ELSEVIER

Physica A 304 (2002) 480–494

PHYSICA A

www.elsevier.com/locate/physa

Multifractal properties of Hao's geometric representations of DNA sequences

Peter Tiňo *

Neural Computing Research Group, Aston University, Aston Triangle, Birmingham B4 7ET, UK

Received 27 August 2001

Abstract

Hao proposed a graphic representation of subsequence structure in DNA sequences and computed fractal dimensions of such representations for factorizable languages. In this study, we extend Hao's work in several directions: (1) We generalize Hao's scheme to accommodate sequences over an arbitrary finite number of symbols. (2) We establish a direct correspondence between the statistical characterization of symbolic sequences via Rényi entropy spectra and the multifractal characteristics (Rényi generalized dimensions) of the sequences' spatial representations. (3) We show that for general symbolic dynamical systems, the multifractal f_H-spectra in the sequence space endowed with commonly used metrics, coincide with the f_H-spectra on Hao's sequence representations. (4) So far the connection between the Hao's scheme and another well-known subsequence visualization scheme—Jeffrey's chaos game representation (CGR)—has been characterized only in very vague terms. We show that the fractal dimension results for Hao's visualization frames directly translate to Jeffrey's CGR scheme. © 2002 Elsevier Science B.V. All rights reserved.

PACS: 05.90.+m

Keywords: Genome; Symbolic dynamics; Multifractal

1. Introduction

Since the complete genome of a living organism was sequenced in 1995, there has been an exponentially growing amount of nucleotide sequence data available in public databanks. Statistical techniques for analyzing the internal structure of long DNA

* Tel.: +44-121-359-3611x4285; fax: +44-121-333-6215.
E-mail address: tinop@aston.ac.uk (P. Tiňo).

PII: S0378-4371(01)00574-X

sequences are becoming ever more important. Yet, such techniques are not well-suited for generating a good visual representation of DNA sequences [1].

Among the first to strike a cord in the direction of visualizing subsequence structure in long DNA sequences was Jeffrey [2]. He proposed to visualize the subsequence structures by driving a simple iterative function system (IFS) [3] with DNA sequences. The process resembles the chaos game algorithm for approximating the IFS attractor [3], but instead of choosing an IFS map (symbol) at each time step at random, the IFS maps are chosen according to the sequence of symbols in the DNA sequence we are interested in. In particular, the DNA sequence is visualized through points in a unit square, where the four corners of the square correspond to the four DNA bases. The first point, representing the first base in the DNA sequence, is plotted half-way between the center of the square and the corner representing that base. The second point is plotted half-way between the previous point and the corner representing the second base, etc. The result, the *chaos game representation* (CGR) of the DNA sequence, is an image where sparse areas correspond to rare subsequences and dense regions represent frequent subsequences. Basu et al. [4], Solovyev et al. [5], Fiser et al. [6] and others (see also references in Ref. [6]) generalized CGR to accommodate larger alphabets while remaining in low-dimensional visualization spaces.

Berthelsen et al. [7] converted DNA sequences into data-driven pseudo-random walks in two- or four-dimensional spaces. Movements in dimensions two and four are driven by the base and dimer sequences, respectively. The authors estimated the fractal dimension of the pseudorandom walks for various DNA sequences and compared them with the fractal dimensions of pseudorandom walks of artificial sequences whose base and dimer statistics matched those of the DNA sequences. Compared with the artificial sequences, the estimated fractal dimensions of the DNA-driven pseudorandom walks were significantly lower indicating an information content in DNA sequences not explained by the base or dimer frequencies. For pointers to other work in this direction see Refs. [7,8].

In a series of recent papers [1,9–11] Hao and collaborators introduced and rigorously studied a scheme for visualizing subsequence frequencies in DNA sequences. The subsequence counters are arranged in a frame in the way that makes the computation of frequencies and detection of dominant/missing subsequences very efficient [1]. In particular, it may be biologically appealing to pick up and characterize the set of missing, or under-represented short sequences in a complete genome [11]. One such possible characterization would be to derive a factorizable language defined by a set of forbidden words. Fractal dimensions of languages defined by tagged strings (that emerged from applying Hao's visualization method to complete genomes) were calculated in Refs. [9–11].

There are two interesting issues related to the Hao's visualization scheme that deserve a deeper investigation. First, even though Hao's scheme involves color coding of the counter sites according to the (relative) subsequence frequencies, the fractal dimensions reflect only the topological properties of the studied sequences. Probabilistic subsequence structure in DNA sequences induces a probability measure on the frame. Such measure in general lives on a fractal support and can be very complicated. It is natural to attempt, as we do in this paper, to characterize the measures on the visualization frames using tools of multifractal analysis. Second, so far the connection between the Hao's scheme and Jeffrey's CGR has been characterized only in very vague terms (see e.g. Refs. [1,9]). We show, that the connection is quite profound: fractal dimension results for Hao's visualization frames directly translate to Jeffrey's CGR scheme.

The paper has the following organization. In the next section, we introduce and generalize Hao's visualization frames and Jeffrey's CGR. Section 3 brings a brief introduction to statistical quantities on symbolic sequences and scaling characteristics of multifractal measures. Formal properties of geometric representations of symbolic sequences and symbolic dynamical systems are studied in Sections 4 and 5, respectively. Finally, the discussion briefly sums up key results of the paper in the context of previous work.

2. Geometric representations of symbolic sequences

Consider a finite alphabet $\mathscr{A} = \{1,2,\dots,A\}$. The sets of all finite[1] and infinite sequences over $\mathscr{A}$ are denoted by $\mathscr{A}^+$ and $\mathscr{A}^\omega$, respectively. The set of all sequences consisting of a finite, or an infinite number of symbols from $\mathscr{A}$ is then $\mathscr{A}^\infty = \mathscr{A}^+ \cup \mathscr{A}^\omega$. The set of all sequences over $\mathscr{A}$ with exactly n symbols (n-blocks) is denoted by $\mathscr{A}^n$.

For each sequence $S = s_1 s_2 \dots s_n \in \mathscr{A}^+$, S^R denotes the reversed sequence $S^R = s_n s_{n-1} \dots s_1$. Definition of the reverse operator can be extended to sets of sequences: for any $Q \subseteq \mathscr{A}^+$, $Q^R = \{S^R \mid S \in Q\}$.

Let $S = s_1 s_2 \dots \in \mathscr{A}^\infty$ and $i \leqslant j$. By S_i^j we denote the string $s_i s_{i+1} \dots s_j$, with $S_i^i = s_i$.

DNA sequences are streams over a four-symbol alphabet $\{a,t,g,c\}$. For our purposes, we identify the letters a, t, g and c with symbols 1, 2, 3 and 4, respectively.

2.1. Hao's frame representations of DNA sequences

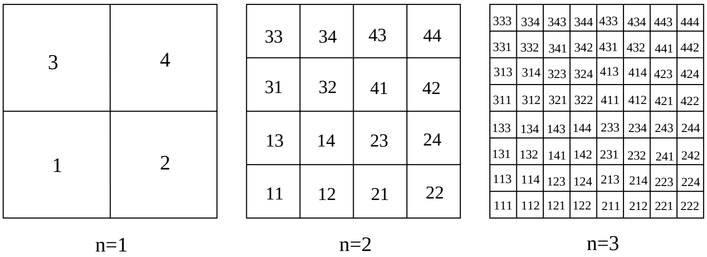
To visualize the n-block frequencies in DNA sequences, Hao et al. [1] arrange 4^n counters, associated with 4^n different n-blocks over $\mathscr{A} = \{1,2,3,4\}$, into a $2^n \times 2^n$ array of counters, as shown in Fig. 1. For each block length $n \geqslant 1$, the array may be expressed a $2^n \times 2^n$ matrix obtained as a direct product of n copies of the base matrix

$$M = \begin{bmatrix} 3 & 4 \\ 1 & 2 \end{bmatrix} ,$$

$M^{(n)} = M \otimes M \otimes \cdots \otimes M$.

Given a DNA sequence $S = s_1 s_2 \dots s_N$, one slides a window of length n along the sequence and subsequently color codes each counter according to the relative frequency with which the corresponding n-block occurred in S. The $2^n \times 2^n$ array of counters is called an n-frame. The size of the frame is independent of n. In this study, without loss of generality, the frames are arranged on the unit square $X = [0,1]^2$.

[1] Excluding the empty word.



n=1

3	4
1	2

n=2

33	34	43	44
31	32	41	42
13	14	23	24
11	12	21	22

n=3

333	334	343	344	433	434	443	444
331	332	341	342	431	432	441	442
313	314	323	324	413	414	423	424
311	312	321	322	411	412	421	422
133	134	143	144	233	234	243	244
131	132	141	142	231	232	241	242
113	114	123	124	213	214	223	224
111	112	121	122	211	212	221	222

Fig. 1. The arrangement of string counters in Hao's frames for block lengths $n = 1, 2, 3$.

2.2. *Jeffrey's chaos game representation of DNA sequences*

Iterated function system (IFS) [3] used by Jeffrey [2] to construct chaos game representations (CGRs) of DNA sequences is a collection of four maps $i = 1, 2, 3, 4$,

$$i(x) = \tfrac{1}{2}(x + t_i)\,, \tag{1}$$

$$t_1 = (0, 0), \quad t_2 = (0, 1), \quad t_3 = (1, 0), \quad t_4 = (1, 1) \tag{2}$$

operating on the unit square $X = [0, 1]^2$.

The chaos game representation $CGR(S)$ of a sequence $S = s_1 s_2 \ldots s_N$, $s_j \in \{1, 2, 3, 4\}$, is obtained as follows:

1. Start in the middle of the unit square, $x_0 = \{\frac{1}{2}\}^2$.
2. For $1 \leqslant m \leqslant N$, plot the point $x_m = i(x_{m-1})$, provided the mth symbol s_m is i.
3. $CGR(S) = \{x_m\}_{m=1}^{N}$.

2.3. *Geometric block representations of symbolic sequences*

In Ref. [12] we generalized the concept of chaos game representation of DNA sequences to allow for an arbitrary finite number of symbols. In our recent study [13] we showed how such representations may be utilized to efficiently construct predictors analogous to variable memory length Markov models [14]. Since our main concern was not a visualization of the subsequence structure in symbolic sequences, our spatial representations might live in a higher-dimensional space.

Here, we present further generalization of our spatial representation scheme that allows one to visualize in a low-dimensional space (if one wishes to do so), the block structure in symbolic sequences over alphabets containing more than 4 symbols. Using an analogy with Hao's scheme (see Fig. 1), this is achieved by allowing more than 2 counters along each side of the base frame corresponding to the block length $n = 1$. Hence, instead of a 2×2 base matrix M we will have an $L \times L$ base matrix, with $L \geqslant 2$.

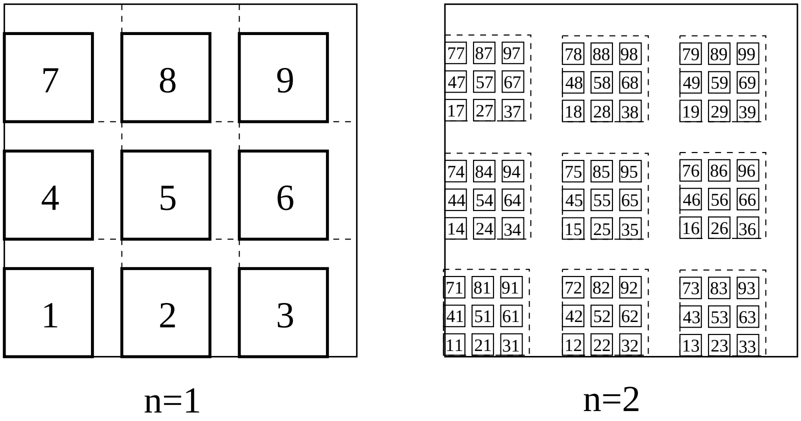


Fig. 2. Representative regions for 1- and 2-blocks under the IFS (3). Here, $\mathscr{A} = \{1, 2, \ldots, 9\}$, $L = 3$ and $\varepsilon = \frac{3}{4}$.

The basis of our scheme is an IFS acting on the d-dimensional unit hypercube $X = [0, 1]^d$, where[2] $d = \lceil \log_L A \rceil$. To keep the notation simple, we slightly abuse mathematical notation and, depending on the context, regard the symbols $1, 2, \ldots, A$, as integers, or as maps on X. The maps $i = 1, 2, \ldots, A$, constituting the IFS are affine contractions

$$i(x) = \frac{\varepsilon}{L} x + t_i, \quad t_i \in \left\{0, \frac{1}{L}, \frac{2}{L}, \ldots, \frac{L-1}{L}\right\}^d \tag{3}$$

with a scale parameter $\varepsilon \in (0, 1]$. The shift vectors t_i are unique, i.e., $t_i \neq t_j$, for $i \neq j$.

The attractor of the IFS (3) is the unique set[3] $K \subseteq X$, for which $K = \bigcup_{i=1}^{A} i(K)$ [3].

For a string $u = u_1 u_2 \ldots u_n \in \mathscr{A}^n$ and a point $x \in X$, the point

$$\begin{aligned} u(x) &= u_n(u_{n-1}(\ldots(u_2(u_1(x)))\ldots)) \\ &= (u_n \circ u_{n-1} \circ \cdots \circ u_2 \circ u_1)(x) \end{aligned} \tag{4}$$

is considered a *geometric representation of the string u under the IFS* (3). For a set $Y \subseteq X$, $u(Y)$ is then $\{u(x) \mid x \in Y\}$. As an illustration, we show in Fig. 2 the sets $u(X)$ for all 1- and 2-blocks u over the alphabet $\mathscr{A} = \{1, 2, \ldots, 9\}$, for the case of $L = 3$ and $\varepsilon = \frac{3}{4}$.

From now on, the center of the hypercube X, $\{\frac{1}{2}\}^d$, will be denoted by x_*. Given a sequence $S = s_1 s_2 \ldots \in \mathscr{A}^\infty$, its (*generalized*) *chaos game representation* under the

[2] For $x \in \mathfrak{R}$, $\lceil x \rceil$ is the smallest integer y, such that $y \geqslant x$.

[3] For $\varepsilon \in (0, 1)$ known as the Sierpinski sponge [15].

IFS (3) is formally defined as the sequence of points

$$CGR_{L,\varepsilon}(S) = \{S_1^m(x_*)\}_{m \geqslant 1} . \tag{5}$$

Note that when $\mathscr{A} = \{1, 2, 3, 4\}$, the sequence $CGR_{2,1}(S)$ coincides with the original Jeffrey's chaos game representation of S (see Section 2.2).

Using the IFS (3), we represent the n-block structure in symbolic sequences through *geometric n-block representations*. The geometric n-block representation of a sequence $S = s_1 s_2 \ldots \in \mathscr{A}^\infty$ is the sequence of points

$$GBR_{L,\varepsilon}^n(S) = \{S_m^{m+n-1}(x_*)\}_{m \geqslant 1} . \tag{6}$$

The reversed n-block structure in the sequence S is geometrically interpreted via the *reversed geometric n-block representation*

$$RGBR_{L,\varepsilon}^n(S) = \{(S_m^{m+n-1})^R(x_*)\}_{m \geqslant 1} . \tag{7}$$

The sequences $GBR_{L,\varepsilon}^n(S)$ and $RGBR_{L,\varepsilon}^n(S)$ contain a representative point for each n-block in S.

An important observation is that when $\mathscr{A} = \{1, 2, 3, 4\}$, the reversed geometric n-block representation $RGBR_{2,1}^n(S)$ mimics the coding scheme of Hao's frames: the representative point of each n-block lies in the center of the corresponding counter shown in Fig. 1. In this sense, the reversed geometric n-block representations can be considered generalizations of Hao's n-frames.

3. Statistics on sequential and spatial structures

In this section, we briefly recall some of the basic tools for describing "non-trivial" spatial measures and subsequence distributions in symbolic sequences.

3.1. Statistics on symbolic sequences

Let $S = s_1 s_2 \ldots \in \mathscr{A}^\infty$ be a sequence generated by a stationary information source [16]. Denote the (empirical) probability of finding an n-block $w \in \mathscr{A}^n$ in S by $P_n(w)$. A string $w \in \mathscr{A}^n$ is said to be an allowed n-block in the sequence S, if $P_n(w) > 0$. The set of all allowed n-blocks in S is denoted by $[S]_n$.

A measure of n-block uncertainty in S is given by the block entropy[4]

$$H_n(S) = -\sum_{w \in [S]_n} P_n(w) \log P_n(w) .$$

The limit of the average uncertainty per symbol $h_n(S) = H_n(S)/n$ is the entropy rate $h(S) = \lim_{n \to \infty} h_n(S)$. The entropy rate quantifies the predictability of an added symbol (independent of block length).

[4] If information is measured in bits, then $\log \equiv \log_2$.

The block entropies H_n and entropy rates h_n are special cases of Rényi entropies and entropy rates [17]. The β-order Rényi entropy of the n-block distribution ($\beta \in \Re$)

$$H_{\beta,n}(S) = \frac{1}{1-\beta} \log \sum_{w \in [S]_n} P_n^{\beta}(w)$$

and the β-order Rényi entropy rate

$$h_{\beta,n}(S) = \frac{H_{\beta,n}(S)}{n} \tag{8}$$

reduce to the block entropy $H_n(S)$ and entropy rate $h_n(S)$ when $\beta = 1$ [18]. The formal parameter β can be thought of, for example, as the inverse temperature in the statistical mechanics of spin systems [19]. In the infinite temperature regime, $\beta = 0$, the Rényi entropy rate $h_{0,n}(S)$ is just a logarithm of the number of allowed n-blocks, divided by n. The limit $h_{(0)}(S) = \lim_{n\to\infty} h_{0,n}(S)$ gives the asymptotic exponential growth rate of the number of allowed n-blocks, as the block length increases.

The entropy rates $h(S) = h_{(1)}(S) = \lim_{n\to\infty} h_{1,n}(S)$ and $h_{(0)}(S)$ are also known as the metric and topological entropies, respectively.

Varying the parameter β amounts to scanning the original n-block distribution P_n: the most probable and the least probable n-blocks become dominant in the positive zero ($\beta = \infty$) and the negative zero ($\beta = -\infty$) temperature regimes, respectively. Varying β from 0 to ∞ amounts to a shift from all allowed n-blocks to the most probable ones by accentuating still more and more probable subsequences. Varying β from 0 to $-\infty$ accentuates less and less probable n-blocks with the extreme of the least probable ones. For a non-extensive formalism involving Tsallis entropy spectra, see e.g. Ref. [20].

3.2. Scaling behavior on multifractals

Loosely speaking, a multifractal is a fractal set supporting a probability measure [21]. The degree of fragmentation of the fractal support K is usually quantified through its fractal dimension $D(K)$ [3]. Denote by $N(\ell)$ the minimal number of hyperboxes of side length ℓ needed to cover K. The fractal (box-counting) dimension $D(K)$ relates the side length ℓ with $N(\ell)$ via the scaling law $N(\ell) \approx \ell^{-D(K)}$.

For $0 < c \leqslant \frac{1}{2}$, the nth order approximation $D_{n,c}(K)$ of the fractal dimension $D(K)$ is given by the box-counting technique with boxes of side $\ell = c^n$:

$$N(c^n) = (c^n)^{-D_{n,c}(K)} .$$

Just as the Rényi entropy spectra describe (non-homogeneous) statistics on symbolic sequences, generalized Rényi dimensions D_β capture multifractal probabilistic measures μ [22]. Generalized dimensions $D_\beta(K)$ of an object K describe a measure μ on K through the scaling law

$$\sum_{B \in \mathscr{B}_\ell,\ \mu(B)>0} \mu^{\beta}(B) \approx \ell^{(\beta-1)D_\beta(K)} , \tag{9}$$

where $\mathscr{B}_\ell$ is a minimal set of hyperboxes with sides of length ℓ disjointly[5] covering K.

[5] At most up to Lebesgue measure zero borders.

In particular, for side lengths $\ell = c^n$, with $0 < c \leqslant \frac{1}{2}$, the nth order approximation $D_{\beta,n,c}(K)$ of $D_\beta(K)$ is given by

$$\sum_{B \in \mathscr{B}_{c^n},\ \mu(B)>0} \mu^\beta(B) = \ell^{(\beta-1)D_{\beta,n,c}(K)} . \tag{10}$$

The infinite temperature scaling exponent $D_0(K)$ is equal to the box-counting fractal dimension $D(K)$ of K. Dimensions D_1 and D_2 are, respectively, known as the information and correlation dimensions [21]. Of special importance are the limit dimensions D_∞ and $D_{-\infty}$ describing the scaling behavior in regions where the probability is most concentrated and rarefied, respectively.

4. Representing *n*-block structure in symbolic sequences

Consider a sequence $S \in \mathscr{A}^\infty$. Let the measures μ_n and v_n on $X = [0,1]^d$ describe the relative frequencies of points from the sequences $GBR^n_{L,\varepsilon}(S)$ and $RGBR^n_{L,\varepsilon}(S)$ (see Section 2.3), respectively, on the Lebesgue subsets of X. In the following, we establish the relationship between the Rényi entropy spectra of the sequence S and the generalized dimension spectra of its geometric block representations.

Theorem 1. *For any sequence $S \in \mathscr{A}^\infty$, and any $n = 1,2,\ldots,$ the nth order approximations of the generalized dimensions of its n-block geometric representations are equal, up to a scaling constant* $\log(L/\varepsilon)$, *to the sequence n-block Rényi entropy rate estimates*:

$$D_{\beta,n,\varepsilon/L}(GBR^n_{L,\varepsilon}(S)) = D_{\beta,n,\varepsilon/L}(RGBR^n_{L,\varepsilon}(S)) = \frac{h_{\beta,n}(S)}{\log L/\varepsilon} .$$

In particular, for any $S \in \mathscr{A}^\omega$,

$$\lim_{n\to\infty} D_{\beta,n,\varepsilon/L}(GBR^n_{L,\varepsilon}(S)) = \lim_{n\to\infty} D_{\beta,n,\varepsilon/L}(RGBR^n_{L,\varepsilon}(S)) = \frac{h_{(\beta)}(S)}{\log L/\varepsilon} ,$$

provided the limits exist.

Proof. Denote the contraction factor of IFS (3) by c, i.e., $c = \varepsilon/L$.

There is a one-to-one correspondence between the allowed n-blocks $w \in [S]_n$ and the boxes $w(X)$ of side length $\ell = c^n$.

From (10)

$$\sum_{w \in [S]_n} \mu_n^\beta(w(X)) = c^{\,n(\beta-1)D_{\beta,n,c}(GBR^n_{L,\varepsilon}(S))}$$

and so

$$\begin{aligned} D_{\beta,n,c}(GBR^n_{L,\varepsilon}(S)) &= \frac{1}{n(1-\beta)\log c^{-1}} \log \sum_{w \in [S]_n} \mu_n^\beta(w(X)) \\ &= \frac{1}{n(1-\beta)\log c^{-1}} \log \sum_{w \in [S]_n} P_n^\beta(w) = \frac{h_{\beta,n}(S)}{\log c^{-1}} . \end{aligned}$$

The entropies introduced in Section 3.1 do not contain any notion of causality

$$\sum_{w\in[S]_n} \mu_n^\beta(w(X)) = \sum_{w\in[S]_n} P_n^\beta(w) = \sum_{w\in[S]_n^R} P_n^\beta(w^R) = \sum_{w\in[S]_n^R} v_n^\beta(w^R(X)) ,$$

where $[S]_n^R$ is the set of all allowed reversed n-blocks in the sequence S. Hence, the generalized dimensions for the two n-block representations agree, i.e.,

$$D_{\beta,n,c}\,(GBR_{L,\varepsilon}^n(S)) = D_{\beta,n,c}\,(RGBR_{L,\varepsilon}^n(S)) . \qquad \square$$

We note that, with the exception of sequences $S = ws^\omega$, $w \in \mathscr{A}^+$, $s \in \mathscr{A}$, the limit n-block representations, $\lim_{n\to\infty} GBR_{L,\varepsilon}^n(S)$, do not exist. However, the dimension estimates $D_{\beta,n,\varepsilon/L}(GBR_{L,\varepsilon}^n(S))$ may still converge. On the other hand, the reversed limit block representations, $\lim_{n\to\infty} GBR_{L,\varepsilon}^n(S)$, do exist for all $S \in \mathscr{A}^\omega$.

Next, we show, that the chaos game representations $CGR_{L,\varepsilon}(S)$ (Section 2.3) share the nth order generalized dimension estimates with the block representations $GBR_{L,\varepsilon}^n(S)$ and $RGBR_{L,\varepsilon}^n(S)$.

Denote by $CGR_{L,\varepsilon}^n(S)$ the sequence $CGR_{L,\varepsilon}(S)$ without the first $n-1$ points.

Theorem 2. *For any sequence $S \in \mathscr{A}^\infty$, and any $n = 1, 2, \ldots,$ the nth order approximations of the generalized dimensions of its chaos game representation are related to the sequence n-block Rényi entropy rate estimates through*

$$D_{\beta,n,\varepsilon/L}(CGR_{L,\varepsilon}^n(S)) = \frac{h_{\beta,n}(S)}{\log L/\varepsilon} .$$

Furthermore, for each $S \in \mathscr{A}^\omega$;

$$D_{\beta,n,\varepsilon/L}(CGR_{L,\varepsilon}(S)) = \frac{h_{\beta,n}(S)}{\log L/\varepsilon} .$$

Proof. Under any IFS consisting of contractive mappings (such as the IFS (3)), if $v \in \mathscr{A}^+$ is a suffix of a string $u = rv$ $(r, u \in \mathscr{A}^+)$, then $u(X) \subset v(X)$. This follows from the fact that compositions of contractions are themselves contractions: $r(X) \subset X$, and so $u(X) = v(r(X)) \subset v(X)$.

Since the n-block S_i^{i+n-1} is a suffix of the initial $(i+n-1)$-block S_1^{i+n-1}, it follows that $S_1^{i+n-1}(X) \subseteq S_i^{i+n-1}(X)$. Hence, we can directly apply arguments in the proof of Theorem 1 and conclude that for the IFS (3)

$$D_{\beta,n,\varepsilon/L}(CGR_{L,\varepsilon}^n(S)) = D_{\beta,n,\varepsilon/L}(GBR_{L,\varepsilon}^n(S)) = \frac{h_{\beta,n}(S)}{\log L/\varepsilon} .$$

To justify the second statement, we write the sequence $S \in \mathscr{A}^\omega$ as $S = wS'$, where $w \in \mathscr{A}^{n-1}$ and $S' \in \mathscr{A}^\omega$. Since n is finite, the empirical n-block probabilities $P_n(v)$, $v \in \mathscr{A}^n$, are the same for both sequences S and S'. $\square$

Theorems 1 and 2 imply, that provided $\log \equiv \log_2$, when $L/\varepsilon = 2$, for infinite sequences $S \in \mathscr{A}^\omega$, the generalized dimension estimates of geometric representations exactly equal the corresponding sequence Rényi entropy rate estimates. In particular, given an infinite sequence $S \in \mathscr{A}^\omega$, when $L = 2$ and $\varepsilon = 1$, as n grows, the box-counting fractal dimension and the information dimension estimates $D_{0,n,1/2}$ and $D_{1,n,1/2}$ of both the original Jeffrey's chaos game representation [2,23,24] and Hao's frame representation tend to the sequence topological and metric entropies, respectively. This follows from the fact that when $L = 2$ and $\varepsilon = 1$, Hao's n-frame representation is equivalent to the reversed n-block geometric representation $RGBR^n_{L,\varepsilon}$ (see Section 2).

The ideas behind Jeffrey's chaos sequence representation of symbolic sequences were independently studied in the image compression community. Quadtree [3] is an addressing scheme used in computer science for addressing small squares in the unit square X (representing the computer display). The square is broken into four quadrants $i = 0, 1, 2, 3$. Points in the quadrant i have addresses beginning with i. Each quadrant i is split into four subquadrants ij, $j = 0, 1, 2, 3$. Points in the subquadrant ij have addresses beginning with ij, etc. Hence, the quadtree scheme is equivalent to both Hao's frames and our reversed n-block representations of symbolic sequences over a four-letter alphabet, with $L = 2$ and $\varepsilon = 1$. Staiger [25] showed that the Hausdorff dimension of pictures addressed by sequences obeying a given regular expression is just a logarithm of the maximum modulus of the connection matrix of the underlying finite automaton (which is in fact the topological entropy of the set of sequences specified by the underlying automaton).

For a fixed block length n, if we identified the n-blocks with symbols in a larger alphabet $\mathscr{A}' = \mathscr{A}^n$ and assigned to symbols from $\mathscr{A}'$ frequencies of the corresponding n-blocks, Theorems 1 and 2 would be related to the results of Mandelbrot [26] and Gutiérrez and Rodriguez [27] concerning the $f(\alpha)$ multifractal spectra (see next section) of single-scaled multinomial measures. [6]

5. Geometric block representations of symbolic dynamical systems

While in the previous section we analyzed geometric and measure scaling properties of various geometric representations of a given (possibly infinite) symbolic sequence, in this section we investigate properties of the reversed block representations (and hence generalized Hao's frames) in the context of general symbolic dynamical systems.

Denote the longest common prefix of two sequences S and S' by $\#(S, S')$. The set of infinite sequences $S = s_1 s_2 \ldots \in \mathscr{A}^\omega$ over the alphabet $\mathscr{A} = \{1, 2, \ldots, A\}$, endowed with a metric

$$d_\gamma(S, S') = \gamma^{\#(S,S')}, \quad 0 < \gamma < 1 \tag{11}$$

forms a metric space $(\mathscr{A}^\omega, d_\gamma)$.

[6] The n-block Rényi entropy rates would be substituted by entropy rates of the n-block escort distributions (e.g. Ref. [28]).

Each sequence $S \in \mathscr{A}^{\omega}$ is coded by a point (the “address” of S)

$$\chi(S) = \lim_{n \to \infty} (S_1^n)^R (x_*) \tag{12}$$

on the attractor K of the iterative function system (3). In this section, we need to make sure that the IFS (3) satisfies the open set condition (e.g. Refs. [27,29]). In other words, the IFS (3) needs to be non-overlapping [3]. This can be achieved by setting ε strictly less than one. In that case, the map χ is one-to one.

Let $\sigma : \mathscr{A}^{\omega} \to \mathscr{A}^{\omega}$ be a shift map given by $\sigma(s_1 s_2 s_3 \ldots) = s_2 s_3 \ldots$. Consider a shift dynamical system on a compact[7] and shift-invariant subset $Q \subseteq \mathscr{A}^{\omega}$. Let τ be a measure supported on Q and preserved by σ.

Even in the very simple case of the full shift $Q = \mathscr{A}^{\omega}$ with a measure τ on $(\mathscr{A}^{\omega}, d_{\gamma})$ defined by a Bernoulli source with unequal symbol probabilities, the distribution of sequences S in $(\mathscr{A}^{\omega}, d_{\gamma})$, as well as the distribution of points $\chi(S)$ in the Euclidean space[8] (K, d_E), are singular [30]. Hence, one cannot describe the distributions by means of densities. Multifractal analysis proves useful in characterizing the complicated geometrical properties of the measure τ on $(\mathscr{A}^{\omega}, d_{\gamma})$ and its pushed forward (via the map χ) counterpart ν on (K, d_E).

The basic idea is to classify the singularities of the measure τ by “strength”. The strength, also known as the Hölder exponent, is measured as a singularity exponent

$$\alpha(S) = \lim_{B \to \{S\}} \frac{\log \tau(B)}{\log \mathrm{Diam}(B)},$$

where $B \to \{S\}$ means that $B \subset \mathscr{A}^{\omega}$ is a ball containing the sequence S and that its diameter

$$\mathrm{Diam}(B) = \sup_{U,V \in B} \{d_{\gamma}(U,V)\}$$

tends to zero.

Usually, points of equal strength lie on interwoven fractal sets

$$K_{\alpha} = \{S \in \mathscr{A}^{\omega} \mid \alpha(S) = \alpha\}.$$

The geometry of the singular distribution τ can then be characterized by giving the “size” of the sets K_{α}, more precisely their Hausdorff dimension $\dim_H(K_{\alpha})$ [31],

$$f_H(\alpha) = \dim_H(K_{\alpha}). \tag{13}$$

In the limit of infinite block lengths, the reversed block representation of the shift-invariant set Q becomes[9]

$$\lim_{n \to \infty} RGBR^n_{L,\varepsilon}(Q) = \{\chi(\sigma^i(S)) \mid S \in Q,\ i = 0,1,2,\ldots\}$$

[7] With respect to the topology induced by d_{γ}.

[8] d_E denotes the Euclidean metric.

[9] $\sigma^0(S) = S$, $\sigma^n(S) = \sigma(\sigma^{n-1}(S))$, $n = 1,2,\ldots$

and so it is natural to connect the multifractal analysis of the invariant measure τ of the shift dynamical system (Q, σ) on the metric space $(\mathscr{A}^{\omega}, d_{\gamma})$ with the pushed forward (via the map χ) measure v on the metric space (K, d_E).

Our strategy is as follows: First, we shall show that the two metric spaces $(\mathscr{A}^{\omega}, d_{\gamma})$ and $(\chi(\mathscr{A}^{\omega}), d_E)$, that correspond to each other through the coding map χ, are metrically equivalent. Under metric equivalence, Hölder exponents $\alpha(S)$ of the measure τ coincide with the Hölder exponents $\alpha(\chi(S))$ of the pushed forward measure v. Since the Hausdorff dimension is a metric equivalence invariant [31], we get that the multifractal spectra $f_H(\alpha)$ for measures τ and v are the same. In other words, in the limit of infinite block lengths, the multifractal f_H-spectrum in the sequence space $(\mathscr{A}^{\omega}, d_{\gamma})$ is the same as the f_H-spectrum in the space $(\chi(\mathscr{A}^{\omega}), d_E)$ of Hao's frames.

Two metrics ρ_1 and ρ_2 on $\mathscr{A}^{\omega}$ are equivalent if there exist constants $0 < c_1 < c_2 < \infty$, such that for all $S, S' \in \mathscr{A}^{\omega}$;

$$c_1 \rho_1(S, S') \leqslant \rho_2(S, S') \leqslant c_2 \rho_1(S, S') .$$

The two metric spaces $(\mathscr{A}^{\omega}, d_{\gamma})$ and $(\chi(\mathscr{A}^{\omega}), d_E)$ are equivalent, if the bijective map $\chi : \mathscr{A}^{\omega} \to \chi(\mathscr{A}^{\omega})$ induces a metric $\tilde{d}_{\gamma}$ in $\mathscr{A}^{\omega}$,

$$\tilde{d}_{\gamma}(S, S') = d_E(\chi(S), \chi(S'))$$

that is equivalent to the metric d_{γ}.

Theorem 3. *The metric spaces $(\mathscr{A}^{\omega}, d_{\gamma})$ and $(\chi(\mathscr{A}^{\omega}), d_E)$ are equivalent if and only if $\gamma = \varepsilon/L$.*

Proof. We will show that

$$\forall S, S' \in \mathscr{A}^{\omega},\ \exists 0 < c_1 < c_2 < \infty \quad \text{such that}$$
$$c_1 d_E(\chi(S), \chi(S')) \leqslant d_{\gamma}(S, S') \leqslant c_2 d_E(\chi(S), \chi(S')) . \tag{14}$$

The minimal distance $d_E(S, S')$ can be bounded from below by

$$d_E(\chi(S), \chi(S')) \geqslant \left(\frac{\varepsilon}{L}\right)^{\#(S,S')} \frac{1 - \varepsilon}{L} . \tag{15}$$

Since $d_{\gamma}(S, S') = \gamma^{\#(S,S')}$, we get from (14) and (15)

$$\gamma^{\#(S,S')} \leqslant c_2 \left(\frac{\varepsilon}{L}\right)^{\#(S,S')} \frac{1 - \varepsilon}{L}$$

which means

$$c_2 \geqslant \left(\frac{L\gamma}{\varepsilon}\right)^{\#(S,S')} \frac{L}{1 - \varepsilon} . \tag{16}$$

On the other hand, the maximal distance between the codes $\chi(S), \chi(S')$ of any two infinite sequences S, S' can be bounded from above by

$$d_E(\chi(S), \chi(S')) \leqslant \left(\frac{\varepsilon}{L}\right)^{\#(S,S')} \sqrt{d} \tag{17}$$

and so, from (14) and (17)

$$c_1 \left(\frac{\varepsilon}{L}\right)^{\#(S,S')} \sqrt{d} \leqslant \gamma^{\#(S,S')},$$

which implies

$$c_1 \leqslant \left(\frac{L\,\gamma}{\varepsilon}\right)^{\#(S,S')} \frac{1}{\sqrt{d}}. \tag{18}$$

Inequalities (16) and (18) should hold for all pairs of sequences S, S', i.e., for all prefix lengths $\#(S,S') = 0, 1, 2, \ldots$. On the other hand, c_1, c_2 are bounded positive constants. This is possible only when $\gamma = \varepsilon/L$. □

It is an easy exercise to show that the metric d_γ (Eq. (11)) is equivalent with another commonly used metric in the code space $\mathscr{A}^\omega$, namely

$$\rho_\lambda(S, S') = \sum_{i=1}^{\infty} \frac{|s_i - s_i'|}{\lambda^i}, \quad \lambda > 1 \tag{19}$$

if and only if $\gamma = \lambda^{-1}$. Hence, in the limit of infinite block lengths, provided $\gamma = \varepsilon/L = \lambda^{-1}$, the multifractal f_H-spectrum in the sequence metric space $(\mathscr{A}^\omega, \rho_\lambda)$ coincides not only with the f_H-spectrum in $(\mathscr{A}^\omega, d_\gamma)$, but also with that in the space $(\chi(\mathscr{A}^\omega), d_E)$ of Hao's frames.

On the other hand, the metrics d_γ and ρ_λ on the code space cannot be made equivalent to the Baire's metric (e.g. Ref. [32])

$$\delta(S, S') = \frac{1}{1 + \#(S, S')} \tag{20}$$

and so there is no straightforward correspondence between the f_H-spectra in the space $(\chi(\mathscr{A}^\omega), d_E)$ of Hao's frames and the f_H-spectra in the Baire's space $(\mathscr{A}^\omega, \delta)$.

6. Discussion

We investigated how scaling properties of the Hao's visualization frames [1,9] correspond to the statistical properties of the sequences they represent. To that end, we formulated a framework that enabled us to generalize and relate Hao's visualization frames with Jeffrey's chaos game representation (CGR) of DNA sequences [2]. In our framework, we allowed for alphabets with an arbitrary finite number of symbols and contraction ratios of iterative function system (IFS) maps other than $\frac{1}{2}$.

We studied geometric representations of symbolic sequences, as well as general symbolic dynamical systems. We have shown that, for both Hao's frames and Jeffrey's CGR, the generalized dimension estimates of the geometric sequence representations directly correspond to the sequence Rényi entropy rates. In particular, by considering finer and finer scales, the box-counting fractal dimension and the information dimension estimates of the geometric sequence representations tend to the (scaled) sequence topological and metric entropies, respectively. It follows that the fractal dimension results for Hao's visualization frames [9–11] directly translate to Jeffrey's CGR scheme. This contrasts with the intuition expressed in Ref. [1] that:

> ... the (Hao's frame) portraits differ essentially from (Jeffrey's) CGR, in which the number of points plotted always equals to the number of nucleotides in the (DNA) sequence and no coarse-graining is made. Thus a very long sequence would fill out most of the plane and no fine details would be resolvable ...

We have shown that, for long sequences, the amount of "structural fragmentation" (measured by the fractal dimension) of Hao's frames is the same as that of Jeffrey's CGR. In fact, when using CGR without the first $n-1$ points, with a proper $2^n \times 2^n$ grid, one can detect missing, or under-represented n-blocks using techniques analogous to those used in Hao's visualization frames.

In the limit of infinite block lengths, the multifractal f_H-spectrum in the code metric spaces $(\mathscr{A}^\omega, d_\gamma)$ and $(\mathscr{A}^\omega, \rho_\lambda)$ coincides with the f_H-spectrum of the (generalized) Hao's frames, provided λ^{-1}, γ and the geometric representations' contraction coefficient ε/L are all equal.

Hao's subsequence visualization scheme is a special case of Moran-like constructions [29,33,34]. The setting of Moran-like geometric constructions is much more general than our setting of geometric sequence representations. Multifractal analysis of Moran-like constructions is primarily concerned with validity of the multifractal formalism [35] (i.e., the $\cap$-shape of the f_H-spectra for dimensions, differentiability of the spectra, Legendre relations between the multifractal quantities, etc., see Ref. [36]). The emphasis lies mostly on Bernoulli source driven constructions, although some analogous results were obtained for constructions driven by following the arcs of a digraph [37]. It is known that the Hausdorff and box-counting dimensions of the limit sets of Moran-like constructions coincide [34] and so the nth order dimension approximations calculated in this study tend to the geometric representations' both box-counting and Hausdorff dimensions.

Acknowledgements

Matter The author is thankful to Rudolf Riedi, Luis Barreira and Lars Olsen for helpful comments on the multifractal formalism. This work was supported by the VEGA grant 1/7611/2000.

References

[1] B.-L. Hao, H.C. Lee, S. Zhang, Chaos, Solitons Fractals 11 (2000) 825–836.
[2] J. Jeffrey, Nucl. Acids Res. 18 (1990) 2163–2170.
[3] M.F. Barnsley, Fractals Everywhere, Academic Press, New York, 1988.
[4] S. Basu, A. Pan, C. Dutta, J. Das, J. Mol. Graph Model 15 (1997) 279–289.
[5] V.V. Solovyev, S.V. Korolev, H.A. Lim, Int. J. Genomic Res. 1 (1993) 109–128.
[6] A. Fiser, G.E. Tusnady, I. Simon, J. Mol. Graphics 12 (1994) 302–304.
[7] C.L. Berthelsen, J.A. Glazier, M.H. Skolnik, Phys. Rev. A 45 (1992) 8902–8913.
[8] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, Z.D. Goldberger, S. Havlin, R.N. Mantegna, S.M. Ossadnik, C.-K. Peng, M. Simons, Physica A 205 (1994) 214–253.
[9] B.-L. Hao, Physica A 282 (2000) 225–246.
[10] B.-L. Hao, H. Xie, Z. Yu, G. Chen, Physica A 288 (2000) 10–20.
[11] Z. Yu, B.-L. Hao, H. Xie, G. Chen, Chaos, Solitons Fractals 11 (2000) 2215–2222.
[12] P. Tiň o, IEEE Trans. Systems, Man, Cybern. Part A: Systems Humans 29 (1999) 386–392.
[13] P. Tiň o, G. Dorffner, Mach. Learning 45 (2001) 187–218.
[14] P. Buhlmann, A.J. Wyner, Ann. Stat. 27 (1999) 480–513.
[15] R. Kenyon, Y. Peres, Ergodic Theory Dyn. Systems 16 (1996) 307–323.
[16] A.I. Khinchin, Mathematical Foundations of Information Theory, Dover Publications, New York, 1957.
[17] A. Renyi, Acta Math. Hung. 10 (1959) 193.
[18] P. Grassberger, Information and complexity measures in dynamical systems, in: H. Atmanspacher, H. Scheingraber (Eds.), Information Dynamics, Plenum Press, New York, 1991, pp. 15–33.
[19] J.P. Crutchfield, K. Young, Computation at the onset of chaos, in: W.H. Zurek (Ed.), Complexity, Entropy, and the Physics of Information, SFI Studies in the Sciences of Complexity, Vol 8, Addison-Wesley, Reading, MA, 1990, pp. 223–269.
[20] C. Tsallis, Fractals 3 (1995) 541–547.
[21] C. Beck, F. Schlogl, Thermodynamics of Chaotic Systems, Cambridge University Press, Cambridge, UK, 1995.
[22] J.L. McCauley, Chaos, Dynamics and Fractals: an Algorithmic Approach to Deterministic Chaos, Cambridge University Press, Cambridge, UK, 1994.
[23] J.L. Oliver, P. Bernaola-Galván, J. Guerrero-Garcia, R. Román Roldan, J. Theor. Biol. 160 (1993) 457–470.
[24] R. Roman-Roldan, P. Bernaola-Galvan, J.L. Oliver, Pattern Recognition Lett. 15 (1994) 567–573.
[25] L. Staiger, Quadtrees and the hausdorff dimension of pictures, in: Workshop on Geometrical Problems of Image Processing, A. Hübler (Ed.), Georgental, GDR, 1989, pp. 173–178.
[26] B.B. Mandelbrot, Pure Appl. Geophys. 131 (1989) 5–42.
[27] J.M. Gutierrez, M.A. Rodriguez, Chaos, Solitons Fractals 11 (2000) 675–683.
[28] R.S. Johal, R. Rai, Physica A 282 (2000) 525–535.
[29] P. Moran, Proc. Cambridge Philos. Soc. 42 (1946) 15–23.
[30] R.H. Riedi, Fractals 5 (1997) 153–168.
[31] K.J. Falconer, Fractal Geometry: Mathematical Foundations and Applications, Wiley, New York, 1990.
[32] H. Fernau, IIFS and codes, in: Developments In Theoretical Computer Science, Proceedings of the Seventh International Meeting of Young Computer Scientists, Gordon & Breach Science Publishers, Switzerland, 1994, pp. 141–152.
[33] Y. Pesin, Dimension Theory in Dynamical Systems: Rigorous Results and Applications, University of Chicago Press, Chicago, 1997.
[34] Y. Pesin, H. Weiss, Commun. Math. Phys. 182 (1996) 105–153.
[35] Y. Pesin, H. Weiss, J. Stat. Phys. 86 (1997) 233–275.
[36] L. Barreira, Y. Pesin, J. Schmeling, Chaos: an Interdisciplinary J. Nonlinear Sci. 7 (1996) 27–53.
[37] G.A. Edgar, R.D. Mauldin, Proc. London Math. Soc. 65 (1992) 604–628.