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:

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

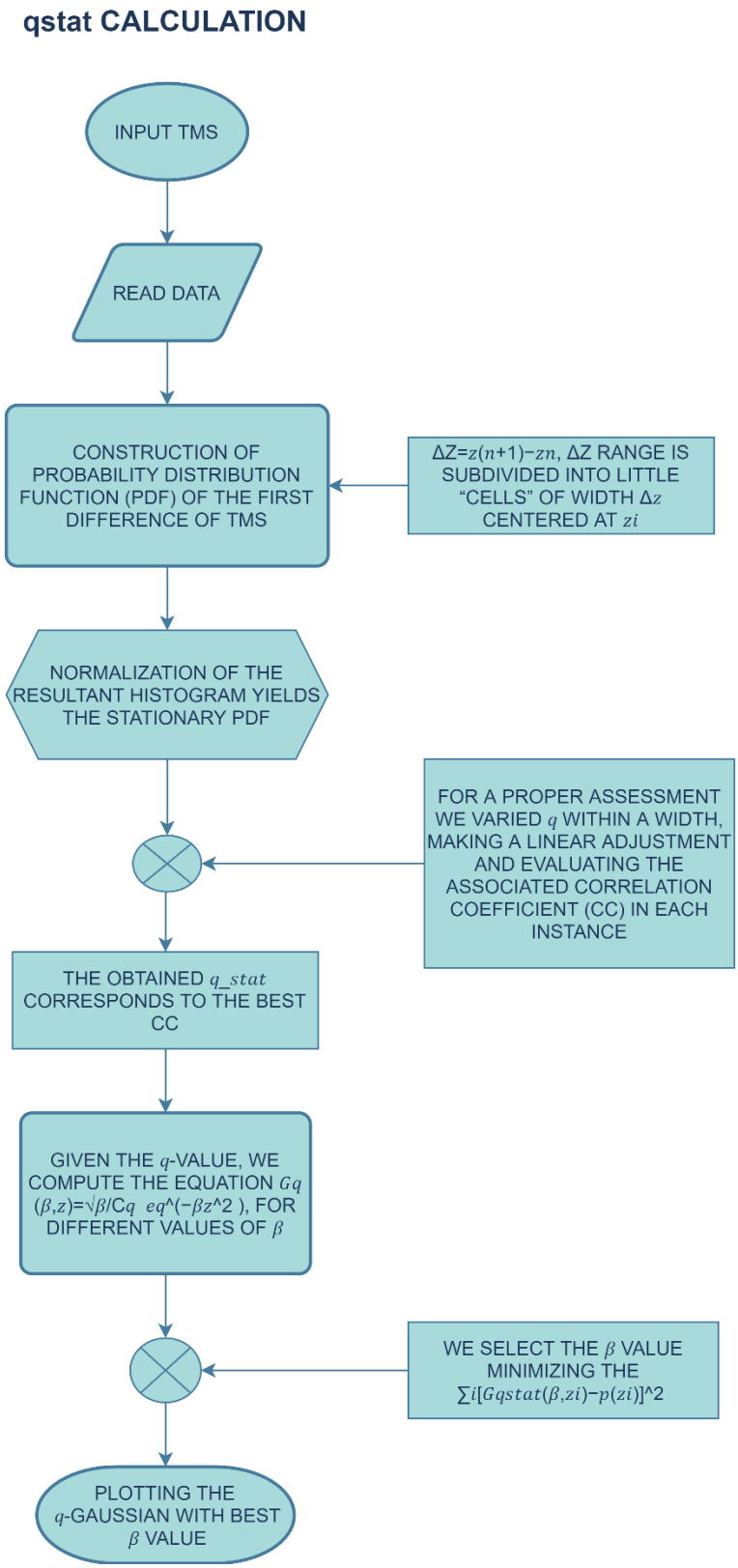
for discrete energy states $\{E_i\}$ and by

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

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”



(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:

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

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:

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

where *sen* stands for sensitivity.

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

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

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:

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

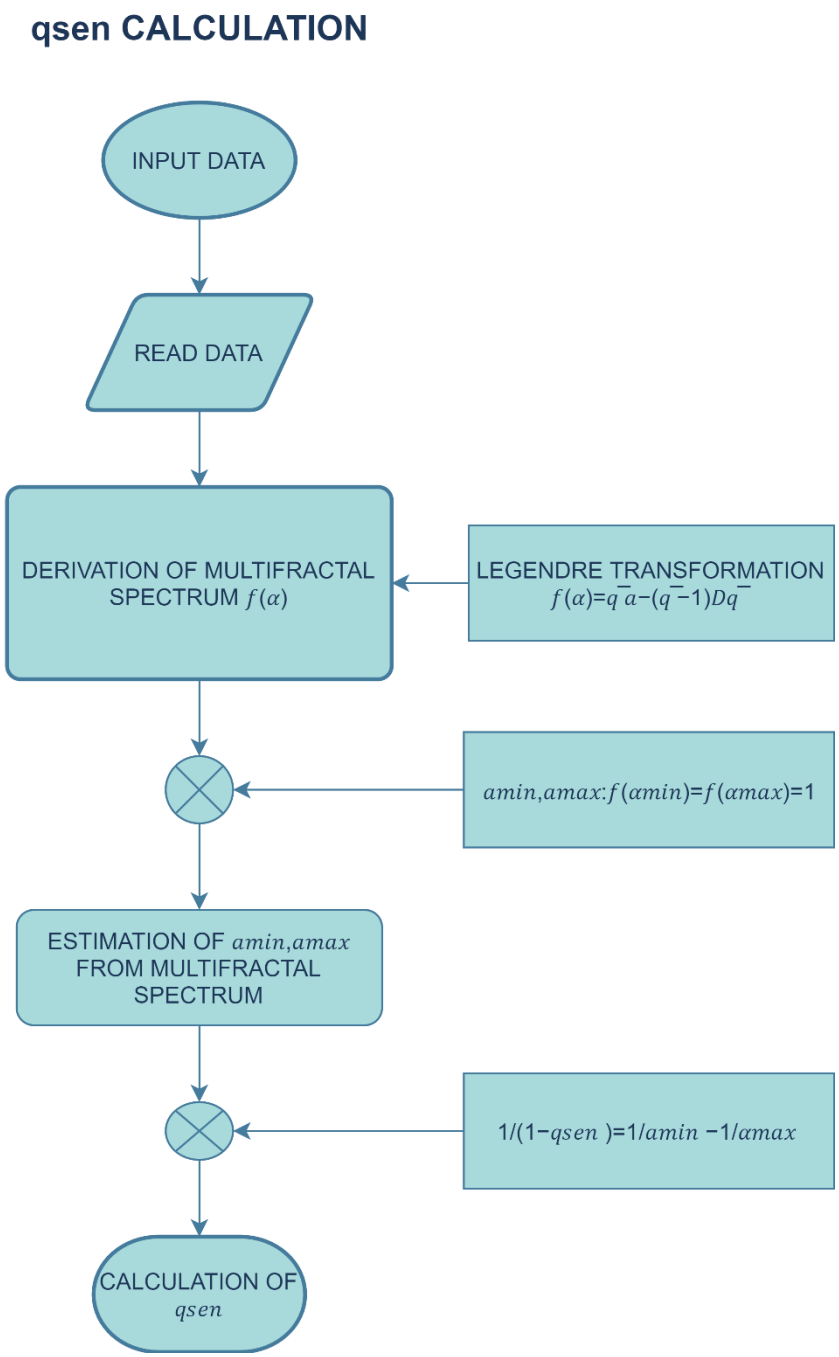
In particular, α_0 corresponds to the largest fractal dimension, which in this case is $f(\alpha) = 1$. It is important to note here that the singularity exponents α of the singularity spectrum $f(\alpha)$ corresponds to the Holder exponent and reveal the intensity of the topological singularity of the 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 $(\bar{q} \to +\infty)$ and high $(\bar{q} \to -\infty)$ 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"



(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.

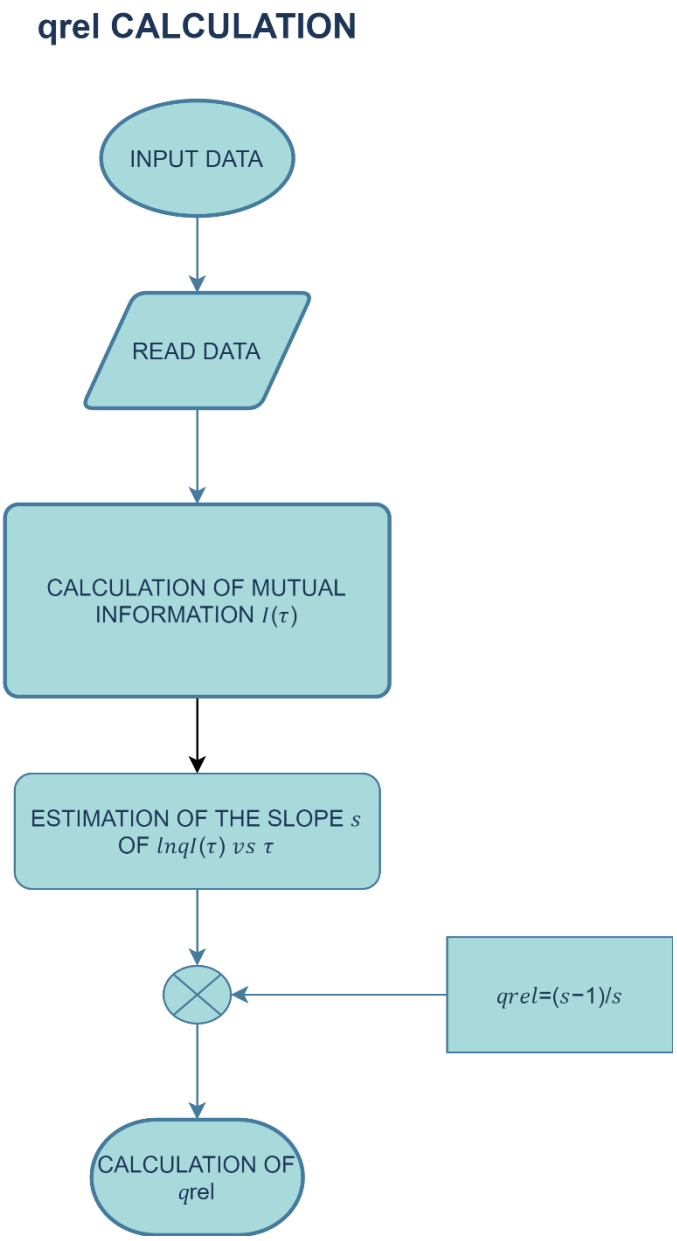
The solution of this equation is given by:

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

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”



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:

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

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 $\Theta(a)=1$ if $a>0$ and $\Theta(a)=0$ if $a \leq 1$, 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:

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

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.

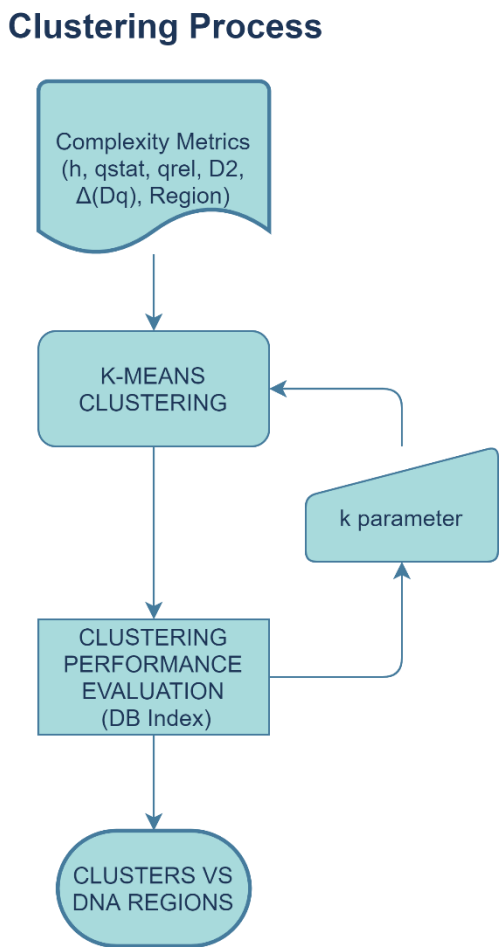
The k-means algorithm determines a set of k clusters and assigns each Examples to exact one cluster. The clusters consist of similar Examples. The similarity between Examples is based on a distance measure between them. A cluster in the k-means algorithm is determined by the position of the center in 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:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(\|x_i - u_j\| \right)^2 = 1, \tag{S21}$$

where $\|x_i - u_j\|$ 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”



(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:

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

where $p(C_j \vee x_1, x_2, \dots, x_d)$ 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:

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

and rewrite the posterior as:

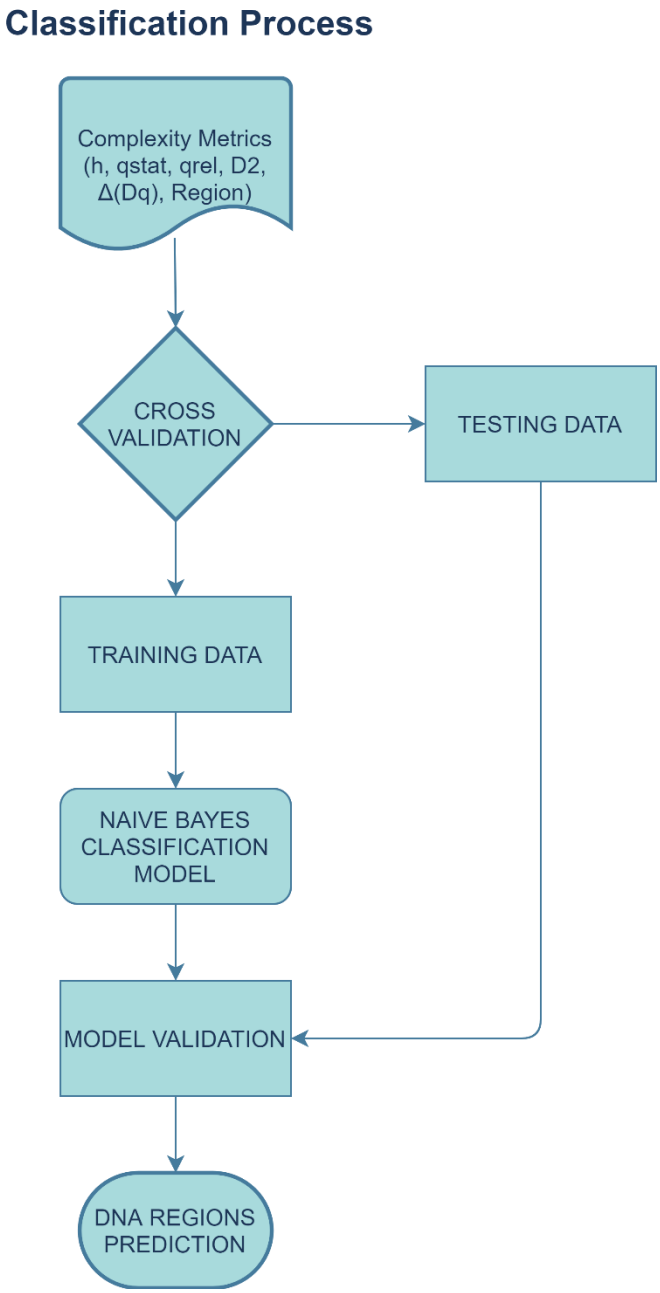
$$p(C_j \vee X) \propto p(C_j) \prod_{k=1}^{d} p(x_k \vee C_j) \tag{S24}$$

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 $p(x_k \vee C_j)$ 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.