, thus, leaving the classification task unaffected.

In the following we show the flow chart for the classification method using in this study:

Figure s5: “The flow chart of the classification process, Related to Figures 15-15”



6. Evaluation - Split Test

For the classifier’s evaluation we used a 60/40 train/test set split. The split of the dataset is a simple way to use one dataset to both train and estimate the performance of the classifier. We split the dataset into a training dataset and a test dataset. Our model randomly selects 60% of the instances for training and use the remaining 40% as a test dataset.

References

Broomhead, D.S., and King, G.P. (1986). Extracting qualitative dynamics from experimental data. Physica D: Nonlinear Phenomena 20, 217–236.

Davies, P. (2004). The Cosmic Blueprint (Templeton Foundation Press).

Ferri, G.L., Reynoso Savio, M.F., and Plastino, A. (2010). Tsallis' *q*-triplet and the ozone layer. Physica A: Statistical Mechanics and Its Applications 389, 1829–1833.

Karakatsanis, L.P., Pavlos, G.P., Iliopoulos, A.C., Pavlos, E.G., Clark, P.M., Duke, J.L., and Monos, D.S. (2018). Assessing information content and interactive relationships of subgenomic DNA sequences of the MHC using complexity theory approaches based on the non-extensive statistical mechanics. Physica A: Statistical Mechanics and its Applications 505, 77-93.

Nicolis, G., and Prigogine, I. (1989). Exploring Complexity: An Introduction (Freeman, W.H., New York).

Nicolis, G. (1993). Physics of far-from-equilibrium systems and self-organization. In The new physics, P. Davies, ed. (Cambridge University Press), pp. 316-347.

Pavlos, G.P., Karakatsanis, L.P., Xenakis, M.N., Pavlos, E.G., Iliopoulos, A.C., and Sarafopoulos, D.V. (2014). Universality of non-extensive Tsallis statistics and time series analysis: Theory and applications. Physica A: Statistical Mechanics and Its Applications 395, 58–95.

Pavlos, G.P., Karakatsanis, L.P., Iliopoulos, A.C., Pavlos, E.G., Xenakis, M.N., Clark, P., Duke, J., and Monos, D.S. (2015). Measuring complexity, nonextensivity and chaos in the DNA sequence of the Major Histocompatibility Complex. Physica A: Statistical Mechanics and Its Applications 438, 188–209.

Prigogine, I. (1978). Time, structure, and fluctuations. Science 201, 777-785.

Prigogine, I. (1997). The end of certainty: Time, Chaos, and the New Laws of Nature (The Free Press).

Takens, F. (1981). Detecting strange attractors in turbulence. In Dynamical Systems and Turbulence. D. Rand, L.S. Young, eds. (Springer), pp. 366-381.

Tsallis, C. (2002). Entropic nonextensivity: A possible measure of complexity. Chaos, Solitons and Fractals 13, 371-391.

Tsallis, C. (2004). Dynamical scenario for nonextensive statistical mechanics. In Physica A: Statistical Mechanics and its Applications 340, 1–10.

Tsallis, C. (2009). Introduction to Nonextensive Statistical Mechanics: Approaching a complex world (Springer).

Tsallis, C. (2011). The Nonadditive Entropy Sq and Its Applications in Physics and Elsewhere: Some Remarks. Entropy 13, 1765–1804.

Umarov, S., Tsallis, C., and Steinberg, S. (2008). On a q-central limit theorem consistent with nonextensive statistical mechanics. Milan Journal of Mathematics 76, 307–328.