Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2010, Article ID 507056, 31 pages
doi:10.1155/2010/507056

Research Article

Fractals and Hidden Symmetries in DNA

Carlo Cattani

Department of Pharmaceutical Sciences, University of Salerno, Via Ponte Don Melillo, 84084 Fisciano, Italy

Correspondence should be addressed to Carlo Cattani, ccattani@unisa.it

Received 27 January 2010; Accepted 8 March 2010

Academic Editor: Cristian Toma

This paper deals with the digital complex representation of a DNA sequence and the analysis of existing correlations by wavelets. The symbolic DNA sequence is mapped into a nonlinear time series. By studying this time series the existence of fractal shapes and symmetries will be shown. At first step, the indicator matrix enables us to recognize some typical patterns of nucleotide distribution. The DNA sequence, of the influenza virus A (H1N1), is investigated by using the complex representation, together with the corresponding walks on DNA; in particular, it is shown that DNA walks are fractals. Finally, by using the wavelet analysis, the existence of symmetries is proven.

1. Introduction

The main task of this paper is to show the existence of hidden geometries which underly the structure of a DNA sequence. Moreover, it will be shown that this geometry is fractal. In order to achieve this goal the fundamental steps are

(1) the choice of the digital representation of the symbolic sequence of DNA,

(2) the definition of the indicator matrix,

(3) the construction of walks on DNA,

(4) the cluster analysis of wavelet coefficients.

In this paper it will be shown that the distribution of nucleotides A, C, G, T along the sequence must fulfill some hidden geometrical rules, thus implying that the biological activity depends on these geometrical rules. The understanding of the underlying biological function from a possible interpretation of the given sequence of nucleotides [1–6] is still under investigation.

The existence of hidden law, periodicities, and statistical correlations [2, 7–9] might help us to characterize each DNA sequence in order to construct a possible (functional) classification.

From mathematical point of view the DNA sequence is a symbolic sequence (of nucleotides) with some empty spaces (no coding regions). In order to get some numerical information from this sequence it must be transformed into a digital sequence. When the symbolic sequence of A, C, G, T is digitalized into one or more sequences of digits one may benefit from the statistical analysis of the digitalized time series, so that the genome can be characterized by the classical statistical parameters like variance, deviation, or nonclassical like complexity, fractal dimension, or long range dependence.

There follows that the symbolic sequence is transformed into a very large time series (from half million of digits, for the primitive organisms such as fungus and eukaryotes and, to several millions, as for mammals, like the nearly 1.5 billion of nucleotides for the humans DNA). However, these large sequences look like some random sequences, from where it seems to be quite impossible to single out any single correlation (see, e.g., [8] and references therein).

In any case the arbitray choice of the representative digital time series (discrete time signal) for the symbolic sequence of the genome, so that the representation would be the most suitable for the statistical-mathematical analysis, is a difficult task, that was approached with some interesting preliminary results by using a complex representation [10, 11].

The easiest mathematical model for the transformation of a symbolic string into a numerical string is based on the Voss indicator function [12, 13] which is a discrete binary function. In the following a suitable generalization is given and it will be shown that the graphical representation gives rise to some featuring patterns. The existence of patterns and symmetries is shown also through the cluster analysis of the wavelet coefficients [14, 15].

The analysis of DNA by wavelets [7, 9, 16], as seen in [9, 16–19], is an expedient tool to single out local behavior and to characterize singularities as local spikes and jumps [7, 14] or to express the scale invariance of coefficients [20] and thus the multifractal nature of the time series [21–23]. However, as shown below, the wavelet transform features also a decorrelation of the sequence, so that it allows the emergence of the basic rules of the uncorrelated sequence. We will see that the wavelet coefficients of the short Haar wavelet transform are quantized. This can be achieved by a decomposition of the sequence into short segments of equal length and by a wavelet transform to be applied to each segment.

The long range-correlation in the digital representation of the DNA sequence [12, 13, 24–34] is a fundamental problem in DNA analysis. Correlation in a digital signal can be roughly linked with the concept of dependence, in a statistical sense, of elements which are far away from each other. The existence of correlation in DNA has been explained with the so-called process of duplication-mutation. According to [29, 35] in the evolutionary model the actual DNA sequence results from an original short-length chain that was duplicating and modifying some pieces of the sequence. Due to this there followed the characterizing $1/f$ power law decay [13, 26, 27]. The power law for long-range correlations is a measure of the scaling law, showing the existence of self-similar structures similar to the physics of fractals. The long-range correlation, which can be detected by the autocorrelation function, implies the scale independence (scale invariance) which is typical of fractals.

The power law for long-range correlations is a measure of the scaling law, showing the existence of self-similar structures similar to the physics of fractals. The long-range correlation, which can be detected by the autocorrelation function [12, 13, 30–32, 36, 37], implies the scale independence (scale invariance) which is typical of fractals. However, the

preliminary results in this topics were disputed [26, 27, 38] because of the limited number of available data and because of different approaches to this analysis. On the other hand the existence of patchiness and correlation would imply some important understanding of DNA organization. Therefore, in the following we will discuss the correlation of a DNA virus sequence, roughly 2000 base pairs (bp), with the undertsanding that a well-defined concept of correlation holds only on a long sequence of the order higher than 10^5. However, the most important are the fractal properties and symmetries which are well defined even for a short sequences. The identification (and classifications) of these patches could be the key point for understanding the large-scale structure of DNA.

Due to the recent outbreak of 2009 H1N1 Flu Virus (Swine Flu) in humans, this paper will focus on the DNA influenza virus (for similar results on a mammalian and a fungus see [10, 11, 20]) with particular regards to the A (H1N1) variant provided by the National Center for Biotechnology Information [3–6]:

(1) influenza A virus (A/Puerto Rico/8/34(H1N1)) segment 1, complete sequence, 2341 bp,

(2) influenza A virus (A/Mexico City/MCIG01/2009(H1N1)) segment 4 sequence, submitted 13-Jan-2010, 1719 bp,

(3) influenza A virus (A/Nagasaki/HA-46/2009(H1N1)) segment 4, HA gene for hemagglutinin, complete cds, submitted 07-Jan-2010, 1701 bp,

(4) influenza A virus (A/Tver/IIV2969/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds, submitted 13-Jan-2010, 1744 bp,

(5) influenza A virus (A/Novosibirsk/02/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds, submitted 10-Jan-2010, 1752 bp,

(6) influenza A virus (A/Rio Grande do Sul/7108/2009(H1N1)) segment 4 sequence, submitted 08-Jan 2010, 1701 bp,

(7) influenza A virus (A/Rio de Janeiro/5090/2009(H1N1)) segment 4 sequence, submitted 08-Jan 2010, 1701 bp,

(8) influenza A virus (A/Ankara/03/2009(H1N1)) segment 4 hemagglutinin (HA) gene, complete cds, submitted 04-Jan-2010, 1701 bp.

It will be shown that as the dog and candida DNA the influenza virus is characterized by DNA walks with fractal shape. The most amazing is that the same symmetries seen on the wavelet coefficients of the DNA walks for dog and candida holds true also for viruses. The fractal dimension of the indicator matrix for this virus is 2.03 much higher than dog's [11]; moreover some difference in the DNA walks for segment 4 will be highlighted.

This paper is organized as follows. Section 2 deals with some preliminary remarks on flu epidemiology; DNA and DNA representation together with the indicator matrix is given in Section 3. Here the global fractal estimate is computed and the existence of fractal patterns is shown. The complex (cardinal) representation is given in Section 4 and the DNA (complex) walks are analysed in Section 5. It is proven that DNA complex walks are fractals and they are compared with walks on pseudorandom and deterministic complex sequences. Section 6 deals with correlation, power spectrum, and complexity of DNA. Sections 7 and 8 deal with wavelets analysis and show the existence of simmetries in the wavelet coefficients.

2. Flu Epidemiology

Flu epidemics cause morbidity and mortality worldwide. Each year, only in the USA, more than 200000 patients are infected by influenza and there are approximately 36000 deaths due to influenza virus. Of the three types of influenza virus—A, B and C—the A and B types can cause flu epidemics. Influenza A virus is found not only in humans but also in many other animals. There are over hundreds of subtypes of Influenza A virus. All subtypes have been detected in wild birds, which are considered the source of influenza A viruses in all other animals. For example, pigs may be infected with influenza A viruses from different species (e.g., ducks and humans) at the same time, which may allow the genes of these viruses to mix, creating new variants of the hemagglutinin and/or neuraminidase proteins on the surface of the virus (antigenic shift). If these variants spread to humans, then they would not be recognized by the immune system and so can cause seasonal epidemics of flu. In addition, influenza viruses undergo mutations when they spread from place to place and therefore introduce gradual changes in the hemagglutinin and/or neuraminidase proteins (antigenic drift). It will be shown however, that even if there are some variants of the same virus at different places, still the DNA structure remains the same (at least in the indicator matrix, see below), without significant variations. In other words the DNA sequence might apparently show some differences, but when we pass to the digital representation and the indicator matrix, these differences vanish.

Each year, it is essential to identify new flu virus variants and produce vaccines against them to avoid flu epidemics. Therefore the investigation of DNA sequence of variants might help to better understand the intrinsic nature of variation.

The Centers for Disease Control and Prevention (CDC) and other health organizations are actively investigating the recent outbreak of 2009 H1N1 Flu Virus (Swine Flu) in humans. First cases were reported at the beginning of 2009. CDC has determined that this swine influenza A (H1N1) virus is contagious and is spreading from human to human. Swine Influenza is a respiratory disease of pigs (swine) caused by type A influenza virus that regularly causes outbreaks of flu in pigs. Like all influenza viruses, swine flu viruses change constantly. Pigs can be infected by avian influenza and human influenza viruses as well as swine influenza viruses. When influenza viruses from different species infect pigs, the viruses can reassort (i.e., swap genes) and new viruses that are a mix of swine, human, and/or avian influenza viruses can emerge. There are four main influenza type A virus subtypes that have been isolated in pigs: H1N1, H1N2, H3N2, and H3N1, but most of the recently isolated influenza viruses from pigs have been H1N1 viruses. While swine flu viruses do not normally infect humans, sporadic human infections with swine flu have occurred. Most commonly, these cases occur in persons with direct exposure to pigs; human-to-human transmission of swine flu can also occur, as is the case with the 2009 outbreak.

An influenza A virion is composed of the nucleocapsid, a surrounding layer of the matrix protein (M1) and the membrane envelope. The envelope contains two major surface glycoproteins, that is, hemagglutinin (HA) and neuraminidase (NA), and a minor membrane protein M2. The nucleocapsid consists of individual ribonucleoproteins (vRNPs). Each vRNP contains one of the 8 genomic negative sense RNA segments (vRNA), multiple copies of the major structural protein NP, and a few copies of the RNA dependent-RNA-polymerase complex. All 8 vRNA species must be present in an infectious virion.

A virion attaches to the host cell membrane via HA and enters the cytoplasm by receptor-mediated endocytosis, thereby forming an endosome. A cellular trypsin-like enzyme cleaves HA into products HA1 and HA2. HA2 promotes fusion of the virus

envelope and the endosome membranes. A minor virus envelope protein M2 acts as an ion channel thereby making the inside of the virion more acidic. As a result, the major envelope protein M1 dissociates from the nucleocapsid and vRNPs are translocated into the nucleus via interaction between NP and cellular transport machinery. In the nucleus, the viral polymerase complexes transcribe and replicate the vRNAs. Newly synthesized mRNAs migrate to cytoplasm where they are translated. Posttranslational processing of HA, NA, and M2 includes transportation via Golgi apparatus to the cell membrane. NP, M1, NS1 (nonstructural regulatory protein), and NEP (nuclear export protein, a minor virion component) move to the nucleus, where bind freshly synthesized copies of vRNAs. The newly formed nucleocapsids migrate into the cytoplasm in an NEP-dependent process and eventually interact via M1 with a region of the cell membrane, where HA, NA, and M2 have been inserted. Then the newly synthesized virions bud from infected cell. NA destroys the sialic acid moiety of cellular receptors, thereby releasing the progeny virions.

3. Patterns on the Indicator Matrix

The DNA of each organism of a given species is a long sequence of a specific (large) number of base pairs (bp). The size of the DNA might range from 10^5 to 10^9 number of base pairs. Each base pair is defined on the 4 elements alphabet of nucleotides:

$$A = \text{adenine}, \quad C = \text{cytosine}, \quad G = \text{guanine}, \quad T = \text{thymine}. \tag{3.1}$$

Let

$$\mathcal{A} \stackrel{\text{def}}{=} \{A, C, G, T\} \tag{3.2}$$

be the finite set (alphabet) of nucleotides and $x \in \mathcal{A}$ any member of the alphabet.
A DNA sequence is the finite symbolic sequence

$$\mathcal{S} = \mathbb{N} \times \mathcal{A} \tag{3.3}$$

so that

$$\mathcal{S} \stackrel{\text{def}}{=} \{x_h\}_{h=1,\ldots,N}, \quad N < \infty \tag{3.4}$$

being

$$x_h \stackrel{\text{def}}{=} (h, x) = x(h), \quad (h = 1, 2, \ldots, N;\ x \in \mathcal{A}) \tag{3.5}$$

the value x at the position h.

3.1. Indicator Matrix

The 2D indicator function, based on the 1D-definition given in [12], is the map

$$u : \mathcal{S} \times \mathcal{S} \longrightarrow \{0, 1\} \tag{3.6}$$

such that

$$u(x_h, x_k) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } x_h = x_k, \\ 0 & \text{if } x_h \neq x_k, \end{cases} \quad (x_h \in \mathcal{S},\ x_k \in \mathcal{S}), \tag{3.7}$$

being

$$u(x_h, x_k) = u(x_k, x_h), \quad u(x_h, x_h) = 1. \tag{3.8}$$

According to (3.7), the indicator of an N-length sequence can be easily represented by the $N \times N$ sparse symmetric matrix of binary values $\{0, 1\}$ which results from the *indicator matrix*

$$u_{hk} \stackrel{\text{def}}{=} u_{x_h}(x_k), \quad (x_h \in \mathcal{S},\ x_k \in \mathcal{S};\ h, k = 0, \ldots, N-1), \tag{3.9}$$

being, explicitly,

$$\begin{array}{c|cccccccccc} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \iddots \\ G & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ C & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ A & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & \cdots \\ A & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & \cdots \\ T & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ A & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & \cdots \\ C & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ T & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ G & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ A & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & \cdots \\ \hline u_{hk} & A & G & T & C & A & T & A & A & C & G & \cdots \end{array} \tag{3.10}$$

This squared matrix can be plotted in 2 dimensions by putting a black dot where (Figure 1) $u_{hk} = 1$ and white spot when $u_{hk} = 0$.

3.2. Indicator Matrix for the Influenza Virus A H1N1

The data, under investigation, refer to influenza virus A (H1N1) segment 1 (Puerto Rico) and segment 4 recently submitted (January 2010) in different world regions: Russia

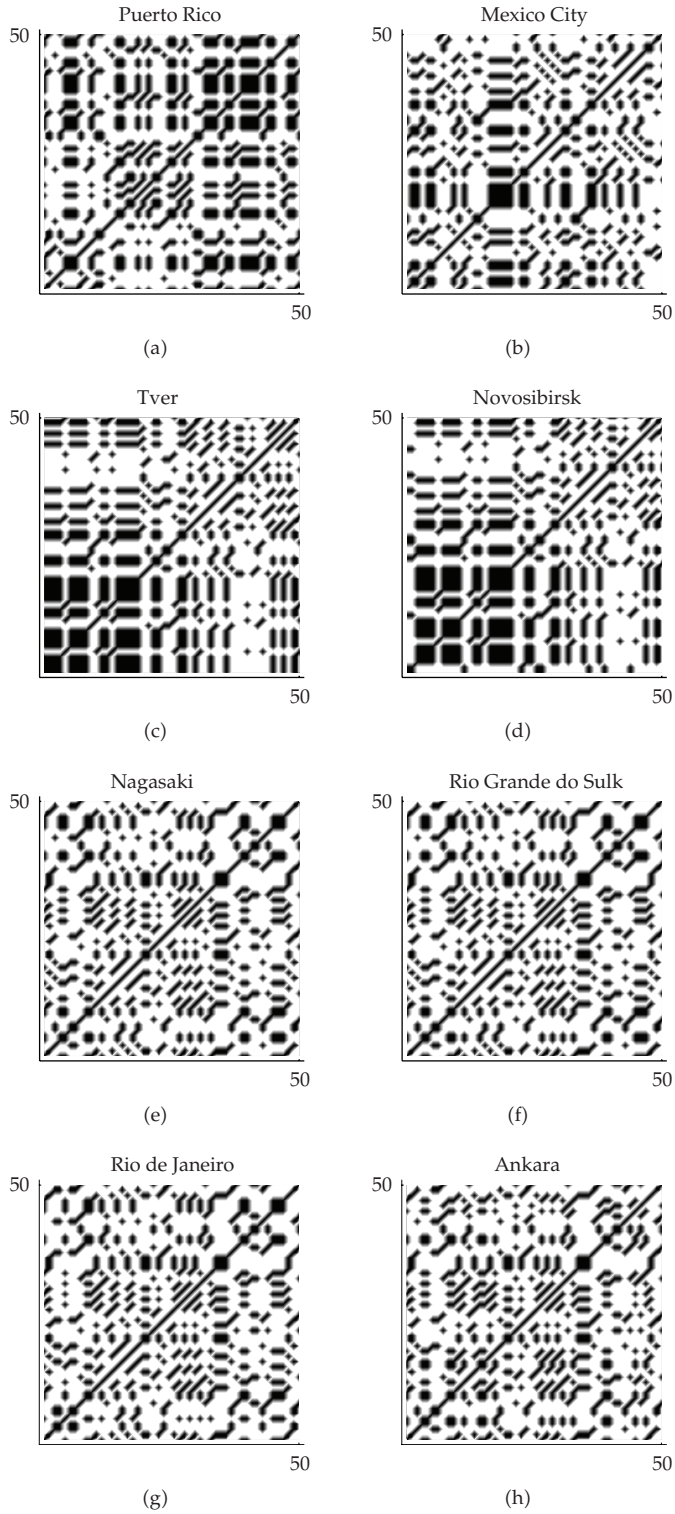

Figure 1: Indicator matrix Influenza A virus (H1N1).

(Tver, Novosibirsk), Japan (Nagasaki), America (Mexico City, Rio Grande do Sul, Rio de Janeiro) and Turky (Ankara).

The plots of indicator matrix (Figure 1) show that

(1) there are some motifs which are repeated at different scales like in a fractal;

(2) empty spaces are more distributed than filled spaces, in the sense that the matrix u_{hk} is a sparse matrix (having more zeroes than ones);

(3) it seems that there are some square-like islands where black spots are more concentrated;

(4) some indicator matrix can be grouped into different sets like, (a) Tver, Novosibirsk; (b) Nagasaki, Rio GRande do Sul, Rio De Janeiro, Ankara; and (c) Mexico City.

From the analyis of Figure 1 we can also notice that even if there is a big distance among different places, the corresponding virus does not change (Nagasaki, Rio do Janeiro and Ankara are nearly the same).

3.3. *Fractal Dimension*

From the indicator matrix we can have an idea of the "fractal-like" distribution of nucleotides. The fractal dimension for the graphical representation of the indicator matrix plots can be computed as the average of the number $p(n)$ of "1" in the randomly taken $n \times n$ minors of the $N \times N$ correlation matrix u_{hk}:

$$D = \frac{1}{N} \sum_{n=2}^{N} \frac{\log p(n)}{\log n}. \tag{3.11}$$

The fractal dimension of the influenza virus A (H1N1) is 2.30±0.1 while that for the dog DNA and candida was 1.66 ± 0.01 (see, e.g., [11]). However some interesting coinciding values can be observed in the following table where the fractal dimension up to 10^{-2} shows the same groups already seen in the matrix shapes of Figure 1:

Puerto Rico			2.322
Mexico City		2.307	
Nagasaki		2.302	
Tver			2.328
Novosibirsk			2.323
Rio Grande do Sul		2.302	
Rio de Janeiro	2.299		
Ankara		2.307	

(3.12)

4. Complex Representation

The (digital) representation of a DNA sequence is defined as the map of $\mathcal{S}$ into $\mathbb{R}^{\ell}$, $\ell \geq 1$. The embedding space of representation is based on the 4 vectors

$$\mathbf{X}_x : \mathcal{A} \longrightarrow \mathbb{R}^{\ell}, \quad (x \in \mathcal{A},\ \ell \geq 1) \tag{4.1}$$

in the real space $\mathbb{R}^{\ell}$, or almost equivalently in the complex space $\mathbb{C}^{\ell}$,

$$\mathbf{X}_x : \mathcal{A} \longrightarrow \mathbb{C}^{\ell}, \quad (x \in \mathcal{A},\ \ell \geq 1) \tag{4.2}$$

so that $\mathbf{X}(x) \equiv \mathbf{X}_x$ is a ℓ-ple which is associated with the symbol $x \in \mathcal{A}$.

There follows that the basic elements of the representation are

$$\begin{aligned} \mathbf{X}_A &= (x_{A1}, x_{A2}, \ldots, x_{A\ell}), \\ \mathbf{X}_C &= (x_{C1}, x_{C2}, \ldots, x_{C\ell}), \\ \mathbf{X}_G &= (x_{G1}, x_{G2}, \ldots, x_{G\ell}), \\ \mathbf{X}_T &= (x_{T1}, x_{T2}, \ldots, x_{T\ell}) \end{aligned} \quad (\ell \geq 1). \tag{4.3}$$

The digital representation in $\mathbb{R}^{\ell}$ or $\mathbb{C}^{\ell}$ of a N-length DNA sequence is the map $\mathcal{R} : \mathbb{N} \times \mathcal{A} \to \mathbb{R}^{\ell}$ or, for a complex representation, $\mathcal{R} : \mathbb{N} \times \mathcal{A} \to \mathbb{C}^{\ell}$ so that for each $x_n \in \mathcal{S}$ it is

$$x_n \stackrel{\mathcal{R}}{\to} \mathbf{Y}(n), \quad \left(x_n \in \mathcal{S};\ \mathbf{Y}(n) \in \mathbb{R}^{\ell}\right), \tag{4.4}$$

being

$$\mathbf{Y}(n) \stackrel{\text{def}}{=} \mathbf{Y}(x_n) \stackrel{(3.5)}{=} \mathbf{Y}(x(n)) \tag{4.5}$$

defined as follows. Each element of the DNA sequence can be considered [39] as the linear combination:

$$\mathbf{Y}(n) \stackrel{\text{def}}{=} u(A, x_n)\mathbf{X}_A + u(C, x_n)\mathbf{X}_C + u(G, x_n)\mathbf{X}_G + u(T, x_n)\mathbf{X}_T, \quad (n = 1, \ldots, N). \tag{4.6}$$

The graph of $\mathbf{Y}_n = \mathbf{Y}(n)$ is $\mathcal{G}$ and if we define

$$a_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} u(A, x_i), \quad c_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} u(C, x_i), \quad g_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} u(G, x_i), \quad t_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} u(T, x_i) \tag{4.7}$$

it is

$$a_n + c_n + g_n + t_n = n, \tag{4.8}$$

so that, as a consequence of (4.6) and the definition (4.7), the following identity holds:

$$n\mathbf{Y}(n) = a_n\mathbf{X}_A + c_n\mathbf{X}_C + g_n\mathbf{X}_G + t_n\mathbf{X}_T. \tag{4.9}$$

We have a degeneracy (or a loop, circuit, or periodicity) if it is (see, e.g., [40]) $\mathbf{Y}(n) = 0$ or, equivalently,

$$a_n \mathbf{X}_A + c_n \mathbf{X}_C + g_n \mathbf{X}_G + t_n \mathbf{X}_T = 0. \tag{4.10}$$

If, we do not have a degeneracy, then there is a one-to-one correspondence between the DNA sequence $\mathcal{S}$ and $\mathcal{G}$, that is, $\mathcal{S} \leftrightarrow \mathcal{G}$.

4.1. Cardinal Complex Representation

In the remaining part of the paper we consider the cardinal representation [11] in $\mathbb{C}^1$, so that the DNA digital representation is the N-length one-dimensional complex signal $\{Y_n\}_{n=0,\ldots,N-1}$. In this case, from (4.6) we have

$$\begin{aligned} Y_n &= u(A, x_n) - u(C, x_n)i - u(G, x_n) + u(T, x_n)i \\ &= [u(A, x_n) - u(G, x_n)] + [u(T, x_n) - u(C, x_n)]i \end{aligned} \tag{4.11}$$

or

$$Y_n = \xi_n + \eta_n i, \quad |\xi_n| + |\eta_n| = 1, \ \xi_n \eta_n = 0 \tag{4.12}$$

with

$$\xi_n \stackrel{\text{def}}{=} u(A, x_n) - u(G, x_n), \qquad \eta_n \stackrel{\text{def}}{=} u(T, x_n) - u(C, x_n), \tag{4.13}$$

so that the representation is a map $\mathcal{S} \to \mathbb{C}^1$ and the time series $\mathbf{Y}(n)$ is a sequence of complex numbers:

$$\{Y_n\}_{n=0,\ldots,N-1}, \qquad Y_n = \xi_n + \eta_n i. \tag{4.14}$$

5. DNA Walks

DNA walk is defined as the series

$$\sum Y_n, \quad n = 0, \ldots, N-1, \tag{5.1}$$

which is the cumulative sum on the DNA sequence representation:

$$\left\{ Y_0, Y_0 + Y_1, \ldots, \sum_{s=0}^{n-1} Y_s, \ldots, \sum_{s=0}^{N-1} Y_s \right\}. \tag{5.2}$$

Taking into account (4.7), (4.11), for the complex cardinal representation, it is

$$\begin{aligned} Z_n &\stackrel{\text{def}}{=} \sum_{s=0}^{n-1} Y_s = \sum_{s=0}^{n-1} \{[u(A,x_s) - u(G,x_s)] + [u(T,x_s) - u(C,x_s)]i\} \\ &= (a_n - g_n) + (t_n - c_n)i, \end{aligned} \tag{5.3}$$

so that the DNA walk is the complex values signal $\{Z_n\}_{n=0,\ldots,N-1}$ with

$$Z_n = (a_n - g_n) + (t_n - c_n)i, \tag{5.4}$$

where the coefficients a_n, g_n, t_n, c_n given by (4.7) and fulfil condition (4.8).

The DNA walk (DNA series) on a complex cardinal representation is a complex series as well. If we map the points

$$P_n = (\Re[Z_n], \Im[Z_n]) = (a_n - g_n, t_n - c_n), \quad n = 0, \ldots, N-1 \tag{5.5}$$

whose coordinates are the real and the imaginary coefficients of each term of the DNA walk sequence, we obtain a cluster showing the existence of some patches or some kind of self-similarity (Figures 2 and 3).

Both figures for the influenza virus (Figures 2 and 3) show that there exists a fractal behavior of the random walk on DNA sequence. Moreover, focussing on some segments of the DNA walks it can be seen that there are some featuring patterns (see, e.g., Figure 3, with respect to the base pairs between 200 and 500).

Let us now compare the DNA walks with walks on pseudorandom and deterministic sequences.

A pseudorandom (white noise) complex sequence similar to the cardinal complex representation (4.11) can be defined as follows:

$$R_n \stackrel{\text{def}}{=} (-1)^{r_n} i^{s_n} \tag{5.6}$$

with r_n, s_2 being random integers and it looks like

$$\{-1, i, 1, 1, -i, i, -1, -i, i, 1, -i, i, -1, i, 1, -i, -i, -i, -1, \ldots\}. \tag{5.7}$$

Its random walk is

$$\sum (-1)^{r_n} i^{s_n}. \tag{5.8}$$

A deterministic walk can be (Figure 4(c))

$$\sum z_n \stackrel{\text{def}}{=} \sum (-i)^n \sin \frac{\pi}{n} + 1 \tag{5.9}$$

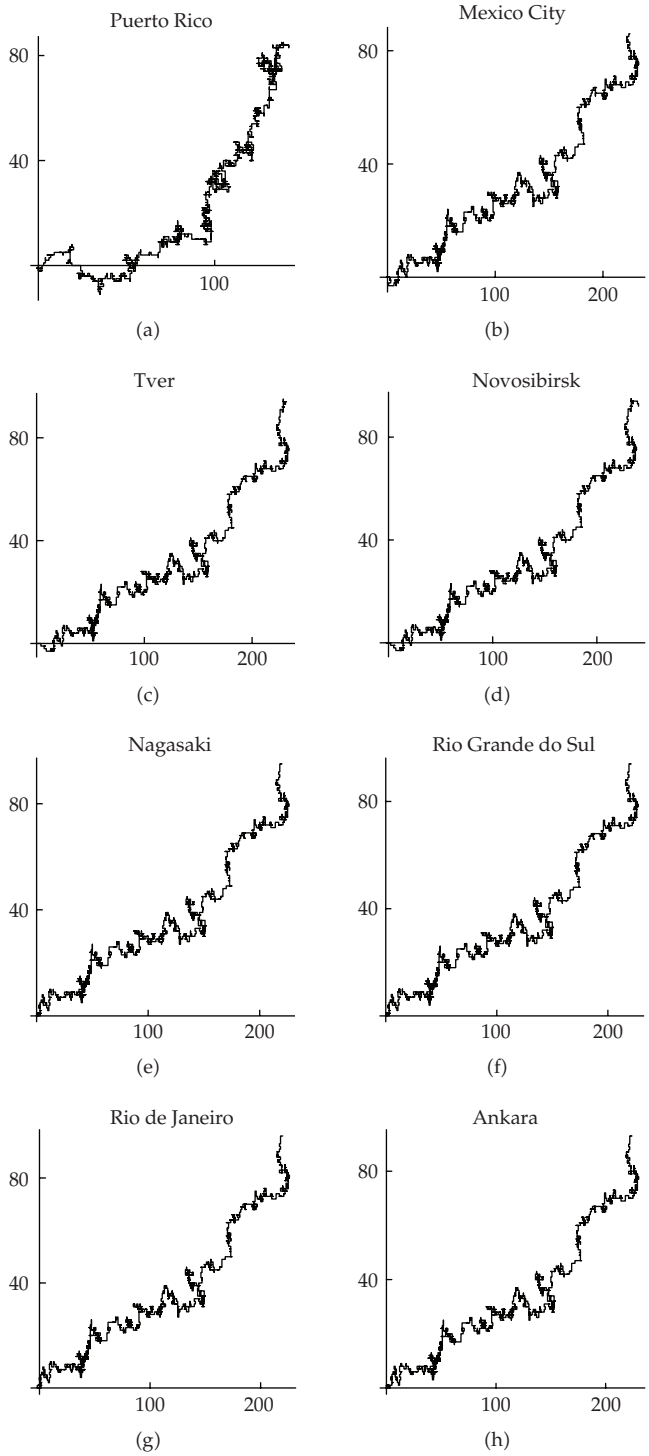

Figure 2: DNA walk of the influenza virus A.

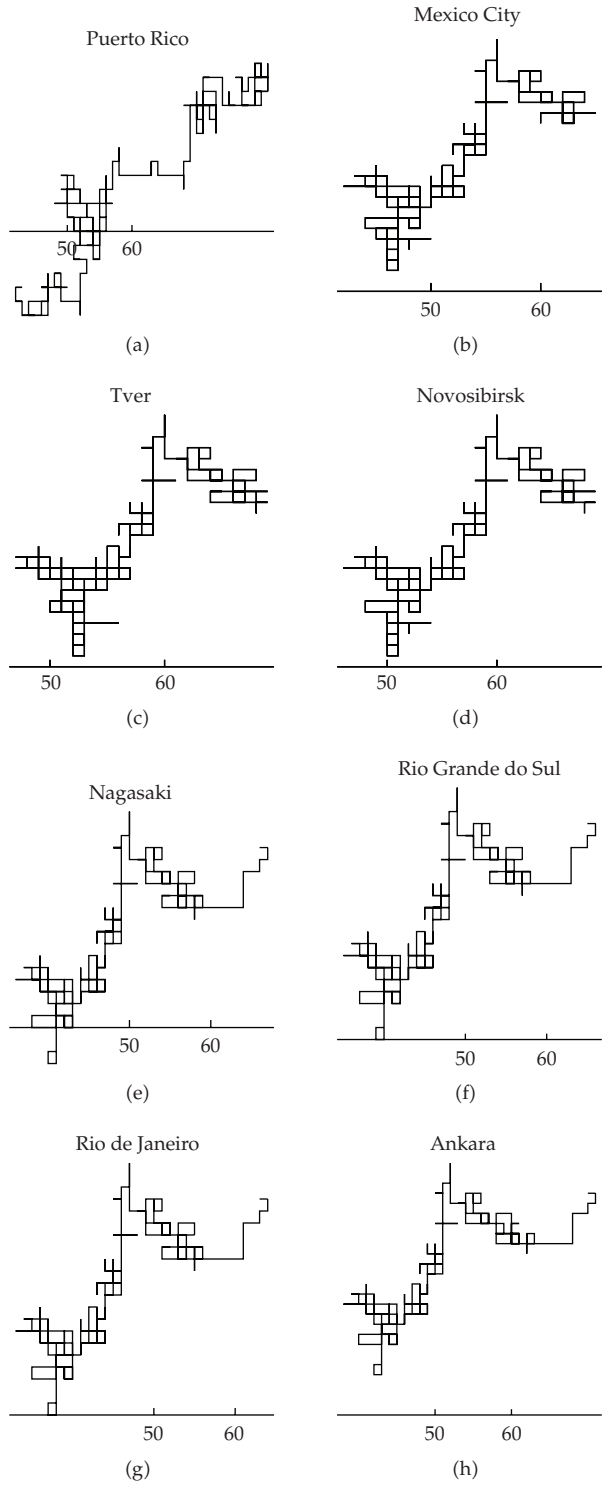

Figure 3: DNA walk of the influenza virus A, on the bp of the interval (200,500).

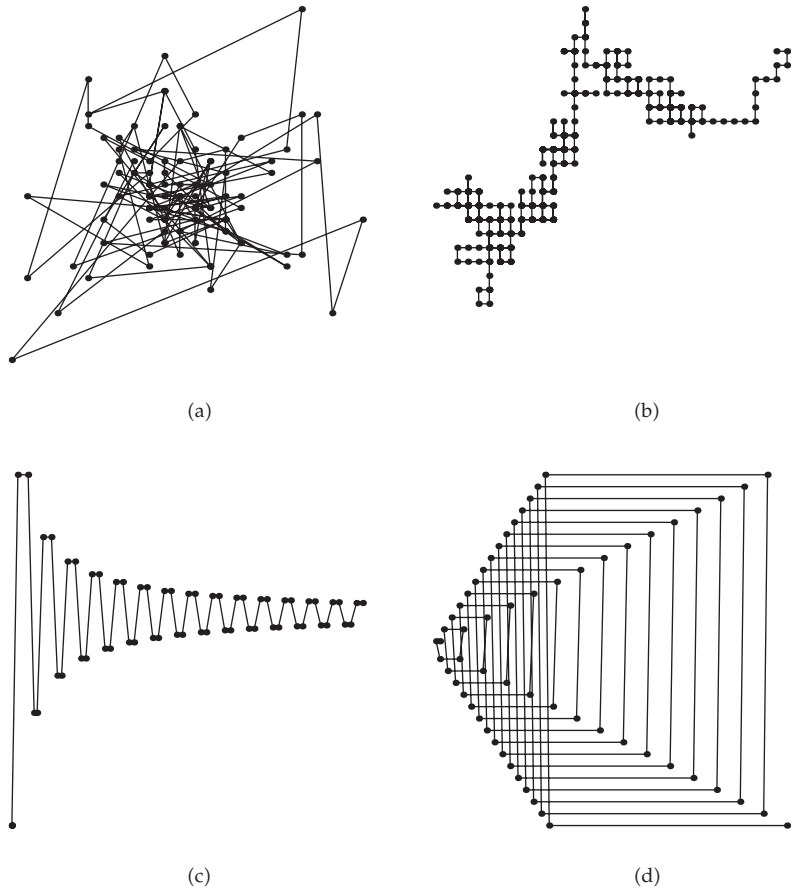

Figure 4: Walks on random (a), flu DNA (b), and deterministic sequences (c, d).

or, more similar to a DNA sequence (Figure 4(d)),

$$\sum z_n \stackrel{\text{def}}{=} \sum i^n n + 1. \tag{5.10}$$

It can be seen (Figure 4) how the fractal shape of DNA walk is completely different from corresponding walks on random and deterministic sequences.

6. Statistical Correlations in DNA

For a given sequence $\{Y_0, Y_1, \ldots, Y_{N-1}\}$ the variance is

$$\sigma^2 \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=0}^{N-1} Y_i^2 - \left(\frac{1}{N}\sum_{i=0}^{N-1} Y_i\right)^2, \tag{6.1}$$

and the variance at the distance $N-k$ is

$$\sigma_k^2 \stackrel{\text{def}}{=} \frac{1}{N-k}\sum_{i=0}^{N-k-1} Y_i^2 - \left(\frac{1}{N-k}\sum_{i=0}^{N-k-1} Y_i\right)^2. \tag{6.2}$$

From the variance follows immediately the standard deviation

$$\sigma = \sqrt{\sigma^2}. \tag{6.3}$$

The autocorrelation at the distance k, $(k = 0, \ldots, N-1)$ is the sequence (see, e.g., [36])

$$c_k \stackrel{\text{def}}{=} \frac{1}{\sigma^2}\left(\frac{1}{N-k}\sum_{i=0}^{N-k-1} Y_i Y_{i+k} - \frac{1}{(N-k)^2}\sum_{i=0}^{N-k-1} Y_i \sum_{i=0}^{N-k-1} Y_{i+k}\right) \tag{6.4}$$

with $k = 0, \ldots, N-1$.

A simplified definition of correlation, in the fragment $F-N$ has been given [41] as follows:

$$c_k \stackrel{\text{def}}{=} \sum_{i=F}^{N-1-k} \frac{1}{N-F-k} u_{x_i}(x_{i+1+k}) \tag{6.5}$$

with the indicator given by (3.7).

The power spectrum can be computed as the Fourier transform of c_k:

$$S_k \stackrel{\text{def}}{=} \widehat{c}_k = \sum_{n=0}^{N-1} c_n e^{-2\pi i n k/N}. \tag{6.6}$$

If $c_k = 0$, there is no linear correlation, $c_k > 0$ means that there is a strong (linear) correlation (anticorrelation when $c_k < 0$), while $c_0 = 1$ does not give any information about correlations. A true random process has a vanishing correlation $c_h = \delta_{0h}$ and its power spectrum S_h is constant. Its integral gives the Brownian motion (random walk) whose power spectrum is proportional to $1/k^2$.

It has been shown [24, 34, 42] that correlations in DNA are linear. However, the main problem of this measure is that it strongly depends on the representation, on the length of the sequence, and, for nonbinary representation, it is affected by spurious results [36]. Moreover, the definition (6.4) holds only for real values of the representation.

6.1. Power Spectrum

Let $\{Y_n\}_{n=0,\ldots,N-1}$ be a given series; the discrete Fourier is the sequence

$$\widehat{Y}_s = \frac{1}{N}\sum_{n=0}^{N-1} Y_n e^{-2\pi i n s/N}, \quad s = 0, \ldots, N-1. \tag{6.7}$$

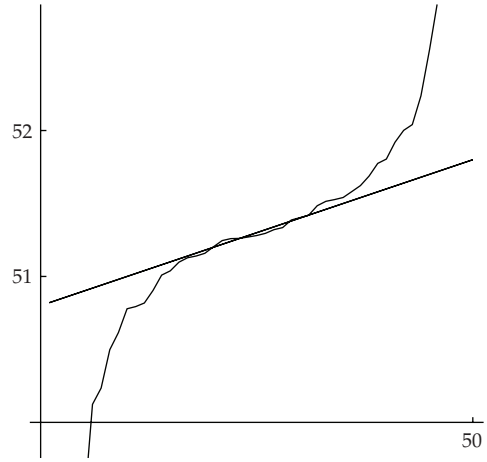

Figure 5: Power spectrum for influenza virus A DNA walk.

The power spectrum of the sequence $\{Y_n\}_{n=0,\ldots,N-1}$, that is, the mean square fluctuation, is defined as [43]

$$S_k \stackrel{\text{def}}{=} \sum_{s=0}^{k-1} \left|\widehat{Y}_s\right|^2. \tag{6.8}$$

The power spectrum of a stationary sequence gives an indirect measure of the autocorrelation. A long-range correlation can be detected if the fluctuations can be described by a power law so that

$$S_k \cong \alpha \frac{k}{\max_{1\le k\le k_{\max}}[\alpha k]}, \quad 1 \le k \le k_{\max} \tag{6.9}$$

with $\alpha > 1/2$.

The fluctuation exponent α, with its values, characterizes a sequence as

(1) anticorrelated: $\alpha < 1/2$,

(2) uncorrelated (white noise): $\alpha \cong 1/2$,

(3) correlated (long range correlated): $\alpha > 1/2$,

(4) $1/f$ noise: $\alpha \cong 1$,

(5) nonstationary, random-walk like: $\alpha > 1$

(6) Brownian noise: $\alpha \cong 3/2$.

For the human DNA there was observed [44] a long range correlation, only for coding regions, with $\alpha = 0.61$. However, the same value can be seen also for dog's and candida DNA [11] for the complete sequence (coding and noncoding regions), even if, by including the noncoding regions, this value is a little bit higher being $\alpha \cong 0.65$ for the dog's, and $\alpha \cong 0.62$ for the candida's DNA, respectively. For the Influenza virus A it is instead $\alpha = 0.02$ (Figure 5).

When the power spectrum is a power-law function

$$S(k) = S_k \propto \frac{1}{k^{\alpha}} \tag{6.10}$$

then this function is scale invariant (like fractals), that is, $f(\lambda x) = \lambda^H f(x)$. It can be shown (see, e.g., [29]) that for the power-law functional dependence, in the $N \to \infty$ limit, it is

$$S(\lambda k) \propto \lambda^{1-a} S(k), \quad S(k) = \frac{1}{k^b} \tag{6.11}$$

with $k \simeq 1 - a$, and a being related to the so-called Hurst exponent. In other words, for a power-law function the power spectrum is scale-invariant (like a fractal).

In particular, from Equation (6.11) there follows that when $b \simeq 1$ and $a \simeq 0$, the correlation function has a slow decay to zero and the spectrum is more properly called $1/f$-noise (or white noise). This spectrum appears in many natural phenomena (noise in electronic devices, traffic flow, signals, radio-antenna, turbolence).

The power spectrum $1/k$ has been observed in DNA sequences [30, 37], however it is not yet clear how the correlation function should be (step-function, power law decay, white-noise), and in particular if there exists a single length scale or a multilength scale [29]. This scale-dependence (or self-similarity) of DNA cannot yet been explained from biological point of views. A possible explanation could be the dynamic process of the evolution or maybe the functional activity inside constrained domain (like the fractal shape of brain, lungs, etc.) which might have some influence on the spatial geometry [38].

The biological explanation of long-range correlations can be explained by the existence of heterogeneity in DNA (i.e., different density distribution of bases). The main questioning is about the power law spectrum: $1/k$ [35] or $1/k^2$ [29]. Indeed it has been observed that the the power spectrum is nearly flat for low and high frequencies and only for the central part has a power low decay.

However, [29, 30] the existence of long range correlation in DNA should be intended from statistical point of view in the sense that far away base pairs tend to have similar variation. In other words, this correlation should be understood as a periodic distribution of base pairs without a causality law between base pairs located at different segments far away from each other.

6.2. Complexity

The existence of repeating motifs, periodicity, and patchiness can be considered as a simple behavior of sequence, while nonrepetitiveness or singularity is taken as a characteristic feature of complexity. In order to have a measure of complexity, for an n-lenght sequence, [44] the following has been proposed:

$$K = \log \Omega^{1/n} \tag{6.12}$$

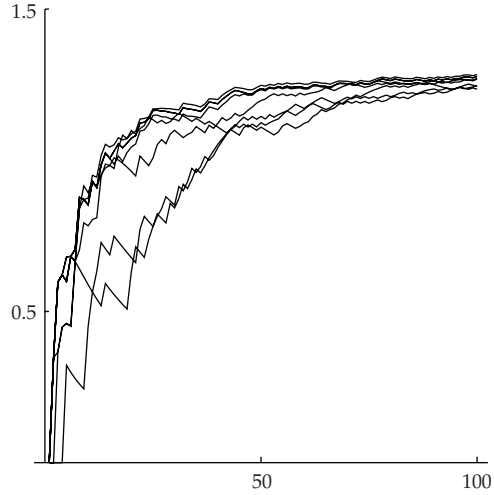

Figure 6: *Complexity for the first 100 base pairs of influenza virus A DNA.*

with

$$\Omega = \frac{n!}{a_n! c_n! g_n! t_n!}. \tag{6.13}$$

By using a sliding n-window [44] over the full DNA sequence one can visualize the distribution of complexity on partial fragment of the sequence. For the whole sequence the asymptotic constant value is (Figure 6)

$$K \cong 1.3, \tag{6.14}$$

that has been observed also for other DNA sequences [11].

Moreover, as can be seen from Figure 6, initially Tver and Novosibirsk DNA shows a lower complexity (see also Figure 1).

7. Wavelet Analysis

Wavelet analysis can be considered as a good tool [19, 24, 29, 42] for studying the heterogeneity in a time series and in particular in a DNA sequence. Heterogeneity can be shortly described as follows: in some fragments of DNA there exists a higher concentration of nucleotides C, G with poor distribution of A, T while, on the contrary, other fragments are more rich of A, T and poor of C, G (see Figure 1). Thus a fundamental problem is to make a partition of a DNA sequence into homogenous segments. This segmentation can be done by minimizing the variance (or maximizing the entropy [36]).

The wavelet transform expresses the signal in terms of dilated and scaled instances of the wavelet basis functions. If we call $W[f]_{x_0}$ the wavelet transform of the signal $f(x)$ computed in $x = x_0$ at the scale 2^{-n} and $h(x_0)$ the local Hölder exponent, it is [24] $W[f]_{x_0} \simeq 2^{-nh(x_0)}$. Therefore, wavelet transform is one of the most expedient tools for

detecting singularities. It can be used to define a generalization of box-counting method, the so-called wavelet transform modulus maxima, in order to focus on scaling behavior [24] and to visualize the multifractal property.

In this section some fundamentals on Haar wavelet theory will be given and applied to the analysis of DNA sequences.

7.1. Haar Wavelet Basis

The *Haar scaling function* $\varphi(x)$ is the characteristic function on $[0,1]$; its family of translated and dilated scaling functions is defined as

$$\begin{aligned} \varphi_k^n(x) &\stackrel{\text{def}}{=} 2^{n/2}\varphi(2^n x - k), \quad (0 \le n,\ 0 \le k \le 2^n - 1), \\ \varphi(2^n x - k) &= \begin{cases} 1, & x \in \Omega_k^n, \\ 0, & x \notin \Omega_k^n \end{cases} \quad \Omega_k^n \stackrel{\text{def}}{=} \left[\frac{k}{2^n},\ \frac{k+1}{2^n}\right). \end{aligned} \tag{7.1}$$

The *Haar wavelet* family $\{\psi_k^n(x)\}$ is the orthonormal basis for the $L^2(\mathbb{R})$ functions [45]:

$$\begin{aligned} \psi_k^n(x) &\stackrel{\text{def}}{=} 2^{n/2}\psi(2^n x - k), \quad \left\|\psi_k^n(x)\right\|_{L^2} = 1, \\ \psi(2^n x - k) &\stackrel{\text{def}}{=} \begin{cases} -1, & x \in \left[\dfrac{k}{2^n}, \dfrac{k+\frac{1}{2}}{2^n}\right), \\ 1, & x \in \left[\dfrac{k+\frac{1}{2}}{2^n}, \dfrac{k+1}{2^n}\right) \\ 0 & \text{elsewhere,} \end{cases} \quad (0 \le n \quad 0 \le k \le 2^n - 1), \end{aligned} \tag{7.2}$$

7.2. Discrete Haar Wavelet Transform

Let $\mathbf{Y} \equiv \{Y_i\}$, $(i = 0, \ldots, 2^M - 1,\ 2^M = N < \infty,\ M \in \mathbb{N})$ be a real and square summable time-series $\mathbf{Y} \in \mathbb{K}^N \subset \ell^2$ (where $\mathbb{K}$ is a real field), sampled at the *dyadic points* $x_i = i/(2^M - 1)$, in the interval restricted, for convenience and without restriction, to $\Omega = [0,1]$. The *discrete Haar wavelet transform* is the $N \times N$ matrix $\mathcal{W}^N : \mathbb{K}^N \subset \ell^2 \to \mathbb{K}^N \subset \ell^2$ which maps the vector $\mathbf{Y}$ into the vector of *wavelet coefficients* $\boldsymbol{\beta}_N = \{\alpha, \beta_k^n\}$:

$$\begin{aligned} \mathcal{W}_N \mathbf{Y} &= \boldsymbol{\beta}_N, \\ \boldsymbol{\beta}_N &\stackrel{\text{def}}{=} \left\{\alpha, \beta_0^0, \ldots, \beta_{2^{M-1}-1}^{M-1}\right\}, \\ \mathbf{Y} &\stackrel{\text{def}}{=} \{Y_0, Y_1, \ldots, Y_{N-1}\} \quad \left(2^M = N\right). \end{aligned} \tag{7.3}$$

Let the direct sum of matrices A, B be defined as

$$A \oplus B = \begin{pmatrix} A & \mathbf{0} \\ \mathbf{0} & B \end{pmatrix}, \tag{7.4}$$

being $\mathbf{0}$ the matrix of zero elements. The $N \times N$ matrix $\mathcal{W}_N$ can be computed by the recursive product [14, 15]

$$\mathcal{W}_N \stackrel{\text{def}}{=} \left[\prod_{k=1}^{M} ((P_{2^k} \oplus I_{2^M - 2^k})(H_{2^k} \oplus I_{2^M - 2^k})) \right], \quad N = 2^M \tag{7.5}$$

of the direct sum of the following elementary matrices.

(1) Identity:

$$I_{2^k} = \begin{pmatrix} 1 & & \mathbf{0} \\ & \underbrace{\ddots}_{2^k} & \\ \mathbf{0} & & 1 \end{pmatrix}, \tag{7.6}$$

which is equivalent to

$$I_2 \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad I_{2^k} \equiv \underbrace{I_2 \oplus \cdots \oplus I_2}_{k} = \begin{pmatrix} I_2 & & \mathbf{0} \\ & \underbrace{\ddots}_{k} & \\ \mathbf{0} & & I_2 \end{pmatrix}. \tag{7.7}$$

(2) Shuffle:

$$P_2 \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad P_4 \equiv \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad P_8 = \begin{pmatrix} 1 & & & & & & & \\ \downarrow & \rightarrow & 1 & & & & & \\ & & \downarrow & \rightarrow & 1 & & & \\ & & & & \downarrow & \rightarrow & 1 & \rightarrow \\ \rightarrow & 1 & & & & & & \\ & \downarrow & \rightarrow & 1 & & & & \\ & & & \downarrow & \rightarrow & 1 & & \\ & & & & & \downarrow & \rightarrow & 1 \end{pmatrix}. \tag{7.8}$$

The arbitrary shuffle matrix $P_{2^k} = (a_{ij})$; $i, j = 1, \ldots, 2^k$ can be defined as $a_{i+1,j+2} = 1$, $i = 1, \ldots, 2^{k-1}$, $j = 1, \ldots, 2^k - 3$, $i = 2^{k-1}, \ldots, 2^k - 1$, $j = 0, \ldots, 2^k - 2$ and zero elsewhere.

(3) Lattice (derived from the recursive inclusion formulas see, e.g., [14, 15]):

$$H_2 \equiv \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \qquad H_4 = H_2 \oplus H_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \ldots, \tag{7.9}$$

and in general

$$H_{2^k} \equiv \underbrace{H_2 \oplus \cdots \oplus H_2}_{k} = \begin{pmatrix} H_2 & & \mathbf{0} \\ & \underbrace{\ddots}_{k} & \\ \mathbf{0} & & H_2 \end{pmatrix}. \tag{7.10}$$

For example, with $N = 4,\ M = 2$, assuming the empty set $I_0 \overset{\text{def}}{=} \emptyset$ as the neutral term for the direct sum $\oplus$ so that $A \oplus I_0 = I_0 \oplus A = A$, it follows from (7.5)

$$\begin{aligned} \mathcal{W}_4 &= \prod_{k=1,2} [(P_{2^k} \oplus I_{4-2^k})(H_{2^k} \oplus I_{4-2^k})] \\ &= [(P_2 \oplus I_2)(H_2 \oplus I_2)]_{k=1} [(P_4 \oplus I_0)(H_4 \oplus I_0)]_{k=2}, \end{aligned} \tag{7.11}$$

that is,

$$\mathcal{W}_4 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}. \tag{7.12}$$

7.3. Haar Wavelet Coefficients and Statistical Parameters

From (7.3) with $M = 2, N = 4$, by explicit computation, we have

$$\alpha = \frac{1}{4}(\Upsilon_0 + \Upsilon_1 + \Upsilon_2 + \Upsilon_3) \tag{7.13}$$

and [10, 11, 20]

$$\begin{aligned}\beta_0^0 &= \frac{1}{2}(Y_2 - Y_0 + Y_3 - Y_1),\\ \beta_0^1 &= \frac{1}{\sqrt{2}}(Y_0 - Y_1),\\ \beta_1^1 &= \frac{1}{\sqrt{2}}(Y_3 - Y_2).\end{aligned} \tag{7.14}$$

When the wavelet coefficients are given, the above equations can be solved with respect to the original data. With $M = 2$, $N = 4$, we have, for example,

$$\begin{aligned} Y_0 &= \alpha - \frac{\beta_0^0 + \sqrt{2}\beta_0^1}{2}, \qquad Y_1 = \alpha - \frac{\beta_0^0 - \sqrt{2}\beta_0^1}{2},\\ Y_2 &= \alpha + \frac{\beta_0^0 - \sqrt{2}\beta_1^1}{2}, \qquad Y_3 = \alpha + \frac{\beta_0^0 + \sqrt{2}\beta_1^1}{2}. \end{aligned} \tag{7.15}$$

Thus the first wavelet coefficient α represents the average value of the sequence and the other coefficients β the finite differences. The wavelet coefficients β's, also called details coefficients, are strictly connected with the first-order properties of the discrete time-series.

7.4. Hurst Exponent

Concerning the variance, from definition (6.1) we obtain by a direct computation its expression in terms of wavelet coefficients:

$$\sigma^2 = \frac{1}{N}\sum_{n=0}^{M-1}\sum_{k=0}^{2^n-1}(\beta_k^n)^2 \qquad \left(N = 2^M\right). \tag{7.16}$$

It has been observed [25] that for scale invariant functions the standard deviation (6.3), as a function of the scale n, is

$$\sigma(2^n) = \sigma\left(2^0\right)2^{n(H-1)} \tag{7.17}$$

with H being Hurst exponent, so that in a log-log plot

$$\log_2\sigma(2^n) = \left[n(H-1) + \log_2\sigma\left(2^0\right)\right] \tag{7.18}$$

we obtain a straight line whose slope gives an estimate of H.

The Hurst exponent, in terms of wavelet coefficients, can be evaluated by the following [11].

Theorem 7.1. *The Hurst exponent is given by*

$$H = \frac{1}{2} + \log_2 \sqrt[n]{\frac{\left[\sum_{k=0}^{2n-1} \left(\beta_k^n\right)^2\right]^{1/2}}{\left|\beta_0^0\right|}}. \tag{7.19}$$

Proof. Taking into account (6.3) and (7.16), definition (7.17) become:

$$\begin{aligned} \frac{1}{2^{n/2}}\left[\sum_{k=0}^{2^n-1}\left(\beta_k^n\right)^2\right]^{1/2} &= \left|\beta_0^0\right| 2^{n(H-1)}, \\ \log_2\left\{\frac{1}{2^{n/2}}\left[\sum_{k=0}^{2^n-1}\left(\beta_k^n\right)^2\right]^{1/2}\right\} - \log_2\left|\beta_0^0\right| &= n(H-1), \end{aligned} \tag{7.20}$$

that is,

$$\begin{aligned} \frac{1}{n}\log_2 \frac{1}{2^{n/2}} \frac{\left[\sum_{k=0}^{2^n-1}\left(\beta_k^n\right)^2\right]^{1/2}}{\left|\beta_0^0\right|} &= (H-1), \\ \log_2 2 + \frac{1}{n}\log_2 \frac{1}{2^{n/2}} \frac{\left[\sum_{k=0}^{2^n-1}\left(\beta_k^n\right)^2\right]^{1/2}}{\left|\beta_0^0\right|} &= H \end{aligned} \tag{7.21}$$

from where (7.19) follows. □

8. Algorithm of the Short Haar Discrete Wavelet Transform

In order to reduce the computational complexity of the wavelet transform (7.3), (7.5), the sequence $\mathbf{Y}$ can be sliced into subsequences and the wavelet transform is applied to each slice. With the reduced Haar transform [10, 11, 20] it is possible to reduce the number of basis functions and the computational complexity.

Algorithm 8.1. Let $\mathbf{Y} = \{Y_i\}_{i=0,\ldots,N-1}$ of N data, segmented into $\sigma = N/p$, $(1 \le \sigma \le N)$ segments *of* $p = 2^m$ data*:*

$$\mathbf{Y} = \{Y_i\}_{i=0,\ldots,N-1} = \bigoplus_{s=0}^{\sigma-1} \mathbf{Y}^s, \quad \mathbf{Y}^s \equiv \left\{Y_{sp}, Y_{sp+1}, \ldots, Y_{sp+p-1}\right\} \quad \left(s = 0, \ldots, N-p;\ 1 \le p \le N\right), \tag{8.1}$$

and the p-parameters short (reduced or windowed) discrete Haar wavelet transform $\mathcal{W}^{p,\sigma}\mathbf{Y}$ of $\mathbf{Y}$ is defined as

$$\begin{aligned}
\mathcal{W}^{p,\sigma} &\equiv \bigoplus_{s=0}^{\sigma-1} \mathcal{W}^{p}, \qquad \mathbf{Y} = \bigoplus_{s=0}^{\sigma-1} \mathbf{Y}^{s}, \\
\mathcal{W}^{p,\sigma}\mathbf{Y} &= \left(\bigoplus_{s=0}^{\sigma-1} \mathcal{W}^{p}\right)\mathbf{Y} = \left(\bigoplus_{s=0}^{\sigma-1} \mathcal{W}^{p}\mathbf{Y}^{s}\right), \\
\mathcal{W}^{2^m}\mathbf{Y}^{s} &= \left\{\alpha_0^{0(s)}, \beta_0^{0(s)}, \beta_0^{1(s)}, \beta_1^{1(s)}, \ldots, \beta_{2^{m-1}-1}^{m-1(s)}\right\} \quad (2^m = p).
\end{aligned} \tag{8.2}$$

For example, the reduced wavelet transform $\mathcal{W}^{4,2}$ (to be compared with $\mathcal{W}^{8}$) is

$$\mathcal{W}^{4,2} = \mathcal{W}^{4} \oplus \mathcal{W}^{4}, \tag{8.3}$$

that is,

$$\mathcal{W}^{4,2} = \begin{pmatrix}
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\
-\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\
-\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
0 & 0 & 0 & 0 & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
0 & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{pmatrix}. \tag{8.4}$$

8.1. Clusters of Wavelet Coefficients

Significant information on a time-series can be derived not only from the wavelet coefficients but also from clusters of wavelet coefficients.

For the $N = 2^M$-length real vector $\mathbf{Y}$ the wavelet transform $\mathcal{W}^{N}\mathbf{Y}$ represents a point in the N-dimensional Euclidean space

$$\mathbb{R}^N : \left(\alpha, \beta_0^0, \beta_0^1, \ldots, \beta_{2^{M-1}-1}^{M-1}\right) \tag{8.5}$$

of the wavelet coefficients.

For the $N = 2^M$-length complex vector $\mathbf{Y}$ the wavelet transform is applied to the real $\mathcal{W}^N\Re(\mathbf{Y})$ and to the imaginary part $\mathcal{W}^N\Im(\mathbf{Y})$ and gives either 1 point in

$$\mathbb{C}^N = \mathbb{R}^{2N} : \left(\alpha, \beta_0^0, \beta_0^1, \ldots, \beta_{2^{M-1}-1}^{M-1}, \alpha^*, \beta_0^{*0}, \beta_0^{*1}, \ldots, \beta_{2^{M-1}-1}^{*M-1}\right) \tag{8.6}$$

or 2 points in

$$\mathbb{R}^N \times \mathbb{R}^N : \left(\alpha, \beta_0^0, \beta_0^1, \ldots, \beta_{2^{M-1}-1}^{M-1}\right) \times \left(\alpha^*, \beta_0^{*0}, \beta_0^{*1}, \ldots, \beta_{2^{M-1}-1}^{*M-1}\right) \tag{8.7}$$

or a cluster of N points in the product of 2 dimensional spaces:

$$\prod_{i=1}^{N} \mathbb{R}_i^2 : (\alpha, \alpha^*) \times \left(\beta_0^0, \beta_0^{*0}\right) \times \left(\beta_0^1, \beta_0^{*1}\right) \times \cdots \times \left(\beta_{2^{M-1}-1}^{M-1}, \beta_{2^{M-1}-1}^{*M-1}\right), \tag{8.8}$$

where the star denotes the wavelet coefficients of $\Im(\mathbf{Y})$. In each 2-dimensional (phase) space $\mathbb{R}_i^2$ there is only one point and these single points do not give any significant information about the existence of some autocorrelation of data. By using, instead, the p-parameter short Haar wavelet transform we can analyse the cluster of points

$$(\mathcal{W}^p\Re(\mathbf{Y}^s), \mathcal{W}^p\Im(\mathbf{Y}^s)), \qquad s = 0, \ldots, \sigma = \frac{N}{p} \tag{8.9}$$

in the $2p$-dimensional space $\mathbb{R}^p \times \mathbb{R}^p$, that is,

$$(\alpha, \alpha^*), \left(\beta_0^0, \beta_0^{*0}\right), \ldots, \left(\beta_{2^{p-1}-1}^{p-1}, \beta_{2^{p-1}-1}^{*p-1}\right). \tag{8.10}$$

For a complex sequence $\{Y_k\}_{k=0,\ldots,N-1} = \{x_k + iy_k\}_{k=0,\ldots,N-1}$ we can consider the correlations (if any) between the wavelet coefficients of the real part $\{x_k\}_{k=0,\ldots,N-1}$ against the imaginary coefficients $\{y_k\}_{k=0,\ldots,N-1}$. This can be realized by the cluster algorithm of Table 1.

This algorithm enables us to construct clusters of wavelet coefficients and to study the correlation between the real and imaginary coefficients of the DNA representation and DNA walk, as given in the following section.

8.2. Cluster Analysis of the Wavelet Coefficients of the Complex DNA Representation

The cluster algorithm of Table 1, applied to the complex representation sequence (4.11), which is in the form

$$\{-1, i, 1, 1, -i, i, -1, -i, i, 1, -i, i, -1, i, 1, -i, -i, -i, -1, \ldots\} \tag{8.11}$$

shows that the values of the wavelet coefficients belong to some discrete finite sets (Figure 7).

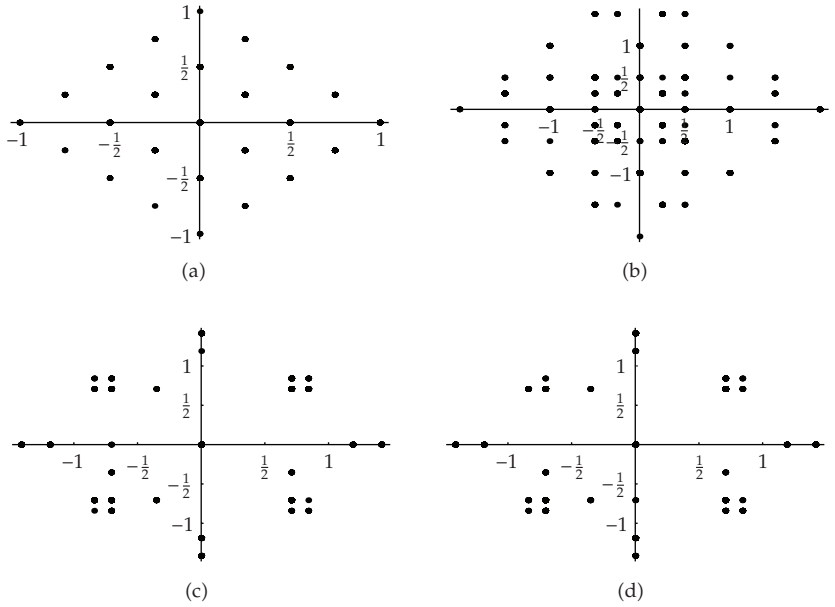

Figure 7: Cluster analysis of the 4th short Haar wavelet transform of the complex representation of the DNA influenza virus A H1N1 (Mexico City) in the planes: (a) (α, α^*); (b) $(\beta_0^0, \beta_0^{*0})$; (c) $(\beta_0^1, \beta_0^{*1})$; (d) $(\beta_1^1, \beta_1^{*1})$.

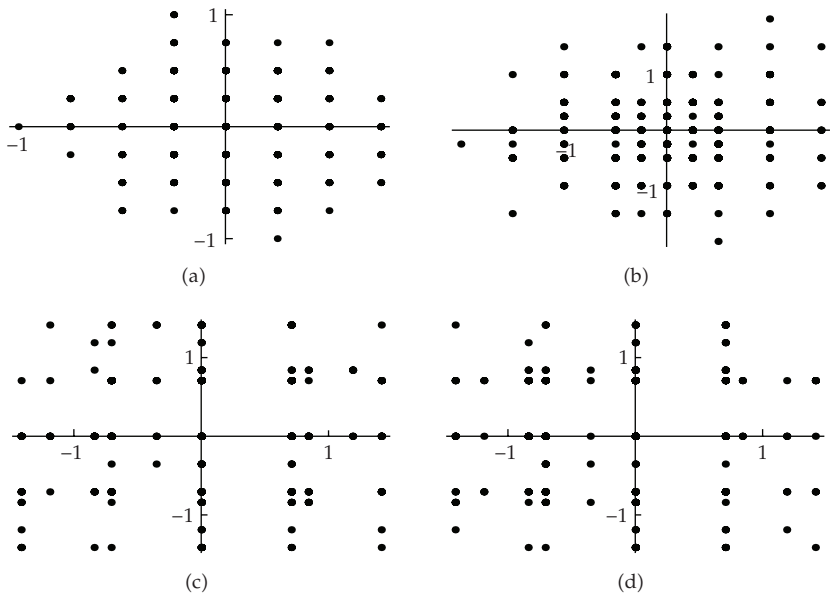

Figure 8: Cluster analysis of the 4th short Haar wavelet transform of the pseudo-random sequence (5.6) $(n \leq 1200)$: (a) (α, α^*); (b) $(\beta_0^0, \beta_0^{*0})$; (c) $(\beta_0^1, \beta_0^{*1})$; (d) $(\beta_1^1, \beta_1^{*1})$.

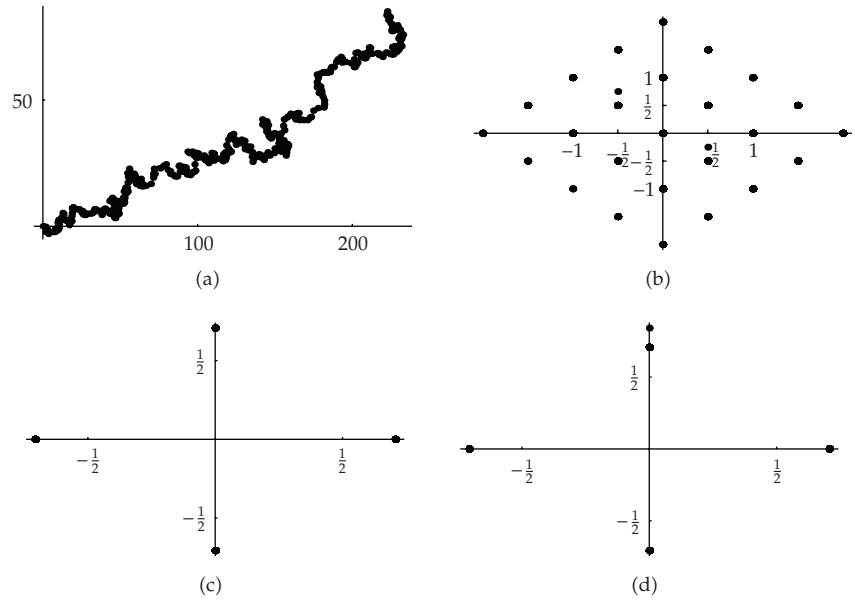

Figure 9: Cluster analysis of the 4th short Haar wavelet transform of the walk on (whole sequence) DNA for the influenza virus A H1N1 (Mexico City) in the planes: (a) (α, α^*); (b) $(\beta_0^0, \beta_0^{*0})$; (c) $(\beta_0^1, \beta_0^{*1})$; (d) $(\beta_1^1, \beta_1^{*1})$.

Table 1

$\{Y_0, Y_1, \ldots, Y_p\} \oplus \{Y_{p+1}, Y_{p+2}, \ldots, Y_{2p}\} \oplus \cdots$	Complex values sequence
$\Downarrow$	
$x_i = \Re(Y_i),\ \ y_i = \Im(Y_i)$	Real values sequences
$\Downarrow$	
$\left.\begin{array}{r}\{x_0, x_1, \ldots, x_p\} \oplus \{x_{p+1}, x_{p+2}, \ldots, x_{2p}\} \oplus \cdots \\ \{y_0, y_1, \ldots, y_p\} \oplus \{y_{p+1}, y_{p+2}, \ldots, y_{2p}\} \oplus \cdots\end{array}\right\rangle$	Real sequences
$\Downarrow$	
$\left.\begin{array}{r}\mathcal{W}^p\{x_0, x_1, \ldots, x_p\} \oplus \mathcal{W}^p\{x_{p+1}, x_{p+2}, \ldots, x_{2p}\} \oplus \cdots \\ \mathcal{W}^p\{y_0, y_1, \ldots, y_p\} \oplus \mathcal{W}^p\{y_{p+1}, y_{p+2}, \ldots, y_{2p}\} \oplus \cdots\end{array}\right\rangle$	Wavelet transform
$\Downarrow$	
$\left.\begin{array}{r}\{\alpha, \beta_0^0, \beta_0^1, \beta_1^1, \ldots\}_1 \oplus \{\alpha, \beta_0^0, \beta_0^1, \beta_1^1, \ldots\}_2 \oplus \cdots \\ \{\alpha^*, \beta_0^{*0}, \beta_0^{*1}, \beta_1^{*1}, \ldots\}_1 \oplus \{\alpha^*, \beta_0^{*0}, \beta_0^{*1}, \beta_1^{*1}, \ldots\}_2 \oplus \cdots\end{array}\right\rangle$	Wavelet coefficients
$\Downarrow$	
$\left.\begin{array}{r}\{(\alpha, \alpha^*)\}_1 \oplus \{(\alpha, \alpha^*)\}_2 \oplus \{(\alpha, \alpha^*)\}_3 \cdots \\ \{(\beta_0^0, \beta_0^{*0})\}_1 \oplus \{(\beta_0^0, \beta_0^{*0})\}_2 \oplus \{(\beta_0^0, \beta_0^{*0})\}_3 \cdots \\ \{(\beta_0^1, \beta_0^{*1})\}_1 \oplus \{(\beta_0^1, \beta_0^{*1})\}_2 \oplus \{(\beta_0^1, \beta_0^{*1})\}_3 \cdots \\ \vdots \quad \vdots \quad \vdots\end{array}\right\rangle$	Clusters

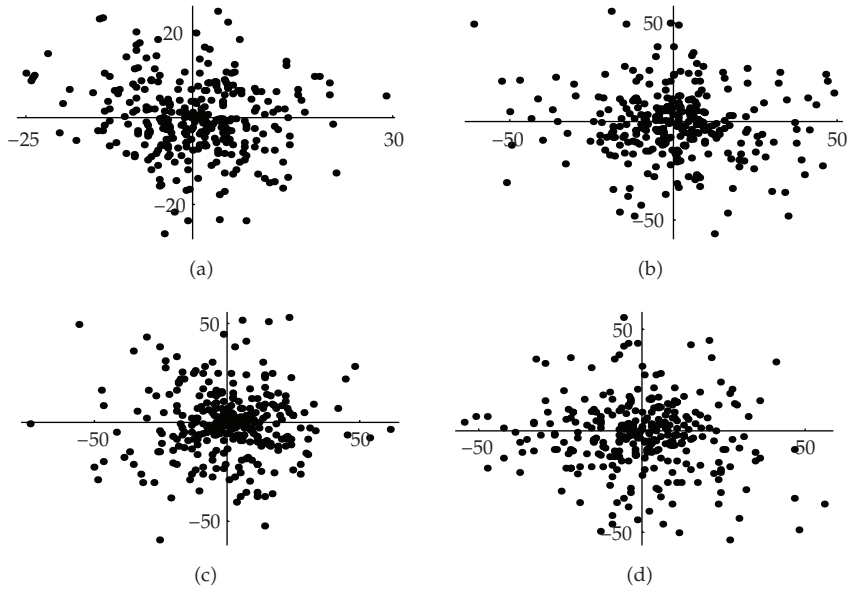

Figure 10: Cluster analysis of the 4th short Haar wavelet transform of the random walk ($n \leq 1200$) of the random sequence (5.6) in the planes: (a) (α, α^*); (b) $(\beta_0^0, \beta_0^{*0})$; (c) $(\beta_0^1, \beta_0^{*1})$; (d) $(\beta_1^1, \beta_1^{*1})$.

For each complex DNA representation there are 2 sets of wavelet coefficients which correspond to the real and complex coefficient of the complex values of (5.1) and (5.4). However, even if the real and complex coefficients of the DNA walk show some nonlinear patterns (Figures 2 and 3 the detail coefficients range in some discrete sets of values. It can be seen by a direct computation that the jumps from one value to another belong to some discrete sets (see, e.g., Figure 7).

If we compare the clusters of Figure 7 with the clusters (Figure 8) of the pseudorandom sequence (5.6), which is similar to the above sequence, we can see that the set of wavelet coefficients is larger (still discrete) than the set for the DNA although the detail coefficients have (more or less) the same values.

As can be seen from Figure 7, the real and imaginary coefficients of the complex DNA representation increase with a given law and the distribution of the nucleotides must follow this rule. Moreover, it should be noticed that all wavelet coefficients are distributed on symmetric grids (Figure 7). Even if the DNA representation looks like the pseudorandom sequence (5.6), the wavelet (detail) coefficients are quantized and symmetrically distributed in the sense that the detail coefficients of both the representation and the DNA walks (see [11]) have discrete finite values (Figure 7), being, in particular,

$$\left|\beta_0^0 \pm \beta_0^{*0}\right| \leq 2. \tag{8.12}$$

This is not true for the pseudorandom series, because the wavelet coefficients of the sequence are still quantized (see Figure 8) while the wavelet coefficients of the corresponding random walk are randomly distributed in the phase plane (Figure 10). It is very interesting to compare also the DNA walk (Figure 9) with the random walk (Figure 10) and random walk on deterministic sequence. DNA walk shows a clear symmetry which is missing in the others.

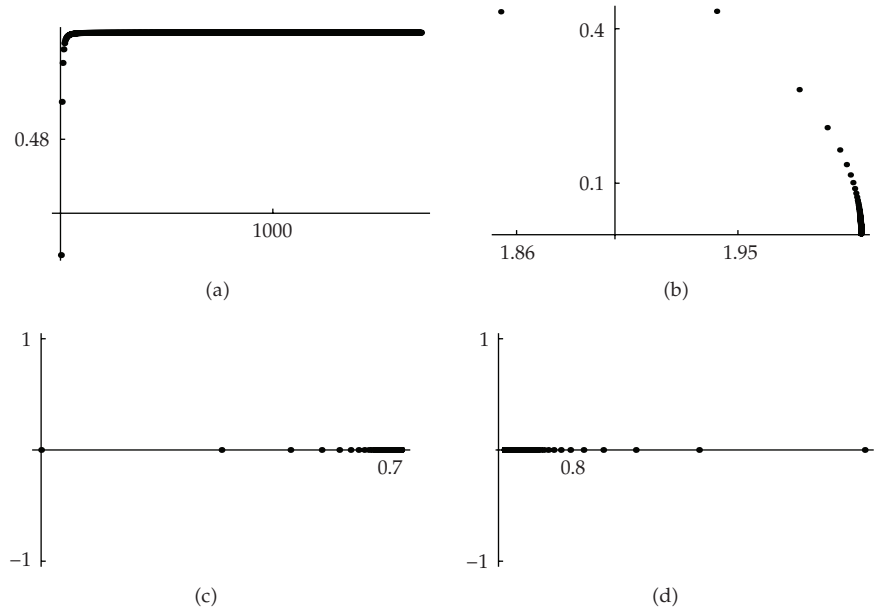

Figure 11: Cluster analysis of the 4th short Haar wavelet transform of the walk ($n \leq 1200$) of the deterministic sequence (5.9) in the planes: (a) (α, α^*); (b) $(\beta_0^0, \beta_0^{*0})$; (c) $(\beta_0^1, \beta_0^{*1})$; (d) $(\beta_1^1, \beta_1^{*1})$.

9. Conclusion

In this paper some fractal shapes and symmetries in DNA sequences and DNA walks have been shown and compared with random and deterministic complex series. DNA sequences are structured in such a way that there exists some fractal behavior which can be observed both on the correlation matrix and on the DNA walks. Wavelet analysis confirms by a symmetrical clustering of wavelet coefficients the existence of scale symmetries.

References

[1] J. P. Fitch and B. Sokhansanj, "Genomic engineering: moving beyond DNA sequence to function," *Proceedings of the IEEE*, vol. 88, no. 12, pp. 1949–1971, 2000.
[2] H. Gee, "A journey into the genome: what's there," *Nature*, 2001, http://www.nature.com/nsu/010215/010215-3.html.
[3] National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/genbank/.
[4] Genome Browser, http://genome.ucsc.edu/ .
[5] European Informatics Institute, http://www.ebi.ac.uk/.
[6] Ensembl, http://www.ensembl.org.
[7] C. Cattani and J. J. Rushchitsky, *Wavelet and Wave Analysis as Applied to Materials with Micro or Nanostructure*, vol. 74 of *Series on Advances in Mathematics for Applied Sciences*, World Scientific, Singapore, 2007.
[8] P. D. Cristea, "Large scale features in DNA genomic signals," *Signal Processing*, vol. 83, no. 4, pp. 871–888, 2003.
[9] K. B. Murray, D. Gorse, and J. M. Thornton, "Wavelet transforms for the characterization and detection of repeating motifs," *Journal of Molecular Biology*, vol. 316, no. 2, pp. 341–363, 2002.
[10] C. Cattani, "Complex representation of DNA sequences," in *Proceedings of the 2nd International Conference Bioinformatics Research and Development (BIRD '08)*, M. Elloumi, et al., Ed., vol. 13 of *Communications in Computer and Information Science*, pp. 528–537, Springer, Vienna, Austria, July 2008.

[11] C. Cattani, "Wavelet algorithms for DNA analysis," in *Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications*, M. Elloumi and A. Y. Zomaya, Eds., Wiley Series in Bioinformatics, chapter 35, Wiley-Blackwell, New York, NY, USA, 2010.
[12] R. F. Voss, "Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences," *Physical Review Letters*, vol. 68, no. 25, pp. 3805–3808, 1992.
[13] F. Voss, "Long-range fractal correlations in DNA introns and exons," *Fractals*, vol. 2, pp. 1–6, 1992.
[14] C. Cattani, "Haar wavelet-based technique for sharp jumps classification," *Mathematical and Computer Modelling*, vol. 39, no. 2-3, pp. 255–278, 2004.
[15] C. Cattani, "Haar wavelets based technique in evolution problems," *Proceedings of the Estonian Academy of Sciences, Physics and Mathematics*, vol. 53, no. 1, pp. 45–63, 2004.
[16] A. A. Tsonis, P. Kumar, J. B. Elsner, and P. A. Tsonis, "Wavelet analysis of DNA sequences," *Physical Review E*, vol. 53, no. 2, pp. 1828–1834, 1996.
[17] M. Altaiski, O. Mornev, and R. Polozov, "Wavelet analysis of DNA sequences," *Genetic Analysis*, vol. 12, no. 5-6, pp. 165–168, 1996.
[18] A. Arneodo, Y. D'Aubenton-Carafa, E. Bacry, P. V. Graves, J. F. Muzy, and C. Thermes, "Wavelet based fractal analysis of DNA sequences," *Physica D*, vol. 96, no. 1–4, pp. 291–320, 1996.
[19] M. Zhang, "Exploratory analysis of long genomic DNA sequences using the wavelet transform: examples using polyomavirus genomes," in *Proceedings of the 6th Genome Sequencing and Analysis Conference*, pp. 72–85, 1995.
[20] C. Cattani, "Harmonic wavelet approximation of random, fractal and high frequency signals," *Telecommunication Systems*, vol. 43, no. 3-4, pp. 207–217, 2010.
[21] M. Li, "Fractal time series—a tutorial review," *Mathematical Problems in Engineering*, vol. 2010, Article ID 157264, 26 pages, 2010.
[22] M. Li and J.-Y. Li, "On the predictability of long-range dependent series," *Mathematical Problems in Engineering*, vol. 2010, Article ID 397454, 9 pages, 2010.
[23] M. Li and S. C. Lim, "Power spectrum of generalized Cauchy process," *Telecommunication Systems*, vol. 43, no. 3-4, pp. 219–222, 2010.
[24] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, "Characterizing long-range correlations in DNA sequences from wavelet analysis," *Physical Review Letters*, vol. 74, no. 16, pp. 3293–3296, 1995.
[25] B. Audit, C. Vaillant, A. Arneodo, Y. D'Aubenton-Carafa, and C. Thermes, "Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes," *Journal of Molecular Biology*, vol. 316, no. 4, pp. 903–918, 2002.
[26] B. Borstnik, D. Pumpernik, and D. Lukman, "Analysis of apparent $1/f^{\alpha}$ spectrum in DNA sequences," *Europhysics Letters*, vol. 23, pp. 389–394, 1993.
[27] S. V. Buldyrev, A. L. Goldberger, S. Havlin, et al., "Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis," *Physical Review E*, vol. 51, no. 5, pp. 5084–5091, 1995.
[28] H. Herzel, E. N. Trifonov, O. Weiss, and I. Große, "Interpreting correlations in biosequences," *Physica A*, vol. 249, no. 1–4, pp. 449–459, 1998.
[29] W. Li, "The study of correlation structures of DNA sequences: a critical review," *Computers and Chemistry*, vol. 21, no. 4, pp. 257–271, 1997.
[30] W. Li and K. Kaneko, "Long-range correlations and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence," *Europhysics Letters*, vol. 17, pp. 655–660, 1992.
[31] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, et al., "Long-range correlations in nucleotide sequences," *Nature*, vol. 356, no. 6365, pp. 168–170, 1992.
[32] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotides," *Physical Review E*, vol. 49, no. 2, pp. 1685–1689, 1994.
[33] O. Weiss and H. Herzel, "Correlations in protein sequences and property codes," *Journal of Theoretical Biology*, vol. 190, no. 4, pp. 341–353, 1998.
[34] Z.-G. Yu, V. V. Anh, and B. Wang, "Correlation property of length sequences based on global structure of the complete genome," *Physical Review E*, vol. 63, no. 1, Article ID 011903, 8 pages, 2001.
[35] P. P. Vaidyanathan and B.-J. Yoon, "The role of signal-processing concepts in genomics and proteomics," *Journal of the Franklin Institute*, vol. 341, no. 1-2, pp. 111–135, 2004.
[36] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, "Compositional segmentation and long-range fractal correlations in DNA sequences," *Physical Review E*, vol. 53, no. 5, pp. 5181–5189, 1996.
[37] W. Li, "The complexity of DNA: the measure of compositional heterogenity in DNA sequence and measures of complexity," *Complexity*, vol. 3, pp. 33–37, 1997.
[38] S. Karlin and V. Brendel, "Patchiness and correlations in DNA sequences," *Science*, vol. 259, no. 5095, pp. 677–680, 1993.

[39] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, 2000.
[40] S. S.-T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, and Y.-K. Ho, "DNA sequence representation without degeneracy," *Nucleic Acids Research*, vol. 31, no. 12, pp. 3078–3080, 2003.
[41] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, and L. Marcourt, "Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences," *Journal of Theoretical Biology*, vol. 206, no. 3, pp. 323–326, 2000.
[42] A. Arneodo, Y. D'Aubenton-Carafa, B. Audit, E. Bacry, J. F. Muzy, and C. Thermes, "What can we learn with wavelets about DNA sequences?" *Physica A*, vol. 249, no. 1–4, pp. 439–448, 1998.
[43] E. Coward, "Equivalence of two Fourier methods for biological sequences," *Journal of Mathematical Biology*, vol. 36, no. 1, pp. 64–70, 1997.
[44] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, "Visualization and analysis of DNA sequences using DNA walks," *Journal of the Franklin Institute*, vol. 341, no. 1-2, pp. 37–53, 2004.
[45] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics, Philadelphia, Pa, USA, 1992.