Spatial Representation of Symbolic Sequences through Iterative Function Systems

Peter Tiňo

Abstract— **Jeffrey proposed a graphic representation of DNA sequences using Barnsley's iterative function systems. In spite of further developments in this direction, the proposed graphic representation of DNA sequences has been lacking a rigorous connection between its spatial scaling characteristics and the statistical characteristics of the DNA sequences themselves. We 1) generalize Jeffrey's graphic representation to accommodate (possibly infinite) sequences over an arbitrary finite number of symbols, 2) 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) show that for general symbolic dynamical systems, the multifractal f_H-spectra in the sequence space coincide with the f_H-spectra on spatial sequence representations.**

Keywords— **Multifractal theory, Iterative function systems, Chaos game representation, Entropy spectra.**

I. Introduction

PERCEIVED order, disorder and recurrence are common features observed in complex symbolic sequences. Such sequences can be found in finance, or nature, or can be produced by chaotic dynamical systems via symbolic dynamics. This observation has lead researches to design simple but informative sequence representations for quick detection of subsequence topological and metric structures. For example, Mayer-Kress et al. [1] proposed a system for auditory representation of chaotic sequences, Berthelsen, Glazier and Skolnik [2] 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 [2], [3].

In [4] Jeffrey investigated a graphic representation of DNA sequences using iterative function systems [5]. A DNA sequence is represented by points within 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.

Jeffrey [4] expressed a need to have a mathematical description of the CGR and concluded that in an intuitive sense, the CGR represents statistical properties of DNA sequences.

The same sentiment was expressed by Berthelsen, Glazier and Skolnik [2]: while the CGR method provided an interesting tool for visualizing subsequence structure through geometric patterns, it lacks a mathematical characterization in terms of their fractal dimension.

Oliver et al. [6] and Román-Roldan et al. [7] used CGR (under the name chaos sequence representation) as a basis for computation of information theory based features of DNA strands such as subsequence entropic profiles with respect to block lengths and sequence non-randomness measures.

Basu et al. [8], Solovyev et al. [9], Fiser et al. [10] and others (see also references in [10]) generalized CGR to accommodate larger alphabets while remaining in low-dimensional visualization spaces.

In spite of further developments in GCR-based methods [11], [12], [6], [7], [13], the GCR methodology for graphic representation of DNA sequences has been lacking a rigorous connection between its spatial scaling characteristics and the statistical characteristics of the DNA sequences themselves.

In this contribution, we rigorously analyze the properties of geometric sequence representations introduced by Jeffrey [4] within the framework of multifractal theory [14].

After an intuitive explanation of the ideas behind Jeffrey's CGR in section 2, in the subsequent section (section 3) we formally define three types of geometric sequence representations and show how they relate to each other. Section 3 also 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 then studied in sections 4 and 5 respectively. In section 6, we mention related work predominantly done in the image compression and dynamical systems' commu-

This work was supported by the Austrian Science Fund (FWF) within the research project "Adaptive Information Systems and Modeling in Economics and Management Science" (SFB 010). The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Science and Transport. P. Tiňo is with the Austrian Research Institute for Artificial Intelligence, Schottengasse 3, A-1010 Vienna, Austria. E-mail: petert@ai.univie.ac.at. He is also with the Department of Computer Science and Engineering, Slovak University of Technology, Ilkovicova 3, 812 19 Bratislava, Slovakia

nities. The conclusion sums up key results of the paper and discusses possible applications of geometric sequence representations in modeling complex symbolic sequences.

II. Preliminary examples

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

$$i(x) = \frac{1}{2}(x + t_i), \quad t_i \in \{0,1\}^2, \quad t_i \neq t_j \text{ for } i \neq j, \qquad (1)$$

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

We identify the DNA bases A,C,T and G with the four maps 1,2,3 and 4 respectively. The chaos game representation $CGR(S)$ of a sequence $S = s_1 s_2 ...,\ \ s_j \in \{1,2,3,4\}$ is obtained as follows:

1. Start in the middle of the unit square, $x_0 = \{\frac{1}{2}\}^2$.
2. Plot the point $x_n = i(x_{n-1}),\ n \geq 1$, provided the n-th base s_n is i.

Each map i maps the unit square X into one of its four quadrants. Each dimer $ij \in \{1,2,3,4\}^2$ corresponds to the composition $(j \circ i)$ of maps i, j, that maps X into one of its 16 subquadrants, etc...

If a DNA sequence S were generated by a Bernoulli source[1] with equal symbol probabilities, the set $CGR(S)$ would "uniformly" sample the unit square X. Natural DNA sequences convey a genetic information and therefore it is reasonable to expect that some base subsequences will be highly accentuated, while others may be completely missing. Frequent subsequences with long common suffices form densely populated areas in the unit square. Missing subsequences manifest themselves through uninhabited "white" regions in X.

As an example, we generated a long sequence S_1 of 20,000 symbols from a Bernoulli source with equal symbol probabilities $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$. The representation $CGR(S_1)$ can be seen in figure 1a. Then, we generated another sequence S_2 (also containing 20,000 symbols) from a Bernoulli source with symbol probabilities $p_1 = p_2 = p_3 = \frac{10}{31}$, $p_4 = \frac{1}{31}$. This time, subsequences containing the symbol 4 are very rare, which is demonstrated by white regions in $CGR(S_2)$ shown in figure 1b. Points in $CGR(S_2)$ approximate a "noisy" Sierpinski triangle [5]. Chaos game representations of chaotic symbolic sequences and sequences generated by stochastic automata can be found in [15].

III. Formal definitions

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

[1] each symbol in S is generated with respect to a given distribution on the alphabet $\{1,2,3,4\}$, independently of all the other symbols in S

[2] excluding the empty word

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

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

A. Geometric representations of symbolic sequence structure

In this paper we study iterative function systems (IFSs) [5] acting on the N-dimensional unit hypercube $X = [0,1]^N$, where[3] $N = \lceil \log_2 A \rceil$. To keep the notation simple, we slightly abuse mathematical notation and, depending on the context, regard the symbols $1, 2, ..., A$, as integers, or as maps on X. The maps $i = 1, 2, ..., A$, constituting the IFS are affine contractions

$$i(x) = kx + (1-k)t_i, \quad t_i \in \{0,1\}^N, \quad t_i \neq t_j \text{ for } i \neq j, \quad (2)$$

with contraction coefficient $k \in (0, \frac{1}{2}]$.

The attractor of the IFS (2) is the unique set $K \subseteq X$, known as the Sierpinski sponge [16], for which $K = \bigcup_{i=1}^{A} i(K)$ [5].

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

$$\begin{aligned} u(x) &= u_n(u_{n-1}(...(u_2(u_1(x)))...)) \\ &= (u_n \circ u_{n-1} \circ ... \circ u_2 \circ u_1)(x) \end{aligned} \qquad (3)$$

is considered a geometrical representation of the string u under the IFS (2). For a set $Y \subseteq X$, $u(Y)$ is then $\{u(x) |\ x \in Y\}$.

From now on, the center $\{\frac{1}{2}\}^N$ of the hypercube X will be denoted by x_*. Given a sequence $S = s_1 s_2 ... \in \mathcal{A}^\infty$, its *chaos game representation* is formally defined as the sequence of points[4]

$$CGR_k(S) = \left\{ S_1^i(x_*) \right\}_{i \geq 1}. \qquad (4)$$

The chaos game representation $CGR_{\frac{2}{5}}$ of a sequence with 20,000 symbols generated by a Bernoulli source with equal symbol probabilities $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$ can be seen in figure 2. It approximates the attractor of the IFS (2) with contraction coefficient $k = \frac{2}{5}$.

When $k = \frac{1}{2}$ and $\mathcal{A} = \{1,2,3,4\}$, we recover the IFS used by Jeffrey and others [4], [6], [7] to construct the chaos game representation of DNA sequences.

It will be more instructive first to work with a geometric representation of the n-block structure in symbolic sequences. Using the IFS (2), the *chaos n-block representation* $CBR_{n,k}(S)$ of a sequence $S = s_1 s_2 ... \in \mathcal{A}^\infty$ is the sequence of points

$$CBR_{n,k}(S) = \left\{ S_i^{i+n-1}(x_*) \right\}_{i \geq 1}. \qquad (5)$$

[3] for $x \in \Re$, $\lceil x \rceil$ is the smallest integer y, such that $y \geq x$

[4] the subscript k in $CGR_k(S)$ identifies the contraction coefficient of the IFS used for the geometric sequence representation

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

$$CBR^R_{n,k}(S) = \left\{ \left(S_i^{i+n-1}\right)^R (x_*) \right\}_{i \geq 1}. \qquad (6)$$

The sequences $CBR_{n,k}(S)$ and $CBR^R_{n,k}(S)$ contain a point for each n-block in S.

Let the distance between any two equally long sequences $\{x_i\}_{i\geq 1}$ and $\{y_i\}_{i\geq 1}$ of points in $\Re^N$ be defined as

$$d_S(\{x_i\},\{y_i\}) = \sup_i d_E(x_i, y_i),$$

where d_E is the Euclidean distance. It is easy to see that for large enough n, up to points associated with the initial n-block, the n-block representations $CBR_{n,k}(S)$ closely approximate the original chaos game representations $CGR_k(S)$.

Theorem 1: Consider a sequence $S = s_1 s_2 ... \in \mathcal{A}^\infty$. Denote by $CGR_{n,k}(S)$ the sequence $CGR_k(S)$ without the first $n-1$ points. Then,

$$d_S(CBR_{n,k}(S), CGR_{n,k}(S)) \leq k^n \sqrt{N}.$$

Proof: First, note that if $v \in \mathcal{A}^+$ is a suffix of a string $u = rv$, $r, u \in \mathcal{A}^+$, then $u(X) \subset v(X)$. To see this, recall that compositions of contractions are themselves contractions and so $r(X) \subset X$. Therefore $u(X) = v(r(X)) \subset v(X)$.

The set $v(X)$ is an N-dimensional hypercube of side length $k^{|v|}$, where $|v|$ is the length of the string v. It follows that

$$\mathrm{Diam}(v(X)) = \max_{x,y \in v(X)} \{d_E(x,y)\} = k^{|v|}\sqrt{N}.$$

Now, for any position $i \geq 1$, the n-block

$$S_i^{i+n-1} = s_i s_{i+1} ... s_{i+n-1}$$

at the position i is a suffix of the initial $(i+n-1)$-block

$$S_1^{i+n-1} = s_1 s_2 ... s_{i+n-1}.$$

Hence, $S_1^{i+n-1}(X) \subseteq S_i^{i+n-1}(X)$, and

$$\mathrm{Diam}(S_i^{i+n-1}(X)) = k^n \sqrt{N}.$$

Consequently,

$$\begin{aligned} d_S(CBR_{n,k}(S), CGR_{n,k}(S)) &\leq \max_{w \in \mathcal{A}^n} \{\mathrm{Diam}(w(X))\} \\ &= k^n \sqrt{N}. \qquad (7) \end{aligned}$$

■

B. Statistics on symbolic sequences

Let $S = s_1 s_2 ... \in \mathcal{A}^\infty$ be a sequence generated by a stationary information source [17]. Denote the (empirical) probability of finding an n-block $w \in \mathcal{A}^n$ in S by $P_n(w)$. A string $w \in \mathcal{A}^n$ 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

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

If information is measured in bits, then $\log \equiv \log_2$. 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).

The block entropies H_n and entropy rates h_n are special cases of Rényi entropies and entropy rates [18]. 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

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

reduce to the block entropy $H_n(S)$ and entropy rate $h_n(S)$ when $\beta = 1$ [19]. The formal parameter β can be thought of as the inverse temperature in the statistical mechanics of spin systems [20]. 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.

C. 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 M is usually quantified through its fractal dimension $D(M)$ [5]. Denote by $N(\ell)$ the minimal number of hyperboxes of side length ℓ needed to cover

M. The fractal (box-counting) dimension $D(M)$ relates the side length ℓ with $N(\ell)$ via the scaling law $N(\ell) \approx \ell^{-D(M)}$.

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

$$N(k^n) = (k^n)^{-D_{n,k}(M)}.$$

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 $D_\beta(M)$ of an object M describe a measure μ on M through the scaling law

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

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

In particular, for $0 < k \leq \frac{1}{2}$, the n-th order approximation $D_{\beta,n,k}(M)$ of $D_\beta(M)$ is given by

$$\sum_{B \in \mathcal{B}_\ell,\ \mu(B)>0} \mu^\beta(B) = \ell^{(\beta-1)D_{\beta,n,k}(M)}, \tag{10}$$

where $\ell = k^n$.

The infinite temperature scaling exponent $D_0(M)$ is equal to the box-counting fractal dimension $D(M)$ of M. 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 of regions where the probability is most concentrated and rarefied, respectively.

IV. Chaos representations of single sequences

Intuitively, for any $k \in (0, \frac{1}{2}]$, the degree of fragmentation of the chaos n-block representation $CBR_{n,k}(S)$ of a sequence S is closely related to the growth rate of allowed n-blocks in S, i.e. to the topological entropy $h_0(S)$ of S (see section III-B).

If S is a periodic sequence of period r, then $CBR_{n,k}(S)$ will have precisely r *distinct* points for all $n \geq r$. Hence, for large block lengths n, the n-th order approximation $D_{n,k}(CBR_{n,k}(S))$ of the fractal dimension of $CBR_{n,k}(S)$ tends to zero. On the other hand, the number of points in the chaos n-block representation of a "chaotic" sequence S with positive topological entropy increases with increasing n. As the block length n grows, the fractal dimension approximation $D_{n,k}(CBR_{n,k}(S))$ does not decrease to zero. Actually, as we will show, it approaches the (properly scaled) topological entropy of the sequence S.

Let the measures μ_n and ν_n on $X = [0,1]^N$ describe the relative frequencies of points from the sequences $CBR_{n,k}(S)$ and $CBR^R_{n,k}(S)$, respectively, on the Lebesgue subsets of X. The next theorem establishes the relationship between the Rényi entropy spectra of a sequence S and the generalized dimension spectra of its chaos block representations.

[5] at most up to Lebesgue measure zero borders

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

$$D_{\beta,n,k}(CBR_{n,k}(S)) = D_{\beta,n,k}(CBR^R_{n,k}(S)) = \frac{h_{\beta,n}(S)}{\log \frac{1}{k}}.$$

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

$$\begin{aligned}\lim_{n\to\infty} D_{\beta,n,k}(CBR_{n,k}(S)) &= \lim_{n\to\infty} D_{\beta,n,k}(CBR^R_{n,k}(S)) \\ &= \frac{h_{(\beta)}(S)}{\log \frac{1}{k}},\end{aligned}$$

provided the limits exist.

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

From (10)

$$\sum_{w \in [S]_n} \mu_n^\beta(w(X)) = \ell^{(\beta-1)D_{\beta,n,k}(CBR_{n,k}(S))},$$

and so

$$\log \sum_{w \in [S]_n} \mu_n^\beta(w(X)) = n(1-\beta)D_{\beta,n,k}(CBR_{n,k}(S)) \log k^{-1}$$

which means that

$$D_{\beta,n,k}(CBR_{n,k}(S)) = \frac{1}{n(1-\beta)\log k^{-1}} \log \sum_{w \in [S]_n} \mu_n^\beta(w(X)) =$$

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

The entropies introduced in section III-B do not contain any notion of causality, or time location. Since

$$\begin{aligned}\sum_{w \in [S]_n} \mu_n^\beta(w(X)) &= \sum_{w \in [S]_n} P_n^\beta(w) \\ &= \sum_{w \in [S]^R_n} P_n^\beta(w^R) = \sum_{w \in [S]^R_n} \nu_n^\beta(w^R(X)),\end{aligned}$$

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

$$D_{\beta,n,k}(CBR_{n,k}(S)) = D_{\beta,n,k}(CBR^R_{n,k}(S)).$$

■

Note that, with the exception of sequences $S = ws^\omega$, $w \in \mathcal{A}^+$, $s \in \mathcal{A}$, the limit n-block representations $\lim_{n\to\infty} CBR_{n,k}(S)$ do not exist. However, the dimension estimates $D_{\beta,n,k}(CBR_{n,k}(S))$ may still converge. On the other hand, the reversed limit block representations $\lim_{n\to\infty} CBR^R_{n,k}(S)$ do exist for all $S \in \mathcal{A}^\omega$.

Theorem 1 and arguments in its proof tell us that the chaos game representation $CGR_{k,n}(S)$ (without the points corresponding to the first n-block) of a sequence $S \in \mathcal{A}^\infty$ creates the same relative frequency of points on boxes $w(X)$, $w \in \mathcal{A}^n$, of side length k^n as does the chaos n-block representation $CBR_{n,k}(S)$. Theorems 1 and 2 imply the following corollary.

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

$$D_{\beta,n,k}(CGR_{n,k}(S)) = \frac{h_{\beta,n}(S)}{\log \frac{1}{k}}.$$

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

$$D_{\beta,n,k}(CGR_k(S)) = \frac{h_{\beta,n}(S)}{\log \frac{1}{k}}.$$

Proof: The first statement is an immediate corollary of theorems 1 and 2. To justify the second statement, write the sequence $S \in \mathcal{A}^\omega$ as $S = wS'$, where $w \in \mathcal{A}^n$ and $S' \in \mathcal{A}^\omega$. Since n is finite, the empirical n-block probabilities $P_n(w)$, $w \in \mathcal{A}^n$, are the same for both sequences S and S'. ■

Hence, for infinite sequences $S \in \mathcal{A}^\omega$, when $k = \frac{1}{2}$, the generalized dimension estimates of geometric chaos game representations exactly equal the corresponding sequence Rényi entropy rate estimates. In particular, given an infinite sequence $S \in \mathcal{A}^\omega$, as n grows, the box-counting fractal dimension and the information dimension estimates $D_{0,n,\frac{1}{2}}$ and $D_{1,n,\frac{1}{2}}$ of the original Jeffrey chaos game representation [4], [6], [7] tend to the sequence topological and metric entropies respectively.

Besides the generalized dimensions D_β, a coarse grained dimension spectrum f_G, also known as the large deviation spectrum, is widely used to characterize multifractal measures [23]. The coarse grained spectrum is related to the generalized dimension spectrum D_β through

$$T_\beta = f_G^*(\beta) = \inf_{\alpha \in \Re} (\beta\alpha - f_G(\alpha)),$$

where $T_\beta = (\beta - 1)D_\beta$. Stated differently, T_β is the Legendre transform of f_G [21].

Write $T_{\beta,n,k} = (\beta - 1)D_{\beta,n,k}$. Then, according to theorem 2, as the block length grows, the n-th order approximations $T_{\beta,n,k}(CBR_{n,k}(S))$ and $T_{\beta,n,k}(CBR^R_{n,k}(S))$ approach the scaled Rényi entropy

$$\frac{1-\beta}{\log k} h_{(\beta)}(S). \tag{11}$$

It follows that the Legendre transform of the large deviation spectrum f_G of the reversed limit block representation $\lim_{n\to\infty} CBR^R_{n,k}(S)$ coincides with the scaled Rényi entropy rates (11).

V. Chaos block representations of symbolic dynamical systems

In the previous section we analysed geometric and measure scaling properties of the chaos sequence/block representations of a given (possibly infinite) symbolic sequence. In this section we investigate properties of the chaos block representations in the limit of infinite block lengths, in the context of general symbolic dynamical systems.

The set of infinite sequences $S = s_1 s_2 ... \in \mathcal{A}^\omega$ over the alphabet $\mathcal{A} = \{1, 2, ..., A\}$, endowed with a metric[6]

$$d_\lambda(S, S') = \sum_{i=1}^{\infty} \frac{|s_i - s_i'|}{\lambda^i}, \quad \lambda > 2$$

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

Each sequence $S \in \mathcal{A}^\omega$ is coded by a point

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

on the attractor K of the iterative function system (2).

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

Even in the very simple case of the full shift $Q = \mathcal{A}^\omega$ with a measure τ on $(\mathcal{A}^\omega, d_\lambda)$ defined by a Bernoulli source with unequal symbol probabilities, the distribution of sequences S in $(\mathcal{A}^\omega, d_\lambda)$, as well as the distribution of points $\chi(S)$ on[8] (K, d_E), are singular [23]. In this case, one cannot describe the distributions by means of densities. Multifractal analysis proves useful in characterizing the complicated geometrical properties of the measure τ on $(\mathcal{A}^\omega, d_\lambda)$ 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 $B \subset \mathcal{A}^\omega$ is a ball containing the sequence S and that its diameter

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

tends to zero.

Usually, points of equal strength lie on interwoven fractal sets

$$K_\alpha = \{S \in \mathcal{A}^\omega \,|\, \alpha(S) = \alpha\}.$$

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 $\dim_H(K_\alpha)$ [24],

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

[6]Usually, the metric d_λ is defined for $\lambda > 1$. In this paper we confine ourselves to the family of metrics d_λ, $\lambda > 2$.

[7]with respect to the topology induced by d_λ

[8]recall that d_E denotes the Euclidean metric

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

$$\lim_{n\to\infty} CBR^R_{n,k}(Q) = \{\chi(\sigma^i(S)) |\ S \in Q,\ i = 0, 1, 2, ...\}$$

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 $(\mathcal{A}^\omega, d_\lambda)$ with the pushed forward (via the map χ) measure ν on the metric space (K, d_E). To this end, we shall prove the metric equivalence of the two metric spaces $(\mathcal{A}^\omega, d_\lambda)$ and $(\chi(\mathcal{A}^\omega), d_E)$ that correspond to each other through the coding map χ. Under such metric equivalence, Hölder exponents $\alpha(S)$ of the measure τ coincide with the Hölder exponents $\alpha(\chi(S))$ of the pushed forward measure ν. Since the Hausdorff dimension is a metric equivalence invariant [24], we get that the multifractal spectra $f_H(\alpha)$ for measures τ and ν are the same. In other words, in the limit of infinite block lengths, the multifractal f_H-spectrum in the sequence space $(\mathcal{A}^\omega, d_\lambda)$ is the same as the f_H-spectrum in the space $(\chi(\mathcal{A}^\omega), d_E)$ of reversed chaos block representations.

A. *Metric equivalence between the spaces $(\mathcal{A}^\omega, d_\lambda)$ and $(\chi(\mathcal{A}^\omega), d_E)$*

The two metrics d_λ and $\tilde{d}_\lambda$ on $\mathcal{A}^\omega$ are equivalent if there exist constants $0 < c_1 < c_2 < \infty$, such that for all $S, S' \in \mathcal{A}^\omega$,

$$c_1 \tilde{d}_\lambda(S, S') \leq d_\lambda(S, S') \leq c_2 \tilde{d}_\lambda(S, S').$$

The two metric spaces $(\mathcal{A}^\omega, d_\lambda)$ and $(\chi(\mathcal{A}^\omega), d_E)$ are equivalent, if there is a bijective map $h : \mathcal{A}^\omega \to \chi(\mathcal{A}^\omega)$, such that the induced metric $\tilde{d}_\lambda$

$$\tilde{d}_\lambda(S, S') = d_E(h(S), h(S'))$$

is equivalent to the metric d_λ.

Theorem 3: The metric spaces $(\mathcal{A}^\omega, d_\lambda)$ and $(\chi(\mathcal{A}^\omega), d_E)$ are equivalent if and only if $\lambda = \frac{1}{k}$, $k \in (0, \frac{1}{2})$.

Proof: Sequences $S \in \mathcal{A}^\omega$ are coded as points on $\chi(\mathcal{A}^\infty) \subseteq K$ via the map χ (eq. (12)). The map χ is one-to-one. We will show that

$$\forall S, S' \in \mathcal{A}^\omega,\quad \exists\ 0 < c_1 < c_2 < \infty \ \text{ such that}$$

$$c_1 d_E(\chi(S), \chi(S')) \leq d_\lambda(S, S') \leq c_2 d_E(\chi(S), \chi(S')). \quad (14)$$

Suppose S and S' have the longest common prefix of length L, i.e. $S = ws_1S_1,\ S' = ws_2S_2,\ w \in \mathcal{A}^L,\ S_1, S_2 \in \mathcal{A}^\omega$ and $s_1, s_2 \in \mathcal{A}$ are different symbols.

First, we shall take care of the second part of the inequality (14) by concentrating on sequences S, S' with maximal distance $d_\lambda(S, S')$ in $(\mathcal{A}^\omega, d_\lambda)$, but minimal code distance $d_E(\chi(S), \chi(S'))$ in (K, d_E).

Since the longest common suffix of S, S' has length L, the distance $d_\lambda(S, S')$ can be upper bounded by

$$d_\lambda(S, S') \leq \sum_{i=L+1}^{\infty} \frac{A-1}{\lambda^i} = \frac{1}{\lambda^L}\frac{A-1}{\lambda - 1}. \quad (15)$$

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

$$d_E(\chi(S), \chi(S')) \geq k^L(1 - 2k), \quad (16)$$

because the two sequences S, S' differ in the $(L+1)$-st symbol and so the points $\chi(S)$, $\chi(S')$ lie in distinct subcubes $(ws_1)^R(X)$, $(ws_2)^R(X)$ of the cube $w^R(X)$ with side length k^L.

From (14), (15) and (16), we get for the constant c_2

$$\frac{1}{\lambda^L}\frac{A-1}{\lambda-1} \leq c_2 k^L (1 - 2k),$$

which means (for $k \in (0, \frac{1}{2})$)

$$c_2 \geq \frac{1}{(\lambda k)^L}\frac{A-1}{\lambda-1}\frac{1}{1-2k}. \quad (17)$$

We now take a closer look at the first part of the inequality (14).

The distance $d_\lambda(S, S')$ is minimal when the sequences S, S' differ by a minimum amount just in their $(L+1)$-st symbols. In this case

$$d_\lambda(S, S') = \frac{1}{\lambda^{L+1}}. \quad (18)$$

The maximal distance between the codes $\chi(S), \chi(S')$ of any two infinite sequences S, S' having the longest common prefix w of length L can be bounded from above by[10] $\mathrm{Diam}(w(X))$ (see eq. (7))

$$d_E(\chi(S), \chi(S')) \leq k^L\sqrt{N}, \quad (19)$$

and so, from (14), (18) and (19),

$$c_1 k^L \sqrt{N} \leq \frac{1}{\lambda^{L+1}},$$

which implies

$$c_1 \leq \frac{1}{(\lambda k)^L}\frac{1}{\lambda\sqrt{N}}. \quad (20)$$

The inequalities (17) and (20) should hold for all prefix lengths $L = 0, 1, 2, ...$ and c_1, c_2 are bounded positive constants. This is possible only when $\lambda = \frac{1}{k}$. ■

Theorem 3 is an extension of the known metric equivalence between the sequence metric space $(\{0, 1, 2, ..., A-1\}^\omega, d_A)$ and the $(A+1)$-ary Cantor set $\mathcal{K}$ in $([0, 1], d_E)$. The set $\mathcal{K}$ consists of all points whose $(A+1)$-ary representation does not contain the digit A (see for example [5]).

VI. Related work

The ideas behind Jeffrey's chaos sequence representation of symbolic sequences were independently studied in the image compression community. Quadtree [5] 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

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

[10] recall that $N = \lceil \log_2 A \rceil$

$i = 0, 1, 2, 3$. 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... Using our notation, points whose addresses start with $v \in \{0, 1, 2, 3\}^n$ lie in the square $v^R(X)$, where the maps $0, 1, 2$ and 3 are defined in (1). So the quadtree scheme is equivalent to our reversed n-block representations of symbolic sequences over the alphabet $\{0, 1, 2, 3\}$, with contraction ratio $k = \frac{1}{2}$. 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).

Culik II and Dube [26], [27] investigated several methods for fractal image description (compression) based on iterative function systems driven by a prescribed set $\mathcal{L}$ of symbolic sequences. Typically, the set $\mathcal{L}$ was taken to be a regular language.

Recently, there has been an extensive research activity in fractal and multifractal analysis of strange sets arising in chaotic dynamical systems. The strange sets are modelled using Moran-like constructions [14], [28]. Briefly, the Moran-like geometric constructions iteratively construct limit sets using a collection of basic sets that may have a very complicated geometry. The basic sets have symbolic addresses and are increasingly refined (with increasing address length) according to a given symbolic dynamical system. As the construction proceeds, the diameter of the basic sets diminishes to zero. In the limit of infinite block lengths, our reversed chaos block representations of symbolic dynamical systems are special cases of Moran-like geometric constructions. Actually, they correspond to Moran constructions [29] driven by general symbolic dynamical systems.

The setting of Moran-like geometric constructions is much more general than our setting of chaos block representations and the multifractal analysis of Moran-like constructions is primarily concerned with validity of the multifractal formalism [30] (i.e. the $\bigcap$-shape of the f_H-spectra for dimensions, differentiability of the spectra, Legendre relations between the multifractal quantities, etc..., see [31]).

An important result of Pesin and Weiss [28] states that the Hausdorff and box-counting dimensions of the limit sets of Moran-like constructions coincide. Consequently, for geometric representations considered in this paper, the n-th order dimension approximations $D_{0,n,k}$ tend to the representations' both box-counting and Hausdorff dimensions.

VII. Conclusion

We investigated how the scaling properties of the chaos game sequence representations proposed by Jeffrey [4] correspond to the statistical properties of the sequences they represent. Although closely related to the original Jeffrey's approach, our investigation was performed in a more general setting:

1. We allowed alphabets with an arbitrary finite number of symbols $A \geq 2$.
2. Besides the chaos game representation, spatial representations of sequence n-block structure were studied and shown to be closely related to the original chaos game representation.
3. Contraction ratios k other than $\frac{1}{2}$ were allowed ($k \in (0, \frac{1}{2}]$).
4. We studied geometric representations of single sequences, as well as general symbolic dynamical systems.

We have shown that

1. 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.
2. In the limit of infinite block lengths
- the Legendre transform of the large deviation spectra f_G of the reversed chaos block representations is equal to the (scaled) spectrum of sequence Rényi entropy rates
- the multifractal f_H-spectrum for dimensions in the sequence space $(\mathcal{A}^\omega, d_\lambda)$ coincides with the f_H-spectrum of the geometric reversed chaos block representation, provided λ is the inverse of the geometric representations' contraction coefficient k.

For alphabets $\mathcal{A}$ of up to eight symbols, theorem 2 and corollary 1 suggest using the geometric representations studied in this paper as illustrative, visual codings of the n-block statistical structures in sequences over $\mathcal{A}$. For each n-block $w \in \mathcal{A}^n$, the cube $w(X)$ is colored according to its probability $P_n(w)$. This approach was taken in [15] to monitor the training process of recurrent neural networks and stochastic machines on chaotic symbolic sequences. At certain stages of the training process the models M were used to generate sequences $M(S)$ of length equal to the length of the training sequence S. Then, the n-block representations of the sequences S and $M(S)$ were compared.

The geometric representations of symbolic sequences S can also be used to construct context sensitive predictive models similar in spirit to variable memory length Markov models [32], [33]. The history of the input stream S is translated into clusters in the chaos representations $CGR_k(S)$ and $CBR_{n,k}(S)$. Points lying in a close neighborhood code histories with a long common suffix (e.g. histories that are likely to produce similar continuations), whereas histories with different suffices (and potentially different continuations) are mapped to points lying far from each other. We then apply a vector quantizer to the sequence representations and interpret codebook vectors as prediction contexts. Dense areas correspond to contexts with common suffices and are given more attention by the vector quantizer. Consequently, more information processing states of the predictive model are devoted to these potentially "problematic" contexts. This directly corresponds to the idea of variable length Markov models, where the length of the past history considered in order to predict the future is not fixed, but context dependent.

The results in modeling long, complex symbolic sequences are very promising [34]. In comparison with the traditional variable length Markov models, the context sensitive prediction models based on chaos sequence representations are cheaper to construct and have a comparable or better modeling performance[11].

Acknowledgments

Many thanks go to Rudolf Riedi, Luis Barreira and Lars Olsen for helpful comments on the multifractal formalism. I am also thankful to anonymous reviewers and Georg Dorffner for helpuful comments on the manuscript.

References

[1] G. Mayer-Kress, R. Bargar, and I. Choi, "Musical structures in data from chaotic attractors," Tech. Rep. CCSR-92-14, Center for Complex Systems Research, Beckman Institute, University of Illinois at Urbana-Champaign, 1992.

[2] C.L. Berthelsen, J.A. Glazier, and M.H. Skolnik, "Global fractal dimension of human dna sequences treated as pseudorandom walks," *Physical Review A*, vol. 45, no. 12, pp. 8902–8913, 1992.

[3] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, Z.D. Goldberger, S. Havlin, R.N. Mantegna, S.M. Ossadnik, C.-K. Peng, and M. Simons, "Statistical mechanics in biology: how ubiquitous are long-range correlations?," *Physica A*, vol. 205, pp. 214–253, 1994.

[4] J. Jeffrey, "Chaos game representation of gene structure," *Nucleic Acids Research*, vol. 18, no. 8, pp. 2163–2170, 1990.

[5] M.F. Barnsley, *Fractals everywhere*, Academic Press, New York, 1988.

[6] J.L. Oliver, P. Bernaola-Galván, J. Guerrero-Garcia, and R. Román Roldan, "Entropic profiles of dna sequences through chaos-game-derived images," *Journal of Theor. Biology*, , no. 160, pp. 457–470, 1993.

[7] R. Roman-Roldan, P. Bernaola-Galvan, and J.L. Oliver, "Entropic feature for sequence pattern through iteration function systems," *Pattern Recognition Letters*, vol. 15, pp. 567–573, 1994.

[8] S. Basu, A. Pan, C. Dutta, and Das J, "Chaos game representation of proteins," *Journal of Molecular Graph. Model.*, vol. 15, no. 5, pp. 279–289, 1997.

[9] V.V. Solovyev, S.V. Korolev, and H.A. Lim, "A new approach for the classification of functional regions of dna sequences based on fractal representation," *International Journal of Genomic Research*, vol. 1, no. 1, pp. 109–128, 1993.

[10] A. Fiser, G.E. Tusnady, and I. Simon, "Chaos game representation of protein structures," *Journal of Molecular Graphics*, vol. 12, no. 4, pp. 302–304, 1994.

[11] K.A. Hill and S.M. Singh, "The evolution of species-type specificity in the global dna sequence organization of mitochondrial genomes," *Genome*, vol. 40, pp. 342–356, 1997.

[12] W. Li, "The study of correlation structures of dna sequences: a critical review," *Computer and Chemistry*, vol. 21, no. 4, pp. 257–272, 1997.

[13] R. Roman-Roldan, P. Bernaola-Galvan, and J.L. Oliver, "Application of information theory to dna sequence analysis: a review," *Pattern Recognition*, vol. 29, no. 7, pp. 1187–1194, 1996.

[14] Y. Pesin, *Dimension Theory in Dynamical Systems: Rigorous Results and Applications*, University of Chicago Press, Chicago, 1997.

[15] P. Tiňo and M. Köteleš, "Extracting finite state representations from recurrent neural networks trained on chaotic symbolic sequences," *IEEE Transactions on Neural Networks*, in press, 1999.

[16] R. Kenyon and Y. Peres, "Measures of full dimension on affine invariant sets," *Ergodic Theory and Dynamical Systems*, vol. 16, pp. 307–323, 1996.

[17] A.I. Khinchin, *Mathematical Foundations of Information Theory*, Dover Publications, New York, 1957.

[11] The performance criterion is the estimated cross-entropy between the training and model generated sequences

[18] A. Renyi, "On the dimension and entropy of probability distributions," *Acta Math. Hung.*, , no. 10, pp. 193, 1959.

[19] P. Grassberger, "Information and complexity measures in dynamical systems," in *Information Dynamics*, H. Atmanspacher and H. Scheingraber, Eds., pp. 15–33. Plenum Press, New York, 1991.

[20] J.P. Crutchfield and K. Young, "Computation at the onset of chaos," in *Complexity, Entropy, and the physics of Information, SFI Studies in the Sciences of Complexity, vol 8*, W.H. Zurek, Ed. 1990, pp. 223–269, Addison-Wesley.

[21] C. Beck and 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, 1994.

[23] R.H. Riedi, "Conditional and realtive multifractal spectra," *Fractals*, vol. 5, no. 1, pp. 153–168, 1997.

[24] K.J. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, John Wiley and Sons, New York, 1990.

[25] L. Staiger, "Quadtrees and the hausdorff dimension of pictures," in *Workshop on Geometrical Problems of Image Processing, Georgental, GDR*, 1989, pp. 173–178.

[26] K. Culik(II) and S. Dube, "Affine automata and related techniques for generation of complex images," *Theoretical Computer Science*, vol. 116, no. 2, pp. 373–398, 1993.

[27] K. Culik(II) and S. Dube, "Rational and affine expressions for image description," *Discrete Appl. Math.*, vol. 41, pp. 85–120, 1993.

[28] Y. Pesin and H. Weiss, "On the dimension of deterministic and cantor-like sets, symbolic dynamics, and the eckmann-ruelle conjecture," *Comm. Math. Phys*, vol. 182, no. 1, pp. 105–153, 1996.

[29] P. Moran, "Additive functions of intervals and hausdorff dimension," *Proceedings of the Cambridge Philosophical Society*, vol. 42, pp. 15–23, 1946.

[30] Y. Pesin and H. Weiss, "A multifractal analysis of equilibrium measures for conformal expanding maps and moran-like geometric constructions," *J. Stat.Phys.*, vol. 86, no. 1–2, pp. 233–275, 1997.

[31] L. Barreira, Y. Pesin, and J. Schmeling, "On a general concept of multifractality: Multifractal spectra for dimensions, entropies, and lyapunov exponents. multifractal rigidity," *Chaos: an Interdisciplinary Journal of Nonlinear Science*, vol. 7, no. 1, pp. 27–53, 1996.

[32] D. Ron, Y. Singer, and N. Tishby, "The power of amnesia," *Machine Learning*, vol. 25, 1996.

[33] M.J. Weinberger, J.J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 643–652, 1995.

[34] P. Tiňo and G. Dorffner, "Constructing finite-context sources from fractal representations of symbolic sequences," Tech. Rep. TR-98-18, Austrian Research Institute for Artificial Intelligence, Austria, 1998.