Nonlinear Dyn
https://doi.org/10.1007/s11071-018-4245-7

ORIGINAL PAPER

Kolmogorov complexity as a data similarity metric: application in mitochondrial DNA

Rómulo Antão · Alexandre Mota · J. A. Tenreiro Machado

Received: 24 October 2017 / Accepted: 27 March 2018

Abstract The problem of developing a similarity index for different objects is discussed. The limitations of current metrics are evaluated and discussed. The normalized compression distance, based on the non-computable Kolmogorov complexity, is examined and compared with two alternative measures. A case study consisting of a phylogenetic tree of different mammals is constructed applying this technique with a mitochondrial DNA database.

Keywords Kolmogorov complexity · Normalized compression distance · Mitochondrial DNA

1 Introduction

Comparing and clustering information is an innate feature of human beings and is performed with considerable success when a set of particular characteristics are identified and used to define a similarity index. However, with the advent of the digital era, it became clear that this task could be optimized and automated. Therefore, it is necessary to define measures for comparing the digital data representations of objects. Mathematical modeling provided the concepts and tools to formalize such metrics, and a new discipline of information studies took the first steps by the hand of Shannon [45], the information theory. In the framework of the seminal concept of entropy, Shannon formalized also the notion of mutual information between two objects, for measuring their independence. One distinctive feature of the information theory is that it ignores the objects' nature, since it is unrelated to their meaning, structure or content. Shannon formalization takes only in consideration the probability distribution of each object on the set of possible outcomes from a source that may be unknown for the observer, or may not exist at all [18]. This limitation kept open the challenge of seeking for an universal distance index that could be used without any *a priori* or estimated knowledge about the objects under study.

Data mining [25], machine learning [38] and clustering algorithms [11] are some examples of application domains that require a precise method to evaluate, distinguish and quantify the relationship among objects. An approach often adopted is to find a set of classification variables that allow the characterization of objects according to some feature. However, such methodology requires domain-specific knowledge about the

R. Antão · A. Mota
Department of Electronics, Telecommunications and Informatics, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
e-mail: romuloantao@ua.pt

A. Mota
e-mail: alex@ua.pt

J. A. T. Machado (✉)
Department Electrical Engineering, Institute of Engineering, Polytechnic of Porto, Rua Dr. António Bernardino de Almeida, 431, 4249-015 Porto, Portugal
e-mail: jtm@isep.ipp.pt

Published online: 06 April 2018

objects, since the choice of classification variables is strongly limited by our subjective interpretation. Consequently, we rely on numerical formalizations that are meaningful and systematic. An example of this high-level feature representation is the classification of DNA sequences [8], using Shannon block entropy for extracting frequency statistics of sub-strings composed by the four nucleotide bases in the whole genome sequence (adenine (A), thymine (T), cytosine (C) and guanine (G)). Other examples are the extraction of information like the pitch, harmony or *beats-per-second* of music files using the Fourier transform [23,36], or the representation of the colors under maps as RGB or YUV [30]. Data classification can then be performed using distance measuring algorithms, depending on the nature of the data points.

An universally metric not requiring preprocessing of the information must be sought to reduce the time-consuming parameter optimization process. Also, we must have in mind increasing the sensitivity of the algorithm to specific features while avoiding over-fitting a particular scenario. One answer to this problem came, late in 2004, when Paul Vitányi and Ming Li proposed *The Similarity Metric* with the normalized compression distance (NCD) [30]. The paper discusses the use of compression algorithms to define a new universal, robust and parameter-free similarity index, representing a promising breakthrough in the similarity metrics paradigm. This approach is deeply inspired in the normalized information distance (NID), a technique firstly presented in [5]. However, the NID is non-computable because it is based on the Kolmogorov complexity theory. As it was proved [11,30], the Kolmogorov complexity can be fairly approximated using standard compression algorithms. Therefore, since the NCD does not use background knowledge of the input data, it does not evaluate specific features and is (up to some point) independent of the adopted compression algorithm. In this line of thought, the NCD can be, in theory, fairly applied to any type of digital data.

Following these ideas the paper is organized as follows. In Sect. 2 the problem of data similarity is introduced. The concept of distance measuring is presented, and some common approaches found in the literature with their downsides and limitations are discussed. In Sect. 3 the ideas underlying an universal distance metric, with focus on the NCD, are formulated. In Sect. 4 an example of application of this method, the phylogenetic tree of 18 mammals is constructed and analyzed. The NCD is compared with two distinct measures. Finally, in Sect. 5 the main conclusions are drawn.

2 Distance metrics problems

Assessing the similarity degree between several features is a procedure that requires the use of a given distance metric to establish comparison standpoints. According to [11], a distance function D, evaluated over two objects x and y can be considered a metric if it returns a nonnegative real value and satisfies the following conditions:

$$\begin{aligned} &\mathcal{C}_1: \; D(x,y) = 0 \; if \; x = y \quad \text{(identity axiom)} \\ &\mathcal{C}_2: \; D(x,y) = D(y,x) \quad \text{(symmetry axiom)} \\ &\mathcal{C}_3: \; D(x,y) \leq D(x,z) + D(z,y) \\ &\qquad \text{(triangle inequality)} . \end{aligned} \tag{1}$$

In the literature we find several different metrics that can be applied to data sets and provide meaningful results for our interpretation. When considering numerical vectors $(x_1, \ldots, x_n)^T \in \mathbb{R}^n$, the L_1 and L_2 norms:

$$\begin{aligned} L_1: \; &\sum_{k=1}^{n} |x_k|, \\ L_2: \; &\sum_{k=1}^{n} x_k^2, \end{aligned} \tag{2}$$

representing specific instances of the Minkowski distance, are often used [13]. In the particular case of DNA analysis, the L_2 norm supports many different algorithms [7,17,27] that, with the application of the Parseval theorem, enable a straightforward comparison of sequences by manipulating their data in the frequency domain. However, instead of measuring distance between individual points, we can adopt a collection of data points as a population and measure the distance between them using the Chi-square distance [21], evaluate categorical data points using the Hamming distance [20] to check how many attributes must be changed to match one another, or compare sequences using the edit distance [4] that evinces how many modifications (such as character insert, modify or delete) are needed to change one string into another.

There is currently some consensus regarding the application of several different metrics in specific domains of application, but the truth is that the choice of the optimal distance metric still poses relevant challenges [52], namely:

卡方分布：k个独立的标准正态分布变量的平方和服从自由度为k的卡方分布。
汉明距离（英语：Hamming distance）是两个字符串对应位置的不同字符的个数。换句话说，它就是将一个字符串变换成另外一个字符串所需要替换的字符个数

- Problems related to the probability distributions of the data to be evaluated. While, for example, it is proven that Gaussian and exponential distributions are better evaluated using the L_2 and L_1 norms [19], respectively, there exist many other distributions and the corresponding optimal metrics are unknown.
- Correlation between the similarity level given by a metric and our own concept of observed similarity. In the particular case of image comparisons, the results obtained with L_1 and L_2 metrics were compared with the opinion of several humans [43]. The conclusions of the study revealed that the L_1 norm may better capture human notion of similarity, but still there is no clear superiority over L_2 since the results vary from case to case.
- Relative comparisons between data, as required in clustering algorithms. Many measures may not give comparable results since they are unbounded, and their value is independent of the size of objects under evaluation. For example, the distance between two objects, using the L_2 norm, of 10,000 elements that differ only in 100 points has the same distance as two sets of 1000 elements each that differ also in 100 points. Even though, the data sets in the first case have clearly an higher similarity than the second one. To overcome this representation problem, normalized versions of comparative metrics have to be used, as is discussed in [51].

Due to all these constraints, assessing the similarity of several objects is not a straightforward process. Nonetheless, as it was discussed in Sect. 1, we can find a non-probabilistic information measure that provides a distinct perspective toward object similarity assessment.

3 Background on the Kolmogorov complexity and information distance metric

The Kolmogorov complexity, also known as algorithmic entropy, provides a measure of information that, unlike Shannon's, does not rely on (often untenable) probabilistic assumptions of the data sequences. Contrary to Shannon's theory, the information measurement is focused on an individual finite object described by a string and accounts with the phenomenon that 'regular' strings are compressible [26]. Represented in the literature as $K(x)$, the Kolmogorov complexity of a string x can be defined as the length of the shortest binary program that, given an empty string ψ at its input, can compute x on its output in an universal computer and then halts. Loosely speaking, the Kolmogorov complexity of a file can be interpreted as the length of its ultimate compressed version.

Based on this concept, the information distance of two strings (or files) $\{x, y\} \in \Sigma$ can also be computed by means of the conditional Kolmogorov complexity $K(x|y)$ [5,15]. This formulation can be read as the length of the shortest program to compute x if y is provided as an auxiliary input. Intuitively, the more similar the two strings the less complex this task should be and, therefore, the smaller the size of the shortest program able to perform the algorithm. Therefore, the following inequality always holds

$$K(x|y) \leq K(x). \tag{3}$$

Built upon this concept, an universal distance metric was formulated [5]

$$\mathrm{NID}(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}, \quad \{x, y\} \in \Sigma, \tag{4}$$

and denoted as normalized information distance (NID).

Considering equation (3), the NID may only take values in the range [0, 1], with $NID(x, x) \approx 0$ and $\mathrm{NID}(x, \psi) \approx 1$, where ψ is an empty object that has no similarity to x. In [5], its is shown that the NID is also a distance since it satisfies the measure inequalities defined in (1), up to some additive precision. However, being based in the Kolmogorov complexity, this measure is as well non-computable, as is proved by contradiction in [5]. Nonetheless, the NID served as a solid foreground for a new computable distance metric, the normalized compression distance (NCD), proposed in [30]. Its computability comes with the cost of a fair approximation of the Kolmogorov complexity by a standard compressor $C(\cdot)$. For the sake of parsimony, the demonstration of the equivalence between the NID and the NCD is not presented here (see the mathematical details in [11]).

The formalization of the NCD is given by:

$$\mathrm{NCD}(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}. \tag{5}$$

The NCD is a nonnegative number, with $0 < \mathrm{NCD}(x, y) < 1 + \epsilon$, and represents the distance between the two files x and y. The smaller the result, the more similar the files. The parameter $\epsilon > 0$ in the

upper bound is related to 'imperfections' in the used compression techniques. The values of $C(x)$ and $C(y)$ are the sizes in bits of the individually compressed files x and y, respectively. In addition, $C(xy)$ is the compressed size of the two concatenated files x and y, took by the compressor as a single file.

Expression (5) can be easily interpreted assuming, for example, that $C(y) \geq C(x)$. The distance $\mathrm{NCD}(x, y)$ assesses the improvement due to compressing y using data from the previously compressed x, and compressing y from scratch, expressed as the ratio between their compressed sizes in bits.

Approximating the NID by the NCD poses several operational problems. Firstly, due to the non-computability of the Kolmogorov complexity, we cannot say how far is the NCD from the real value of NID [9]. Furthermore, such approximation may yield arithmetic problems, particularly in the case of small size strings where numerical indeterminate forms may arise [11]. The effect of string size is discussed in detail in [10].

Other restriction imposed to compression algorithms as an approximation of the Kolmogorov complexity is that they must be 'normal' compressors, so that the NCD satisfies the distance inequalities [11]. By 'normal' compressor is understood a technique C that given the object XX (i.e., X concatenated with himself) generates a compressed object that has almost the same size as the compression version of X. This restriction is a known limitation for the universality of the NCD since in specific domains, as in image compression, the best performing lossless algorithms such as JBIG, JPEG2000 and JPEG-LS do not satisfy such propriety [40]. Nevertheless, some remarkable and interesting results were already obtained in several other domains using the NCD, as image OCR [12], genomic analysis [3], malware recognition [6] and image distinguishability [48].

4 Application to mitochondrial DNA

In the last years, the advances in genomics revealed that the mitochondrial DNA (mtDNA) genomes can be used by researchers to get a better insight about phylogenetic relations between species, discover new evolutionary paths or even correct previously accepted relationships [22,39,41,47,50]. A complete database of genomes from several species can be found in the National Center for Biotechnology Information website (https://www.ncbi.nlm.nih.gov). The information of each mtDNA genome is coded in a 4 symbol alphabet, corresponding each one to the nucleobases {A,C,T,G} = {adenine, cytosine, thymine, guanine} restricted to the base pairing rule (A with T and C with G) [33–35,37]. Depending on the species under analysis, their mtDNA sequence can have up to 18k bases, presented as a 18k length string.

If we are to compare several mtDNA sequences, then we need a distance for evaluating the amount of shared information between them. Among the plethora of methods found in the literature, the *k-mer* Statistic (Shannon block entropy) is one technique used to classify DNA sequences [30]. This approach uses the frequency statistic of i sub-strings of length k in each mtDNA sequence, constructing a vector where the ith entry is the frequency with which the ith block overlaps the mtDNA under test. After gathering these data from two different mtDNA sequences, their similarity can be evaluated using the L_2 norm of the difference between those vectors. However, as was verified in [30] the success of this method strongly depends on the adopted sub-string length (k) that significantly influences the resulting shape for the phylogenetic tree.

4.1 Analysis by means of NCD

For evaluating the universality of the NCD compression-based metric, a Python script (open-source available at https://github.com/romantao/ncd_python) implements the NCD algorithm. To simplify the analysis, the program allows the user to define the set of files to be compared in a pairwise way and to choose which 'normal' and lossless compressor will be adopted to approximate the Kolmogorov complexity. Additionally, it provides additional scripts to aid the visualization and comparison of the obtained results by means of different methods.

Currently, two compressing engines, namely the `zlib` and `bzip2`, are available. The first is based on the Lempel–Ziv–Markov LZ77 scheme (`zlib`, dictionary based) and the second follows the Burrows–Wheeler algorithm (`bzip2`, block-sorting based) [1]. One important condition for the fidelity of this approximation is the invariability of the compressor with respect to the size of the objects. While some well-known compressors do not verify this condition due to

Table 1 The set $\mathcal{M}$ of 18 mammals

i	Mammal	mtDNA size (bytes)
1	*Papio anubis* (Baboon)	16,521
2	*Ursus arctos* (Brown Bear)	17,021
3	*Felis catus* (Cat)	17,009
4	*Pan troglodytes* (Chimpanzee)	16,563
5	*Canis lupus familiaris* (Dog)	16,727
6	*Equus asinus* (Donkey)	16,670
7	*Glis glis* (Fat Dormouse)	16,602
8	*Gorilla gorilla* (Gorilla)	16,364
9	*Equus caballus* (Horse)	16,660
10	*Homo sapiens* (Human)	16,569
11	*Rhinoceros unicornis* (Indian Rhinoceros)	16,829
12	*Mus musculus* (Mouse)	16,295
13	*Pongo abelii* (Orangutan)	16,389
14	*Ursus Maritimus* (Polar Bear)	17,018
15	*Oryctolagus cuniculus* (Rabbit)	17,254
16	*Rattus norvegicus* (Rat)	16,300
17	*Sciuridae* (Squirrel)	16,495
18	*Ceratotherium simum* (White Rhinoceros)	16,832

the algorithm window size, in their range of use they can be considered a good approximation of the Kolmogorov complexity [9].

In this scenario 18 text files, containing the mtDNA sequence of 18 different mammals, were used. The set of mammals, $\mathcal{M}$, is listed in Table 1.

Before applying the NCD to the mtDNA data set, it is important to verify if the conditions $\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$ presented in (1) are followed by the `zlib` and `bzip2` compressors when applied to $\mathcal{M}$.

Figure 1 depicts the test for the identity axiom $\mathcal{C}_1$. The results of $\mathrm{NCD}(x, x)$ evince that self-comparisons of the mtDNA files based on the NCD with `zlib` are closely validate since it provides values close to 0. On the other hand, the `bzip2` leads to poor results with values far from 0. We conclude that `zlib` is superior to `bzip2` in the perspective of $\mathcal{C}_1$.

Figure 2 illustrates the test for the symmetry axiom $\mathcal{C}_2$. We verify that both compressors lead to values of $\mathrm{NCD}(x, y) - \mathrm{NCD}(y, x)$ very close to zero. In this case, we conclude that both the `zlib` and the `bzip2` are adequate.

Figure 3 shows the results when testing the triangle inequality $\mathcal{C}_3$. The results demonstrate that $\mathrm{NCD}(x, z) - \mathrm{NCD}(x, y) - \mathrm{NCD}(y, z) < 0$, for both compressors, confirming their validity.

Based on the test results, the `zlib` compressor was chosen to approximate the NID complexity. In this line of thought, the NCD algorithm was applied to all pairs of files using the `zlib` compressor. The corresponding $\mathbf{M_{NCD}}$ symmetric matrix, 18×18 dimensional, with the distances between the mtDNA sequences is represented in Table 2. To visualize the whole species relationships, the best-fitting two dimensions unrooted tree is used, as represented in Figure 4. Due to the non-ideal nature of the compressor, the main diagonal, with values smaller than 4×10^{-2}, was set to zero for the tree generation.

The PHYLIP (PHYLogeny Inference Package) [2] set of programs for inferring phylogenies was adopted. In particular, were used the neighbor-joining method [44] and the UPGMA (unweighted pair group method with arithmetic mean) an agglomerative (bottom-up) hierarchical clustering [46], followed by the DRAWTREE [2] that plots unrooted phylogenies.

We observe in the tree structure that the NCD successfully identified the closest mammal species. It is possible to distinguish in $\mathcal{M}$ three taxa, namely the primates (blue), Glires (green) and Ferungulata (red). It is also interesting to observe, for example, the localization of the rabbit in the tree since, for long time the rabbit and the rest of the rodents were classified in different groups, being their apparent similarity justified by convergent evolution. However, recent mtDNA analysis suggested that they have a common ancestor (see Reference [49]), as is also verified in the presented results.

Despite being a sub-optimal metric, the NCD with the `bzip2` compressor was also assessed. Figure 5 compares the two measures, namely NCD with `zlib` vs NCD with `bzip2`. We verify that NCD with `zlib` and NCD with `bzip2` follow a linear relationship and, for that reason, the emerging trees reveal an identical interdependence relationships between the 18 species.

4.2 Analysis by means of procrustes and Jensen–Shannon divergence

In this subsection we evaluate the performance of the NCD with the `zlib` compression engine. For this purpose we consider 2 alternative distances, namely a Procrustes analysis based on 'images' of the mtDNA and a statistical analysis based on triple symbols of the DNA that are explored in the follow-up.

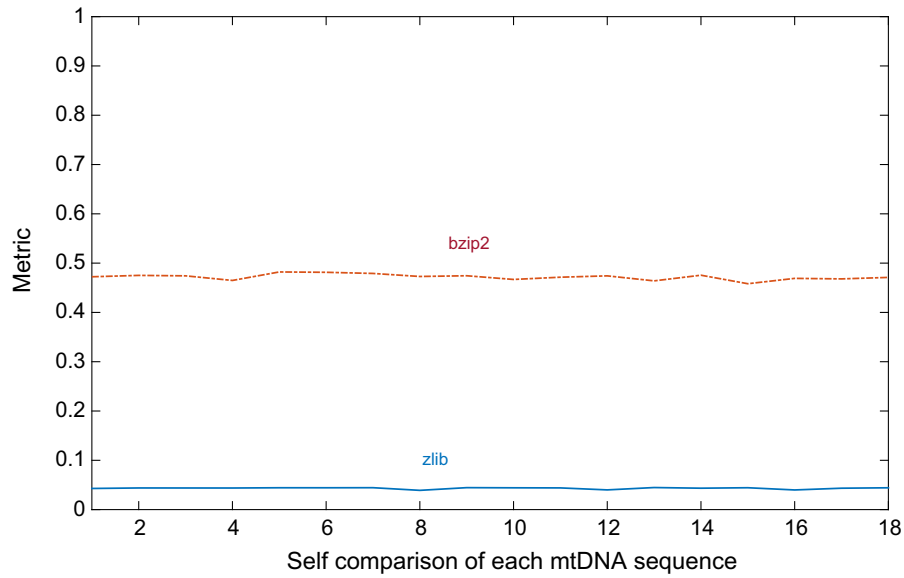


Fig. 1 Validation of the identity axiom $\mathcal{C}_1$ with the NCD using `zlib` and `bzip2` compressors for set $\mathcal{M}$

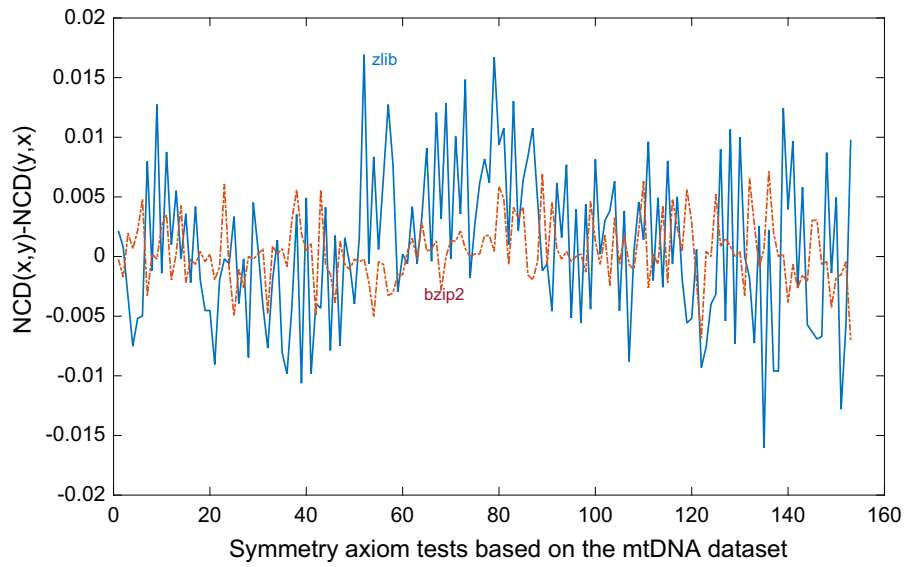


Fig. 2 Validation of the symmetry axiom $\mathcal{C}_2$ with the NCD using `zlib` and `bzip2` compressors for set $\mathcal{M}$

4.2.1 The procrustes analysis

In this case we consider the 4 DNA nucleobases, the base pairing C-G and A-T and a time walk along the string. The sequential reading of the symbols of the set {C,G,A,T} leads to a path in a pseudo-state plane variables xy, having x-axis for {C,G} and y-axis for {A,T}. The symbols C and A correspond to a positive quantum unit $\delta = +1$, while the symbols G and T represent the negative quantum unit $\delta = -1$. The algorithm $\mathcal{A}$ for 'reading' the mtDNA sequence is as follows [33]:

1. Start at DNA reading at initial point (i.e., the beginning of the pseudo-time)
2. One 'step forward' in the DNA strand is modeled by one discrete time increment
3. The observation of symbol
   (a) C reads as $\delta x = +1$
   (b) G reads as $\delta x = -1$
   (c) A reads as $\delta y = +1$
   (d) T reads as $\delta y = -1$
4. The DNA reading stops when reaching its endpoint (i.e., the end of the pseudo-time period).

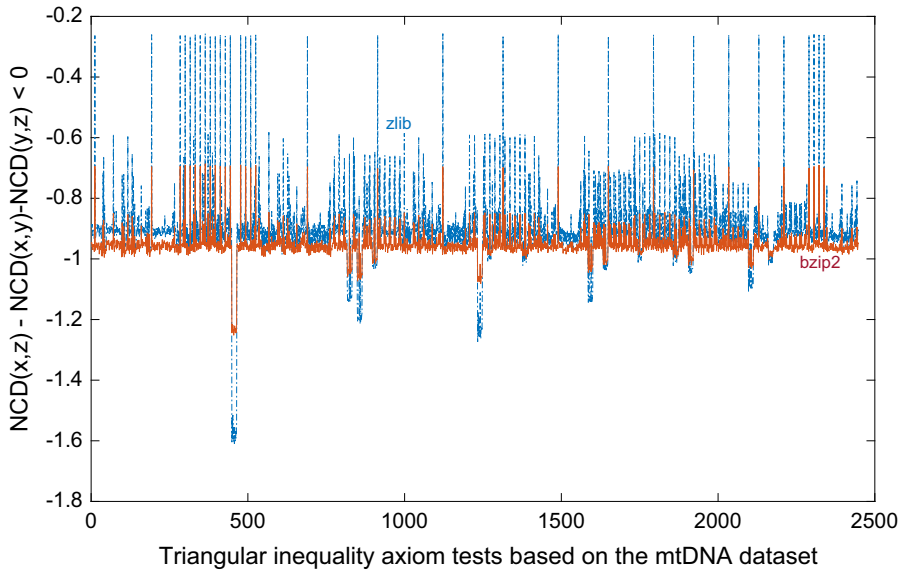


Fig. 3 Validation of the triangle inequality $\mathcal{C}_3$ with the NCD using `zlib` and `bzip2` compressors for set $\mathcal{M}$

Table 2 The symmetric matrix $\mathbf{M_{NCD}}$ with the NCD using `zlib` for set $\mathcal{M}$ (values $\times 10^{-2}$ are rounded to 2 decimal places for visualization)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1																		
2	93																	
3	93	89																
4	90	93	92															
5	94	89	89	92														
6	94	91	90	93	90													
7	94	92	91	93	91	91												
8	90	93	93	71	93	92	93											
9	93	91	89	92	89	59	91	93										
10	89	93	93	65	92	92	93	72	92									
11	93	91	90	93	90	85	92	93	85	92								
12	93	93	92	93	91	91	91	93	92	94	92							
13	90	94	93	82	94	93	93	83	93	82	93	94						
14	93	26	89	93	89	90	92	93	91	94	90	93	93					
15	93	91	91	93	91	91	92	92	91	93	92	92	93	92				
16	93	93	92	93	91	91	91	93	91	94	91	85	92	93	91			
17	94	93	91	93	91	91	91	93	91	92	91	92	93	92	91	92		
18	93	90	89	92	89	85	92	92	84	92	75	92	93	91	91	92	91	

For example, Fig. 6 shows the representation of the mtDNA of the *Homo sapiens* in the xy plane.

The pairs of matrix representations of the mtDND walks in xy are compared by means of Procrustes analysis. Procrustes [16,24] determines a linear transformation (i.e., a combination of translation, reflection, orthogonal rotation and scaling) of the points in one matrix that best conform to the points in the second matrix. The goodness-of-fit criterion is the sum of squared errors. Based on this index a symmetric matrix $\mathbf{M_P}$ of item-to-item comparison can be built and the corresponding tree visualized.

4.2.2 The Jensen–Shannon divergence

The Kullback–Leibler divergence (D_{KL}) between two probability distributions, P_1 and P_2, is given by [28, 29]:

$$D_{KL}\left(P_1||P_2\right) = -\sum_{k=1}^{n} P_1(k) \log \frac{P_2(k)}{P_1(k)}. \tag{6}$$

The Jensen–Shannon divergence (D_{JS}) uses a symmetric version of the D_{KL} and measures the similarity between two probability distributions. The D_{JS} is given by [14,31]:

$$D_{JS}\left(P_1||P_2\right) = \frac{1}{2}\left[D_{KL}\left(P_1||M\right) + D_{KL}\left(P_2||M\right)\right], \tag{7}$$

where $M = \frac{1}{2}\left(P + Q\right)$ is the mixture distribution of P_1 and P_2.

The D_{JS} can be successfully applied in complex systems and, in particular, to DNA analysis [32,42].

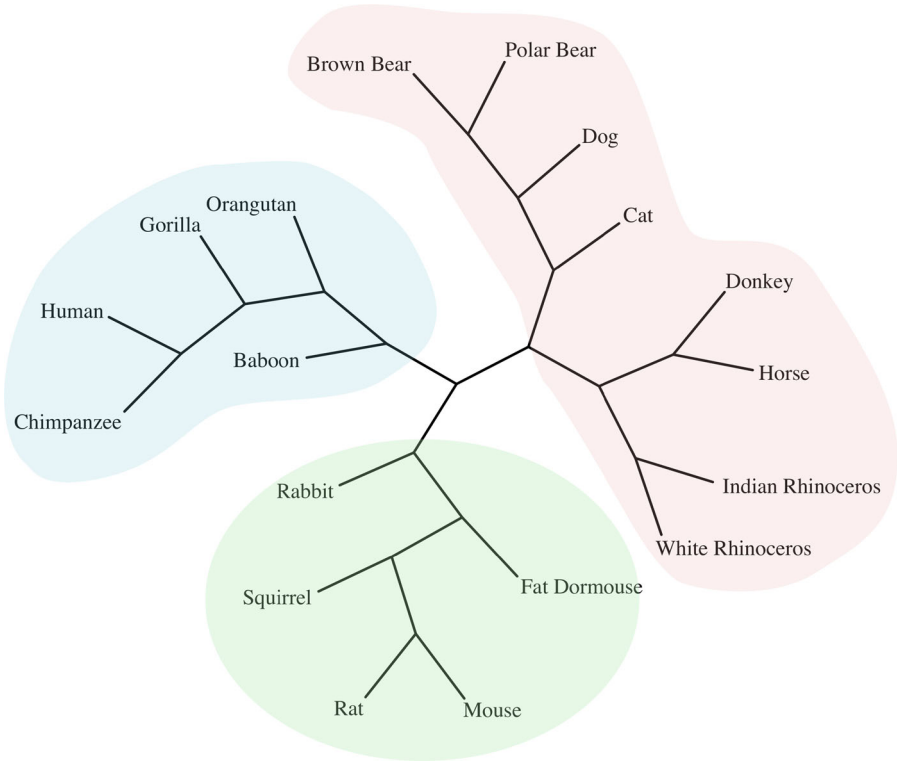


Fig. 4 Evolutionary unrooted tree built by PHYLIP based on matrix $\mathbf{M_{NCD}}$ with the NCD using `zlib` and set $\mathcal{M}$ (primates in blue, Glires in green and Ferungulata in red). (Color figure online)

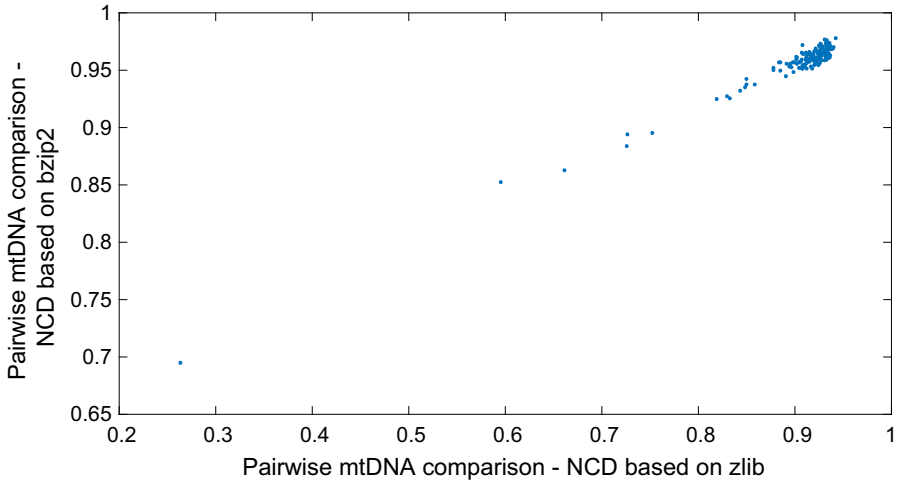


Fig. 5 Comparison of the two measures: NCD with `zlib` versus NCD with `bzip2` for set $\mathcal{M}$

Based on the D_{JS} index, a symmetric matrix $\mathbf{M_{JS}}$ of item-to-item distances and the corresponding tree are constructed.

The probability distribution $P_i, i = 1, \ldots, 18$, for the i-th species is estimated by means of the histogram having 4^3 bins counting sets of three consecutive symbols. Furthermore, it is considered a sliding window of 1 symbol, so that a mtDNA with N symbols leads to $N - 1$ samples.

4.2.3 Comparison of the measures

Figure 7 depicts the Procrustes vs NCD with `zlib` and the D_{JS} vs NCD with `zlib`. We verify that they follow a nonlinear relationship, which seems natural since they derive from very different logical schemes. Nevertheless, they have a monotonic evolution versus `zlib` and we can study its behavior when applied in the set $\mathcal{M}$ of the 18 mammals listed in Table 1. Figure 7

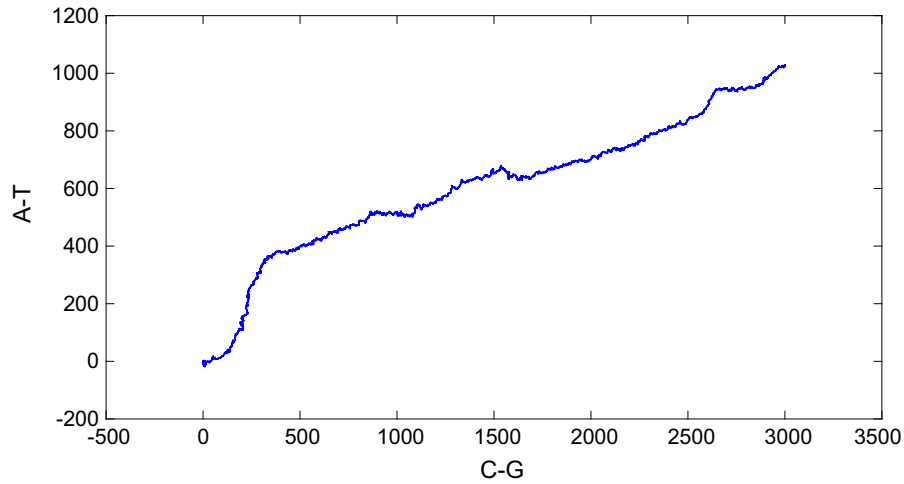
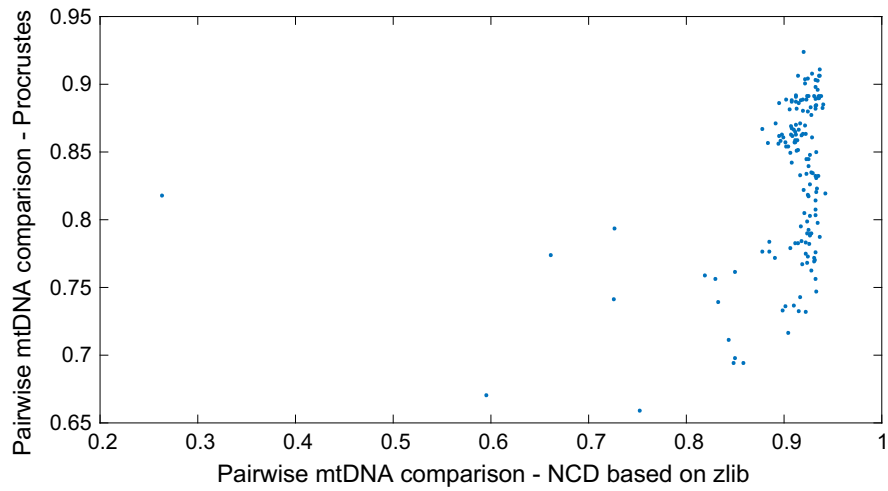
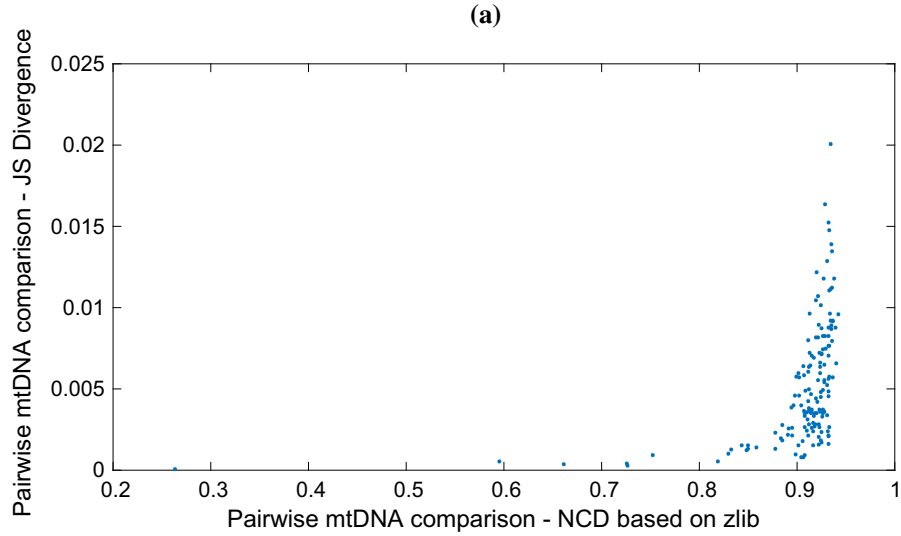


Fig. 6 The representation of the mtDNA of the *Homo sapiens* in the xy plane according with algorithm $\mathcal{A}$: the nucleobases C or G read as $\delta x = +1$ or $\delta x = -1$, and the nucleobases A or T read as $\delta y = +1$ or $\delta y = -1$, respectively



(a)



(b)

Fig. 7 Comparison of distances: **a** Procrustes vs NCD with `zlib`, **b** D_{JS} vs NCD with `zlib`, for set $\mathcal{M}$

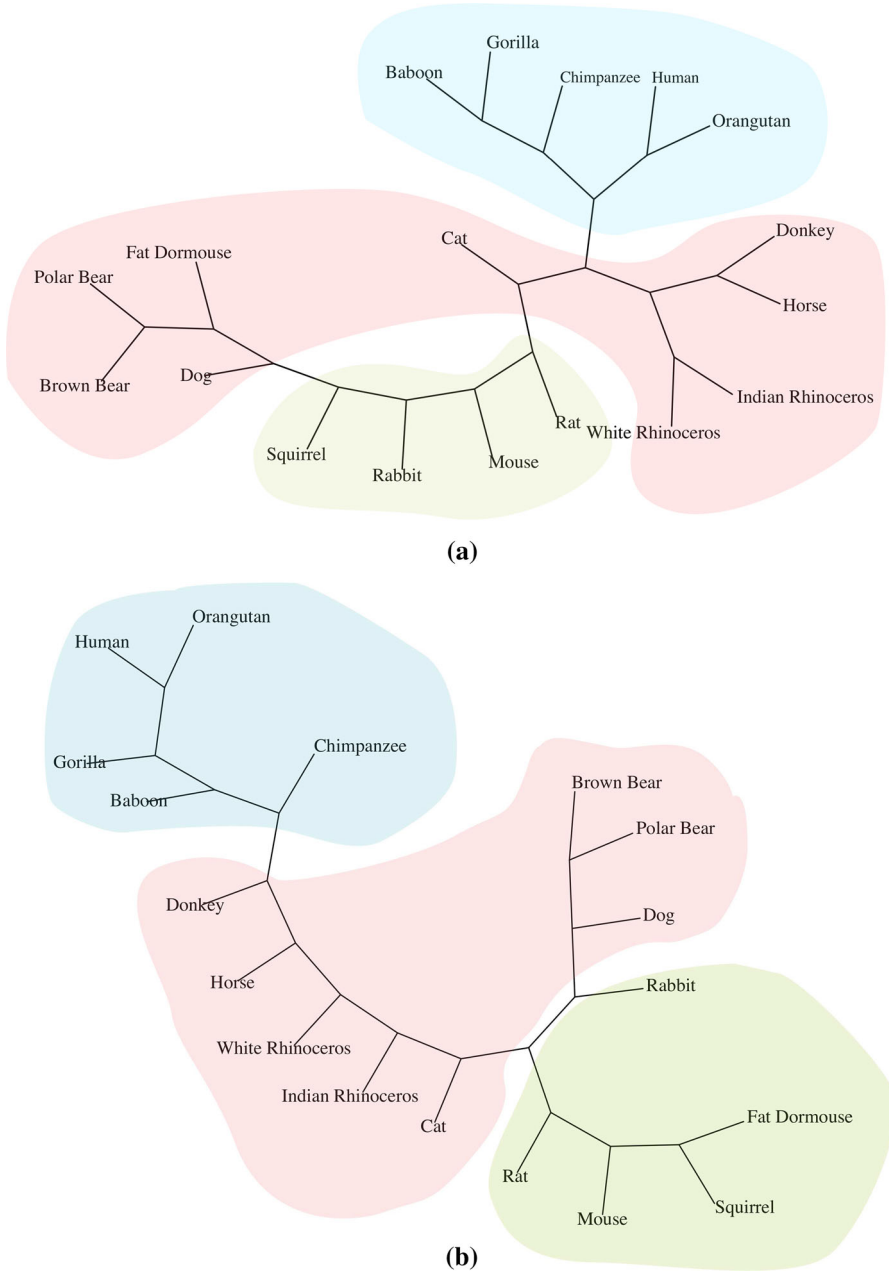


Fig. 8 Evolutionary unrooted trees built by PHYLIP based on matrices: **a** $\mathbf{M_P}$, **b** $\mathbf{M_{JS}}$, and set $\mathcal{M}$ (primates in blue, Glires in green and Ferungulata in red). (Color figure online)

confirms that the trees produced by NCD with `zlib`, Procrustes and D_{JS} are compatible and that NCD is an useful distance.

Figure 8 shows the evolutionary unrooted trees built by PHYLIP based on matrices $\mathbf{M_P}$ and $\mathbf{M_{JS}}$ for the set $\mathcal{M}$ (primates in blue, Glires in green and Ferungulata in red).

In the final test we assess the sensitivity of the three distances by testing the mtDNA of two very close species, the chimpanzee and the human (positions 4 and 10 in Table 1) . We 'corrupt' the mtDNA file of the chimpanzee, and we calculate the distance toward the unchanged mtDNA file of the human. The relative

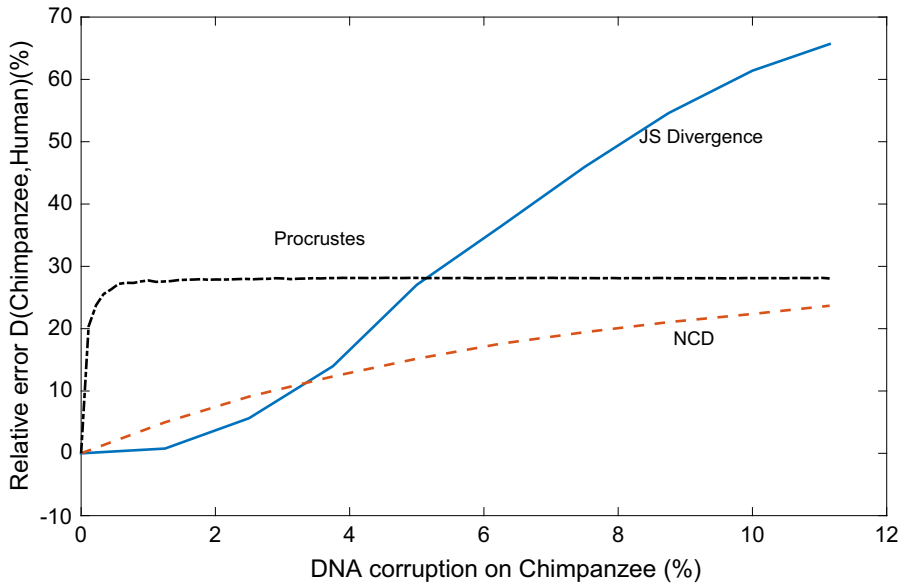


Fig. 9 Average value $\overline{E}$ of the distance error between the chimpanzee and the human versus the percentage of random noise N for the NCD with `zlib`, Procrustes and D_{JS}

error E is calculated as

$$E(N) = \frac{\left|D_{4,10}(N) - D_{4,10}(0)\right|}{D_{4,10}(0)}, \tag{8}$$

where N represents the percentage of noise, and $D_{4,10}(N)$ and $D_{4,10}(0)$ denote the distances between the chimpanzee and the human with and without noise, respectively.

The experiment is repeated for a statistical sample of 100 trials and each percentage of noise N, and the average error value, $\overline{E}$, is determined. Figure 9 shows $\overline{E}$ versus N and we verify that, for the three measures, $\overline{E}$ increases with N. However, both Procrustes and D_{JS} reveal a significant nonlinear behavior. Procrustes has a sudden increase for low values of N followed by a saturation, while the D_{JS} reveals two distinct regimes with an inflection point for $N \approx 4\%$. On the other hand, the NCD with `zlib` reveals a smooth evolution, increasing almost proportionally with N. This test shows that the NCD with `zlib` exhibits a predictable and robust behavior.

5 Conclusions

The non-computable Kolmogorov complexity approximation using lossless and 'normal' compression algorithms provided a similarity metric based on the non-computable formalization of the NID. The NCD is a powerful tool due to its capability of mining patterns in different domains without requiring some initial background or knowledge. The underlying concepts of this theory were reviewed and applied in genomics for a set of 18 mtDNA files. In the first phase, the NCD with two alternative compression engines, `zlib` and `bzip2`, was tested for the three distance axioms. In the second phase, the NCD with `zlib` was compared with two distinct distances following the Procrustes and the Jensen–Shannon divergence. In the third phase the NCD sensitivity to random noise was also assessed. The results for the NCD are solid and support further research using the NCD not only with larger data sets, but also with different types of natural and artificial phenomena described by means of data files following a given logical scheme.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Engineering and technology history wiki: History of lossless data compression algorithms. http://ethw.org/History_of_Lossless_Data_Compression_Algorithms. Accessed 19 Oct 2017
2. Phylip. http://evolution.genetics.washington.edu/phylip.html

3. On the Approximation of the Kolmogorov Complexity for DNA Sequences (2017). https://doi.org/10.1007/978-3-319-58838-4_29
4. Aziz, M., Alhadidi, D., Mohammed, N.: Secure approximation of edit distance on genomic data. BMC Med Genomics **10**(Suppl 2), (2017). https://doi.org/10.1186/s12920-017-0279-9
5. Bennett, C.H., Gács, P., Li, M., Vitányi, P., Zurek, W.H.: Information distance. IEEE Trans. Inf. Theory **44**(4), 1407–1423 (1998)
6. Borbely, R.S.: On normalized compression distance and large malware. J. Comput. Virol. Hacking Tech. **12**(4), 235–242 (2016). https://doi.org/10.1007/s11416-015-0260-0
7. Yin, C., Chen, Y., Sdddd, Y.: A measure of DNA sequence similarity by fourier transform with applications on hierarchical clustering complexity for DNA sequences. J. Theor. Biol. **359**, 18–28 (2014). https://doi.org/10.1016/j.jtbi.2014.05.043
8. Carbone, A.: Information measure for long-range correlated sequences: the case of the 24 human chromosomes. Scientific Reports **3** (2013). https://doi.org/10.1038/srep02721
9. Cebrián, M., Alfonseca, M., Ortega, A.: Common pitfalls using the normalized compression distance: what to watch for in a compressor. Commun. Inf. Syst. **5**(4), 367–384 (2005)
10. Cebrián, M., Alfonseca, M., Ortega, A.: Common pitfalls using the normalized compression distance: what to watch out for in a compressor. Commun. Inf. Syst. **5**(4), 367–384 (2005). https://doi.org/10.4310/CIS.2005.v5.n4.a1
11. Cilibrasi, R., Vitany, P.M.B.: Clustering by compression. IEEE Trans. Inf. Theory **51**(4), 1523–1545 (2005). https://doi.org/10.1109/TIT.2005.844059
12. Cohen, A.R., Vitányi, P.M.B.: Normalized compression distance of multisets with applications. IEEE Trans. Pattern Anal. Mach. Intell. **37**(8), 1602–1614 (2015). https://doi.org/10.1109/TPAMI.2014.2375175
13. Deza, M.M., Deza, E.: Encyclopedia of Distances. Springer, Berlin (2009)
14. Endres, D., Schindelin, J.: A new metric for probability distributions. IEEE Trans. Inf. Theory **49**(7), 1858–1860 (2003). https://doi.org/10.1109/TIT.2003.813506
15. Fortnow, L., Lee, T., Vereshchagin, N.: Kolmogorov complexity with error. In: Durand, B., Thomas, W. (eds.) STACS 2006–23rd Annual Symposium on Theoretical Aspects of Computer Science, Marseille, France, February 23–25, 2006. Lecture Notes in Computer Science, pp. 137–148. Springer, Berlin (2006)
16. Gower, J.C., Dijksterhuis, G.B.: Procrustes Problems. Oxford University Press, Oxford (2004)
17. Glunčić, M., Paar, V.: Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Research **41**(1) (2013). https://doi.org/10.1093/nar/gks721
18. Grünwald, P.D., Vitányi, P.M.B.: Kolmogorov complexity and information theory. J. Logic Lang. Inf. **12**, 497–529 (2003)
19. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (eds.): Feature Extraction: foundations and Applications. Springer, Berlin (2008)
20. Hautamaki, V., Pollanen, A., Kinnunen, T., Aik, K., Haizhou, L., Franti, L.: A Comparison of Categorical Attribute Data Clustering Methods, pp. 53–62. Springer, Berlin (2014). https://doi.org/10.1007/978-3-662-44415-3_6
21. Hu, L.Y., Huang, M.W., Ke, S.W., Tsai, C.F.: The distance function effect on k-nearest neighbor classification for medical datasets. Springer Plus **5**, 1304 (2016). https://doi.org/10.1186/s40064-016-2941-7
22. Kalinowski, S.T., Leonard, M.J., Andrews, T.M.: Nothing in evolution makes sense except in the light of DNA. CBE Life Sci. Educ. **2**(9), 87–97 (2010). https://doi.org/10.1187/cbe.09-12-0088
23. Kawakatsu, H.: Methods for evaluating pictures and extracting music by 2D DFA and 2D FFT. Procedia Comput. Sci. **60**, 834–840 (2015). https://doi.org/10.1016/j.procs.2015.08.246
24. Kendall, D.G.: A survey of the statistical theory of shape. Stat. Sci. **4**(12), 87–99 (1989)
25. Klenk, S., Thom, D., Heidemann, G.: The Normalized Compression Distance as a Distance Measure in Entity Identification. Springer, Berlin (2009)
26. Kolmogorov, A.: Three approaches to the quantitative definition of information. Int. J. Comput. Math. **2**(1–4), 157–168 (1968)
27. Kubicova, V., Provaznik, I.: Relationship of bacteria using comparison of whole genome sequences in frequency domain. Inf. Technol. Biomed. **3**, 397–408 (2014). https://doi.org/10.1007/978-3-319-06593-9_35
28. Kullback, S.: Information Theory and Statistics. Wiley, New York (1959)
29. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
30. Li, M., Chen, X., Li, X., Ma, B., Vitány, P.M.B.: The similarity metric. IEEE Trans. Inf. Theory **50**(12), 3250–3264 (2004). https://doi.org/10.1109/TIT.2004.838101
31. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory **37**(1), 145–151 (1991). https://doi.org/10.1109/18.61115
32. Machado, J.A.T.: Fractional order generalized information. Entropy **16**(4), 2350–2361 (2014). https://doi.org/10.3390/e16042350
33. Machado, J.A.T.: Bond graph and memristor approach to DNA analysis. Nonlinear Dyn. **88**(2), 1051–1057 (2017). https://doi.org/10.1007/s11071-016-3294-z
34. Machado, J.T.: Fractional order description of DNA. Appl. Math. Model. **39**(14), 4095–4102 (2015). https://doi.org/10.1016/j.apm.2014.12.037
35. Machado, J.T., Costa, A., Quelhas, M.: Entropy analysis of DNA code dynamics in human chromosomes. Comput. Math. Appl. **62**(3), 1612–1617 (2011). https://doi.org/10.1016/j.camwa.2011.03.005
36. Machado, J.T., Costa, A.C., Lima, M.F.M.: Dynamical analysis of compositions. Nonlinear Dyn. **65**(4), 399–412 (2011). https://doi.org/10.1007/s11071-010-9900-6
37. Machado, J.T., Costa, A.C., Quelhas, M.D.: Fractional dynamics in DNA. Commun. Nonlinear Sci. Numer. Simul. **16**(8), 2963–2969 (2011). https://doi.org/10.1016/j.cnsns.2010.11.007
38. MacKay, D.J.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge (2003)

39. Moscato, P., Buriol, L., Cotta, C.: On the analysis of data derived from mitochondrial DNA distance matrices: Kolmogorov and a traveling salesman give their opinion (2002)
40. Pinho, A., Ferreira, P.: Image similarity using the normalized compression distance based on finite context models. In: Proceedings of IEEE International Conference on Image Processing (2011). https://doi.org/10.1109/ICIP.2011.6115866
41. Rajarajeswari, P., Apparao, A.: Normalized distance matrix method for construction of phylogenetic trees using new compressor - DNABIT compress. J. Adv. Bioinf. Appl. Res. **2**(1), 89–97 (2011)
42. Ré, M.A., Azad, R.K.: Generalization of entropy based divergence measures for symbolic sequence analysis. PLoS ONE **9**(4), e93,532 (2014). https://doi.org/10.1371/journal.pone.0093532
43. Russel, R., Sinha, P.: Perceptually based comparison of image similarity metrics. Perception **40**, 1269–1281 (2011). https://doi.org/10.1068/p7063
44. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**(4), 406–425 (1987)
45. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
46. Sokal, R.R., Michener, C.D.: A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. **38**(22), 1409–1438 (1958)
47. Starr, T.N., Picton, L.K., Thornton, J.W.: Alternative evolutionary histories in the sequence space of an ancient protein. Nature **549**, 409–413 (2017). https://doi.org/10.1038/nature23902
48. Vázquez, P.P., Marco, J.: Using normalized compression distance for image similarity measurement: an experimental study. J. Comput. Virol. Hacking Tech. **28**(11), 1063–1084 (2012). https://doi.org/10.1007/s00371-011-0651-2
49. Walsh, B.: Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals. Genetics **158**(2), 897–912 (2001)
50. Wang, W., Wang, T.: Conditional LZ complexity and its application in mtDNA sequence analysis. MATCH Commun. Math. Comput. Chem. **66**, 425–443 (2011)
51. Yianilos, P.N.: Normalized forms of two common metrics. Tech. Rep. Report 91-082-9027-1, NEC Research Institute (1991)
52. Yu, J., Amores, J., Sebe, N., Tian, Q.: A new study on distance metrics as similarity measurement. In: IEEE International Conference on Multimedia and Expo (2006). https://doi.org/10.1109/ICME.2006.262443