PERGAMON

Chaos, Solitons and Fractals 11 (2000) 825–836

CHAOS
SOLITONS & FRACTALS

www.elsevier.nl/locate/chaos

# Fractals related to long DNA sequences and complete genomes

Bai-lin Hao [a,*,1], H.C. Lee [b], Shu-yu Zhang [c]

[a] *Centre for Theoretical Sciences, P.O. Box 2-131, Hsinchu, Taiwan 300, Taiwan*
[b] *Department of Physics and Centre for Complex Systems, National Central University, Chung-li, Taiwan 320, Taiwan*
[c] *Institute of Physics, Academia Sinica, Beijing 100080, China*

Accepted 3 July 1998

**Abstract**

In visualizing very long DNA sequences, including the complete genomes of several bacteria, yeast and segments of human genes, we encounter fractal-like patterns underlying these biological objects of prominent importance. The method used here to visualize genomes of organisms may well be used as a convenient tool to trace, e.g., evolutionary relatedness of species. We describe the method and explain the origin of the observed fractal-like patterns. © 2000 Elsevier Science Ltd. All rights reserved.

## 1. Introduction

The heredity information of all organisms is encoded in a universal way [2] in long chains of nucleic acids composed of four nucleotides represented by the letters *g* (guanine), *c* (cytosine), *a* (adenine), and *t* (thymine), respectively. In the cell of a *prokaryote*, the huge kingdom that includes all bacteria, there is only one linear or circular DNA sequence in the cell (a *prokaryote* cell does not have a nucleus), whereas in the cell of a *eukaryote*, the kingdom of higher life forms including yeast and human, the DNA sequence is further organized into chromosomes that are enclosed in a cell nucleus.

Since the complete genome of the first free-living organism was sequenced in 1995 [1], the pace of nucleotide sequence data being deposited in public databanks [2] has been growing at an exponential rate. Statistical methods have been fruitfully applied to the analysis of nucleotide sequences long before the first complete sequence was obtained and will continue to be useful for this purpose. Yet much as the statistical approach is well suited to reveal details, it cannot generate a good visual representation of a sequence. In fact a method capable of providing a global visualization of very long – longer than, say, 1 Mb – DNA sequences is lacking. Here we propose a simple "deterministic" method that will do just this. The method is based on counting and coarse-graining the frequency of appearance of strings of a given length. It shows distinctive patterns for different genomes. When applying the method to all the known complete genomes and some long DNA sequences, it reveals fractal-like patterns in the sequences. In what follows we first describe the method, then explain the origin of the observed fractal-related patterns. Possible biological implications of the patterns will be elucidated in separate publications elsewhere.

* Corresponding author. Address: Institute of Theoretical Physics, P.O. Box 2735, Beijing 100080, China. Tel.: +86-10-6254-1807; fax: +86-10-6256-2587.

*E-mail address:* hao@itp.ac.cn (B.-l. Hao).

[1] On leave from The Institute of Theoretical Physics, Academia Sinica, Beijing 100080, China.

[2] There are minor exceptions, e.g., in the genetic code of mammalian mitochondria. However, we will ignore these kinds of biological subtleties in this paper.

0960-0779/00/$ - see front matter 
PII: S0960-0779(98)00182-9

## 2. Graphical representation of counters

We call any string composed of $K$ contiguous 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 frequency of appearance of each kind of $K$-strings in a given DNA sequence $4^K$ counters are needed, which we arrange into a $2^K \times 2^K$ array of sites, each site assigned to a counter, as shown in Fig. 1 for $K = 1$ to 3.

The array may be expressed as a $2^K \times 2^K$ matrix that is the direct product of $K$ copies of the $2 \times 2$ matrix

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

that represents the $K = 1$ square in Fig. 1

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

For convenience in 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 binary subscripts of a general element in $M^{(K)}$,

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

are then given by $I = i_1 i_2, \ldots, i_K$ and $J = j_1 j_2, \ldots, j_K$.

The general idea is that to each of the $2^K \times 2^K$ $K$-strings there is a counter that corresponds to the $(I, J)$ element of the matrix $M^{(K)}$. Now to each counter we assign an integer-valued *index* and a pair of integer-valued "coordinates" $(x, y)$ whose binary value is just $(I, J)$. In the coordinate system used here $x$ runs vertically from top to bottom and $y$ runs horizontally from left to right. To compute these integers we first define a mapping that maps the four letters $g$, $c$, $a$ and $t$ to the base 4 number system:

$$\alpha : \{g, c, a, t\} \mapsto \{0, 1, 2, 3\}.$$

Consider now an input DNA sequence of length $N$,

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

where $s_i \in \{g, c, a, t\}$. We slide a window of width $K$ along the sequence so that $K$-strings come into view one at a time. For a linear or circular sequence – many bacteria have circular DNA – of total length $N$, $N - K + 1$ or $N$ $K$-strings will produced. The *index* for the counter of the first $K$-string $s_1 s_2, \ldots, s_K$ is

$$index = \sum_{i=1}^{K} 4^{K-i} \alpha(s_i),$$

and its coordinates are

$$x = \sum_{i=1}^{K} 2^{K-i} [\alpha(s_i) >> 1],$$

$$y = \sum_{i=1}^{K} 2^{K-1} [\alpha(s_i) \& \mathbb{E}],$$

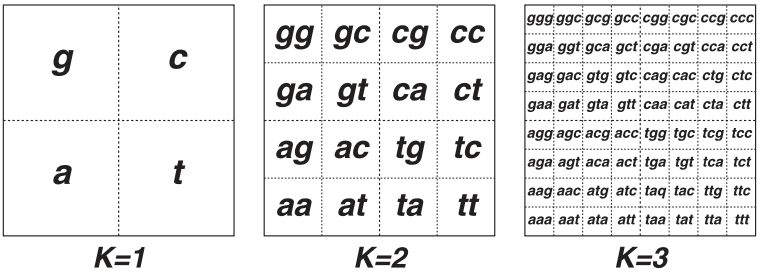


Fig. 1. The arrangement of string counters for $K = 1$ to 3 in squares of the same size.

where $\gg 1$ means the bitwise operation "right-shift by one bit", & is the bitwise operator "logical and" and $\mathbb{E}$ is the binary number 1.

Next we compute the *index* and coordinates $(x, y)$ for the counter of the second $K$-string that appears in the sliding window, $s_2 s_3, \ldots, s_{K+1}$. These are obtained by

$$\begin{aligned} index' &= 4[\mathrm{mod}(index, 4^{K-1})] + \alpha(s_{K+1}), \\ x' &= 2[\mathrm{mod}(x, 2^{K-1})] + [\alpha(s_{K+1}) \& \mathbb{E}], \\ y' &= 2[\mathrm{mod}(y, 2^{K-1})] + [\alpha(s_{K+1}) \gg 1]. \end{aligned}$$

Note that in the above only four integers, *index*, $x$, $y$ and $\alpha(s_{K+1})$ are used as input, as opposed to the $K$ integers needed to compute the *index* and coordinates for the first counter. Furthermore, this implies that after the first $K$-string is read the sliding window can be shrunk to having a width of only one; for subsequent computations only the last letter of the next string need be read. The efficiency of this algorithm, which depends linearly on sequence length $N$ but is independent of $K$, becomes critical when one is dealing with large $K$ and very long sequences. In particular, for a fixed sequence but increasing $K$, even as the amount of information being extracted increases as $4^K$, as does the storage requirement, the computation time will remain essentially constant.

We call the $2^K \times 2^K$ array of counters a $K$-frame and call the $K$-frame of a complete genome a *portrait* of the organism. We use the same frame size for different $K$ and for different DNA sequences. This is a crucial point in order that regularities in portraits are shown off. As the count for a particular type of $K$-string may vary from zero to a big number, we use a crude color code to show the result of the counting. For simplicity one and the same color code is used for all portraits in this paper, although in practice one or another characteristic feature of a portrait may be enhanced by judicious adjustments of color contrast. The color code we use is biased to highlight under-represented strings, in particular, white color indicates those that are absent. In addition, in order that relative rather than absolute abundance of $K$-strings may be compared across different sequences, the string counts are normalized to correspond to a sequence length of one million for coloring.

Fig. 2 shows the portraits of the bacteria *Escherichia coli* (*E. coli*) [3] (a most common and best studied intestine bacterium with 4.639 Mega basepairs (Mbp)), *Archaeoglobus fulgidus* [4] (a bacterial thermophile, 2.178 Mbp), *Synechocystis PCC6803* [5,6] (a photosynthesizing bacterium, 3.573 Mbp), and *Aquifex aeolicus* [7] (a recently sequenced thermophilic bacterium whose phylogenesis has raised new debates in the systematics of micro-organisms [8], 1.551 Mbp) in $K = 8$ frames. One sees that they are distinctive and the upper two share a common set of smaller and smaller red squares the meaning of which will be explained in subsequent sections.

Fig. 3 shows the portraits of *Methanococcus jannaschii* [9] (a thermophilic and methanogenic bacterium), *Helicobacter pylori* [10] (a bacterium causing duodenum and stomach ulcer), *Borrelia burgdorferi* [11] (a bacterium causing the Lyme disease) and *Haemophilus influenzae* [1] (a bacterium related to flu) in $K = 9$ frames. Since there are $4^9 = 262144$ possible different types of 9-strings and the length of the genomes are around 1 Mbp, which is shorter than the lengths of the genomes in Fig. 2, here many more strings are necessarily absent, which results in a preponderance of white squares.

Fig. 4 gives four more portraits of bacteria in $K = 8$ frames. They are *Methanobacterium thermoautotrophicum* [12] (1.751 Mbp, another thermophilic and methanogenic bacterium closely related to *M. jannaschii*), *Mycoplasma genitalium* [13] (0.580 Mbp, the smallest genome of a free-living organism sequenced so far), *Mycoplasma pneumoniae* [14] (0.816 Mbp, a bacterium related to pneumonia) and *Bacillus subtilis* [15] (4.215 Mbp, a bacterium used for producing Natto from soya beans in Japan).

In the last group of color portraits we show some important long nucleotide sequences that are not complete genomes. The upper-left square in Fig. 5 is the portrait of *Rhizobium* sp. *NGR234* [16] (0.536 Mbp, the complete sequence of its so-called *plasmid* that endows the bacterium with nitrogen-fixing ability in their symbiotic life with leguminous plants). The upper-right square shows the human immunoglobulin light chain [17] 1.025 Mbp long that is the concatenation of 36 segments fetched from GenBank. This is a very small part of the 3 billion nucleotides of the human genome of which only 2% have been sequenced and analyzed as of September 1997 [18]. Its fractal-like patterns will be explained in the sequel.

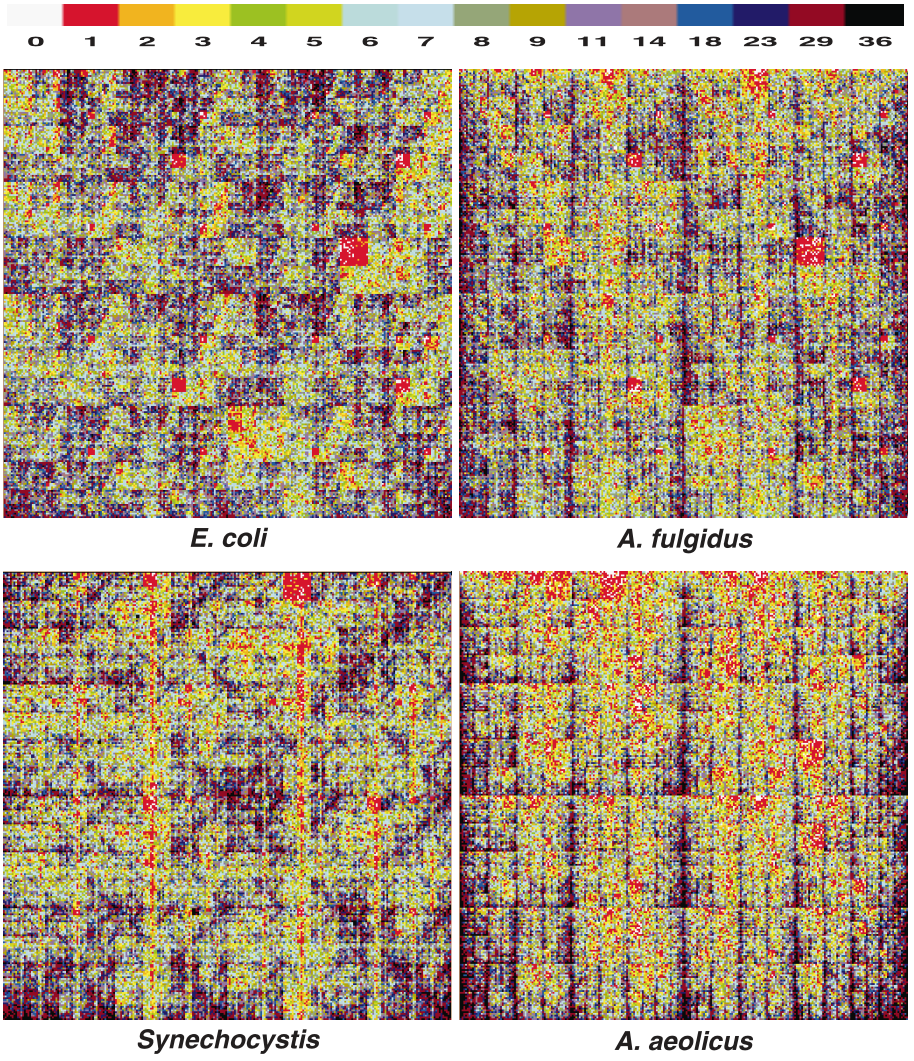


Fig. 2. Portraits of 4 bacteria in $K = 8$ frames.

Of all eukaryotic genomes only the 16 chromosomes of the unicellular baker's yeast (*Saccharomyces cereisiae*) have been completely sequenced and made public, e.g., see Ref. [19]. We show the portrait of its chromosome 15 (1.091 Mbp) in the lower-left of Fig. 5. The genome of a multicell model organism, the soil worm *Caenorhabitidis elegans*, has been 70% sequenced. In the lower-right of Fig. 5 is shown the concatenation of all the known segments of its chromosome I (9.048 Mbp), the longest sequence displayed in this paper. Besides the eye-catching similitude of the two chromosome portraits, there are subtle differences. However, for the time being the number of completely sequenced eukaryote genomes is too small for anything to be said from a comparison of their portraits. We will therefore concentrate on inspecting the patterns in the portraits of the prokaryote genomes.

## 3. Self-similar dark lines in portraits

We start from the horizontal or vertical or diagonal lines seen at different scales in many portraits. Some lines are darker than others, some have sharp color contrast on two sides and so on. These are caused by the difference in the $g$, $c$, $a$ and $t$ contents in the sequences.

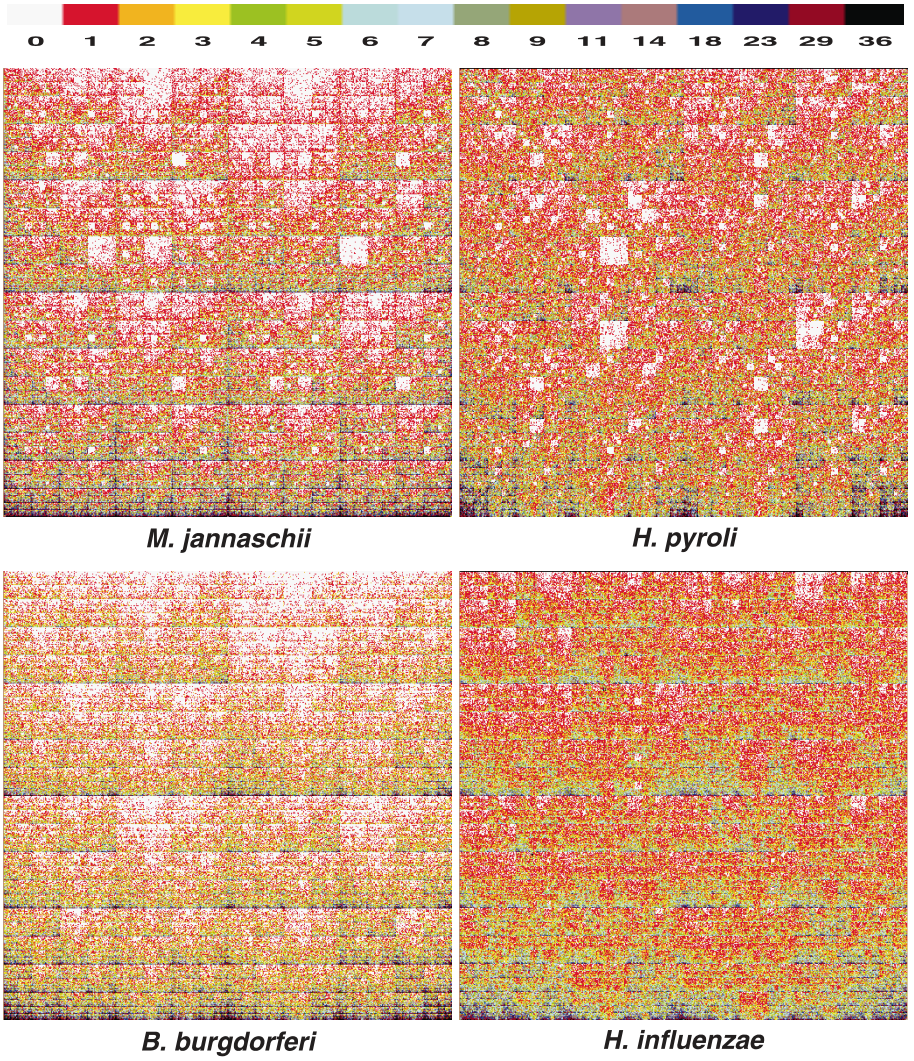


Fig. 3. Portraits of another 4 bacteria in $K = 9$ frames.

Fig. 6 shows some straight lines in an otherwise blank frame (that is not the portrait of anything) whose four corners are labeled by the four letters $g$, $c$, $a$, and $t$. In a, say, $K = 8$ portrait, the counters for the 8-strings *gggggggg*, *cccccccc*, *aaaaaaaa* and *tttttttt* would sit in tiny squares located at these corners (upper-left, upper-right, lower-left, and lower-right), respectively. There are 6 solid lines that connect the four corners of the frame. Let us denote by $\{u, v\}^n$ the set of all $n$-strings composed of the two letters $u$ and $v$, and by $w\{u, v\}^{n-1}$ the set of all $n$-strings composed of a leading letter $w$ concatenated with the $(n-1)$-strings in $\{u, v\}^{n-1}$. Then the horizontal line connecting, say, the corners $a$ and $t$ traverses all the sites of the counters of $\{a, t\}^K$ and none other. Similarly the $g - t$ diagonal traverses all the sites of the counters of $\{g, t\}^K$. Because the frame is discretely partitioned the number of counters along a side line and along a diagonal are the same, both are $2^K$. The frame has a hierarchical structure. Take, for example, the dash line $\alpha - \beta$ parallel to the $a - c$ diagonal. It traverses the sites of the counters of all $g\{a, c\}^{K-1}$ strings and none other. Similarly, the $\delta - \gamma$ dash line traverses the sites for the counters of all $t\{a, c\}^{K-1}$ strings. In the frame the counter sites have a simple symmetry. For example, reflection with respect to the $a - c$ line interchanges $g$ and $t$ in the designation of each counter and leaves $c$ and $a$ unchanged.

Let us have a closer look at the counters located just above and below the $a - c$ diagonal. The upper line traverses the $2^K - 1$ counters for:

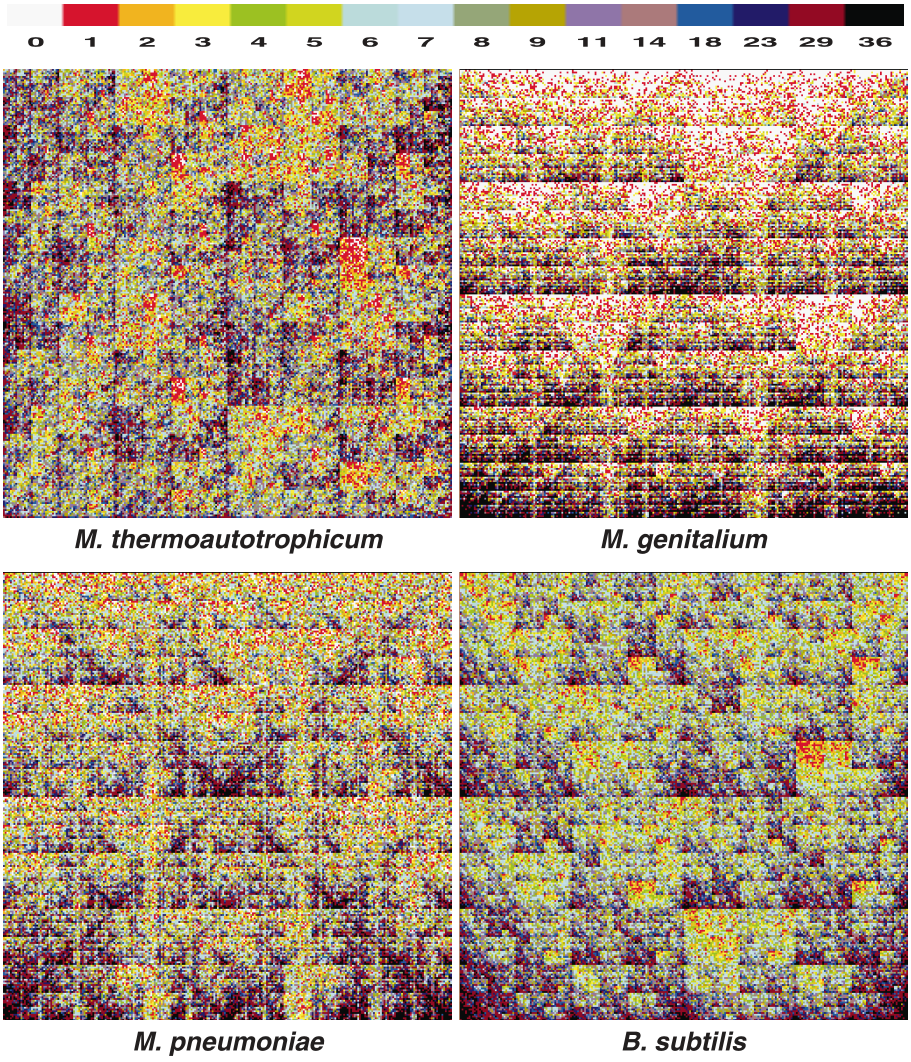


Fig. 4. Portraits of another 4 bacteria in $K = 8$ frames. Note the common crossing patterns in the two Mycoplasma.

1. The one $(K-1)$-string $gt\cdots t$;
2. $2^{K-2}$ strings of $\{a,c\}^{K-1}g$ type;
3. $2^{K-3}$ strings of $\{a,c\}^{K-2}gt$ type;
4. $2^{K-4}$ strings of $\{a,c\}^{K-3}gtt$ type and so on.

The lower line traverses the counters for the same set of $2^K-1$ strings as above except with the letters $g$ and $t$ interchanged.

Analogous regularities hold with respect to the $g-t$, $g-a$ and $c-t$ lines, but the situation for horizontal lines is different. Take, for example, the half-line just above the $\alpha - O$ dash line. It traverses the $2^{K-1}$ counters for strings of $g\{g,c\}^{K-1}$ type, but the half-line just below the $\alpha - O$ dash line traverses the $2^{K-1}$ counters for strings of $a\{a,t\}^{K-1}$ type. Similarly, crossing the $O-\gamma$ dash line from below replaces counters for $t\{a,t\}^{K-1}$ type strings by counters for $c\{g,c\}^{K-1}$ type strings. In most bacterial genomes, the contents of $g$ and $c$ are equal to within a few percent, as are the contents of $a$ and $t$. However the $g+c$ to $a+t$ ratio may vary widely. Consider a sequence where this ratio is significantly different from one. Since going from the line below $\alpha-\gamma$ to the line above replaces all $a+t$ labels in the counters by $g+c$ labels, in this case one would expect a significant difference in the colors of the two lines. This analysis applied to any vertical movement in a portrait. Indeed, in the portraits of all the $(g+c)$-poor sequences a sharp contrast is seen

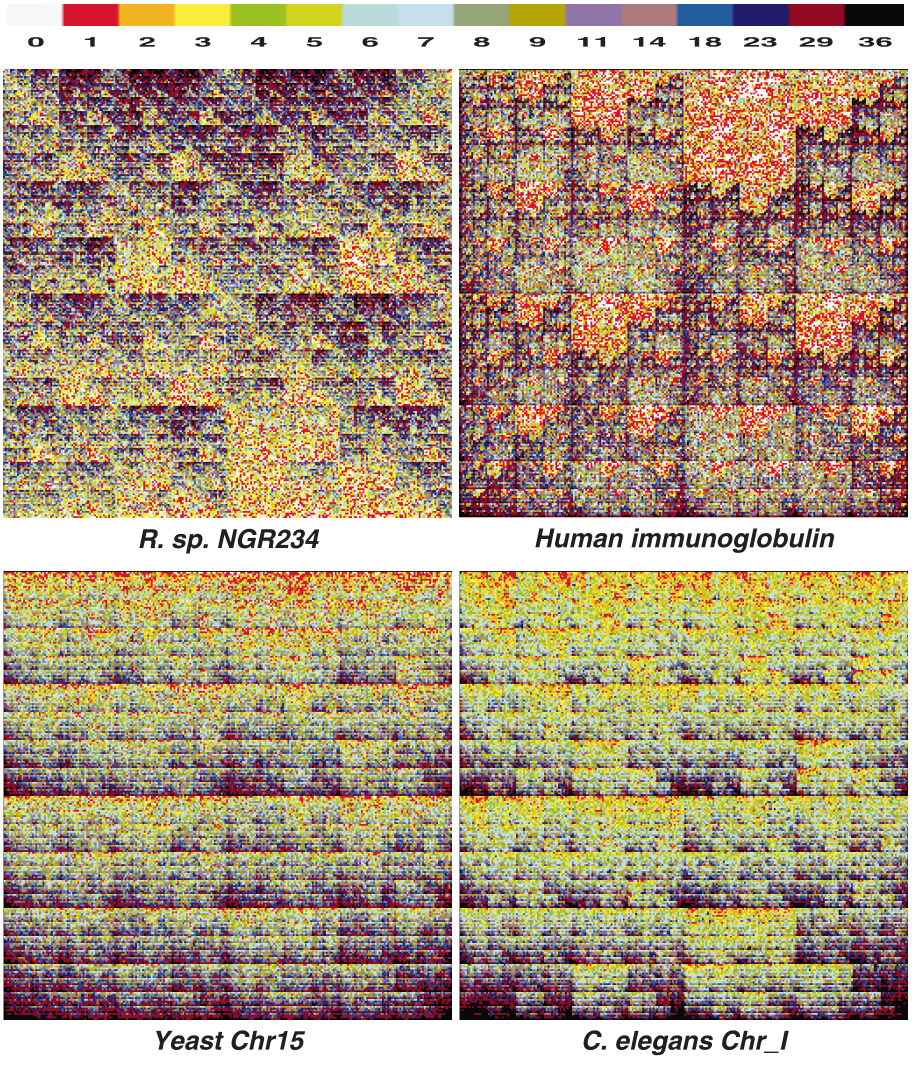


Fig. 5. Portraits of some long DNA sequences in $K = 8$ frames. For explanation see the text.

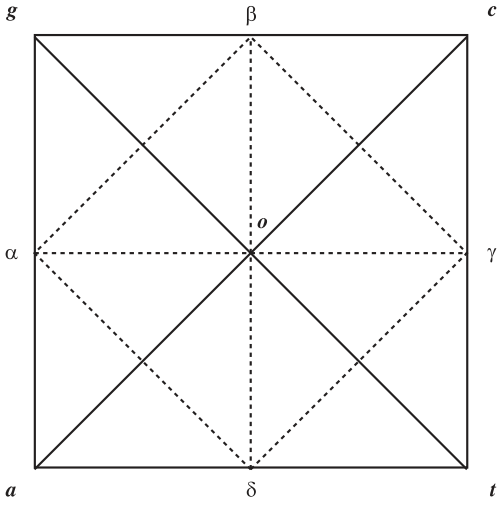


Fig. 6. Some straight lines in a portrait.

upon crossing the $\alpha - \gamma$ line or their parallels – darker above and brighter below. In the portrait of the only $(g + c)$-rich genome (*R. sp. NGR234*) considered in this paper, the color contrast goes in the opposite sense – brighter above and darker below.

## 4. Fractal-like arrays in portraits

Now we turn to the fractal-like arrays of red or white squares of various sizes seen in many portraits of bacterial genomes. We show that this is caused by under-representation of those strings that contain a certain short palindromic sequence of nucleotides. First, a few words on the biological meaning of such palindromes. In the late 1960s it was first discovered in *E. coli* and then in many other bacteria that they are capable of producing enzymes that cut "foreign" DNA double helix at well-defined sites. Designed by Nature as the part of immune system in bacteria, these "cutters" soon became an indispensable instrument in the toolkit of genetic engineers. At present about 3000 such "restriction enzymes", many of which are commercially available, and about 300 "recognition sites" are known [20]. The most common type of recognition sites are palindromes of 4–8 nucleotides. The term palindrome in this context means that the short string remains the same when read in both directions, provided the Watson-Crick conjugation of letters $g \leftrightarrow c$ and $a \leftrightarrow t$ is applied when the reading direction is reversed. Thus *cctagg* and *ctag* are both

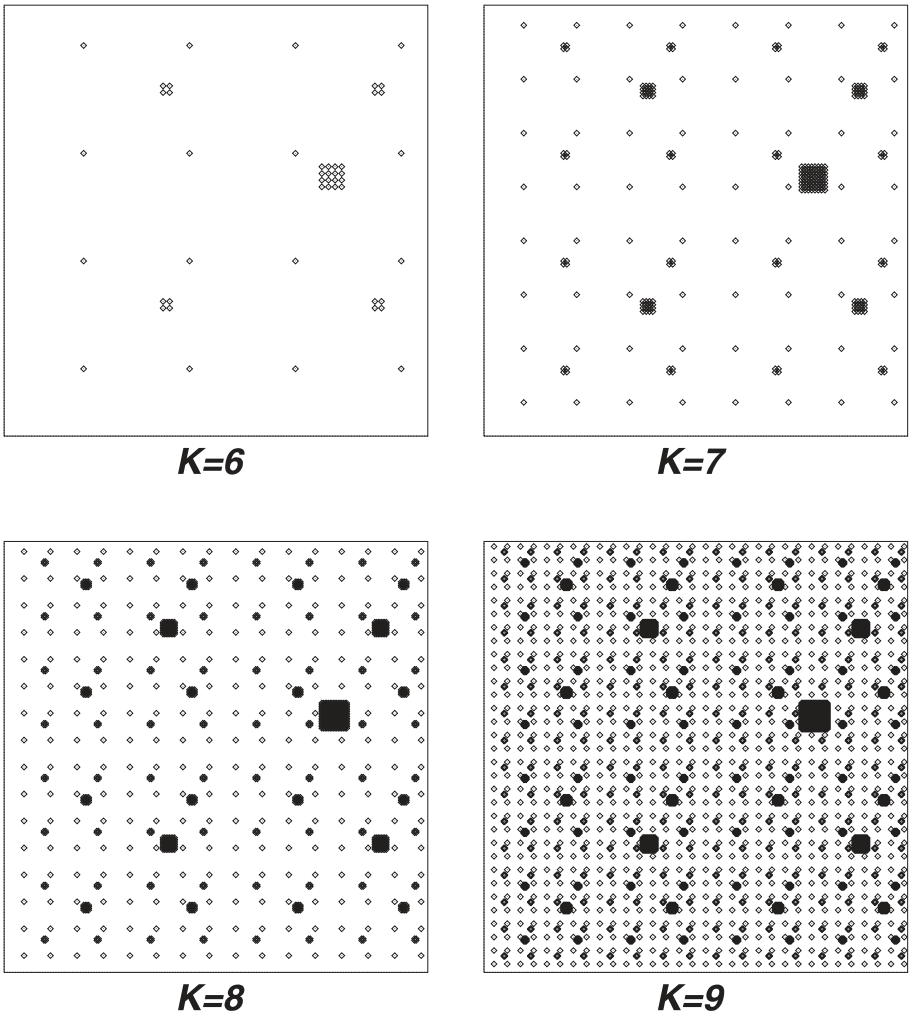


Fig. 7. The location of *ctag-tagged strings in* $K = 6$ *to* 9 *frames*.

palindromes. Among the 256 possible tetranucleotides, 16 are palindromes and all are recognition sites of some restriction enzymes.

Let us call a string containing the tetranucleotide *ctag* a *ctag*-tagged string. In Fig. 7, small diamonds are placed at the counter positions for all *ctag*-tagged $K$-strings in the four otherwise blank portraits of $K = 6$ to 9 frames. The figures exhibit a self-similar structure with more details appearing with increasing $K$. One may call such a figure the "portrait" of a recognition site – in this case *ctag* – in a $K$-frame and note that the predominant structure of such a portrait remains invariant with respect to increasing frame size $K$. Such portraits may be constructed for any short strings of nucleotides. Fig. 8 shows the portraits for four other recognition sites. It is seen that each individual recognition site has its own distinctive pattern. A comparison of portraits for tetranucleotides with the portraits in Figs. 2–5 reveals that 9 of the 13 bacterial genomes avoid some palindromic strings, all of which are known recognition sites of some restriction enzymes [20]. The biological implication of this fact will be discussed elsewhere. We only note that randomization of a DNA sequences with its composition unchanged erases all regular patterns in the portraits except some almost unperciptible contrast caused by different $g + c$ to $a + t$ ratios.

Not all fractal-like patterns are caused by under-representation of recognition sites. In the portrait of *B. subtilis* or *M. thermoautotrophicum* shown in Fig. 4, the area of the largest light-colored patterns is bigger than those in *E. coli*, *A. fulgidus* and *M. jannaschii* that signal the under-representation of *ctag*. In fact, the former are caused by the under-representation of the trinucleotide *cta* which in turn is contained in *ctag*.

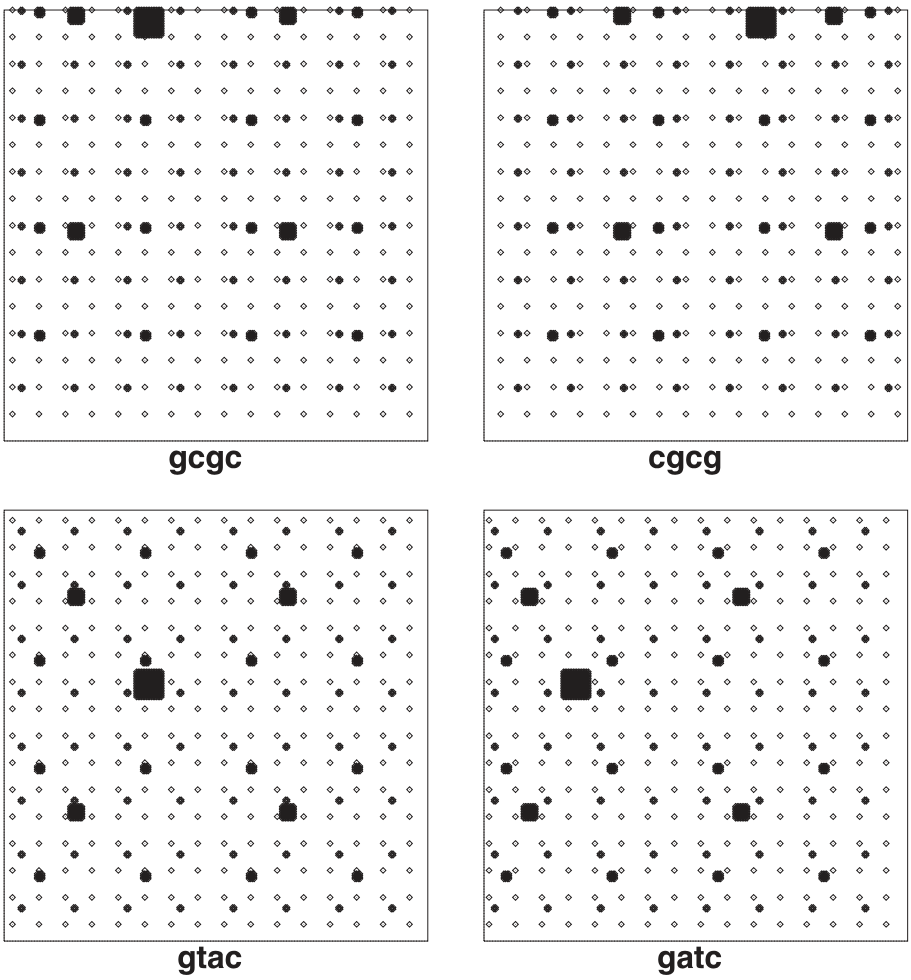


Fig. 8. The portraits of four palindromic tetranucleotides in $K = 8$ frames.

Fig. 9 shows the portraits of some trinucleotides. They may be used to interpret patterns seen in some other portraits of genomes.

Perhaps the most prominent fractal-like patterns are those seen in the one-megabase sequence of human immunoglobulin light chain (upper-right square in Fig. 5). They are caused by the under-representation of *cg*-tagged strings, as is evident from the portrait of the dinucleotide *cg* in the $K = 8$ frame shown in Fig. 10.

## 5. Summary and discussion

We have described a simple method to visualize long DNA sequences by converting a sequence to a portrait and applied the method to a number of sequences including all the published complete bacterial genomes. Prominent fractal-like patterns are seen in these portraits and their origins are explained. Due to its coarse-grained, distinctive, and species-specific nature, this method may well be a useful means for a quick qualitative view of a newly determined long DNA sequence, or even for the detection of evolutionary relatedness of species. A detailed discussion of such biological implications goes beyond the scope of this paper and will be the subject of another publication.

In view of a recent discussion of whether Nature is fractal [21] or not, we would like to make the following remark. Contrary to genuine fractals in mathematical constructions that do possess scaling invariance over infinitely many orders of magnitude, any scaling behaviour observed in Nature must be

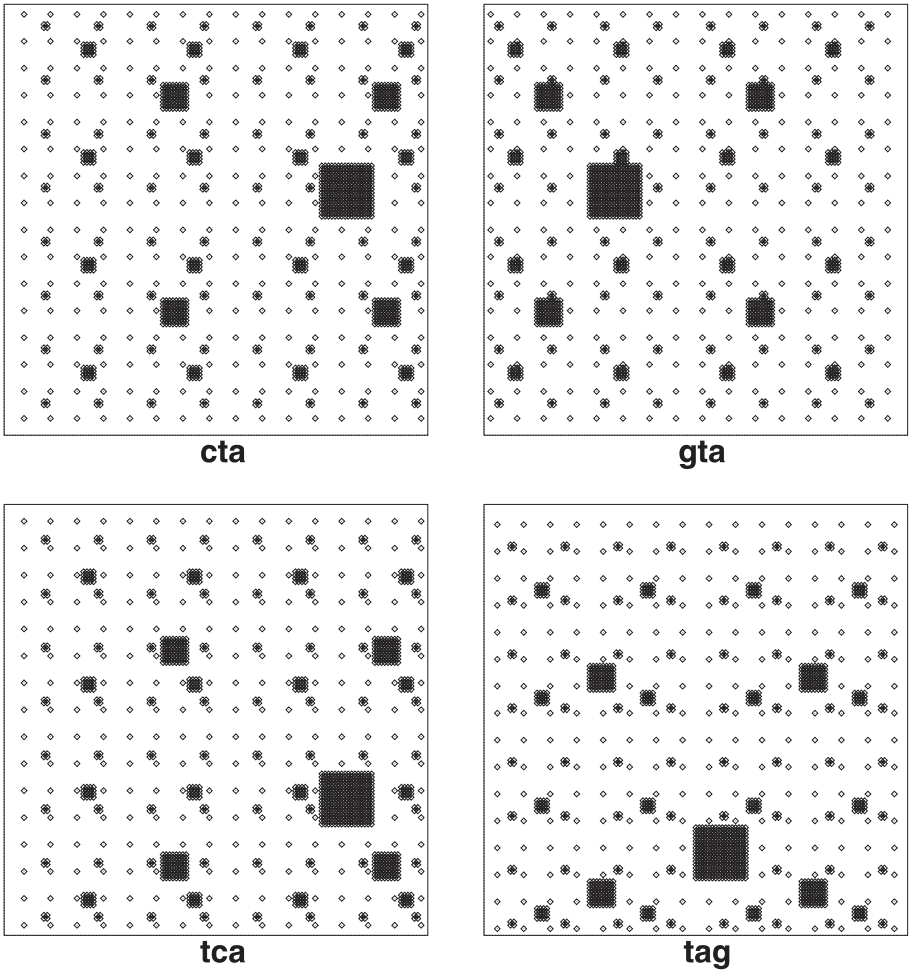


Fig. 9. The portraits of four trinucleotide in $K = 8$ frames.

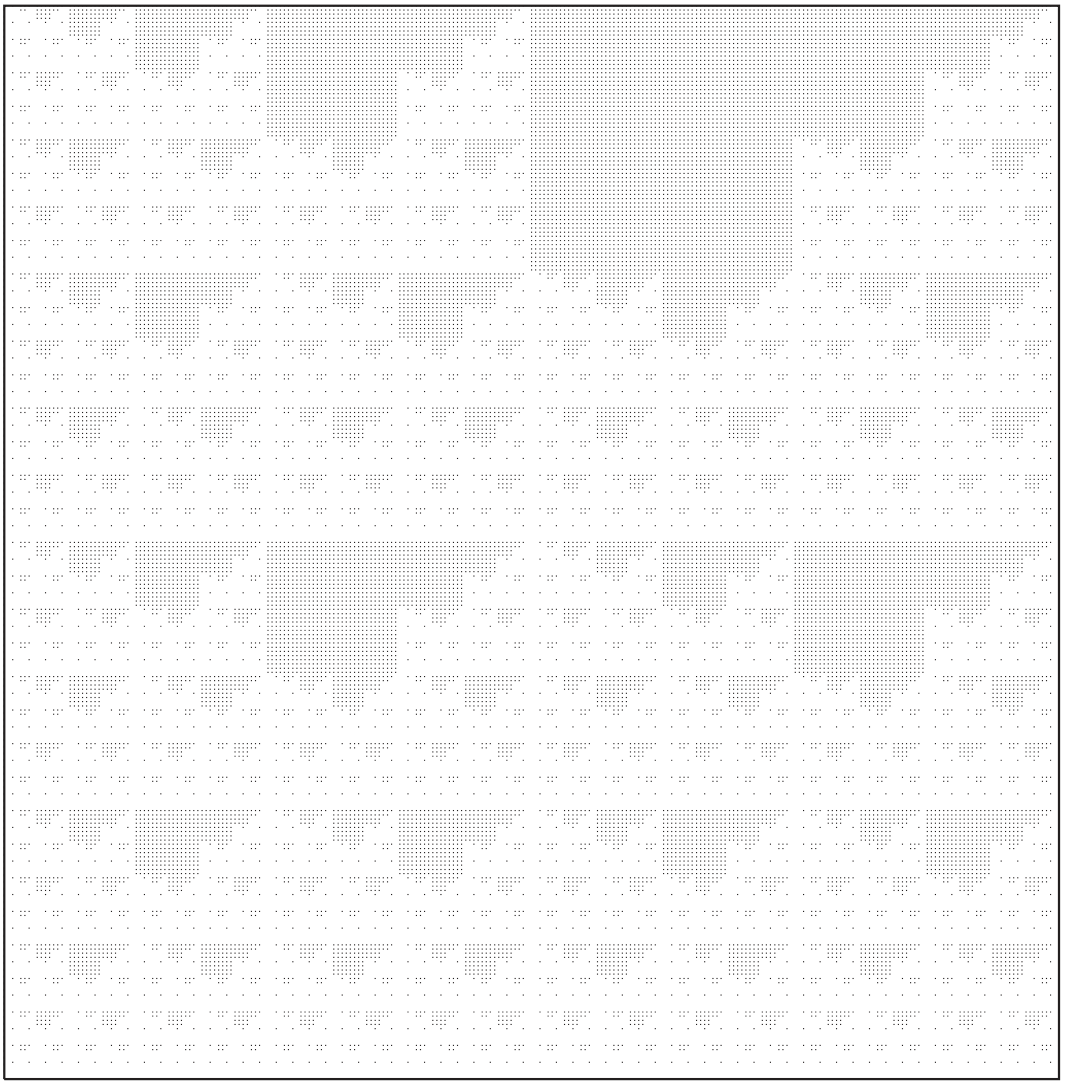

Fig. 10. The portrait of the dinucleotide *cg* in the $K = 8$ frame, to be compared with the characteristic patterns seen in the portrait of the human immunoglobulin sequence.

limited in scale both from above and below. While agreeing with the statement that "exponents" extracted from an "invariance" over 0.5 to 2 orders of magnitude do not make much sense [21], we point out for the fractal-like patterns seen in this paper that there are no "exponents". What we have done is to display real data from Nature in a framework that do have an underlying fractal structure which is well-defined in the $K \to \infty$ limit.

Another remark concerns the so-called chaos game representation of DNA sequences [22]. The portraits shown in this paper differ essentially from the chaos game representation, in which the number of points plotted always equals to the number of nucleotides in the 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.

To conclude we return to the comparison of "deterministic" and statistical approaches to DNA sequences mentioned briefly in the Introduction. Any analysis based on strings "tagged" with a specific "word" already goes beyond statistics. Our concentration on under-represented strings has a deeper reason. It is now known that a simple formal language such as that defined by the symbolic dynamics of unimodal maps may be completely specified by its set of "forbidden words" or "distinct excluded blocks", e.g., see Ref. [23]. Apparently, the genetic language of DNA, if it exists, can neither be defined by a closed grammar

nor by a definite set of forbidden words. However, the progress of molecular biology has been supplying and will continue to supply us with an ever growing collection of allowed words such as the recognition sites of restriction enzymes or binding sites of transcription factors, and studies such as the present work may provide more knowledge on the set of forbidden or avoided "words". The combined use of these two collections may help us to discover more truth from the DNA sequences.

## 6. Note added in proof

The fractal dimensions of the template of various tagged-strings in the infinite $K$ limit may be calculated precisely. See B.-L. Hao, H.-M. Xie, Z.-G. Yu, and G.-Y. Chen, "Avoided strings in bacterial complete genomes and a related combinatorial problem", *Ann. Combin.* (to appear), and Z.-G. Yu, B.-L. Hao, H.-M. Xie, and G.-Y. Chen, "Dimension of fractals related to language defined by tagged-strings in complete genomes", *Chaos, Solitons and Fractals* (accepted for publication).

## Acknowledgements

This work was supported by grants from the National Science Council NCS87-2119-M-007-004 to B.L.H. and NSC87-2112-M-008-002 to H.C.L.

## References

[1] R.D. Fleischmann et al., Whole genome random sequencing and assembly of Hameophilus influenzae Rd., Sci. 269 (1995) 496.
[2] A main source is the International Nucleotide Sequence Data Bank of DDBJ, EMBL, and NCBI, see, e.g., http://www.genome.ad.jp/ or www.ebi.ac.uk or ncbi.nlm.nih.gov. We fetched all the sequence mentioned in this paper from GenBank/NCBI.
[3] F.R. Blattner et al., The complete genome sequence of *Escherichia coli* K-12, Sci. 277 (1997) 1453.
[4] H.P. Klenk et al., The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus, Nature 390 (1997) 364.
[5] T. Kaneko et al., Sequence analysis of the genome of the unicellular Cyanobacterium synechocystis sp. strain PCC6803 I, DNA Res. 2 (1995) 153.
[6] T. Kaneko et al., Sequence analysis of the genome of the unicellular Cyanobacterium synechocystis sp. strain PCC6803 II, DNA Res. 3 (1996) 109.
[7] G. Deckert et al., The complete genome of the hyperthermophilic bacterium Aquifex aeolicus, Nature 392 (1998) 353.
[8] E. Pennisi, Genome data shake tree of life, Sci. 280 (1998) 672.
[9] C.J. Bult et al., Complete genome sequence of the methanogenic archaeon Methanococcus jannaschii, Sci. 273 (1996) 1058.
[10] J.F. Tomb et al., The complete genome sequence of the gastric pathogen Helicobacter pylori, Nature 388 (1997) 539.
[11] C.M. Fraser et al., Genome sequence of a Lyme disease spirochaete Borrelia burgdorferi, Nature 390 (1997) 580.
[12] D.R. Smith et al., Complete genome sequence of Methanobacterium thermoautotrophicum Δ H: functional analysis and comparative genomics, J. Bacteriol 179 (1997) 7135.
[13] C.M. Fraser et al., The minimal gene complement of Mycoplasma genitalium, Sci. 270 (1995) 397.
[14] R. Himmelreich et al., Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae, Nucl. Acids Res. 24 (1996) 4420.
[15] F. Kunt et al., The complete genome sequence of the Gram-positive bacterium Bacillus subtilis, Nature 390 (1997) 249.
[16] C. Freiberg et al., Molecular basis of symbiosis between Rhizobium and legumes, Nature 387 (1997) 394.
[17] K. Kawasaki et al., One-megabase sequence analysis of the human immunoglobulin $\lambda$ gene locus, Gene Res. 7 (1997) 250.
[18] L. Roman, G. Mahairas, L. Hood, Sequencing the human genome, Sci. 278 (1997) 605.
[19] A. Goffeau et al., Life with 6000 genes, Science 274 (1996) 546.
[20] R.J. Roberts, D. Macelis, REBASE – restriction enzymes and methylases, Nucl. Acids Res 26 (1998) 338.
[21] D. Avnir et al., Is the geometry of Nature fractal, Science 279 (1998) 39.
[22] H.J. Jeffrey, Chaos game representation of gene structure, Nucl. Acids Res. 18 (1990) 2163.
[23] H.M. Xie, Grammatical Complexity and One-Dimensional Dynamical Systems, World Scientific, Singapore, 1996.