Uncovering signals from the coronavirus genome

E. Canessa*

Science Dissemination Unit (SDU)

The Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, Italy

Abstract

A signal analysis of the genoma sequenced of coronavirus variants: B.1.1.7, B.1.135, B.1.429-B.1.427, B.1.525 and P1 is presented. We deal with a certain type of finite alternating sum series having independently distributed terms associated with binary $(0, 1)$ indicators for the nucleotide bases A, C, G, T. This method provides additional information to conventional Similarity comparisons and Power Spectrum approaches. It leads to uncover distinctive patterns regarding the intrinsic data organization of genomic sequences according to its progression along the nucleotide bases position. Hence, the method could be useful for surveillance of genoma variants.

Keywords: *SARS-CoV-2, Sequence analysis, Comparative genomic variants, Alternating Series*

*E-mail: canessae@ictp.it

1 Introduction

Chinese scientists were the first to sequence the complete genome of SARS-CoV-2 coronavirus in humans and shared their data with the rest of the world in early 2020 [1, 2, 3, 4]. The virus presented a unique lineage for almost half of its genome, with little genetic relationships to other known viruses specially in the genomic region encoding the spike (S-protein) responsible for the virus entry into the human host cells [5]. The race to find immunity from this SARS-CoV-2 started soon after and since then, genomic sequences from around the Globe have been added into open archives such as the global science GISAID initiative (www.gisaid.org) for further research.

There are thousands primary sources available in GISAID which warrant urgent investigation on biometric analyses, comparison and characterization of these sequences of different coronavirus lineages responsible for the ongoing pandemic. In particular an analytical study of emerging SARS-CoV-2 variants (sharing mutations) is highly needed since some variants appear to be more persistent and contagious. This is a good reason for increasing genomic surveil-

lance on the emerging variants by the development of new tools that can detect and catalog these strains in a timely manner [6].

We propose here a quantitative method for the examination of distinctive patterns of complete coronavirus genome data. We deal with a certain type of alternating finite series having terms converted to binary values $(0, 1)$ for the nucleotide bases (A)denine, (C)ytosine, (G)uanine and (T)hymine according to their progression along the genomic sequences. This mapping into four binary projections of the coronavirus sequence is similarly done as in previous studies on the three-base periodicity characteristic of protein-coding DNA sequences [7, 8].

We apply our finite and alternating sum series method to most variants of SARS-CoV-2 so far sequenced. It is shown that this approach provides additional information to conventional genomic Similarity computations [1] and to the power spectrum in the 'frequency' domain of the associated binary sequences [9, 10]. By this method we uncover distinctive signals of the intrisic gene organization revealed by the genome sequences of the single-stranded RNA coronaviruses.

2 Similarity and Power Spectrum Analysis

The encoded genes in a sequence of four nucleotides, represented by the symbols A, C, G and T, store instructions to assemble and reproduce every living organism. In DNA, nucleotides of one strand are complementary to those of the opposite strand according to the pairing rules $A-T$ and $C-G$. Similarity plots of SARS-like CoVs and bat SARS-like CoVs complete sequences of A, C, G and T revealed apparent recombination events, useful to understand the probable coronavirus pandemic origins [1, 2, 3, 4]. Genome sequences of SARS-CoV-2 from the city of Wuhan in China, exhibit high level of genetic similarity (88%) to bat-derived severe acute respiratory syndrome (SARS)-like coronaviruses: bat-SL-CoVZC45 and bat-SL-CoVZXC21. Similarity plots of this kind based on the nucleotide sequence of only the S-spyke gene of bat SL-CoV WIV16 has been reported in [11].

Similarity plots based on the full-length nucleotide sequence of gene variants of coronavirus are next presented. In Fig.1 we illustrate the genetic Similarity plot between SARS-CoV-2 Wuhan-Hu-1 (MN908947.3) and several representative full-length genome sequences of variants known as B.1.1.7 (UK), B.1.135 (South Africa), B.1.429-B.1.427 (California), B.1.525 (Nigeria) and P1 (Brazil). These complete GISAID samples come from broad geographical regions indicated in brackets, and were collected during different periods of time (from 01 Nov 2020 to 01 March 2021). Only available sequences in FASTA format with high coverage have been considered and grouped.

In these calculations we installed and run under the Linux Ubuntu O.S., the current version 36 of the FASTA sequence comparison software, which includes lalign36. It produces multiple non-overlapping alignments for protein and DNA sequences using the Huang and Miller sim algorithm for the Waterman-Eggert algorithm (github.com/wrpearson/fasta36).

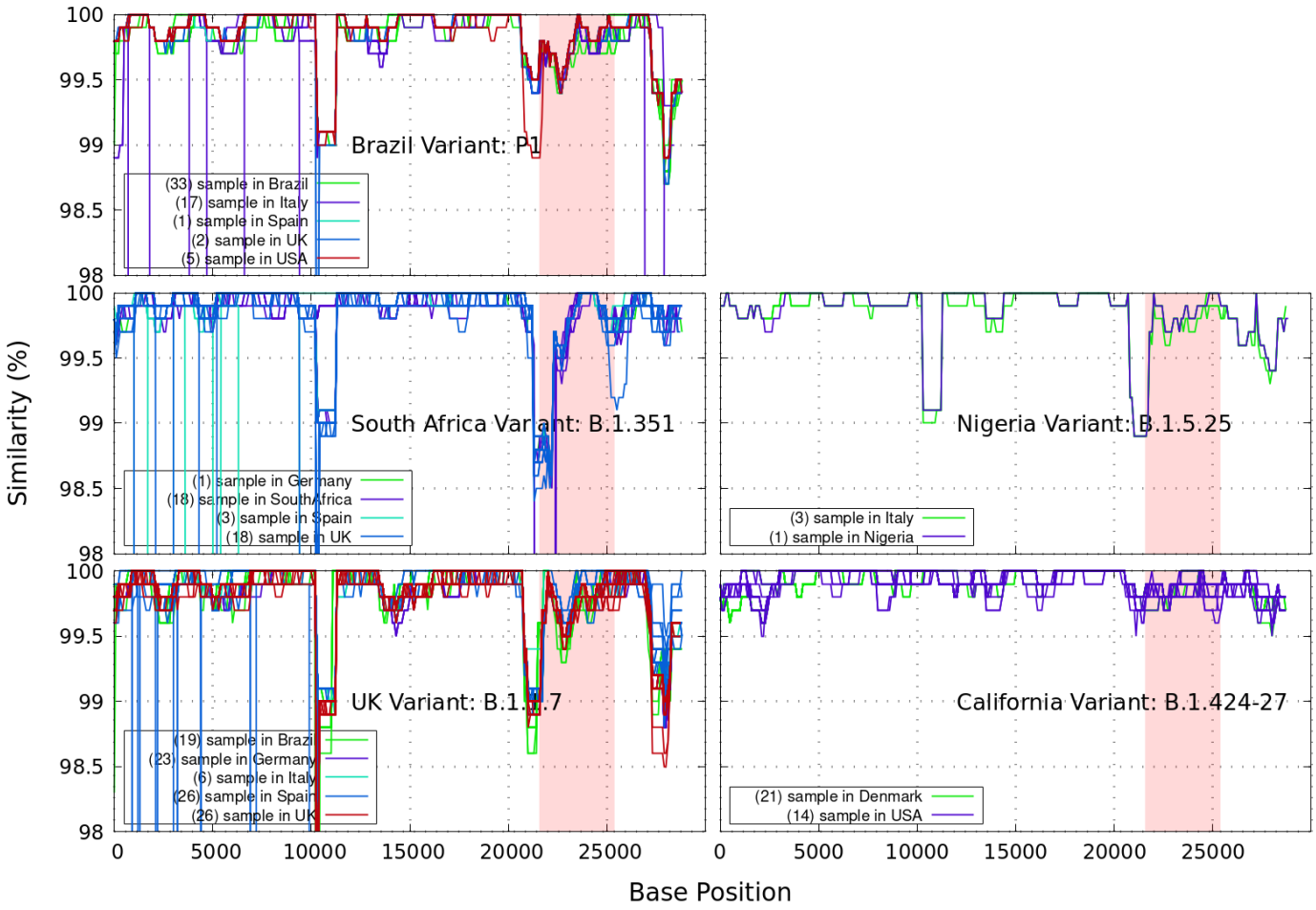

Figure 1: Genetic similarity plot between the query sequence SARS-CoV-2 Wuhan-Hu-1 and several representative full-length genome sequences grouped by variants as found in broad geographical regions, for different number of samples (indicated in brackets for each country), and collected between 01 Nov 2020 and 01 March 2021. A sliding 1000–base pair (bp) window in steps of 100-bp is used. In red is the genomic region encoding the spike (S-protein).

The command line used is: $lalign36 \ -m\ 9i\ query.file\ library.file\ -n\ -f\ -12\ -g\ 0\ -E\ 10.0\ -m\ 0\ -m\ "F11\ fasta.output"$ An iterative script was written to generate the results in Fig.1 within a 1000–base pair (bp) window sliding every 100-bp steps.

The small regions with discordant clustering ($< 2\%$) of the different coronavirus isolated with the first SARS-CoV-2 Wuhan-Hu-1 sequences shown in Fig.1, suggest that these sequences reveal extreme similarities spanning throughout the genomes as expected. The less discordant genetic similarity is found with the California and Nigeria strains. More distant relationships are detected between the query and the sequences for the Brazil and South Africa variants, which are known to share mutations (N501Y) with the UK Variant. Specifically the latter could be a consequence of the principal common valleys found at around the first 10,000 bp and in the S-protein region (positions 21563-25384 –coloured in red in the figure).

Let us analyse next the power spectrum as a function of a discrete 'frequency' f of the different coronavirus sequences with N nucleotides (of the order of $30,000$ bp) as depicted in Fig.1. Since biological sequences are strings of symbolic $\alpha = A, G, C$ and T nucleotides, binary values can be assigned to those sequences in order to apply Discrete Fourier Transform methods. Genetic sequences generate inherent signals since they are functions of an independent variable X, denoting the occurrence of a particular nucleotide in position k of the sequence. This technique has been broadly used in the literature to search for periodicity in DNA sequences [7, 8, 12].

In Fig.2, we examine correlations between the strings of symbols by this Discrete Fourier Transform. The plotted Power Spectrum of the coronavirus sequence is considered as the sum of the partial spectra: $\sum_\alpha |S_\alpha(f)|^2 = (1/N^2)\sum_\alpha^N |X_{\alpha,k} \ \ exp(2\pi i f k)|^2$, with discrete frequencies $f = 1/N,\ 2/N\ \cdots$

For all variants considered, our plots reveal a distinct peak above noise level at around frequency $f = 33.3333$, which identifies base periodicity property in the core genoma of coronavirus. To this end, the three-base periodicity is a distinctive property of protein-coding DNA sequences from a variety of organisms [7, 8].

In the past, the two methods of Similarity and Power Spectrum have been developed for computing different type of protein features. We have seen that when applied to the genoma of coronavirus variants both of these analyses do not give much info on the variants characterization themselves. We shall show next how an alternative method can provide additional insights to conventional studies, which may help to characterize variants of the current pandemic.

3 Finite and alternating sum series of virus genome

In the present paper we deal with the simplest alternating sum of the type

$$E_\alpha(X) = \sum_{k=1}^{N} (-1)^{k-1} X_{\alpha,k} \quad , \qquad (1)$$

where the variable X_α is in correspondence with one of the four nucleotide bases. The individ-

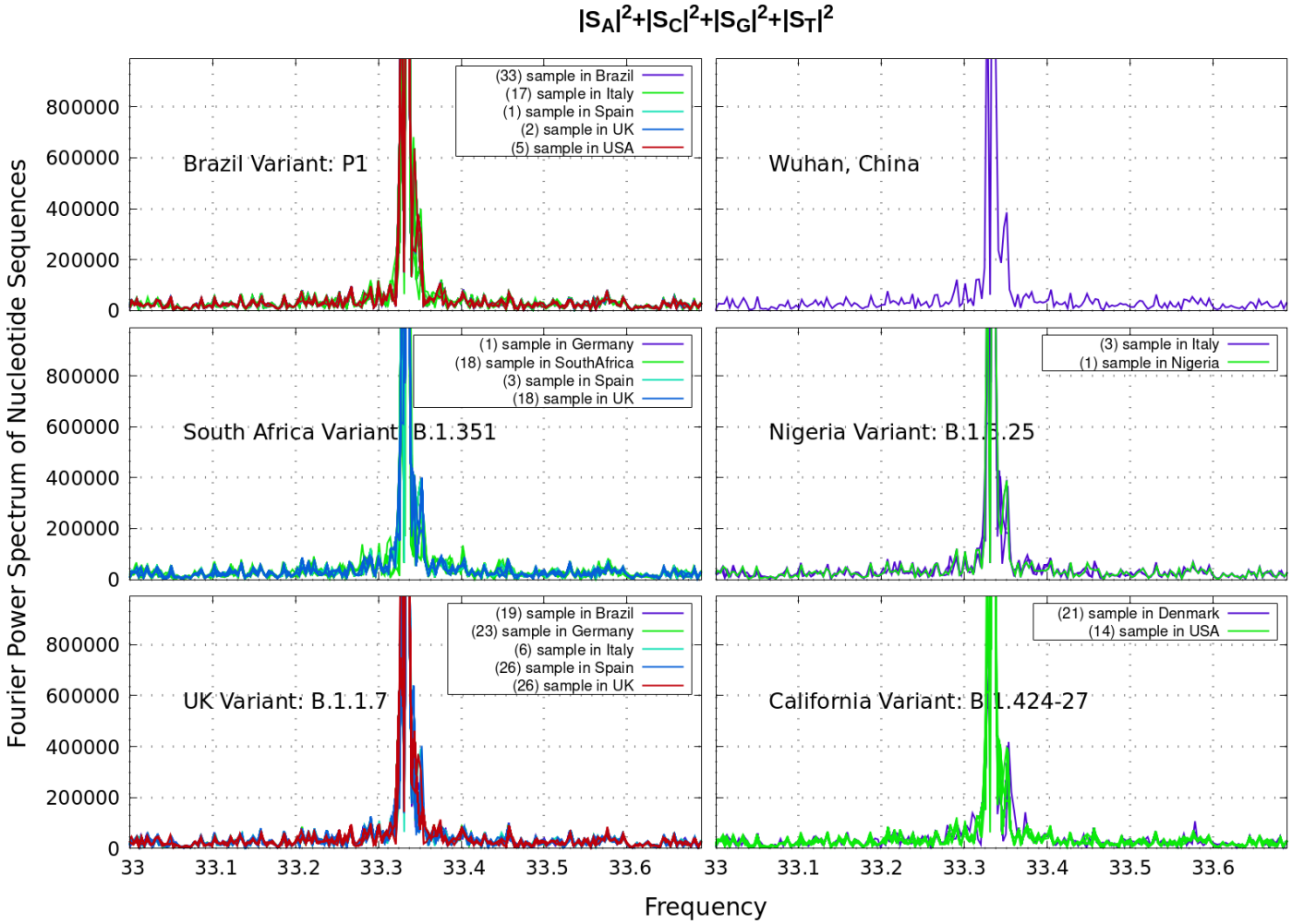

Figure 2: Discrete Fourier Transform identifying base periodicity property of coronavirus sequence. The plots reveal a distinct peak above noise level at around 'frequency' 33.3333 for all variants considered.

ual terms X_k are associated with binary 0 or 1 values according to its presence along the complete genome sequences. This mapping follows previous studies in [7, 8]. In our method, however, the arithmetic progression in Eq.(1) of the genome sequences carries positive and negative signs $(-1)^{k-1}$, and a finite non-zero first moment of the independently distributed variables X_k.

Analysing genomic sequencing via Eq.(1) measures, allows to extract unique features at each bp position with a small degree of noise variations. In figures 3-6 we display nucleotide bases A, C, G and T's imprints for the genomic strand of prevalent coronavirus variants reported from different countries for a number of samples.

Eq.(1) provides an additional tool to annotate the emerging virus patterns and it aids in their screening, comparison and classification. SARS-CoV-2 has been identified as an enveloped, single-stranded positive-sense RNA virus with a genome material encoding 27 proteins from 14 ORFs including 15 non-structural, 8 accessory, and 4 major structural proteins [13]. In the figures, it can be seen that all variants present approximately similar mirrored behaviour for the first two-thirds of the viral RNA sequence. Within the area comprising the S-protein gene sequences (drawn in red) the curves undergo strong deviations among the nucleotide bases A, C, G, T with respect to curves for the first SARS-CoV-2 Wuhan-Hu-1 sequences. Studies in [13] suggest this region encoding to be a potential target to halt the entry of SARS-CoV-2.

In particular, we note distinctive trends specially around the nucleotide region of the S-Protein. The base sequence series for Cytosine in Fig.5 shows that the UK, South Africa, Brazil and Nigeria variants share (great part of their) behaviour, whereas the California and Wuhan display essentially similar patters. To some degree, all signals for Guanine in Fig.6 are essentially similar. Nevertheless, the Nigerian variant here diverges rapidly. It is also observed when considering the Adenine nucleobase results in Fig.3. On the other hand, it is worth noting that the patterns for the base sequence series for Thymine in Fig.4, and for Adenine in Fig.3, display completely different convergences between the variants: UK, South Africa and from Brazil.

The positive and negative terms in the sums in Eq.(1) for our discrete variables partly cancel out, allowing the series 'to converge' in some variants to nearly-zero values for the nucleotide A-class as depicted in Fig.3. Most other patterns for the base sequence series seem to be of non-Cauchy sequence type. Interestingly, the observed distributions of X_k appear to be symmetric about 0 in all figures, which may imply that the mean $E(X)$ is zero. These comparative genomic statistical representations can offer insights of inherent data organization. Curves as in 3-6 could be powerful to targeting and identify variants evolution during the course of pandemic across the world.

4 Future perspective

To recap, we can conclude the following. Our method is effective and easier to apply in protein sequence comparison. It is motivated by the need to identify genetic mechanisms involved in coronavirus spreading. The added value of the

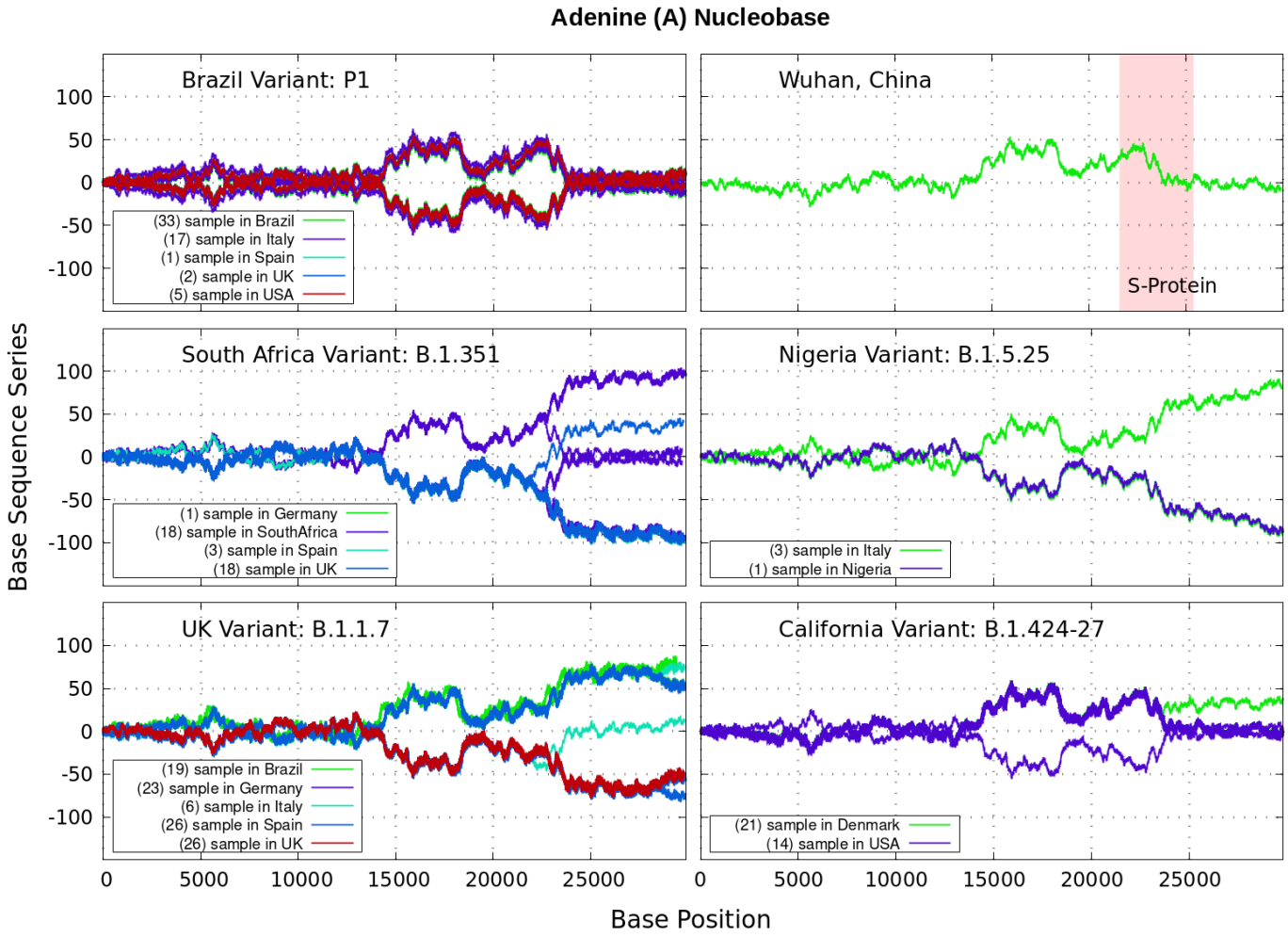

Figure 3: Variant imprints displayed by the nucleotide base: Adenine according to its progression via Eq.(1) along different number of samples (shown in brackets) of the genomic strand of coronavirus available from different countries. In red is the genomic region encoding the spike (S-protein).

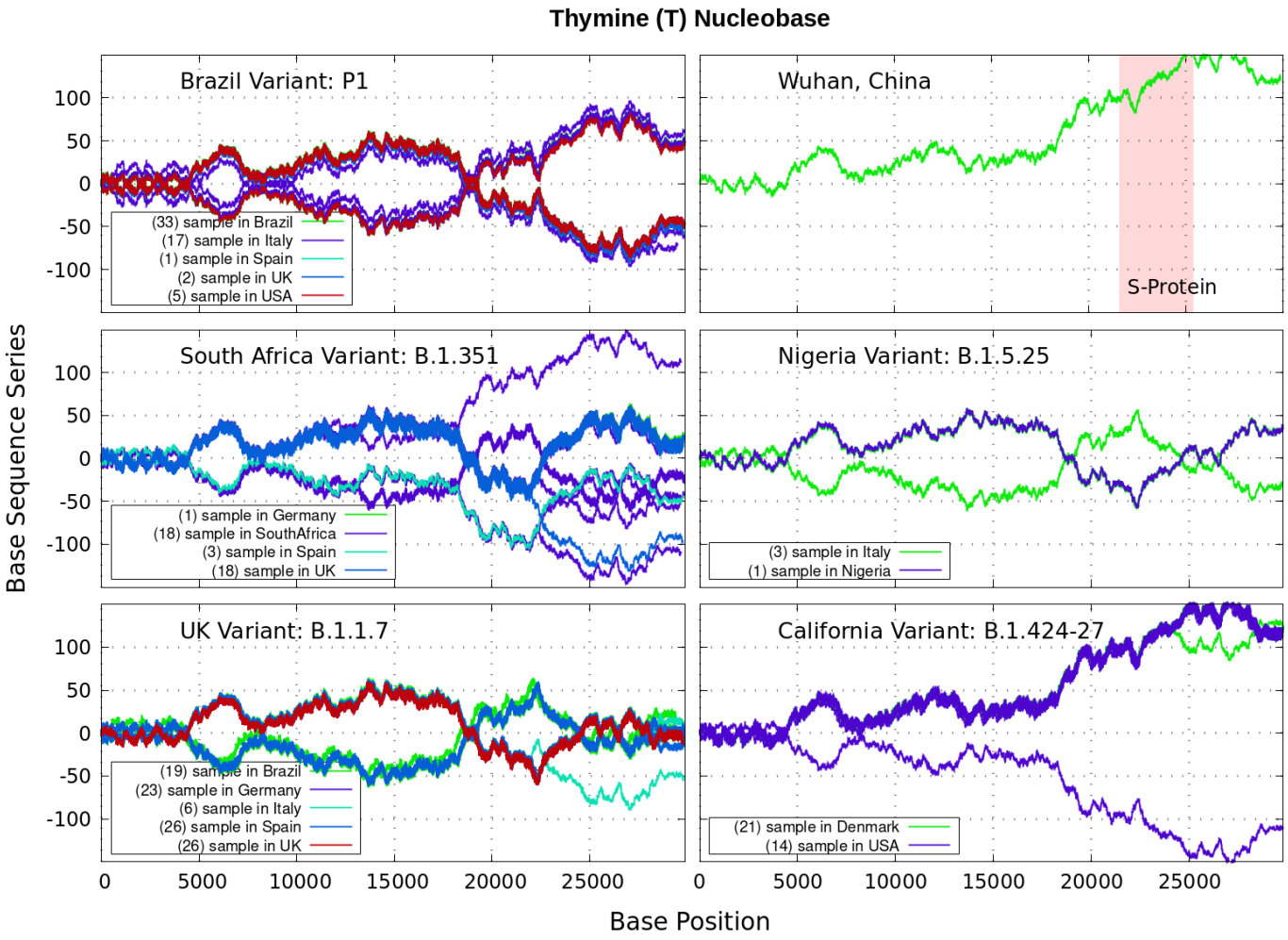

Figure 4: Variant imprints displayed by the nucleotide base: Thymine

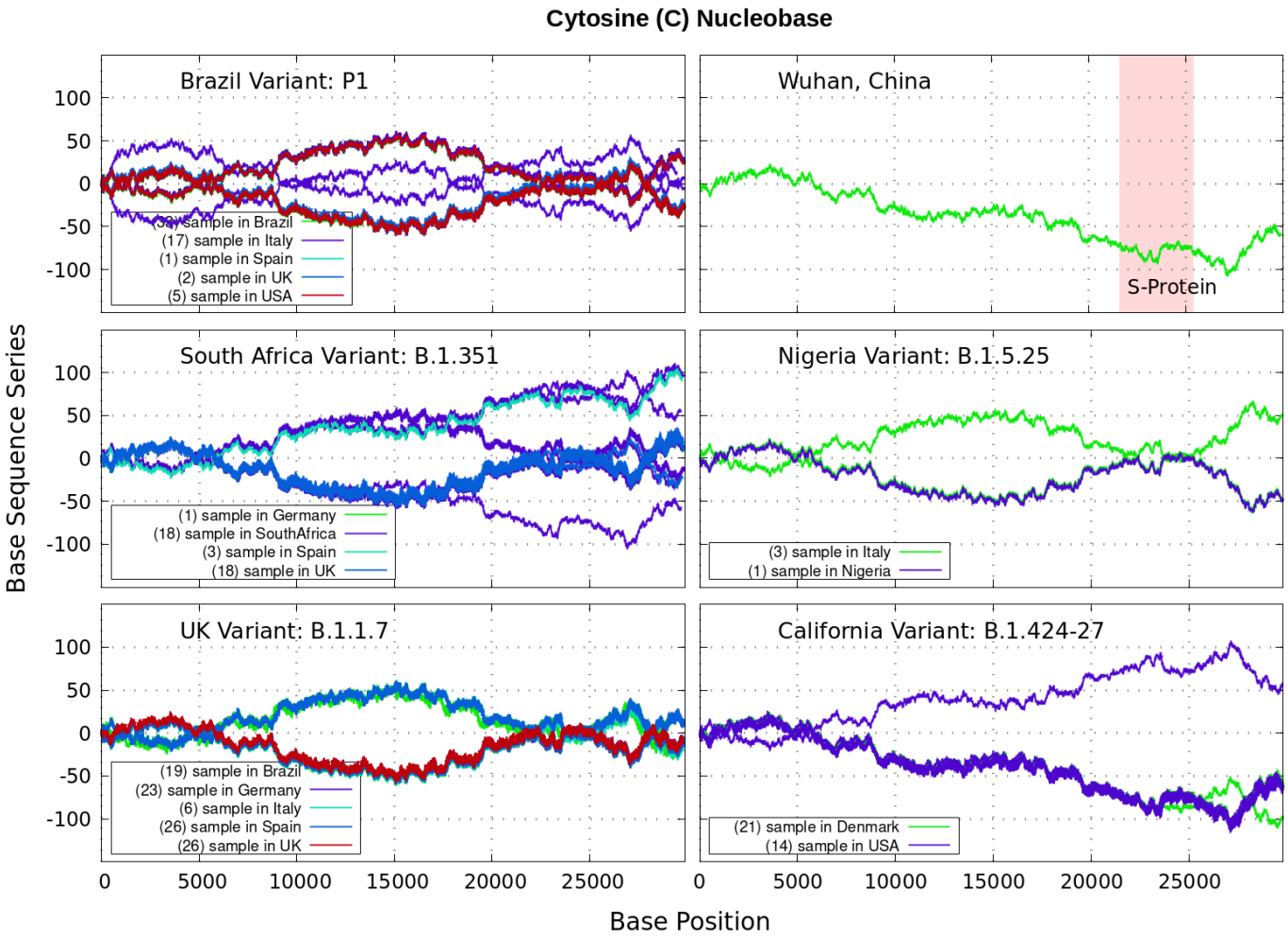

Figure 5: Variant imprints displayed by the nucleotide base: Cytosine

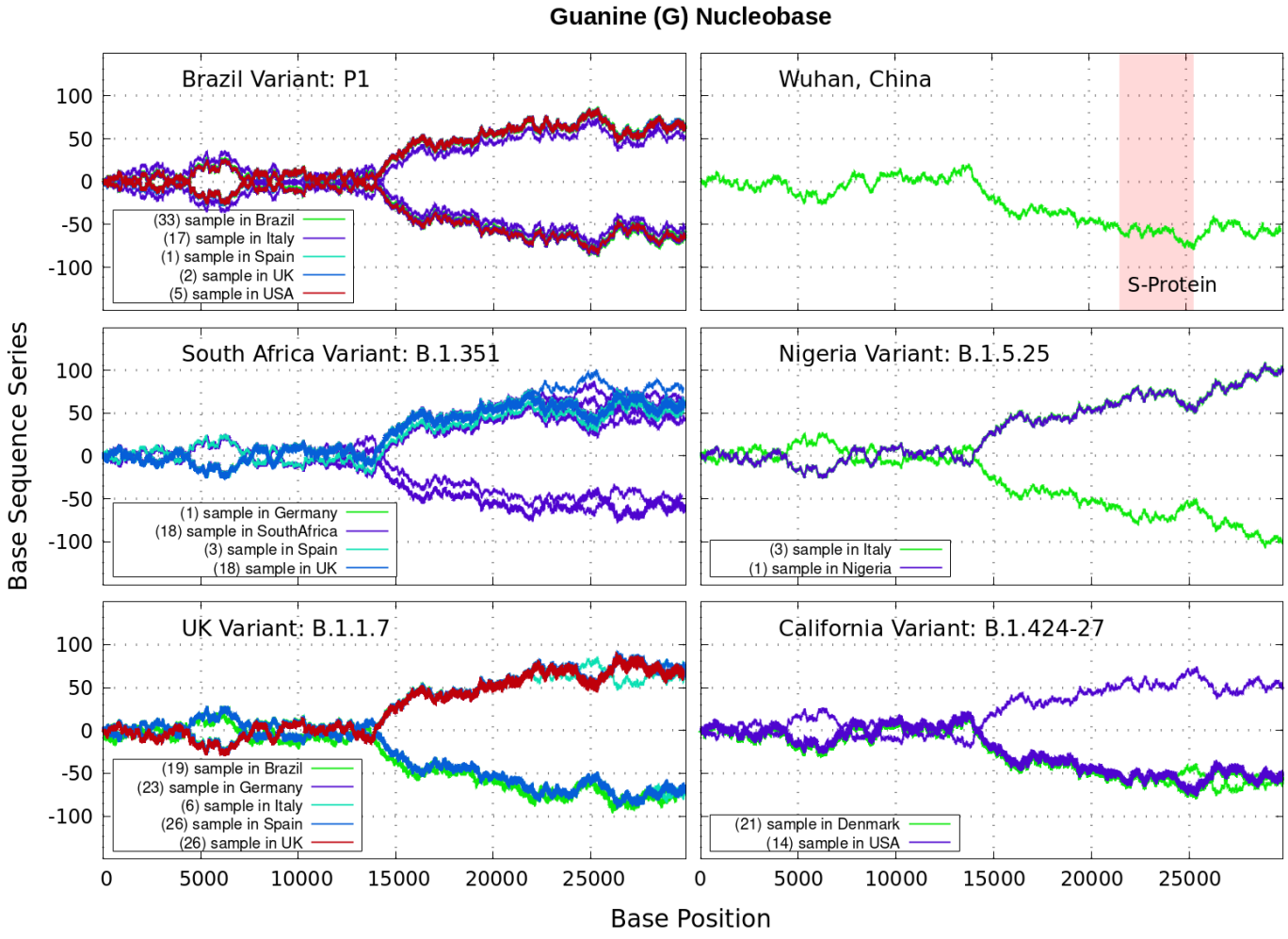

Figure 6: Variant imprints displayed by the nucleotide base: Guanine

alternating sums of the type in Eq.(1), is to have a distintive function representation of naturally occurring genoma sequences of the virus variants $X_{\alpha,k}$. The starting point is a finite alternating series following measures over N intervals. Plus and minus signs are chosen sequentially starting with +1 by default. From the view of statistics, such a sequence is equivalent to a discrete-valued time series for statistical identification and characterization of datasets [14, 15].

We have shown that these alternating sums provide additional information to conventional Similarity comparisons and Power Spectrum approaches. We also verified that the Discrete Fourier transform of the complete alternating sum series (not shown), leads to a peaked structure at a 'frequency' equal to $16,6666$ meaning for the sums a 100/6 characteristic periodicty. Distinctive trends have been identified specially around the nucleotide region of the Spike protein for all variants studied. These emerging variants seem to have only few mutations, *i.e.*, no more than a dozen amino acids changes out of the 1200 building blocks that make up the Spike protein. They give a selective advantage for their replicative capacity [5, 16].

We downloaded worldwide sequencing data of coronavirus variants from GISAID and verified multiple deviations from the originating first Wuhan sequences identified a bit more than a year ago. Our statistical representation of coronavirus genoma variants, taking summands with both signs, can reveal signals from future genome evolution at the level of nucleotides ordering. These observations lie at the heart of future studies. One could investigate further the present numerical results to study the n-moment calculations $E(X^n)$, and the tail of the series, which could light even more underlying properties of the sequences of viral strains related to the pathogenesis of SARS-CoV-2.

Declaration of competing interest

None.

Acknowledgments

The author gratefully acknowledges the GISAID Initiative and all of the submitting laboratories of the coronavirus genetic sequences, on which this analysis is based.

References

[1] Lu R., Zhao X. *et al.*, 'Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding'. *The Lancet* **395** (2020) 565-574. doi: 10.1016/s0140-6736(20)30251-8

[2] Zhou P., Yang X-L. *et al.*, 'A pneumonia outbreak associated with a new coronavirus of probable bat origin'. *Nature* **579** (2020) 270-273. doi: 10.1038/s41586-020-2012-7

[3] Wu F., Zhao S. *et al.*, 'A new coronavirus associated with human respiratory disease in China'. *Nature* **579** (2020) 265-269. doi: 10.1038/s41586-020-2008-3 *Erratum: Nature* doi:10.1038/s41586-020-2202-3

[4] Li X., Giorgi E.E. *et al.*, 'Emergence of SARS-CoV-2 through recombination and strong purifying selection'. *Science Adv.* **6** (2020) 1-11. doi: 10.1101/2020.03.20.000885

[5] Du L., He Y. *et al.*, 'The spike protein of SARS-CoV —a target for vaccine and therapeutic development'. *Nature Reviews Microbiology* **7** (2009) 226-236. doi: 10.1038/nrmicro2090

[6] Koch L., Poteski C. and Trenkmann M. –Coordinating editors, 'Milestones in Genomic Sequencing'. *Nature Milestones* February 2021. www.nature.com/collections/genomic-sequencing-milestones

[7] Chechetkin V.R. and Turygin A. Yu, 'Size-dependence of three-periodicity and long-range correlations in DNA sequences'. *Phys. Letters A* **199** (1995) 75-80. doi: 10.1016/0375-9601(95)00047-7

[8] Hong T., Yin C. *et al.*, 'A new method to cluster DNA sequences using Fourier power spectrum'. *J. Theo. Biology* **372** (2015) 135-145. doi: 10.1016/j.jtbi.2015.02.026

[9] Pal J., Ghosh S. *et al.*, 'Use of FFT in Protein Sequence Comparison under Their Binary Representations'. *Comp. Mol. Bioscience* **6** (2016) 33-40. doi: 10.4236/cmb.2016.62003

[10] Touati R., Haddad-Boubaker S. *et al.*, 'Comparative genomic signature representations of the emerging COVID-19 coronavirus and other coronaviruses: High identity and possible recombination between Bat and Pangolin coronaviruses'. *Genomics* **112** (2020) 4189-4202. doi: 10.1016/j.ygeno.2020.07.003

[11] Yang X-L., Hu B. *et al.*, 'Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus'. *J. Virology* **90** (2016) 3253-3256. doi: 10.1128/JVI.02582-15

[12] Voss R.F., 'Evolution of Long-Range Fractal Correlations and 1lf Noise in DNA Base Sequences'. *Phys. Review Letters* **68** (1992) 135-145. doi: 10.1103/PhysRevLett.68.3805

[13] Naqvi A.A.T., Fatima K. *et al.*, 'Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis andtherapies: Structural genomics approach'. *BBA -Molecular Basis of Disease* **1866** (2020) 165878-165895 doi: 10.1016/j.bbadis.2020.165878

[14] Canessa E., 'Modeling of body mass index by Newton's second law'. *J. Theoretical Biology* **248** (2007) 646-56. doi: 10.1016/j.jtbi.2007.06.011

[15] Canessa E., 'Multifractality in time series'. *J. Physics A: Math. Gen.* **33** (2000) 3637-3651 doi: 10.1088/0305-4470/33/19/302

[16] Pardi N., Hogan M.J. *et al.*, 'mRNA vaccines —a new era in vaccinology'. *Nature Reviews Drug Discovery* **17** (2018) 261-279. doi: 10.1038/nrd.2017.243