GenomeBits insight into omicron and delta variants of coronavirus pathogen

Enrique Canessa[1]*, Livio Tenze[1]

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

* canessae@ictp.it

Abstract

Differences and correspondence regarding the intrinsic data organisation of complete delta and omicron genome sequences according to its progression along the nucleotide A,C,G,T bases position are analysed. We found a sort of *'ordered'* to *'disordered'* transition around the S-spike protein region in the curves obtained from finite alternating sum series having independently distributed terms associated with (0,1) binary indicators for the nucleotides. To uncover such underlying features of genome sequences may assist in the development of synthetic proteins.

keywords: SARS-CoV-2 variants; genome sequence analysis; statistical methods; alternating series

Introduction

Delta (lineage AY.4.2) and omicron (B.1.1529) variants of viral strains related to the pathogens of SARS-CoV-2 are today of the greatest concern globally. The first known confirmed delta variant was detected in India in late 2020 and the B.1.1.529 infection appeared from a specimen collected in South Africa a year later, early November 2021.

The delta and omicron variants share some parts of their structures [1, 2]. Both variants have common mutations, i.e., amino acid changes in the building blocks that conform the spike protein (D614G mutation), and different mutations elsewhere (the P323L mutation in the NSP12 polymerase, and the C241U nucleotide mutation in the 5' untranslated region). The latter seems to give a greater advantage for their replicating or transmissibility capacity. Although omicron seems to cause less severe COVID-19 than delta.

The ongoing SARS-CoV-2 research is currently focusing on understanding the essential functions of the conforming proteins in the ribonucleic acid RNA coronaviruses [3]. These use an unusually large collection of RNA synthesizing and processing enzymes to express and replicate genome sequences that are targets for antiviral drug design. One possibility to reveal and undertand the genome organization of viruses is through statistical methods in order to associate and characterise their different variants. These studies may be useful, e.g., for the development of synthetic messenger RNA.

Genome sequence comparisons can be studied using the new GenomeBits method [4, 5]. This is a statistical algorithm based on a finite alternating sum series having independently distributed terms associated with (0,1) binary indicators for the nucleotide bases. In this paper, we apply GenomeBits to uncover distinctive patterns

and common features regarding the intrinsic data organisation of delta and omicron genome sequences from the GISAID archive at www.gisaid.org, according to its progression along the nucleotide A,C,G,T base positions (denoted as bp). We report a kind of *'ordered'* (or constant) to *'disordered'* (or peaked) transition around the S-spike protein region. This can provide additional information to conventional comparative Similarity via alignment methods [6], specially on the single nucleotide structures.

Comparison Methods

Similarity plots

Similarity between full-length genome sequences is the standard method for determining whether there are sequence homology in terms of shared ancestry between them in their evolutionary history using alignment methods [6, 7]. To determine the best parameters to achieve optimal alignments is difficult. There are several user-defined parameters to overcome gaps and mismatches usually found between genome sequences. The computational resources required increase considerably depending on the length and number of sequences being aligned.

Genetic Similarity plots of different query sequences of SARS-CoV-2 genome are shown in Fig 1. We downloaded genome sequence data in FASTA format as the only imput from GISAID collected in December 2021 (whose accession IDs are indicated in the figures). The FASTA format is the standard text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes: (A)denine, (C)ytosine, (G)uanine and (T)hymine (or Uracil RNA genome for single strand folded onto itself), which store instructions to assemble and reproduce every living organism.

In the plot, we verified multiple deviations of both delta and omicron variants against one of the first Wuhan-China sequences identified over a year ago (MN908947) [8]. Multiple deviations from the delta variants from Spain (EPI ISL 7900475 and 7951710) against several omicron variants from Spain, plus the delta variant from Spain EPI ISL 8179449 against different omicron variants from the USA are also show in Fig 1. For the calculations we used "lalign36" sequence comparison software (via the Waterman–Eggert algorithm at github.com/wrpearson/fasta36) [4].

In the figure, regions with clustering ($< 1\%$) of biological sequences of SARS-CoV-2 from the city of Wuhan to the delta lineage suggest some genetic similarities outside the S-spike gene region (bp 21563-25384, coloured in clear blue). More divergent genetic similarities are found with respect to the omicron strains, reduced to $\sim 97\%$, and in between the delta against omicron variants as a consequence of the mutations. Conventional Similarity comparisons via alignment provides only limited information on their progression along the genome sequences of the single nucleotide bases A,C,G,T.

GenomeBits representation

Our new quantitative method for the examination of distinctive patterns of complete genome data consists of a certain type of alternating series having terms converted to (0,1) binary values for the nucleotide variables $\alpha = A, C, T, G$ as observed along the reported genome sequences, namely

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

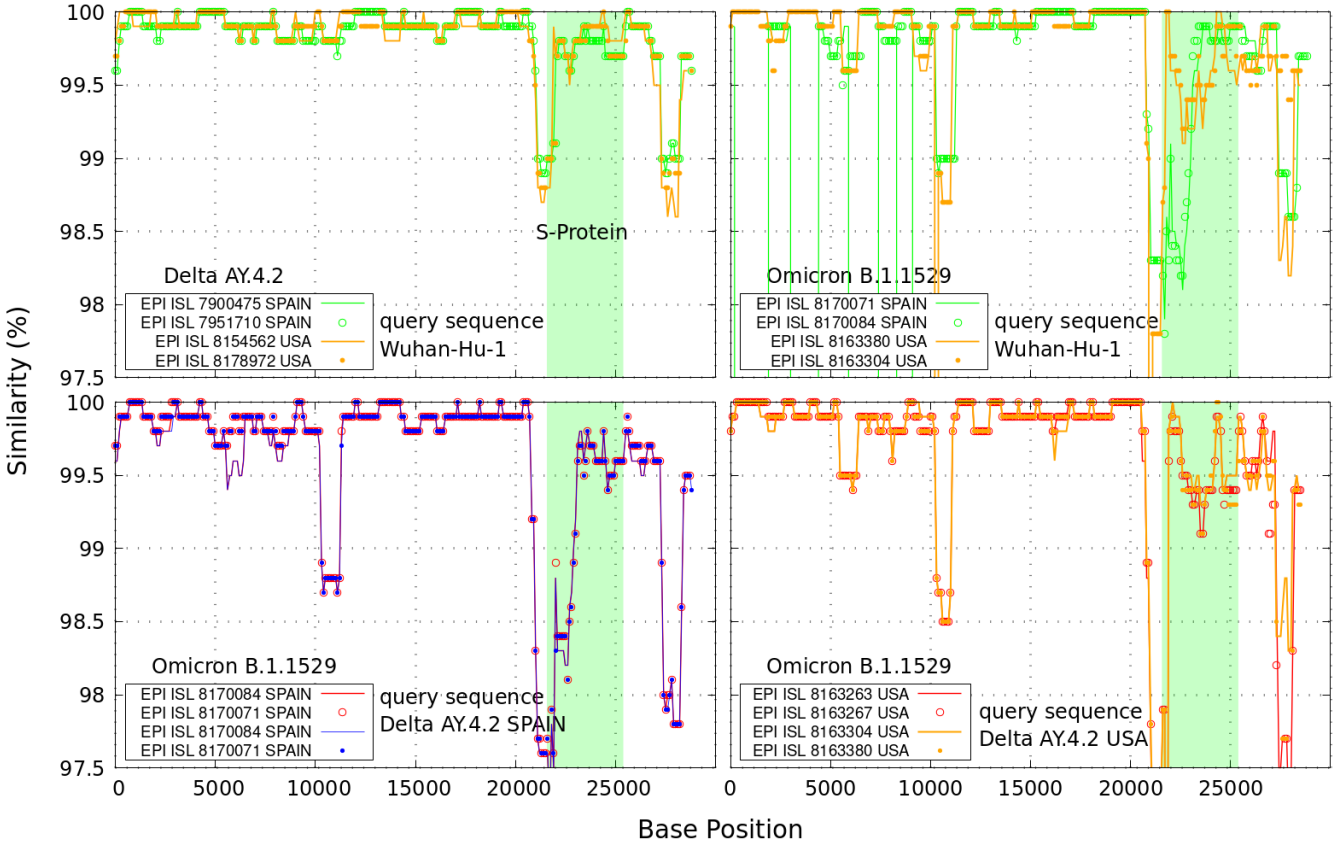

Fig 1. Similarity plots. *Upper curves: genetic similarity curves between the query sequence SARS-CoV-2 Wuhan-Hu-1 and representative delta and omicron complete genome sequences. In clear blue is the genomics region encoding the spike (S-protein). Lower curves: delta genome sequences used as query against omicron data from Spain and USA. A typical sliding 1000 base pair window in steps of 100 bp was used in the calculations.*

where the individual terms X_k are associated with 0 or 1 values according to their position along the genome sequences of length N, satisfying the following relation

$$X_{\alpha,k=N} = |E_{\alpha,N}(X) - E_{\alpha,N-1}(X)| \, . \tag{2}$$

The arithmetic progression carries positive and negative signs $(-1)^{k-1}$ and a finite non-zero first moment of the independently distributed variables $X_{k,\alpha}$. Plus and minus signs are chosen sequentially starting with $+1$ at $k = 1$ by default.

This mapping into four binary projections of genome sequences follows previous studies on the three-base periodicity characteristic of protein-coding DNA sequences [9]. Analysing genomics sequencing via this type of finite alternating sums allows to extract unique features at each bp with a small degree of noise variations. From the view of statistics, our series is equivalent to a discrete-valued time series for the statistical identification and characterisation of (random) data sets [10].

GenomeBits is a user-friendly Graphics User Interface (GUI) to the present signal

analysis method of genome sequences according to its progression along the nucleotide bp [4]. It runs under Linux Ubuntu O.S. and can be downloaded from Github [5]. GenomeBits considers samples with A,C,T,G sequences for (up to two) given Countries corresponding to genomics sequence data from (up to six) given variants/species and it discards uncompleted sequences containing codification errors (usually denoted with "NNNNN" and other letters). In brief, GenomeBits allows with one click to

- run alternating sums in Eq (1) for up to six-times-two inputs of FASTA files containing (i.e., concatenating) more than one genome sequence each.
- Separate concatenated genome sequences and save in single FASTA files (for each Country), which may be containing in a single FASTA input file including more than one genome sequence.
- Get into single files each of the four nucleotide bases represented by the symbols A,C,T,G.
- Get the alternating sums results in single files for each of the four nucleotide bases A,C,T,G associated with 0 or 1 binary values according to its presence along the genome sequences.
- Plot the alternating sums of the binary data files to compare behaviour of the pairs A,T and C,G nucleotide bases versus bp.
- Compare in one plot the alternating sums results versus bp for all four nucleotide bases A,C,T,G.
- Plot the alternating sums curves versus bp for each of the four nucleotide bases A,C,T,G.
- Plot the alternating sums results versus bp for the four nucleotide bases A,C,T,G (if countries given as input).
- Plot in one image all the GenomeBits GUI results versus bp for each nucleotide base A,C,T or G for up to six variants/species and up to 4 FASTA files by Country.
- Visual compare the GenomeBits GUI curves versus bp for (up to six) given variants/species and (up to two) selected Countries, with those results from the original paper in [4].

Results

The GenomeBits statistical representation of coronavirus genome variants, by adding binary values with $\pm$ signs following Eq (1) and using GenomeBits, can reveal interesting imprints of the genome dynamics at the level of nucleotide ordering. By this method of binary projections we are able to uncover distinctive signals of the intrinsic gene organisation embedded in the genome sequences of the single-stranded RNA coronaviruses.

In Fig 2, we show the results obtained for the genes in the sequences of each nucleotide A, C, G and T of the coronavirus variants of concern —AY.4.2 (delta) and B.1.1529 (omicron) reported from Spain and the USA for a number of representative samples as indicated. As reference, we display on the left our results for nucleotides A,C of the strand and on the right the nucleotides T,G ("complementary to those of the opposite strand" —according to the pairing rules A-T and C-G of DNA). The complete

genome sequences consists of N nucleotides on the order of 30,000 base pairs in length, two to three times larger than that of most other RNA viruses [3].

It is interesting to note how in the Figure there are regions where the curves for the delta variant (in blue) mirror those of the omicron variant (in green). This peculiar behaviour becomes clear by averaging both curves as shown by the red lines. The regions of (almost) zero or rather constant average values indicates perfect mirroring matching, which is driven by the $\pm$ signs of the alternating series. It reveals sequence correspondences between the delta and omicron variants.

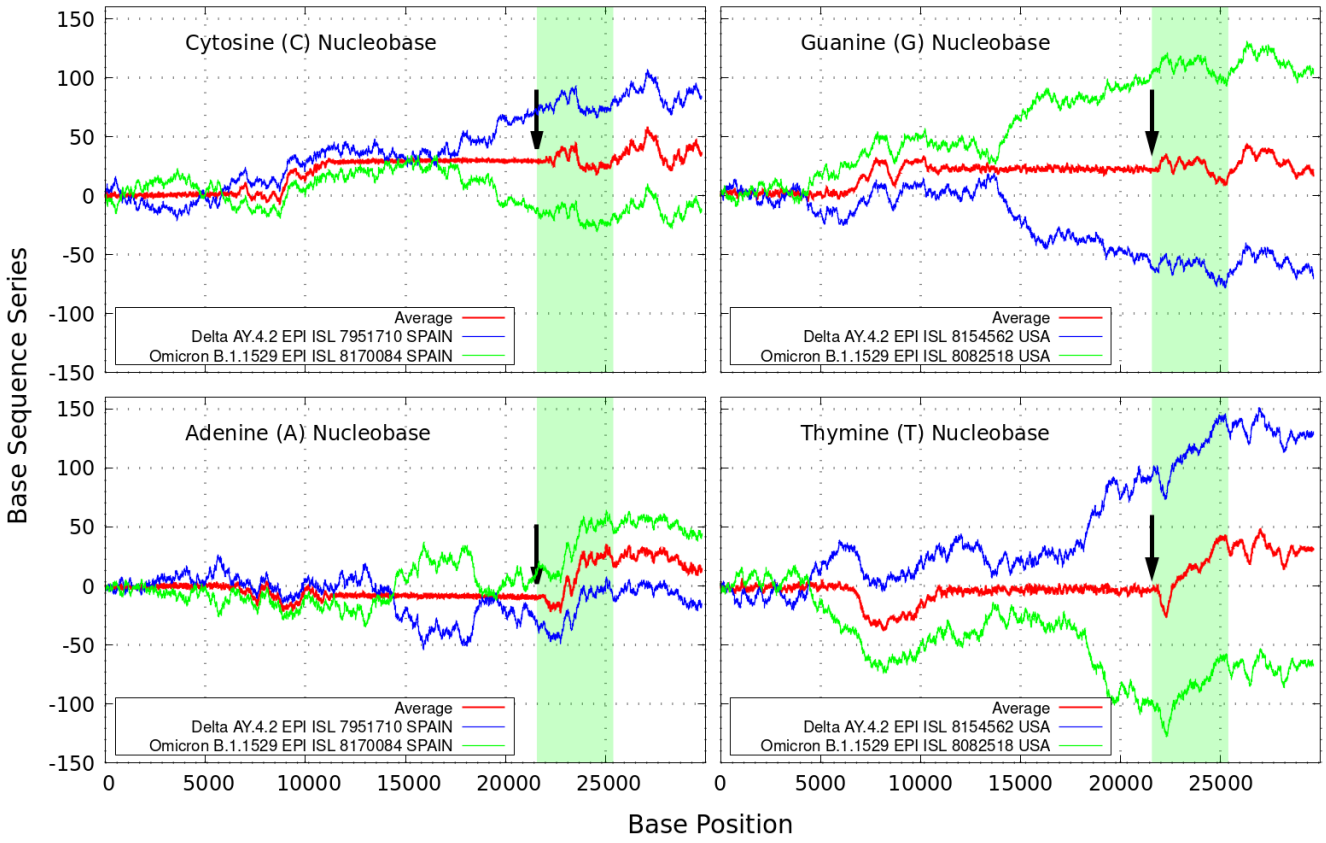

Fig 2. Sequence sum series. *Delta (in blue) and omicron (in green) variant imprints displayed by the nucleotides A,C,G,T according to their progression via Eq (1) along different samples of the genomics strand of coronavirus available from Spain and USA. The arrows indicate a sort of 'ordered' (constant) to 'disordered' (peaked) transition before the coding region of the S-spike genes of the SARS-CoV-2 Wuhan-Hu-1 sequence (drawn in clear blue).*

The regions of main discrepancies as found in the Sequence identities curves of Fig 1, e.g., around $N = 10000$ are also reflected by the red lines of Fig 2. The main difference between both comparative genomics approaches is that changes via Eq (1) can be analysed and characterised at each single A,C,G,T nucleotide level. We found a kind of *'ordered'* (constant) to *'disordered'* (peaked) phase transition phenomena around the NSP5 polymerase within the open reading frames ORF1a region [1, 2], up to the

nucleotide region of the S-Protein (coloured clear blue area).

To some degree, there are also distinctive trends especially around the S-Protein. As seen in Fig 2, the black arrows indicate a phase transition point appearing close to the coding region of the S-spike genes. The peaked curves diverge rapidly and tend to separate denoting bigger dissimilarities for increasing N. It is worth noting that the patterns for the base sequence series for Adenine and Cytosine display completely different convergences between the variants. The positive and negative terms in the sums in Eq (1) for the discrete α variables partly cancel out, allowing the series "to converge" to some non-zero values for all the nucleotide classes. This feature allows to estimate a separation ratio of ~ 2 between nucleotide bases C and T curves, and ~ 4 times higher between the curves for A and T at the black arrows bp.

Remarks

We believe the additional underlying properties of the genome sequences for mutant pathogens as derived from a simple alternating series, and following measures over N intervals, may allow to locate and distinguish polymers of amino acids (proteins states and positions) in a sequence and determine if the altered genes behave similar to those already targeted.

GenomeBits may shed further light on the bioinformatics surveillance behind future infectious diseases. It may be of some relevance to assist in further developments of synthetic mRNA-based vaccine designs [3, 11]. Such comparative genomics statistical representations can offer insights on the inherent data organisation during the natural evolution of pandemic.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Stanford University, USA Coronavirus antiviral & resistance database -Last updated on 1/10/2022. https://covdb.stanford.edu/page/mutation-viewer/

2. Kumar A., Asghar A., Singh H.N. et al. An in silico analysis of early SARS-CoV-2 variant B.1.1.529 (Omicron) genomic sequences and their epidemiological correlates. medRxiv 2021.12.18.21267908; doi: https://doi.org/10.1101/2021.12.18.21267908

3. Malone B., Urakova N., Snijder E.J. et al. Structures and functions of coronavirus replication–transcription complexes and their relevance for SARS-CoV-2 drug design. Nature Rev Mol Cell Bio. 2022; 23:21-39; doi: https://doi.org/10.1038/s41580-021-00432-z

4. Canessa E. Uncovering signals from the coronavirus genome. Genes (Basel) 2021; 12(7):973; doi: https://doi.org/10.3390/genes12070973

5. Canessa E., Tenze L. GenomeBits: A tool for the signal analysis of complete genome sequences https://github.com/canessae/GenomeBits/ (Last visited 1/10/2022).

6. Hu B., Zeng L.-P., Yang X.-L. et al. Discovery of a rich gene pool of bat SARS related coronaviruses provides new insights into the origin of SARS coronavirus. PLOS Pathog 2017; 13(11):e100669; doi: https://doi.org/10.1371/journal.ppat.1006698

7. Lu R., Zhao X., Li J. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet 2020; 395(10224):565; doi: https://doi.org/10.1016/S0140-6736(20)30251-8

8. Zhou P., Yang X.-L., Wang X.-G. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020; 579:270-273; doi: https://doi.org/10.1038/s41586-020-2012-7

9. Chechetkin V., Turygin A. Size-dependence of three-periodicity and long-range correlations in DNA sequences. Phys. Lett. A 1995; 199(1):75; doi: https://doi.org/10.1016/0375-9601(95)00047-7

10. Canessa E. Multifractality in time series. J. Phys. A Math. Gen. 2000; 33(19):3637; doi: https://doi.org/10.1088/0305-4470/33/19/302

11. Elkashif A., Alhashimi M., Sayedahmed E.E. Adenoviral vector-based platforms for developing effective vaccines to combat respiratory viral infections. Clinical & Translational Immunology 2021; 10(e1345):1; doi: https://doi.org/10.1002/cti2.1345