entropy

MDPI

Article

Comparison of Compression-Based Measures with Application to the Evolution of Primate Genomes

Diogo Pratas [1,*], Raquel M. Silva [1,2] and Armando J. Pinho [1,3]

1 Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, 3810-193 Aveiro, Portugal; raquelsilva@ua.pt (R.M.S.); ap@ua.pt (A.J.P.)
2 Department of Medical Sciences and Institute for Biomedicine - iBiMED, University of Aveiro, 3810-193 Aveiro, Portugal
3 Department of Electronics, Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal
* Correspondence: pratas@ua.pt; Tel.: +351-234-370-507

Received: 3 March 2018; Accepted: 21 May 2018; Published: 23 May 2018

Abstract: An efficient DNA compressor furnishes an approximation to measure and compare information quantities present in, between and across DNA sequences, regardless of the characteristics of the sources. In this paper, we compare directly two information measures, the Normalized Compression Distance (NCD) and the Normalized Relative Compression (NRC). These measures answer different questions; the NCD measures how similar both strings are (in terms of information content) and the NRC (which, in general, is nonsymmetric) indicates the fraction of one of them that cannot be constructed using information from the other one. This leads to the problem of finding out which measure (or question) is more suitable for the answer we need. For computing both, we use a state of the art DNA sequence compressor that we benchmark with some top compressors in different compression modes. Then, we apply the compressor on DNA sequences with different scales and natures, first using synthetic sequences and then on real DNA sequences. The last include mitochondrial DNA (mtDNA), messenger RNA (mRNA) and genomic DNA (gDNA) of seven primates. We provide several insights into evolutionary acceleration rates at different scales, namely, the observation and confirmation across the whole genomes of a higher variation rate of the mtDNA relative to the gDNA. We also show the importance of relative compression for localizing similar information regions using mtDNA.

Keywords: data compression; NCD; NRC; DNA sequences; primate evolution

1. Introduction

In 1965 [1], Kolmogorov described three ways to measure the information contained in strings: combinatorial [2,3], probabilistic [4] and algorithmic. The algorithmic approach, also known as Kolmogorov complexity or algorithmic entropy, enables measurement and comparison of the information (or complexity) contained in different natural processes that can be expressed using sequences of symbols (strings) from a finite alphabet [1,5–13].

The Kolmogorov complexity differs from the Shannon entropy [4], because it considers that the source, rather than generating symbols from a probabilistic function, creates structures that follow algorithmic schemes [14,15]. Therefore, to reverse the problem, there is the need to identify the program(s) and parameter(s) that generate the outcome(s) [1,11]. Successful implementations, using small Turing machines [16], have been proposed and implemented (for example [17,18]). Other implementations, using lossless data compressors, have also been proposed [19,20].

The Kolmogorov complexity is non-computable [21], mostly because of the halting problem [22]. Therefore, we have to rely on approximations such as string compressors, $C(x)$. For a definition of safe approximation, see [23]. The normalized version, known as the Normalized Compression (NC), is defined by

$$\mathrm{NC}(x) = \frac{C(x)}{|x| \log_2 |\mathrm{A}|}, \tag{1}$$

where x is a string, $|\mathrm{A}|$ the number of possible different elements in x (size of the alphabet) and $|x|$ the length of x. The NC enables to compare the information contained in the strings independently from their sizes [19].

The usage of compressors has also been applied to approximate the amount of information between two strings, x and y, namely through the Normalized Compression Distance (NCD) [11,24–28]. In order to compute this distance, we may use the conditional compression [29,30], $C(x|y)$, as

$$\mathrm{NCD}(x,y) = \frac{\max\{C(x|y), C(y|x)\}}{\max\{C(x), C(y)\}}, \tag{2}$$

or the conjoint compression [25], $C(x,y)$, as

$$\mathrm{NCD}(x,y) = \frac{C(x,y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}, \tag{3}$$

given that the conditional and conjoint compressions are related through the chain rule [11].

The NCD enables to measure the information between two strings, being robust to some degree of noise [31], as long as the compressor respects the normality properties [11,32].

There are many examples of the NCD applicability, namely in plagiarism analysis [33], stemmatology [34], DNA analysis [27,28,35–37], gene expression analysis [38,39], complex system behaviour studies and time series analysis [40], image distinguishability [41], image similarity [42–44], image distortion analysis [45], visual analysis of document collections [46], classification of file fragments [47], retinal progenitor cell fate prediction and handwritten digits classification [48], clustering music [49], evolving computer-generated music [50], cover songs identification [51], entity identification [52], assessing the impact of students in peer review [53], analysis of internet malware [54,55], analysis of software systems stability [56], measuring the similarity between black-and white-box regression test prioritization techniques [57], optimization of compilers [58], analysis of public opinion [59], deception detection [60], and fawns detection [61].

On the other hand, to measure the information of a string relative to another [62,63], we have to rely on relative semi-distances. These are measures that do not need to respect two distance properties, namely symmetry and the triangle inequality [64]. Several approaches to quantify the relative information have been proposed (e.g., [62,65–69]), such as for handling images [66], texts [68,69], ECG data [70], and genomic sequences [71–73].

The relative information can be approximated using relative compressors [62,68,69,72,74–77], regardless if they are based on dictionaries [62,68,74] or Markovian models [69,72]. These compressors aim to model and organize the data of y (without knowing x). Then, they freeze the model of y [78] and, finally, they measure the number of bits needed to describe x, using exclusively the information from y. We call this operation the compression of x relatively to y and denote it by $C(x\|y)$.

The information of x relative to y, $C(x\|y)$, is a sum of the information content provided, symbol by symbol, after processing the complete x, according to

$$C(x\|y) = \sum_{i=1}^{|x|} C(x_i\|y), \tag{4}$$

where $|x|$ is the size of string x. The information profile of $C(x_i\|y)$ is very important to localize regions that are similar in x relatively to y, namely through the lower information content regions below a certain threshold [64,71]. For overall quantities, we need to compute the Normalized Relative Compression (NRC) as

$$\mathrm{NRC}(x\|y) = \frac{C(x\|y)}{|x| \log_2 |\mathrm{A}|}. \tag{5}$$

The NRC enables to compare and relate symbolic sequences, namely DNA sequences, permitting to test multiple hypotheses and infer patterns in most of the evolution theories [64].

The main theories in evolution shown that species evolve and have a common ancestral [79–82]. This means that the DNA between close species is very similar [81,82]. Therefore, the lower the variation among the species, the lower the information needed from one species DNA sequence to relatively describe the other.

However, we should note that, in practice, genomic sequences are not just a succession of letters with four possible outcomes (A,C,G,T), which indicate the order and nature of nucleotides within a DNA chemical chain. They are also the outcome of machines and algorithms due to the sequencing and assembling phases [83]. In reality, they are the outcome of a probabilistic capture of small pieces from a huge puzzle with lots of repeated, changed and missing pieces [84,85].

Additionally, genomic sequences have other alteration sources, for example environmental factors [86,87], pathogen species [88,89], rearrangements [90,91], and unknown sources [92]. Therefore, dealing with DNA sequences from different species is the equivalent of dealing with heterogeneous, dynamic, incomplete and imperfect information [93].

In eukaryotic cells, DNA sequences can have different sources, namely from the nucleus and organelles [94]. In animals, the double membrane-bound organelles are mitochondria, while in plants are mitochondrion and chloroplasts (plastids). The nuclear sequence (gDNA) is substantially longer than the mitochondrial sequence (mtDNA). A recent hypothesis to the conservation of the mtDNA size has been pointed to the protection against virulence agents [95].

The quantity of mutations that have been accumulated between two genomes provides a record of the time elapsed since their common ancestor [96]. The NCD and NRC are measures that are able to quantify variation, particularly because they measure information between and across DNA sequences. Moreover, when using relative measures, we can extend it to profile the information content enhancing relations and correlations between the data [71].

In this paper, we directly compare the NCD and NRC in synthetic and real data. For the purpose, we use a state-of-the-art compression algorithm and apply it on DNA sequences with different scales and natures. First, we compare the measures in synthetic data, with different characteristics. Then, we apply them on mtDNA, mRNA, and gDNA of seven primates. Finally, we show the importance of relative compression for localizing similar information regions using the mtDNA.

2. Methods

All the results provided in this paper can be replicated using the scripts provided in the repository https://github.com/pratas/APE. These will have as a dependency the GeCo compressor [72]. GeCo is freely available, under GPLv3 license, at https://github.com/pratas/geco. GeCo is a DNA sequence compressor, with state of the art compression capabilities, and uses reasonable computation resources. As an alignment-free tool [97,98], GeCo is able to determine absolute measures, namely for many (semi) distance computations, and local measures, such as the information content associated with each element. GeCo allows both reference-free and relative compression. Therefore, all the compression results provided in this paper were obtained using GeCo, although with custom models and parameters.

2.1. Compressor and Parameters

We use GeCo to compute the NCD (Equation (3)) and NRC (Equation (5)). For these two measures we only use two types of compression, $C(x)$ and $C(x\|y)$, mainly because the conjoint compression in

Equation (3), $C(x, y)$, can be approximated using the compression of x concatenated with y, namely $C(x, y) \approx C(xy) = C(yx)$, as long as the *normality* properties are verified [32,99]. Notice that $C(xy)$ stands for the compression of x and y concatenated together, while $C(x, y)$ represents the compression of x and y, as well as the necessary description to split them apart (split xy into x and y). Since the number of bits needed to represent a program to split xy is asymptotically insignificant, we can consider that the approximation is reasonable.

The GeCo compressor enables parametrization and a choice of the models for the two different modes needed, $C(x)$ and $C(x\|y)$. Despite the compression mode, the compressor uses a cooperation between multiple context models and substitutional tolerant context models [72,100].

The cooperation is supervised by a soft blending mechanism [101,102] that gives importance to the models that had a better performance given an exponential decayment record [102]. Figure 1 depicts an example of the cooperation between five context and tolerant context models, namely $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, $\mathcal{M}_4$, $\mathcal{M}_5$. Each of these models has a probability (P) and a weight (W) associated.

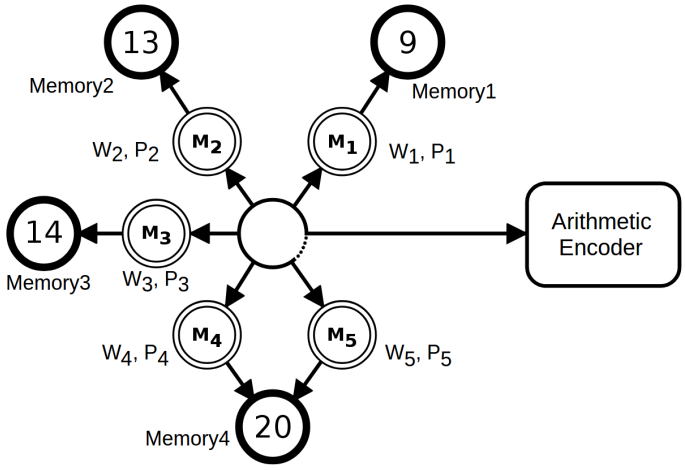

Figure 1. A mixture of five context models. Each model has a weight (W) and associated probabilities (P) that are calculated according to the respective memory model. The tolerant context model (5) uses the same memory of model 4, since they have the same context (depth 20). After, the probabilities are averaged according to the respective weight and redirected to the arithmetic encoder.

The difference between the $C(x)$ and the $C(x\|y)$ is given by the initialization and access to the memory models. In $C(x)$, the memory model starts with uniform counters, each set to zero. Through all the computation of $C(x)$ the memory model is updated. On the other hand, the $C(x\|y)$ is initialized with the memory of y, then the memory model is set static and used through all the computation of x.

The probability of each symbol, x_i, is given by

$$P(x_i) = \sum_{m \in \mathcal{M}} P_m(x_i | x_{i-k}^{i-1}) \; w_{m,i}, \tag{6}$$

where $P_m(x_i | x_{i-k}^{i-1})$ is the probability assigned to the next symbol by a context or substitutional tolerant context model, k the order of the corresponding model m, and where $w_{m,i}$ denote the corresponding weighting factor, with

$$w_{m,i} \propto (w_{m,i-1})^{\gamma_m} P_m(x_i | x_{i-k}^{i-1}), \tag{7}$$

where $\gamma_m \in [0, 1)$ acts as a forgetting factor for each context model. Frequently, we use the same value for each γ_m, since its optimization only provides minimal gains. The sum of the weights, for all the respective models, must be equal to one.

The depth of the model, k, identifies the number of contiguous symbols seen in the past for predicting the next symbol and, hence, x_{i-k}^{i-1} [103]. The alpha is an estimator parameter that allows

balancing between the uniform and the frequency distribution (usually deepest models have lower alphas [104]). The inverted repeats define if a short program, that allows searching for subsequences with similarity in complemented reverse sequences, is running (normally used in deepest models). The tolerance is a short program that enables to set the number of allowed substitutions in a certain context depth [72,100]. The cache-hash enables to keep in memory only the latest entries up to a certain number of hash collisions [99].

We use GeCo with a set of models and parameters according to the length and nature of the data, that have been set according to our experience. Notice that better models may be achieved with optimization, although we believe that the precision is good enough. The parameters are the following:

- $C(x\|y) \rightarrow$ mixture of seven models with a decayment (γ) of 0.95 and a cache-hash of 30:

```
1  tolerant context model: depth: 17, alpha: 0.02, tolerance: 5;
2  context model: depth: 17, alpha: 0.002, inverted repeats: no;
3  tolerant context model: depth: 14, alpha: 0.1, tolerance: 3;
4  context model: depth: 14, alpha: 0.005, inverted repeats: no;
5  context model: depth: 11, alpha: 0.01, inverted repeats: no;
6  context model: depth: 8, alpha: 0.1, inverted repeats: no;
7  context model: depth: 5, alpha: 1, inverted repeats: no;
```

- $C(x)$ and $C(xy) \rightarrow$ mixture of eight models with a decayment (γ) of 0.95 and a cache-hash of 30:

```
1  tolerant context model: depth: 17, alpha: 0.1, tolerance: 5;
2  context model: depth: 17, alpha: 0.005, inverted repeats: no;
3  tolerant context model: depth: 14, alpha: 1, tolerance: 3;
4  context model: depth: 14, alpha: 0.01, inverted repeats: no;
5  context model: depth: 11, alpha: 0.1, inverted repeats: no;
6  context model: depth: 8, alpha: 1, inverted repeats: no;
7  context model: depth: 5, alpha: 1, inverted repeats: no;
8  context model: depth: 3, alpha: 1, inverted repeats: no.
```

2.2. NCD versus NRC in Synthetic Data

Using the mentioned compressor, parameters and modes, we have compared the NCD and the NRC in synthetic data with custom redundancy, rearrangement formats and mutation rates. The sequences have been built using copies with different sizes and mutation rates, according to the legend (right panel) of Figure 2.

As it can be seen (Figure 2), there is an approximate symmetry in the NCD (Figure 2A). On the other hand, the NRC spends approximately the same amount of information to describe the x that have been originally extracted from y, regardless the rate of substitutions applied (when $|x| \leq |y|$).

When the distribution of the synthetic sequences is not uniform, being more similar to real DNA sequences (simulated with XS [105]), we are able to notice a decrease in the amount of information needed to describe x (Figure 2B), as well as higher values in the NCD in comparison with the NRC. This characteristic enables to predict higher NCD values on real datasets. On the other hand, comparing Figure 2A with Figure 2B, we see an increase in the non-linear behaviour of the NCD, while the NRC seems to approximate to a linear behaviour.

The most significant computations in the NCD are the $C(x)$ and $C(yx)$, while for the NRC is the $C(x\|y)$. Namely, because the normalization of the NRC is given by the size of x times a constant (2). Therefore, we now look into to the profiles of each computation in order to compare them. We emphasize that the NCD can be computed using the conjoint (Equation (3)) or conditional (Equation (2)) compression, and, therefore, $C(x|y) = C(yx) - C(y)$, an approximation that can be made through the chain rule [11] (as previously mentioned).

Using this connection, we have simulated two sequences, x and y, with characteristics according to Figure 3. Then, we have computed and compared the profiles of $C(x)$, $C(yx)$ and $C(x\|y)$.

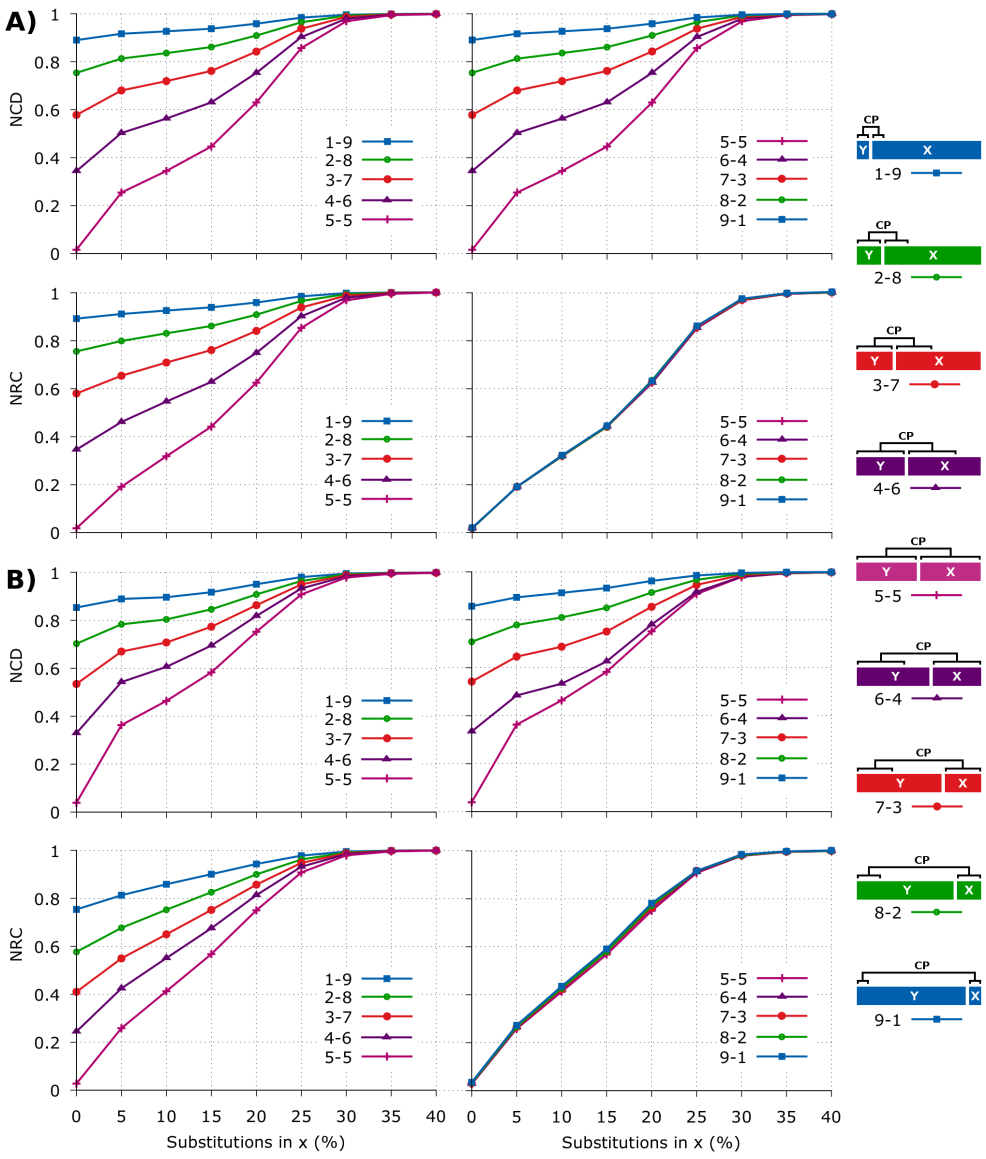

Figure 2. Comparison of the NCD (Equation (3)) and the NRC (Equation (5)) for several synthetic sequences with different substitutions applied on x. The sequences architecture is at right, where "CP" means copy. The substitutions in x are only applied after coping a region of y into x. Each pair, x and y, has a length of 1 MB. (**A**) the distribution of the sequences is uniform. For replication use the script runComparison.sh; (**B**) distribution is not uniform, and the sequences contain multiple repeats [105]. The numbers *a-b* stand for string sizes proportions, for example 1-9 means that y has size 0.1 MB and x 0.9 MB. For replication use runComparisonWithRedundancy.sh.

In Figure 3 we notice that, contrary to $C(x)$ and $C(yx)$, the $C(x\|y)$ is not able to measure similar sub-regions in x, namely because the region M, which is a copy of G, is measured as a high complexity region. This is what it means to use the exclusive information from y. In fact, ignoring constants, we are able to asymptotically generalize this as

$$C(x^n\|y) = nC(x\|y), \tag{8}$$

where x^n is the concatenation of n copies of x.

Additionally, we are able to notice that $C(x\|y)$ better describes the region J than $C(yx)$. This might be due to the parameter alpha that is more appropriate for the relative compression mode and because

of the model memory of $C(x\|y)$ is lower than $C(yx)$. The last is given the fact that $C(yx)$, besides modelling y, adds the events seen in x, while the $C(x\|y)$ only maps y.

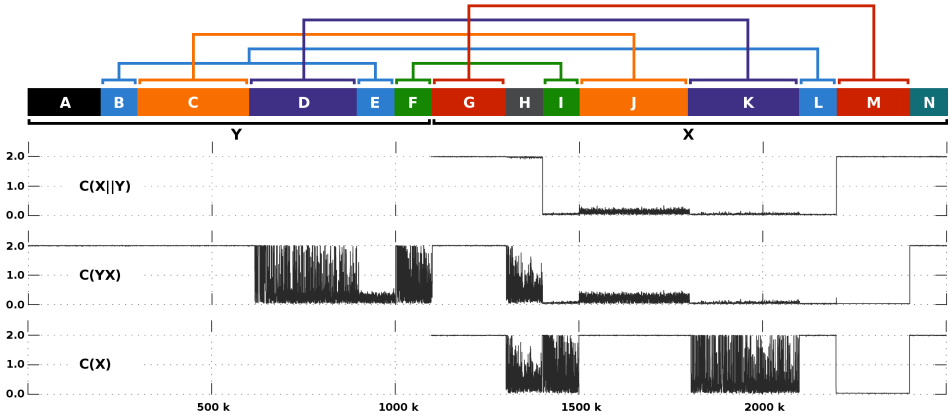

Figure 3. Comparison of the $C(x\|y)$ (top profile), the $C(yx)$ (middle profile) and the $C(x)$ (bottom profile) given several types of rearrangements between x and y. The upper map depicts the multiple block regions that compose x and y. The region A and N identify unmatched sequences with high entropy, while H an unmatched sequence with low entropy. Region E and L are a copy of B (high entropy), both with 1% of substitutional mutations. Region J is a copy of C (high entropy) with 1% of substitutional mutations. Region K is a copy of D, both with low entropy. Region I is a copy of F, both with low entropy. Region M is a copy of G, both with high entropy. The sequences have been generated using XS [105] and GOOSE (https://github.com/pratas/goose). For replication use runLocalMethod.sh.

3. Results

In this section, we use the NC, NCD, and NRC to measure the information in, between and across mtDNA, mRNA, and gDNA of seven primates. For the purpose, we provide the description of the primate dataset and the parameters used in the compressor as well as a benchmark of the compressor in different compression modes, then we make some previsions applying alterations on the datasets and, finally, provide the empirical results.

All the results presented in this paper can be reproduced, under a Linux OS, using the scripts provided at the repository https://github.com/pratas/APE, specifically runNC.sh, runNCD.sh, runNRC.sh, runReferenceFreeComparison.sh, runReferenceFreeConjoint.sh, runRelativeCompressors Comparison.sh, runExpectationNRC.sh, and runRearrange.sh.

3.1. Dataset

We have used the mtDNA, mRNA, and gDNA of seven primates according to the Table 1.

Table 1. Description of the dataset for the mtDNA, mRNA, and gDNA. All the sequences have been downloaded from the National Center for Biotechnology Information (NCBI).

Species	mtDNA		mRNA		gDNA	
	Length	Reference	Length	Version	Length	Version
Human	16,569	NC_012920.1	587,117,742	GRCh38.p7	2,948,627,755	GRCh38.p7
Chimpanzee	16,554	NC_001643.1	351,298,530	3.0	2,845,195,942	3.0
Gorilla	16,412	NC_011120.1	153,150,229	4	2,788,268,060	4
Orangutan	16,499	NC_002083.1	102,315,527	2.0.2	2,722,968,486	2.0.2
Gibbon	16,478	NC_021957.1	110,221,273	3.0	2,611,673,151	3.0
Baboon	16,516	NC_020006.2	312,140,410	3.0	2,727,993,489	3.0
Marmoset	16,499	NC_025586.1	172,133,747	3.2	2,618,690,967	3.2
Total	115,548	-	1,788,377,458	-	19,263,417,850	-

3.2. Parameters

The models and parameters used in the compressor are fundamental to ensure compression efficiency. The optimization of the parameters provides better compression results and, therefore, the minimum of a compression result is an indicator of the best-known set of parameters for a specific dataset. However, the optimization requires computational time that, given the length of the larger sequences, becomes prohibitive. For these results, we use a set of models and parameters built upon our experience. The parameters used for the compression of the primates DNA sequences, for the respective nature of the data, are the following:

- **mtDNA** $\rightarrow$ mixture of five models with a decayment (γ) of 0.95:

  1 tolerant context model: depth: 13, alpha: 0.1, tolerance: 5;
  2 context model: depth: 13, alpha: 0.005, inverted repeats: yes;
  3 context model: depth: 10, alpha: 0.01, inverted repeats: yes;
  4 context model: depth: 6, alpha: 1, inverted repeats: no;
  5 context model: depth: 3, alpha: 1, inverted repeats: no;

- **mRNA** $\rightarrow$ mixture of seven models with a decayment (γ) of 0.88 and a cache-hash of 200:

  1 tolerant context model: depth: 20, alpha: 0.1, tolerance: 5;
  2 context model: depth: 20, alpha: 0.005, inverted repeats: yes;
  3 context model: depth: 14, alpha: 0.02, inverted repeats: yes;
  4 context model: depth: 13, alpha: 0.05, inverted repeats: no;
  5 context model: depth: 11, alpha: 0.1, inverted repeats: no;
  6 context model: depth: 9, alpha: 1, inverted repeats: no;
  7 context model: depth: 4, alpha: 1, inverted repeats: no;

- **gDNA** $\rightarrow$ mixture of six models with a decayment (γ) of 0.88 and a cache-hash of 250:

  1 tolerant context model: depth: 20, alpha: 0.1, tolerance: 5;
  2 context model: depth: 20, alpha: 0.005, inverted repeats: yes;
  3 context model: depth: 14, alpha: 0.02, inverted repeats: yes;
  4 context model: depth: 13, alpha: 0.05, inverted repeats: no;
  5 context model: depth: 11, alpha: 0.1, inverted repeats: no;
  6 context model: depth: 9, alpha: 1, inverted repeats: no.

For the relative compression the parameters and models have been the same, with the exception of the alpha for context 14 (set to 0.01) and that the models were in the beginning loaded with the counts regarding y and, then, set static through all the computation of x (described in Section 2.1). The maximum RAM used was 34.3 GB (for the gDNA dataset, which has an approximate sum of 18 GB). Essentially, by a small precision payoff the RAM might be decreased, despite in this experience we wanted to cope with the full *normality* properties, namely the idempotency [32].

3.3. Comparison of Compressors

The NCD and NRC measures are extremely dependent on the capability for the compressor to reduce losslessly the storage associated with the respective sequences, independently if the running mode is reference-free or relative. The better the compression is, the more reliable are the results. This happens because the compressor acts as a description program of a string given what it knowns in each specific moment with the extra information (from y) when it is available. If the compressor is not efficient while describing the strings, then the computation of the measure will not be accurate, sometimes even showing misleading results. On the other hand, if the compressor is prepared to handle most of the characteristics of the data, such as in this case, genomic rearrangements (translocations, inversions, duplications, fissions, fusions), stochastic variation (especially, high level of substitutions), and high heterogeneity (high alteration between high and low complexity regions), then it is an efficient describer and an obvious candidate to be used in the NCD and NRC measures. Therefore, it is not just to cope with the normality characteristics, but also to be as efficient as possible for the data type.

The interest for DNA sequence compression was started with the Biocompress algorithm in 1993 [106]. The subsequent two and half decades have seen the publication of a considerable number of algorithms for reference-free compression of DNA sequences (e.g., [72,102,107–129]) and for reference-based compression (e.g., [72,75–77,118,126,130–137]). With the development of the next-generation sequencing and the increasing availability of genomic data [138], several compression algorithms have been developed to cope with the specific needs of special file formats, namely the inclusion of multiple information channels, for example, quality-scores, as well as adding increasing levels of redundancy [139–148]. For a review of biological sequence compression algorithms, see [149].

In order to guarantee that GeCo is an efficient and appropriate compressor for the task, we compare GeCo with some of the best known high-ratio reference-free and relative compressors. However, notice that since, in this paper, we also use large-scale sequences (several GB), and to cope with the normality, we excluded the usage of XM [118]. The XM method uses computational resources that are not affordable to use in a large-scale scenario. Nevertheless, it is worth to mention that we have found competitive results and, sometimes, marginal higher compression capabilities in XM relative to GeCo.

Figure 4 depicts a benchmark for several chromosome sequences of a human, chimpanzee and gorilla. As it can be seen, GeCo, with the models and parameters described in the previous subsection, with the exception of a cache-hash set to 50 (more appropriate to the size of the samples), achieved always the best compression results, namely 3.16% and 27.2% compression improvement over MFCompress in reference-free compression and reference-free conjoint compression (respectively), and 26% over GDC2 (the second best). The compression gain is calculated with $100(a-b)/a$, where a stands for the number of bytes compressed with the second best compressor and b for the number of bytes compressed with GeCo.

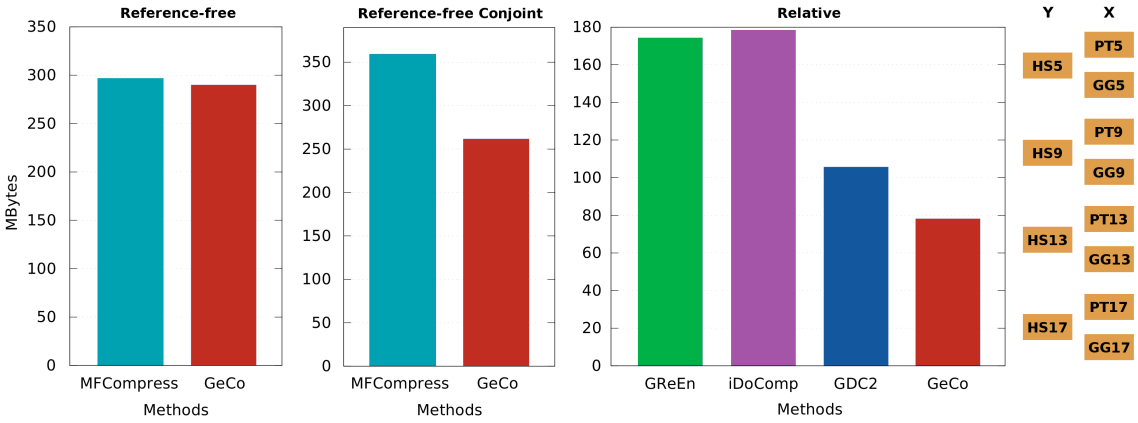

Figure 4. Number of MegaBytes needed for each compression tool to represent a lossless compact form of each dataset. Benchmark for three types of compression is provided: reference-free ($C(x)$), reference-free conjoint ($C(yx)$) and relative ($C(x\|y)$). The reference-free includes the compression of chromosome sequences corresponding to HS5, PT5, GG5, HS9, PT9, GG9, HS13, PT13, GG13, HS17, PT17, GG17. The reference-free conjoint, the HS5_PT5, HS5_GG5, HS9_PT9, HS9_GG9, HS13_PT13, HS13_GG13, HS17_PT17, HS17_GG17. The relative, the PT5-HS5, GG5-HS5, PT9-HS9, GG9-HS9, PT13-HS13, GG13-HS13, PT17-HS17, GG17-HS17. The prefix initials stand for species (HS→human, PT→chimpanzee, GG→gorilla), the "_" stand for concatenation, and "-" for "relative to". For replication use scripts runReferenceFreeComparison.sh, runReferenceFreeConjoint.sh and runRelativeCompressorsComparison.sh.

Although we have used four chromosome sequences of each human, chimpanzee and gorilla references, we believe that the compression gain will increase for chromosome sequences of more distant species, namely baboon and marmoset. The reason is that GeCo uses multiple Markov and Tolerant Markov models in cooperation and, hence, it is more suitable to dissimilar species relative to the other compressors. Of course, the penalty is computational resources, namely memory and

time. Regarding memory, GeCo needed 9 GB of RAM and used five times the time of the average compressors. Nevertheless, since the objective is to measure dissimilarity as best as we can, we believe that, currently, these resources are affordable to be used. Moreover, GeCo enables to set the models according to the available memory of the machine, independently of the size of the dataset to be compressed and respective reference. This includes the capability to be used, with good results, in even larger scale scenarios.

3.4. Expectation

In order to understand how the NRC behaves according to substitutional mutations using very different lengths, we have computed an experiment using the real mtDNA and gDNA of the human as a starting point and, then, we have applied different substitutional mutations rates to several copies of the original. Finally, for the respective scale (mtDNA or gDNA), we have computed the NRC using, as x, every copy and, as y, the human reference. Figure 5 depicts the result.

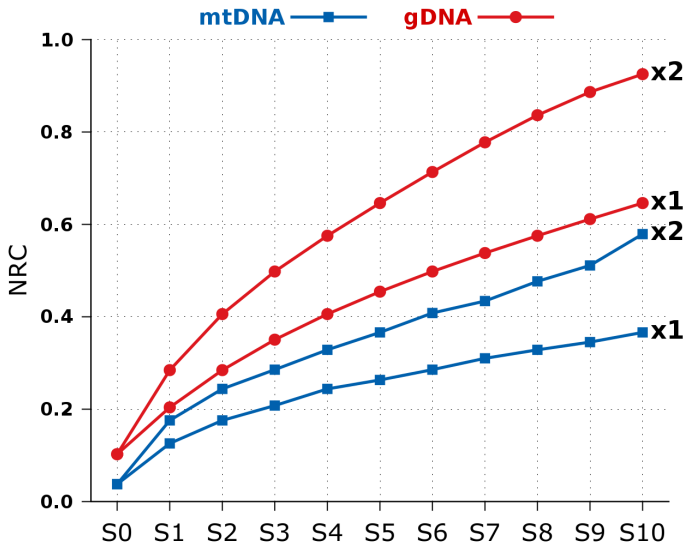

Figure 5. Normalized Relative Compression (Equation (5)) for several substitutional mutations applied to the human mtDNA and gDNA. The "x1" identifies the slope of the mutation rate of 1, while "x2" a 2%. The mutation rate at a given point is identified by multiplying the suffix number by the slope. For replication use runExpectationNRC.sh.

As it can be seen, the scale provides a different slope between the constant increasing of substitutional mutations. In this example, the slope two (x2) in the mtDNA seems more similar to the slope one (x1) in the gDNA. Notice that, unlike in the gDNA, the "x2" of the mutated mtDNA has more noise because the sequence has near only 16,000 symbols, which makes it more sensible and prone to collisions in pseudo-random attributions in the process of synthetically creating substitutional mutations.

3.5. Primate Analysis

In Figure 6, we provide the NCD and NRC for six primates, setting as y the human. The main idea was to use the highest quality sequence (human) as a reference in the NRC and because in Figure 2 we have seen that the NRC provides better comparative results when $|x| < |y|$. As expected the mtDNA and gDNA information values, both in the NCD and NRC, increase as the divergence of the species increase, since the genomes of the primates were ordered given the divergence time depicted in the common literature [150].

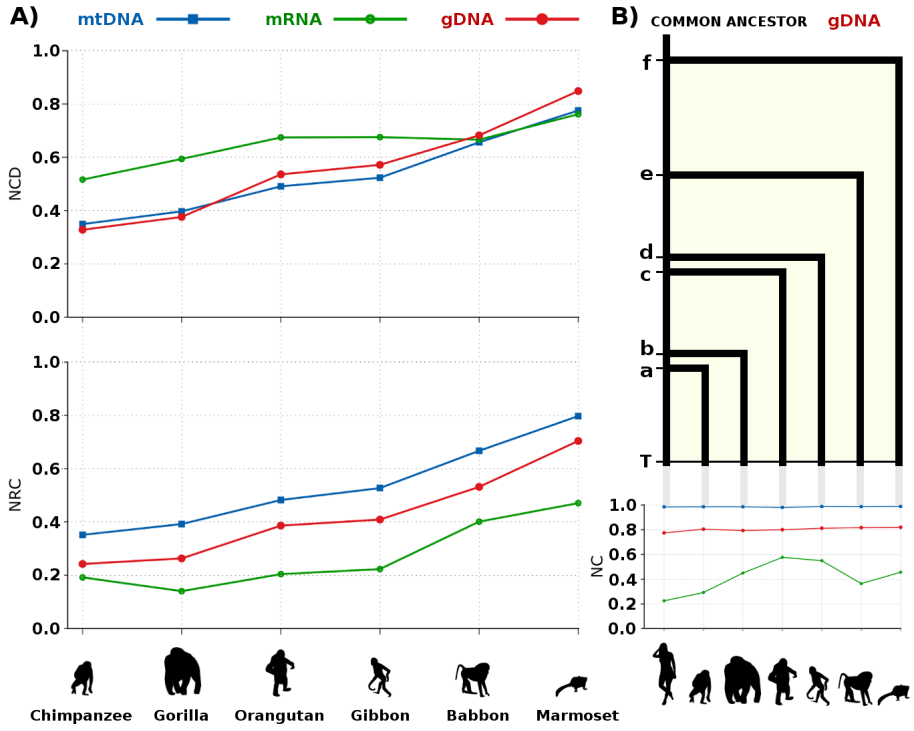

Figure 6. (**A**) Normalized Compression Distance (NCD at the upper plot, using Equation (3)) and Normalized Relative Compression (NRC at the lower plot, using Equation (5)) of mtDNA, mRNA, and gDNA sequences for several anthropoids in relation to the human genome. The gDNA represents the whole genome, including the unplaced and unlocalized sequences, with exception of the Y chromosome (female species); (**B**) Evolutionary tree of the gDNA is up to scale, based on the NRC. Letters a, b, c, d, e, and f represent the divergence time between the respective species, while T stands for the actual time. The NRC of the human relatively to the human has been subtracted from each result (≈0.1). All the genomes have been sequenced in T. The bottom right plot represents the Normalized Compression (NC, using Equation (1)) for each species. For replication use the scripts: runNCD.sh, runNRC.sh and runNC.sh.

In Figure 7, we show the cumulative time and maximum RAM needed to compute the mRNA and the gDNA for the respective measures (computation of Figure 6A). The time needed to compute the NRC is much lower than the NCD, especially in the larger dataset (gDNA) where it spent, roughly, one-fifth of the NCD. Notice that both computations used approximately the same RAM (for example, 34.3 GB in the largest dataset). We recall, that lower RAM may be used although having at some point impact in the precision of the measure. Although in this paper we have not studied this subject, this subject would be a thematic of a very interesting work.

Notice that both NCD and NRC are measuring the complete or whole information of the samples, that in the case of the gDNA is not a sampling of the genome, namely only the genes sequences but rather the entire DNA, such as the coding and non-coding regions. Besides, both are able to efficiently deal in an unsupervised mode with chromosomal rearrangements (translocations, inversions, duplications, fissions, fusions), high stochastic variation (especially high level of substitutions) and high heterogeneity of the data (high alteration between high and low complexity regions). This is something, as far as we know, that alignment-based methods are not able to cope using the reported times [97,98,151]. Moreover, the RAM necessary to work with two genomes is approximately the same as with 20. Besides, the computational time increases linearly with the number of input genomes.

Regarding the meaning results, we notice that the NRC for the mRNA of the gorilla is lower than for the chimpanzee. This could indicate that when compared to the human genome, parts that are more similar to the gorilla than to the chimpanzee. However, we have to take into account that the

gorilla mRNA sequence has approximately half the size of the chimpanzee mRNA sequence and lower redundancy (NCBI source), as depicted by the NC.

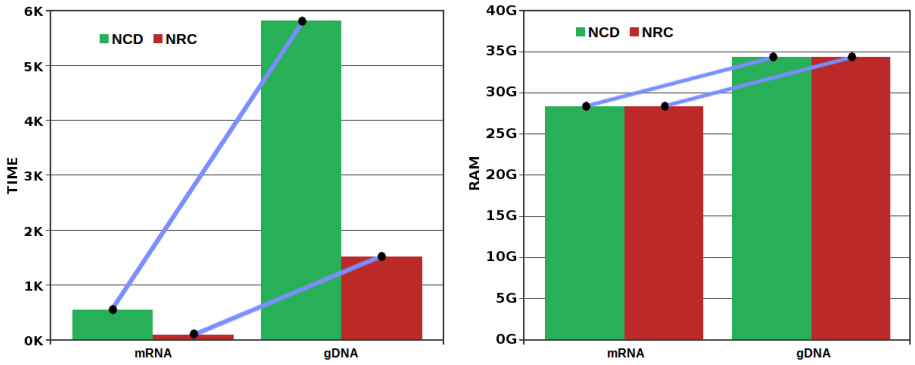

Figure 7. Time (**left**), in minutes, and RAM (**right**), in Gigabytes, needed to compute the NCD and NRC, for the mRNA and the gDNA, in all the measures of Figure 6A. The computation for the mtDNA spent only a few seconds and used less than 0.5 GB of RAM. Given the present orders of magnitude, is asymptotically irrelevant, and, hence, we have excluded from this image. The RAM needed to compute both measures was equivalent for each data type. All the computations were performed in a single core at 2.13 GHz (without parallelization). Unlike the NCD, the NRC can be easily parallelized while maintaining approximately the same RAM using an efficient speed-up.

On the other hand, the NCD shows that the information between the human-chimpanzee is lower than human-gorilla, which is according to the well-known theories in primates evolution [152–154]. There is also the hypothesis of hybridization [154], such as in the colobine monkeys, although unlikely, given that the NRC of the mtDNA and gDNA shows the opposite behaviour. Additionally, we recall that these sequences, besides having duplications [155,156], are outcomes of processing algorithms [157,158], that need to perform alignments with the gDNA and, therefore, they are dependent on pattern matching algorithms.

The mtDNA and gDNA are sequences which, given their nature, offer a more reliable and complete comparison between the measures, despite the difference in scale. As it can be seen in Figure 6, both show an approximate similar behaviour, although the mtDNA in the NCD has a higher slope (taking into account the differences in scales, analogous to the NRC in Figure 5).

In primates, mtDNA is maternally inherited, while nuclear DNA in offspring represents a combination between mother and father. Although there are studies that show that the rates of nucleotide substitution vary among mitochondrial and nuclear DNAs [159], there is evidence that, in mammals, the spontaneous mutation rate in the germline is lower than in somatic cells [160]. The exact rates of evolution between the mtDNA and the whole gDNA in primates are not known, although it is widely believed that they are higher in mtDNA than in gDNA, given several sampling analysis [160]. Despite these studies have previously calculated mutation rates in both mtDNA and gDNA for primate species and showed that mtDNA mutation rates are higher than gDNA, these are mostly based in a subset of genes and include only coding regions. To our best knowledge, the analysis of whole genomes, including non-coding regions, has not been done so far and is presented in this paper using two measures.

Recombination and mobile genetic elements, such as retrotransposons, shape the genomic architecture, namely through large alteration events such as inversions, translocations, fusions, and fissions, being a major factor for genome reshuffling [161]. However, rearranged chromosomes present significantly lower recombination rates than chromosomes that have been maintained since the last common ancestor of great apes [162], mostly because inverted regions have lower recombination rates than collinear and noninverted regions, independently of the effect of centromeres [162]. Therefore, major alterations create an effect of evolutionary deceleration. For a compression-based analysis of inverted repeats in several species, see [163].

When comparing Figure 5 with Figure 6, we notice that the gDNA seems to have near half of the variation of the mtDNA, although with evidence of changing acceleration events, but always maintaining an equilibrium through time. When computing an evolutionary map (Figure 6B), we see that the NRC approximates the events of speciation in chimpanzee (a) with the gorilla (b) and orangutan (c) with the gibbon (d).

One of the applications of the relative compression is to localize and connect similar regions in DNA sequences through the computation of Equation (4). These regions are usually associated with rearrangements [71]. The conjoint information, combined with the information contained is not suitable for this purpose, namely because it does not ignore repetitive regions on the targets, originating a hard problem to delimitate the regions. On the other hand, the relative compression ignores repetitive regions in x, as Equation (8) shows.

In Figure 8, we use an approximation of the relative compression to compute relative similarities between the mtDNA of the seven primates, ordered by time of speciation. The figure shows a decrease in the number of connections while the time of speciation increases. Two reasons may be related to this. First, since the speciation time is higher, the variation accumulated is higher through time and, therefore, the similarity is lower. The second reason might be related to the size of the species, namely because small species seem to have higher evolution rates [164].

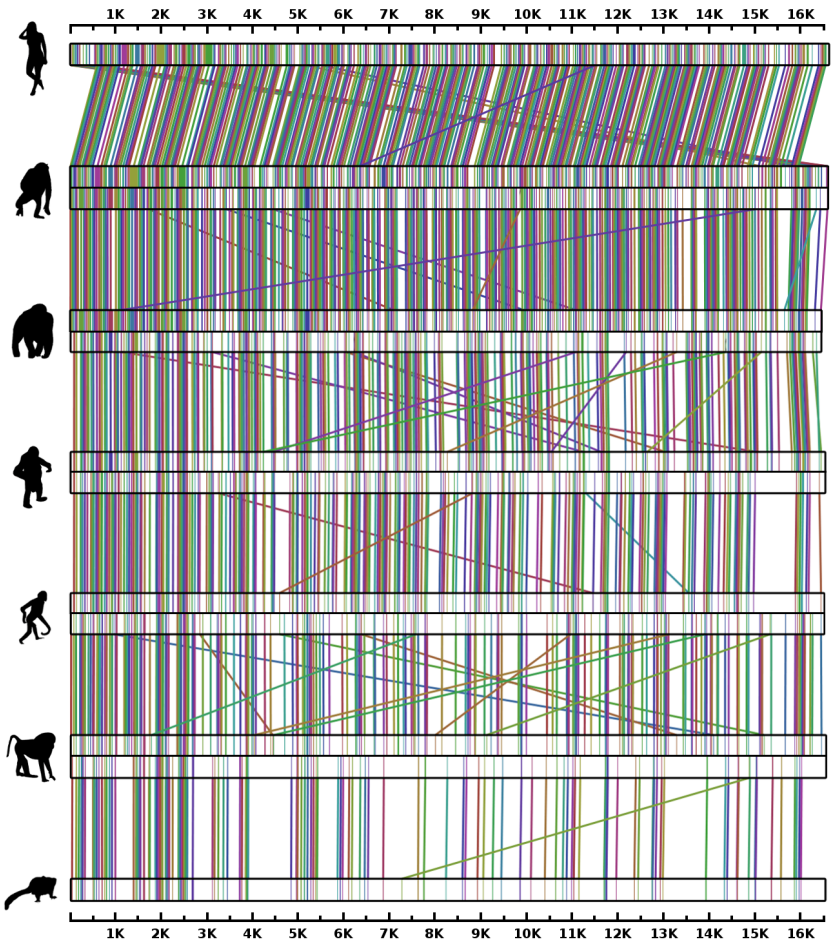

Figure 8. Patterns of similarity between mtDNA from different anthropoids, estimated with relative compression technology. From the top to the bottom: human-chimpanzee, chimpanzee-gorilla, gorilla-orangutan, orangutan-gibbon, gibbon-baboon, baboon-marmoset. The maps are depicted according to the output of the SMASH-contigs tool. This tool uses a simplified version of the computation of Equation (4). For replication use the script: runRearrange.sh.

Interestingly, a region of the gorilla mtDNA has similarity with other regions, from distant region positions, relative to both chimpanzee and orangutan. The same happens in the gDNA, where one of the major rearrangement sources between the gorilla and both chimpanzee and orangutan is a chromosomal translocation between chromosomes 15 and five [153]. We recall that, besides the NUMTS, minimal correlations across both are known [165–167].

4. Discussion and Conclusions

In this paper, we have directly compared two measures, the Normalized Compression Distance (NCD) and the Normalized Relative Compression (NRC). Although related, both are two normalized measures to answer different questions. The NCD measures the information distance between two strings, independently from the direction. The NRC measures the information of a string x relatively to y, which may be different from measuring the information of y relatively to x.

As an analogy, consider that an observer is assisting to the run of one motorcyclist and a cyclist. The observer wants to quantify the amount of information needed to describe one given the other. The NCD aims to quantify the mutual information of both runners, accounting the amount of descriptive surprise that each runner has regarding the other, namely the use of a wheel, helmet, motor, pedal, etc. However, it only considers these elements once, since the amount of information between two elements, for example, wheels is minimal.

On the other hand, the NRC selects a runner as a reference, for example, the motorcyclist. After, models him (creating an informative representation of him), and, then, measures the amount of surprise that is seen on the cyclist, regardless if he has repetitive elements, such as two pedals. In this case, the information describing the repetitive features on the cyclist, that have not been seen in the motorcyclist, may be added to the final count after being multiplied by the first repetitive element seen. Therefore, the relative compression will ignore if the regions on the target are repetitive or not (Equation (8)). The normalization of the relative compression will, then, act as a rescaling operation, decreasing more the importance of the repetitions.

Notice that $C(x\|y) \geq C(x|y)$ and $|x| \geq C(x)$. The $|x|$ usually represents a large quantity and, therefore, the NRC is usually less than the NCD, specially when the $C(x)$ is much smaller than $|x|$. The absence of self-redundancy in the NRC and the relation between x and y seem to be the major factors for the NRC usually to be smaller than the NCD. In the real large examples (mRNA and gDNA), we have consistently found this.

Regarding computational time, the NRC uses, consistently, less time to be computed than the NCD. In the larger dataset (gDNA), we found that the computation of the NRC is roughly one-fifth of the NCD. Regarding memory, both can use the same memory, although the NCD needs more precision to store the memory model of the x and y, unlike the NRC that only needs to store the memory model of y.

The NRC is also able to efficiently detect and connect similar regions of information. The NRC is also able to produce a better adaptation, using lower computational resources, to the nature of the data given its simplicity. However, the NRC requires a specific relative compressor, while the NCD can be computed through the conjoint information, approximated by the concatenation of x and y.

We have compared both measures in genomic sequences with different scales and natures. These included mitochondrial DNA (mtDNA), messenger RNA (mRNA) and whole genome DNA (gDNA) of seven primates. With this approach, we provided several insights into evolutionary acceleration rates between different scales, namely a higher (near the double) variation rate of the mtDNA, relatively to the gDNA in these primates. Notice that the usage of alignment methods to estimate such measures in whole genomes is dependent on the capability to cope with genome rearrangements and high variation. In fact, quantifying dissimilarity of sequences with alignment methods is problematic, due to the need for fine-tuned thresholds, considering relaxed edit distances and, consequently, the increase of computational cost. More important, the choice of the thresholds have the problem of how to quantify dissimilarity without producing overestimated measures, namely measures that cope with an approximation of the information present and not an overestimation of the

same value [73]. Therefore, we believe that compression-based normalized measures, independently of the use of NRC or NCD, are the natural measures to quantify such information.

Finally, we have shown a practical example of the relative compression for localizing similar information regions using the mtDNA of seven primates. This is an application where we have found that the relative mode question is more suitable for what we want to address.

Author Contributions: D.P., R.M.S. and A.J.P. conceived and designed the experiments; D.P. performed the experiments; D.P., R.M.S. and A.J.P. analyzed the data; D.P., R.M.S. and A.J.P. wrote the paper.

Acknowledgments: This work was partially funded by FEDER (Programa Operacional Factores de Competitividade-COMPETE) and by National Funds through the FCT-Foundation for Science and Technology, in the context of the projects UID/CEC/00127/2013, UID/BIM/04501/2013, POCI-01-0145-FEDER-007628, PTCD/EEI-SII/6608/2014, and the grant SFRH/BPD/111148/2015 to RMS.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NC	Normalized Compression
NCD	Normalized Compression Distance
NRC	Normalized Relative Compression
DNA	Deoxyribonucleic acid
mtDNA	mitochondrial DNA
gDNA	nuclear DNA
cpDNA	chloroplast DNA
mRNA	messenger Ribonucleic acid
SNP	Single Nucleotide Polymorphism
GeCo	Genomic Compressor (tool)

References

1. Kolmogorov, A.N. Three approaches to the quantitative definition of information. *Probl. Inf. Transm.* **1965**, *1*, 1–7. [CrossRef]
2. Niven, R.K. Combinatorial entropies and statistics. *Eur. Phys. J. B* **2009**, *70*, 49–63. [CrossRef]
3. Mantaci, S.; Restivo, A.; Rosone, G.; Sciortino, M. A new combinatorial approach to sequence comparison. *Theory Comput. Syst.* **2008**, *42*, 411–429. [CrossRef]
4. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. [CrossRef]
5. Solomonoff, R.J. A formal theory of inductive inference. Part I. *Inf. Control* **1964**, *7*, 1–22. [CrossRef]
6. Solomonoff, R.J. A formal theory of inductive inference. Part II. *Inf. Control* **1964**, *7*, 224–254. [CrossRef]
7. Chaitin, G.J. On the length of programs for computing finite binary sequences. *J. ACM* **1966**, *13*, 547–569. [CrossRef]
8. Wallace, C.S.; Boulton, D.M. An information measure for classification. *Comput. J.* **1968**, *11*, 185–194. [CrossRef]
9. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471. [CrossRef]
10. Hutter, M. Algorithmic information theory: A brief non-technical guide to the field. *arXiv* **2004**, arXiv:cs/0703024. [CrossRef]
11. Li, M.; Vitányi, P. *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed.; Springer: Berlin, Germany, 2008.
12. Levin, L.A. Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problemy Peredachi Informatsii* **1974**, *10*, 30–35.
13. Shen, A.; Uspensky, V.A.; Vereshchagin, N. *Kolmogorov Complexity and Algorithmic Randomness*; American Mathematical Society: Providence, RI, USA, 2017.
14. Hammer, D.; Romashchenko, A.; Shen, A.; Vereshchagin, N. Inequalities for Shannon entropy and Kolmogorov complexity. *J. Comput. Syst. Sci.* **2000**, *60*, 442–464. [CrossRef]

15. Henriques, T.; Gonçalves, H.; Antunes, L.; Matias, M.; Bernardes, J.; Costa-Santos, C. Entropy and compression: Two measures of complexity. *J. Eval. Clin. Pract.* **2013**, *19*, 1101–1106. [CrossRef] [PubMed]
16. Soler-Toscano, F.; Zenil, H.; Delahaye, J.P.; Gauvrit, N. Calculating Kolmogorov complexity from the output frequency distributions of small Turing machines. *PLoS ONE* **2014**, *9*, e96223. [CrossRef] [PubMed]
17. Soler-Toscano, F.; Zenil, H. A computable measure of algorithmic probability by finite approximations with an application to integer sequences. *Complexity* **2017**, *2017*, 7208216. [CrossRef]
18. Gauvrit, N.; Zenil, H.; Soler-Toscano, F.; Delahaye, J.P.; Brugger, P. Human behavioral complexity peaks at age 25. *PLoS Comput. Biol.* **2017**, *13*, e1005408. [CrossRef] [PubMed]
19. Pratas, D.; Pinho, A.J. On the Approximation of the Kolmogorov Complexity for DNA Sequences. In Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis, Faro, Portugal, 20–23 June 2017; Springer: Berlin, Germany, 2017; pp. 259–266.
20. Kettunen, K.; Sadeniemi, M.; Lindh-Knuutila, T.; Honkela, T. Analysis of EU languages through text compression. In *Advances in Natural Language Processing*; Springer: Berlin, Germany, 2006; pp. 99–109.
21. Terwijn, S.A.; Torenvliet, L.; Vitányi, P.M.B. Nonapproximability of the normalized information distance. *J. Comput. Syst. Sci.* **2011**, *77*, 738–742. [CrossRef]
22. Rybalov, A. On the strongly generic undecidability of the halting problem. *Theor. Comput. Sci.* **2007**, *377*, 268–270. [CrossRef]
23. Bloem, P.; Mota, F.; de Rooij, S.; Antunes, L.; Adriaans, P. A safe approximation for Kolmogorov complexity. In Proceedings of the International Conference on Algorithmic Learning Theory, Bled, Slovenia, 8–10 October 2014; Springer: Berlin, Germany, 2014; pp. 336–350.
24. Bennett, C.H.; Gács, P.; Vitányi, M.L.P.M.B.; Zurek, W.H. Information distance. *IEEE Trans. Inf. Theory* **1998**, *44*, 1407–1423. [CrossRef]
25. Li, M.; Chen, X.; Li, X.; Ma, B.; Vitányi, P.M.B. The similarity metric. *IEEE Trans. Inf. Theory* **2004**, *50*, 3250–3264. [CrossRef]
26. Cilibrasi, R.; Vitányi, P.M.B. Clustering by compression. *IEEE Trans. Inf. Theory* **2005**, *51*, 1523–1545. [CrossRef]
27. Ferragina, P.; Giancarlo, R.; Greco, V.; Manzini, G.; Valiente, G. Compression-based classification of biological sequences and structures via the universal similarity metric: Experimental assessment. *BMC Bioinform.* **2007**, *8*, 252. [CrossRef] [PubMed]
28. El-Dirany, M.; Wang, F.; Furst, J.; Rogers, J.; Raicu, D. Compression-based distance methods as an alternative to statistical methods for constructing phylogenetic trees. In Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, China, 15–18 December 2016; pp. 1107–1112.
29. Nikvand, N.; Wang, Z. Generic image similarity based on Kolmogorov complexity. In Proceedings of the 2010 17th IEEE International Conference on Image Processing (ICIP-2010), Hong Kong, China, 26–29 September 2010; pp. 309–312.
30. Pratas, D.; Pinho, A.J. A conditional compression distance that unveils insights of the genomic evolution. In Proceedings of the Data Compression Conference (DCC-2014), Snowbird, UT, USA, 26–28 March 2014.
31. Cebrián, M.; Alfonseca, M.; Ortega, A. The normalized compression distance is resistant to noise. *IEEE Trans. Inform. Theory* **2007**, *53*, 1895–1900. [CrossRef]
32. Cebrián, M.; Alfonseca, M.; Ortega, A. Common pitfalls using the normalized compression distance: What to watch out for in a compressor. *Commun. Inf. Syst.* **2005**, *5*, 367–384.
33. Seaward, L.; Matwin, S. Intrinsic plagiarism detection using complexity analysis. In Proceedings of the SEPLN, San Sebastian, Spain, 8–10 September 2009; pp. 56–61.
34. Merivuori, T.; Roos, T. Some Observations on the Applicability of Normalized Compression Distance to Stemmatology. In Proceedings of the Second Workshop on Information Theoretic Methods in Science and Engineering, Tampere, Finland, 17–19 August 2009.
35. Antão, R.; Mota, A.; Machado, J.T. Kolmogorov complexity as a data similarity metric: Application in mitochondrial DNA. *Nonlinear Dyn.* **2018**, *4*, 1–13. [CrossRef]
36. Pratas, D.; Pinho, A.J.; Garcia, S.P. Computation of the Normalized Compression Distance of DNA Sequences using a Mixture of Finite-context Models. In Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS-2012), Algarve, Portugal, 1–4 February 2012; pp. 308–311.

37. La Rosa, M.; Rizzo, R.; Urso, A.; Gaglio, S. Comparison of genomic sequences clustering using Normalized Compression Distance and evolutionary distance. In Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Zagreb, Croatia, 3–5 September 2008; Springer: Berlin, Germany, 2008; pp. 740–746.
38. Nykter, M.; Yli-Harja, O.; Shmulevich, I. Normalized Compression Distance for gene expression analysis. In Proceedings of the Workshop on Genomic Signal Processing and Statistics (GENSIPS), Newport, RI, USA, 22–24 May 2005.
39. Nykter, M.; Price, N.D.; Aldana, M.; Ramsey, S.A.; Kauffman, S.A.; Hood, L.E.; Yli-Harja, O.; Shmulevich, I. Gene expression dynamics in the macrophage exhibit criticality. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 1897–1900. [CrossRef] [PubMed]
40. Mihailović, D.T.; Mimić, G.; Nikolić-Djorić, E.; Arsenić, I. Novel measures based on the Kolmogorov complexity for use in complex system behavior studies and time series analysis. *Open Phys.* **2015**, *13*. [CrossRef]
41. Tran, N. The normalized compression distance and image distinguishability. In Proceedings of the SPIE Human Vision and Electronic Imaging XII, San Jose, CA, USA, 29 January–1 February 2007; p. 64921D.
42. Coltuc, D.; Datcu, M.; Coltuc, D. On the Use of Normalized Compression Distances for Image Similarity Detection. *Entropy* **2018**, *20*, 99. [CrossRef]
43. Pinho, A.J.; Ferreira, P.J.S.G. Image similarity using the normalized compression distance based on finite context models. In Proceedings of the 2011 18th IEEE International Conference on Image Processing (ICIP-2011), Brussels, Belgium, 11–14 September 2011 .
44. Vázquez, P.P.; Marco, J. Using Normalized Compression Distance for image similarity measurement: An experimental study. *Vis. Comput.* **2012**, *28*, 1063–1084. [CrossRef]
45. Nikvand, N.; Wang, Z. Image distortion analysis based on normalized perceptual information distance. *Signal Image Video Process.* **2013**, *7*, 403–410. [CrossRef]
46. Telles, G.P.; Minghim, R.; Paulovich, F.V. Normalized compression distance for visual analysis of document collections. *Comput. Graph.* **2007**, *31*, 327–337. [CrossRef]
47. Axelsson, S. Using Normalized Compression Distance for classifying file fragments. In Proceedings of the ARES'10 International Conference on Availability, Reliability, and Security, Krakow, Poland, 15–18 February 2010; pp. 641–646.
48. Cohen, A.R.; Vitányi, P. Normalized compression distance of multisets with applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1602–1614. [CrossRef] [PubMed]
49. Cilibrasi, R.; Vitányi, P.; Wolf, R.D. Algorithmic clustering of music based on string compression. *Comput. Music J.* **2004**, *28*, 49–67. [CrossRef]
50. Alfonseca, M.; Cebrián Ramos, M.; Ortega, A. Evolving computer-generated music by means of the Normalized Compression Distance. In Proceedings of the 5th WSEAS Conference on Simulation, Modeling and Optimization (SMO '05), Corfu Island, Greece, 17–19 August 2005.
51. Foster, P.; Dixon, S.; Klapuri, A. Identifying cover songs using information-theoretic measures of similarity. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* **2015**, *23*, 993–1005. [CrossRef]
52. Klenk, S.; Thom, D.; Heidemann, G. The Normalized Compression Distance as a distance measure in entity identification. In Proceedings of the Industrial Conference on Data Mining, Miami, FL, USA, 6–9 December 2009; Springer: Berlin, Germany, 2009; pp. 325–337.
53. Yoshizawa, S.; Terano, T.; Yoshikawa, A. Assessing the impact of student peer review in writing instruction by using the Normalized Compression Distance. *IEEE Trans. Prof. Commun.* **2012**, *55*, 85–96. [CrossRef]
54. Bailey, M.; Oberheide, J.; Andersen, J.; Mao, Z.M.; Jahanian, F.; Nazario, J. Automated classification and analysis of internet malware. In Proceedings of the International Workshop on Recent Advances in Intrusion Detection, Gold Coast, Australia, 5–7 September 2007; Springe: Berlin, Germany, 2007; pp. 178–197.
55. Borbely, R.S. On Normalized Compression Distance and large malware. *J. Comput. Virol. Hacking Tech.* **2016**, *12*, 235–242. [CrossRef]
56. Threm, D.; Yu, L.; Ramaswamy, S.; Sudarsan, S.D. Using Normalized Compression Distance to measure the evolutionary stability of software systems. In Proceedings of the 2015 IEEE 26th International Symposium on Software Reliability Engineering (ISSRE), Gaithersbury, MD, USA, 2–5 November 2015; pp. 112–120.

57. Henard, C.; Papadakis, M.; Harman, M.; Jia, Y.; Le Traon, Y. Comparing white-box and black-box test prioritization. In Proceedings of the 2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE), Austin, TX, USA, 14–22 May 2016; pp. 523–534.
58. Martins, L.G.; Nobre, R.; Cardoso, J.M.; Delbem, A.C.; Marques, E. Clustering-based selection for the exploration of compiler optimization sequences. *ACM Trans. Archit. Code Optim. (TACO)* **2016**, *13*, 8. [CrossRef]
59. Rios, R.A.; Lopes, C.S.; Sikansi, F.H.; Pagliosa, P.A.; de Mello, R.F. Analyzing the Public Opinion on the Brazilian Political and Corruption Issues. In Proceedings of the 2017 Brazilian Conference on Intelligent Systems (BRACIS), Uberlandia, Brazil, 2–5 October 2017; pp. 13–18.
60. Ting, C.L.; Fisher, A.N.; Bauer, T.L. Compression-Based Algorithms for Deception Detection. In Proceedings of the International Conference on Social Informatics, Oxford, UK, 13–15 September 2017; Springer: Berlin, Germany, 2017; pp. 257–276.
61. Cerra, D.; Israel, M.; Datcu, M. Parameter-free clustering: Application to fawns detection. In Proceedings of the 2009 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2009), Cape Town, South Africa, 12–17 July 2009.
62. Ziv, J.; Merhav, N. A measure of relative entropy between individual sequences with application to universal classification. *IEEE Trans. Inf. Theory* **1993**, *39*, 1270–1279. [CrossRef]
63. Cerra, D.; Datcu, M. Algorithmic relative complexity. *Entropy* **2011**, *13*, 902–914. [CrossRef]
64. Pratas, D. Compression and Analysis of Genomic Data. Ph.D. Thesis, University of Aveiro, Aveiro, Portugal, 2016.
65. Helmer, S.; Augsten, N.; Böhlen, M. Measuring structural similarity of semistructured data based on information-theoretic approaches. *VLDB J. Int. J. Very Large Data Bases* **2012**, *21*, 677–702. [CrossRef]
66. Cerra, D.; Datcu, M. Expanding the algorithmic information theory frame for applications to Earth observation. *Entropy* **2013**, *15*, 407–415. [CrossRef]
67. Cerra, D.; Datcu, M.; Reinartz, P. Authorship analysis based on data compression. *Pattern Recognit. Lett.* **2014**, *42*, 79–84. [CrossRef]
68. Coutinho, D.P.; Figueiredo, M. Text Classification Using Compression-Based Dissimilarity Measures. *Int. J. Pattern Recognit. Artif. Intell.* **2015**, *29*, 1553004. [CrossRef]
69. Pinho, A.J.; Pratas, D.; Ferreira, P.J.S.G. Authorship attribution using relative compression. In Proceedings of the Data Compression Conference (DCC-2016), Snowbird, UT, USA, 29 March–1 April 2016.
70. Brás, S.; Ferreira, J.H.T.; Soares, S.C.; Pinho, A.J. Biometric and emotion identification: An ECG compression based method. *Front. Psychol.* **2018**, *9*, 467. [CrossRef] [PubMed]
71. Pratas, D.; Silva, R.M.; Pinho, A.J.; Ferreira, P.J.S.G. An alignment-free method to find and visualise rearrangements between pairs of DNA sequences. *Sci. Rep.* **2015**, *5*, 10203. [CrossRef] [PubMed]
72. Pratas, D.; Pinho, A.J.; Ferreira, P.J.S.G. Efficient compression of genomic sequences. In Proceedings of the Data Compression Conference (DCC-2016), Snowbird, UT, USA, 29 March–1 April 2016; pp. 231–240.
73. Pratas, D.; Pinho, A.J.; Silva, R.M.; Rodrigues, J.M.O.S.; Hosseini, M.; Caetano, T.; Ferreira, P.J.S.G. FALCON-meta: A method to infer metagenomic composition of ancient DNA. *bioRxiv* **2018**, 267179. [CrossRef]
74. Coutinho, D.; Figueiredo, M. An information theoretic approach to text sentiment analysis. In Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM), Barcelona, Spain, 15–18 February 2013; pp. 577–580.
75. Pinho, A.J.; Pratas, D.; Garcia, S.P. GReEn: A tool for efficient compression of genome resequencing data. *Nucleic Acids Res.* **2012**, *40*, e27. [CrossRef] [PubMed]
76. Wandelt, S.; Leser, U. FRESCO: Referential compression of highly similar sequences. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2013**, *10*, 1275–1288. [CrossRef] [PubMed]
77. Liu, Y.; Peng, H.; Wong, L.; Li, J. High-speed and high-ratio referential genome compression. *Bioinformatics* **2017**, *33*, 3364–3372. [CrossRef] [PubMed]
78. Dawy, Z.; Hagenauer, J.; Hoffmann, A. Implementing the context tree weighting method for content recognition. In Proceedings of the Data Compression Conference (DCC-2004), Snowbird, UT, USA, 23–25 March 2004.
79. Darwin, C.; Bynum, W.F. *The Origin of Species by Means of Natural Selection: Or, The Preservation of Favored Races in the Struggle for Life*; John Murray: London, UK, 1859.
80. Huxley, T.H. *Evidence as to Mans Place in Nature by Thomas Henry Huxley*; Williams and Norgate: London, UK, 1863.

81. Delsuc, F.; Brinkmann, H.; Philippe, H. Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* **2005**, *6*, 361–375. [CrossRef] [PubMed]
82. Wolf, Y.I.; Rogozin, I.B.; Grishin, N.V.; Koonin, E.V. Genome trees and the tree of life. *Trends Genet.* **2002**, *18*, 472–479. [CrossRef]
83. Tomkins, J. How genomes are sequenced and why it matters: Implications for studies in comparative genomics of humans and chimpanzees. *Answ. Res. J.* **2011**, *4*, 81–88.
84. O'Rawe, J.A.; Ferson, S.; Lyon, G.J. Accounting for uncertainty in DNA sequencing data. *Trends Genet.* **2015**, *31*, 61–66. [CrossRef] [PubMed]
85. Henn, B.M.; Botigué, L.R.; Bustamante, C.D.; Clark, A.G.; Gravel, S. Estimating the mutation load in human genomes. *Nat. Rev. Genet.* **2015**, *16*, 333–343. [CrossRef] [PubMed]
86. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 3439–3444. [CrossRef] [PubMed]
87. Jeong, C.; Di Rienzo, A. Adaptations to local environments in modern human populations. *Curr. Opin. Genet. Dev.* **2014**, *29*, 1–8. [CrossRef] [PubMed]
88. Beres, S.; Kachroo, P.; Nasser, W.; Olsen, R.; Zhu, L.; Flores, A.; de la Riva, I.; Paez-Mayorga, J.; Jimenez, F.; Cantu, C.; et al. Transcriptome remodeling contributes to epidemic disease caused by the human pathogen *Streptococcus pyogenes*. *MBio* **2016**, *7*, e00403-16. [CrossRef] [PubMed]
89. Fumagalli, M.; Sironi, M. Human genome variability, natural selection and infectious diseases. *Curr. Opin. Immunol.* **2014**, *30*, 9–16. [CrossRef] [PubMed]
90. Rieseberg, L.H. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **2001**, *16*, 351–358. [CrossRef]
91. Roeder, G.S.; Fink, G.R. DNA rearrangements associated with a transposable element in yeast. *Cell* **1980**, *21*, 239–249. [CrossRef]
92. Long, H.; Sung, W.; Kucukyildirim, S.; Williams, E.; Miller, S.F.; Guo, W.; Patterson, C.; Gregory, C.; Strauss, C.; Stone, C.; et al. Evolutionary determinants of genome-wide nucleotide composition. *Nat. Ecol. Evol.* **2018**, *2*, 237–240. [CrossRef] [PubMed]
93. Golan, A. *Foundations of Info-Metrics: Modeling and Inference with Imperfect Information*; Oxford University Press: Oxford, UK, 2017.
94. Gray, M.W. The evolutionary origins of organelles. *Trends Genet.* **1989**, *5*, 294–299. [CrossRef]
95. Seligmann, H. Alignment-based and alignment-free methods converge with experimental data on amino acids coded by stop codons at split between nuclear and mitochondrial genetic codes. *Biosystems* **2018**, *167*, 33–46. [CrossRef] [PubMed]
96. Kimura, M. *The Neutral Theory of Molecular Evolution*; Cambridge University Press: Cambridge, UK, 1983.
97. Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W.M. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* **2017**, *18*, 186. [CrossRef] [PubMed]
98. Ren, J.; Bai, X.; Lu, Y.Y.; Tang, K.; Wang, Y.; Reinert, G.; Sun, F. Alignment-Free Sequence Analysis and Applications. *Annu. Rev. Biomed. Data Sci.* **2018**, *1*, arXiv:1803.09727v1. [CrossRef]
99. Ferreira, P.J.S.G.; Pinho, A.J. Compression-based normal similarity measures for DNA sequences. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-2014, Florence, Italy, 4–9 May 2014; pp. 419–423.
100. Pratas, D.; Hosseini, M.; Pinho, A.J. Substitutional Tolerant Markov Models for Relative Compression of DNA Sequences. In Proceedings of the 11th International Conference on Practical Applications of Computational Biology & Bioinformatics; Porto, France, 21–23 June 2017; Springer: Berlin, Germany, 2017; pp. 265–272.
101. Bell, T.C.; Cleary, J.G.; Witten, I.H. *Text Compression*; Prentice Hall: Upper Saddle River, NJ, USA, 1990.
102. Pinho, A.J.; Pratas, D.; Ferreira, P.J.S.G. Bacteria DNA sequence compression using a mixture of finite-context models. In Proceedings of the IEEE Workshop on Statistical Signal Processing, Nice, France, 28–30 June 2011.
103. Sayood, K. *Introduction to Data Compression*; Morgan Kaufmann: Burlington, MA, USA, 2017.
104. Pratas, D.; Pinho, A.J. Exploring deep Markov models in genomic data compression using sequence pre-analysis. In Proceedings of the 22nd European Signal Processing Conference (EUSIPCO-2014), Lisbon, Portugal, 1–5 September 2014; pp. 2395–2399.
105. Pratas, D.; Pinho, A.J.; Rodrigues, J.M.O.S. XS: A FASTQ read simulator. *BMC Res. Notes* **2014**, *7*, 40. [CrossRef] [PubMed]

106. Grumbach, S.; Tahi, F. Compression of DNA sequences. In Proceedings of the Data Compression Conference (DCC-93), Snowbird, UT, USA, 30 March–1 April 1993; pp. 340–350.
107. Grumbach, S.; Tahi, F. A new challenge for compression algorithms: Genetic sequences. *Inf. Process. Manag.* **1994**, *30*, 875–886. [CrossRef]
108. Rivals, E.; Delahaye, J.P.; Dauchet, M.; Delgrange, O. A guaranteed compression scheme for repetitive DNA sequences. In Proceedings of the Data Compression Conference (DCC-96), Snowbird, UT, USA, 31 March–3 April 1996; p. 453.
109. Loewenstern, D.; Yianilos, P.N. Significantly lower entropy estimates for natural DNA sequences. In Proceedings of the Data Compression Conference (DCC-97), Snowbird, UT, USA, 25–27 March 1997; pp. 151–160.
110. Matsumoto, T.; Sadakane, K.; Imai, H. Biological sequence compression algorithms. *Genome Inform.* **2000**, *11*, 43–52.
111. Chen, X.; Kwong, S.; Li, M. A compression algorithm for DNA sequences. *IEEE Eng. Med. Biol. Mag.* **2001**, *20*, 61–66. [CrossRef]
112. Chen, X.; Li, M.; Ma, B.; Tromp, J. DNACompress: Fast and effective DNA sequence compression. *Bioinformatics* **2002**, *18*, 1696–1698. [CrossRef] [PubMed]
113. Tabus, I.; Korodi, G.; Rissanen, J. DNA sequence compression using the normalized maximum likelihood model for discrete regression. In Proceedings of the Data Compression Conference (DCC-2003), Snowbird, UT, USA, 25–27 March 2003; pp. 253–262.
114. Manzini, G.; Rastero, M. A simple and fast DNA compressor. *Softw. Pract. Exp.* **2004**, *34*, 1397–1411. [CrossRef]
115. Korodi, G.; Tabus, I. An efficient normalized maximum likelihood algorithm for DNA sequence compression. *ACM Trans. Inform. Syst.* **2005**, *23*, 3–34. [CrossRef]
116. Behzadi, B.; Le Fessant, F. DNA compression challenge revisited. In Proceedings of the Combinatorial Pattern Matching, CPM-2005; Jeju Island, Korea, 19–22 June 2005; Springer: Jeju Island, Korea, 2005; Volume 3537, pp. 190–200.
117. Korodi, G.; Tabus, I. Normalized maximum likelihood model of order-1 for the compression of DNA sequences. In Proceedings of the Data Compression Conference (DCC-2007), Snowbird, UT, USA, 27–29 March 2007; pp. 33–42.
118. Cao, M.D.; Dix, T.I.; Allison, L.; Mears, C. A simple statistical algorithm for biological sequence compression. In Proceedings of the Data Compression Conference (DCC-2007), Snowbird, UT, USA, 27–29 March 2007; pp. 43–52.
119. Kaipa, K.K.; Bopardikar, A.S.; Abhilash, S.; Venkataraman, P.; Lee, K.; Ahn, T.; Narayanan, R. Algorithm for dna sequence compression based on prediction of mismatch bases and repeat location. In Proceedings of 2010 IEEE International Conference on the Bioinformatics and Biomedicine Workshops (BIBMW), Hong Kong, China, 18 December 2010; pp. 851–852.
120. Gupta, A.; Agarwal, S. A novel approach for compressing DNA sequences using semi-statistical compressor. *Int. J. Comput. Appl.* **2011**, *33*, 245–251. [CrossRef]
121. Pinho, A.J.; Ferreira, P.J.S.G.; Neves, A.J.R.; Bastos, C.A.C. On the representability of complete genomes by multiple competing finite-context (Markov) models. *PLoS ONE* **2011**, *6*, e21588. [CrossRef] [PubMed]
122. Zhu, Z.; Zhou, J.; Ji, Z.; Shi, Y. DNA sequence compression using adaptive particle swarm optimization-based memetic algorithm. *IEEE Trans. Evol. Comput.* **2011**, *15*, 643–658. [CrossRef]
123. Mohammed, M.H.; Dutta, A.; Bose, T.; Chadaram, S.; Mande, S.S. DELIMINATE–A fast and efficient method for loss-less compression of genomic sequences. *Bioinformatics* **2012**, *28*, 2527–2529. [CrossRef] [PubMed]
124. Pinho, A.J.; Pratas, D. MFCompress: A compression tool for FASTA and multi-FASTA data. *Bioinformatics* **2014**, *30*, 117–118. [CrossRef] [PubMed]
125. Li, P.; Wang, S.; Kim, J.; Xiong, H.; Ohno-Machado, L.; Jiang, X. DNA-COMPACT: DNA Compression Based on a Pattern-Aware Contextual Modeling Technique. *PLoS ONE* **2013**, *8*, e80377. [CrossRef] [PubMed]
126. Dai, W.; Xiong, H.; Jiang, X.; Ohno-Machado, L. An Adaptive Difference Distribution-Based Coding with Hierarchical Tree Structure for DNA Sequence Compression. In Proceedings of the Data Compression Conference (DCC-2013), Snowbird, UT, USA, 20–22 March 2013; pp. 371–380.

127. Guo, H.; Chen, M.; Liu, X.; Xie, M. Genome compression based on Hilbert space filling curve. In Proceedings of the 3rd International Conference on Management, Education, Information and Control (MEICI 2015), Shenyang, China, 29–31 May 2015; pp. 29–31.
128. Xie, X.; Zhou, S.; Guan, J. CoGI: Towards compressing genomes as an image. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12*, 1275–1285. [CrossRef] [PubMed]
129. Benoit, G.; Lemaitre, C.; Lavenier, D.; Drezen, E.; Dayris, T.; Uricaru, R.; Rizk, G. Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph. *BMC Bioinform.* **2015**, *16*, 288. [CrossRef] [PubMed]
130. Fritz, M.H.Y.; Leinonen, R.; Cochrane, G.; Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* **2011**, *21*, 734–740. [CrossRef] [PubMed]
131. Kozanitis, C.; Saunders, C.; Kruglyak, S.; Bafna, V.; Varghese, G. Compressing genomic sequence fragments using SlimGene. *J. Comput. Biol.* **2011**, *18*, 401–413. [CrossRef] [PubMed]
132. Deorowicz, S.; Grabowski, S. Compression of DNA sequence reads in FASTQ format. *Bioinformatics* **2011**, *27*, 860–862. [CrossRef] [PubMed]
133. Wandelt, S.; Leser, U. Adaptive efficient compression of genomes. *Algorithms Mol. Biol.* **2012**, *7*, 30. [CrossRef] [PubMed]
134. Qiao, D.; Yip, W.K.; Lange, C. Handling the data management needs of high-throughput sequencing data: SpeedGene, a compression algorithm for the efficient storage of genetic data. *BMC Bioinform.* **2012**, *13*, 100–107. [CrossRef] [PubMed]
135. Ochoa, I.; Hernaez, M.; Weissman, T. iDoComp: A compression scheme for assembled genomes. *Bioinformatics* **2014**, *31*, 626–633. [CrossRef] [PubMed]
136. Deorowicz, S.; Danek, A.; Niemiec, M. GDC 2: Compression of large collections of genomes. *Sci. Rep.* **2015**, *5*, 1–12. [CrossRef] [PubMed]
137. Saha, S.; Rajasekaran, S. NRGC: A novel referential genome compression algorithm. *Bioinformatics* **2016**, *32*, 3405–3412. [CrossRef] [PubMed]
138. Stephens, Z.D.; Lee, S.Y.; Faghri, F.; Campbell, R.H.; Zhai, C.; Efron, M.J.; Iyer, R.; Schatz, M.C.; Sinha, S.; Robinson, G.E. Big data: Astronomical or genomical? *PLoS Biol.* **2015**, *13*, e1002195. [CrossRef] [PubMed]
139. Hanus, P.; Dingel, J.; Chalkidis, G.; Hagenauer, J. Compression of whole genome alignments. *IEEE Trans. Inf. Theory* **2010**, *56*, 696–705. [CrossRef]
140. Jones, D.C.; Ruzzo, W.L.; Peng, X.; Katze, M.G. Compression of next-generation sequencing reads aided by highly efficient de novo assembly. *Nucleic Acids Res.* **2012**, *40*, e171. [CrossRef] [PubMed]
141. Hach, F.; Numanagić, I.; Alkan, C.; Sahinalp, S.C. SCALCE: Boosting sequence compression algorithms using locally consistent encoding. *Bioinformatics* **2012**, *28*, 3051–3057. [CrossRef] [PubMed]
142. Matos, L.M.O.; Pratas, D.; Pinho, A.J. A compression model for DNA multiple sequence alignment blocks. *IEEE Trans. Inf. Theory* **2013**, *59*, 3189–3198. [CrossRef]
143. Bonfield, J.K.; Mahoney, M.V. Compression of FASTQ and SAM format sequencing data. *PLoS ONE* **2013**, *8*, e59190. [CrossRef] [PubMed]
144. Holley, G.; Wittler, R.; Stoye, J.; Hach, F. Dynamic alignment-free and reference-free read compression. In Proceedings of the International Conference on Research in Computational Molecular Biology, Hong Kong, China, 3–7 May 2017; Springer: Berlin, Germany, 2017; pp. 50–65.
145. Cox, A.J.; Bauer, M.J.; Jakobi, T.; Rosone, G. Large-scale compression of genomic sequence databases with the Burrows-Wheeler transform. *Bioinformatics* **2012**, *28*, 1415–1419. [CrossRef] [PubMed]
146. Popitsch, N.; Haeseler, A. NGC: Lossless and lossy compression of aligned high-throughput sequencing data. *Nucleic Acids Res.* **2013**, *41*, e27. [CrossRef] [PubMed]
147. Wan, R.; Anh, V.N.; Asai, K. Transformations for the compression of FASTQ quality scores of next-generation sequencing data. *Bioinformatics* **2012**, *28*, 628–635. [CrossRef] [PubMed]
148. Huang, Z.A.; Wen, Z.; Deng, Q.; Chu, Y.; Sun, Y.; Zhu, Z. LW-FQZip 2: A parallelized reference-based compression of FASTQ files. *BMC Bioinform.* **2017**, *18*, 179. [CrossRef] [PubMed]
149. Hosseini, M.; Pratas, D.; Pinho, A.J. A survey on data compression methods for biological sequences. *Information* **2016**, *7*, 56. [CrossRef]
150. Prado-Martinez, J.; Sudmant, P.H.; Kidd, J.M.; Li, H.; Kelley, J.L.; Lorente-Galdos, B.; Veeramah, K.R.; Woerner, A.E.; O'Connor, T.D.; Santpere, G.; et al. Great ape genetic diversity and population history. *Nature* **2013**, *499*, 471–475. [CrossRef] [PubMed]

151. Zhang, Q.; Jun, S.R.; Leuze, M.; Ussery, D.; Nookaew, I. Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. *Sci. Rep.* **2017**, *7*, 40712. [CrossRef] [PubMed]
152. Locke, D.; Segraves, R.; Carbone, L.; Archidiacono, N.; Albertson, D.; Pinkel, D.; Eichler, E. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.* **2003**, *13*, 347–357. [CrossRef] [PubMed]
153. Ventura, M.; Catacchio, C.R.; Alkan, C.; Marques-Bonet, T.; Sajjadian, S.; Graves, T.A.; Hormozdiari, F.; Navarro, A.; Malig, M.; Baker, C.; et al. Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res.* **2011**, *21*, 1640–1649. [CrossRef] [PubMed]
154. Roos, C.; Zinner, D.; Kubatko, L.S.; Schwarz, C.; Yang, M.; Meyer, D.; Nash, S.D.; Xing, J.; Batzer, M.A.; Brameier, M.; et al. Nuclear versus mitochondrial DNA: Evidence for hybridization in colobine monkeys. *BMC Evol. Biol.* **2011**, *11*, 77. [CrossRef] [PubMed]
155. Alkan, C.; Kidd, J.M.; Marques-Bonet, T.; Aksay, G.; Antonacci, F.; Hormozdiari, F.; Kitzman, J.O.; Baker, C.; Malig, M.; Mutlu, O.; et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **2009**, *41*, 1061. [CrossRef] [PubMed]
156. Zhang, J. Evolution by gene duplication: An update. *Trends Ecol. Evol.* **2003**, *18*, 292–298. [CrossRef]
157. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [CrossRef] [PubMed]
158. Chevreux, B.; Pfisterer, T.; Drescher, B.; Driesel, A.J.; Müller, W.E.; Wetter, T.; Suhai, S. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **2004**, *14*, 1147–1159. [CrossRef] [PubMed]
159. Wolfe, K.H.; Li, W.H.; Sharp, P.M. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 9054–9058. [CrossRef] [PubMed]
160. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **2010**, *26*, 345–352. [CrossRef] [PubMed]
161. Farré, M.; Ruiz-Herrera, A. Role of chromosomal reorganisations in the human-chimpanzee speciation. In *Encyclopedia of Life Sciences (eLS)*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
162. Farré, M.; Micheletti, D.; Ruiz-Herrera, A. Recombination rates and genomic shuffling in human and chimpanzee—A new twist in the chromosomal speciation theory. *Mol. Biol. Evol.* **2013**, *30*, 853–864. [CrossRef] [PubMed]
163. Hosseini, M.; Pratas, D.; Pinho, A.J. On the role of inverted repeats in DNA sequence similarity. In Proceedings of the International Conference on Practical Applications of Computational Biology & Bioinformatics, Porto, Portugal, 21–23 June 2017; Springer: Berlin, Germany, 2017; pp. 228–236.
164. Fleagle, J.G. *Primate Adaptation and Evolution*; Academic Press: Cambridge, MA, USA, 2013.
165. Richly, E.; Leister, D. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* **2004**, *21*, 1081–1084. [CrossRef] [PubMed]
166. Calabrese, F.; Balacco, D.; Preste, R.; Diroma, M.; Forino, R.; Ventura, M.; Attimonelli, M. NumtS colonization in mammalian genomes. *Sci. Rep.* **2017**, *7*, 16357. [CrossRef] [PubMed]
167. Damas, J.; Samuels, D.C.; Carneiro, J.; Amorim, A.; Pereira, F. Mitochondrial DNA rearrangements in health and disease—A comprehensive study. *Hum. Mutat.* **2014**, *35*, 1–14. [CrossRef] [PubMed]