Zuo-Bing Wu

Laboratory of Nonlinear Mechanics, Institute of Mechanics, Academia Sinica, Beijing, China
Institute of Theoretical Physics, Academia Sinica, Beijing, China

Metric representation of DNA sequences

A metric representation of DNA sequences is borrowed from symbolic dynamics. In view of this method, the pattern seen in the chaos game representation of DNA sequences is explained as the suppression of certain nucleotide strings in the DNA sequences. Frequencies of short nucleotide strings and suppression of the shortest ones in the DNA sequences can be determined by using the metric representation.

Keywords: DNA sequences / Metric representation / Suppression of short nucleotide strings

EL 3929

The DNA sequence of an organism determines heredity and variation. Understanding the one-dimensional symbolic sequence composed of the four letters “A”, “C”, “G” and “T” (or “U”) has been a main theme in bioinformatics. Chaos game representation (CGR) [1, 2], which generates a two-dimensional square from one-dimensional sequence, provides a technique to visualize the composition of the DNA sequences. The pattern formation in CGR has been explained in terms of the mononucleotide, dinucleotide, and trinucleotide frequencies [3]. Shortsequence representation partially quantifies nonuniform distribution of data points [4–6]. A generalization of CGR to analyze amino acid sequences has been given previously [7]. Recently, a visualization scheme of the string composition of long DNA sequences and complete genomes has been proposed [8]. Although the algorithm of string counting is much simpler, the resulting figures look much like what seen in CGR.

In this paper, we propose a metric representation (MR) of DNA sequences inspired by symbolic dynamics [9] and apply it to explore the suppression of short nucleotide strings in the DNA sequences. In what follows we describe the MR without reference to symbolic dynamics. The proposed MR provides a mathematical framework to explain the patterns seen both in CGR and [8].

For our purpose, a DNA molecule may be understood as a one-dimensional symbolic sequence, s_1, $s_2 \cdots s_i \cdots s_N$ (s_i ϵ {A, C, G, T}). Consider a subsequence $\Sigma = s_1 s_2 \cdots s_i \cdots s_m$ ($1 \leq m \leq N$). In order to have a two-dimensional representation, we first generate two sequences from Σ. The first sequence is obtained from Σ by replacing C by A, and G by T. Let us call it the Σ_α sequence. The second, the Σ_β sequence, is obtained from Σ by replacing T by A, and G by C. Note that one can always recover the original sequence Σ from Σ_α and Σ_β by looking at the corresponding letters and using the rules embodied in the following table:

Σ_α	A	T	A	T
Σ_β	C	A	A	C
Σ	C	T	A	G

Next, we introduce an ordering for the two sets of sequences. Suppose two Σ_α sequences have the same string $s_{i+1} \cdots s_m$ but differ at s_i, we define that the sequence with s_i = T is larger than that with s_i = A. Similarly, if two Σ_β sequences share a common string $s_{j+1} \cdots s_m$ but differ at s_j, we define that the one with s_j=C is lager than that with s_j = A.

In short, in order to establish an MR for a one-dimensional DNA sequence in a two-dimensional plane (α, β), we define ordering rules

$$A = C < G = T \tag{1}$$

in the α-direction and

$$A = T < C = G \qquad (2)\text{resolution}$$

in the β-direction. According to the ordering rules, we put Σ into correspondence with a point in the 2-D (α, β) plane, where the two real numbers $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ are calculated from the Σ_α and Σ_β sequences as follows. For $\Sigma_\alpha = s_1 s_2 \cdots s_i \cdots s_m$, we define

$$\alpha = 2\sum_{j=1}^{m} \mu_{m-j+1} 3^{-j} + 3^{-m} = 2\sum_{i=1}^{m} \mu_i 3^{-(m-i+1)} + 3^{-m} \tag{3}$$

where $\mu_i = 0$ if s_i = A or $\mu_i = 1$ if s_i = T. Similarly, the number β is defined from Σ_β as

Correspondence: Dr. Zuo-Bing Wu, Laboratory of Nonlinear Mechanics, Institute of Mechanics, Academia Sinica, Beijing 100080, China
E-mail: wuzb@lnm.imech.ac.nc

Abbreviations: CGR, chaos game representation; **MR**, metric representation

0173-0835/00/1212-2321 $17.50+.50/0

$$\beta = 2\sum_{j=1}^{m} \nu_{m-j+1}3^{-j} + 3^{-m} = 2\sum_{i=1}^{m} \nu_i 3^{-(m-i+1)} + 3^{-m} \qquad (4)$$

where $\nu_i = 0$ if s_i = A or $\nu_i = 1$ if s_i = C. As essentially we deal with three letters in Eqs. (1) and (2), the MR of Eqs. (3) or (4) resemble the number system in base 3. In this MR, for example, the four sequences A^∞, C^∞, G^∞ and T^∞ correspond to the four vertices {(0,0), (0,1), (1,1), (1,0)} of a unit square in the (α, β) plane. A conceptual sketch of the MR is shown in Fig. 1.

Moreover, we introduce a left shift operator *S* on the sequence $s_1s_2\cdots s_i\cdots s_m$, defined as $S(s_1s_2\cdots s_i\cdots s_m) = s_1s_2\cdots s_i\cdots s_ms_{m+1}$, where *S* denotes any of the letters, A, C, G or T, and s_{m+1} is the same as *S*. Under the action of the left shift operators, we have

$$\{A(A^\infty, C^\infty, G^\infty, T^\infty)\} = \{A^\infty, C^\infty A, G^\infty A, T^\infty A\} \Rightarrow \left\{\left(0,0\right), \left(0,\tfrac{1}{3}\right), \left(\tfrac{1}{3},\tfrac{1}{3}\right), \left(\tfrac{1}{3},0\right)\right\}$$

$$\{C(A^\infty, C^\infty, G^\infty, T^\infty)\} = \{A^\infty C, C^\infty, G^\infty C, T^\infty C\} \Rightarrow \left\{\left(0,\tfrac{2}{3}\right), \left(0,1\right), \left(\tfrac{1}{3},1\right), \left(\tfrac{1}{3},\tfrac{2}{3}\right)\right\}$$

$$\{G(A^\infty, C^\infty, G^\infty, T^\infty)\} = \{A^\infty G, C^\infty G, G^\infty, T^\infty G\} \Rightarrow \left\{\left(\tfrac{2}{3},\tfrac{2}{3}\right), \left(\tfrac{2}{3},1\right), \left(1,1\right), \left(1,\tfrac{2}{3}\right)\right\}$$

and

$$\{T(A^\infty, C^\infty, G^\infty, T^\infty)\} = \{A^\infty T, C^\infty T, G^\infty T, T^\infty\} \Rightarrow \left\{\left(\tfrac{2}{3},0\right), \left(\tfrac{2}{3},\tfrac{1}{3}\right), \left(1,\tfrac{1}{3}\right), \left(1,0\right)\right\}$$

Each left shift operator shrinks the area of the square to $1/3^2$ and forms a fundamental zone. The four fundamental zones can be named after the letters, A, C, G, T, respectively. In each of the fundamental zones, *e.g.*, in zone A, we reapply the left shift operators and get

$$\{A(A^\infty, C^\infty A, G^\infty A, T^\infty A)\} = \{A^\infty, C^\infty A^2, G^\infty A^2, T^\infty A^2\} \Rightarrow \left\{\left(0,0\right), \left(0,\tfrac{1}{9}\right), \left(\tfrac{1}{9},\tfrac{1}{9}\right), \left(\tfrac{1}{9},0\right)\right\}$$

$$\{C(A^\infty, C^\infty A, G^\infty A, T^\infty A)\} = \{A^\infty C, C^\infty AC, G^\infty AC, T^\infty AC\} \Rightarrow \left\{\left(0,\tfrac{2}{3}\right), \left(0,\tfrac{7}{9}\right), \left(\tfrac{1}{9},\tfrac{7}{9}\right), \left(\tfrac{1}{9},\tfrac{2}{3}\right)\right\}$$

$$\{G(A^\infty, C^\infty A, G^\infty A, T^\infty A)\} = \{A^\infty G, C^\infty AG, G^\infty AG, T^\infty AG\} \Rightarrow \left\{\left(\tfrac{2}{3},\tfrac{2}{3}\right), \left(\tfrac{2}{3},\tfrac{7}{9}\right), \left(\tfrac{7}{9},\tfrac{7}{9}\right), \left(\tfrac{7}{9},\tfrac{2}{3}\right)\right\}$$

$$\{T(A^\infty, C^\infty A, G^\infty A, T^\infty A)\} = \{A^\infty T, C^\infty AT, G^\infty AT, T^\infty AT\} \Rightarrow \left\{\left(\tfrac{2}{3},0\right), \left(\tfrac{2}{3},\tfrac{1}{9}\right), \left(\tfrac{7}{9},\tfrac{1}{9}\right), \left(\tfrac{7}{9},0\right)\right\}$$

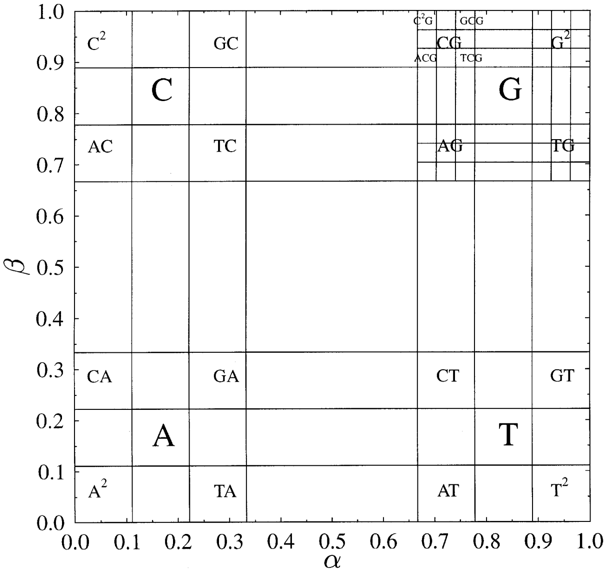

Figure 1. Conceptual sketch of metric representation. The left shift operators, A, C, G, T shrink the area of every square to $1/3^2$. Each action of a right shift operator divides the orginal square into four smaller ones.

The left shift operators transfer zone A to the four different fundamental zones, as well as shrink their area again by a factor of $1/3^2$. For the left shift operator A, the shrunk zone is still kept in the fundamental zone A and may be encoded by A^2. For the left shift operator $S \in$ {C, G, T}, the shrunk zones are transferred to others, which bear the same name as *S* itself. The shrunk zone is encoded by A*S*. Thus, under the action of the four left shift operators, the four fundamental zones in Fig. 1 shrink to 16 zones, which are redistributed in the four fundamental zones and encoded by all the dinucleotides. Furthermore, we can continue the contracting process and obtain 64 zones encoded by all the trinucleotides, 256 zones encoded by all the tetranucleotides, and so on. According to Eqs. (3) and (4), the one-dimensional symbolic sequence $s_1s_2\cdots s_i\cdots s_N$ is subdivided into four kinds of subsequences, corresponding to points located in four fundamental zones. All subsequences with the same ending *k*-nucleotide string correspond to points in the zone encoded by that *k*-nucleotide string. Their distribution in the zone conforms to ordering rules (1) and (2).

In order to keep the shrunk zones within the same fundamental zone, we define a right shift operator $M \in$ {A, C, G, T} for a *k*-nucleotide string $s_k\cdots s_2s_1$ as $M\ (s_k\cdots s_2s_1) = s_{k+1}s_k\cdots s_2s_1$, where *M* and s_{k+1} are the same letter. The contracting process of lthe four fundamental zones shown in Fig. 1 can be described by the right shift operators as follows. Under the action of a right shift operator, M, each one of the four fundamental zones shrinks to four smaller

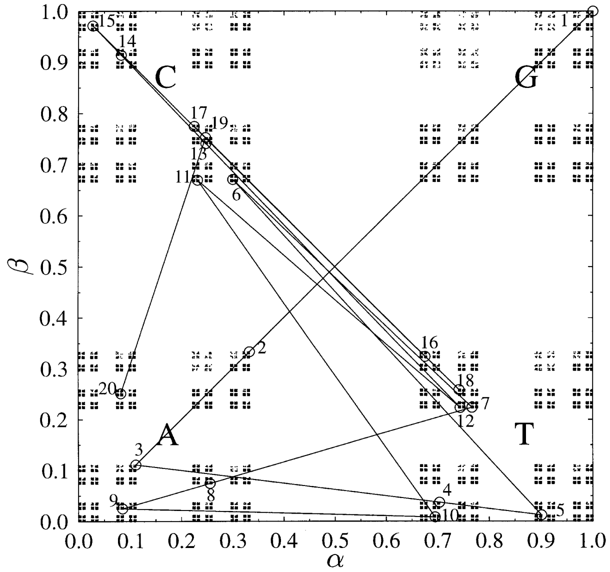

Figure 2. Metric representation of HUMHBB. The circles display walks of the first 20 bases in the representation. The points corresponding to the subsequences ···TCTA and ···A^2TC are marked by 8 and 11, respectively.

zones, which are still kept in the same fundamental zone. For example, the fundamental zone G shrinks to four zones encoded by AG, CG, G^2 and TG. When the action of a right shift operator is repeated, the zone CG shrinks to four zones encoded by ACG, C^2G, GCG and TCG. Furthermore, we obtain 64 zones encoded by all the tetranucleotides in the fundamental zone G and so on. In the process, we obtain tandemly reiterated sequences (regardless of shifts), *e.g.*, 6 (C_4^2) repeated dinucleotide sequences $(CA)^\infty$, $(GA)^\infty$, $(TA)^\infty$, $(GC)^\infty$, $(TC)^\infty$, $(TG)^\infty$. Their MRs {(0,$^1/_4$), ($^1/_4$, $^1/_4$), ($^1/_4$,0), ($^1/_4$,1), ($^1/_4$, $^3/_4$), (1,$^3/_4$)} distribute inside the corresponding zones {CA, GA, TA, GC, TC, TG}. In the same way, we also obtain 20 ($C_4^1C_3^1 + C_4^3/3$) repeated trinucleotide sequences and 48 ($C_4^1C_3^1 + C_4^2 + C_4^1P_3^2 + P_4^4/4$) repeated tetranucleotide sequences. Their MRs also distribute inside the corresponding zones.

The MR of the entire 73308 bases of HUMHBB (human β-globin region, chromosome 11) is shown in Fig. 2, where the points corresponding to the first 20 bases GAATTCTAATCTCCCTCTCA are drawn. We can check the ordering rules in terms of the distribution of the 20 points with their subsequences. For example, for the subsequences···TCTA and ···A^2TC, the former ordering is larger than the latter in the α-direction, and in the β-direction the ordering is reverse. Moreover, we have (α, β) [···TCTA] = (0.255601, 0.074531) and (α, β) [···A^2TC] =(0.231689, 0.669427). The ordering for the two subsequences is therefore identical with the relative position of corresponding two points (8 and 11) in Fig. 2.

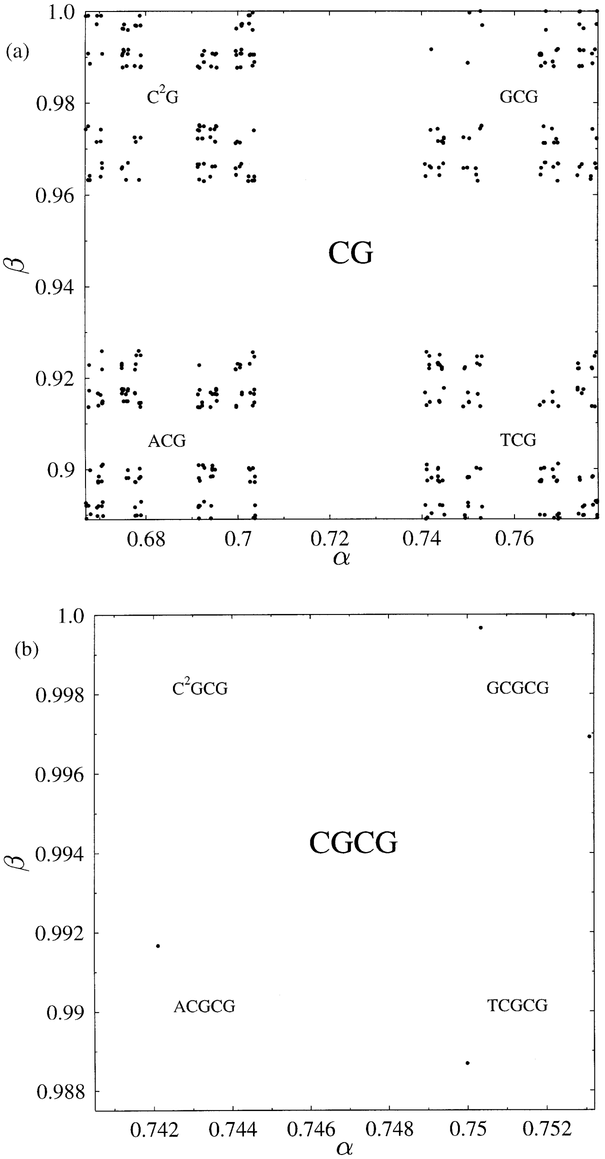

Figure 3. (a) An enlarged zone CG in Fig 2; (b) an enlarged zone CGCG in (a).

An outstanding characteristic in Fig. 2 is that the frequency of points in zone CG is less than others. It points out that the number of subsequences with the ending dinucleotide CG is less than those with the other 15 dinucleotides. When the left shift operator S ϵ {A, C, G, T} handles the subsequences, similar result can be obtained. In Fig. 2, the frequencies of points in zones CGA, CGC, CG^2 and CGT are less than others. Under the action of the left shift operator, such phenomenon can go on, *i.e.*, the pattern information in zone CG is transmitted to smaller zones. Thus, the global similarity feature in MR is produced by the left shift operators.

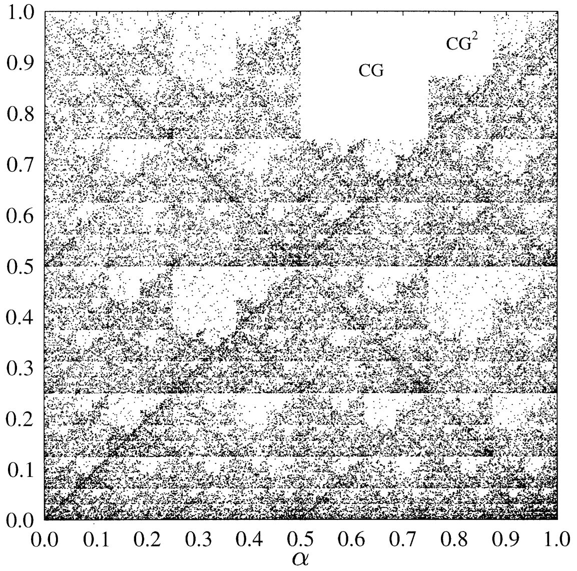

Figure 4. Chaos game representation of HUMHBB excluding the subsequences with the ending strings CG and CG^2.

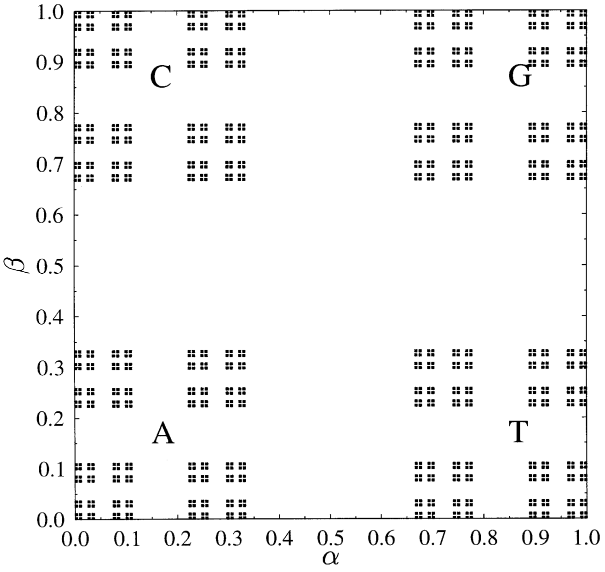

Figure 5. Metric representation of the DNA sequence of yeast's chromosome 1 with 230209 bases.

In the similarity of MR, the original zone is CG, where the existence of *k*-nucleotide strings in the DNA sequence will be discussed. In Fig. 2, since all 256 zones encoded by the tetranucleotides have some points, we may conclude that all tetranucleotides cannot be suppressed in the DNA sequence. In order to explore the existence of 5-nucleotide strings in zone CG, we first take the action of the four right shift operators on zone CG to arrive zones ACG, C^2G, GCG, TCG and enlarge it in Fig. 3a; then, we continue the action of the right shift operators on zone CGCG to zones ACGCG, C^2GCG, GCGCG, TCGCG and further enlarge it in Fig. 3b. Only five subsequences with the ending tetranucleotide string CGCG in Fig. 3b, *i.e.*, one subsequence$\cdots$ACGCG, three subsequences$\cdots$GCGCG and one subsequence$\cdots$TCGCG, exist in the DNA sequence. No subsequence$\cdots C^2GCG$ exists in the DNA sequence, *i.e.*, the 5-nucleotide string C^2GCG is suppressed. Eliminating the enlarging process, we can directly obtain the above result in Fig. 3a and other results, such as the 5-nucleotide strings CGC^2GH and CGTCG, are suppressed. The subsequences containing any suppressed *k*-nucleotide strings are still forbidden in the DNA sequence. For example, the subsequences $\cdots C^2GCG\cdots$, $\cdots CGC^2G\cdots$ and $\cdots$CGTCG$\cdots$ are also suppressed in the DNA sequence. MR can thus be used to determine the existence of a given *k*-nucleotide string in a DNA sequence.

According to ordering rules (1) and (2), we can also present another MR of the subsequences $\Sigma = s_1 s_2 \cdots s_i \cdots s_m$ as follows:

$$\alpha = \sum_{j=1}^{m} \mu_{m-j+1} 2^{-j} + 2^{-(m+1)} = \sum_{i=1}^{m} \mu_i 2^{-(m-i+1)} + 2^{-(m+1)} \tag{5}$$

and

$$\beta = \sum_{j=1}^{m} \nu_{m-j+1} 2^{-j} + 2^{-(m+1)} = \sum_{i=1}^{m} \nu_i 2^{-(m-i+1)} + 2^{-(m+1)} \tag{6}$$

where the meanings of μ_i, ν_i, α, β are the same as those given in Eqs. (3) and (4). Carefully, we find that the MR is equivalent to the CGR of DNA sequences. Similar results in MR can thus be obtained in the CGR. In [1, 2], the main results or open questions are the pattern recognition and the similarity in CGR, *i.e.*, the repetition of the double-scoop pattern in Fig. 3 of [1]. Using Eqs. (5) and (6), we have $\{A^\infty, C^\infty, G^\infty, T^\infty\} \Rightarrow \{(0,0),(0,1),(1,1),(1,0)\}$ and $\{C(A^\infty,C^\infty,G^\infty,T^\infty)\}) = \{A^\infty C, C^\infty, G^\infty C, T^\infty C\} \Rightarrow \{(0,{}^1/_2),(0,1),({}^1/_2,1),({}^1/_2,{}^1/_2)\}$. For the sequences of $\{GC(A^\infty, C^\infty, G^\infty, T^\infty)\} = \{A^\infty CG, C^\infty G, G^\infty CG, T^\infty CG\}$, the MRs $\{({}^1/_2,{}^3/_4), ({}^1/_2,1), ({}^3/_4,1), ({}^3/_4,{}^3/_4)\}$ constitute the four vertices of the zone CG. In the same way, for the sequences of $\{G^2C\ (A^\infty,C^\infty,G^\infty,T^\infty)\} = \{A^\infty CG^2, C^\infty G^2, G^\infty CG^2, T^\infty CG^2\}$, the MRs $\{({}^3/_4,{}^7/_8), ({}^3/_4,1), ({}^7/_8,1), ({}^7/_8,{}^7/_8)\}$ constitute the four vertices of zone CG^2, which is taken as an example of *S*(CG), $S \in \{A,C,G,T\}$. According to ordering rules (1) and (2), the ordering of any subsequences with the ending doublet CG is no less than that of $A^\infty CG$

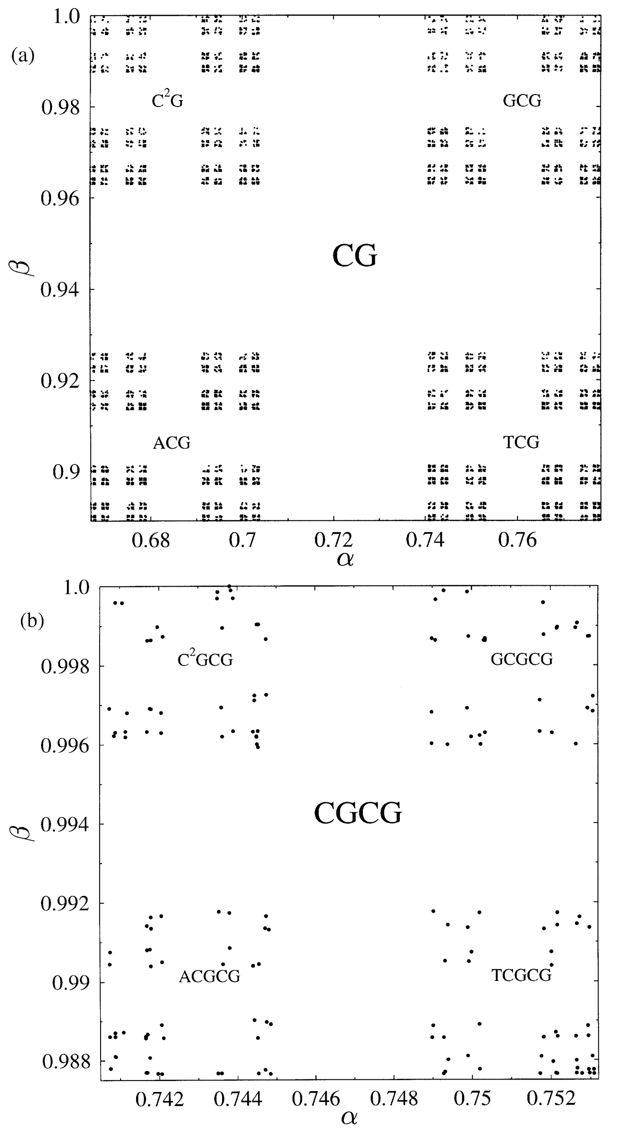

Figure 6. (a) An enlarged zone CG in Fig. 5; (b) an enlarged zone CGCG in (a).

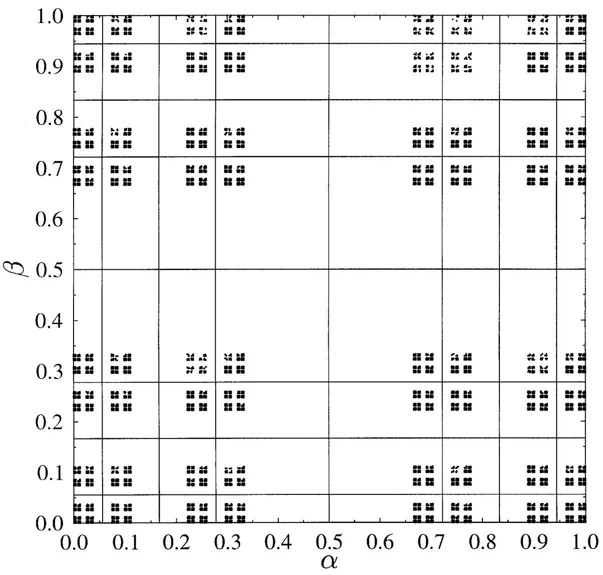

Figure 7. The bound lines for trinucleotides surrounding metric representation of HUMHBB. Each bound line exists in the center between nearby zones.

(A$^\infty$CG) and no larger than that of T$^\infty$CG (C$^\infty$G) in the α (β) direction. In Fig. 4, we draw the CGR of HUMHBB, excluding the subsequences with the ending nucleotide strings CG and CG^2. The numerical experiment confirms the above analytic result. Thus, we point out that the double-scoop pattern consists of the zone CG and its left shift mappings G(CG), G^2(CG), ⋯; the repetition of the double-scoop pattern is the result of A, C, G, T and other, longer string, left shift mapping of the zones CG, CG^2, CG^3, ⋯.

In general, the global similarity structure in CGR corresponds to the appearance of the lower frequencies of strings than others in the DNA sequence [3]. However, when the finite distinction of points in the plane is taken into account, the suppression of *k*-nucleotide strings in the DNA sequence, which is more exact than the target for frequencies of strings, will be considered. In order to compare with the MR of HUMHBB, we take the DNA sequence of yeast's chromosome 1 as an example. Its MR, with its entire 230 209 bases is drawn in Fig. 5. No apparent pattern feature exists in the figure. The frequencies of CG and TA are extracted as f_{CG} = 0.030794 and f_{TA} = 0.070297, respectively. We thus still take zone CG holding the lower frequency and repeat the above process. In Fig. 6a, all 64 zones in zone CG contain some points; therefore all 5-nucleotide strings with the ending CG in the sequence are allowed. In Fig. 6b, some of the 64 zones in zone CGCG do not contain any points, thus, the suppression of nucleotide strings exists in the 7-letter length, such as AC^{3}GCG,GC^{3}GCG,G^{2}C^{2}GCG,G^{3}CGCG, C^{2}ACGCG,CTACGCG and TGTCGCG. Thus, the suppression of some nucleotide strings in a DNA sequence leads to the pattern feature. The shorter the length of the suppressed nucleotide strings, the more apparent the pattern feature is.

One main difference of the MR from the CGR is the apparent division of point sets. The MR method can be used to exactly determine the frequencies of given *k*-nucleotide strings and the occurrence of the shortest suppressed ones. First, for a given nucleotide string in length *k*, the left and right bound lines along the α-axis, which surround the zones encoded by all *k*-nucleotide strings in the length, are obtained. Then, using Eqs. (3) and (4), the number α for the given k-nucleotide string as well as its left and right bound lines in the above set are determined. Following the same steps, we also obtain the number β for the given k-nucleotide string and determine its bottom and top bound lines. Furthermore, the number of points n falling in the square composed of the four bound lines and the frequency $f = n/(N-k+1)$, where N is the total number of all bases in the sequence, are extracted. For the mononucleotides, we have $\alpha(A) = {}^1/_3$ and $\alpha(T) = 1$. All points in zones A and T are included in $[\alpha(A^\infty), \alpha(T^\infty A)] = [0,{}^1/_3]$ and $[\alpha(A^\infty T), \alpha(T^\infty)] = [{}^2/_3,1]$, respectively. The bound lines can thus be taken as 0, $^1/_2$, and 1 along the α axis. For the dinucleotides, we have $\alpha(A^2) = {}^1/_9$, $\alpha(TA) = {}^1/_3$, $\alpha(AT) = {}^7/_9$ and $\alpha(T^2)] = 1$. All points in the zones A^2, TA, AT and T^2 are included in $[\alpha(A^\infty), \alpha(T^\infty A^2)] = [0,{}^1/_9]$, $[\alpha(A^\infty TA), \alpha(T^\infty A)] = [{}^2/_9,{}^1/_3]$, $[\alpha(A^\infty T), \alpha(T^\infty AT)] = [{}^2/_3,{}^7/_9]$ and $[\alpha(A^\infty T^2), \alpha(T^\infty)] = [{}^8/_9,1]$, respectively. The bound lines can thus be taken as 0, $^1/_6$,$^1/_2$,$^5/_6$ and 1 along the α axis. For the trinucleotides, we have $\alpha(A^3) = {}^1/_{27}$, $\alpha(TA^2) = {}^1/_9$, $\alpha(ATA) = {}^7/_{27}$, $\alpha(T^2A) = {}^1/_3$, $\alpha(A^2T) = {}^{19}/_{27}$, $\alpha(TAT) = {}^7/_9$, $\alpha(AT^2) = {}^{25}/_{27}$ and $\alpha(T^3) = 1$. All points in zones A^3, TA^2, ATA, T^2A, A^2T, TAT, AT^2 and T^3 are included in $[\alpha(A^\infty), \alpha(T^\infty A^3)] = [0,{}^1/_{27}]$, $[\alpha(A^\infty TA^2), \alpha(T^\infty A^2)] = [{}^2/_{27},{}^1/_9]$, $[\alpha(A^\infty TA), \alpha(T^\infty ATA)] = [{}^2/_9,{}^7/_{27}]$, $[\alpha(A^\infty T^2A), \alpha(T^\infty A)] = [{}^{20}/_{27},{}^1/_3]$, $[\alpha(A^\infty T), \alpha(T^\infty A^2T)] = [{}^2/_3,{}^{19}/_{29}]$, $[\alpha(A^\infty TAT), \alpha(T^\infty AT)] = [{}^{20}/_{27},{}^7/_9]$, $[\alpha(A^\infty T^2), \alpha(T^\infty AT^2)] = [{}^8/_9,{}^{25}/_{27}]$ and $[\alpha(A^\infty T^3), \alpha(T^\infty)] = [{}^{26}/_{27},1]$, respectively. The bound lines can thus be taken as 0, $^1/_{18}$,$^1/_6$,$^5/_{18}$,$^1/_2$,$^{13}/_{18}$,$^5/_6$,$^{17}/_{18}$ and 1 along the α-axis. In terms of the bound lines along the α- and β-axes given in Fig. 7, the frequencies of all trinucleotides can be extracted by counting the number of points falling in the squares composed of the bound lines. Continuing the process, we can obtain the bound lines 0, $^1/_{54}$,$^1/_{18}$,$^5/_{54}$,$^1/_6$,$^{13}/_{54}$,$^5/_{18}$,$^{17}/_{54}$,$^1/_2$,$^{37}/_{54}$,$^{13}/_{18}$,$^{41}/_{54}$,$^5/_6$,$^{49}/_{54}$,$^{17}/_{18}$,$^{53}/_{54}$ and 1 along the α-axis for the tetranucleotides and frequencies of all tetranucleotides, *etc.*. Repeating the above process for a given short nucleotide string, we can obtain its frequency in DNA sequences. Moreover, for suppression of the shortest nucleotide strings, we can determine its occurrence, *i.e.*, the length, number and corresponding bound lines of the shortest nucleotide strings. According to the compositional feature of MR in Fig. 1, the shortest suppressed nucleotide strings in DNA sequences can also be obtained. For example, in HUMHBB, the length and number of the shortest suppressed nucleotide strings are extracted as 5 and 4, respectively. Their corresponding bound lines in the MR are also obtained. By using the compositional feature in MR, the shortest suppressed nucleotide strings CGCGA, CGC^2G, C^2GCG and CGTCG are determined. All tetranucleotide strings are thus allowed in HUMHBB. In the DNA sequence of yeast's chromosome 1, the length and number of the shortest suppressed nucleotide strings are extracted as 7 and 110, respectively. All 6-nucleotide strings are thus allowed in the DNA sequence.

On the principle of the compositional feature in MR, the suppressed nucleotide strings of any length (including the shortest ones) in DNA sequences can be determined. But, in general, a real number in computers has significant digits with limited space, *e.g.*, 16's space. Since $3^{-31} = 1.6 \times 10^{-15}$, the longest valid length of suppressed nucleotide strings in DNA sequences determined by using the MR is 31. Thus, using the MR, we can determine frequencies of short nucleotide strings and the occurrence of the shortest suppressed ones in DNA sequences.

In summary, we have proposed an MR of DNA sequences in view of symbolic dynamics. Using this method, the patterns seen in CGR and [8] are explained as the suppression of certain nucleotide strings in the sequences. Moreover, the MR divides the point sets in CGR, so that each zone in the plane has a clear boundary. By using the MR, frequencies of short nucleotide strings and suppression of the shortest ones in the DNA sequences can be determined.

This work was supported in part by the Nonlinear Science Project of China. The author thanks Prof. Bai-Lin Hao for reading the manuscript and making valuable suggestions.

Received November 9, 1999

References

[1] Jeffrey, H. J., *Nucleic Acids Res.* 1990, *18*, 2163–2170.

[2] Jeffrey, H. J., *Comput. Graph.* 1992, *15*, 25–33.

[3] Goldman, N., *Nucleic Acids Res.* 1993, *21*, 2487–2491.

[4] Burge, C., Campbell, A. M., Karlin, S., *Proc. Natl. Acad. Sci. USA* 1992, *89*, 1358–1362.

[5] Cardon, L., Burge, C., Clayton, D. A., Karlin, S., *Proc. Natl. Acad. Sci. USA* 1994, *91*, 3799–3803.

[6] Hall, K. A., Singh, S. M., *Genome* 1997, *40*, 342–356.

[7] Pleißner, K.-P., Wernisch, L., Oswald, H., Fleck, E., *Electrophoresis* 1997, *18*, 2709–2713.

[8] Hao, B.-L., Lee, H.-C., Zhang, S.-Y., *Chaos, Solitons and Fractals* 2000, *11*, 825–836.

[9] Hao, B.-L., Zheng, W.-M., *Applied Symbolic Dynamics and Chaos*, World Scientific, Singapore 1998.