Article

GenomeBits Characterization of MPXV

Enrique Canessa

International Centre for Theoretical Physics—ICTP, 34151 Trieste, Italy; canessae@ictp.it

Abstract: Genome sequences of the monkeypox virus (MPXV) causing the current outbreak are being reported from an increasing number of countries. We present a letter-to-numerical sequence study via GenomeBits signal mapping in order to characterize the evolution of the MPXV via simple statistical analysis. Histograms, empirical and theoretical cumulative distribution curves and the resulting scatter plots for the base nucleotides A and C versus their complementary base nucleotides T and G are discussed. GenomeBits may help the surveillance of emergent infectious diseases.

Keywords: monkeypox virus; MPXV surveillance; genome sequence; GenomeBits

1. Introduction

According to WHO, the monkeypox virus (MPXV) poses new threats to humanity [1]. As surveillance expands in non-endemic countries, the cases of monkeypox being identified continue to rise with the risk of becoming an outbreak. Proper precautions and steps are necessary to avoid another global pandemic in 2022–2023.

MPXV is an infectious disease caused by the Zoonotic Orthopoxvirus closely related to cowpox and smallpox viruses—first identified in a monkey in Denmark at the end of the 1950s. Although it is mainly transmitted by monkeys and rodents, the human-to-human spread is also prevalent since the first case was recorded in central Africa in 1970. As of today, individuals infected by the MPXV are reported across continents. MPXV is transmitted through close physical contact with an infected person, animal or object, and its principal long-term effect is the presence of human skin lesions.

Understanding this pathogen better is still a global challenge for scientific research. A global network of permanent bioinformatics surveillance, with the capacity to rapidly deploy genomic tools and statistical studies for the MPXV characterization of the evolution of lineages and mutations, is still much needed. There are hundreds of complete genome sequence samples reported from confirmed cases from around the world [2–4], which indicate a close match to exported cases from Africa found in 2017 [5].

In this paper, we present a novel method for characterizing the MPXV genome from available data sequences. We apply the novel GenomeBits statistical algorithm, whose salient feature is to map the nucleotide bases $\alpha = A, C, T, G$ (as observed along a genome sequence) into a finite alternating ($\pm$) sum series of distributed terms of binary (0,1) indicators [6]. We show how GenomeBits allow infectious MPXV surveillance by tracking and analyzing viral sequences in relation to their evolution status. In particular, we derive scatter plots for the nucleotides A-T and C-G to uncover correlations between MPXV sequences of complementary nucleotide bases.

2. Method

The new quantitative method presented for the characterization of distinctive patterns of complete genome sequences of length $N$ considers

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

(1)
where the independently distributed terms $X_{k,\alpha}$ are associated with (0,1) according to their position along the sequences, satisfying the following relation $X_{k,\alpha} = |E_{\alpha,N}(X) - E_{\alpha,N-1}(X)|$. The arithmetic progression carries positive and negative signs $(-1)^{k-1}$ depending on the nucleotide base position $k$.

This mapping into four binary projections of genome sequences follows previous studies on the three-base periodicity characteristic of protein-coding DNA sequences [7]. However, as a principal difference from other binary representations (see also, e.g., Refs [8,9]), in Equation (1), the terms in the sums change sign. If a term $X_{\alpha,k}$ in the sum series is positive at a given $k$, then the next $X_{\alpha,k+1}$ term is negative, and vice versa.

The binary indicator for the sequences is chosen sequentially, starting with $+1$. This is illustrated in Table 1, where a mapping example for converting the brief genome fragment TTTAGTACATTAAT (consisting of 15 nucleotides) into the alternating binary array via Equation (1) is given. As illustrated in Table 1, the positive and negative terms in the sums of Equation (1) partly cancel out, allowing the sum series to diverge from zero rapidly and to become a non-Cauchy sequence type [6].

Table 1. Genome sequence-to-GenomeBits mapping example via Equation (1) for $N = 15$.

| Base Position $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|-------------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|
| Sequence (String) | T | T | T | T | A | G | T | A | C | A | T | A | A | T | T |
| genome sequence   | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| GenomeBits sums   | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |

Letter sequence-to-numerical GenomeBits signal mapping. Single letters represent nucleotide codes: (A)denine, (C)ytosine, (G)uanine and (T)hymine. Signs are assigned sequentially with $+$ at base position $k = 1$.

There is a user-friendly Graphics User Interface (GUI) for the GenomeBits signal analysis method of genome sequences in FASTA format. The GUI can be downloaded freely from GitHub (see https://github.com/canessae/GenomeBits/, accessed on 24 November 2022) [10]. It runs under Linux Ubuntu OS. For large genomic data, say, of the order of $N = 197,204$ nucleotide bases as, for example, for ON563414.3 monkeypox virus isolate MPXV_USA_2022_MA001 complete genome [3], the one used in the present study, the GUI requires little processing time and discards uncompleted sequences containing codification errors (usually denoted with “NNNNN” letters).

3. Discussion

Next, we show some interesting statistical imprints of the intrinsic gene organization at the level of nucleotides from the above formula. In Figure 1, the results obtained via Equation (1) for the binary sequences of each A, C, G and T nucleotides of the MPXV reported from the USA are shown. For comparison, the curves (a) on the top display results for nucleotides A and T of the strand and on the bottom (b) are the nucleotides C and G (complementary to those of the opposite strand according to the pairing rules A-T and C-G of DNA). It is interesting to note how in the figure there are regions where the curves seem to mirror each other. This peculiar behavior reveals coding regions of correspondence between single-nucleotide structures.
Figure 1. GenomeBits imprints of the genome sequence monkeypox virus isolated in the United States MPXV_USA_2022_MA001 (GenBank ID ON563414.3): (a) Results according to the pairing rules A-T and (b) pairing rules C-G.

Moreover, the fact that for $\alpha = (A)$, the total number of zeros is 66.4% of the whole sequence and the total number of ones is 33.6% is also interesting. For $\alpha = (C)$, these correspond to 83.5% and 16.5%, respectively. For $\alpha = (G)$, one finds 83.5% and 16.5% and, similarly, for $\alpha = (T)$, the values become 66.6% and 33.4%.

A first step in assessing the probability distribution of datasets is usually obtained by computing a basic histogram (i.e., counting the number of data points in a given bin). A standard cumulative, normalized histogram as a step function in order to visualize the empirical cumulative distribution function (CDF) for a sample is also useful [11]. In this case, one gets the probability of, and observation from, the sample not exceeding a particular $x$-value at a fixed $y$-value. However, our goal here is to compare two different datasets or, alternatively, two different histograms of equally sized bins.

To achieve this, a pair of histograms above and to the right of a scatter plot of such two datasets can help to visualize and highlight correlations at certain colored contact points. Scatter plots with marginal distributions of this kind can be easily made using Python programming through the built-in Matplotlib’s routines: scatter(x,y), hist(x), gridspec and invert_xaxis (see, e.g., Ref. [12]). In Figures 2 and 3, we illustrate the (blue and green bar-graph) histograms, the empirical (red full line) and theoretical (red dotted lines) CDF curves, and the resulting scatter plots for the nucleotides A and its complementary T and C and its complementary G plotted in Figure 1. We have used 50 bins, and the fits (blue line) have been calculated via an expected Gaussian distribution.

In Figures 2 and 3, we have drawn scatter diagrams in an attempt to obtain correlations and anomalies between GenomeBits sequences of complementary nucleotide bases. The clearer colors of the hexagon markers add a new dimension to these plots. These show how the mapped binary data is stratified along the sequences and which areas show somehow correlated points that fall along complex, rounded shapes. The markers are unequally distributed in the scatter plot and far from forming a straight line. The patterns found that summarize the underlying data do not occur by random chance.
Figure 2. Scatter plot of GenomeBits (A)denine and (T)hymine curves with histograms for the monkeypox virus (GenBank ID ON563414.3).

From these results, we believe GenomeBits may help to shed light on the surveillance behind infectious diseases. By GenomeBits, we have also uncovered distinctive signals for the intrinsic gene insights in the coronavirus genome sequences for past variants of concern: Alpha, Beta and Gamma, and the variants of interest: Epsilon and Eta [6]. A study of GenomeBits into the most recent delta and omicron variants of the coronavirus pathogen can be found in Ref. [13] (see also [14]). The GenomeBits method focuses on single-nucleotide structures and may also provide useful information to study the evolution of the MPXV outbreak via simple statistical analysis.
Figure 3. Scatter plot of GenomeBits (C)ytosine and (G)uanine curves with histograms for the monkeypox virus (GenBank ID ON563414.3).

4. Conclusions

Our main contributions here are the novel and exciting scatter plot patterns between A-T and G-C of MPXV sequences as extracted from the GenomeBits algorithm. Histograms, empirical and theoretical cumulative distribution curves, and the resulting scatter plots for pairs of base nucleotides have been presented. These statistical imprints of genome sequences by data mining can be a valuable tool to identify insightful patterns and convey cluster anomalies present within two datasets of nucleotides. Scatter plots, in general, such as those in Figures 2 and 3, offer a simple and fast form of data visualization (with little cognitive effort) of a pair of variables when plotted on a plane as a set of colored points. In the presence of high-dimensional datasets, scatter plots of all possible combinations of two variables are usually combined to build a scatter plot matrix and so extract patterns present in these datasets. The event patterns allow for visual cues and comparisons, and their structures are expected to be different along mutations of the virus based on measurements of their genome sequences.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.
**Data Availability Statement:** All custom code used in the manuscript is available at https://github.com/canessae/GenomeBits/, accessed on 24 November 2022.

**Conflicts of Interest:** The author declares no conflict of interest.

**Abbreviations**

The following abbreviations are used in this manuscript:

MPXV  Monkeypox Virus

**References**

1. WHO. Disease Outbreak News. Multi-Country Monkeypox Outbreak in Non-Endemic Countries. Available online: https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON385 (accessed on 24 November 2022).
2. GISAID. Global Science Initiative—Open Access Archive to Genomic Data of Viruses. Available online: http://www.gisaid.org (accessed on 24 November 2022).
3. GenBank. NIH Genetic Sequence Database—An Annotated Collection of All Publicly Available DNA Sequences. Available online: https://www.ncbi.nlm.nih.gov/genbank/ (accessed on 24 November 2022).
4. Gigante, C.M.; Korber, B.; Seabolt, M.H.; Wilkins, K.; Davidson, W.; Rao, A.K.; Zhao, H.; Smith, T.G.; Hughes, C.M.; Minbai, F; et al. Multiple lineages of monkeypox virus detected in the United States, 2021–2022. Science 2022, 378, 560–565. [CrossRef] [PubMed]
5. Falendysz, E.A.; Lopera, J.G.; Doty, J.B.; Nakazawa, Y.; Crill, C.; Lorenzsonn, F.; Kalemba, L.S.N.; Ronderos, M.D.; Mejia, A.; Malekani, J.M.; et al. Characterization of Monkeypox virus infection in African rope squirrels (Funisciurus sp.). PLoS Negl. Trop. Dis. 2017, 11, e0005809. [CrossRef] [PubMed]
6. Canessa, E. Uncovering signals from the coronavirus genome. Genes 2021, 12, 973. [CrossRef] [PubMed]
7. Chechetkin, V.; Turygin, A. Size-dependence of three-periodicity and long-range correlations in DNA sequences. Phys. Lett. A 1995, 199, 75. [CrossRef]
8. Voss, R.F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett. 1992, 68, 3805. [CrossRef] [PubMed]
9. Dong, R.; He, R.L.; Yau, S.S.-T. A novel approach to clustering genome sequences using inter-nucleotide covariance. Front. Genet. 2019, 10, 234. [CrossRef] [PubMed]
10. Canessa, E.; Tenze, L. GenomeBits: A Tool for the Signal Analysis of Complete Genome Sequences. Available online: https://github.com/canessae/GenomeBits/ (accessed on 24 November 2022).
11. Statistics Example Code: Histogram_demo_cumulative.py. Available online: https://matplotlib.org/2.0.2/examples/statistics/histogram_demo_cumulative.html (accessed on 24 November 2022).
12. Matplotlib. Plotting Histogram Plot just above Scatter Plot. Available online: https://stackoverflow.com/questions/37008112 (accessed on 24 November 2022).
13. Canessa, E.; Tenze, L. GenomeBits insight into omicron and delta variants of coronavirus pathogen. PLoS ONE 2022, 17, e0271039. [CrossRef] [PubMed]
14. Canessa, E. Wave-like behaviour in (0,1) binary sequences. Sci. Rep. 2022, 12, 13971. [CrossRef] [PubMed]