EMERGING MINDS
JOURNAL FOR STUDENT RESEARCH

ISSN 2976 - 9094
EMJSR 2023, (1):74-80
https://doi.org/10.59973/emjsr.22

Received: 2023-07-18
Accepted: 2023-07-20
Published: 2023-07-20

Information Theory: Applications to the Study of Mutation Dynamics

C. N. Marembo[1, 2*]

[1]University of Portsmouth, School of Mathematics and Physics, Portsmouth, P01 2HE, United Kingdom
[2]Information Physics Institute, Gosport, Hampshire, United Kingdom

*Corresponding author (Email: cnmarembo@gmail.com)

Abstract – This study investigates the mutation dynamics of viral genomes using computational methods and information theory. The analysis focuses on Influenza-A virus genomes collected from Tianjin, China, between November 2009 and February 2011. The GENetic Information Entropy Spectra (GENIES) software is employed to calculate the information entropy (IE) of viral genomes and to compare them against a reference genome. The analysis reveals frequent mutation sites, with adenine (A) exhibiting the highest mutation frequency. The study provides valuable insights into the mutation patterns and dynamics of the analysed genomes, however, limitations in data size and the capabilities of the software are acknowledged, highlighting the need for further research and larger datasets to validate and expand upon these findings. Overall, this computational approach demonstrates the potential of using information theory and GENIES to enhance our understanding of viral mutation dynamics, with implications for vaccine design and preparedness for future viral strains.

Keywords – Information Theory; Mass-energy-information; Information entropy (IE); Mutation dynamics.

1. Introduction

As global temperatures continue to rise, the likelihood of the transmission of viruses tends to increase. Warmer climates mean that vector-borne diseases are typically more infectious, as well as being transmissible earlier in their lives [1]. In recent years, the rise of the SARS-COVID-19 pandemic devastated global populations, earning the virus notoriety among the deadliest in history. It is due to the higher chances of disease that the need of the human race to manufacture vaccines and treatments has become considerably greater. Vaccines, however, are not easy to engineer, are of varied effectiveness, and are also most effective on the particular strains they were designed to disable. A mutation is a change within a DNA or RNA sequence, sometimes provoking a change in an organism's reproductive capabilities, deadliness, and resilience [2-5]. The Influenza virus is an example of how mutations can allow the virus to continually infect populations seasonally, while also evading current vaccines and treatments [6]. Henceforth, understanding viruses and their mutation dynamics is of increasing significance. By investigating the mutations of genomes, in addition to the corresponding physical attributes or phenotypes, it may be possible to engineer more effective vaccines for better preparation for future strains that could otherwise devastate the world, exceeding previous plagues and pandemics.

Information theory is the mathematical study of how information can be encoded, stored and transmitted. Information Theory was introduced by the works of R. Hartley and H. Nyquist [7], and C. Shannon [8] in the 1920s and 1940s, respectively. A particular measure that will be of use within this investigation is information entropy; Information entropy measures the average amount of information associated with an event relative to all possible outcomes, providing a normalised understanding of the information content. The significance of information entropy lies within its parallel nature with entropy in thermodynamics, whereby Gibbs entropy and Shannon entropy formulae are near-identical, supposedly reinforcing Shannon information theory as a reliable mode of analysis.

In the case of this project, the information of concern is genetic data, specifically, nucleotide bases within an organism's genome. By encoding the genetic information, it is supposed, that it could be possible to assess the differences between strains of viruses mathematically. Thus, the aim of predicting future mutations by statistical means becomes more feasible and holds high potential in this regard. Note, this application of information theory to genetics is not the first, hence this work serves the purpose of expanding upon and supporting past works, aiming to build more understanding of the field, its applications and limitations at present [9]. Past studies include Reichert et al. [10] who developed a method rooted in information theory to assess the alignment quality of code sequences in 1973, while Veluchamy et al. [11] explored genome clusters in evolutionary biology as recently as 2021. These works serve as a foundation for our research, aiming to further enhance our understanding of the field [12].

In this study, we employ the use of GENetic Information Entropy Spectra (GENIES) and leverage the methods proposed by M. Vopson [13] to investigate correlations and patterns within viral genomes. By uncovering these insights, we aim to provide valuable observations that can be further explored to gain a deeper understanding of mutation dynamics and their implications for viral evolution.

Overall, this report aims to demonstrate the significance of information theory in studying mutation dynamics in viral genomes, and how it can contribute to the development of more effective strategies in vaccine design and preparedness for future viral strains.

2. Theory

The proposed method of analysing the genetic information of genomes requires the understanding and application of Information Theory. In this case, the information concerned, is the probability, p, of an event occurring. Shannon postulates an axiomatic approach for calculating information as a function of probability, which is ideal as the probability is limited to real numbers between 0 and 1.

For a measure of information, $I(P)$, we define four fundamental axioms:

1. That Information, as a quantity, is non-negative:

$$I(P) \geq 0 \tag{1}$$

2. If an event has a probability, 1, of occurring, no information is observed:

$$I(1) = 0 \tag{2}$$

3. If two independent events occur (whose joint probability is the product of their individual probabilities), then the information from observing the events is the sum of the two:

$$I(p_1 \times p_2) = I(p_1) + I(p_2) \tag{3}$$

4. Information should be a continuous and essentially monotonic function of probability, meaning slight changes in probability should result in proportional changes in information.

Therefore, the following is derived:

1.

$$I(p^2) = I(p \times p) = I(p) + I(p) = 2I(p) \tag{4}$$

2. Thus,

$$I(p^n) = nI(p) \tag{5}$$
(By induction)

From these various identities, among others, information can be described as below:

$$I(p) = -log_b(p) = log_b\left(\frac{1}{p}\right) \tag{6}$$

For some positive arbitrary base, b, which determines the units of the measure of information. For example, base-2 i.e. $\log_2(\mathrm{p})$ gives units of bits, as in binary; Base-e i.e. $\log_{\mathrm{e}}(\mathrm{p}) = \ln(\mathrm{p})$ are nats (for natural logarithm); Base-10 are Hartleys, after R. Hartley. Having defined information, information entropy can be introduced. For a number, N, of independent occurring events, a, the associated events and probabilities can be denoted as $[a_1, a_2, a_3, \ldots, a_i]$, $[p_1, p_2, p_3, \ldots, p_i]$, respectively.

Therefore, a string of N, number of events gives information:

$$I = \sum_{i=1}^{n} (N \times p_i) \times log\left(\frac{1}{p_i}\right) \tag{7}$$

Therefore, the average information per symbol would require dividing both sides by N, such that:

$$\frac{I}{N} = \frac{1}{N}\sum_{i=1}^{n} (N \times p_i) \times log\left(\frac{1}{p_i}\right) = \sum_{i=1}^{n} (p_i \times log\left(\frac{1}{p_i}\right)) \tag{8}$$

For a probability distribution, $P = [p_1, p_2, p_3, \ldots, p_i]$, the entropy, H, is given by:

$$H(P) = \sum_{i=1}^{n} (p_i \times log\left(\frac{1}{p_i}\right)) \tag{9}$$

Or for a continuous distribution:

$$H(P) = \int P(x) \times log\left(\frac{1}{p(x)}\right) dx \tag{10}$$

While Eq. (7) shows how the encoding of information can be done, the physical significance or reliability of the equations may at first, seem arbitrary or incomplete. Information has been postulated as a further component of the proven mass-energy equivalence principle, therefore even referred to as the Mass-Energy-Information principle [8].

The cohesion of information, entropy and genomics may seem unlikely but may prove a robust and powerful tool for better understanding organisms from the minutia, which could provide insight into the significance of such components and their ability to characterise the organism with varying distinctions. In the context of this project, the information which is measured is the presence of specific nucleotide bases in the DNA sequences of viruses. The nucleotide bases concerned are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). By encoding the genome, one can gain a measure for the information and therefore information entropy. Using the information entropy, one can then calculate the ratio of entropies of various genomes to their mutated counterparts, where the ratio will be 1 where the IE is equal, i.e. there is no difference between the information entropies; While a ratio not equal to 1 implies a mutation.

$$Information\ Entropy\ Ratio\ (IER) = \frac{H(P)_{Reference}}{H(P)_{Mutated}} \tag{11}$$

Equation (11) gives the most important measure used during the course of this project, as it is most conclusive and describes both systems simultaneously, whilst also eliminating the use of units. GENIES outputs this in the form of spectra which are the main findings of interest as we can visually compare genomes.

3. Methodology

The methodology employed in this study is entirely computational, utilising specialised software called GENetic Information Entropy Spectra (GENIES). The computational approach is necessary due to the large data sets and complex calculations involved in analysing genomic information. The primary focus of the methodology is to calculate information entropies and compare those between different genomes. The GENIES software is used to perform these calculations, providing information entropy spectra and ratios. The spectra visually represent the differences in information entropy between genomes, allowing for the identification of common areas of mutation.

The genomes analysed in this project are variants of Influenza-A, specifically collected from Tianjin, China, between November 2009 and February 2011 [6]. The genomes were obtained from the National Centre for Biotechnology Information (NCBI) and selected based on specific criteria. These criteria include restricting the samples to those from human hosts, fully sequenced genomes, and samples collected over a period of more than a year by a single submitter.

To analyse the genomes using GENIES, the first variant of Influenza A Virus (Accession ID: JQ714202.1) is treated as the reference genome for subsequent variants. Note, the parameters for GENIES were M-Block Size: 3, M-Block Step Size: 1, Window Size: 33 and Window Step Size: 1. These parameters carry the significance of determining how many of the characters within the genome are properly read and collected. The above parameters were chosen as they are similar to the defaults of GENIES upon booting, M. Vopson [13] also reports ideal parameters with which the chosen ones comply.

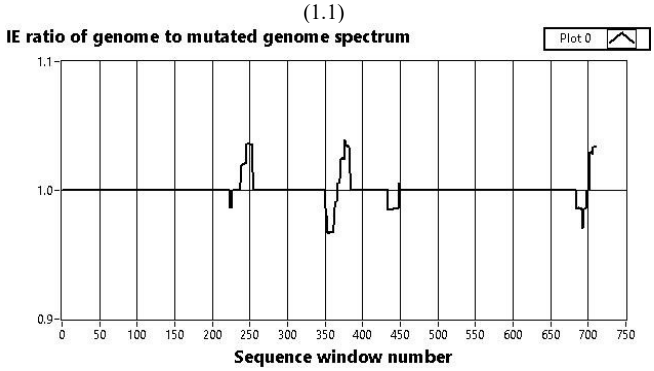
The GENIES software handles all the necessary calculations and data processing for this project. Users only need to input the formatted genomes and collect the generated data, minimising the likelihood of errors. For each combination of the reference sequence and variant, the information entropy spectra and ratios are collected and compared. The comparisons can be performed side by side or by superimposing the ratios and spectra, enabling simultaneous observation of the differences between all variants. GENIES also provides detailed information on the index positions of mutations and the specific mutations that occur, allowing for a more in-depth analysis of recurring patterns. Furthermore, the frequency of each base change undergone during mutations is plotted, providing additional insights into the mutation dynamics.

The methodology ensures the comprehensive analysis of genomic data using the GENIES software, allowing for the identification of mutation patterns and tendencies within the analysed genomes. The utilisation of computational methods minimises errors and provides efficient handling of large datasets, leading to reliable and meaningful results.

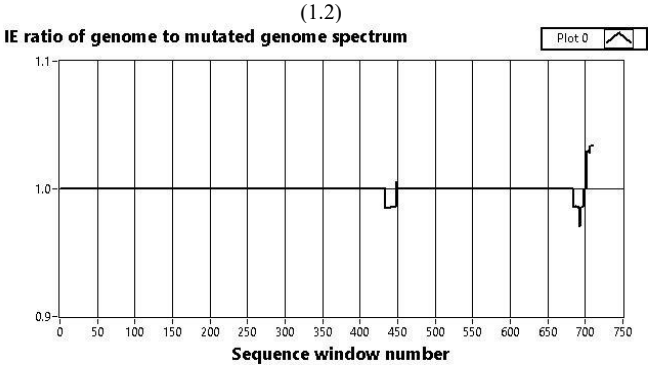
4. Results

The analysis of the obtained results reveals interesting findings within the specific set of samples analysed. The spectra generated from the analysis highlight certain sites where mutations appear to occur more frequently than others.

(1.1)



(1.2)



(1.3)

(1.4)

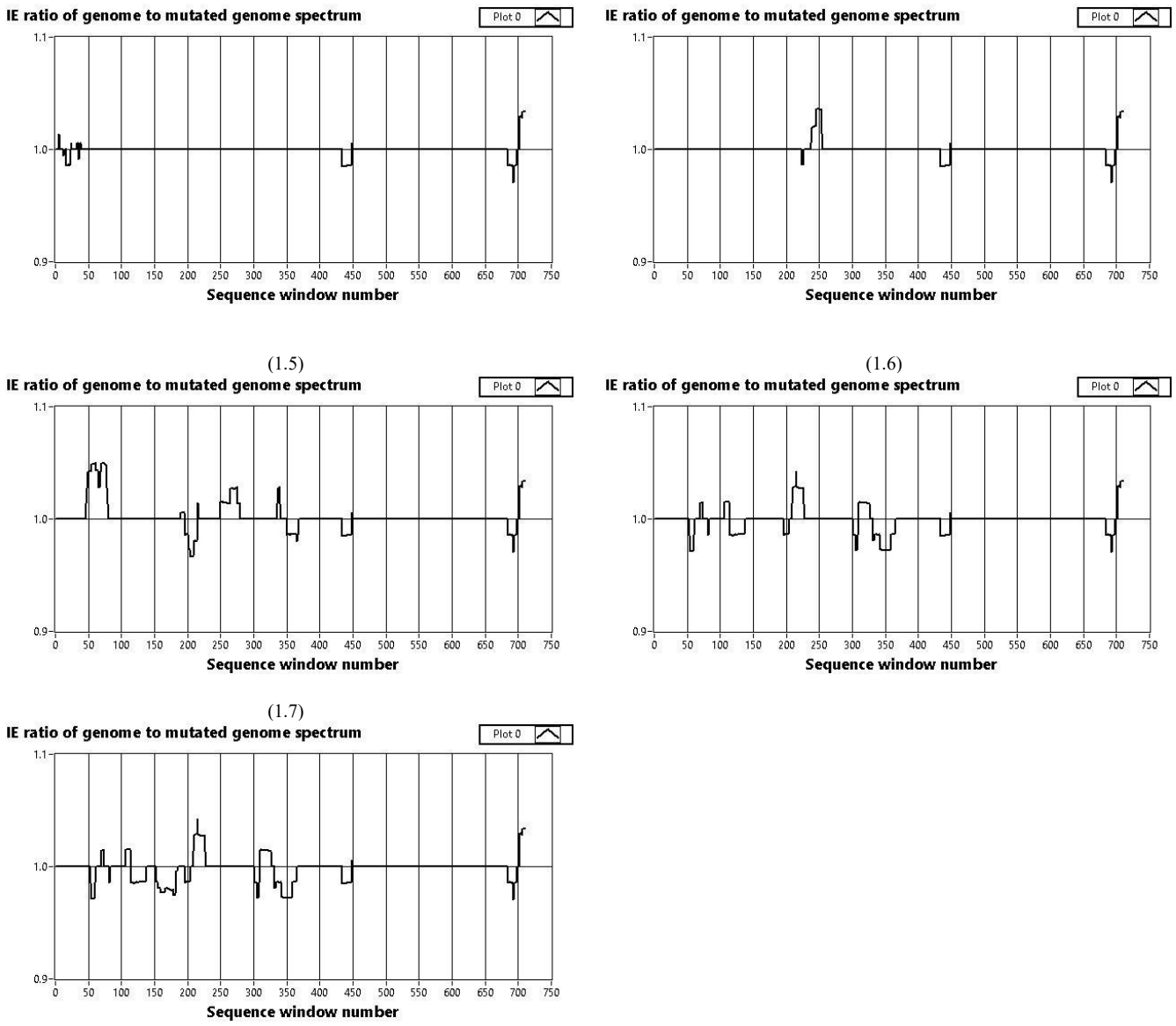


Figure 1. The Information Entropy Ratio (IER) spectra of Influenza A Viruses, with Accession ID: JQ714202.1 as the reference genome against Accession IDs: **(1.1)** JQ714203.1; **(1.2)** JQ714204.1; **(1.3)** JQ714205.1; **(1.4)** JQ714206.1, **(1.5)** JQ714207.1, **(1.6)** JQ714208.1, **(1.7)** JQ714209.1. Note, the length of the genomes used was 730 characters.

Among the frequent peaks observed in the spectra, three notable sequence window numbers stand out: 448, 712, and 716 (Fig. 1). These indices correspond to locations where mutations are most commonly observed. To further investigate the distribution of mutations, the frequencies of these mutation indices were plotted in a histogram. This histogram provides a spectrum of the modal frequencies across all selected variants. Additionally, GENIES provides detailed information on the specific mutations that occur at each index. Upon examining this information, it was observed that the nucleotide base adenine (A) exhibits the highest frequency of mutation, accounting for approximately 52.5% of all observed changes. The second most frequent mutation involves cytosine (C) changing to thymine (T), accounting for 22.5% of the mutations. Other notable mutations include guanine (G) changing to adenine (A) at a frequency of 15%, and thymine (T) changing to cytosine (C) at a frequency of 10%. It is important to note that the total sample size of mutations across all seven spectra was relatively small, with 40 observed mutations in total. This highlights the need for further investigation to reduce uncertainty and obtain a more comprehensive understanding of the strains.

Despite the limited sample size, the observed distinctions in mutation frequencies provide valuable insights into the mutation dynamics. These findings represent a positive step forward in the analysis of mutation mechanisms. Nevertheless, further research and extensive trials are required to solidify these initial observations.

The similarities observed between the various genomes analysed were expected, as the samples were submitted to the National Centre for Biotechnology Information (NCBI) Virus by the same organisation. The close

genetic relationship between the genomes, as indicated by the mutation patterns, suggests that these samples may have been exposed to similar conditions. This information is crucial for improving our understanding of mutation dynamics and mechanisms. Building upon these findings, future investigations could explore genomes from a wider geographical area, considering both the differences and similarities in mutation sites. Such studies could shed light on the role of environmental factors, including geographical regions and collection dates, in shaping the dynamics of the virus. The overall methodology employed in this study is repeatable and holds the potential for yielding more promising results, given extensive trials and further research.

5. Conclusion

The project has identified several limitations that should be considered when interpreting the results. One of the major limitations is the capability of the GENIES software and the challenges associated with handling large amounts of data. GENIES is effective for comparing genomes of the same length, but it faces difficulties when mutations involve changes in the length of genomes through insertions or deletions. As a result, not all mutations within a strain can be fully analysed using the current analysis approach. To address this limitation, potential solutions include implementing sorting algorithms to detect common areas between genomes of different lengths and recording areas of differences as potential sites of insertion, deletion, transition, or transversion. Moreover, modifying the source code of GENIES to allow the analysis of multiple genomes simultaneously could enhance its effectiveness in identifying areas of interest more conclusively by plotting mutation frequencies.

Furthermore, it is important to note that the certainty of the observations made in this project is limited due to the relatively small sample size of eight strains. Further research and data collection using larger datasets of genomes are necessary to validate and expand upon these findings. By including a broader variety of genomes, a more comprehensive analysis can provide a more precise understanding of the mutation dynamics.

Among the findings, this project highlights that certain indices of mutation are more common within Influenza-A genomes. Specifically, the most frequent nucleotide base change observed is adenine (A) to guanine (G), accounting for over half of the mutations within the selected samples. This information contributes to our understanding of the common behaviours within genomes and emphasises the potential of the GENIES software to analyse genomes and provide detailed information, such as the frequency of specific mutations.

In summary, while this project has identified limitations and uncertainties, it demonstrates the potential of using computational methods and the GENIES software to analyse and gain insights into the mutation dynamics of viral genomes. Further research with larger datasets and improvements in analysis techniques can lead to more comprehensive and reliable conclusions in the future.

Acknowledgments

Melvin M. Vopson, my academic supervisor, who inspired and supported this project.
My Mother, Brother and Sisters, for their continued support and encouragement.

References

[1] P. M. Polgreen and E. L. Polgreen, Infectious Diseases, Weather, and Climate, Clinical Infectious Diseases, Volume 66, Issue 6, (2018), https://doi.org/10.1093/cid/cix1105
[2] L. Herrero, E. Madzokere, What's the difference between mutations, variants and strains?, (2021), https://www1.racgp.org.au/newsgp/clinical/what-s-the-difference-between-mutations-variants-a
[3] M. Sobhanie, How do virus mutations happen, and what do they mean?, (2021), https://wexnermedical.osu.edu/blog/virus-mutations-what-do-they-mean
[4] R. Sanjuán, P. Domingo-Calap, Mechanisms of viral mutation. Cell. Mol. Life Sci. 73, (2016), https://doi.org/10.1007/s00018-016-2299-6
[5] Cann, A. J. (2012). Genomes. Principles of Molecular Virology (Fifth Edition), Academic Press, (2016), https://doi.org/10.1016/B978-0-12-801946-7.00003-1
[6] X. Li, M. Kong and J. Chen, et al., Epidemiology and full genome sequence analysis of H1N1pdm09 from Northeast China, (2013), https://doi.org/10.1007/s11262-013-0931-1
[7] R. V. L. Hartley, Bell System Technical Journal, 7: 3, Transmission of Information (1928), https://doi.org/10.1002/j.1538-7305.1928.tb01236.x
[8] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948), https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
[9] J. A. Tenreiro Machado, "Shannon Entropy Analysis of the Genome Code", Mathematical Problems in Engineering, 2012, https://doi.org/10.1155/2012/132625
[10] T. A. Reichert, D. N. Cohen and A. K. C. Wong, An application of information theory to genetic mutations and the matching of polypeptide sequences. Journal of Theoretical Biology. 42(2), (1973), https://doi.org/10.1016/0022-5193(73)90088-X

[11] A. Veluchamy, et al., Information theoretic perspective on genome clustering. Saudi Journal of Biological Sciences. 28(3), (2021), https://doi.org/10.1016/j.sjbs.2020.12.039
[12] C. Adami, The use of information theory in evolutionary biology. Annals of the New York Academy of Sciences, (2012), https://doi.org/10.1111/j.1749-6632.2011.06422.x
[13] M. M. Vopson and S. C. Robson, A new method to study genome mutations using the information entropy. Physica A: Statistical, (2021), https://doi.org/10.1016/j.physa.2021.126383
[14] T. M. Nieuwenhuizen, The mass-energy-information equivalence principle. AIP Advances. 9(9), 095206, (2019), https://doi.org/10.1063/1.5123794