Evaluation of Statistical Complexity in Viral Genome Sequences

Jorge Miguel Silva
jorge.miguel.ferreira.silva@ua.pt

Diogo Pratas
pratas@ua.pt

Sérgio Matos
aleixomatos@ua.pt

Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Portugal.

Department of Virology, University of Helsinki, Finland.

Department of Electronics, Telecommunications and Informatics, University of Aveiro, Portugal.

Abstract

In algorithmic information theory, the Kolmogorov complexity of an object is the length of the shortest computer program that produces the object as output. It is a measure of the computational resources needed to specify the object. However, Kolmogorov complexity is non-computable as such, it can only be approximately attainable. One of the most notable approximations are data compressors, since the bitstream produced by a lossless data compression algorithm allows the reconstruction of the original data with the appropriate decoder, and therefore can be seen as an upper bound of the algorithmic complexity of the sequence. In this paper, we evaluate the usage of the Normalized Compression (NC) as the compression measure for analysing various Virus DNA sequences and evaluate how it changes when substitutions and permutations are performed on the DNA sequences. Finally, we draw conclusions regarding the nature of these sequences.

1 Introduction

Shannon is considered the founder of the information theory field since he was responsible for the creation of a quantitative model of communication as a statistical process underlying information theory. He defined the notion of average information (also called Shannon entropy) as the summation of the product between the probability of each character and the logarithm of this probability [9].

The notion of algorithmic information was produced by Solomonoff, Kolmogorov, and Chaitin, who introduced the concept of Kolmogorov complexity (*algorithmic complexity*). This notion became widely adopted and is currently the standard to perform information quantification. It differs from Shannon's entropy because it considers that the source creates structures which follow algorithmic schemes [3], rather than perceiving the machine as generating symbols from a probabilistic function. In essence, Solomonoff, Kolmogorov, and Chaitin, showed that among all the algorithms that decode strings from their codes, there is an optimal one. This algorithm, for all strings, allows codes as short as allowed by any other up to an additive constant that depends on the algorithms, but not on the strings themselves. Concretely, the algorithmic information quantifies the amount of information of a string s by determining its complexity $K(s)$ (Equation 1), which is defined by the shortest length l of the binary program p that computes s on a universal Turing machine U and halts [4].

$$K(s) := \min_{p}\{l(p) : U(p) = s\}. \tag{1}$$

Latter Bennett introduced the notion of Logical Depth, which adds to Kolmogorov complexity the notion of time, and therefore it quantifies information as the time required by a standard universal TM U to generate a given string from an input that is algorithmically random [1].

Despite the progress in the field of information theory, quantifying information is still an open questions in computer science, since there is no computable measure that encapsulates all concepts surrounding information. Thus, one usually chooses between two options in order to quantify information: *Shannon entropy* or an approximation of *algorithmic complexity*.

Shannon entropy poses some problems since it is not invariant to the description of the object and its probability distribution. Furthermore, it lacks an invariance theorem, forcing us to decide on a characteristic shared by the objects of interest [10]. On the other hand, *algorithmic complexity* is only approximately attainable, since the Kolmogorov complexity is non-computable [5]. These approximations are computable variants of the Kolmogorov complexity and are bounded by time and resources.

The most notable approximations were made by data compressors, as the bit stream produced by a lossless data compression algorithm allows the reconstruction of the original data with the appropriate decoder, and therefore can be seen as an upper bound of the algorithmic complexity of the sequence [2].

In this paper we make use of Compression based approach to assess statistical complexity (Martin-Löf randomness) of Virus DNA sequences. Our methodology consists in applying the Normalized Compression (NC) computing the selection of the best Markov model that minimizes the complexity quantity of the DNA Virus sequence through the variation of the order-depth. We also evaluate how the NC behaves when substitutions and permutations are performed on the DNA sequences.

2 Methods

2.1 Normalized Compression and Markov Models

To assess Martin-Löf randomness we evaluated the data compression of the Virus DNA sequences. The Normalized Compression (NC) was used as the compression measure (Equation 2) where x is the string, $C(x)$ represents the number of bits needed by the lossless compression program to represent the string x, $|A|$ is the size of the alphabet (an approximation of Σ) and $|x|$ is the length of the string x [7]. Furthermore, $C(x)$ is computed with the probability P of each string symbol x_i occurring.

$$NC(x) = \frac{C(x)}{|x| \times \log_2 |A|}, \text{ where } C(x) = \sum_{i=1}^{|x|} -\log_2 P(x_i). \tag{2}$$

The NC is computed using a Markov model, a finite-context model that predicts the next outcome of a sequence given a past context k [6]. Specifically, a Markov model reads the input using a given context k and updates its internal model. This internal model is used to compute the probability of any character being read at a given point. In this case, the DNA sequence is provided to the Markov Model with context k. This model is used to determine the NC by computing the normalized summation of the probability of each character occurring on the tape. For each DNA sequence, the Markov model with context $k = \{2, ..., 10\}$ that minimizes the NC is selected.

2.2 Substitutions and Permutations of the DNA Sequences

For each DNA sequence, the substitution rate probability of the string was increased from 0% to 100% and simultaneously, randomly permuted in blocks of increasingly smaller sizes. The substitution was fixed to be a specific nucleotide in order to continually decrease the complexity of the sequence. At each point the NC was computed using the best Markov model (within defined the context range) that minimizes the complexity quantity of that given string and the obtained results where plotted as a heat map.

3 Results

For this case study we analised the DNA sequences of the *Ebola* Virus; *Microplitis* Virus; *Human parvovirus B19 isolate BX1* and the *Torque teno indri* Virus.

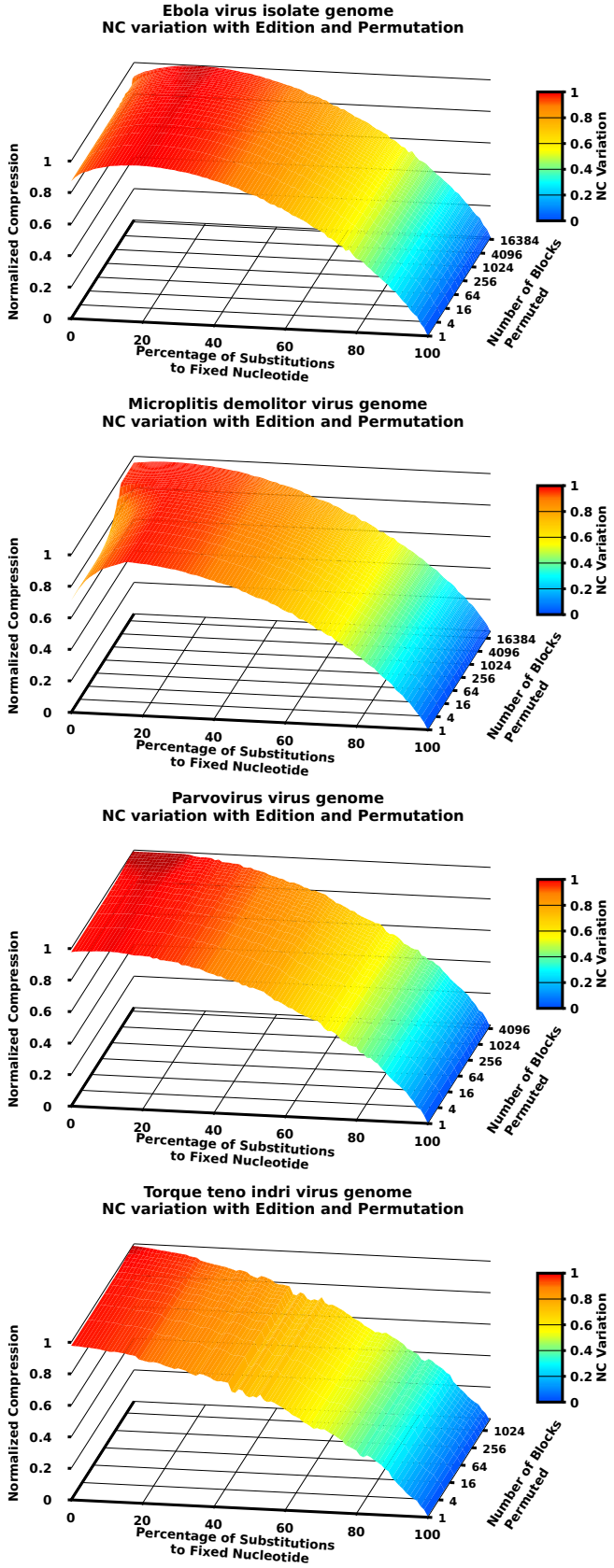

Figure 1: Heat Maps of Normalized Compression with an increase in permutation and edition rate for the Ebola, *Microplitis*, *Human parvovirus B19 isolate BX1* and *Torque teno indri* Virus respectively.

As shown in Figure 1, the NC behaves as expected since the NC decreases logarithmically as the substitution rate of the strings increases in genomic data for all cases.

The same occurs regarding random permutation of the Viral genomic sequences. Although less noticeable for the *Human parvovirus B19 isolate BX1* and *Torque teno indri* Virus due to being more statistically complex in nature than the *Microplitis* and *Ebola* Virus, in all Virus genomic sequences it can be clearly seen an increase in the NC with an increase inf the number of blocks permuted.

4 Conclusion

In this paper we evaluate the usage of the Normalized Compression (NC) as the compression measure for analysing various Virus DNA sequences and valuate how it changes when substitutions and permutations are performed on the DNA sequences. The results suggest that computation of the NC in this manner is both a sensitive and tolerant way of dealing with substitutions and permutations. This methodology is interesting for evaluating the Martin-Löf randomness of Virus DNA sequences since besides being an ultra-fast method for obtaining information regarding the DNA, it can cope with the presence of substitutions and permutations in a string. It is important to note that viral genome sequences studied seem to be highly statistically complex, however, other DNA sequences have been previously compressed to a very efficient degree with the usage of lossless compression algorithms which are hybrids between statistical and algorithmic schemes (GeCo [8]). Virus could potentially follow algorithmic schemes which make simple statistical compressors unable to encompass these patterns. As such, compressors which follow algorithmic modeling could potentially provide a better compression of data and consequentially provide a better data representation.

5 Acknowledgments

Funded by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UID/CEC/00127/2019 and FCT research grant SFRH/BD/141851/2018.

References

[1] Charles H. Bennett. Logical depth and physical complexity. *The Universal Turing Machine A Half-Century Survey*, pages 207–235, 1995.

[2] Peter Bloem, Francisco Mota, Steven de Rooij, Luís Antunes, and Pieter Adriaans. A safe approximation for kolmogorov complexity. In *International Conference on Algorithmic Learning Theory*, pages 336–350. Springer, Springer International Publishing, 2014.

[3] Daniel Hammer, Andrei Romashchenko, Alexander Shen, and Nikolai Vereshchagin. Inequalities for Shannon Entropy and Kolmogorov Complexity. *Journal of Computer and System Sciences*, 60 (2):442–464, 2000. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1999.1677. URL `http://www.sciencedirect.com/science/article/pii/S002200009991677X`.

[4] Andrei N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1): 1–7, 1965.

[5] Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008.

[6] Armando J. Pinho, Paulo J. S. G. Ferreira, António J. R. Neves, and Carlos A. C. Bastos. On the representability of complete genomes by multiple competing finite-context (markov) models. *PLOS ONE*, 6(6):1–7, 06 2011. doi: 10.1371/journal.pone.0021588. URL `https://doi.org/10.1371/journal.pone.0021588`.

[7] Diogo Pratas and Armando J. Pinho. On the approximation of the kolmogorov complexity for DNA sequences. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 259–266. Springer, 2017.

[8] Diogo Pratas, Armando J. Pinho, and Paulo J. S. G. Ferreira. Efficient compression of genomic sequences. In *2016 Data Compression Conference (DCC)*, pages 231–240. IEEE, 2016.

[9] Claude Elwood Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x`.

[10] Hector Zenil, Santiago Hernández-Orozco, Narsis A. Kiani, Fernando Soler-Toscano, Antonio Rueda-Toicen, and Jesper Tegnér. A decomposition method for global evaluation of shannon entropy and local estimations of algorithmic complexity. *Entropy*, 20(8), 2018. ISSN 1099-4300. doi: 10.3390/e20080605. URL `http://www.mdpi.com/1099-4300/20/8/605`.