PERGAMON

Chaos, Solitons and Fractals 11 (2000) 2215–2222

CHAOS
SOLITONS & FRACTALS

www.elsevier.nl/locate/chaos

Dimensions of fractals related to languages defined by tagged strings in complete genomes ☆

Zu-Guo Yu [a,b,*], Bai-lin Hao [b], Hui-min Xie [b,c], Guo-Yi Chen [b]

[a] Department of Mathematics, Xiangtan University, Hunan 411105, People's Republic of China
[b] Institute of Theoretical Physics, Academia Sinica, P.O. Box 2735, Beijing 100080, People's Republic of China
[c] Department of Mathematics, Suzhou University, Jiangsu 215006, People's Republic of China

Accepted 28 July 1999

Abstract

A representation of frequency of strings of length K in complete genomes of many organisms in a square has led to seemingly self-similar patterns when K increases. These patterns are caused by under-represented strings with a certain "tag"-string and they define some fractals in the $K \to \infty$ limit. The Box and Hausdorff dimensions of the limit set are discussed. Although the method proposed by Mauldin and Williams to calculate Box and Hausdorff dimension is valid in our case, a different and sampler method is proposed in this paper.

1. Introduction

In the past decade or so there has been a ground swell of interest in unraveling the mysteries of DNA. The heredity information of organisms (except for so-called RNA-viruses) is encoded in their DNA sequence which is a one-dimensional unbranched polymer made of four different kinds of monomers (nucleotides): adenine (a), cytosine (c), guanine (g), and thymine (t). As long as the encoded information is concerned we can ignore the fact that DNA exists as a double helix of two "conjugated" strands and only treat it as a one-dimensional symbolic sequence made of the four letters from the *alphabet* $\Sigma = \{a, c, g, t\}$. Since the first complete genome of a free-living bacterium *Mycoplasma genitalium* was sequenced in 1995 [3], an ever-growing number of complete genomes has been deposited in public databases. The availability of complete genomes opens the possibility to ask some global questions on these sequences. One of the simplest conceivable questions consists in checking whether there are short strings of letters that are absent or under-represented in a complete genome. The answer is in the affirmative and the fact may have some biological meaning [5].

The reason why we are interested in absent or under-represented strings is twofold. First of all, this is a question that can be asked only nowadays when complete genomes are at our disposal. Second, the question makes sense as one can derive a *factorizable* language from a complete genome which would be entirely defined by the set of forbidden words.

We start by considering how to visualize the avoided and under-represented strings in a bacterial genome whose length is usually the order of a million letters.

Hao et al. [5] proposed a simple visualization method based on counting and coase-graining the frequency of appearance of strings of a given length. When applying the method to all known complete

☆ This work was partially supported by Chinese Natural Science Foundation and Chinese Postdoctoral Science Foundation.
* Corresponding author.

0960-0779/00/$ - see front matter
PII: S0960-0779(99)00141-1

genomes, fractal-like patterns emerge. The fractal dimensions are basic and important quantities to characterize the fractal. One will naturally ask the question: what are the fractal dimensions of the fractals related to languages defined by tagged strings? In this paper we will answer the question.

2. Graphical representation of counters

We call any string made of K letters from the set $\{g, c, a, t\}$ a K-string. For a given K there are in total 4^K different K-strings. In order to count the number of each kind of K-strings in a given DNA sequence 4^K counters are needed. These counters may be arranged as a $2^K \times 2^K$ square, as shown in Fig. 1 for $K = 1$ to 3.

In fact, for a given K the corresponding square may be represented as a direct product of K copies of identical matrices:

$$M^{(K)} = M \otimes M \otimes \cdots \otimes M,$$

where each M is a 2×2 matrix:

$$M = \begin{bmatrix} g & c \\ a & t \end{bmatrix},$$

which represents the $K = 1$ square in Fig. 1. For convenience of programming, we use binary digits 0 and 1 as subscripts for the matrix elements, i.e., let $M_{00} = g$, $M_{01} = c$, $M_{10} = a$, and $M_{11} = t$. The subscripts of a general element of the $2^K \times 2^K$ direct product matrix $M^{(K)}$,

$$M^{(K)}_{I,J} = M_{i_1 j_1} M_{i_2 j_2} \cdots M_{i_K j_K},$$

are given by $I = i_1 i_2 \cdots i_K$ and $J = j_1 j_2 \cdots j_K$. These may be easily calculated from an input DNA sequence

$$s_1 s_2 s_3 \cdots s_K s_{K+1} \cdots,$$

where $s_i \in \{g, c, a, t\}$. We call this $2^K \times 2^K$ square a K-frame. Put in a frame of fixed K and described by a color code biased towards small counts, each bacterial genome shows a distinctive pattern which indicates on absent or under-represented strings of certain types [5]. For example, many bacteria avoid strings containing the string *ctag*. Any string that contains *ctag* as a substring will be called a *ctag*-tagged string. If we mark all *ctag*-tagged strings in frames of different K, we get pictures as shown in Fig. 2. We also note that bacterium *Aquifex aeolius* [1] avoid strings containing the string *gcgc*. The large scale structure of these pictures persists but more details appear with growing K. Excluding the area occupied by these tagged strings, one gets a fractal F in the $K \to \infty$ limit. It is natural to ask what are the fractal dimensions of F for a given tag.

In fact, this is the dimension of the complementary set of the tagged strings. The simplest case is that of g-tagged strings. As the pattern has an apparently self-similar structure the dimension is easily calculated to be

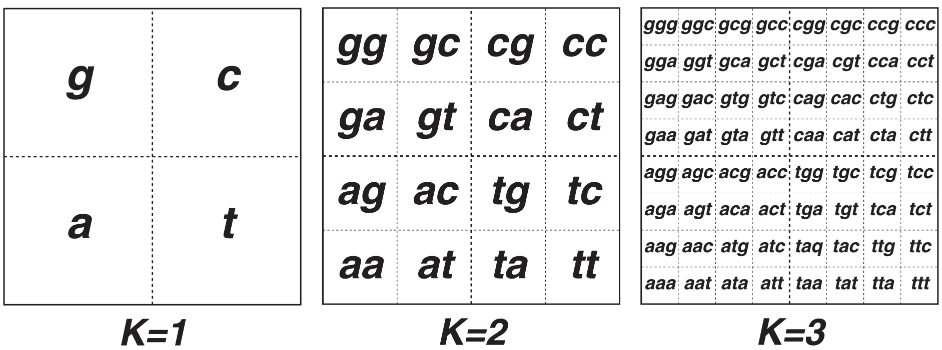

Fig. 1. The arrangement of string counters for $K = 1$ to 3 in squares of the same size.

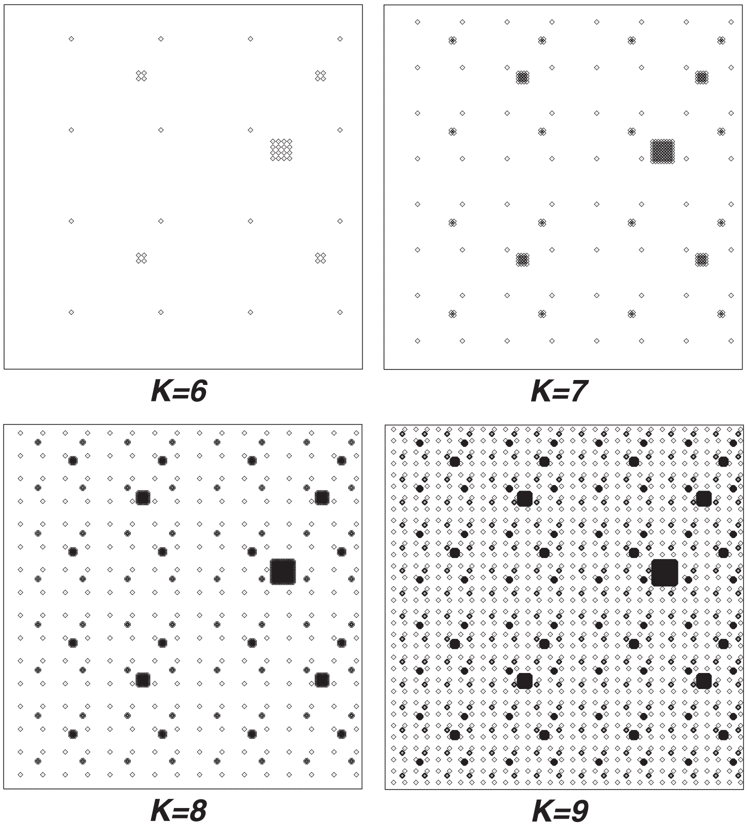

Fig. 2. *ctag*-tagged strings in $K = 6$ to 9 frames.

$$\dim_{\mathrm{H}}(F) = \dim_{\mathrm{B}}(F) = \frac{\log 3}{\log 2},$$

where $\dim_{\mathrm{H}}(F)$ and $\dim_{\mathrm{B}}(F)$ are the Hausdorff and Box dimensions [2] of F.

In formal language theory, we starts with alphabet $\Sigma = \{a, c, g, t\}$. Let Σ^* denotes the collection of all possible strings made of letters from Σ, including the empty string ϵ. We call any subset $L \subset \Sigma^*$ a *language* over the alphabet Σ. Any string over Σ is called a *word*. If we denote the given tag as w_0, for our case,

$$L = \{\text{word which does not contain } w_0 \text{ as factor}\}.$$

F is called the fractal related to language L.

3. Box dimension of fractals

When we discuss the Box dimension, we can consider more general case, i.e., the case of more than one tag. We denote the set of tags as B, and assume that there has not one element being factor of any other element in B. We define

$$L_1 = \{\text{word which does not contain any of element of } B \text{ as factor}\}.$$

Now let a_K be the number of all strings of length K that belong to language L_1. As the linear size δ_K in the K-frame is $1/2^K$, the Box dimension of F may be calculated as

$$\dim_{\mathrm{B}}(F) = \lim_{K\to\infty} \frac{\log a_K}{-\log \delta_K} = \lim_{K\to\infty} \frac{\log a_K^{1/K}}{\log 2}. \qquad (1)$$

Now we define the generating function of a_K as

$$f(s) = \sum_{K=0}^{\infty} a_K s^K,$$

where s is a complex variable.

First L_1 is a dynamic language, from Theorem 2.5.2 of Ref. [6,9], we have

$$\lim_{K\to\infty} a_K^{1/K} \quad \text{exists, we denote it as } l. \tag{2}$$

From (1), we have

$$\dim_B(F) = \frac{\log l}{\log 2}. \tag{3}$$

For any word $w = w_1 w_2 \ldots w_n, w_i \in \Sigma$ for $i = 1, \ldots, n$, we denote

$$\text{Head}(w) = \{w_1, w_1w_2, w_1w_2w_3, \ldots, w_1w_2\ldots w_{n-1}\},$$
$$\text{Tail}(w) = \{w_n, w_{n-1}w_n, w_{n-2}w_{n-1}w_n, \ldots, w_2w_3\ldots w_n\}.$$

For given two words u and v, we denote overlap$(u, v) = \text{Tail}(u) \cap \text{Head}(v)$. If $x \in \text{Head}(v)$, then we can write $v = xx'$. We denote $x' = v/x$ and define

$$u : v = \sum_{x\in \text{overlap}(u,v)} s^{|v/x|},$$

where $|v/x|$ is the length of word v/x. From Golden–Jackson Cluster method [4,8], we can know that

$$f(s) = \frac{1}{1 - 4s - \text{weight}(\mathscr{C})},$$

where $\text{weight}(\mathscr{C}) = \sum_{v\in B} \text{weight}(\mathscr{C}[v])$ and $\text{weight}(\mathscr{C}[v])$ $(v \in B)$ are solutions of the linear equations:

$$\text{weight}(\mathscr{C}[v]) = -s^{|v|} - (v : v)\text{weight}(\mathscr{C}[v]) - \sum_{\substack{u\in B\\ u\neq v}} (u : v)\text{weight}(\mathscr{C}[u]).$$

It is easy to see that $f(s)$ is a rational function. Its maximal analytic disc at center 0 has radius $|s_0|$, where s_0 is the minimal module zero point of $1/f(s)$. On the other hand, according to the Cauchy criterion of convergence we have $1/l$ is the radius of convergence of series expansion of $f(s)$. Hence $|s_0| = 1/l$. From (3), we obtain the following result.

Theorem 3.1. *The Box dimension of F is*

$$\dim_B(F) = -\frac{\log |s_0|}{\log 2},$$

where s_0 is the minimal module zero point of $1/f(s)$ and $f(s)$ is the generating function of language L_1.

In particular, the case of a single tag – B contains only one word – is easily treated and some of the results are shown in Table 1.

4. Hausdorff dimension of fractals

We obtained the Box dimension of F in the previous section. Now one will naturally ask whether the Hausdorff dimension of F equals to the Box dimension of it. In this section we will discuss the Hausdorff dimension of F. Now we only discuss the case of B contains only one word w_0. From the K-frames $(K = |w_0|, |w_0| + 1, \ldots)$, we can find:

Table 1
Generating function and dimension for some single tags

Tag	$f(s)$	D
g	$\frac{1}{1-3s}$	$\frac{\log 3}{\log 2}$
gc	$\frac{1}{1-4s+s^2}$	1.89997
gg	$\frac{1+s}{1-3s-3s^2}$	1.92269
gct	$\frac{1}{1-4s+s^3}$	1.97652
gcg	$\frac{1+s^2}{1-4s+s^2-3s^3}$	1.978
ggg	$\frac{1+s+s^2}{1-3s-3s^2-3s^3}$	1.98235
ctag	$\frac{1}{1-4s+s^4}$	1.99429
ggcg	$\frac{1+s^3}{1-4s+s^3-3s^4}$	1.99438
gcgc	$\frac{1+s^2}{1-4s+s^2-4s^3+s^4}$	1.99463
gggg	$\frac{1+s+s^2+s^3}{1-3s-3s^2-3s^3-3s^4}$	1.99572

Proposition 4.1.

$$\frac{\log 3}{\log 2} \leqslant \dim_{\mathrm{H}}(F) \leqslant \dim_{\mathrm{B}}(F) \leqslant \frac{\log(4^{|w_0|}-1)}{\log 2} < 2.$$

Now we denote $\alpha = -\log|s_0|/\log 2$ and $\alpha_K = \log a_K^{1/K}/\log 2$.

For any word $w = w_1w_2\cdots w_K$, we denote $F_{w_1w_2\cdots w_K}$ the corresponding close square in K-frame and denote

$$F_K = \bigcap_{w=w_1w_2\cdots w_K\in L} F_{w_1w_2\cdots w_K},$$

then $F = \lim_{K\to\infty} F_K$.

We first prove $\dim_{\mathrm{H}}(F) = \dim_{\mathrm{B}}(F)$ under a condition using elementary method.

Lemma 4.1. *Suppose* $E \subset \mathbb{R}^2$ *with* $|E| < 1/2$, *let*

$$B_1 = \{w = w_1w_2\cdots w_K \in L:\ |F_{w_1w_2\cdots w_K}| < |E| \leqslant |F_{w_1w_2\cdots w_{K-1}}| \text{ and } E\cap F_{w_1w_2\cdots w_K} \neq \emptyset\},$$

then $\#B_1 \leqslant 2\pi$.

Proof. Note that for each $w = w_1 w_2 \cdots w_K \in B_1$

$$\frac{|E|}{|F_{w_1 w_2 \cdots w_K}|} \leqslant \frac{|F_{w_1 w_2 \cdots w_{K-1}}|}{|F_{w_1 w_2 \cdots w_K}|} = \frac{1}{2},$$

then $|E| \leqslant \frac{1}{2}|F_{w_1 w_2 \cdots w_K}|$. The interiors of $F_{w_1 w_2 \cdots w_K}$ with $w = w_1 w_2 \cdots w_K \in B_1$ are non-overlapping and all lie in a disc with radius $2|E|$, and all $F_{w_1 w_2 \cdots w_K}$ are squares, hence

$$(2|E|)^2 \pi \geqslant \left(\frac{1}{\sqrt{2}} |F_{w_1 w_2 \cdots w_K}| \right)^2 \# B_1 \geqslant \frac{1}{2} (2|E|)^2 \# B_1,$$

hence $\# B_1 \leqslant 2\pi$. □

For any $w = w_1 \cdots w_{|w|}, r \in \Sigma$, we denote $w * r = w_1 \cdots w_{|w|} r$ and define $v_w = v_{w_1} v_{w_2} \cdots v_{w_{|w|}}$, where

$$v_{w_j} = \begin{cases} 2^\alpha/4 & \text{if } \#\{r \in \Sigma : \ w_1 w_2 \cdots w_{j-1} r \in L\} = 4 \\ 2^\alpha/3 & \text{if } \#\{r \in \Sigma : \ w_1 w_2 \cdots w_{j-1} r \in L\} = 3 \end{cases}$$

We assume

$$(C_1) \qquad v_w = v_{w_1} v_{w_2} \cdots v_{w_{|w|}} < M \text{ (a constant)} \quad \text{for any } w \in L.$$

Now we have the following result.

Theorem 4.1. *Under condition* (C_1), *we have*

$$\dim_{\mathrm{H}}(F) = \dim_{\mathrm{B}}(F) = \alpha \quad \textit{and} \quad 0 < \mathscr{H}^\alpha(F) < \infty,$$

where $\mathscr{H}^\alpha(F)$ *is the Hausdorff measure of* F.

Proof. We first prove that

$$\mathscr{H}^\alpha(F) < \infty. \tag{4}$$

Since $\alpha_K \to \alpha$ as $K \to \infty$, for any small $\varepsilon > 0$, there exists an integer $N > 0$ such that for any $K > N$, we have $\alpha > \alpha_K - \varepsilon$. Hence

$$\sum_{w = w_1 w_2 \cdots w_K \in L} |F_{w_1 w_2 \cdots w_K}|^\alpha = a_K \left(\frac{1}{2}\right)^{K\alpha} < a_K \left(\frac{1}{2}\right)^{K(\alpha_K - \varepsilon)} = \left(\frac{1}{2}\right)^{-K\varepsilon} \leqslant \left(\frac{1}{2}\right)^{-(N+1)\varepsilon} < \infty.$$

Hence $\mathscr{H}^\alpha(F) < \infty$.

Now we want to prove $\mathscr{H}^\alpha(F) > 0$. We denote

$$\Sigma^\infty = \{\tau = \tau_1 \tau_2 \cdots : \ |\tau| = \infty \text{ and } \tau_1 \cdots \tau_K \in L \text{ for } K = 1, 2, \ldots\}.$$

For any $\tau = \tau_1 \tau_2 \cdots \in \Sigma^\infty$, we denote $\tau|_K = \tau_1 \tau_2 \cdots \tau_K$, and define a probability measure $\widetilde{\mu}$ on Σ^∞ by

$$\widetilde{\mu}([w]) = \left(\frac{1}{2}\right)^{|w|\alpha} v_w, \quad \text{where } [w] = \{\tau \in \Sigma^\infty : \ \tau|_{|w|} = w\}.$$

We can see

$$\sum_{w*r \in L, r \in \Sigma} \widetilde{\mu}([w * r]) = \sum_{w*r \in L, r \in \Sigma} \left(\frac{1}{2}\right)^{(|w|+1)\alpha} v_{w*r} = \left(\frac{1}{2}\right)^{|w|\alpha} v_w \sum_{w*r \in L, r \in \Sigma} \left(\frac{1}{2}\right)^\alpha v_r = \left(\frac{1}{2}\right)^{|w|\alpha} v_w = \widetilde{\mu}([w]).$$

There exists a natural continuous map f from Σ^∞ to F. Now we transfer $\widetilde{\mu}$ to a probability measure on F, let $\mu = \widetilde{\mu} \circ f^{-1}$. We will show that there is some constant $M_1 > 0$ such that if E is a Borel subset of $\mathbb{R}^2$ with $|E| < 1/2$, then $\mu(E) \leqslant M_1 |E|^\alpha$. Of course, this inequality implies $\mathscr{H}^\alpha(F) \geqslant 1/M_1 > 0$.

Set

$$B_1 = \{w = w_1w_2\cdots w_K \in L:\ |F_{w_1w_2\cdots w_K}| < |E| \leqslant |F_{w_1w_2\cdots w_{K-1}}| \text{ and } E \cap F_{w_1w_2\cdots w_K} \neq \emptyset\}.$$

Then

$$\mu(E) \leqslant \sum_{w\in B_1} \widetilde{\mu}([w]) \leqslant \#B_1|F_{w_1w_2\cdots w_K}|^\alpha v_w \leqslant \#B_1|E|^\alpha v_w \leqslant 2\pi M|E|^\alpha = M_1|E|^\alpha. \qquad \square$$

Theorem 4.2. *If the length of tag* $|w_0| \geqslant 3$ *and for any* $w \in L$, v_w *has the form*

$$v_w = \left(\frac{2^\alpha}{3}\right)\left(\frac{2^\alpha}{4}\right)^{i_1}\left(\frac{2^\alpha}{3}\right)\left(\frac{2^\alpha}{4}\right)^{i_2}\left(\frac{2^\alpha}{3}\right)\cdots$$

or

$$v_w = \left(\frac{2^\alpha}{4}\right)^{i_1}\left(\frac{2^\alpha}{3}\right)\left(\frac{2^\alpha}{4}\right)^{i_2}\left(\frac{2^\alpha}{3}\right)\left(\frac{2^\alpha}{4}\right)^{i_3}\cdots,$$

where i_1, i_2 *and* i_3 *are positive integers, then* $\dim_H(F) = \dim_B(F) = \alpha$ *and* $0 < \mathscr{H}^\alpha(F) < \infty$.

Proof. Since $|w_0| \geqslant 3$, we have $\alpha > \log 12/2\log 2$, hence

$$\left(\frac{2^\alpha}{3}\right)\left(\frac{2^\alpha}{4}\right) > 1.$$

From the other condition, we know that there exists $M_1 = \max\{(2^\alpha/3), 1\}$ such that $v_w \leqslant M_1$ for any $w \in L$. Then from Theorem 4.1, we can obtain our result of this theorem. $\square$

Example. $w_0 = ctg$ or $w_0 = ctag$, the result $\dim_H(F) = \dim_B(F)$ holds.

If we do not have condition (C_1), in the following we still can obtain $\dim_H(F) = \dim_B(F)$.

We define $B_2 = \{u \in \Sigma^* \mid |u| = |w_0|, u \neq w_0\}$. One can know the set B_2 contains $N_1 = 4^{|w_0|} - 1$ elements, hence we can write $B_2 = \{u_1, u_2, \ldots, u_{N_1}\}$. Now we can define an $N_1 \times N_1$ matrix $\mathscr{A}$ by

$$\mathscr{A} = [t_{i,j}]_{i,j\leqslant N_1},$$

where $t_{i,j} = (1/2)^\beta$ if $u_i = r_1x$ and $u_j = xr_2$ with $|x| = |w_0| - 1, r_1, r_2 \in \Sigma$, and $t_{i,j} = 0$ otherwise, and where β satisfies $\Phi(\beta) = 1$ with $\Phi(\beta)$ being the largest nonnegative eigenvalue of $\mathscr{A}$. Then from the results of Ref. [7], we have the following theorem.

Theorem 4.3. *If* $B = \{w_0\}$, *then*

$$\dim_H(F) = \dim_B(F) = \beta \quad \textit{and} \quad 0 < \mathscr{H}^\alpha(F) < \infty.$$

From Theorems 3.1 and 4.1, we have the following result.

Corollary 4.1. *If* $B = \{w_0\}$, *then*

$$\beta = \dim_H(F) = \dim_B(F) = \alpha.$$

Remark. When B contains more than one word, we can also construct a matrix $\mathscr{A}$ similarly, then from the results of Ref. [7], we can obtain the same conclusions of Theorem 4.3 and Corollary 4.1 for this case. From Corollary 4.1, we have two methods to calculate the Hausdorff and Box dimensions of F, i.e., calculate α and β, respectively.

Acknowledgements

The first author would like to express his thanks to Prof. Zhi-Ying Wen for encouragement, and to Dr. Hui Rao and De-Jun Feng for many usful discussions.

References

[1] Deckert et al. G. The complete genome of the hyperthermophilic bacterium. Nature 1998;392:353–58.
[2] Falconer KJ. Fractal geometry: mathematics foundations and applications. New York: Wiley, 1990.
[3] Fraser et al. CM. The minimal gene complement of Mycoplasma genitalium. Science 1995;270:397.
[4] Goulden I, Jackson DM. An inversion theorem for cluster decomposition of sequences with distinguished subsequences. J London Math Soc 1979;20:567–76.
[5] Hao B-l, Lee H-C, Zhang S-y. Fractals related to long DNA sequences and complete genomes. Chaos, Solitons & Fractals 1999, to appear.
[6] Hao B-l, Zheng W-m. Appl symbolic dyn and chaos. Singapore: World Scientific, 1998.
[7] Mauldin RD, Williams SC. Hausdorff dimension in graph directed constructions. Trans Amer Math Soc 1988;309:811–29.
[8] Noonan J, Zeilberger D. The Goulden–Jackson cluster method: extensions, applications and implementations. http://www.math.temple.edu/~zeilberg.
[9] Xie H-m. Grammatical complexity and one-dimensional dynamical systems. Singapore: World Scientific, 1996.