Fourier spectral density of the coronavirus genome

H.S.Tan

Physics Division, National Center for Theoretical Sciences,
Hsinchu 30013, Taiwan.

haisiong.hstan@gmail.com

Abstract

We present an analysis of the coronavirus RNA genome via a study of its Fourier spectral density based on a binary representation of the nucleotide sequence. We find that at low frequencies, the power spectrum presents a small and distinct departure from the behavior expected from an uncorrelated sequence. We provide a couple of simple models to characterize such deviations. Away from a small low-frequency domain, the spectrum presents largely stochastic fluctuations about fixed values which vary inversely with the genome size generally. It exhibits no other peaks apart from those associated with triplet codon usage. We uncover an interesting, new scaling law for the coronavirus genome: the complexity of the genome scales linearly with the power-law exponent that characterizes the enveloping curve of the low-frequency domain of the spectral density.

Contents

1 Introduction

Motivated by our search for deeper organizational principles governing genetic information [1], the study of a DNA/RNA genome via its Fourier spectral density has given us several interesting insights into the code of life. An example of a seminal paper in this subject is that of Voss in [2] where the author found that the spectral density of the genome of many different species follows a power law of the form $1/k^\beta$ in the low-frequency domain, with the exponent β potentially related to the organism's evolutionary category. In [2], β was found to be close to 1, a phenomenon shared by a wide variety of physical systems especially those that carry long-range correlations or characterized by a myriad of length scales. It was also found that the power spectra may contain defining peaks or resonances, for example at period 9 for primates, vertebrates and invertebrates, or period 10-11 for yeast, bacteria and archaea as shown in [3] where the peaks were remarkably related to aspects of protein structuring and folding. Over the years, these methods and results have been extended in various ways [4], such as wavelet-type analysis [5, 6, 7] of the sequences, using features of the spectra to classify and cluster genomes with the aid of neural networks [8], prediction of coding regions [9] and periodic structures [10], etc.

In this paper, we study the Fourier spectral density of the genome of coronaviruses — a positive-sense single-stranded RNA genome with size ranging from roughly 26 to 32 kilobases, based on the dataset of [11] which covers all four genera of coronaviruses. In addition, motivated by the recent COVID-19 pandemic, we include the genomes of SARS-CoV-2, a bat coronavirus Bat-RaTG13 of close genome identity, and the MERS coronavirus.

Across the 30 different genome sequences, we find that their Fourier spectra take on the same form. There is a low frequency domain ($k \lesssim 10$ in units of inverse genome length) where a sinc-squared-like oscillatory form is enveloped by a roughly $1/k^2$ decay curve. This is followed by stochastic white-noise type fluctuations about fixed mean values which tend to vary inversely with the genome size. We find that a random, uncorrelated sequence — with the probability of

[1] See for example [1] for an inspiring read.

occurrence for each nucleotide being its frequency ratio in the sequence — yields similar behavior in the low-frequency domain. We develop a few models to characterize the typical spectrum, and in the process stumble upon a linear scaling law between a measure of the complexity of each genome and the power-law exponent that describes the enveloping curve of the low-frequency domain. The complexity measure that we use here is intimately related to the Shannon entropy of the sequence, and thus this relation concretely realizes a way by which information-theoretic content is carried within the genome's spectral density.

Now, power-law decay of the form $1/k^\alpha$ have previously been discussed in literature for other types of genomes (see for example [12, 13]). We would like to emphasize that here, we do not employ either the Fast Fourier Transform or non-overlapping averaging procedures to smoothen the data in the low-frequency domain. These are common techniques used for easing computations in past related works, but may compromise the sensitivity by which we characterize the spectral curves. We also perform the spectral density analysis at the level of the coding region (a few thousand nucleotides) for the Spike protein, an essential protein that binds to the host cell's receptor. We find that interestingly, the general features of the spectrum persist at the protein level, but not the scaling law mentioned above.

Our paper is organized as follows. In Section 2, we present some background theory for our work, followed by Section 3 where we present the results and a few graphical plots for visualization, before concluding in Section 4. The Appendix A collects a table listing all the GenBank accession numbers [14] of the genomes, and another gathers several graphs useful for interpreting our various results.

2 Theoretical preliminaries

In this Section, we present some essential mathematical concepts that form the basis for our study. Our analysis of the RNA genomes can only begin after transformation of the genome sequence consisting of the four nucleotides (Adenine, Cytosine, Guanine and Uracil) into a numerical string. The spectral density of interest here is the absolute square of the discrete Fourier transform of a nucleotide indicator function $\phi(i)$ defined as follows

$$S_{\alpha\beta}(k) = \frac{1}{M^2} \sum_{l=1}^{M} \sum_{j=1}^{M} \phi_\alpha(l) \phi_\beta(j) e^{\frac{2\pi i k}{M}(l-j)}, \tag{1}$$

where M denotes the length of the genome, and α, β denote particular choices of nucleotides. In the continuum limit and after averaging over some distribution of genomes, this approaches the Fourier transform of the correlation function $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{2\pi i k(x_1-x_2)} \langle \phi_\alpha(x_1)\phi_\alpha(x_2)\rangle dx_1 dx_2$.

Now, a basic premise lies in the choice of the indicator function $\phi(i)$. While various propositions have been explored in the literature, in this paper, following [2], we use a simple binary-valued model where for each nucleotide, $\phi(i)$ is equal to 1 if the nucleotide is found at position 'i' and 0 otherwise. For all our genome data, we find that (1) exhibits a clear specific oscillatory form that resembles a sinc(-squared) function in the low-frequency domain (up to $k \sim 10$). In the following, we furnish a potential simple explanation of such low-frequency behavior. For simplicity and definiteness, we will mainly focus on the spectral density sum

$$S(k) = \sum_\alpha S_{\alpha\alpha}(k) = S_{AA}(k) + S_{CC}(k) + S_{GG}(k) + S_{UU}(k), \tag{2}$$

for the rest of the paper, but we have checked that the general features described above pertain to the cross-spectra $S_{\alpha\beta}$ (with $\beta \neq \alpha$) as well as the individual autocorrelations $S_{\alpha\alpha}$ for all the four nucleotides.

Apart from computing various quantities at the level of the entire RNA genome, we also examine the spectral density associated with the coding region for the Spike protein. For the coronaviruses, apart from the Spike protein, the genome encodes several proteins each carrying unique functions, such as the envelope, membrane, nucleocapsid, etc. In particular, the Spike protein plays an essential role in host cell receptor binding during the process of viral infection, and is thus a common target for developments of antibodies and vaccines (see for example, [11]). Now the coding region associated with this protein is only of the order of 10^3 nucleotides, so a priori it is not clear if the spectral density can be meaningfully analyzed. We find however that the general features of the spectral density persist for the Spike protein's coding region too.

2.1 A reference curve: the uncorrelated background

Consider the case of an uncorrelated numerical sequence, where the probablity of a nucleotide of type α occurring at some position is a constant, independent of others and the position itself. Given N_α such nucleotides in the sequence, we can estimate this constant to be $\frac{N_\alpha}{M}$, with the expectation value of the spectral density being

$$S_{\alpha\alpha}^{(uncorrelated)}(k) \sim \frac{1}{M^2}\sum_{l=1}^{M}\sum_{j=1}^{M}\frac{N_\alpha^2}{M^2}e^{\frac{2\pi ik}{M}(l-j)} = \frac{N_\alpha^2}{M^4}\frac{\sin^2(\pi k)}{\sin^2\frac{\pi k}{M}}. \tag{3}$$

We would find that up to $k \sim 10$, (3) models the spectral density rather well. For the local maxima of (3), they approximately occur at half-integer values of k and thus the upper envelope of the oscillation is manifest as a $1/k^2$ decay function in this domain which follows from expanding (3) about $k = 0$. The decaying behavior of the envelope curve typically stops at about $k \sim 30$, and thereafter the spectral density appears to be characterized by stochastic fluctuations about some fixed mean.

Although (3) appears to model observed datasets well, the goodness of fit doesn't extend beyond the $k \sim 10$ range, nor is it clear from the data whether deviations from (3) are unimportant random fluctuations or otherwise within the low-frequency domain. To gain further insights, we present a few simple models which characterize the observed deviations from (3). The models' parameters can potentially be used for clustering coronavirus genomes if future studies prove that these values persist for a larger sets of data, or more interestingly, they could potentially demonstrate correlation with other features of the genome that would help us recognize the presence of long-range correlations. From now on, we refer to (3) as the 'uncorrelated background'.

2.2 Three simple models

In the following, we present three models for the observed spectral density that characterize deviations from the uncorrelated background. The first two concerns the description of the low-frequency domain ($k \lesssim 10$) whereas the third involves a more global description.

(A) Power-law decay of the enveloping curve

Motivated by previous works on this subject, we consider fitting a power-law decay via least-square regression to the enveloping curve (for $k \in [1, 10]$) of the form

$$S(k) \sim \frac{1}{k^{\delta}}, \tag{4}$$

for some power exponent δ. The power-law description is convenient and has proven to be a popularly studied model for spectral density of genomes in general (see for example [13]). It is crucial to bear in mind that it is a coarse-grained description which doesn't extend to the origin, and valid only for the low-frequency domain. We would later find that this is the parameter that remarkably scales linearly with a measure of the genome complexity. For all our datasets, $\epsilon = \delta - 2 \sim 10^{-2}$. It is not a priori clear how large ϵ has to be in order for the deviation to be significant, and more sequences corresponding to each type of coronavirus should be studied in order to determine the range of ϵ and its statistical distribution. Although we leave this for future work, we found evidence that the variation in the ϵ correlates with a measure of the complexity of the genome (which at the limit of infinite genome size approaches the Shannon entropy) in a way that is distinctly different from a completely random sequence.

It is useful to compute the expected δ for the hypothetical uncorrelated background (3) which is parametrized by the genome size M and the sum of squares of nucleotide number $\sum_{\alpha} N_{\alpha}^2$. For the general spectral density $S(k)$, from least-square regression of the log-log relation, we obtain $\epsilon \equiv 2 - \delta$ to be

$$\epsilon[S] = \frac{\langle \left(\log(k) \log(S(k)k^2)\right) \rangle - \langle \log(k) \rangle \langle \log(S(k)k^2) \rangle}{[\langle (\log(k))^2 \rangle - \langle \log(k) \rangle^2]}, \tag{5}$$

where $\langle \ldots \rangle = \frac{1}{9} \sum_k (\ldots)$ denotes averaging over the nine local maxima points in the domain $k \in [1, 10]$. For the uncorrelated case of (3), we find that the factor $\sum_{\alpha} N_{\alpha}^2$ cancels away in (5) and numerically, $\delta \approx 1.956$ for all the datasets at the level of the genome and that of the protein coding region. This defines a background value for the detection of a deviation away from the completely random sequence.

(B) Linearized correlation function

In contrast to an empirical power-law fitting of only the enveloping curve, one could adopt a bottom-up approach by postulating certain forms of the correlation function, and then performing the discrete Fourier transform. Consider the case where the correlation function is a linear function of the nucleotide separation, we can write

$$S(k) = \frac{R_0}{M^2} \sum_{j=1}^{M} \sum_{l=1}^{M} \left(1 - \kappa \frac{|l-j|}{M} \right) e^{\frac{2\pi i k (l-j)}{M}} \tag{6}$$

for some constant κ, and $R_0 = \sum_{\alpha} \frac{N_{\alpha}^2}{M^2}$. A straightforward calculation yields

$$S(k) = \frac{R_0}{M^2} \left(\frac{\sin^2(\pi k)}{\sin^2 \frac{\pi k}{M}} + \kappa \frac{\cos(\pi k)}{2M \sin^3 \frac{k\pi}{M}} \left[(M-1) \sin \frac{k\pi(M+1)}{M} + (M+1) \sin \frac{k\pi(1-M)}{M} \right] \right). \tag{7}$$

This function is invariant under the reflection $k \leftrightarrow M - k$, which is an exact discrete symmetry for the spectral density $S(k)$ (or the individual $S_{\alpha\alpha}(k)$) more generally. The parameter κ admits the physical interpretation of the presence of long-range correlation/anti-correlation depending on whether it's positive/negative, and we would find that apart from one exception, all our datasets can be matched to a positive κ of the order 10^{-2}. We find that if the curve-fitting is performed taking into account only the first ten local maxima as in the case for δ-parameter, the local minima points at integral k-values are not well captured by the fitted curve, so we also include them in the curve-fitting.

Beyond the specific linear form of the correlation function postulated in (6), it is also representative of a large class of correlation functions of the form

$$R_0 F\left(\tilde{\kappa}\frac{|\tau|}{M}\right) \approx R_0\left(1 + \partial_{\tilde{\tau}} F(\tilde{\tau})|_{\tilde{\tau}=0}\tilde{\kappa}\frac{|\tau|}{M} + \mathcal{O}(\tilde{\tau}^2)\right),$$

where $\tau \equiv l - j$, $\tilde{\kappa}$ is a small constant and $\tilde{\tau} = \tilde{\kappa}\frac{|\tau|}{M}$. This first-order truncation is identical to (6) with $F'(0)\tilde{\kappa} = \kappa$. Thus, (6) could approximate correlation functions of the general form $F\left(\tilde{\kappa}\frac{|\tau|}{M}\right)$ where $\tilde{\kappa}$ is a small dimensionless parameter, and

$$F(0) = 1, \qquad |F'(0)|\tilde{\kappa} \ll 1, \qquad F'(0)\tilde{\kappa} < 0.$$

For example, if the correlation function turns out to be an exponentially decaying function of the form $e^{-\frac{B|\tau|}{M}}$ with $B \ll 1$, then to a good approximation we can identify $\kappa \sim B$.

(C) **A Lorentzian function**

The power-law decay in (A) parametrizes the decay of the envelope whereas the model in (6) could account for non-vanishing local minima in the low-frequency domain. Beyond this region, we seek an interpolating curve that extends throughout the spectrum including the origin. For this purpose, we consider fitting a Lorentzian function of the following form to the spectrum

$$L(k) = N\frac{b^2}{k^2 + b^2} + \overline{m}, \tag{8}$$

where $N = \frac{\sum_\alpha N_\alpha^2}{M^2} - \overline{m}$ and $\overline{m}$ is the mean value near the spectrum's midpoint, about which stochastic fluctuations are observed.[2] This is a simple coarse-grained model which averages over the oscillations in the low-frequency domain and describes the overall decay of the spectrum via a smooth curve. Like the κ parameter in the model (6), the curve-fitting is performed with the set of extremal points in the low-frequency domain, with the initial and final conditions taken into account by first fixing $N, \overline{m}$ with their observed values for each genome sequence. As a useful reference, we also fit the Lorentzian function to the uncorrelated background (3) and finding $b^2 \approx 0.0765$ with $\overline{m} \sim 10^{-10}$ at the genome level, and $\overline{m} \sim 10^{-8}$ at the protein coding region level.

2.3 A measure of complexity and Shannon entropy

Scaling laws manifest in the Fourier spectral density have often motivated the study of features of the genome that reflect various properties of it being a complex system, such as the fractal dimension (of a suitably defined matrix representation of the correlation function), etc. A measure of the complexity of the genome considered in the past literature (see for example [15, 16, 17]) is defined as follows .

$$\Omega = \frac{1}{M}\log\left(\frac{M!}{N_A!N_C!N_G!N_U!}\right), \tag{9}$$

[2] We only consider half of the spectrum, since the other half naturally arises from the discrete symmetry $S(k) = S(M-k)$.

where N_α is the number of the α-nucleotide. The logarithmic argument counts the number of distinguishable permutations given a fixed number of each nucleotide. At large M, this admits a natural interpretation of the Shannon entropy of the genome sequence. To see this, we can invoke Stirling's formula to express the large-M limit of Ω as

$$\lim_{M\to\infty} \Omega = - \sum_{\alpha\in\{A,C,G,U\}} \left(\frac{N_\alpha}{M}\right) \log\left(\frac{N_\alpha}{M}\right), \tag{10}$$

which is a function of only the fractional distribution of nucleotides. In this form (10), the measure of complexity Ω is clearly the Shannon entropy which measures the information entropy associated with a genome sequence where the probability of nucleotide-α occurring in any position is $\frac{N_\alpha}{M}$. We would find later that interestingly, the model parameter δ (but not κ) scales linearly with Ω across the dataset of 30 types of coronavirus genomes. Also, when restricted to the Spike protein's level, the measure of complexity appears to scale linearly with the overall measure at the genome level. But the model parameter δ that is computed at the level of the Spike protein does not correlate with Ω at either the genome/protein level, and neither does κ.

3 Results and graphs

Our genome dataset[3] consisting of 30 types of coronaviruses spread across four genera mainly follows from reference [11] plus a few other additions : SARS-CoV-2, MERS-CoV and Bat-RaTG13. Bat-RaTG13 is a bat coronavirus that was most recently found to have 96% genome identity with SARS-CoV-2 and featured in papers discussing a possible bat origin of the latter [18]. We included it here to see how the model parameters for this genome compare to that of SARS-CoV-2 relative to the other coronaviruses. In the following, we outline the essential results, using the example of the SARS-CoV-2 reference genome for various graphical illustrations.

We find that the Fourier spectral density is characterized by the following features:

(a) In a small low-frequency regime ($k \lesssim 10$), the uncorrelated background (3) is a good approximation (see Fig. 1) for all genome sequences we examined. After curve-fitting to the datasets, we find the following range of values for the model parameters:[4]

$$\begin{aligned} &\delta_w \in [1.935, 2.079], \quad \kappa_w \in [-0.0056, 0.0583], \\ &\delta_s \in [1.744, 1.934], \quad \kappa_s \in [0.0155, 0.0818]. \end{aligned} \tag{11}$$

From visual inspection of the relevant graphical plots, we find no obvious correlation among these model parameters, nor between them and the genome/Spike protein sizes. But we find that δ_w and Ω_w appear to be related. Linear regression yields the following best-fit line (see Fig. 2)

$$\Omega_w \sim \alpha + \beta\delta_w, \tag{12}$$

with the line parameters being (with the 95% confidence intervals in brackets)

$$\alpha \approx 2.28(2.14, 2.41), \qquad \beta \approx -0.47(-0.53, -0.40), \tag{13}$$

[3] We list their names and GenBank accession IDs of the genomes in the Appendix for reference.

[4] Whenever appropriate, we use the subscripts to label the level (w = whole genome, s = Spike protein) at which various quantities are computed.

Since we checked that for all the 30 coronaviruses, the assumption of a completely uncorrelated background yields $\delta \approx 1.956$, this leads to a convenient definition of a reference complexity value

$$\Omega^u \approx 1.361,$$

which lies at the intersection between the uncorrelated vertical line and the observed one with finite slope. The difference between the observed complexity measure and Ω^u in turn enacts a measure of the deviation from complete randomness of the sequence.

There is also a similar relation between δ_w and Ω_s, consistent with the following linear relation that we found:

$$\Omega_w = \mathcal{C}\Omega_s, \;\; \mathcal{C} \approx 1.004(1.003, 1.006).$$

It would be interesting to study this for other coding and non-coding regions as it is suggestive of some level of self-similarity for this complexity measure.

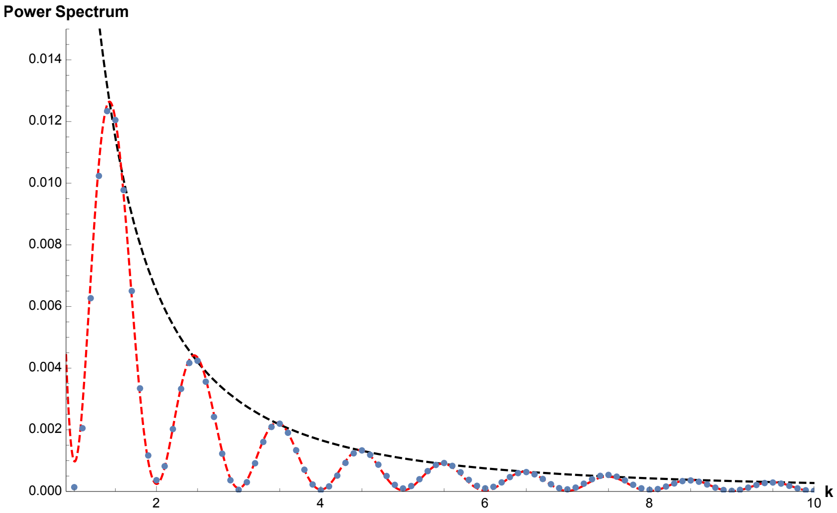


Figure 1: Plot of the power spectral density (SARS-CoV-2 genome) in the range where we perform the curve-fitting. The black dashed line is the curve $\sim 1/k^\delta, \delta = 1.968$ obtained from the set of local maxima, while the red dashed line is equation (7) obtained by fitting to all maxima and minima, with the best-fit value $\kappa = 0.0362$.

(b) After $k \sim 10$, the genome displays much more scatter about the uncorrelated background, and the models of deviation are no longer effective descriptions (see Fig. 3). Stochastic fluctuations about a fixed mean appear to set in and there are no isolated peaks apart from two prominent ones at $k \sim \frac{M}{3}, \frac{2M}{3}$ which have been seen and interpreted in past literature [2, 19] to correspond to the universal triplet codon usage. We applied an (overlapping) moving average (of window size ~ 100 nucleotides) to smooth out the data, and checked that there is no apparent regime where some non-trivial scaling law holds (see Fig. 4 and 5).

At the level of both the genome and protein coding region, the fixed mean parameter $\overline{m}$ appears to correlate with the genome size. It appears to generally decrease with the size of the sequence,at both levels of the genome and the Spike protein (see Figures 6a and 6b) in Appendix B). At the genome level, it is of the order $\sim 10^{-5}$ which is about 10^5 larger than the value expected for the uncorrelated background, whereas at the spike protein level, $\overline{m} \sim 10^{-4}$ which is 10^4 times larger than the uncorrelated background. The Lorentzian function that is fitted to the data with initial and final conditions fixed by R_0 and $\overline{m}$ is parametrized by the

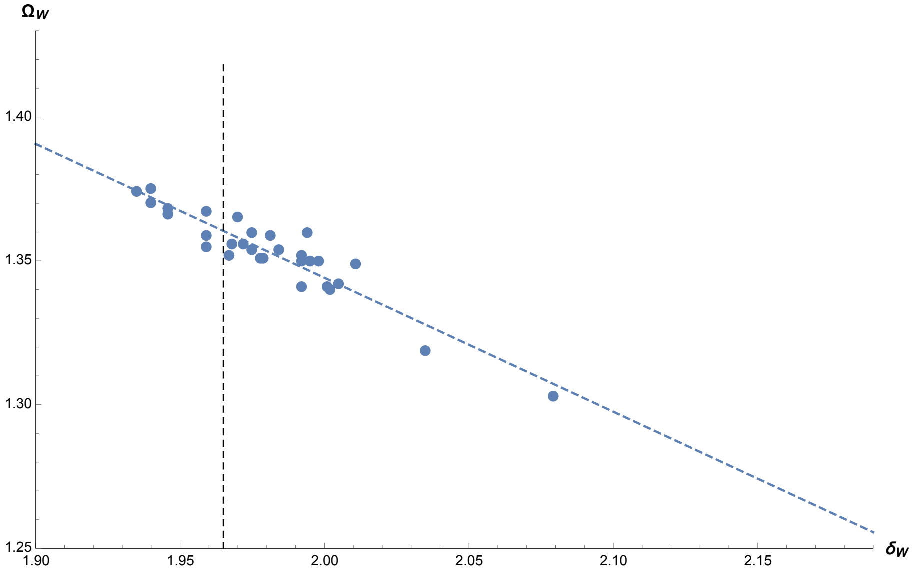


Figure 2: Plot showing linear regression fit for (δ_w, Ω_w) parameters. In the absence of any correlation, we would instead observe a vertical line at $\delta_w = 1.956$ — the value that corresponds to (3).

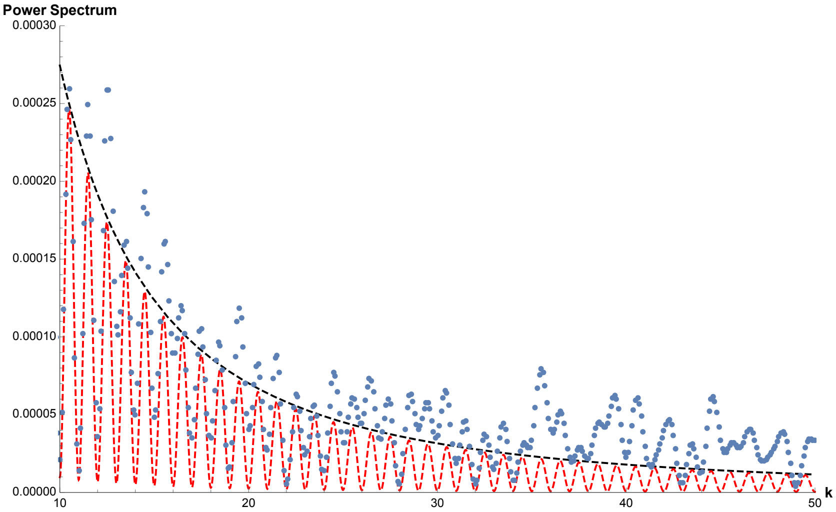


Figure 3: Plot of the spectral density (SARS-CoV-2 genome) showing how after about $k = 10$, the data points appear to be noisy and such stochastic fluctuations appear to persist throughout apart from a couple of isolated peaks. Neither the envelope curve of $1/k^\delta$ nor equation (7) continue to be effective descriptions.

half-width parameter b. We find that this parameter generally increases with κ at both genome and Spike protein levels (see Figures 7a and 7b in Appendix B).

(c) Finally, although for simplicity, we have kept to analyzing the spectral density corresponding to the sum of all the nucleotides, the general qualitative features described in (a) and (b) above apply to the spectral density for each individual nucleotide as well as the cross-spectra.

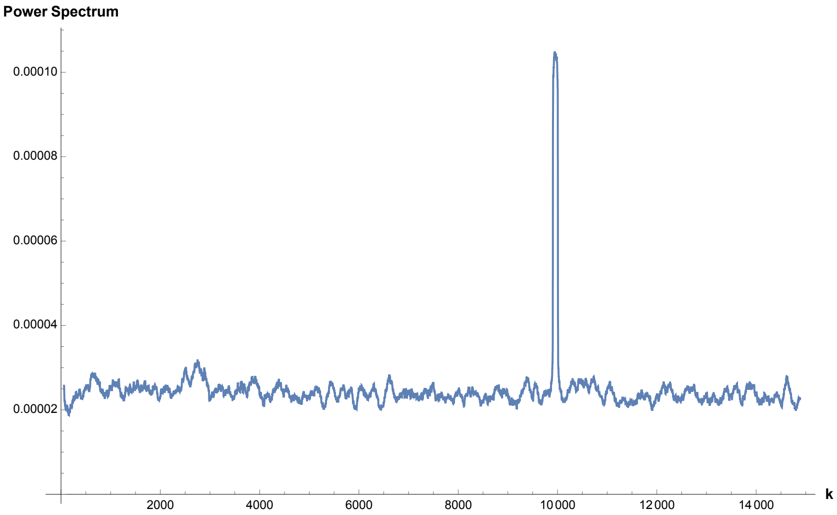


Figure 4: The smoothened data presents a stochastic fluctuation about a fixed mean $\sim 2\times10^{-5}$ and there is only an isolated peak at $M/3$ due to the triplet codon-usage. Only half of the spectrum is shown here since the other half is a reflection of it due to the discrete symmetry $S(k) \leftrightarrow S(M-k)$ the spectrum as mentioned in Section 2.

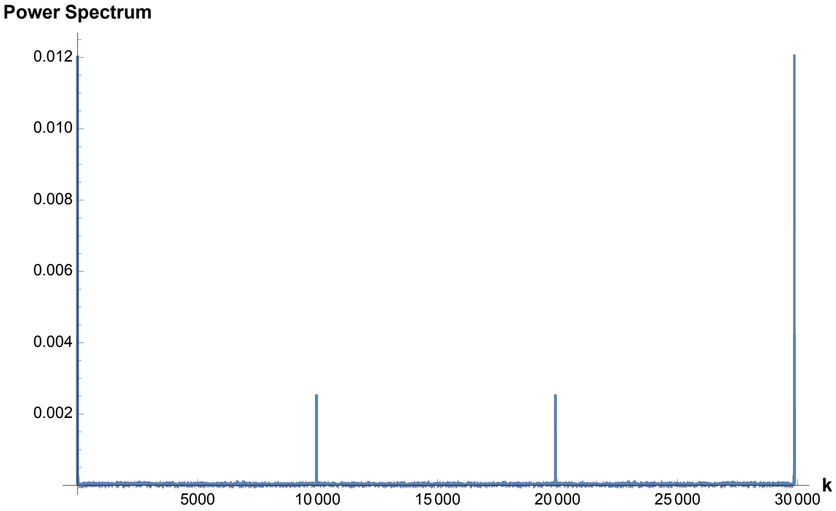


Figure 5: Plot of the Fourier spectral density (SARS-CoV-2 genome) which is mostly featureless with noise apart from prominent peaks at $\frac{M}{3}$, $\frac{2M}{3}$ which correspond to triplet-codon usage.

4 Discussion

We have presented a study of the Fourier spectral density of the coronavirus genome at the level of the entire genome as well as the coding region for the Spike protein. The power spectrum profile can be well-described by considering aspects of deviation from the hypothetical case of a random, uncorrelated sequence (eqn. (3)). We summarize the essential general features below:

(i) There is a low-frequency domain ($k \lesssim 10$) which exhibits a clear oscillatory form that is close to (3). In this domain, we find that the enveloping curve connecting the local maxima is well-described by a power decay law of the form $1/k^{\delta}$. We noted that the power exponent δ shows a correlation with a measure of complexity of the sequence (eqn. (9)) which in the limit of large genome size is the sequence's Shannon entropy. The deviation from the uncorrelated background can be described by a linear relation between δ and Ω. This behavior does not however persist at the level of the Spike protein's coding region.

(ii) Beyond the low-frequency domain, the spectrum displays stochastic fluctuations about certain fixed values $\overline{m}$, and we find no other resonances apart from the peaks at $\frac{M}{3}, \frac{2M}{3}$ which are associated with the universal triplet codon usage. Relative to the uncorrelated case, $\overline{m}$ is about 10^5 higher at the genome level and about 10^4 higher at the Spike protein level. It also generally decreases with the size of the genome or the protein coding region.

(iii) Upon fitting the Lorentzian function to the spectrum with initial and final conditions determined by R_0 and $\overline{m}$ respectively, we find that its half-width parameter is correlated with κ — the dimensionless constant that defines the linearized correlation function in the low-frequency domain, and generally increases with it. This is observed at both the genome and Spike protein's levels.

Let us conclude by briefly pointing out several future directions and applications. Now, it has been noted in literature for some time that DNA viruses and unicellular organisms tend to have mutation rates which vary inversely with the genome size ('Drake's rule' [20, 21, 22]). This correlation has been studied for RNA viruses recently (see for example [23]) although we are unaware of any evidence for the case of coronaviruses[5] which is the only RNA virus family which has a 3'-exonuclease proofreading mechanism that enhances replication fidelity. The parameter $\overline{m}$ that we have introduced here appears to vary inversely with genome size, and thus it may be worthwhile to explore its role in models that attempts to explain viral mutation rates. In [25], a negative association between molecular evolution rate and genome size was established for RNA viruses. It would be interesting to compute the parameter $\overline{m}$ for the viral sequences studied in [25, 26]. Another potential application of our work which has immediate relevance is to study the distribution of $\overline{m}$ for SARS-CoV-2 genomes specifically to explore if they could describe current evolution of the virus (see for example [27]).

The Lorentzian function that we fit broadly to the spectrum as a whole is a coarse-grained description that does not model the transition from the low-frequency spectrum to the other part of the spectrum that appears to be dominated by stochastic fluctuations. It would be interesting to develop theoretical models that could possibly account for such a transition and in the process, construct a clearer understanding for the parameter $\overline{m}$ or why the information-theoretic measure (9) is relevant for the low-frequency domain.

A complementary approach towards understanding correlation effects is to study directly the correlation function itself (see for example [28]), although this is more computationally intensive. It would be interesting to study what forms of correlation functions could lead to the enveloping curve being of the form $1/k^{\delta}$. A few related models were proposed in [29, 30], and it may be worthwhile to revisit them in light of the newfound relation with the measure of complexity.

Finally, it would be interesting to perform a more extensive study of the models here with a larger set of viral genomes so that we have a fuller understanding of their statistical distribution

[5] Some recent results concerning mutation rates of coronaviruses were published in [24].

and whether they can be useful in clustering and classifying purposes.[6] Motivated by the COVID-19 pandemic, notwithstanding our limited dataset, in Table 1 below, we show the viral genome that is the closest neighbor to SARS-CoV-2 for each of the four model parameters at both levels of the genome and Spike protein coding region. From Table 1, we see that Bat-RTG13 features most frequently and that apart from TGEV and HKU1 which infect pigs and humans respectively, the others are bat coronaviruses. Collectively, they appear to be broadly compatible with the plausibility of the bat origin of SARS-CoV-2, while to our knowledge, the association of SARS-CoV-2 with TGEV and HKU1 has never been made in literature.

	Genome	**Spike protein**
δ	TGEV	Bat-RTG13
κ	Bat-RTG13	HKU3
$\overline{m}$	Bat-RTG13	Bat-CoV-512, HKU5
b	HKU9	HKU1

Table 1: We display the coronavirus that has a genome closest to SARS-CoV-2 in terms of each of the model parameters.

Acknowledgments

I thank Neal Snyderman and Rajesh Parwani for stimulating discussions.

A GenBank accession numbers

This appendix collects the GenBank accession ID and names of the 30 coronaviruses used in this work, which largely follows [11], our additions being SARS-CoV-2, MERS-CoV and Bat-RaTG13. These genomes can be freely downloaded from `https://www.ncbi.nlm.nih.gov`. For each genome, we exclude the poly(A) tail for our analysis.

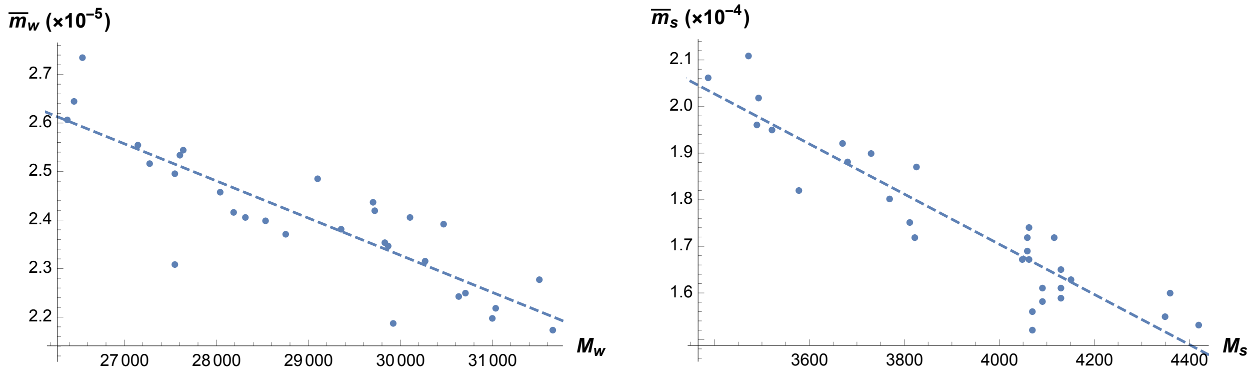
[6] See [31] for a recent attempt in this direction for coronaviruses.

Virus	GenBank accession number
Alphacoronavirus	
Transmissible gastroenteritis virus (TGEV)	DQ811785
Porcine respiratory coronavirus (PCRV)	DQ811787
Feline coronavirus (FCoV)	NC_002306
Human coronavirus 229E (HCoV-229E)	NC_002645
Human coronavirus NL63 (HCoV-NL63)	NC_005831
Porcine epidemic diarrhea virus (PEDV)	NC_003436
Scotophilus bat coronavirus 512	NC_009657
Rhinolophus bat coronavirus HKU2	NC_009988
Miniopterus bat coronavirus HKU8	NC_010438
Miniopterus bat coronavirus 1A	NC_010437
Betacoronavirus	
Human coronavirus OC43	NC_006213
Bovine coronavirus (BCoV)	NC_003045
Porcine hemagglutinating encephalomyelitis virus (PHEV)	KY419103
Equine coronavirus (ECoV)	EF446615
Human coronavirus HKU1	NC_006577
Mouse hepatitis virus (MHV)	AC_000192
SARS coronavirus (SARS-CoV)	NC_004718
SARS coronavirus-2 (SARS-CoV-2)	NC_045512
Bat SARS coronavirus HKU3	GQ153539
Bat coronavirus RaTG13	MN996532
MERS-coronavirus	NC_019843
Tylonycteris bat coronavirus HKU4	NC_009019
Pipistrellus bat coronavirus HKU5	NC_009020
Rousettus bat coronavirus HKU9	NC_009021
Gammacoronavirus	
Infectious bronchitis virus (IBV)	NC_001451
Beluga whale coronavirus (SW1)	NC_010646
Turkey coronavirus (TCoV)	NC_010800
Deltacoronavirus	
Munia coronavirus HKU13	NC_011550
Thrush coronavirus HKU12	NC_011549
Bulbul coronavirus HKU11	FJ376620

Table 2: The GenBank accession ID for the thirty coronavirus RNA genomes covered in this work.

B Some Graphs

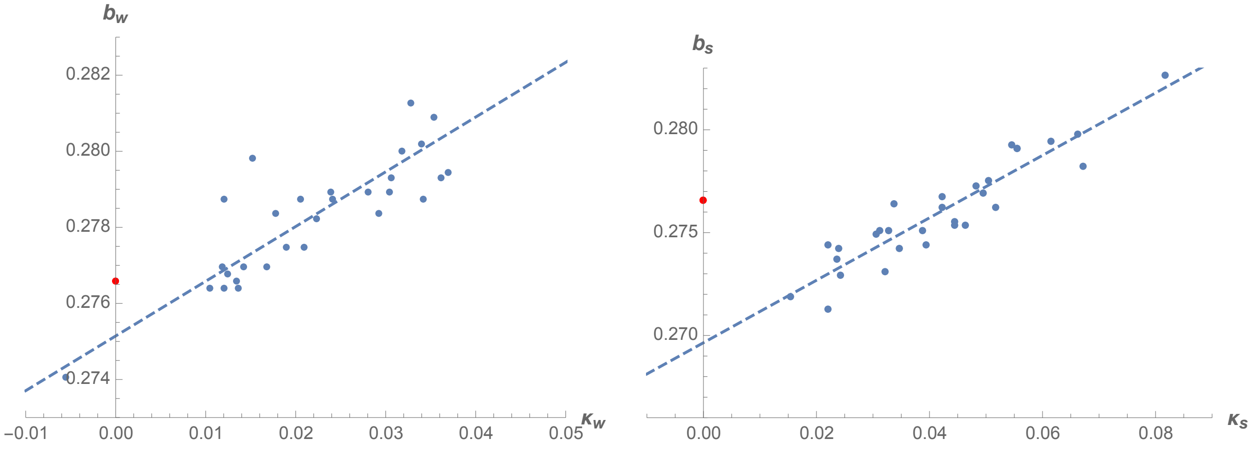
In this Section, we collect several graphs useful for visualizing two particular trends observed: (i) the parameter $\overline{m}$ tends to vary inversely with size of genome/Spike protein coding region, (ii) the linearized correlation function parameter κ and the half-width parameter b appears to be correlated.



(a) At the genome level.

(b) At the Spike protein level.

Figure 6: Plots showing how $\overline{m}$ generally decreases with sequence size at both genome and Spike protein levels. The dashed line is obtained from least-square regression. For the genome, we have $\overline{m}_w = a_w + b_w M_w$, with $a_w/10^{-5} \approx 4.62(4.11, 5.14)$, $b_w/10^{-10} \approx -7.65(-9.42, -5.89)$, whereas for the Spike protein, we have $\overline{m}_s = a_s + b_s M_s$, with $a_s/10^{-4} \approx 3.86(3.52, 4.19)$, $b_s/10^{-8} \approx -5.38(-6.24, -4.52)$.



(a) At the genome level.

(b) At the Spike protein level.

Figure 7: Plots showing how the half-width parameter b generally increases with κ. At the genome level, the best-fit line is $b_w = g_w + h_w\kappa_w$ with $g_w \approx 0.275(0.274, 0.276)$, $h_w \approx 0.144(0.116, 0.172)$. At the Spike protein level, the best-fit line is $b_s = g_s + h_s\kappa_s$, with $g_s \approx 0.270(0.269, 0.271)$, $h_s \approx 0.152(0.128, 0.176)$. The red point pertains to the uncorrelated background.

References

[1] E. Schrdinger, "What Is Life? : The Physical Aspect of the Living Cell". Based on lectures delivered under the auspices of the Dublin Institute for Advanced Studies at Trinity College, Dublin, in February 1943.

[2] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," Phys. Rev. Lett. **68**, 3805-3808 (1992) doi:10.1103/PhysRevLett.68.3805

[3] H. Hanspeter, O. Weiss, and E. Trifonov, "10-11 bp periodicities in complete genomes reflect protein structure and DNA folding," Bioinformatics (Oxford, England). 15. 187-93. 10.1093/bioinformatics/15.3.187. (1999)

[4] Li W, Holste D., "Universal 1/f noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome." Phys Rev E Stat Nonlin Soft Matter Phys. 2005;71(4 Pt 1):041910. doi:10.1103/PhysRevE.71.041910

[5] A. Alain, C. Vaillant, B. Audit, Argoul, Francoise, d'Aubenton-Carafa, Yves and C. Thermes, "Multi-scale coding of genomic information: From DNA sequence to genome structure and function," Physics Reports. 498. 10.1016/j.physrep.2010.10.001. (2001)

[6] M. Altaiski, O. Mornev and R. Polozov, "Wavelet analysis of DNA sequences," Genetic Analysis: Biomolecular Engineering, vol. 12, 165-168 (1996)

[7] A. Arneodo, E. Bacry, P. V. Graves and J. F. Muzy, "Characterizing long-range correlations in DNA sequences from wavelet analysis," Phys. Rev. Lett. **74**, 3293-3296 (1995)

[8] C. C. Jeng, I. Yang, K. Hsieh, and C. Lin, "Bacteria Classification on Power Spectrums of Complete DNA Sequences by Self-Organizing Map," Neural Information Processing, Letters and Reviews, vol. 9, (2006)

[9] S. Buldyrev, A. Goldberger, S. Havlin, R. Mantegna, M. Matsa, C. Peng, M. Simons and H. Stanley, "Long-Range Correlation Properties of Coding and Noncoding DNA Sequences: Genbank Analysis." Phys. Rev. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics. **51**, 5084-91. 10.1103/PhysRevE.51.5084.

[10] A. Fukushima, T. Ikemura, M. Kinouchi, et al. "Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis," Gene. 2002;300(1-2):203-211. doi:10.1016/s0378-1119(02)00850-8 (2002)

[11] P. C. Woo, Y. Huang, S. K. Lau, and K. Y. Yuen, "Coronavirus genomics and bioinformatics analysis," Viruses, 2(8), 18041820. https://doi.org/10.3390/v2081803 (2010)

[12] d. S. Vieira, "Statistics of DNA sequences: a low-frequency analysis," Phys. Rev. E. **60**(5 Pt B):5932-5937 (1999) doi:10.1103/physreve.60.5932

[13] W. Li, T. G. Marr and K. Kaneko, "Understanding long-range correlations in DNA sequences," Physica D: Nonlinear Phenomena, vol. 75, 392-416 (1994)

[14] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] [cited 2020 Jun 30]. Available from: `https://www.ncbi.nlm.nih.gov/`

[15] P. Salamon and A. Konopka, “A Maximum Entropy Principle for the Distribution of Local Complexity in Naturally Occurring Nucleotide Sequences,” Computers and Chemistry. 16. 117-124. 10.1016/0097-8485(92)80038-2.

[16] C. Cattani, “Fractals and Hidden Symmetries in DNA,” Mathematical Problems in Engineering, ”Nonlinear Time Series: Computations and Applications,” vol. 2010, 507056, 2010.

[17] M. Carli, “Visualization and analysis of DNA sequences using DNA walks,” Journal of the Franklin Institute, Engineering and Applied Mathematics, vol. 341, 37-53 (2004)

[18] P. Zhou, X. Yang, X. Wang, et al. “A pneumonia outbreak associated with a new coronavirus of probable bat origin,” Nature 579, 270273 (2020). https://doi.org/10.1038/s41586-020-2012-7

[19] W. Lee and L. Luo, “Periodicity of base correlation in nucleotide sequence,” Phys. Rev. E. **56**, 848-851 (1997) doi:10.1103/PhysRevE.56.848

[20] J. W. Drake, “A constant rate of spontaneous mutation in DNA-based microbes,” Proc. Natl. Acad. Sci. USA 88: 71607164 (1991)

[21] Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow, 1998 Rates of spontaneous mutation. Genetics 148: 1667 1686.

[22] Lynch, M., 2010 Evolution of the mutation rate. Trends Genet. 26: 345352.

[23] Bradwell, Katie and Combe, Marine and Domingo-Calap, Pilar and Sanjun, Rafael. (2013). Correlation Between Mutation Rate and Genome Size in Riboviruses: Mutation Rate of Bacteriophage Q. Genetics. 10.1534/genetics.113.154963.

[24] Zhao Z, Li H, Wu X, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evolutionary Biology. 2004 Jun;4:21. DOI: 10.1186/1471-2148-4-21.

[25] Sanjun, R., 2012 From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS Pathog. 8: e1002685.

[26] K. M. Peck and A. S. Lauring, “Complexities of Viral Mutation Rates,” J Virol. 2018;92(14):e01031-17, (2018)

[27] Xiaolu Tang, Changcheng Wu, Xiang Li, Yuhe Song, Xinmin Yao, Xinkai Wu, Yuange Duan, Hong Zhang, Yirong Wang, Zhaohui Qian, Jie Cui, Jian Lu, On the origin and continuing evolution of SARS-CoV-2, National Science Review, Volume 7, Issue 6, June 2020, Pages 10121023, https://doi.org/10.1093/nsr/nwaa036

[28] Bernaola-Galvn P, Carpena P, Romn-Roldn R, Oliver JL., “Study of statistical correlations in DNA sequences,” Gene. 2002;300(1-2):105-115. doi:10.1016/s0378-1119(02)01037-5

[29] W. Li, “Spatial $1/f$ spectra in open dynamical systems,” Europhysics letter, 10(5), 395-400 (1989)

[30] W. Li, “Expansion-modification systems: a model for spatial $1/f$ spectra,” Physical Review A, 43(10), 5240-5260 (1991)

[31] S. Hassan, R. Ranjeet and V. Sharma, “A Quantitative Genomic View of the Coronaviruses: SARS-COV2,” 10.20944/preprints202003.0344.v1. (2020)