Tsonis et al. Human Genomics (2021) 15:26
https://doi.org/10.1186/s40246-021-00327-2

Human Genomics

PRIMARY RESEARCH Open Access

An application of slow feature analysis to the genetic sequences of coronaviruses and influenza viruses

Anastasios A. Tsonis[1,2*], Geli Wang[3], Lvyi Zhang[3], Wenxu Lu[3], Aristotle Kayafas[4] and Katia Del Rio-Tsonis[4*]

Abstract

Background: Mathematical approaches have been for decades used to probe the structure of nucleotide sequences. This has led to the development of Bioinformatics. In this exploratory work, a novel mathematical method is applied to probe the genetic structure of two related viral families: those of coronaviruses and those of influenza viruses. The coronaviruses are SARS-CoV-2, SARS-CoV-1, and MERS. The influenza viruses include H1N1-1918, H1N1-2009, H2N2-1957, and H3N2-1968.

Methods: The mathematical method used is the slow feature analysis (SFA), a rather new but promising method to delineate complex structure in nucleotide sequences.

Results: The analysis indicates that the nucleotide sequences exhibit an elaborate and convoluted structure akin to complex networks. We define a measure of complexity and show that each nucleotide sequence exhibits a certain degree of complexity within itself, while at the same time there exists complex inter-relationships between the sequences within a family and between the two families. From these relationships, we find evidence, especially for the coronavirus family, that increasing complexity in a sequence is associated with higher transmission rate but with lower mortality.

Conclusions: The complexity measure defined here may hold a promise and could become a useful tool in the prediction of transmission and mortality rates in future new viral strains.

Keywords: Nucleotide complexity, Slow feature analysis, Coronaviruses, Influenza viruses

Background

Since the early 1970s, scientists have attempted to discover some kind of order or hidden structures in nucleotide sequences. With the advent of sequencing techniques in the late 1970s, scientists had the opportunity to probe nucleic acid sequences for such order [1–3]. Soon, mathematical approaches were employed to shed light in this endeavor, leading to the full-blown field of Bioinformatics [4–7]. We report, for the first time, the application of slow feature analysis (SFA) to genetic sequences. SFA is a procedure for extracting slowly varying, driving signals from a given nonstationary time series and is used here to delineate signals or structure in nucleotide sequences, which would not otherwise be detected. Descriptions of this procedure, which have been successfully applied in many scientific areas, have been reported previously in detail [8–10].

Methods

SFA is an approach that is designed to optimally identify low-frequency behavior in a time series, thereby

* Correspondence: aatsonis@uwm.edu; delriok@miamioh.edu
[1]Department of Mathematical Sciences, Atmospheric Sciences Group, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA
[4]Department of Biology and Center for Visual Sciences, Miami University, Oxford, OH 45056, USA
Full list of author information is available at the end of the article

delineating its complex structure more effectively. This analysis is rooted, theoretically, in the time-embedding theorems. In this method, a one-dimensional time series is embedded in a multi-dimensional space consisting of the original time series and lagged copies thereof. SFA further uses a nonlinear expansion to map this multi-dimensional input signal onto an even larger feature space, and then solves a linear problem to find a linear combination of feature-space variables that minimizes their time derivative (rate of change) [11]. The objective of SFA is to find the optimally filtered signals that vary as slowly as possible but still carry significant information. To ensure this, the output signals need to be uncorrelated and have unit variance [12]. This approach has been successfully applied in many areas, including climate science [13, 14].

In mathematical terms [8], the goal of SFA is, given an n-dimensional input signal $\mathbf{x}(t)$, to find a set of real-valued input-output functions $g_j(\mathbf{x})$ such that the output signals

$$y_j(t) \coloneqq g_j(\mathbf{x}(t))$$

minimize $\Delta(y_i) \coloneqq \langle \dot{y}_j^2 \rangle_t$
under the constraints

$$\begin{aligned} \langle y_j \rangle_t &= 0 && \text{(zero mean)},\\ \langle y^2{}_j \rangle_t &= 1 && \text{(unit variance)},\\ \forall i < j : \langle y_i y_j \rangle_t &= 0 && \text{(decorrelation and order)} \end{aligned}$$

with $\langle\cdot\rangle_t$ and $\dot{y}$ indicating temporal averaging and the derivative of y, respectively.

The Δ-value is a measure of the temporal slowness of the signal $y(t)$. It is given by the mean square of the signal's time derivative. Small Δ-values correspond to slowly varying signals. The first two constraints avoid the trivial constant solution, while the last constraint guarantees that the output functions g_j are distinct and hence extract different information from the input signal. For a tutorial on this method, the reader could consult reference [8] or a more recent presentation in [15]. In that tutorial, a simple example of a two-dimensional input signal $x_1(t)=\sin(t)+\cos(11t)^2$ and $x_2(t)=\cos(11t)$ is considered. Both components are quickly varying, but hidden in the signal is the slowly varying "feature" $y(t)=x_1(t)-x_2(t)^2=\sin(t)$, which can be extracted with a polynomial of degree two, namely $h(\mathbf{x})=x_1-x_2{}^2$.

In the situation with one observable (time series of some variable) from an unknown system where the actual state space is not known (as is the case here), embedding is necessary (and essential) to delineate the underlying dynamics much like in attractor reconstructions. The SFA algorithm can be summarized as follows. Consider a time series $\{x(t)\}_{t=t_1,\ldots,t_n}$, where t denotes time and n indicates the length of the time series. First, we embed $\{x(t)\}$ into an m-dimensional state space using time-delayed copies of $x(t)$:

$$\mathbf{X}(t) = \{x_1(t), x_2(t), \ldots, x_m(t)\}_{t=t_1,\ldots,t_N},$$

where $x_1(t) = x(t)$; $x_2(t) = x_1(t-\tau)$; $x_3(t) = x_1(t-2\tau)$, and so on, τ is the delay, and $N = n - m + 1$. Then, nonlinear expansions (usually second-order polynomials) are used to generate a k-dimensional function state space:

$$\begin{aligned} \mathbf{H}(t) = \{&x_1(t), \ldots, x_m(t), x_1^2(t), \ldots, x_1(t)x_m(t)\\ &, \ldots, x_{m-1}^2(t), \ldots, x_m^2(t)\}_{t=t_1,\ldots,t_N}, \end{aligned}$$

which can also be written as $\mathbf{H}(t) = \{h_1(t), h_2(t), \ldots, h_k(t)\}_{t=t_1,\ldots,t_N}$, where

$$k = m + m(m+1)/2.$$

The expanded signal $\mathbf{H}(t)$ is then centered and normalized to zero mean and unit variance. This process is referred to as whitening or sphering. Thus, we have

$$\mathbf{H}'(t) = \left\{h_1'(t), h_2'(t), \ldots, h_k'(t)\right\}_{t=t_1,\ldots,t_N},$$

where

$$\overline{h_j'} = 0 \text{ (zero mean)},$$

$$h_j' h_j'^{T} = 1 \text{ (unit variance)},$$

$h_j'(t) = [h_j(t) - \overline{h_j}]/S$, and $S = \frac{1}{k}\sqrt{\sum_{j=1}^{k}(h_j(t)-\overline{h})^2}$.

Using the Schmidt algorithm, $\mathbf{H}'(t)$ is orthogonized into:

$$\mathbf{Z}(t) = \{z_1(t), z_2(t), \ldots, z_k(t)\}_{t=t_1,\ldots,t_N},$$

where the transformed signal matrix $\mathbf{Z}$ is column orthogonal:

$$\overline{z_i}(t) = \overline{z_j}(t) = 0, \quad z_i^T(t)\cdot z_j(t) = 0, \quad z_j^T(t)\cdot z_j(t) = 1,$$

The final step of SFA is to find the set of coefficients $(a_1, a_2, \ldots, a_k)$ such that the time series

$$y(t) = a_1 z_1(t) + a_2 z_2(t) + \ldots + a_k z_k(t)$$

varies as slowly as possible. This set is given by the eigenvector $\mathbf{W}_1$ of the time-derivative covariance matrix

$$\mathbf{B} = \dot{\mathbf{Z}}^T \dot{\mathbf{Z}}$$

corresponding to the smallest eigenvalue λ_1. Here

$$\dot{\mathbf{Z}}(t) = \{\dot{z}_1(t), \dot{z}_2(t), \ldots, \dot{z}_k(t)\}_{t=t_1,\ldots,t_N}$$

and

$$\dot{z}_j(t_i) = z_j(t_{i+1}) - z_j(t_i).$$

Using $\mathbf{W}_1$, the optimally filtered slow-feature signal (also known as a driving force factor, which can be composed of one or more components) can be written as:

$$y(t) = r\mathbf{W}_1 \cdot \mathbf{Z}(\mathrm{t}) + \mathrm{c}, \quad (1)$$

where r and c are constants derived to best match y(t) and the original time series $x(t)$.

Once the optimally filtered (low-frequency) SFA signal has been identified, its significant periodicities can be found from the time-averaged wavelet power spectrum. Wavelet analysis has been widely used to analyze localized structures and spectral properties of time series. For example, [16] provides a detailed description of the wavelet analysis, along with a very useful toolkit to conduct step-by-step wavelet analysis, including a statistical significance test based on the red-noise surrogate data (see http://paos.colorado.edu/research/wavelets/). We here used the Morlet wavelet with the wavenumber set to 4 to match the smoothness of the SFA-derived slow-feature signal, focusing, once again, on the spectral peaks statistically significant at the 5% level. Note also that SFA is applicable to non-stationary data, so no data preprocessing is required.

The combination of the SFA and wavelet analyses we use in the present study has been shown to be more effective in diagnosing low-frequency periodicities in data sets of a limited length than direct spectral analysis methods. Note that the driving force may not necessarily consist of just one component, but several components, which, as we will see below, correspond to forcings or signals at certain time scales. The success of SFA in delineating these slow signals lies in the fact that embedding the time series in high enough dimensions and the subsequent dynamical procedure removes the noise and small-scale features that may obscure or suppress those slow signals, thereby delineating more accurately the complex structure of a sequence.

Analysis and results

We first analyzed the nucleotide sequences from three viruses from the same family: SARS-CoV-2, SARS-CoV-1, and MERS. Those sequences are approximately 30,450 bases long and part of the now world-infamous coronavirus family. Since a nucleotide sequence is a string of the bases A, T (U in RNA), C, and G, we first transformed it to a time series of integers in the interval [1–4] (i.e., A➔1, T/U ➔2, C ➔3, G ➔4). Here, we need to stress that a time series represents a particular type of process, where some quantity is sampled in time, t. A nucleotide sequence is a very similar object, but the "sampling" is over space. In a time series, we are interested in the dependency of observations at different time scales, whereas in nucleotide sequences, we are interested in dependencies in different space scales. As such, the mathematical tools to identify structures in time can in principle be applied to identify structure in space, as long as t is thought as a parameter identifying the scale. Transforming a nucleotide sequence into a time series has been used in the past to identify interesting properties in nucleotide sequences (such as the well-known period 3; see [4, 5] and references therein). Note also that the above transformation of A➔1, T/U ➔2, C ➔3, and G ➔4 may, depending on frequency distribution of A, T/U, C, and G in the sequence, result in a nonstationary time series. However, unlike other spectral methods, SFA is not affected by non-stationarity in the data.

Once we have a time series, we apply SFA, and once we have the SFA signal (which as we mentioned above may be comprised of several components, see Eq. 1), we extract the SFA components by wavelet analysis. Figure 1 shows the SFA signal for SARS-CoV-2 virus for m=15 and τ=1. As explained above, this signal is normalized to zero mean and unit variance. Figure 2 shows the wavelet of the time series in Fig. 1. In order to extract the peak "periods" of the driving force signal, we used the Morlet wavelet to compute the time-averaged power spectrum of the wavelet transform [16]. The black solid line in Fig. 3 is the time-averaged power spectrum of the wavelet transform of the driving force, and the dashed line represents the 95% confidence level, estimated using AR-1 surrogate data [16]. The dots show the periods of the oscillatory components of the driving force that are significant above the 95% level.

The significant peak periodicities for SARS-CoV-2 are as follows[1]:

[1]Note here that before we applied SFA to actual nucleotide sequences, and in order to test the efficiency of SFA when the time series is a string of integers, we considered artificial sequences of known periodicities. SFA was able to reproduce the known periodicities.

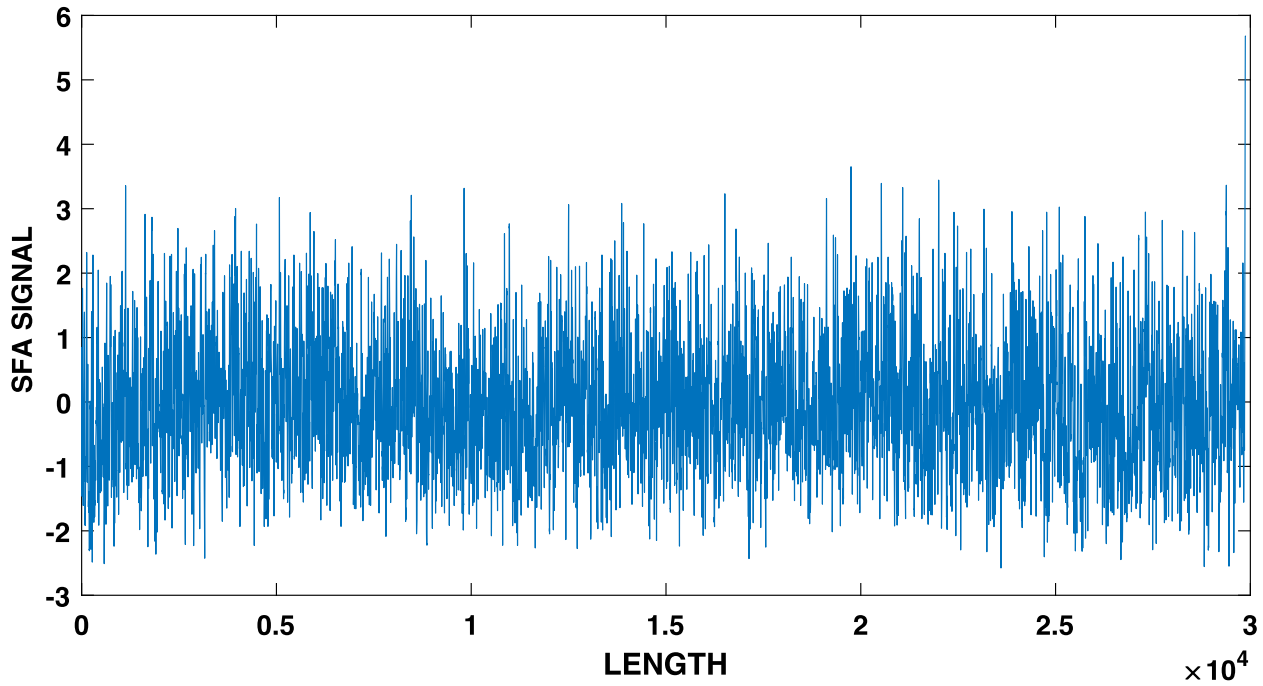

Fig. 1 The SFA signal of the nucleotide sequence of SARS-CoV-2. Note the oscillatory components at many scales

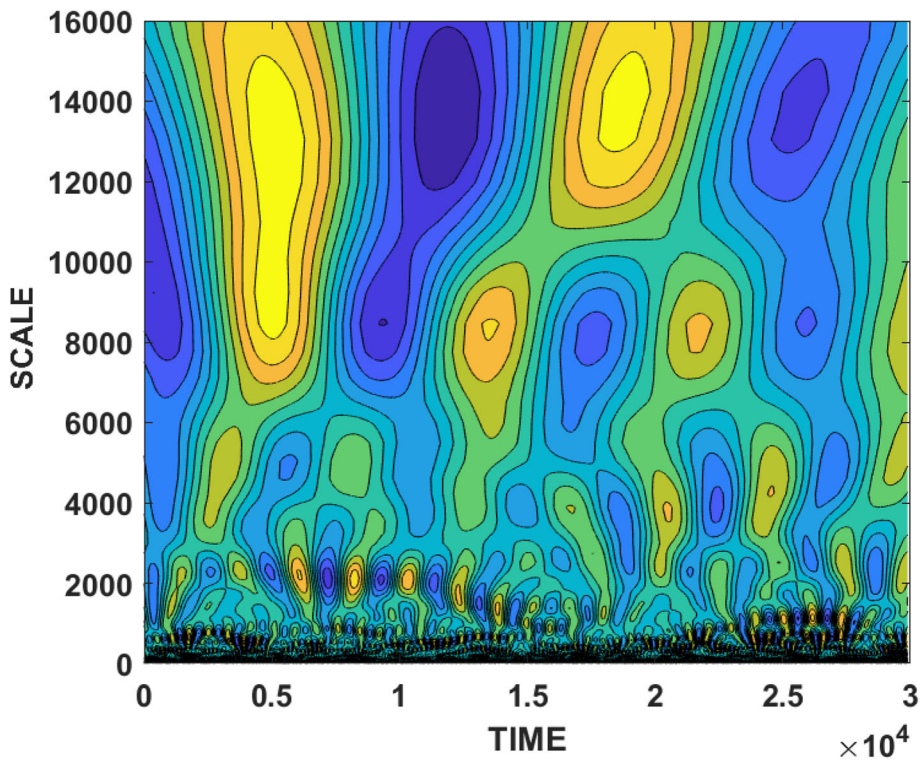

Fig. 2 The wavelet of the signal extracted from Fig. 1

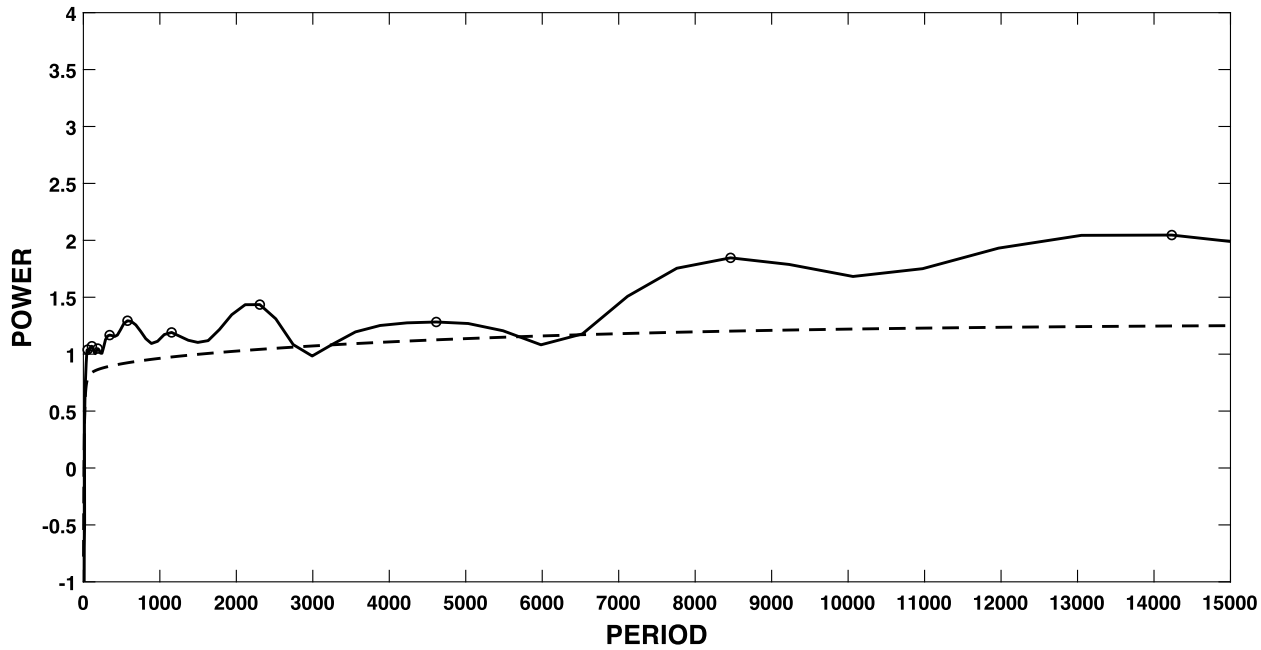

Fig. 3 The time-averaged power spectrum of the wavelet transform extracted from Fig. 2. The dashed line represents the 95% confidence level. The dots show the periods of the oscillatory components of the driving force that are significant above the 95% level

$$
\begin{aligned}
P_1 &= 55.5956928123500\\
P_2 &= 111.191385624700\\
P_3 &= 187.000875157807\\
P_4 &= 342.961117205042\\
P_5 &= 576.789548058258\\
P_6 &= 1153.57909611652\\
P_7 &= 2307.15819223303\\
P_8 &= 4614.31638446607\\
P_9 &= 8462.69356236189\\
P_{10} &= 14232.4973599616
\end{aligned}
\tag{2}
$$

Given the above periodicities recovered from SARS-CoV-2, we next construct Table 1, which shows the ratios between these peaks. We observe the following EXACT relations between peak periods:

$$
\begin{aligned}
P_2 &= 2P_1\\
P_{10} &= 128P_2\\
P_{10} &= 256P_1\\
P_8 &= 2P_7\\
P_8 &= 4P_6\\
P_8 &= 8P_5\\
P_7 &= 2P_6\\
P_7 &= 4P_5\\
P_6 &= 2P_5
\end{aligned}
\tag{3}
$$

And the following almost exact relationships based on the criterion:

$$|\ \text{P–nearest integer}\ |\ /\text{nearest integer} < 0.25\%$$

Table 1 Ratios between the peaks in (2)

Period	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	1	2	3.363585661	6.168843302	10.37471644	20.74943287	41.49886575	82.9977315	152.2185107	256
P2		1	1.681792831	3.084421651	5.187358219	10.37471644	20.74943287	41.49886575	76.10925536	128
P3			1	1.834008086	3.084421651	6.168843302	12.3376866	24.67537321	45.254834	76.10925536
P4				1	1.681792831	3.363585661	6.727171322	13.45434264	24.67537321	41.49886575
P5					1	2	4	8	14.67206469	24.67537321
P6						1	2	4	7.336032346	12.3376866
P7							1	2	3.668016173	6.168843302
P8								1	1.834008086	3.084421651
P9									1	1.681792831
P10										1

$$\begin{aligned} P_9 &= 152P_1 \\ P_9 &= 76P_2 \\ P_{10} &= 76P_3 \\ P_8 &= 83P_1 \end{aligned} \tag{4}$$

Keeping only those relationships, we remain with Table 2, which could be thought as portraying the degree of structure or complexity in the SARS-CoV-2 sequence. We observe in the exact relationships multiples of a power of 2 and in the almost exact relationships multiples of 19 (152=2×76=8×19) and 83. Clearly, a sophisticated and rather convoluted structure, with numerous processes embedded in the sequence, is present. Keep in mind that the factors 19 and 83 (odd numbers) will appear in the rest of the sequences studied here. We define the number of entries above the diagonal in Table 2 as the degree complexity, C. In this case, C=13.

In the Supplementary material, Figures S1, S2, S3 are similar to Figs. 1, 2, and 3, and Tables ST1 and ST2 are similar to Tables 1 and 2 but for SARS-CoV-1. Figures S4, S5, and S6 are similar to Figs. 1, 2, and 3, and Tables ST3 and ST4 are similar to Tables 1 and 2 but for MERS.

According to Figure S3, the peak periods for SARS-CoV-1 are as follows:

$$\begin{aligned} P_1 &= 50.9814750936898 \\ P_2 &= 111.191385624700 \\ P_3 &= 528.918347647618 \\ P_4 &= 748.003500631229 \\ P_5 &= 2115.67339059047 \\ P_6 &= 3558.12433999040 \\ P_7 &= 8462.69356236189 \\ P_8 &= 13051.2576239846 \end{aligned}$$

and according to Tables ST1 and ST2, we now have five exact periodicities

$$\begin{aligned} P_5 &= 4P_3 \\ P_6 &= 32P_2 \\ P_7 &= 4P_5 \\ P_7 &= 16P_3 \\ P_8 &= 256P_1 \end{aligned} \tag{5}$$

and three almost exact

$$\begin{aligned} P_5 &= 19P_2 \\ P_7 &= 166P_1 \\ P_7 &= 76P_2 \end{aligned} \tag{6}$$

Again here, we observe in the exact relationships, multiples of a power of 2, and in the almost exact between P_7 and P_1, P_7 and P_2, and between P_5 and P_2. Note again the multiples of 19 and 83. Note also that from the almost exact relationships, it follows that P_7/P_5=4, which is one of the exact relationships. Here, the degree of complexity is C=8.

According to Figure S6, the peak periods for MERS are as follows:

$$\begin{aligned} P_1 &= 101.962950187380 \\ P_2 &= 132.229586911904 \\ P_3 &= 242.510131659000 \\ P_4 &= 628.993462278030 \\ P_5 &= 1779.06216999520 \\ P_6 &= 2515.97384911212 \\ P_7 &= 4614.31638446607 \\ P_8 &= 8462.69356236189 \\ P_9 &= 13051.2576239846 \end{aligned}$$

and according to Tables ST3 and ST4, we now have three exact periodicities

Table 2 Same as Table 1 but keeping only the exact and almost exact relationships, see relationships (3) and (4)

Period	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	1	2						83	152	256
P2		1							76	128
P3			1							76
P4				1						
P5					1	2	4	8		
P6						1	2	4		
P7							1	2		
P8								1		
P9									1	
P10										1

$$\begin{aligned} P_9 &= 128P_1 \\ P_8 &= 64P_2 \\ P_6 &= 4P_4 \end{aligned} \tag{7}$$

and three almost exact relationships

$$\begin{aligned} P_8 &= 83P_1 \\ P_7 &= 19P_3 \\ P_6 &= 19P_2 \end{aligned} \tag{8}$$

P_9, P_8, and P_6 are multiples (again in a power of 2) of P_1, P_2, P_4 (ordered in a bottom-top "symmetric" way), P_6, P_7, and P_8 are multiples of P_2, P_3, and P_1 but not of 2, but again of 19 and 83 (it is interesting to note that the odd multiples of 19 and 83 appear in all three sequences). Here, the degree of complexity is C=6.

By comparing Tables 2, ST2, and ST4 (and their associated C), one may argue that there is more embedded complexity and intricate patterning in SARS-CoV-2 than SARS-CoV-1 and MERS.

Other important relationships

Keeping in mind that all three sequences belong to the same coronavirus family, there are similarities and interrelationships between the sequences. For example, it is easy to observe that:

$$\begin{aligned} P_7 \text{ (MERS) is the same as } P_8 \text{ (SARS–CoV–2)} \\ P_8 \text{ (MERS) is the same as } P_9 \text{ (SARS–CoV–2)} \\ P_8 \text{ (MERS) is the same as } P_7 \text{ (SARS–CoV–1)} \\ P_9 \text{ (MERS) is the same as } P_8 \text{ (SARS–CoV–1)} \\ P_2 \text{ (SARS–CoV–2) is the same as } P_2 \text{ (SARS–CoV–1)} \\ P_9 \text{ (SARS–CoV–2) is the same as } P_7 \text{ (SARS–CoV–1)} \end{aligned} \tag{9}$$

In general, SFA reveals a consistent picture between these sequences with very intricate structure with details at many scales, indicating very elaborate and sophisticated embedded processes, with complexity increasing from MERS to SARS-CoV-1 to SARS-CoV-2.

Extension of the analysis to the influenza viruses of H1N1-1918, H1N1-2009, H2N2-1957, and H3N2-1968

In an effort to provide further support for the efficiency and consistency of SFA in the analysis of nucleotide sequences, we consider four other viral sequences from a different viral family, that of the influenza viruses or the *Orthomyxoviridae* family [17].

In the Supplementary material, Figures S7, S8, and S9 and Tables ST5 and ST6 correspond to H1N1-1918 and are similar to Figs. 1, 2, and 3 and Tables 1 and 2. Figures S10, S11, and S12 and Table ST7 and ST8 correspond to H1N1-2009 and are again similar to Figs. 2 and 3 and Table 2. Figures S13, S14, and S15 and Tables ST9 and ST10 correspond to H2N2-1957 and are similar to Figs. 2 and 3 and Table 2. The same goes for Figures S16, S17, and S18 and Tables ST11 and ST12, which correspond to H3N2-1968. From these figures and tables, it follows that:

a) Peak SFA periodicities for H1N1-1918

$$\begin{aligned} P_1 &= 60.6275329147499 \\ P_2 &= 157.248365569507 \\ P_3 &= 288.394774029129 \\ P_4 &= 576.789548058258 \\ P_5 &= 970.040526635999 \\ P_6 &= 1631.40720299807 \\ P_7 &= 2515.97384911212 \\ P_8 &= 5031.94769822424 \end{aligned}$$

Exact relationships

$$\begin{aligned} P_4 &= 2P_3 \\ P_5 &= 16P_1 \\ P_7 &= 16P_2 \\ P_8 &= 32P_2 \\ P_8 &= 2P_7 \end{aligned} \tag{10}$$

Almost exact relationships

$$P_8 = 83P_1 \tag{11}$$

Complexity measure, C=6

b) Peak SFA periodicities for H1N1-2009

$$\begin{aligned} P_1 &= 55.5956928123500 \\ P_2 &= 157.248365569507 \\ P_3 &= 288.394774029129 \\ P_4 &= 576.789548058258 \\ P_5 &= 1057.83669529524 \\ P_6 &= 2515.97384911212 \\ P_7 &= 5984.02800504983 \end{aligned}$$

Exact relationships

$$\begin{aligned} P_4 &= 2P_3 \\ P_6 &= 16P_2 \end{aligned} \tag{12}$$

Almost exact relationships (note 38=2×19)

$$\begin{aligned} P_5 &= 19P_1 \\ P_7 &= 38P_2 \end{aligned} \tag{13}$$

Complexity measure, C=4

c) Peak SFA periodicities for H2N2-1957

$$
\begin{aligned}
P_1 &= 72.0986935072823\\
P_2 &= 203.925900374759\\
P_3 &= 288.394774029129\\
P_4 &= 576.789548058258\\
P_5 &= 889.531084997600\\
P_6 &= 2515.97384911212\\
P_7 &= 5984.02800504983
\end{aligned}
$$

Exact relationships

$$
\begin{aligned}
P_3 &= 4P_1\\
P_4 &= 8P_1\\
P_4 &= 2P_3
\end{aligned}
\tag{14}
$$

Almost exact relationships

$$
P_7 = 83P_1 \tag{15}
$$

Complexity measure, C=4

d) Peak SFA periodicities for H3N2-1968

$$
\begin{aligned}
P_1 &= 72.0986935072823\\
P_2 &= 187.000875157807\\
P_3 &= 628.993462278030\\
P_4 &= 889.531084997600\\
P_5 &= 2515.97384911212\\
P_6 &= 5487.37787528068
\end{aligned}
$$

Exact relationships

$$
P_5 = 4P_3 \tag{16}
$$

Almost exact periodicities

$$
P_6 = 76P_1 \text{ (note } 76 = 2 \times 38 = 4 \times 19) \tag{17}
$$

Complexity measure, C=2

Inter-relationships

As in the case of the coronaviruses, the influenza virus sequence analysis also revealed plenty of inter-relationships as expected, since the four viruses belong to the same family.

$$
\begin{gathered}
P_2(\text{H1N1–1918}) = P_2(\text{H1N1–2009})\\
P_3(\text{H1N1–1918}) = P_3(\text{H1N1–2009}) = P_3(\text{H2N2–1957})\\
P_4(\text{H1N1–1918}) = P_4(\text{H1N1–2009}) = P_4(\text{H2N2–1957})\\
P_7(\text{H1N1–1918}) = P_6(\text{H1N1–2009}) = P_6(\text{H2N2–1957}) = P_5(\text{H3N2–1968})\\
P_1(\text{H2N2–1957}) = P_1(\text{H3N2–1968})\\
P_7(\text{H1N1–2009}) = P_7(\text{H2N2–1957})
\end{gathered}
\tag{18}
$$

Interestingly, we found that many relationships exist between the two viral families investigated here. If we compare the results in this section to the previous section, we can infer that:

$$
\begin{gathered}
P_1(\text{SARS–CoV–2}) = P_1(\text{H1N1–2009})\\
P_5(\text{SARS–CoV–2}) = P_4(\text{H1N1–1918}) = P_4(\text{H1N1–2009}) = P_4(\text{H2N2–1957})\\
P_3(\text{SARS–CoV–2}) = P_2(\text{H3N2–1968})\\
P_6(\text{MERS}) = P_7(\text{H1N1–1918}) = P_6(\text{H1N1–2009}) = P_6(\text{H2N2–1957}) = P_5(\text{H3N2–1968})\\
P_4(\text{MERS}) = P_3(\text{H3N2–1968})
\end{gathered}
\tag{19}
$$

More on this is discussed next.

Discussion

If we consider the peak SFA periodicities from a sequence as nodes of a community, and their relationships as links between the nodes, then, a visualization of the results for the SARS-CoV-2 community would look like the top left panel of Fig. 4. Since there are 10 peak periodicities, we have 10 nodes. Then, from Eqs. 3 and 4, we have 13 (recall that C=13) links between them (showing in blue). The rest of the panels correspond to the rest of the sequences in both families. The red lines give the links between the communities within a family (from Eq. 9 for the coronavirus family and from Eq. 18 for the influenza family). The black lines are the links between the two families (Eq. 19). This picture is a perfect example of complex networks, which are often characterized by a community structure, where in each community the nodes are connected in a certain way (meaning the community obeys its own dynamics), but where there exist also some connections (or interactions) between the communities (see for example [18, 19]). We note two interesting observations: (1) the influenza virus family is much more connected (more red links) than the coronavirus family, possibly indicating that the influenza strains are less mutated than the coronavirus strains, and (2) SARS-CoV-1 has no direct links to the influenza family.

This result supports our claims that SFA has the potential and efficiency to delineate the complex mathematical structure of genetic sequences and that it could become a useful tool in such analyses. We need to stress here that, given the mathematics behind SFA, while we can make direct comparisons of the complexity measure "C" within a certain family (where more or less the number of bases is the same), we cannot compare complexities based on "C" between different families. This is due to the differences in nucleotide length between viral families. The coronavirus family sequence length is approximately 30,450 bases, whereas the influenza family sequence length is approximately 13,500 bases. As such, SFA may "see" longer oscillations in the coronavirus family than in the influenza family. Thus, there will be more entries above the diagonal (in tables such as Table 2), and therefore, higher complexity in the coronavirus family.

Finally, it is interesting to note that the complexity measure "C" in the case of the coronavirus family relates to mortality and severity of symptoms as well as to the

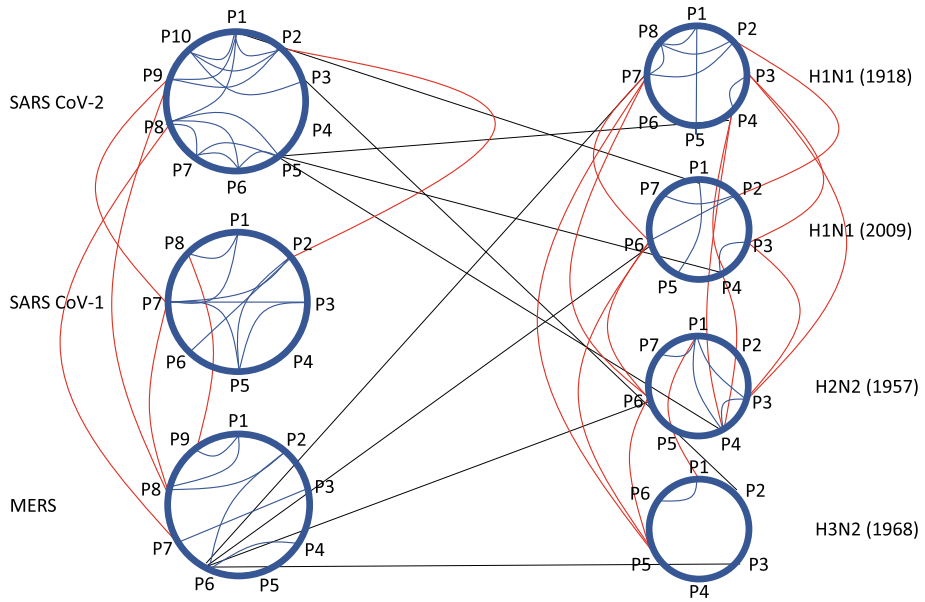

Fig. 4 A complex network visualization of the relationships (connections) between individual nucleotide sequences (blue), between sequences within each individual family (the coronavirus family and the influenza family; red), and between the two families (black) resulted from the SFA. This picture is akin to structures of complex networks where in each community the nodes are connected in a certain way (meaning the community obeys its own dynamics), but where there are also connections between the communities

rate of transmission. As "*C*" increases, the transmission rate to humans increases, but mortality rate decreases. It is reported that symptoms of the SARS-CoV-2 are milder than SARS-CoV-1 and MERS; however, the viral transmission rate (from human-to-human) is greater than the other family members. The mortality rate of SARS-CoV-2 is lower (3.4%) than that of SARS-CoV-1 (9.6%) and MERS (35%) [20]. This relationship is not as clear, however, in the case of the influenza virus family. Unfortunately, in this case, the outbreaks span over a century, and the actual numbers are skewed by several factors such as deaths by secondary infection (due to the unavailability of antibiotics), hygiene, lack of experience and lack of proper healthcare, especially in the early outbreaks, and other problems. For example, H1N1-1918 (C=6) infected 30% of the planet's population and H1N1-2009 (C=4) infected 10% of the population. This is consistent with "increasing C ➔ higher infection rate", but it is not consistent with "increasing C➔ less mortality rate". H1N1-1918 killed about 8% of the infected, whereas H1N1-2009 killed only 0.0025% of the infected [21–28]. But how can we compare the conditions in 1918 and 2009? To complicate comparisons further, there is hardly any reliable data of infection rates for H2N2 and H3N3. In any case, the complexity measure "*C*" may hold a promise and could become a useful tool in the prediction of transmission and mortality rates in future new viral strains.

Conclusions

In this exploratory work, a relatively recent mathematical method (SFA) is applied to probe the structure of the nucleotide sequences of two related viral families: those of coronaviruses and those of influenza viruses. The coronaviruses are SARS-CoV-2, SARS-CoV-1, and MERS. The influenza viruses include H1N1-1918, H1N1-2009, H2N2-1957, and H3N2-1968. The analysis indicates that the nucleotide sequences exhibit an elaborate and convoluted structure akin to complex networks. We define a measure of complexity and show that each nucleotide sequence exhibits a certain degree of complexity within itself, while at the same time there exists complex inter-relationships between the sequences within a family and between the two families. From these relationships, we find evidence, especially for the coronavirus family, that increasing complexity in a sequence is associated with higher transmission rate but with lower mortality. As such, the complexity measure defined here may hold a promise and could become a useful tool in the prediction of transmission and mortality rates in future new viral strains.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40246-021-00327-2.

Additional file 1. Supplementary figures and tables

Additional file 2. Nucleotide sequences used in the study

Acknowledgements
Not applicable

Authors' contributions
AAT designed the research, did some of the analysis, contributed to the interpretation of the results, and wrote the first draft of the paper. GW, LZ, and WL contributed largely to SFA analysis. AK and KDRT contributed to interpretation of the results and writing of the manuscript. All authors have approved the paper for submission. There are no competing interests.

Funding
None

Availability of data and materials
All nucleotide sequences used in this analysis are public domain and can be downloaded from the National Center for Biotechnology Information https://www.ncbi.nlm.nih.gov. For convenience, we have supplied all the nucleotide sequences used here.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Mathematical Sciences, Atmospheric Sciences Group, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA. [2]Hydrologic Research Center, San Diego, CA 92127, USA. [3]Key Laboratory of Middle Atmosphere and Global Environment Observation (LAGEO), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China. [4]Department of Biology and Center for Visual Sciences, Miami University, Oxford, OH 45056, USA.

Received: 3 March 2021 Accepted: 19 April 2021
Published online: 07 May 2021

References

1. Shepherd JCW. From primeval message to present day gene. CSH Symp Quant Biol. 1982;47:1099–108.
2. Ohno S. Codon preference is but an illusion created by the construction principle of coding sequences. Proc Natl Acad Sci USA. 1998;85:4378–82.
3. Yomo T, Ohno S. Concordant evolution and noncoding regions made it possible by the universal rule of TA/CG deficiency-TG/Ct excess. Proc Natl Acad Sci USA. 1989;86:8452–6.
4. Tsonis AA, Elsner JB, Tsonis PA. Periodicity in DNA sequences: Implications in gene evolution. J Theor Biol. 1991;151:323–31.
5. Tsonis AA, Kumar P, Elsner JB, Tsonis PA. Wavelet analysis of DNA sequences. Phys Rev E. 1996;53:1828–34.
6. Lask AM. Introduction to Bioinformatics, 3rd edition, Oxford University press; 2008. p. 474.
7. Pevsner J. Bioinformatics and Functional Genomics, 3rd edition, Willey-Blackwell; 2015. p. 1124.
8. Wiskott L, Sejnowski TJ. Slow Feature Analysis: Unsupervised learning of invariance. Neural Comput. 2002;14:715–70.
9. Wiskott L. Estimating driving forces of nonstationary time series with slow feature analysis, (2003), http://arxiv.org/abs/cond-mat/0312317/
10. Berkes P, Wiskott L. Slow feature analysis yields a rich repertoire of complex cells. J Vis. 2005;5(6):579–602.
11. Blaschke T, Berkes P, Wiskott L. What is the relationship between slow feature analysis and independent component analysis? Neural Comput. 2006;18(10):2495–508. https://doi.org/10.1162/neco.2006.18.10.2495.
12. Franzius M, Wilbert N, Wiskott L. Invariant object recognition and pose estimation with slow feature analysis. Neural Comput. 2011;23(9):2289–323. https://doi.org/10.1162/NECO_a_00171.
13. Yang P, Wang G, Zhang F, Zhou X. Causality of global warming seen from observations: a scale analysis of driving force of the surface air temperature time series in the Northern Hemisphere. Clim Dyn. 2015. https://doi.org/10.1007/s00382-015-2761-4.
14. Tsonis AA, Pan X, Wang G, Nicolis C. On the min-max estimation of mean daily temperatures. Clim Dyn. 2019;53:1981–9. https://doi.org/10.1007/s00382-019-04757-6.
15. Wiskott L, et al. Slow feature analysis. Scholarpedia. 2011;6(4):5282.
16. Torrence C, Compo GP. A practical guide to wavelet analysis. Bull Amer Meteor Soc. 1998;79:61–78. https://doi.org/10.1175/1520-0477.
17. Ng WM, Stelfox AJ, Bowden TA. Unraveling virus relationships by structure-based phylogenetic classification. Virus Evol. 2020;6(1). https://doi.org/10.1093/ve/veaa003.
18. Newman MEJ, Girwan M. Finding and evaluating community structure in networks. Phys Rev E. 2004;69:026113. https://doi.org/10.1103/PhysRevE.69.26113.
19. Newman MEJ. Modularity and community structure in networks. Proc Natl Acad Sci USA. 2006;103:8577–82.
20. Fani M, Teimoori A, Ghafari S. Comparison of the COVID-2019 (SARS-CoV-2) pathogenesis with SRS-CoV and MERS-CoV infections. Future Virol. 2020. https://doi.org/10.2217/fvl-2020-0050.
21. Andreasen V, Viboud C, Simonsen L. Epidemiologic characterization of the 1918 influenza pandemic summer wave in Copenhagen: implications for pandemic control strategies. J Infect Dis. 2008;197(2):270–728.
22. Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli R. Features, evaluation and treatment Coronavirus (COVID-19): StatPearls Publishing; 2020.
23. Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng P-Y, et al. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. Lancet Infect Dis. 2012;12(9):687–95.
24. Hassan SA, Sheikh FN, Jamal S, Ezeh JK, Akhtar A. Coronavirus (COVID-19): a review of clinical features, diagnosis, and treatment. Cureus. 2020;12(3):e7355.
25. Liu J, Xie W, Wang Y, Xiong Y, Chen S, Han J, et al. A comparative overview of COVID-19, MERS and SARS: review article. Int J Surg. 2020;81:1–8.
26. Taubenberger JK. The origin and virulence of the 1918 'Spanish' Influenza Virus. Proc Am Philos Soc. 2006;150(1):86–112.
27. Viboud C, Grais RF, Lafont BAP, Miller MA, Simonsen L, Multinational Influenza Seasonal Mortality Study Group. Multinational impact of the 1968 Hong Kong influenza pandemic: evidence for a smoldering pandemic. J Infect Dis. 2005;192(2):233–48.
28. Viboud C, Simonsen L, Fuentes R, Flores J, Miller MA, Chowell G. Global mortality impact of the 1957-1959 influenza pandemic. J Infect Dis. 2016;213(5):738–45.

Publisher's Note