Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2012, Article ID 673934, 21 pages
doi:10.1155/2012/673934

Research Article

On the Existence of Wavelet Symmetries in Archaea DNA

Carlo Cattani

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

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

Received 13 September 2011; Revised 27 October 2011; Accepted 29 October 2011

Academic Editor: Sheng-yong Chen

This paper deals with the complex unit roots representation of archea DNA sequences and the analysis of symmetries in the wavelet coefficients of the digitalized sequence. It is shown that even for extremophile archaea, the distribution of nucleotides has to fulfill some (mathematical) constraints in such a way that the wavelet coefficients are symmetrically distributed, with respect to the nucleotides distribution.

1. Introduction

In some recent papers the existence of symmetries in nucleotide distribution has been studied for several living organisms [1–6] including mammals, fungi [1–4], and viruses [5, 6]. Thus showing that any (investigated) DNA sequence, when converted into a digital sequence, features some fractal shape of its DNA walk and an apparently random-like distribution. However, when the short wavelet transform maps the digital sequence into the space of wavelet coefficients, and these coefficients are clustered then they are located along some symmetrical shapes.

One of the main tasks of this paper is to show that although the distribution of nucleotide, in any DNA sequence, can be considered as randomly given, when we compare a random sequence (and the corresponding random walk) with a DNA sequence (and walk) it can be seen that there exists some distinctions. So that the nucleotides distribution seems to side with a random distribution with some constraints. These constraints (rules) are singled out in the following, by showing the existence of hidden geometry which underlies the structure of a DNA sequence.

In other words, nucleotides are distributed along any DNA sequence at first apparently randomly but at second analysis according to some (statistical) mathematical constraints which does not allow a given nucleotide to be arbitrarily followed by any other remaining nucleotides.

It is interesting to notice that even in the primitives organisms which billions of years ago have been colonizing the earth under extreme conditions of life, their DNA has to fulfill the same constraints of the more evolved DNAs.

In order to achieve this goal some fundamental steps have to be taken into consideration and discussed.

(1) Since DNA is a sequence of symbols, a map of these symbols into numbers has to be defined. In the following we will consider the complex unit roots map, which has the advantage of being unitary and distributed along the unit circle.

(2) The indicator matrix is defined on the the indicator map. This matrix is important in order to draw the dot plot of the DNA sequence and from this plot we can see that apparently nucleotides seem to be randomly distributed. However, we will show by wavelet analysis that they look randomly distributed, while they are not.

(3) The Ulam spiral adapted to DNA sequences is defined in order to single out some geometrical patterns.

(4) Random walks on DNA, or short DNA walks, show that the random walks look like fractals.

(5) The analysis of clusters of wavelet coefficients show that DNA walks have to fulfill some geometrical constraints.

In all DNA sequences, analyzed so far, for different kinds of living organisms, this geometrical symmetry [1–6] has been detected. In the following this analysis is extended also to archaea, since they might be considered at the early

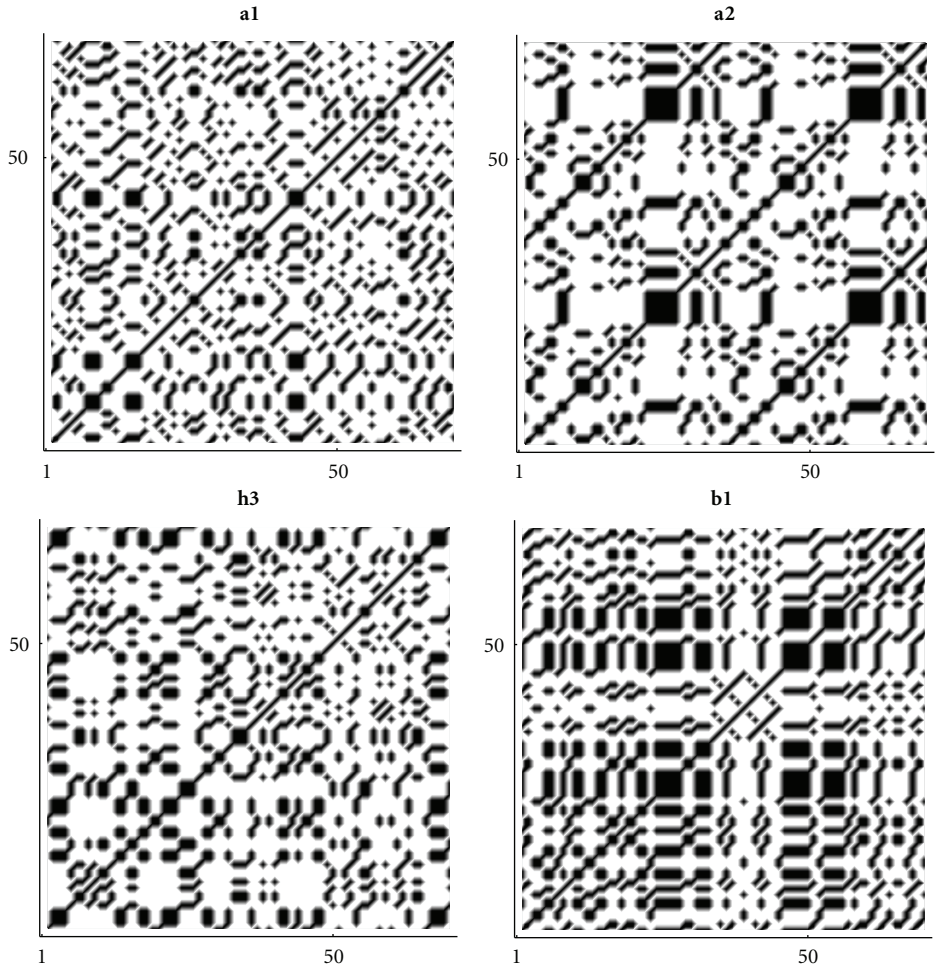

Figure 1: Indicator matrix for: (**a1**) pseudorandom 70-length sequence; (**a2**) pseudo-periodic 70-length sequence with period $\pi = 35$; (**b1**) 70-length DNA sequence of Mycoplasma KS1 bacter; (**h3**) 70-length DNA sequence of Acidilobus Archaea.

stage of life and their DNA is compared with more evolved microorganisms as bacteria.

It will be shown that, inspite of the many similarities with random sequences, only the wavelet analysis makes it possible to single out some distinctions. In particular, the wavelet coefficients of all (analyzed) organisms tend to fulfill a minimum principle for the energy of the signal. Also the archaea which often live in extreme environments have to fulfill the same geometrical rule of any other living organism.

The analysis of DNA by wavelets [7–9], as seen in [8–12], helps to single out local behavior and singularities [7, 13] or to express the scale invariance of coefficients [14]. Also multifractal nature of the time series [15–17] can be easily detected by wavelet analysis.

Some previous paper have studied various sequences of DNA such as leukemia tet variants, influenza viruses such as the A (H1N1) variant, mammalian, and a fungus (see [1–3, 14]) provided by the National Center for Biotechnology Information [18–21]. In all these papers it was observed that DNA has to fulfill not only some chemical steady state given by the chemical ligands but also some symmetrical distribution of nucleotide along the sequence. In other words, base pairs have to be placed exactly in some positions.

According to previous results, it will be shown that as any other living organisms also these elementary organisms have DNA walks with fractal shape and wavelet coefficients bounded on a short-range wavelet transform. In other words, also anaerobic organism which should be understood as the most elementary at the first step of life have the same symmetries on wavelet coefficients as for more evolved organism, so that life has to fulfill some constrained distribution of nucleotides in order to give rise to some organism even at the most elementary step.

In particular, in Section 2, some remarks about the analysed data are given. Section 3 deals with some elementary plots which can easily visualize the distribution of nucleotides. The Ulam spiral plot is also proposed for the first time and it is observed a different distribution of weak/strong

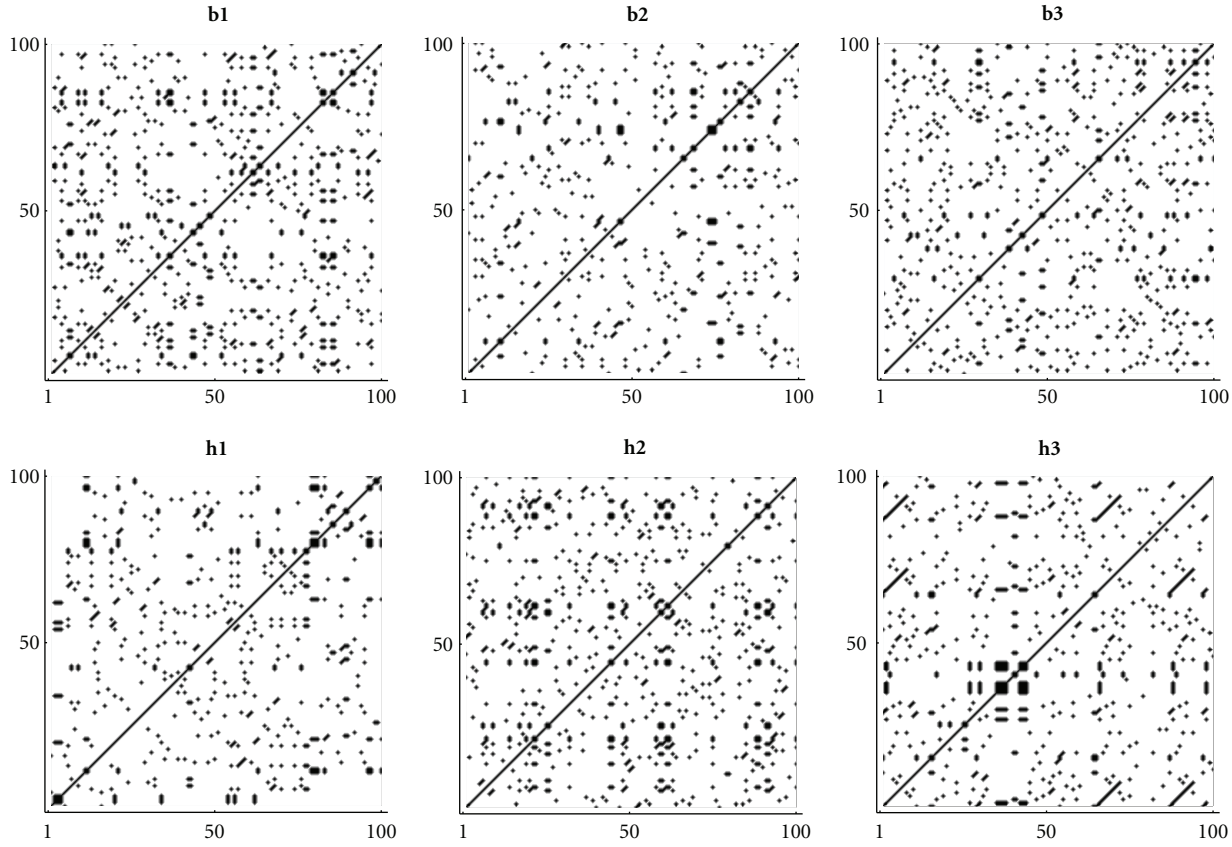

Figure 2: Indicator matrix for the first 100 amino acids of (**h1**) *Aeropyrum pernix* K1, (**h2**) *Acidianus hospitalis* W1, (**h3**) *Acidilobus saccharovorans* 345-15 (**b1**) *Mycoplasma putrefaciens* KS1, (**b2**) *Mortierella verticillata*, and (**b3**) *Blattabacterium* sp.

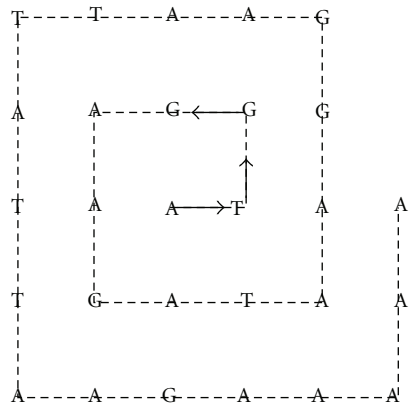

Figure 3: Distribution of nucleotides on a rectangular spiral.

hydrogen bonds. Section 4 provides some definitions about parameters of complexity. We will notice that all these parameters give rise to the same classification of organism. Section 4 proposes a complex numerical representation of DNA chains and random walks, while in final Section 6 the short wavelet trasform is given in order to single out some symmetries at the lower order of transform.

2. Materials and Methods

In the following we will take into consideration some genome, complete sequences of DNA, concerning the following archaea:

h1: *Aeropyrum pernix* K1, complete genome. DNA, circular, 1669696 bp, [18–21], accession BA000002.3. Lineage: *Archaea*; Crenarchaeota; Thermoprotei; Desulfurococcales; Desulfurococcaceae; Aeropyrum; *Aeropyrum pernix*; *Aeropyrum pernix* K1.
This organism, which was the first strictly aerobic hyperthermophilic archaeon sequenced, was isolated from sulfuric gases in Kodakara-Jima Island, Japan in 1993.

h2: *Acidianus hospitalis* W1, complete genome. DNA, circular, 2137654 bp, [18–21], accession CP002535. Lineage: *Archaea*; Crenarchaeota; Thermoprotei; Sulfolobales; Sulfolobaceae; Acidianus; *Acidianus hospitalis*; *Acidianus hospitalis* W1.

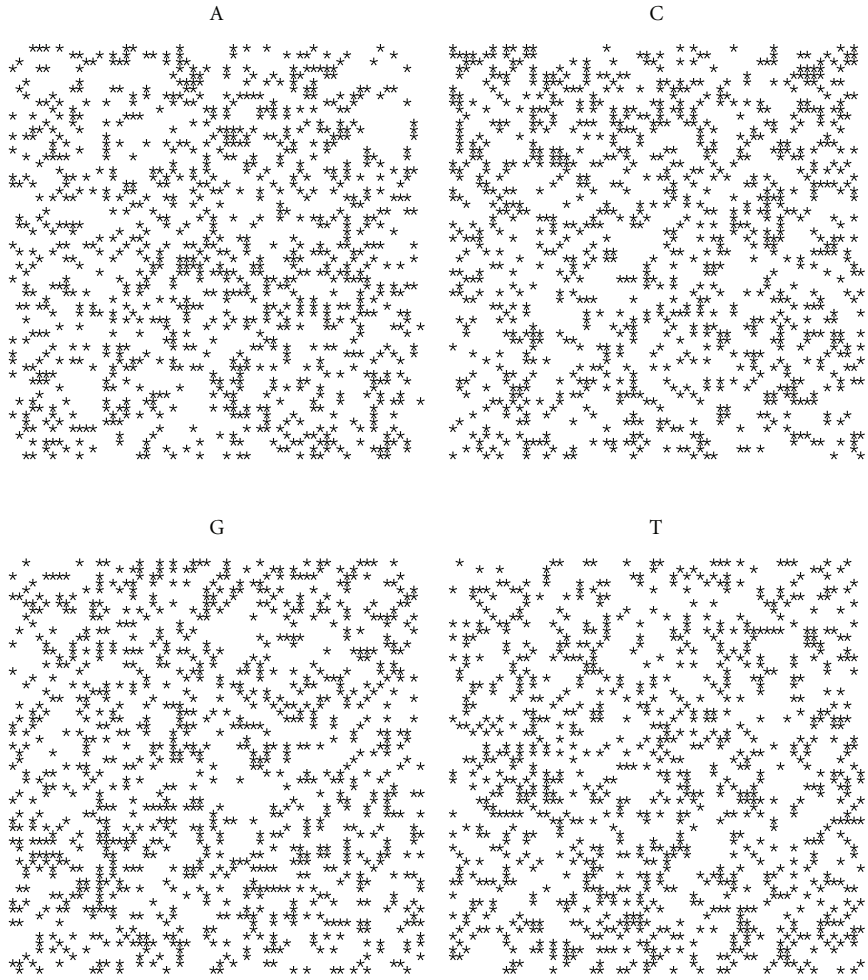

Figure 4: Spiral distribution of the first 3752 nucleotides for the random sequence.

h3: *Acidilobus saccharovorans* 345-15. complete genome. DNA, circular, 2137654 bp, [18–21], accession CP001742.1.
Lineage: *Archaea*; Crenarchaeota; Thermoprotei; Acidilobales; Acidilobaceae; Acidilobus; *Acidilobus saccharovorans*; *Acidilobus saccharovorans* 345-15.
Anaerobic bacteria found in hot springs.

to be compared with the following (aerobic/anaerobic) bacteria/fungi:

b1: *Mycoplasma putrefaciens* KS1 chromosome, complete genome. DNA, circular, length 832603 bp, [18–21], accession NC 015946,. Lineage: *Bacteria*; Tenericutes; Mollicutes; Mycoplasmatales; Mycoplasmataceae; Mycoplasma; *Mycoplasma putrefaciens*; *Mycoplasma putrefaciens* KS1.

b2: *Mortierella verticillata mitochondrion*, complete genome. dsDNA, circular, length 58745 bp, [18–21], accession NC 006838. Lineage: Eukaryota; Opisthokonta; *Fungi*; Fungi incertae sedis; Basal fungal lineages; Mucoromycotina; Mortierellales; Mortierellaceae; Mortierella; *Mortierella verticillata*.

b3: *Blattabacterium* sp. (Periplaneta Americana) str. BPLAN, complete genome. DNA, circular, length 636994 nt, [18–21], accession NC 013418. Lineage: *Bacteria*; Bacteroidetes/Chlorobi group; Bacteroidetes; Flavobacteria; Flavobacteriales; Blattabacteriaceae; Blattabacterium; Blattabacterium sp. (Periplaneta Americana); *Blattabacterium* sp. (Periplaneta Americana) str. BPLAN.

Moreover we will compare DNA sequences with artificial sequences of nucleotides randomly taken (see Section 4).

2.1. Archaea. Archaea are a group of elementary single-cell microorganisms, having no cell nucleus or any other membrane-bound organelles within their cells. They are similar to bacteria, since they have the same size and shape (apart few exceptions) and the generally similar cell structure. However, the evolutionary history of archaea and their biochemistry has significant differences with regard to other forms of life.

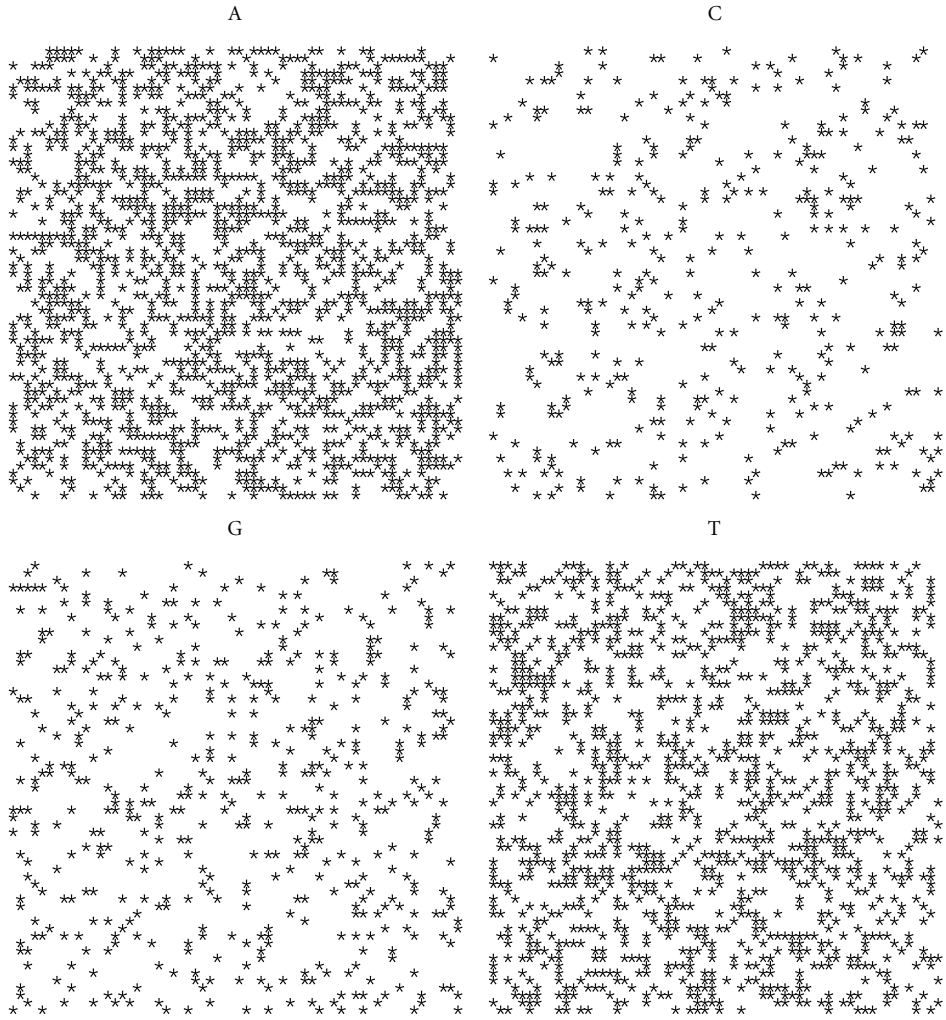

Figure 5: Spiral distribution of the first 3752 nucleotides for *Mycoplasma putrefaciens* KS1.

Therefore they are considered as members of a phylogenetic group distinct from bacteria and eukaryota.

Archaea during their evolution have been spreading all over the Earth in almost all habitats [22, 23] existing in a broad range of habitats, being one of the major contribution (20%) to earth's biomass. The most peculiar feature of archaea is that they can live in some environments with extreme life conditions (thus being considered as extremophiles [22, 24]). Indeed, some archaea survive to high temperatures, over 100°C, while others can live in very cold habitats or highly saline, acidic, or alkaline water. Nevertheless some archaea are living in mild conditions.

It has been also recognized that the archaea may be the most ancient organisms on the Earth, so that archaea, and eu-karyotes are probably diverged early from an ancestral colony of organisms.

We will see, in the following, that archaea DNA it looks very close to random sequences so that we can assume that the ancestral organism were evolving by random permutations from a primitive assembly of nucleotides. So that the evolution can be seen as a tendency to a steady state far from the randomness. Therefore, the bacteria's DNA (and other eukaryotes' [1–6]), as a result of the evolution, shows the existence of some hidden stability.

3. Correlation Plots

In this section we will consider some elementary plots from where it is possible to visualize autocorrelation, distribution law of nucleotides and to measure some fundamental parameters by using frequency count.

Let

$$\mathcal{A} \stackrel{\text{def}}{=} \{\text{A}, \text{C}, \text{G}, \text{T}\} \tag{1}$$

be the finite set (alphabet) of nucleotides (nucleic acids): adenine (A), cytosine (C), guanine (G), thymine (T), and

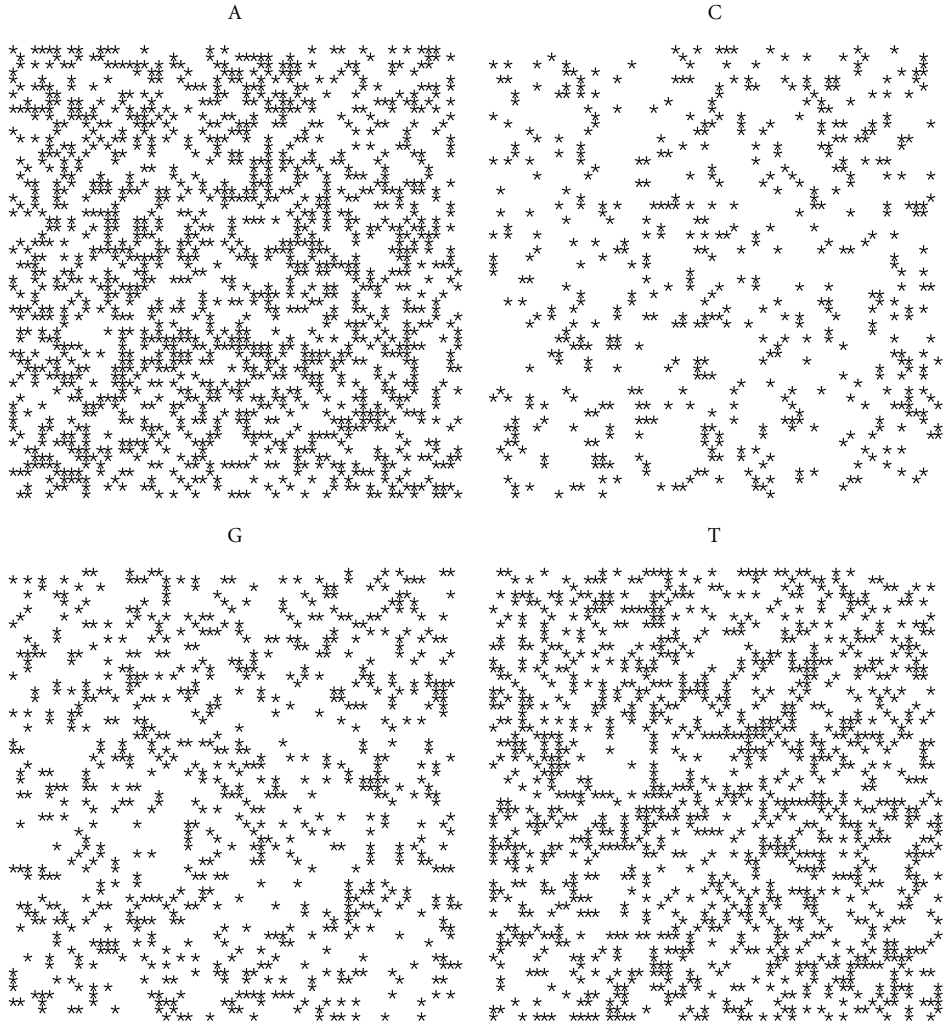

Figure 6: Spiral distribution of the first 3752 nucleotides for *Mortierella verticillata*.

$x \in \mathcal{A}$ any member of the alphabet. Nucleic acids are further grouped according to their ligand properties as

(a) purine {A, G}, pyrimidine {C, T},

(b) amino {A, C}, keto {G, T},

(c) weak hydrogen bonds {A, T}, strong hydrogen bond {G, C}.

A DNA sequence is the finite symbolic sequence

$$\mathcal{S} = \mathbb{N} \times \mathcal{A} \tag{2}$$

so that

$$\mathcal{S} \stackrel{\text{def}}{=} \{x_h\}_{h=1,\ldots,N}, \quad N < \infty \tag{3}$$

with

$$x_h \stackrel{\text{def}}{=} (h, x) = x(h), \quad (h = 1, 2, \ldots, N;\ x \in \mathcal{A}) \tag{4}$$

being the nucleotide x at the position h.

In general we can define an ℓ-length alphabet as follows: let the ℓ-length DNA word be defined by the ℓ-combination of the 4 nucleotides (1). For each fixed length ℓ there are 4^ℓ words, however not all of them can be considered, from biological point of view, as independent instances (see, e.g., Table 1), for this we define the ℓ-length alphabet as the set of ℓ-length independent words:

$$\mathcal{A}_\ell \stackrel{\text{def}}{=} \{a_1, a_2, \ldots, a_{M_\ell}\}, \qquad M_\ell \stackrel{\text{def}}{=} |\mathcal{A}_\ell| \le 4^\ell \tag{5}$$

with $|\cdots|$ cardinality of the set and

$$\ell \stackrel{\text{def}}{=} \text{length}\left(a_j\right), \quad (j = 1, \ldots, M_\ell). \tag{6}$$

For instance with $\ell = 1$, the alphabet is $\mathcal{A}_1 = \mathcal{A} = \{A, C, G, T\}$, with $\ell = 3$ the alphabet is given by the 20 amino acids

$$\mathcal{A}_3 = \{M, E, Q, D, R, T, N, H, V, G, L, S, P, F, I, C, A, K, Y, W\} \tag{7}$$

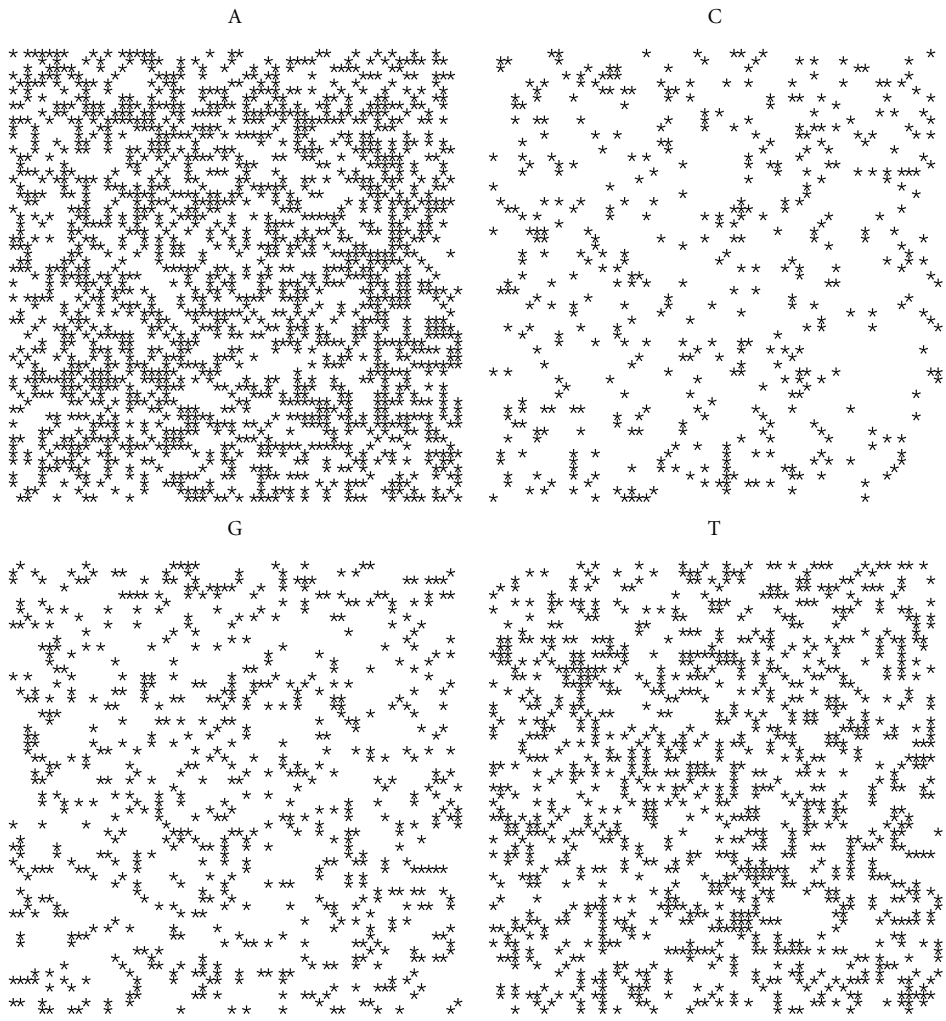

Figure 7: Spiral distribution of the first 3752 nucleotides for *Blattabacterium* sp.

each amino acid being represented by a 3-length word of Table 1.

Let $\mathcal{S}_N$ be an N-length ordered sequence of nucleotides $\{A, C, G, T\}$ and $\mathcal{A}_\ell$ the chosen alphabet, a DNA sequence of words is the finite symbolic sequence

$$\mathcal{D}_\ell(S_N) = \mathbb{N} \times \mathcal{A}_\ell \tag{8}$$

so that

$$\mathcal{D}_\ell(S_N) \stackrel{\text{def}}{=} \{x_h\}_{h=1,\ldots,N}, \quad (x \in \mathcal{A}_\ell;\ N < \infty) \tag{9}$$

with

$$x_h \stackrel{\text{def}}{=} (h, x), \quad (h = 1, 2, \ldots, N;\ x \in \mathcal{A}_\ell) \tag{10}$$

being the word x at the position h.

3.1. Indicator Matrix. The 2D indicator function, based on the 1D definition given in [25], is the map

$$u : \mathcal{S} \times \mathcal{S} \longrightarrow \{0, 1\} \tag{11}$$

such that

$$u(x_h, x_k) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } x_h = x_k, \\ 0 & \text{if } x_h \neq x_k, \end{cases} \quad (x_h \in \mathcal{S},\ x_k \in \mathcal{S}), \tag{12}$$

with

$$u(x_h, x_k) = u(x_k, x_h), \qquad u(x_h, x_h) = 1 \tag{13}$$

and, where for short, we have assumed

$$\mathcal{S} \stackrel{\text{def}}{=} \mathcal{D}_1(S_N). \tag{14}$$

According to (12), 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 (see also [3–5])

$$u_{hk} \stackrel{\text{def}}{=} u(x_h, x_k), \quad (x_h \in \mathcal{S},\ x_k \in \mathcal{S};\ h, k = 1, \ldots, N), \tag{15}$$

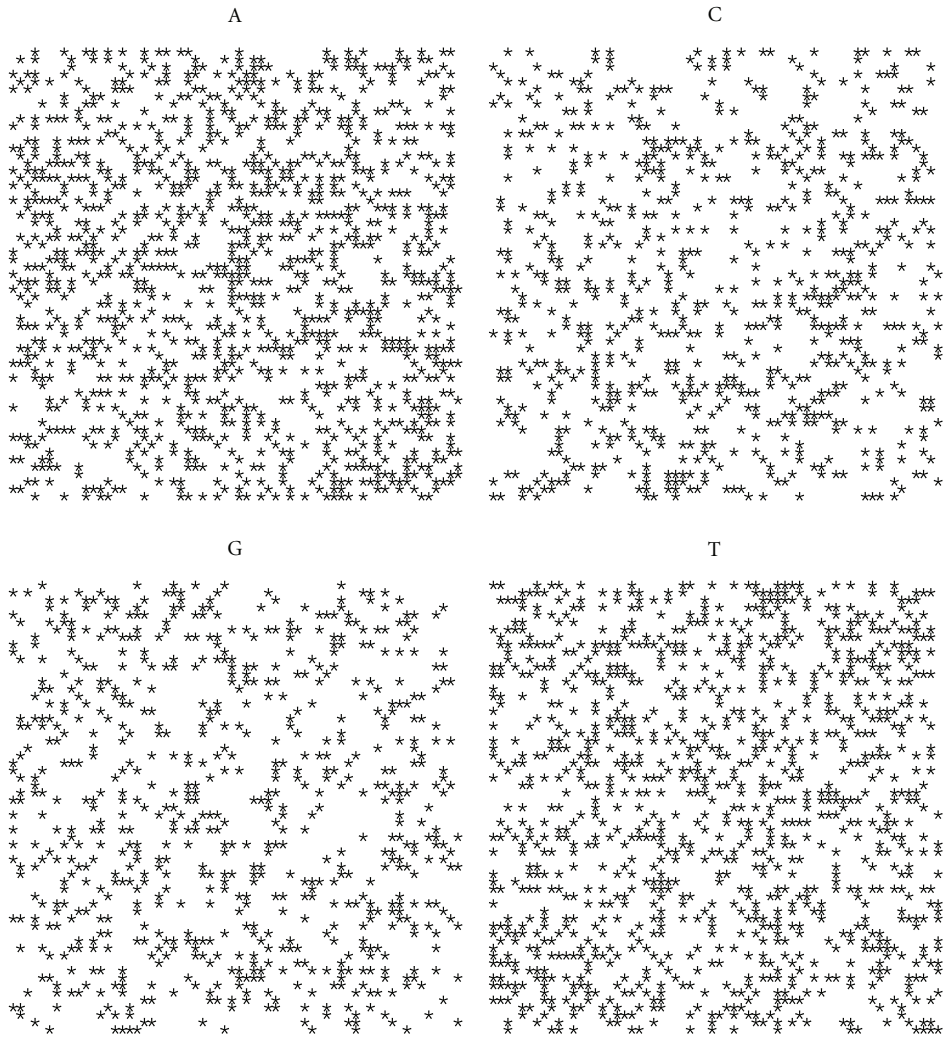

Figure 8: Spiral distribution of the first 3752 nucleotides for *Aeropyrum pernix* K1.

being, explicitly

$$\begin{array}{c|cccccccccc} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \iddots \\ G & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ C & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ A & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & \cdots \\ A & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & \cdots \\ T & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ A & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & \cdots \\ C & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ T & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ G & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ A & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & \cdots \\ \hline u_{hk} & A & G & T & C & A & T & A & A & C & G & \cdots \end{array} \tag{16}$$

This squared matrix can be plotted in 2 dimensions by putting a black dot where $u_{hk} = 1$ and white spot when $u_{hk} = 0$ (Figure 1) thus giving rise to the two-dimensional dot plot, which is a special case of the *recurrence plot* [26].

A simple generalization of this matrix can be considered for the alphabets $\mathcal{A}_\ell$, as follows. By choosing the 3 alphabet of amino acids, the 2D indicator function is the map

$$u : \mathcal{D}_3(S_N) \times \mathcal{D}_3(S_N) \longrightarrow \{0, 1\} \tag{17}$$

such that

$$u(x_h, x_k) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } x_h = x_k, \\ 0 & \text{if } x_h \neq x_k, \end{cases} \quad (x_h \in \mathcal{D}_3(S_N),\ x_k \in \mathcal{D}_3(S_N)), \tag{18}$$

with

$$u(x_h, x_k) = u(x_k, x_h), \qquad u(x_h, x_h) = 1. \tag{19}$$

Table 1: Correspondence codons to amino acids.

	Amino acid		Codon
1	M	Methionine	ATG
2	E	Glutamic acid	GAA, GAG
3	Q	Glutamine	CAA, CAG
4	D	Aspartic acid	GAT, GAC
5	R	Arginine	CGT, CGC, CGA, CGG, AGA, AGG
6	T	Threonine	ACT, ACC, ACA, ACG
7	N	Asparagine	AAT, AAC
8	H	Histidine	CAT, CAC
9	V	Valine	GTT, GTC, GTA, GTG
10	G	Glycine	GGT, GGC, GGA, GGG
11	L	Leucine	TTA, TTG, CTT, CTC, CTA, CTG
12	S	Serine	TCT, TCC, TCA, TCG, AGT, AGC
13	P	Proline	CCT, CCC, CCA, CCG
14	F	Phenylalanine	TTT, TTC
15	I	Isoleucine	ATT, ATC, ATA
16	C	Cysteine	TGT, TGC
17	A	Alanine	GCT, GCC, GCA, GCG
18	K	Lysine	AAA, AAG
19	Y	Thyroxine	TAT, TAC
20	W	Tryptophan	TGG
		Stop	TAA, TAG, TGA

According to (12), the indicator, on the 3-alphabet of amino acids of an N-length sequence can be easily represented by the $N \times N$ sparse symmetric matrix of binary values $\{0, 1\}$:

$$u_{hk} \stackrel{\text{def}}{=} u(x_h, x_k), \quad (x_h \in \mathcal{D}_3(S_N),\ x_k \in \mathcal{D}_3(S_N);\ h, k = 1, \ldots, N), \tag{20}$$

being, explicitly

⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋰
M	1	0	0	0	0	0	0	0	0	1	⋯
Q	0	0	0	0	0	0	0	0	1	0	⋯
R	0	0	1	1	0	0	0	1	0	0	⋯
T	0	0	0	0	0	1	1	0	0	0	⋯
T	0	0	0	0	0	1	1	0	0	0	⋯
E	0	0	0	0	1	0	0	1	0	0	⋯
R	0	0	1	1	0	0	0	0	0	0	⋯
R	0	0	1	1	0	0	0	1	0	0	⋯
K	0	1	0	0	0	0	0	0	0	0	⋯
M	1	0	0	0	0	0	0	0	0	1	⋯
u_{hk}	M	K	R	R	E	T	T	R	Q	M	⋯

(21)

With the graphical representation of this matrix we can also show the correlation of amino acids.

3.2. Test Sequences. In the following, in order to single out the main features of biological sequences, we will compare the DNA sequence with some test sequences.

(1) Pseudorandom N-length sequence of nucleotides is the sequence $\{\mathcal{R}_i\}_{i=1,\ldots,N}^{\ell}$ where r_i is a symbol randomly chosen in the alphabet $\mathcal{A}_\ell$, like for example, $(\ell = 1)$:

$$\{\text{A}, \text{C}, \text{A}, \text{G}, \text{T}, \text{A}, \text{T}, \text{G}, \text{G}, \text{A}, \text{T}, \text{T}, \text{A}, \text{C}, \text{C}, \text{G}, \ldots\}. \tag{22}$$

(2) Pseudoperiodic N-sequence of nucleotides with period π is the direct sum of a given π-length pseudorandom sequence, such that $N = k\pi$, $(k \in \mathbb{N})$ and $\mathcal{R}_i = \mathcal{R}_{i+\pi}$, for example,

$$\{\text{A}, \text{C}, \text{A}, \text{G}, \text{A}, \text{C}, \text{A}, \text{G}, \text{A}, \text{C}, \text{A}, \text{G}, \text{A}, \text{C}, \text{A}, \text{G}, \ldots\}, \quad (\pi = 4). \tag{23}$$

When $\pi = 1$ we have a pseudorandom sequence.

If we plot the indicator matrix of some bacteria and compare it with a pseudorandom and periodic sequence, we can see that (Figure 1)

(1) the main diagonal is a symmetry axis for the plot;

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

(3) periodicity is detected by parallel lines to the main diagonal (Figure 1(**a2**));

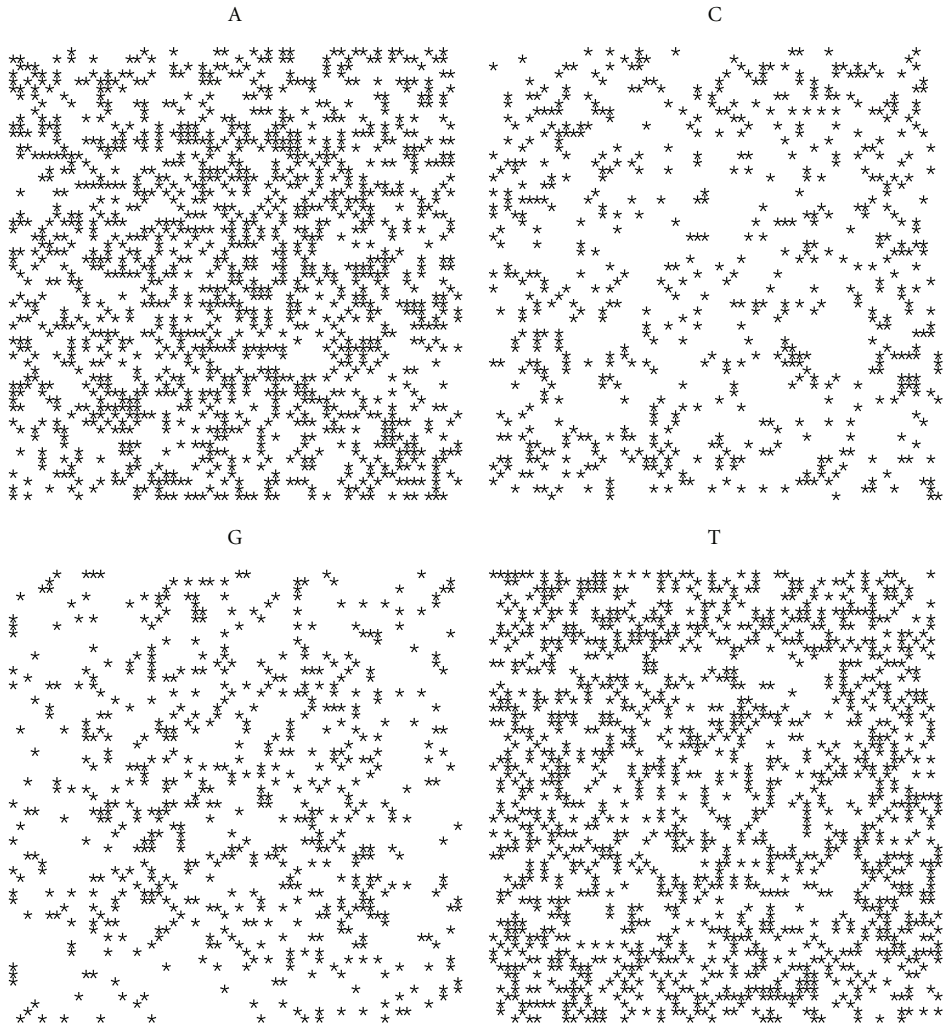

Figure 9: Spiral distribution of the first 3752 nucleotides for *Acidianus hospitalis* W1.

(4) empty spaces are more distributed than filled spaces, in the sense that the matrix u_{hk} is a sparse matrix (having more 0's than 1's);

(5) it seems that there are some square-like islands where black spots are more concentrated; these islands show the persistence of a nucleotide (Figures 1(**a2**) and 1(**b1**));

(6) the dot plot of archaea is very similar to the dot plot of a random sequence (Figures 1(**a1**) and 1(**h3**)).

It can be noticed that DNA sequences of a living organism resemble (Figure 1) random sequences, with some short range influence, built on the same alphabet. This has been taken as an axiom of nucleotides distribution, so that DNA sequences are often considered as Markov chain [27]. However, there are some hidden rules in combining the nucleotides and these rules lead, during the evolution, to a steady distribution. In fact, the more primitive the sequence is, the more randomly distributed the nucleotides are. It seems that as a consequence of the evolution, nucleotides move from a disordered aggregation toward a more organized structure, shown by the growing islands in the dot plot. The biological evolution is such that the challenge for the self-organization might follow from random permutations of a primitive disordered sequence so that the organization, that is, the complexity, is only the result of many arbitrary permutations of randomness. During the challenge for complexity, DNA sequence becomes "less random" and it loses some kind of energy.

From the graphical representation of the indicator matrix for bacteria and amino acids we can see a more sparse matrix, but with some typical plots (Figure 2).

3.3. Spiral Plot. In this section we consider a 2D distribution of nucleotides, following the idea given by Ulam for the distribution of primes, along an Ulam-like spiral [28]. In order to find some patterns in their distribution, nucleotides are arranged along a rectangular spiral. This is equivalent to

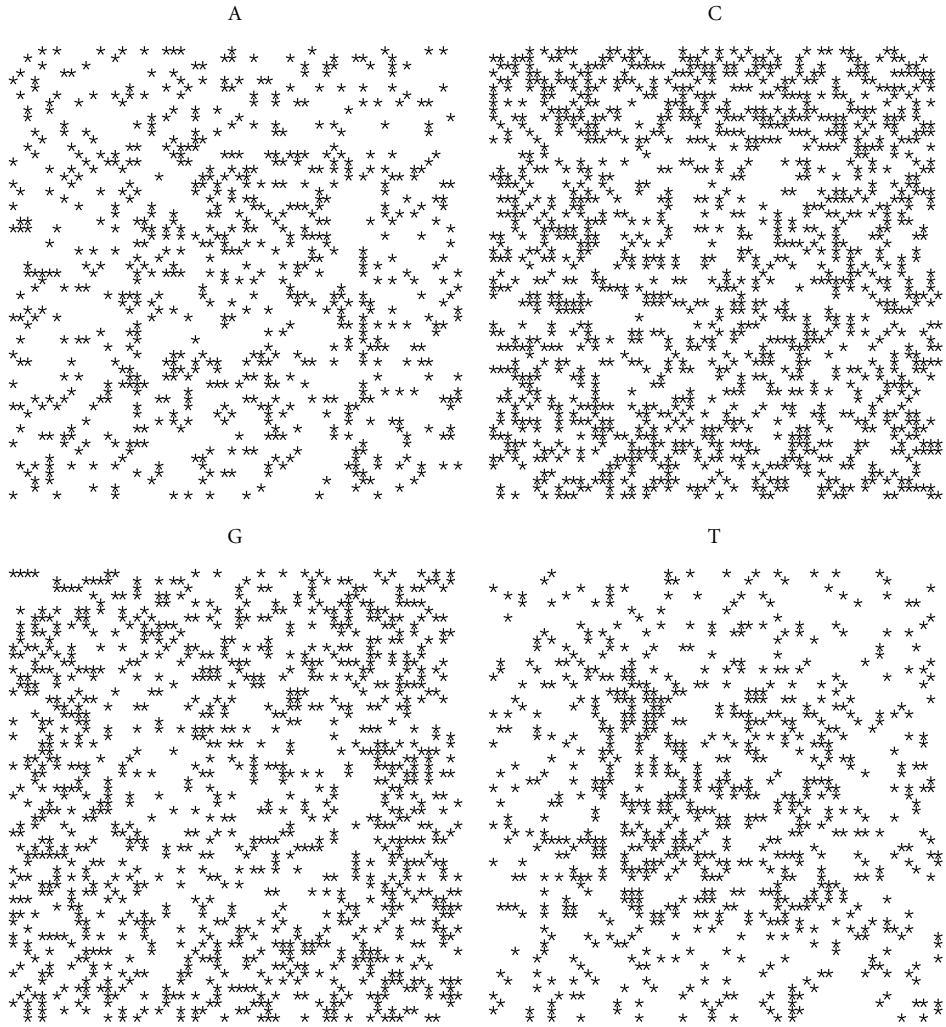

Figure 10: Spiral distribution of the first 3752 nucleotides for *Acidilobus saccharovorans* 345-15.

mapping the 1D sequence of integers into a 2D sequence as follows:

$$\begin{array}{lll} X_1 & 1 & \{0,0\} \\ X_2 & 2 & \{1,0\} \\ X_3 & 3 & \{1,1\} \\ X_4 & 4 & \{0,1\} \\ X_5 & 5 & \{-1,1\} \\ X_6 & 6 & \{-1,0\} \\ X_7 & 7 & \{-1,-1\} \\ X_8 & 8 & \{0,-1\} \\ X_9 & 9 & \{1,-1\} \\ X_{10} & 10 & \{2,-1\} \\ X_{11} & 11 & \{2,0\} \\ \vdots & \vdots & \vdots \end{array} \tag{24}$$

For instance the sequence

$$\{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{G},\mathrm{A},\mathrm{A},\mathrm{G},\mathrm{A},\mathrm{T},\mathrm{A},\mathrm{A},\mathrm{G},\ldots\} \tag{25}$$

distributed along the spiral looks like Figure 3.

For each nucleotide we can draw a spiral containing the distribution of only one acid nucleic. To each organism there correspond four plots, for A, C, G, T, respectively.

Let us first note that on a random sequence (Figure 4) the four distribution are equivalent.

By comparing the spirals of bacteria, random and archaea (Figures 4, 5, 6, 7, 8, 9, 10) we can see that there is a different distribution of each nucleotide. However the more evolved organism tends to have a higher percentage of weak hydrogen bonds (Figures 5, 6 and 7), so that we can assume the following.

Conjecture 1. *During the evolution, the distribution of nucleotides changes in a such way that strong hydrogen bonds tend to become weak.*

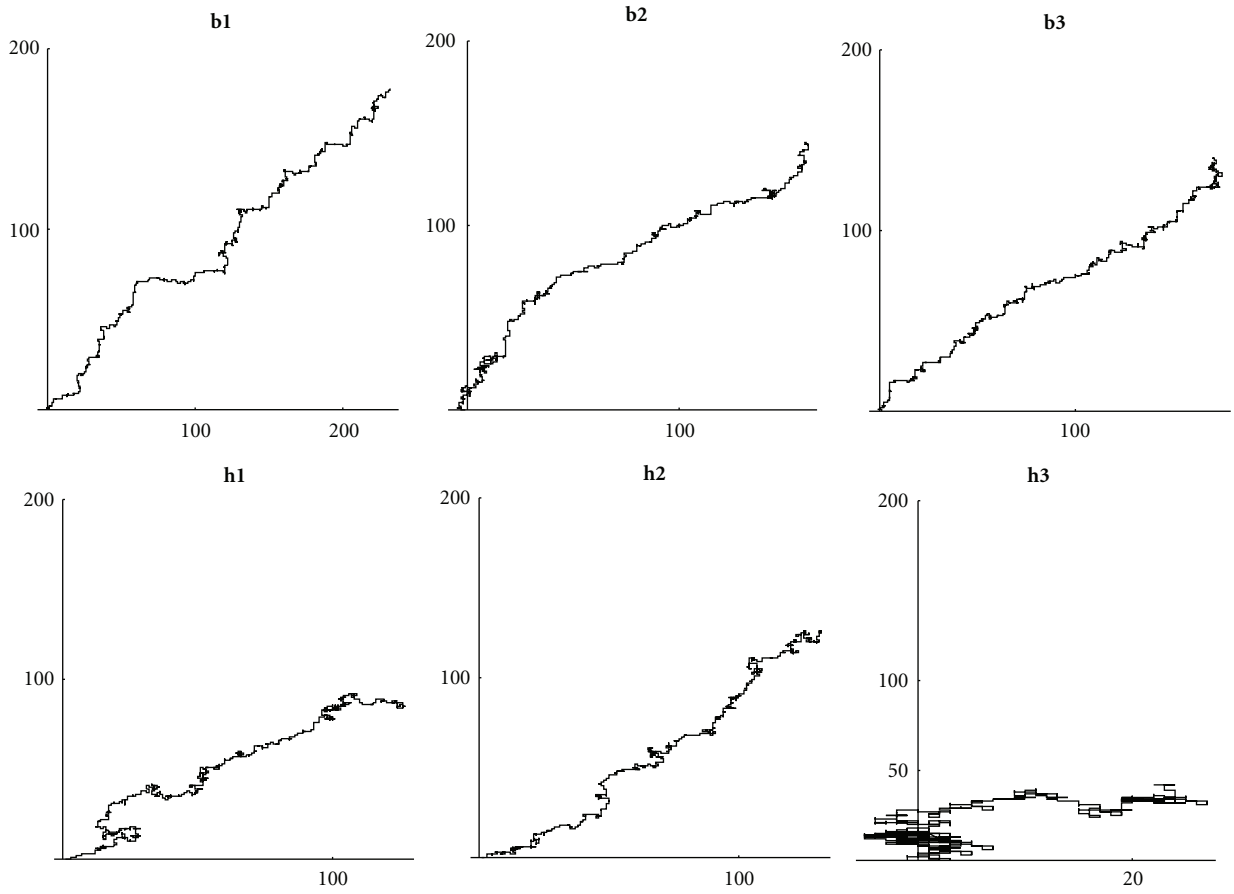

Figure 11: Walks on the first 200 nucleotides: (**b1**) *Mycoplasma putrefaciens*, (**b2**) *Mortierella verticillata*, (**b3**) *Blattabacterium*, (**h1**) *Aeropyrum pernix*, (**h2**) *Acidianus hospitalis*, and (**h3**) *Acidilobus saccharovorans*.

It should be noticed that along these spirals, there is a one-to-one map λ between $\mathbb{N}$ and the points of the spiral (with integer coordinates) in $\mathfrak{R}^2$

$$\lambda : \mathbb{N} \longmapsto \gamma \subset \mathfrak{R} \times \mathfrak{R} \tag{26}$$

so that

$$\lambda(n) = (a,b), \quad (n \in \mathbb{N};\ (a,b) \in \gamma \subset \mathfrak{R} \times \mathfrak{R};\ a \in \mathbb{Z},\ b \in \mathbb{Z}),$$
$$\lambda^{-1}(a,b) = n. \tag{27}$$

This bijective map can be considered also between $\mathbb{N}$ and the complex space $\mathbb{C}$ so that each natural number corresponds to a complex number (with integer coefficients)

$$\lambda(n) = z \stackrel{\text{def}}{=} a + ib, \quad (n \in \mathbb{N};\ a,b \in \mathbb{Z};\ z \in \mathbb{C}). \tag{28}$$

Since these spirals seem to fill in a finite region of the plane we can evaluate the complexity of each curve by typical fractal measures.

4. Parameters of Complexity

In this section we define some parameters, based on frequency distribution, which can measure the complexity of a DNA by computing the complexity of its representation in the complex plane (for a more detailed analysis see [29] and references therein).

Let $\mathcal{S}_N$ be an N-length-ordered sequence of nucleotides, and

$$p_x(h), \quad x \in \mathcal{A}_1 = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\} \tag{29}$$

be the probability to find the nucleotide x at the position $h, 1 \le h \le N$. According to (12) we define

$$\begin{aligned} a_h &\stackrel{\text{def}}{=} \sum_{j=1}^{h} u_{Aj}, & c_h &\stackrel{\text{def}}{=} \sum_{j=1}^{h} u_{Cj}, \\ g_h &\stackrel{\text{def}}{=} \sum_{j=1}^{h} u_{Gj}, & t_h &\stackrel{\text{def}}{=} \sum_{j=1}^{h} u_{Tj}, \\ & & & (1 \le h \le N) \end{aligned} \tag{30}$$

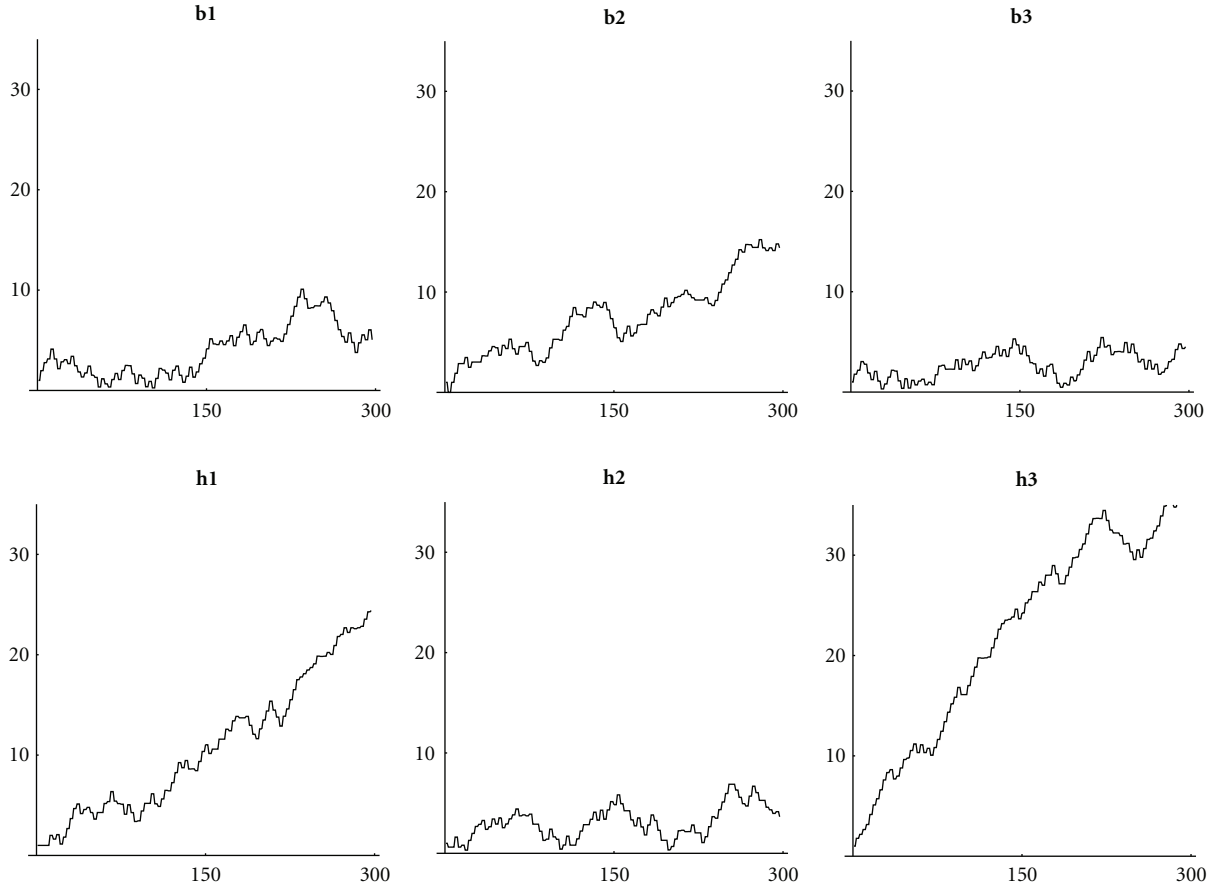

Figure 12: Absolute value of walks on the first 100 amino acids: (**b1**) *Mycoplasma putrefaciens*, (**b2**) *Mortierella verticillata*, (**b3**) *Blattabacterium*, (**h1**) *Aeropyrum pernix*, (**h2**) *Acidianus hospitalis*, (**h3**) *Acidilobus saccharovorans*.

as the number of nucleotides in the h-length segment of $\mathcal{S}_N$, so that

$$a_h + c_h + g_h + t_h = h. \tag{31}$$

The corresponding frequencies are

$$v_x(h) \stackrel{\text{def}}{=} \frac{1}{h}\sum_{j=1}^{h} u_{xj}, \quad x \in A_1, \ (1 \le h \le N), \tag{32}$$

so that

$$\begin{aligned} v_A(h) &= \frac{a_h}{h}, \qquad v_C(h) = \frac{c_h}{h}, \\ v_G(h) &= \frac{g_h}{h}, \qquad v_T(h) = \frac{t_h}{h}. \end{aligned} \tag{33}$$

We can assume that for large sequences

$$p_x(h) \cong v_x(h). \tag{34}$$

4.1. Randomness. Since for a random sequence the frequencies of nucleotides coincide for large n,

$$v_A(n) \cong v_C(n) \cong v_G(n) \cong v_T(n) \tag{35}$$

Table 2: Randomness.

Mycoplasma putrefaciens	0.696				
Mortierella verticillata			0.779		
Blattabacterium		0.743			
Aeropyrum pernix					0.982
Acidianus hospitalis			0.828		
Acidilobus saccharouorans				0.934	
pseudorandom					0.999

we can define as randomness index the following:

$$\mathcal{R} \stackrel{\text{def}}{=} 1 - \sigma(v_A(n), v_C(n), v_G(n), v_T(n)) \tag{36}$$

with σ being the variance, so that $\mathcal{R} = 1$ for random sequence and $\mathcal{R} = 0$ for a nonrandom sequence. Over the first 10000 nucleotides we have the randomness value of Table 2.

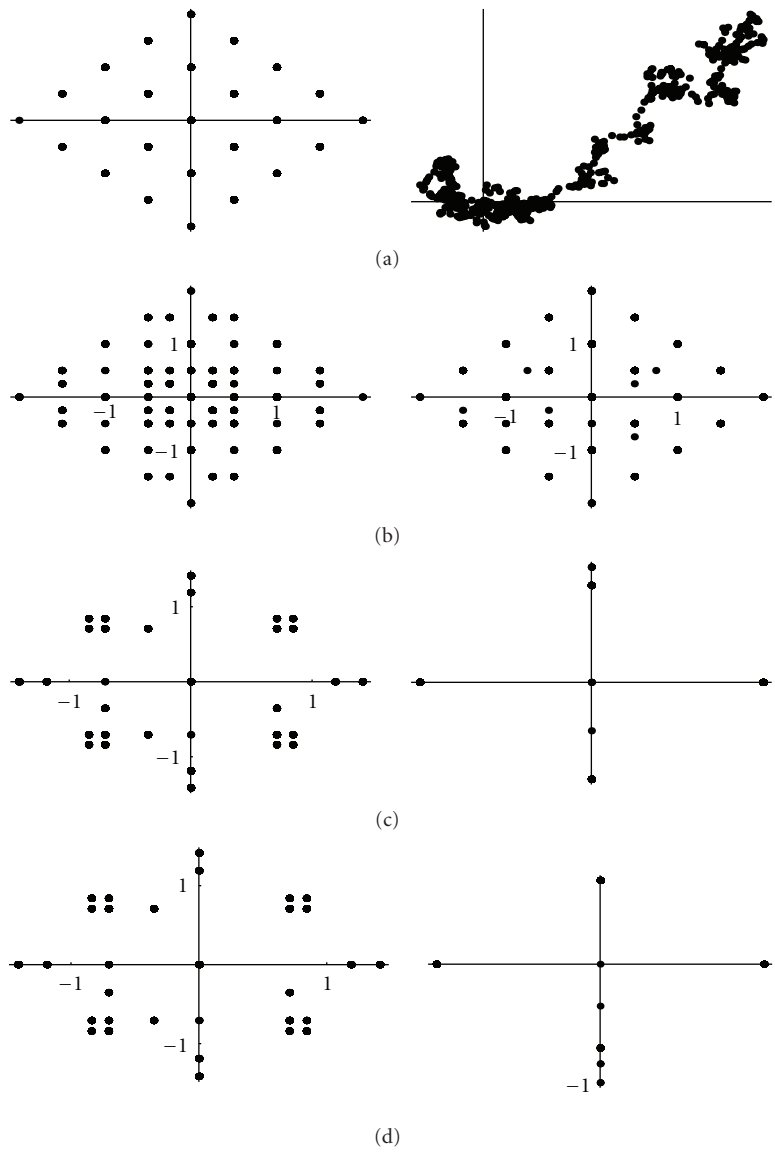

Figure 13: Cluster analysis of the 4th short Haar wavelet transform of a 4000-length random sequence (left) and its 2000-length random walk (right): (a) (α, α^*); (b) $(\beta_0^0, \beta^{*0}_0)$; (c) $(\beta_0^1, \beta^{*1}_0)$; (d) $(\beta_1^1, \beta^{*1}_1)$.

However, if we compute the randomness index over the frequencies of amino acids in the $\mathcal{A}_3$ alphabet then we can observe a different distribution of values. Over the first 30000 nucleotides corresponding to 10000 amino acids, we have the randomness value of Table 3.

So that we can comment that the arising complexity of the words and alphabets shows a different randomness in each alphabet.

4.2. Complexity. As a simple measure of complexity [30–32], for an n-length sequence, the following has been proposed [33]:

$$K = \frac{1}{n} \log \frac{n!}{a_n! c_n! g_n! t_n!}. \tag{37}$$

In Table 4 the complexity of the first 100-length segment of the DNA sequences is computed. It is interesting to notice the more similarities between the archaea Acidilobus with the pseudorandom sequence than with the pseudoperiodic.

Table 3: Randomness of amino acids distribution.

Mycoplasma putrefaciens		0.946			
Mortierella verticillata	0.938				
Blattabacterium			0.953		
Aeropyrum pernix				0.962	
Acidianus hospitalis	0.916				
Acidilobus saccharouorans			0.950		
pseudorandom					0.963

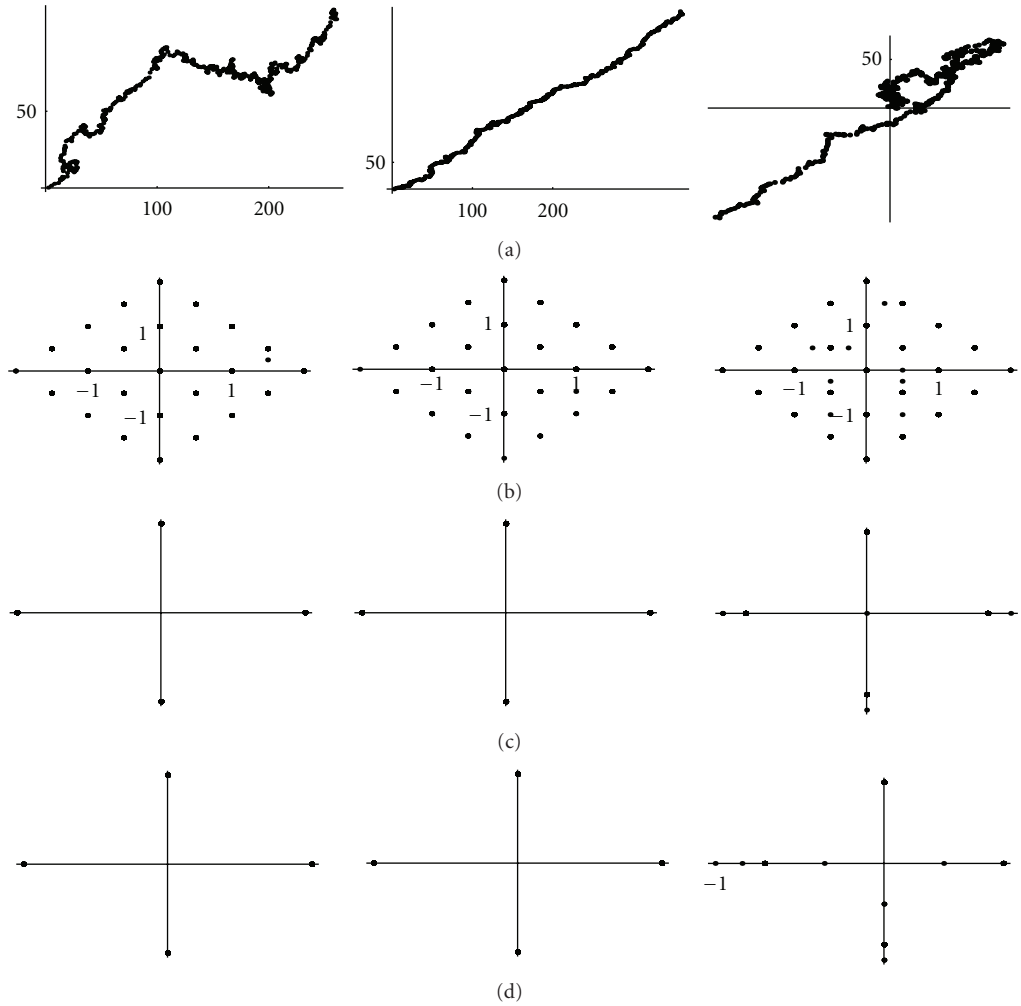

Figure 14: Cluster analysis of the 4th short Haar wavelet transform of the complex representation for a DNA walk on the first 2000 nucleotides of (**h1**) Aeropyrum, (**h2**) Acidianus, (**h3**) Acidilobus saccharovorans in the planes: (a) (α, α^*); (b) $(\beta_0^0, \beta^{*0}_0)$; (c) $(\beta_0^1, \beta^{*1}_0)$; (d) $(\beta_1^1, \beta^{*1}_1)$.

Table 4: Complexity.

Mycoplasma putrefaciens	1.151				
Mortierella verticillata			1.285		
Blattabacterium		1.197			
Aeropyrum pernix			1.212		
Acidianus hospitalis				1.231	
Acidilobus saccharouorans					1.296
Pseudorandom				1.295	

Nucleotide distribution in primitive biosequences is more likely random than pseudodeterministic. Moreover, the evolution reduces the complexity of the sequence.

4.3. Fractal Dimension. The fractal dimension is computed on the dot plot, by the box counting algorithm [34, 35], as the average of the number $p(n)$ of 1's in the randomly taken $n \times n$ minors of the $N \times N$ indicator matrix u_{hk} or equivalently the number $p(n)$ of black dots in the randomly taken $n \times n$ squares over the dot plot

$$D = \frac{1}{2N} \sum_{n=2}^{N} \frac{\log p(n)}{\log n}. \tag{38}$$

The explicit computation enables us to compare the fractal dimension on the first 100-length segments of DNA chains, with an approximation up to 10^{-3} (see Table 5).

If we compare the fractal dimensions of the bacteria with pseudorandom and pseudoperiodic we can see that the fractal dimension of nucleotide distribution ranges, for all variants, in the interval [1.28–1.30]. As expected, the more "random" sequences have higher fractal dimension.

4.4. Entropy. Another fundamental parameter, related to the information content of a sequence which measures the heterogeneity of data, is the information entropy (or Shannon entropy) [36–42]. Based on the axiom that less information

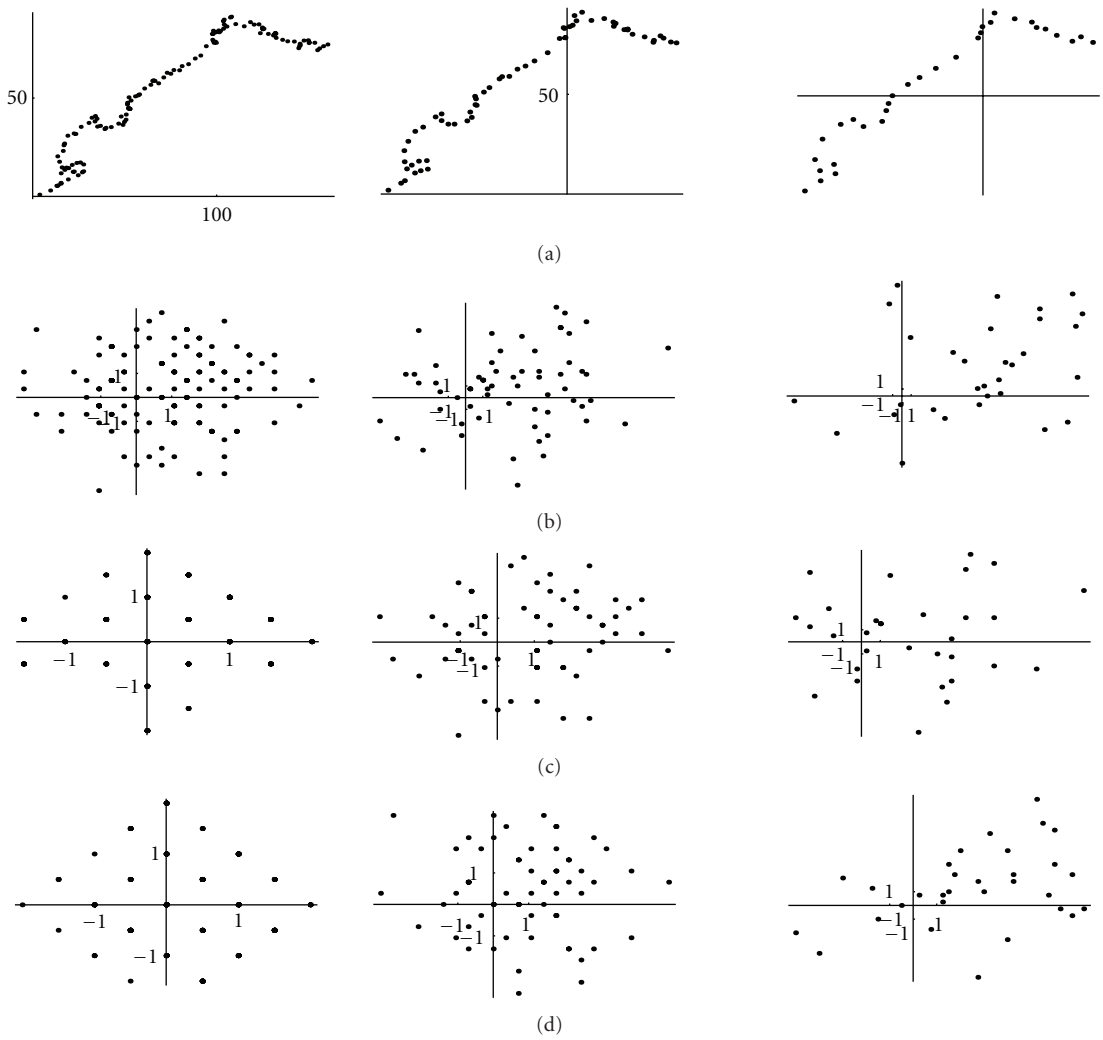

Figure 15: Cluster analysis of the 8th (left), 16th (middle column), 32th (right) short Haar wavelet transform of the DNA walk on the first 1000 nucleotides of h1 (Aeropyrum) in the planes: (a) (α, α^*); (b) $(\beta_0^0, \beta^{*0}_0)$; (c) $(\beta_0^1, \beta^{*1}_0)$; (d) $(\beta_1^1, \beta^{*1}_1)$.

Table 5: Fractal dimensions.

Mycoplasma putrefaciens	1.283				
Mortierella verticillata			1.296		
Blattabacterium		1.287			
Aeropyrum pernix			1.288		
Acidianus hospitalis				1.290	
Acidilobus saccharouorans					1.297
pseudorandom					1.298
pseudoperiodic	1.285				

implies a larger uncertainty and vice versa that more information leads us to a more deterministic model, the entropy concept has been recently offering some interesting interpretations about uncertainty in DNA. In fact, DNA as any other signal has been considered as a sequence of symbols carrying chemical-functional information.

The normalized Shannon entropy [39, 40, 42] is defined, over the alphabet $\mathcal{A}_\ell$, as

$$H(n) = -\frac{1}{\log \ell} \sum_{x \in \mathcal{A}_\ell} p_x(n) \times \begin{cases} \log p_x(n) & \text{if } p_x(n) \neq 0, \\ 0 & \text{if } p_x(n) = 0, \end{cases} \tag{39}$$

where $p_x(n)$ should be computed for large sequences. According to (32), (34), we will approximate its value with

$$p_x(n) \cong \frac{1}{n} \sum_{i=1}^{n} u_{xi}, \quad (x \in \mathcal{A}_\ell,\ 1 \leq n \leq N). \tag{40}$$

However, the entropy is a parameter very similar to the complexity. In fact, it can be easily seen that (for the proof see [29]) the entropy H and the measure of complexity K differ for a factor. There follows that the entropy does not give any

Table 6: Shannon entropy.

Mycoplasma putrefaciens	0.877				
Mortierella verticillata				0.976	
Blattabacterium		0.911			
Aeropyrum pernix			0.922		
Acidianus hospitalis				0.937	
Acidilobus saccharovorans					0.984
pseudorandom					0.984

new information comparing with the previous parameters. As expected also the table of entropies classifies bacteria and archaea in the same way (Table 6).

5. Complex Root Representation of DNA Words

The complex (digital) representation of a DNA sequence of words is the map of the symbolic sequence of words into a set of complex numbers and it is defined as

$$\mathcal{D}_\ell(S_N) \xrightarrow{\rho} \mathbb{C} \tag{41}$$

such that for each $x_h \in \mathcal{D}_\ell(S_N)$ it is $\rho(x_h) \in \mathbb{C}$.

The complex root representation of the sequence S_N is the sequence $\mathcal{D}_\ell(S_N)$ of complex numbers $\{y_h\}_{h=1,\dots,N}$ defined as

$$y_h = \rho(x_h) \stackrel{\text{def}}{=} e^{2\pi i(j-1)/|\mathcal{A}_\ell|}, \quad (j = 1,\dots,|\mathcal{A}_\ell|,\ h = 1,\dots,N) \tag{42}$$

with $i = \sqrt{-1}$ being the imaginary unit. There follows that, independently on the alphabet, it is

$$|y_h| = \left|e^{2\pi i(j-1)/|\mathcal{A}_\ell|}\right| = 1, \quad (\forall \ell;\ h = 1,\dots,N) \tag{43}$$

being all complex roots, of the unit, located on the unit circle of the complex plane $\mathbb{C}^1$.

For instance, with $\mathcal{A}_1 = \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, the cardinality of the alphabet is $|\mathcal{A}_1| = 4$ and

$$\begin{aligned} \rho(\mathrm{A}) &= e^{0/4} = 1, \quad j = 1, \\ \rho(\mathrm{C}) &= e^{\pi i/2} = i, \quad j = 2, \\ \rho(\mathrm{G}) &= e^{\pi i} = -1, \quad j = 3, \\ \rho(\mathrm{T}) &= e^{\pi i 3/2} = -i, \quad j = 4. \end{aligned} \tag{44}$$

Analogously, with $\mathcal{A}_3 = \{M, E, \dots, W\}$ it is $|\mathcal{A}_3| = 20$ and the 20 complex roots of unit

$$\rho(x_n) = e^{2\pi i(n-1)/20}, \quad (n = 1,\dots,20;\ x_n \in \mathcal{A}_3) \tag{45}$$

so that explicitly is

$$\begin{aligned} \rho(M) &= e^{2\pi i 0/20} = 1, && j = 1, \\ \rho(E) &= e^{\pi i/10} = \frac{1}{4}\left[\sqrt{2\left(5+\sqrt{5}\right)} + i\left(\sqrt{5}-1\right)\right], && j = 2, \\ \rho(Q) &= e^{\pi i/5} = \frac{1}{4}\left[1+\sqrt{5}+i\sqrt{2\left(5-\sqrt{5}\right)}\right], && j = 3, \\ &\vdots && \vdots \\ \rho(W) &= e^{\pi i 19/10} = \frac{1}{4}\left[\sqrt{2\left(5+\sqrt{5}\right)} - i\left(\sqrt{5}-1\right)\right], && j = 20. \end{aligned} \tag{46}$$

Therefore the complex representation of a DNA sequence is a sequence of complex numbers

$$y_h = \xi_h + \eta_h i, \quad \xi_h = \Re(y_h), \quad \eta_h = \Im(y_h) \tag{47}$$

with y_h given by (42).

An n-length pseudorandom (white noise) complex sequence belonging to the unit circle can be defined directly by using some random exponents

$$R_n \stackrel{\text{def}}{=} (-1)^{r_n} i^{s_n}, \quad |R_n| = 1, \tag{48}$$

with $r_n,\ s_n$ being random values in the set $\{0, \mathbb{N}\}$.

5.1. Random Walks. Random walk on the complex sequence $\mathbf{Y}_N$ is defined as the series $\mathbf{Z}_N = \{z_n\}_{n=1,\dots,N}$

$$z_n \stackrel{\text{def}}{=} \sum_{k=1,\dots,n} y_k, \quad n = 1,\dots,N \tag{49}$$

which is the cumulative sum

$$\left\{y_1, y_1 + y_2, \dots, \sum_{s=1}^{n} y_s \dots, \sum_{s=1}^{N} y_s\right\}. \tag{50}$$

When $y_k = \rho(x_k)$ with $x_k \in \mathcal{A}_\ell$ and $\mathbf{X}_k \in S_N$ we will properly call these walks as DNA walk. When the y_k are randomly generated we will call them random walks.

By remembering the definition of frequencies, DNA walk is the complex value signal $\{Z_n\}_{n=0,\dots,N-1}$ with

$$z_n = (\Re[z_n], \Im[z_n]) = (a_n - g_n) + (t_n - c_n)i, \quad z_n \in \mathbb{C}_1, \tag{51}$$

where the coefficients $a_n,\ g_n,\ t_n,\ c_n$ given by (12) fulfill the condition (31).

If we compare the DNA walks (Figure 11) some primitive archaea such as h3 are very similar to a random walk (Figure 13). In particular archaea seem to grow less than other bacteria (with the exception of b2).

It is interesting also to notice that the random walks on amino acids (Figure 12) show that more evolved organisms have some "periodic" behavior, while the absolute value of walks on archaea is growing fast.

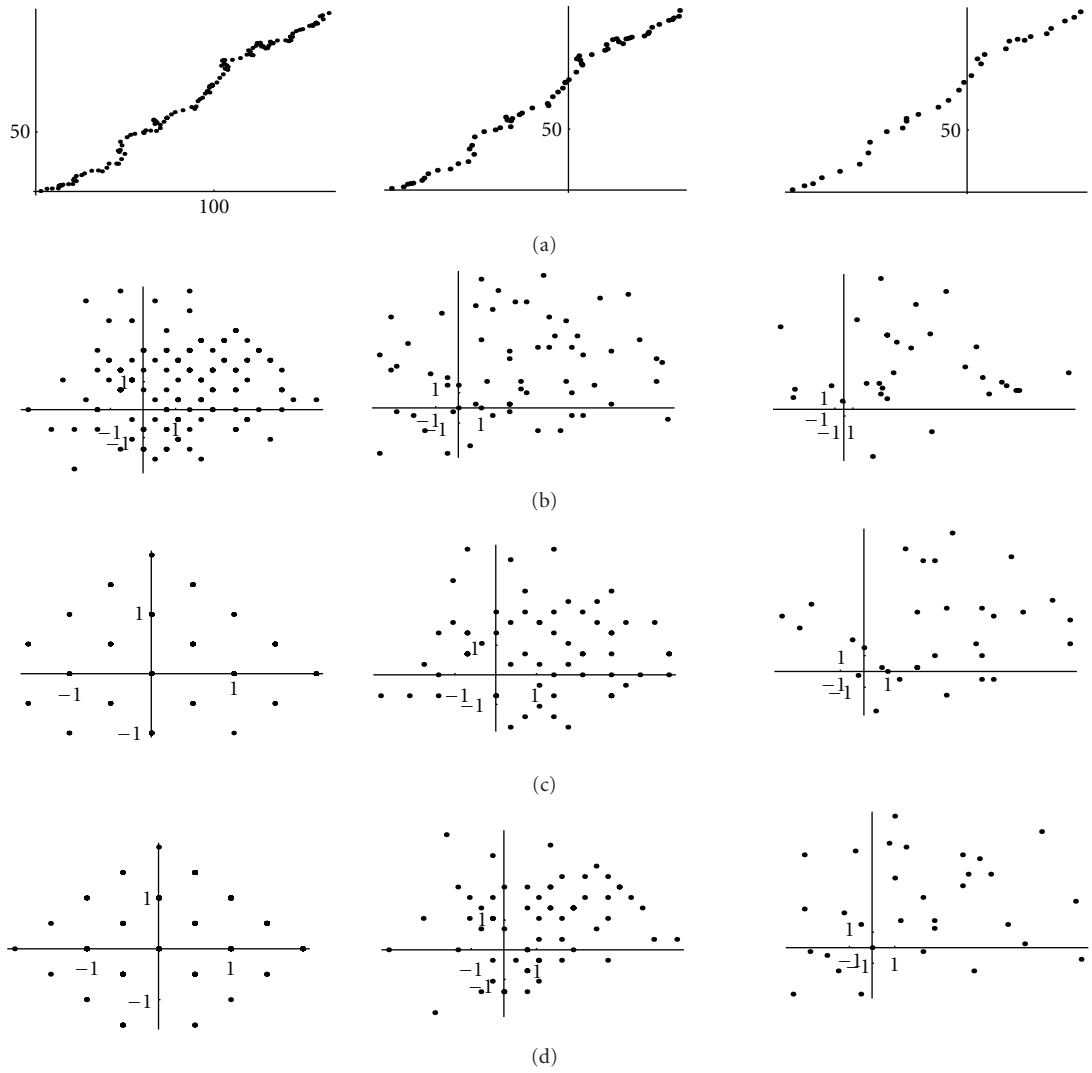

Figure 16: Cluster analysis of the 8th (left), 16th (middle column), 32th (right) short Haar wavelet transform of the DNA walk on the first 1000 nucleotides of h2 (Acidianus) in the planes: (a) (α, α^*); (b) $(\beta_0^0, \beta^{*0}_0)$; (c) $(\beta_0^1, \beta^{*1}_0)$; (d) $(\beta_1^1, \beta^{*1}_1)$.

6. Wavelet Analysis

Wavelet analysis is a powerful method extensively applied to the analysis of biological signals [12, 43–45] aiming to single out the most significant parameters of complexity and heterogeneity in a time series and, in particular, in a DNA sequence. This method is based on the analysis of wavelet coefficients which are obtained by the wavelet transform.

We will consider in the following the Haar wavelet basis (see, e.g., [3, 4, 29]) made by scaling functions:

$$\varphi_k^n(x) \stackrel{\text{def}}{=} 2^{n/2}\varphi(2^n x - k), \quad (0 \le n,\ 0 \le k \le 2^n - 1),$$

$$\varphi(2^n x - k) = \begin{cases} 1, & x \in \Omega_k^n,\ \Omega_k^n \stackrel{\text{def}}{=} \left[\dfrac{k}{2^n}, \dfrac{k+1}{2^n}\right), \\ 0, & x \notin \Omega_k^n, \end{cases} \tag{52}$$

and the *Haar wavelets*:

$$\psi_k^n(x) \stackrel{\text{def}}{=} 2^{n/2}\psi(2^n x - k), \quad \left\|\psi_k^n(x)\right\|_{L^2} = 1,$$

$$\psi(2^n x - k) \stackrel{\text{def}}{=} \begin{cases} -1, & x \in \left[\dfrac{k}{2^n}, \dfrac{k+1/2}{2^n}\right), \\ 1, & x \in \left[\dfrac{k+1/2}{2^n}, \dfrac{k+1}{2^n}\right), \\ & (0 \le n,\ 0 \le k \le 2^n - 1), \\ 0, & \text{elsewhere.} \end{cases} \tag{53}$$

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

$$\mathbf{Y} \equiv \{Y_i\}, \quad \left(i = 0, \ldots, 2^M - 1,\ 2^M = N < \infty,\ M \in \mathbb{N}\right) \tag{54}$$

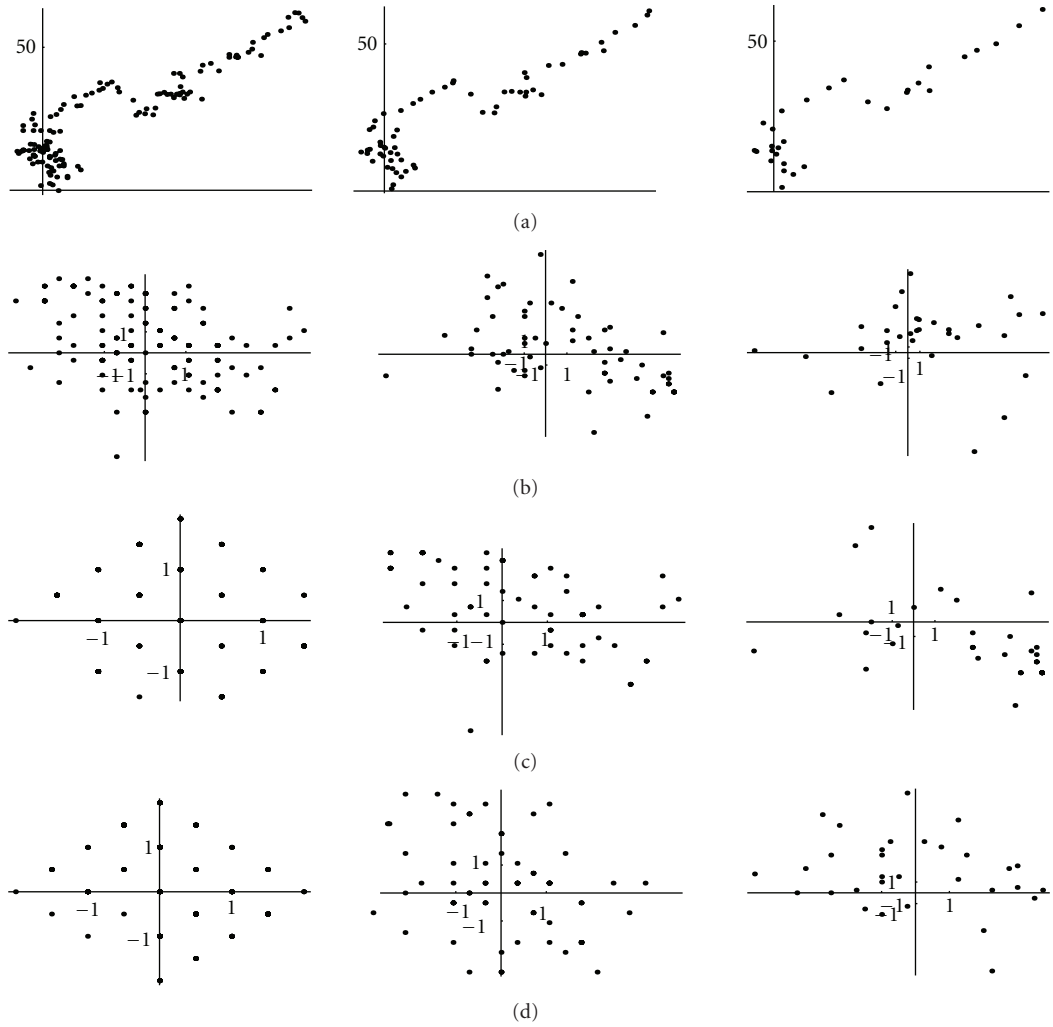

FIGURE 17: Cluster analysis of the 8th (left), 16th (middle column), 32th (right) short Haar wavelet transform of the DNA walk on the first 1000 nucleotides of h3 (Acidilobus saccharovorans) in the planes: (a) (α, α^*); (b) $(\beta_0^0, \beta^{*0}{}_0)$; (c) $(\beta_0^1, \beta^{*1}{}_0)$; (d) $(\beta_1^1, \beta^{*1}{}_1)$.

into the vector of *wavelet coefficients* $\boldsymbol{\beta}_N = \{\alpha, \beta_k^n\}$:

$$\begin{aligned} &\mathcal{W}_N \mathbf{Y} = \boldsymbol{\beta}_N, \\ &\boldsymbol{\beta}_N \stackrel{\text{def}}{=} \left\{\alpha, \beta_0^0, \ldots, \beta_{2^{M-1}-1}^{M-1}\right\}, \\ &\mathbf{Y} \stackrel{\text{def}}{=} \{Y_0, Y_1, \ldots, Y_{N-1}\}, \quad \left(2^M = N\right). \end{aligned} \tag{55}$$

The matrix $\mathcal{W}_N$ can be easily computed by some recursive product [3, 4, 13, 29, 46] so that with $N = 4,\ M = 2$, we have [3, 4, 29]

$$\mathcal{W}_4 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}. \tag{56}$$

From (55) with $M = 2,\ N = 4$, by explicit computation, we have

$$\alpha = \frac{1}{4}(Y_0 + Y_1 + Y_2 + Y_3) \tag{57}$$

and [1–3, 14]

$$\begin{aligned} \beta_0^0 &= \frac{1}{2}(Y_2 - Y_0 + Y_3 - Y_1), \\ \beta_0^1 &= \frac{1}{\sqrt{2}}(Y_0 - Y_1), \\ \beta_1^1 &= \frac{1}{\sqrt{2}}(Y_3 - Y_2). \end{aligned} \tag{58}$$

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.

In the following we will consider the short wavelet transform which consists in the subdivision of the DNA sequence

into 4-length segments and apply the wavelet transform to each segment. As a result, from the $N = 2^M$-length complex vector $\mathbf{Y}$, which is subdivided into 2^{M-2} segments, the 4-parameter short Haar wavelet transform gives the cluster of points

$$(\mathcal{W}^p \Re(\mathbf{Y}^s), \mathcal{W}^p \Im(\mathbf{Y}^s)), \quad s = 0, \ldots, \ \sigma = \frac{N}{p}, \ p = 4 \tag{59}$$

in the 8-dimensional space $\mathbb{R}^4 \times \mathbb{R}^4$, that is,

$$(\alpha, \alpha^*), \left(\beta_0^0, \beta^{*0}_0\right), \ldots, \left(\beta_{2^{p-1}-1}^{p-1}, \beta^{*p-1}_{2^{p-1}-1}\right), \quad p = 4. \tag{60}$$

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. It has been observed [3, 4, 29] that some symmetry arises from the plots of wavelet coefficients of DNA walks.

6.1. Cluster Analysis of the Wavelet Coefficients of the Complex DNA Representation. Let us first compute the clusters of wavelet coefficients for the random sequence (48). As can be seen the wavelet coefficients both for the sequence and for its series range in some discrete set of values (see Figure 13).

The cluster algorithm applied to the complex representation sequence shows that the values of the wavelet coefficients belong to some discrete finite sets (Figure 14).

It should be noticed that this symmetry on detail coefficients is lost for wavelet transform on longer segments (Figures 15, 16 and 17).

There follows that DNA sequences have to be considered as Markov chain with short range dependence; in other words any acid nucleic is attached to the chain on the base of a correlation of the previous acid nucleic. In other words, if we look for a dependence rule on the DNA nucleotides this dependence might be summarized by a function as

$$x_{n+1} = f(x_n), \quad (n = 1, \ldots, N). \tag{61}$$

7. Conclusions

In this paper archaea DNAs have been studied by focussing on the main parameters for complexity. It has been shown that more or less the main indices for complexity and heterogeneity, such as entropy, fractal dimension, and complexity do not differ too much when we have to classify the complexity of the sequence. However, some DNA sequences look more close to random sequences than others, thus suggesting that the evolution involves a process of complexity reduction: the more evolved a sequence is, the more far from a random distribution it is. In any case seems to be apparently impossible to distinguish between a random sequence and a DNA chain. By using the short wavelet transform instead we have shown that on short range (4-nucleotides) a DNA sequence shows some symmetries that slowly disappear by increasing the length of the analysed segment. Moreover, more evolved organisms have a more symmetrical distribution of wavelet coefficients.

References

[1] C. Cattani, "Complexrepresentation of DNA sequences," in *Proceedings of the Bioinformatics Research and Development Second International Conference*, M. Elloumi et al., Ed., Springer, Vienna, Austria, July 2008.

[2] C. Cattani, "Complex representation of DNA sequences," *Communications in Computer and Information Science*, vol. 13, pp. 528–537, 2008.

[3] 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, pp. 799–842, John Wiley & Sons, New York, NY, USA, 2010.

[4] C. Cattani, "Fractals and hidden symmetries in DNA," *Mathematical Problems in Engineering*, vol. 2010, Article ID 507056, pp. 1–31, 2010.

[5] C. Cattani and G. Pierro, "Complexity on acute myeloid leukemia mRNA transcript variant," *Mathematical Problems in Engineering*, vol. 2011, pp. 1–16, 2011.

[6] C. Cattani, G. Pierro, and G. Altieri, "Entropy and multi-fractality for the myeloma multiple TET 2 gene," *Mathematical Problems in Engineering*, vol. 2011, pp. 1–17, 2011.

[7] C. Cattani and J. J. Rushchitsky, *Wavelet and Wave Analysis as applied to Materials with Micro or Nanostructure, Series on Advances in Mathematics for Applied Sciences*, vol. 74, World Scientific, Singapore, 2007.

[8] 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.

[9] 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.

[10] M. Altaiski, O. Mornev, and R. Polozov, "Wavelet analysis of DNA sequences," *Genetic Analysis—Biomolecular Engineering*, vol. 12, no. 5-6, pp. 165–168, 1996.

[11] 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.

[12] 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.

[13] C. Cattani, "Haar wavelet-based technique for sharp jumps classification," *Mathematical and Computer Modelling*, vol. 39, no. 2-3, pp. 255–278, 2004.

[14] C. Cattani, "Harmonic wavelet approximation of random, fractal and high frequency signals," *Telecommunication Systems*, vol. 43, no. 3-4, pp. 207–217, 2010.

[15] M. Li, "Fractal time series-a tutorial review," *Mathematical Problems in Engineering*, vol. 2010, Article ID 157264, pp. 1–26, 2010.

[16] M. Li and J. Y. Li, "On the predictability of long-range dependent series," *Mathematical Problems in Engineering*, vol. 2010, Article ID 397454, pp. 1–9, 2010.

[17] M. Li and S. C. Lim, "Power spectrum of generalized Cauchy process," *Telecommunication Systems*, vol. 43, no. 3-4, pp. 219–222, 2010.

[18] National Center forBiotechnology Information, http://www.ncbi.nlm.nih.gov/genbank.

[19] Genome Browser, http://genome.ucsc.edu.

[20] European Informatics Institute, http://www.ebi.ac.uk.

[21] Ensembl, http://www.ensembl.org.

[22] J. L. Howland, *The Surprising Archaea*, Oxford University Press, New York, NY, USA, 2000.
[23] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 11, pp. 5088–5090, 1977.
[24] M. T. Madigan and B. L. Marrs, "Extremophiles," *Scientific American*, vol. 276, no. 4, pp. 82–87, 1997.
[25] 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.
[26] J. P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *Europhysics Letters*, vol. 5, pp. 973–977, 1987.
[27] J. Szczepański and T. Michałek, "Random fields approach to the study of DNA chains," *Journal of Biological Physics*, vol. 29, no. 1, pp. 39–54, 2003.
[28] M. Stein and S. M. Ulam, "An observation on the distribution of primes," *American Mathematical Monthly*, vol. 74, no. 1, p. 4344, 1967.
[29] C. Cattani, "Complexity and Simmetries in DNA sequences," in *Handbook of Biological Discovery*, M. Elloumi and A. Y. Zomaya, Eds., Wiley Series in Bioinformatics, chapter 22, pp. 700–742, John Wiley & Sons, New York, NY, USA, 2012.
[30] M. A. Gates, "Simpler DNA sequence representations," *Nature*, vol. 316, no. 6025, p. 219, 1985.
[31] M. A. Gates, "A simple way to look at DNA," *Journal of Theoretical Biology*, vol. 119, no. 3, pp. 319–328, 1986.
[32] E. Hamori and J. Ruskin, "H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences," *Journal of Biological Chemistry*, vol. 258, no. 2, pp. 1318–1327, 1983.
[33] 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.
[34] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, "Compositional segmentation and long-range fractal correlations in DNA sequences," *Physical Review E*, vol. 55, no. 5, pp. 5181–5189, 1996.
[35] C. L. Berthelsen, J. A. Glazier, and M. H. Skolnick, "Global fractal dimension of human DNA sequences treated as pseudorandom walks," *Physical Review A*, vol. 45, no. 12, pp. 8902–8913, 1992.
[36] P. R. Aldrich, R. K. Horsley, and S. M. Turcic, "Symmetry in the language of gene expression: a survey of gene promoter networks in multiple bacterial species and non-σ regulons," *Symmetry*, vol. 3, pp. 1–20, 2011.
[37] R. Ferrer-I-Cancho and N. Forns, "The self-organization of genomes," *Complexity*, vol. 15, no. 5, pp. 34–36, 2010.
[38] T. Misteli, "Self-organization in the genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 17, pp. 6885–6886, 2009.
[39] C. E. Shannon, "A mathematical theoryof communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.
[40] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 623–656, 1948.
[41] R. V. Solé, "Genome size, self-organization and DNA's dark matter," *Complexity*, vol. 16, no. 1, pp. 20–23, 2010.
[42] R. M. Yulmetyev, N. A. Emelyanova, and F. M. Gafarov, "Dynamical Shannon entropy and information Tsallis entropy in complex systems," *Physica A*, vol. 341, no. 1–4, pp. 649–676, 2004.
[43] 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.
[44] 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.
[45] W. Li, "The study of correlation structures of DNA sequences: a critical review," *Computers and Chemistry*, vol. 21, no. 4, pp. 257–271, 1997.
[46] C. Cattani, "Haar wavelets based technique in evolution problems," *Proceedings of the Estonian Academy of Sciences: Physics & Mathematics*, vol. 53, no. 1, pp. 45–63, 2004.