Metagenomic composition analysis of sedimentary ancient DNA from the Isle of Wight

Diogo Pratas
IEETA
University of Aveiro
Aveiro, Portugal
pratas@ua.pt

Armando J. Pinho
IEETA/DETI
University of Aveiro
Aveiro, Portugal
ap@ua.pt

***Abstract*—The DNA from several organisms is sequenced conjointly in metagenomics. This allows searching for exogenous microorganisms contained in the samples, with the goal of studying the evolution and co-evolution of host-pathogen, namely for building better diagnostics and therapeutics. However, the quantity and quality of the DNA present in the samples is very poor, pushing the responsibility of analysis improvements into the development of better computational methods. Here, we develop a new processing paradigm to infer the metagenomic composition analysis based on the relative compression of whole genome sequences. Using this method, we present the metagenomic composition analysis of a sedimentary ancient DNA sample, dated to 8,000 years before the present, from the Isle of Wight, United Kingdom. The results show several viruses and bacteria expressing high levels of similarity relative to the samples, namely a circular virus similar to the Avon-Heathcote estuary virus 14 sequenced in New Zealand.**

***Index Terms*—metagenomics, ancient DNA, data compression**

I. Introduction

The DNA from several organisms is sequenced conjointly in metagenomics. Metagenomics allows looking into ancestral samples for determining composition, such as contamination and, more important, ancient pathogens [1]. The latter may help in inferring ancestral death causes, investigate the changing landscape of infectious disease through time, predicting microbial communities or developing novel diagnostics and therapeutics [2].

However, sequencing ancient remains from the Neolithic period is a very difficult process, given the low quantity and quality of DNA present in the samples. These are, usually, correlated with post-mortem degradation. [3].

Additionally, there are strong computational challenges, such as: dealing with a large volume of raw data (*random* shuffled reads), with variable and very short read sizes, high degree of contamination [4], very unbalanced composition in terms of sample sizes and lack of a complete catalog of extant and extinct species.

The identification of metagenomic composition, in ancestral samples, is being addressed using large-scale approaches. Namely, because the use of 16S rRNA gene sequences (small-scale) has, in aDNA, distinguishability problems and high ambiguity on related organisms [5].

The are many approaches to detect and quantify metagenomic sample composition (see, for example, [6]–[9]), namely through the measurement of similarity of the reads according to a database with reference genomes. Recently, we have proposed a compression-based approach [10] that is fast, ultra-sensitive and without overestimating similarity. The later is fundamental to provide quality, rigor, and consistency.

In this paper, we follow the line of the compression approach [10], improving and adding better protocols to new case analysis, ancient DNA from submerged marine sediments, dated to 8,000 years before the present, from the Isle of Wight, United Kingdom.

These sediments include 8,000-year-old wheat [11] that was challenged on the base of a lack of signal of cytosine deamination relative to other datasets. Newer approaches, with more rigorous controls, shown that these data meet the criteria of authentic ancient DNA [12].

Here, rather than looking into eukaryotic genomes, we infer the metagenomic composition of microorganisms, specifically viruses, bacteria, archaea, and fungi. From the sediments, we use two datasets, ERR567364 and ERR567365, with sizes 5.8 and 4.6 GB, respectively.

Further, we present our method, providing background and introducing to the theory behind the Normalized Relative Similarity (NRS). We provide filtering protocols to ensure data quality. Then, we reveal the metagenomic composition of the sedimentary ancient DNA, with an extension to localize the regions with similarity relatively to the samples.

II. Method

The Kolmogorov complexity, also known as the algorithmic information, enables to measure and compare the information contained in different natural processes, namely DNA sequences, that can be expressed using sequences of symbols (strings) from a finite alphabet [13]–[15].

The Kolmogorov complexity differs from the Shannon entropy [16] because it considers that the source, rather than generating symbols from a probabilistic function, it creates structures that follow algorithmic schemes. In order to reverse

This work was partially funded by FEDER (Programa Operacional Factores de Competitividade - COMPETE) and by National Funds through the FCT, in the context of the projects UID/CEC/00127/2013 & PTCD/EEI-SII/6608/2014

ISBN 978-90-827970-1-5 © EURASIP 2018

the problem, there is the need to identify the program and parameters that generate the outcome [13].

The usage of small Turing machines [17] and lossless data compressors [18] are two of the most successful implementations to approximate the Kolmogorov complexity.

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) [14], [18], [19], regardeless if it is computed through the conditional compression [20] or the conjoint compression [18].

The need for a computational measure able to measure the complexity of a string given exclusively other has led to a new concept, that of relative algorithmic information [21]–[25].

Several approaches, to quantify the relative information, have been proposed (e.g., [21], [24]–[28]) such as for handling images [27], texts [24], [25], ECG (electrocardiographic) data [29] and genomic sequences [30].

The relative information can be approached using relative compressors regardless if they are based on dictionaries [21], [24] or context models [25], [31]. These string compressors aim to model and organize the data of a string, y, without knowing the other string, x. Then, freeze the model of y and, finally, measure the number of bits needed to describe x. We call this operation, the compression of x relative to y, as $C(x\|y)$.

Using the $C(x\|y)$ we are able to define the Relative Similarity as

$$\mathrm{RS}(x\|y) = |x| \log_2 |\Theta| - C(x\|y) \tag{1}$$

where $|x|$ is the number of symbols in x and Θ the cardinality of the alphabet. Finally, the Normalized Relative Similarity is defined as

$$\mathrm{NRS}(x\|y) = \frac{|x| \log_2 |\Theta| - C(x\|y)}{|x| \log_2 |\Theta|} = 1 - \frac{C(x\|y)}{|x| \log_2 |\Theta|}. \tag{2}$$

Note that, when x is equal to y, the NRS is approximately one and, when x has completely different nature from y, the NRS is approximately zero.

According to Fig. 1, to compute the NRS (Eq. 2), we use as a y all the FASTQ reads and as a x, individually, each microorganism genome extracted from multiple databases. Since we only have a y, we freeze the model of y and compress each x, without loading again the models of y. This approach allows saving substantially in computational time.

We use a relative compressor, to compute $C(x\|y)$, based on [31], that applies soft-blending, with a decaying forgetting factor, between three context models (CM) and two tolerant CMs. The decaying factor used is 0.95 and a cache-hash of 200. The models have the following parameters:

- **1, tolerant CM**: depth: 20, alpha: 0.1, tolerance: 5;
- **2, CM**: depth: 20, alpha: 0.01, inverted repeats: yes;
- **3, tolerant CM**: depth: 14, alpha: 1, tolerance: 3;
- **4, CM**: depth: 14, alpha: 0.02, inverted repeats: no;
- **5, CM**: depth: 13, alpha: 0.1, inverted repeats: no.

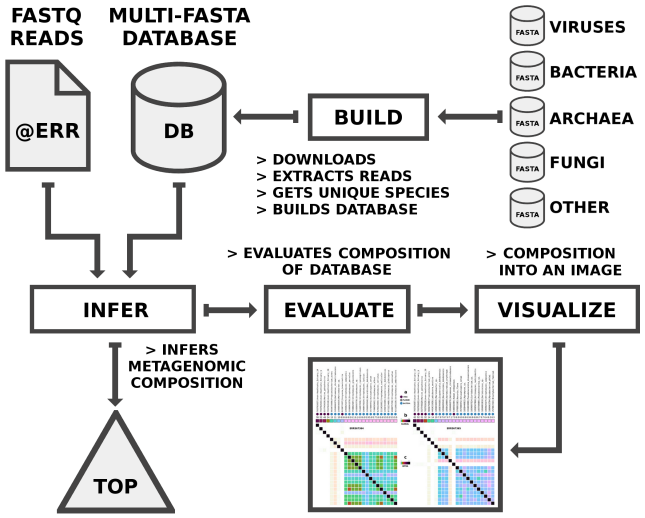

Fig. 1. Pipeline for the analysis of the metagenomic composition using as input the ancient samples (ERR567364 and ERR567365) and a database containing several reference organisms. The "BUILD" phase is according to the subsection "Building the database (DB)". The "INFER" phase is given by the computation of Eq. 2. The "EVALUATE" phase is a detecting control of intra-database similarities.

For more information on the parameters, meaning and the genomic compression field, see [31], [32].

With the mentioned relative compressor each value of the NRS is computed and sorted in a ranking, where higher values stand for higher relative similarity. After, each ranked sequence above a certain threshold is evaluated according to its *self*-redundancy and preserved for further visualization. This approach works as a control, enabling to detect low complexity regions in x, for example, repetitive patterns that may be related to errors in the sequencing or assembling process, which may have a contradictory meaning through the NRS calculation.

Then, we evaluate the reference genomes that had high relative similarity during the metagenomic inference, calculating the NRS between each combination. This process enables to detect patterns of similarity at an intra-database level. These may identify, for example, that some of the high NRS values are related to homologous parts between species. Notice that here the time complexity is quadratic, however, the number of elements is, usually, very low. Moreover, the size of the genomes from the database is much smaller, frequently an order of magnitude, than the input FASTQ samples.

However, before the inference and evaluation, we need to build the database (DB) and trim/clean the reads to ensure high quality and rigorous analysis. In the next subsections, we show the procedure.

A. Building the database (DB)

For building the database (DB), we have downloaded the entire NCBI database for viruses, bacteria, archaea, and fungi, resulting in four datasets of several gigabytes (GB). For each dataset, we have extracted only the sequences labeled as

ISBN 978-90-827970-1-5 © EURASIP 2018

"complete genomes". For downloading, extracting reads by patterns, getting unique species sequences and building the database, we have used our home-made software GOOSE (https://github.com/pratas/goose).

B. Trimming and cleaning the FASTQ reads

Trimming and cleaning the reads is essential to prevent reads with a short size, sequencing errors or with a low quality given the process of DNA sequencing, that may create inconsistent statistics. For cleaning the reads of ERR567364 and ERR567365 we have used GOOSE (G). We have trimmed reads with a local sequencing score below the quality of 15 in a window of 5. Then, we filtered the reads with size below 35 bases, with more than 5 unknown bases in each read, and with an average (global) in the quality of the scores below 15. The commands used were the following:

```
zcat <file>.fastq.gz \
| ./G-FastqMinimumLocalQualityScoreForward \
-k 5 -w 15 -m 33 \
| ./G-FastqMinimumReadSize 35 \
| ./G-FastqExcludeN 5 \
| ./G-FastqMinimumQualityScore 15 \
> reads.fq
```

where <file> represents ERR567364 and ERR567365. With this approach, in ERR567364, from 16,501,223 we have trimmed 840,341 and cleanned 24,196 reads ($\approx 5\%$ reads filtered). While on ERR567365, from 12,634,237 we have trimmed 797,554 and cleanned 26,868 reads ($\approx 7\%$ reads filtered).

III. Results

The sedimentary DNA datasets used, in this study, were accessed through the EBI (https://www.ebi.ac.uk/ena/data/view/PRJEB6766), using the run accession identifiers ERR567364 and ERR567365. The sequencing machine used by [11] was the Illumina MiSeq (Library layout: SINGLE). The database (DB) of the reference microbial genomes was downloaded from the NCBI (ftp://ftp.ncbi.nih.gov/refseq/).

All the results presented in this paper can be fully replicated, under a Linux machine, using the two scripts provided at the repository: https://github.com/pratas/shikra. All the computations ran in a Linux desktop computer with an Intel®Core™i7-6700 CPU @3.40 GHz, with 32 GB of RAM (without an SSD). Using this machine, the computation of the metagenomic composition analysis of the datasets ERR567364 and ERR567365 cost near 98 and 67 minutes of real time, respectively.

Fig. 2a and b depict the results of the metagenomic composition analysis with the highest ranked NRS genomes, with the exception to the *Enterobacteria phage* phiX174 sensu lato, a virus with 5,386 bases and without relevant similarities relative to the potential genomes contained in the samples, that was, clearly, identified as a contamination (NRS of 97.691). We did not include this phage in the image given space constraints.

Besides this phage, we also found high similarities relative to several viral and bacterial species. In the bacteria domain, we have found similarity in several *Streptomyces*. These express similar correlations between multiple species in this genus as the Fig. 2-c depicts. There are other bacteria, namely *Marinobacter sp.*, *Halomonas chromatiredunces*, *Propionibacterium acnes*. The large majority bacteria are from water and soil and, hence, they are according to the source of the samples (underwater soil).

Regarding the viruses domain, we notice similarity relative to *Labidocera aestiva* and *Sicyonia brevirostris* viruses, which usually infects marine copepods and brown rock shrimps, respectively. Another virus is the *Marine gokushovirus*, a single strand (ssDNA) virus with near 4,080 bases and very hard to isolate.

Curiously, we have detected 7% of similarity relative to the human endogenous virus, that is absent of relevant intra-database similarity. Contamination is the probable cause.

The highest NRS values from the figure are representative of a class of ssDNA circular virus, namely the *Avon-Heathcote estuary* virus [33]. The majority of the elements do not share relevant similarity among them (Fig. 2-c). These genomes were collected in Christchurch, New Zealand. When searching for similar viruses at the NCBI, the top five, with the respective identity, are:

- 81% - Beak and feather disease virus;
- 78% - Bat circovirus;
- 78% - Barbel circovirus;
- 77% - Lake Sarah-associated circular virus-27;
- 76% - Canine circovirus strain AZ4438-13;

Usually, these viruses work as the following procedure. They penetrate into the host cell where, uncoating, the viral ssDNA genome penetrates into the nucleus. The viral ssDNA is converted into dsDNA. The viral mRNAs are created through the dsDNA transcription. Then, the viral mRNAs are translated to produce the viral proteins. Notice that the replication may be oriented by the replication associated protein (Rep). This may occur by rolling circle producing ssDNA genomes. The newly synthesized ssDNA can be converted to dsDNA, to serve as a template in the transcription-replication processes, or to be encapsidated by capsid protein and, finally, form complete viruses that are released by cell lysis.

Specifically, we now center on the high NRS associated to the Avon-Heathcote estuary virus 14 [33]. Fig. 3 depicts the profiles of the relative information content, computed with the compression of this virus relative to each of the samples. The lower information content in the plot represents the regions with higher similarity. When we associate these results with the top map, containing the regions of the virus genes in scale, we notice that the gene VM18_gp2 has very high similarity relative to the samples. This is a gene which encodes the replication associated protein (Rep). We also dismiss the possibility of higher redundancy contained in the virus, namely through high copy number, with the *self*-compression of the virus, $C(virus)$.

The other gene (VM18_gp1), associated with the putative capsid proteins (Cap), has alternated regions with very high and very low similarity. This gene, although with several

ISBN 978-90-827970-1-5

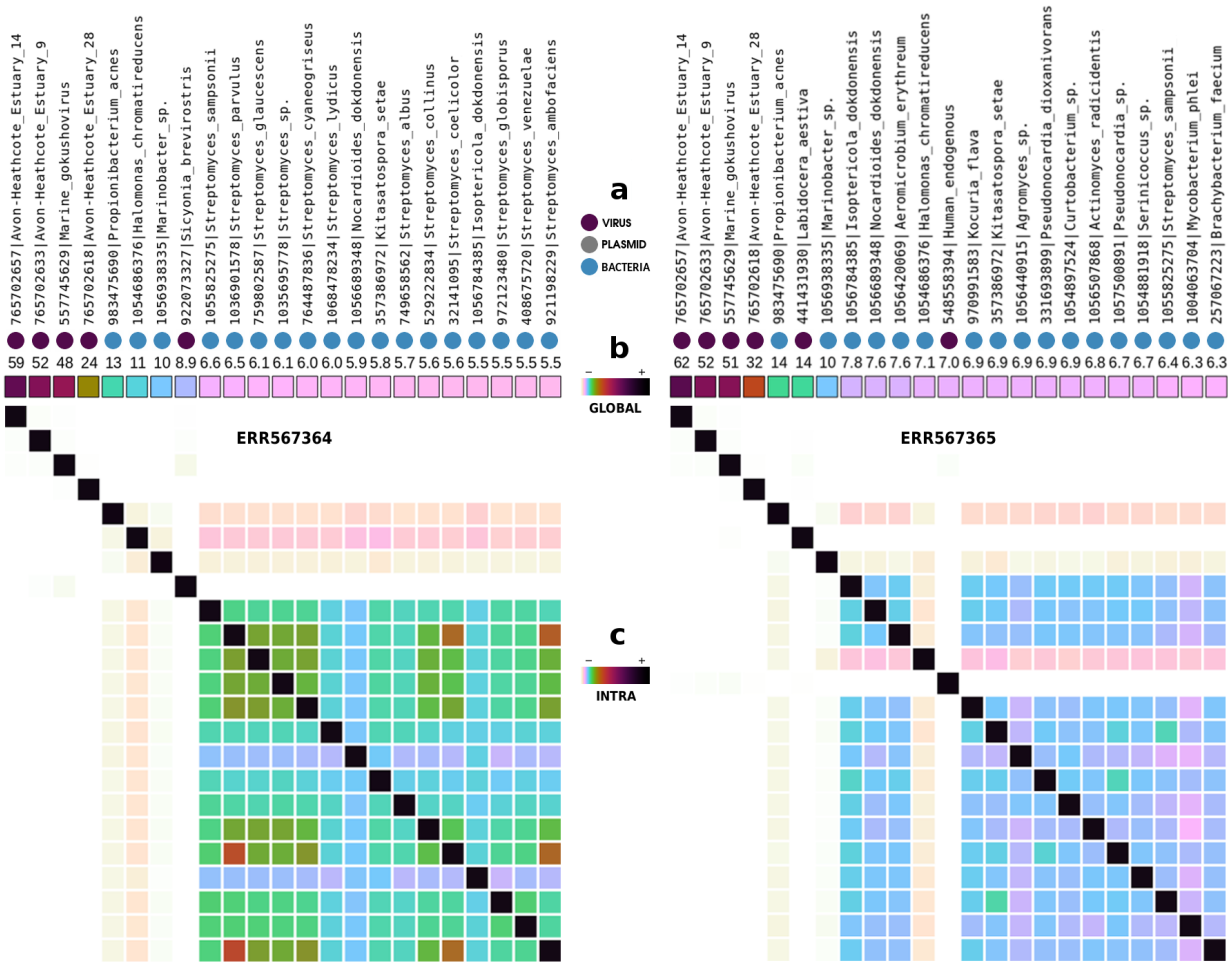

Fig. 2. Metagenomic composition analysis of two datasets corresponding to samples of sedimentary ancient DNA (ERR567364 at left and ERR567365 at right). **a)** identification of the name, GI and domain of the highest ranked NRS genomes; **b)** NRS values in percentage according to a); **c)** NRS between the reference genomes from the database (database intra-similarity).

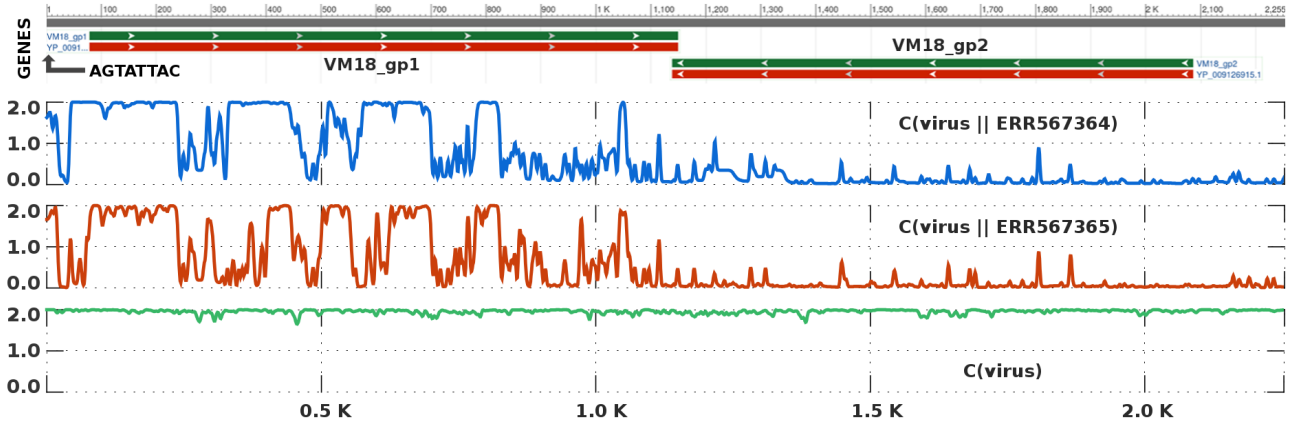

Fig. 3. Information and relative information content of the Avon-Heathcote estuary virus 14. The y-axis stand for information, while the x for the length. The top map depicts the position and genes of the virus (VM18_gp1 and VM18_gp2); the top profile (blue) shows the information content of the virus relative to the ERR567364 sample; the middle profile (brown) shows the information content of the virus relative to the ERR567365 sample; the bottom profile (green) depicts the (*self*) information content of the virus. All the profiles have been low-pass filtered with a Blackman window length of 5 bases.

similar parts has a very different evolution from the reference. These findings are according with other new circular ssDNA viruses identified in marine invertebrates that revealed high sequence diversity and consistent predicted intrinsic disorder

ISBN 978-90-827970-1-5 © EURASIP 2018

patterns within putative structural proteins [34].

One of the known hosts of these viruses is *Paphies subtriangulata*, a bivalve clam. In future works, we plan to identify the host of this newly detected virus. Given the diverse genomic characteristics of the virus, localization, and water temperature, it may express new functionalities and specific adaptation to the local environment and hosts, namely parallelism in the co-evolution with the host.

IV. Conclusions

We have developed a new processing paradigm to infer the metagenomic composition analysis based on the relative compression of whole genome sequences, adding better quality control protocols and introducing a way to localize where the similarities occur.

Using this method, we have presented the metagenomic composition analysis of two sedimentary ancient DNA samples, dated to 8,000 years before the present, from the Isle of Wight, United Kingdom.

We have found several viruses and bacteria expressing high levels of similarity relative to the samples. From these, the most similar is a circular virus known as Avon-Heathcote estuary virus 14. In this virus, we have localized both genes where particularly the VM18_gp2 gene is very similar relative to the samples. This gene is used in the process of protein replication.

Acknowledgment

We thank João Zilhão, Martin Kircher, and Gabriel Renaud for helpful comments and explanations.

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 and PTCD/EEI-SII/6608/2014.

References

[1] S. Pääbo, H. Poinar, D. Serre, *et al.* "Genetic analyses from ancient DNA," *Annu. Rev. Genet.*, vol. 38, pp. 645–679, 2004.

[2] C. J. Houldcroft, M. A. Beale, and J. Breuer, "Clinical and biological insights from viral genome sequencing," *Nature Reviews Microbiology*, vol. 15, no. 3, pp. 183, 2017.

[3] A. W. Briggs, U. Stenzel, P. L. F. Johnson, *et al.*, "Patterns of damage in genomic DNA sequences from a Neandertal," *Proceedings of the National Academy of Sciences*, vol. 104, no. 37, pp. 14616–14621, 2007.

[4] A. Sajantila, "Editors' pick: Contamination has always been the issue!," *Investigative genetics*, vol. 5, no. 17, pp. 2, 2014.

[5] L. S. Weyrich, S. Duchene, J. Soubrier, *et al.*, "Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus," *Nature*, 2017.

[6] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome biology*, vol. 15, no. 3, pp. 1, 2014.

[7] J. Ren, N. A. Ahlgren, Y. Y. Lu, *et al.*, "VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data," *Microbiome*, vol. 5, no. 1, pp. 69, 2017.

[8] G. Louvel, C. D. Sarkissian, K. Hanghøj, *et al.* "metaBIT, an integrative and automated metagenomic pipeline for analysing microbial profiles from high-throughput sequencing shotgun data," *Molecular ecology resources*, vol. 16, no. 6, pp. 1415–1427, 2016.

[9] A. Herbig, F. Maixner, K. I. Bos, *et al.* "MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman," *bioRxiv preprint*, 2017.

[10] D. Pratas, A. J. Pinho, R. M. Silva, *et al.* "FALCON-meta: a method to infer metagenomic composition of ancient DNA," *bioRxiv*, p. 267179, 2018.

[11] O. Smith, G. Momber, R. Bates, *et al.* "Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago," *Science*, vol. 347, no. 6225, pp. 998–1001, 2015.

[12] L. Kistler, O. Smith, R. Ware, *et al.* "Thermal age, cytosine deamination and the veracity of 8,000 year old wheat DNA from sediments," *bioRxiv*, p. 032060, 2015.

[13] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Problems of Information Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

[14] M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications*, Springer, 3rd edition, 2008.

[15] J. Rissanen, *Optimal estimation of parameters*, Cambridge University Press, 2012.

[16] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423,623–656, 1948.

[17] F. Soler-Toscano, H. Zenil, J.-P. Delahaye, *et al.* "Calculating Kolmogorov complexity from the output frequency distributions of small Turing machines," *PLoS ONE*, vol. 9, no. 5, pp. e96223, 2014.

[18] M. Li, X. Chen, X. Li, *et al.* "The similarity metric," *IEEE Trans. on Information Theory*, vol. 50, no. 12, pp. 3250–3264, Dec. 2004.

[19] C. H. Bennett, P. Gács, M. Li, *et al.*, "Information distance," *IEEE Trans. on Information Theory*, vol. 44, no. 4, pp. 1407–1423, July 1998.

[20] N. Nikvand and Z. Wang, "Generic image similarity based on Kolmogorov complexity," in *Proc. of the IEEE Int. Conf. on Image Processing, ICIP-2010*, Hong Kong, Sept. 2010, pp. 309–312.

[21] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. on Information Theory*, vol. 39, no. 4, pp. 1270–1279, July 1993.

[22] D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Physical Review Letters*, vol. 88, no. 4, pp. 048702–1–048702–4, Jan. 2002.

[23] D. Cerra and M. Datcu, "Algorithmic relative complexity," *Entropy*, vol. 13, pp. 902–914, 2011.

[24] D. P. Coutinho and M. Figueiredo, "Text classification using compression-based dissimilarity measures," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 29, no. 5, 2015.

[25] A. J. Pinho, D. Pratas, and P. J. S. G. Ferreira, "Authorship attribution using relative compression," in *Proc. of the Data Compression Conf., DCC-2016*, Snowbird, Utah, Mar. 2016.

[26] S. Helmer, N. Augsten, and M. Böhlen, "Measuring structural similarity of semistructured data based on information-theoretic approaches," *The VLDB Journal - The International Journal on Very Large Data Bases*, vol. 21, no. 5, pp. 677–702, 2012.

[27] D. Cerra, M. Datcu, and P. Reinartz, "Authorship analysis based on data compression," *Pattern Recognition Letters*, vol. 42, pp. 79–84, 2014.

[28] R. Cilibrasi et al., *Statistical inference through data compression*, Ph.D. thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam, 2007.

[29] J. M. Carvalho, S. Brãs, J. Ferreira, *et al.* "Impact of the Acquisition Time on ECG Compression-Based Biometric Identification Systems," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2017, pp. 169–176.

[30] D. Pratas, R. M. Silva, A. J. Pinho, *et al.* "An alignment-free method to find and visualise rearrangements between pairs of DNA sequences," *Scientific Reports*, vol. 5, pp. 10203, May 2015.

[31] D. Pratas, A. J. Pinho, and P. J. S. G. Ferreira, "Efficient compression of genomic sequences," in *Proc. of the Data Compression Conf., DCC-2016*, Snowbird, Utah, 2016, pp. 231–240.

[32] M. Hosseini, D. Pratas, and A. J Pinho, "A survey on data compression methods for biological sequences," *Information*, vol. 7, no. 4, pp. 56, 2016.

[33] A. Dayaram, S. Goldstien, G. R. Argüello-Astorga, *et al.* "Diverse small circular DNA viruses circulating amongst estuarine molluscs," *Infection, Genetics and Evolution*, vol. 31, pp. 284–295, 2015.

[34] K. Rosario, R. O. Schenck, R. C. Harbeitner, *et al.* "Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins," *Frontiers in microbiology*, vol. 6, pp. 696, 2015.