Fourier spectral density of the coronavirus genome

H.S.Tan

June 2020

Division of Physics and Applied Physics, School of Physical and Mathematical Sciences,
Nanyang Technological University,
21 Nanyang Link, Singapore 637371

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 and exhibits no other peaks apart from one 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 low-frequency domain of the spectral density.

Contents

1 Introduction **2**

2 Theoretical preliminaries **3**

2.1 A reference curve: the uncorrelated background 4

2.2 Two simple models 4

2.3 A measure of complexity and Shannon entropy 5

2.4 On the Spike protein 6

3 Results and graphs **6**

4 Concluding remarks **9**

A GenBank accession numbers, and a list of fitted model parameters **11**

1 Introduction

Motivated by our search for deeper organizational principles governing genetic information, 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 [1] 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 [1], 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 [2] 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, such as wavelet-type analysis [3, 4, 5] of the sequences, using features of the spectra to classify and cluster genomes with the aid of neural networks [6], prediction of coding regions [7] and periodic structures [8], 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 [9] 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 ($k < 10$ in units of inverse genome length) where a sinc-squared-like oscillatory form is enveloped by a roughly $1/k^2$ decay curve preceding stochastic white-noise type fluctuations about fixed mean values. We develop a putative explanation for this behavior, explore a couple of simple models to characterize each 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 parametrizes its low-frequency decay behavior. 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 laws of the form $1/k^{\alpha}$ have previously been discussed in literature for other types of genomes (see for example [10, 11]). 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 two tables - one listing all the GenBank accession numbers [12] of the genomes, and another tabulating the numerical 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

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

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 $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{2\pi ik(x_1-x_2)}\langle\phi_{\alpha}(x_1)\phi_{\alpha}(x_2)\rangle dx_1 dx_2$. In this vein, the spectral density is thus often regarded as containing information about correlation effects within the genome.

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 [1], 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.

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 occuring 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. 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 two 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.

2.2 Two simple models

(A) Motivated by previous works on this subject, we consider fitting a power-law decay 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. With our datasets, we would find that the spectral density sum δ is mostly smaller than the unbiased value of 2. For the uracil nucleotide (S_{UU}), we find that δ is always larger than 2 whereas it is the converse for the other three nucleotides. 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 [11]). 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.

(B) 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

$$S(k) = \frac{R_0}{M^2}\sum_{j=1}^{M}\sum_{l=1}^{M}\left(1+\kappa\frac{|l-j|}{M}\right)e^{\frac{2\pi ik(l-j)}{M}} \tag{5}$$

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

$$S(k) = \frac{R_0}{M^4}\left(\frac{\sin^2(\pi k)}{\sin^2\frac{\pi k}{M}} + \kappa\frac{\cos(\pi k)}{2M\sin^3\frac{k\pi}{M}}\left[(M-1)\sin\frac{k\pi(M+1)}{M} + (M+1)\sin\frac{k\pi(1-M)}{M}\right]\right). \tag{6}$$

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. For the uncorrelated ($\kappa = 0$) limit, the zeros occur at integral k, and local maxima at half-integral values are enveloped with a $1/k^2$ curve. 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}.

Beyond the specific linear form of the correlation function postulated in (5), 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 (5) with $F'(0)\tilde{\kappa} = \kappa$. Thus, (5) 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 $\kappa \sim B$.

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 [13, 14, 15]) is defined as follows .

$$\Omega = \frac{1}{M}\log\left(\frac{M!}{N_A!N_C!N_G!N_U!}\right), \tag{7}$$

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

$$\lim_{M\to\infty}\Omega = -\sum_{\alpha\in\{A,C,G,U\}}\left(\frac{N_\alpha}{M}\right)\log\left(\frac{N_\alpha}{M}\right), \tag{8}$$

which is a function of only the fractional distribution of nucleotides. In this form (8), 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 occuring 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.

2.4 On the Spike protein

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, [9, 16].

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. 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[1] consisting of 30 types of coronaviruses spread across four genera mainly follows from reference [9] 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 [17]. 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 ($k \leq 10$), the equation (3) is a good approximation (see Fig. 1). After curve-fitting to the datasets, we find the following range of values for the model parameters:

$$\begin{aligned} &\delta_w \in [1.935, 2.079], \;\; \kappa_w \in [-0.0056, 0.0583], \\ &\delta_s \in [1.744, 1.934], \;\; \kappa_s \in [0.0155, 0.0818], \end{aligned} \tag{9}$$

where the subscripts label the level (w = whole genome, s = Spike protein) at which they are computed. In Table 2, we tabulate their values for all 30 genomes. Most δ_w's are less than 2 (the uncorrelated benchmark), while all except that of Bat-CoV512 coronavirus have positive κ_w. At the Spike protein level, all $\delta_s < 2$ and $\kappa > 0$.

At the whole genome level, we find that in the (δ_w, κ_w) space, there are two particular pairs which are close to each other : (SARS-CoV-2, Bat-RTG13) and (SW1, HKU11). The former pair is known for its sequence similarity (hence the possible bat origin of SARS-CoV-2 as studied in [17]). At the Spike protein level, SARS-CoV-2 and Bat-RTG13 are of the closest similarity again, and there are a few other close pairings: (FCoV, HKU4), (HCoV-OC43, PEDV) and (HKU5, HKU13). It would be interesting to study if there exist any particular relationships between members of each of these pairs.

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

(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. 2). Stochastic fluctuations about a fixed mean $\sim 2 \times 10^{-5}$ 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 [1, 18] 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. 3 and 4).

(c) 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. 5)

$$\Omega_w \sim \alpha + \beta \delta_w, \tag{10}$$

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

$$\alpha \approx 2.28(2.14, 2.41), \qquad \beta \approx -0.47(-0.53, -0.40), \tag{11}$$

with the R^2-value being 0.87. There is also a similar relation between δ_w and Ω_s, consistent with a linear relation that we found to exist between Ω_s and Ω_w (see Fig. 6).

$$\Omega_w = \mathcal{C}\Omega_s, \;\; \mathcal{C} \approx 1.004(1.003, 1.006), \tag{12}$$

with the R^2-value being 0.99998. 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.

(d) 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 nucleotides as well as the cross-spectra.

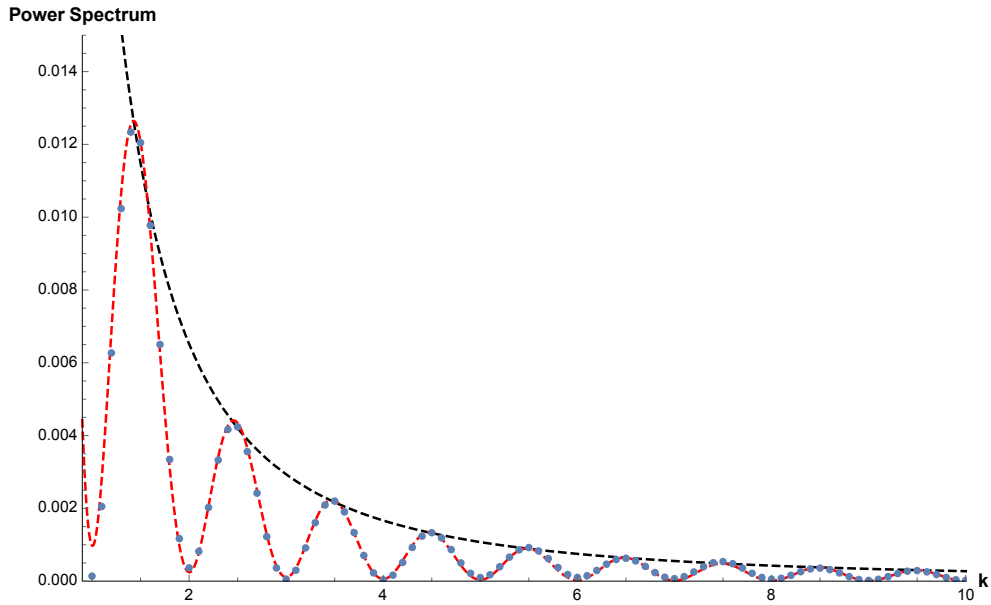

Figure 1: Plot of the power spectral density in the range where we perform the curve-fitting. The black dashed line is the curve $\sim 1/k^{\delta}, \delta = 1.968$ obtained from the set of local maxima, while the red dashed line is equation (6) obtained by fitting to all maxima and minima, with the best-fit value $\kappa = 0.0362$.

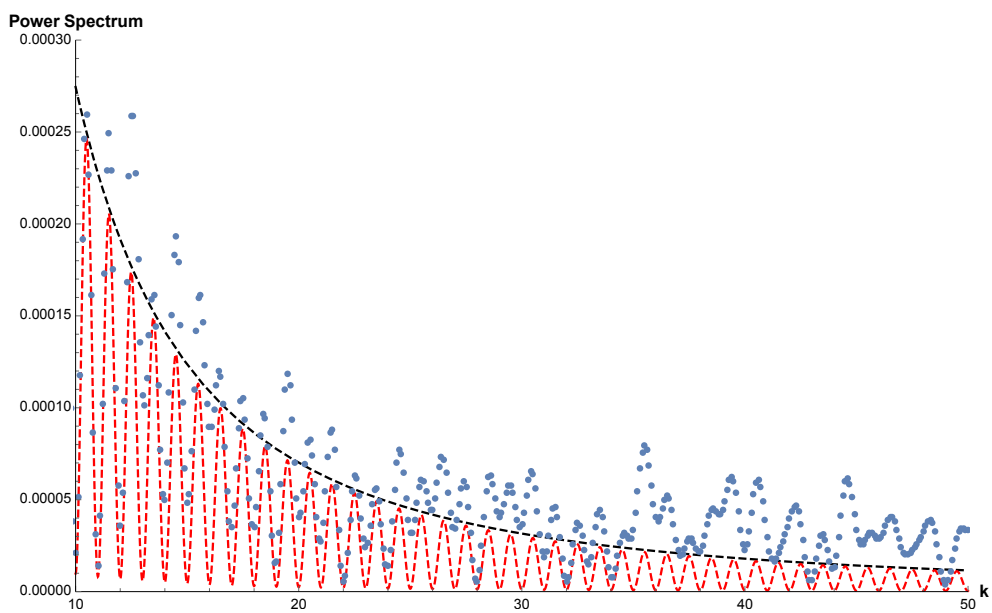

Figure 2: Plot showing how after about $k = 10$, the data points appear to be noisy and such stochastic fluctuations appear to persist throughout apart from a couple of isolated peaks. Neither the envelope curve of $1/k^{\delta}$ nor equation (6) continue to be effective descriptions.

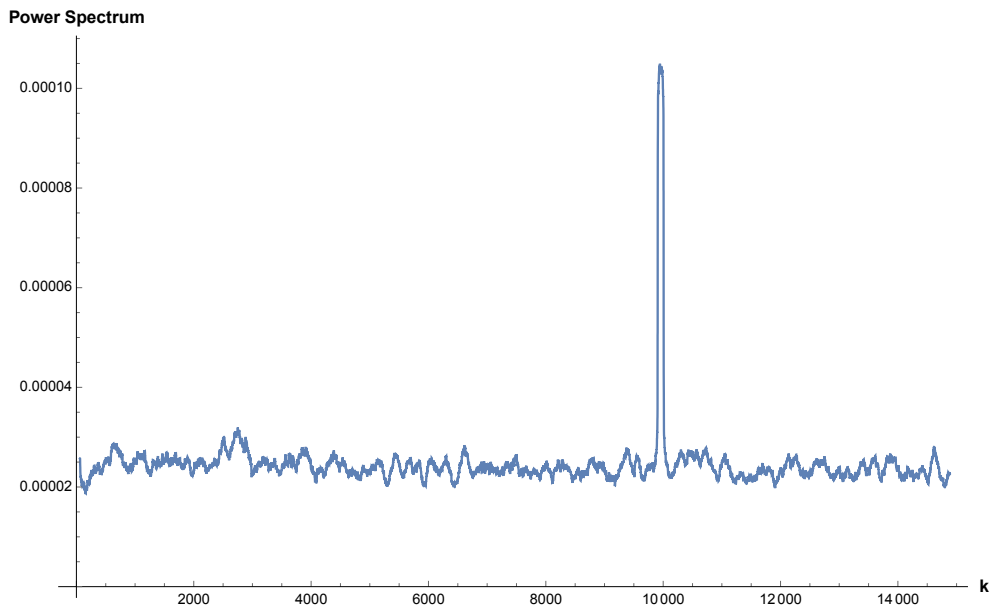

Figure 3: The smoothened data presents a stochastic fluctuation about a fixed mean $\sim 2\times10^{-5}$ and there is only an isolated peak at $M/3$ due to the triplet codon-usage. Only half of the spectrum is shown here since the other half is a reflection of it due to the discrete symmetry $S(k) \leftrightarrow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

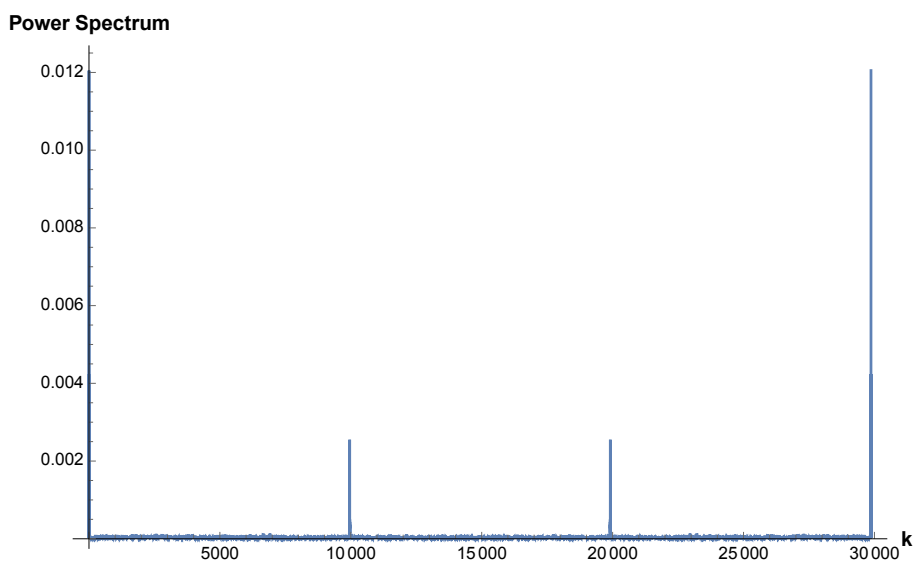

Figure 4: Plot of the Fourier spectral density which is mostly featureless with noise apart from prominent peaks at $\frac{M}{3}, \frac{2M}{3}$ which correspond to triplet-codon usage.

4 Concluding remarks

We found that the RNA genome of coronaviruses has a Fourier spectral density that at low frequencies ($k < 10$) presents a small and distinct deviation from the curve derived by assuming an

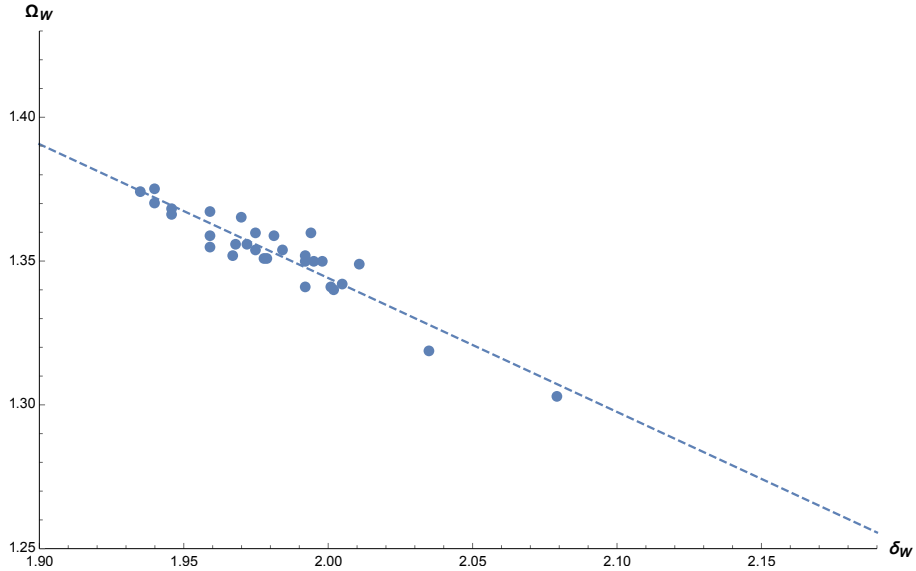

Figure 5: Plot showing linear regression fit for (δ_w, Ω_w) parameters. In the absence of any correlation, we would instead observe a vertical line at $\delta_w = 2$.

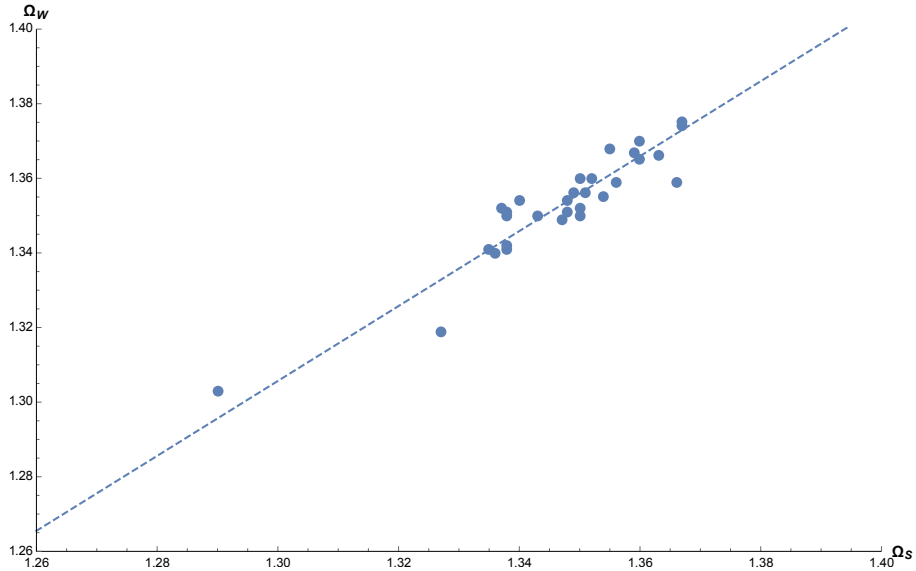

Figure 6: Plot showing linear regression fit for (Ω_s, Ω_w) parameters.

uncorrelated sequence where each nucleotide type occurs at each position with some fixed probability. Away from this low-frequency domain, the spectrum appears to be largely featureless with stochastic fluctuations about some fixed mean and absent of prominent peaks apart from those at $\frac{M}{3}, \frac{2M}{3}$ which follow from triplet codon usage. We presented two simple approaches towards characterizing the low-frequency behavior:

(i) by fitting a power-law decay parametrized by an exponent δ to the enveloping curve

(ii) by assuming a linear (anti-)correlation function from the outset and deriving the modified spectral density curve which is parametrized by a dimensionless constant κ

We computed these model parameters and found a linear correlation between δ and a measure of complexity of the genome [13]. We have also carried out this analysis for the coding region associated with the Spike protein and found similar behavior, but at the protein level, there is no correlation between δ_s and the coding region's complexity. To our knowledge, this is the first scaling law observed for any genome that involves a measure of its complexity and a characterizing parameter of the spectral density. Since the complexity measure that we use here can be interpreted as the Shannon entropy in the limit of large genome size, this is a concrete realization of how information-theoretic content is present in the spectral density of the viral genome.

It would be interesting to explore further if the model parameters δ, κ (or more generally other descriptive parameters of the low-frequency domain arising from other models) can be used for clustering purposes (see [19] for a recent attempt in this direction for coronaviruses). They could also potentially be useful in deciphering other properties of the viral genome, such as those related to the receptor-binding domain usage [16], mutation rates [20] etc that may be related to long-range correlation effects within the sequence.

A complementary approach towards understanding correlation effects is to study directly the correlation function itself (see for example [21]), 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 [22, 23], and it may be worthwhile to revisit them in light of the newfound relation with the measure of complexity.

A GenBank accession numbers, and a list of fitted model parameters

This appendix collects two tables: Table 1 listing the GenBank accession ID and names of the 30 coronaviruses used in this work, and Table 2 listing various numerical parameters that characterize each genome sequence. The dataset largely follows [9], our additions being SARS-CoV-2, MERS-CoV and Bat-RaTG13. These genomes can be freely downloaded from `https://www.ncbi.nlm.nih.gov`.

References

[1] 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

[2] 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)

[3] 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)

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

[5] 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)

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

[6] 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)

[7] 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.

Virus	M	Ω_w	δ_w	κ_w	M_{spike}	Ω_s	δ_s	κ_s
SARS-CoV	29727	1.368	1.946	0.0342	3768	1.355	1.816	0.0443
HKU12	26380	1.355	1.959	0.0120	3579	1.354	1.778	0.0395
MERS-CoV	30107	1.367	1.959	0.0189	4062	1.359	1.892	0.0422
SARS-CoV-2	29870	1.356	1.968	0.0362	3822	1.349	1.879	0.0554
Bat-RaTG13	29835	1.356	1.972	0.0370	3810	1.351	1.884	0.0483
IBV	27608	1.351	1.979	0.0209	3489	1.338	1.906	0.0421
HCoV-229E	27277	1.350	1.992	0,0240	3522	1.350	1.798	0.0671
HCoV-OC43	30713	1.340	2.002	0.0317	4062	1.336	1.850	0.0239
HCoV-NL63	27553	1.319	2.035	0.0328	4071	1.327	1.744	0.0504
HKU1	29926	1.303	2.076	0.0583	4071	1.290	1.795	0.0547
MunCoV	26536	1.374	1.935	0.0104	3471	1.367	1.829	0.0445
FCoV	29355	1.354	1.984	0.0178	4359	1.348	1.865	0.0311
PEDV	28033	1.365	1.970	0.0280	4152	1.360	1.845	0.0244
PRCV	27550	1.351	1.978	0.0239	3669	1.348	1.795	0.0517
BatCoV-512	28179	1.359	1.959	-0.0056	4116	1.366	1.821	0.0387
HKU2	27149	1.349	2.011	0.0206	3387	1.347	1.838	0.0322
HKU8	28757	1.359	1.981	0.0124	4128	1.356	1.867	0.0155
BatCoV1A	28309	1.350	1.995	0.0119	4128	1.343	1.869	0.0219
BCoV	31028	1.341	2.001	0.0340	4092	1.335	1.820	0.0237
MHV	31511	1.366	1.946	0.0137	4131	1.363	1.854	0.0662
HKU4	30270	1.350	1.998	0.0224	4059	1.338	1.86	0.0307
HKU5	30465	1.375	1.940	0.0167	4059	1367	1.832	0.0462
HKU9	29097	1.360	1.994	0.0306	3825	1.352	1.853	0.0329
TCoV	27631	1.352	1.992	0.0292	3681	1.337	1.895	0.0818
SW1	31654	1.354	1.975	0.0143	4419	1.340	1.779	0.0221
HKU11	26459	1.360	1.975	0.0134	3492	1.350	1.909	0.0615
PHEV	30632	1.342	2.005	0.0353	4050	1.338	1.817	0.0496
ECoV	30992	1.341	1.992	0.0151	4092	1.338	1.896	0.0336
TGeV	28536	1.352	1.967	0.0304	4350	1.350	1.777	0.0347
HKU3	29704	1.370	1.940	0.0120	3729	1.360	1.934	0.0555

Table 2: All the model parameters, genome length M and complexity Ω, at the level of the entire sequence (with subscript M) and the coding region for the Spike protein (with subscript s). The poly(A) tail is excluded from each genome.

[8] 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)

[9] 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)

[10] 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

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

[12] 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/`

[13] 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.

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

[15] 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)

[16] Cao, Y., Li, L., Feng, Z. et al., "Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations," Cell Discov 6, 11 (2020). https://doi.org/10.1038/s41421-020-0147-1

[17] 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

[18] 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

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

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

[21] 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

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

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