Computational Biology and Chemistry 53 (2014) 49–58

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Analysis of correlation structures in the *Synechocystis* PCC6803 genome

Zuo-Bing Wu*

State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Accepted 11 July 2014
Available online 19 August 2014

Keywords:
Synechocystis PCC6803 genome
Correlation structures
Recurrence plot
Reconstructed phase space

ABSTRACT

Transfer of nucleotide strings in the *Synechocystis* sp. PCC6803 genome is investigated to exhibit periodic and non-periodic correlation structures by using the recurrence plot method and the phase space reconstruction technique. The periodic correlation structures are generated by periodic transfer of several substrings in long periodic or non-periodic nucleotide strings embedded in the coding regions of genes. The non-periodic correlation structures are generated by non-periodic transfer of several substrings covering or overlapping with the coding regions of genes. In the periodic and non-periodic transfer, some gaps divide the long nucleotide strings into the substrings and prevent their global transfer. Most of the gaps are either the replacement of one base or the insertion/reduction of one base. In the reconstructed phase space, the points generated from two or three steps for the continuous iterative transfer via the second maximal distance can be fitted by two lines. It partly reveals an intrinsic dynamics in the transfer of nucleotide strings. Due to the comparison of the relative positions and lengths, the substrings concerned with the non-periodic correlation structures are almost identical to the mobile elements annotated in the genome. The mobile elements are thus endowed with the basic results on the correlation structures.

1. Introduction

The rapid accumulation of complete DNA sequences of many organisms provides an opportunity to systematically analyze their components, structures and functions. On the one hand, from the point of view of statistics and geometry, nontrivial statistical characteristics, such as the long-range correlations, the short-range correlations and the fractal features or genomic signatures were determined (Jeffrey, 1990; Li and Kaneko, 1992; Peng et al., 1992; Karlin et al., 1997; Deschavanne et al., 1999; Hao, 2000; Holste et al., 2003; Garte, 2004; Messer et al., 2005; Katsaloulis et al., 2006). Meanwhile, lots of graphical methods such as dot plot, dot matrix and recurrence analysis to compare the genomes and visualize their similarity were developed (Mount, 2004; Cao et al., 2005; Conte et al., 2012; Frahm and Shepelyansky, 2012; Kandiah and Shepelyansky, 2013). On the other hand, it was found that the transposable elements as the mobile DNA sequences can move in the genomes and make many replicas (Bennetzen, 2000; Feschotte et al., 2002; Kazazian, 2004). Understanding their origin, evolution, and effects on genome structures and gene functions is of fundamental importance for biology (Peyrard, 2004; Bergman and Quesneville, 2007; Lönnig and Saedler, 2002; Delihas, 2011).

The *Synechocystis* sp. PCC6803 (*synecho*) is one of unicellular cyanobacteria, which presumably are the oldest organisms capable of oxygenic photosynthesis. The transformable ability of the *synecho* facilitates its biotechnological applications (Thiel, 1994). Since the entire *synecho* genome was determined (Kaneko et al., 1996), a series of studies on its physical and genetic maps, and functions has been completed (Katani and Tabata, 1999; Bhaya et al., 2000; Kucho et al., 2005; Tajima et al., 2011). In particularly, 10–11 bp oscillations in the statistical correlation analysis were found to reflect protein structure and DNA folding (Herzel et al., 1999). The whole genome offers meaningful information for understanding the metabolic network and transcriptional organization of this organism in the bioengineering application (Hong and Lee, 2007; Fu, 2009; Knoop et al., 2010; Mitschke et al., 2011).

So far, the statistical analysis of the *synecho* genome has exhibited well global properties of base pairs with a correlation distance. However, correlation properties of nucleotide strings repeated in the genome are neglected, such as the transfer of nucleotide strings may happen at many positions of the genome and generate periodic correlation structures (Wu, 2013). Besides the repetition of basic periodic nucleotide strings, the transfer of non-periodic nucleotide strings with the same increasing periods would form the

* Tel.: +86 10 82543955.
E-mail address: wuzb@lnm.imech.ac.cn

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.009
1476-9271/

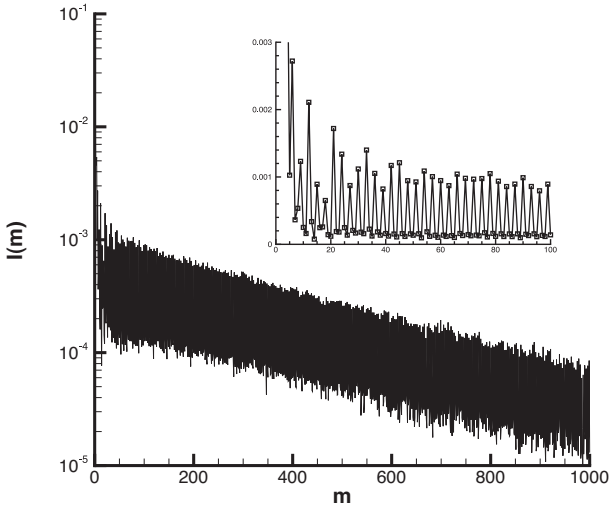

Fig. 1. Mutual information function $I(m)$ in the log scale for the *synecho* genome. A local blow-up region $m \in [0, 100]$ with the linear scale is redrawn in the figure.

periodic correlation structures. It inspires to explore more extensive correction of nucleotide strings and the intrinsic mechanism. In this paper, by using the recurrence plot method and the phase space reconstruction technique, we identify transfer of nucleotide strings in the *synecho* genome and make a detail analysis of the periodic and non-periodic correlation structures.

2. Methods

For a given genome $s_1 s_2 \cdots s_i \cdots s_N$ ($s_i \in A, C, G, T$), the mutual information function is defined (Shannon, 1948; Li, 1990; Herzel and Gloße, 1995) as

$$I(m) = \sum_{\xi,\eta=1}^{4} p_{\xi\eta}(s_\xi, s_\eta) \log_2 \frac{p_{\xi\eta}(s_\xi, s_\eta)}{p_\xi(s_\xi) p_\eta(s_\eta)}, \tag{1}$$

where $p_{\xi\eta}$ is the relative frequency of the pair of s_ξ and $s_\eta = s_{\xi+m}$ in a distance m and p_ξ is the relative frequency of s_ξ. The *synecho* genome denoted as BA000022 is obtained from the GenBank (ftp.ncbi.nih.gov) and has 3,573,470 bases. The symmetrical distribution of four bases along the single strand is $p_\xi = 26.1\%/26.2\%$ and $23.8\%/23.9\%$ for $s_\xi = A/T$ and C/G. Fig. 1 displays the mutual information function of the genome ($m \le 1000$) and reflects correlations with the exponential decay in the long-range scale. To investigate correlations in the short-range scale, we concentrate on the mutual information function for $m \in [1, 100]$ and redraw the local blow-up region in Fig. 1. It is evident that the fundamental vibration frequency is 3 bp, which are due to the genetic code. The correlation analysis of the genome provides the global correlation properties of the two base pairs (s_ξ and s_η) with correlation distance m in the short- and long-range scales, but correlation properties of two nucleotide strings with the ending bases s_ξ and s_η in the genome are neglected.

In what follows, we give a brief presentation of the recurrence plot method based on the metric representation, which is detailed in Eckmann et al. (1987), Wu (2000), and Wu (2004). Firstly, the genome is partitioned into N subsequences $\Sigma_k = s_1 s_2 \cdots s_k (1 \le k \le N)$ and mapped in a metric plane (α, β). The metric mapping (α_k, β_k) of a subsequence is defined as

$$\begin{aligned} \alpha_k &= 2\sum_{j=1}^{k} \mu_{k-j+1} 3^{-j} + 3^{-k} = 2\sum_{i=1}^{k} \mu_i 3^{-(k-i+1)} + 3^{-k}, \\ \beta_k &= 2\sum_{j=1}^{k} \nu_{k-j+1} 3^{-j} + 3^{-k} = 2\sum_{i=1}^{k} \nu_i 3^{-(k-i+1)} + 3^{-k}, \end{aligned} \tag{2}$$

where μ_i is 0 if $s_i \in \{A, C\}$ or 1 if $s_i \in \{G, T\}$ and ν_i is 0 if $s_i \in \{A, T\}$ or 1 if $s_i \in \{C, G\}$. The points (α_k, β_k) concentrate in local zones of the metric plane ($[0, 1] \times [0, 1]$). The subsequences with the same ending l-nucleotide string labeled by Σ^l correspond to points in the zone encoded by the l-nucleotide string. With two subsequences $\Sigma_i \in \Sigma^l$ and Σ_j ($j \ge l$), we calculate the distance between the points (α_i, β_i) and (α_j, β_j) in the plane. When the distance is not longer than the zone size $\epsilon_l = 3^{-l}$, i.e., $\Sigma_j \in \Sigma^l$, the point (i, j) is plotted in a recurrence plot plane. Repeating the above process and shifting forward, we obtain the recurrence plot of the genome. In comparison with the definition (1), it is clear that the mutual information function only corresponds to the recurrence plot with $l = 1$. The recurrence plot basically depends on the length N of the genome and the zone size ϵ_l. At the fixed length N, the recurrence plot with a small l is easier to investigate global properties than that with a large l, but to find local properties such as the transfer of long nucleotide strings, latter is better. Although the density of points in the recurrence plot decreases monotonically as l increases, their distributions in the plane are fixed. From the recurrence plot plane, we calculate the maximal value of x to satisfy $\Sigma_{i+x}, \Sigma_{j+x} \in \Sigma^l$ ($x = 0, 1, 2, \cdots x_{max}$). The transferred nucleotide string has the length $L = l + x_{max}$ and is placed at the positions $(i - l + 1, i + x_{max})$ and $(j - l + 1, j + x_{max})$, which implies the transferring distance $d_T = j - i$ starting from the diagonal line in the plane. Then, to depict the correlation structures in the plane, on the one hand, we propose a correlation intensity at a given transferring distance d_T

$$\Xi(d_T) = \sum_{i=1}^{N-d_T} \Theta(\epsilon_l - |\Sigma_i - \Sigma_{i+d_T}|), \tag{3}$$

where Θ is the Heaviside function; on the other hand, we define an iterative transferring distance x_k of the given nucleotide string

$$\begin{aligned} x_k &= d_T(k) - d_T(k-1), \quad k > 1, \\ x_1 &= d_T(1), \end{aligned} \tag{4}$$

where $d_T(k)$ is the kth transferring distance of the given nucleotide string starting from the diagonal line in the plane. Applying the phase space reconstruction technique (Packard et al., 1980), we generate two-dimensional vectors from the one-dimensional iterative transferring distance x_k

$$\mathbf{y}_k = (x_k, x_{k+1}), \quad k = 1, 2 \cdots. \tag{5}$$

3. Analytical results

3.1. Recurrence analysis of coding and non-coding regions

By using the above method in Section 2, points in the recurrence plot of the genome for $l = 15$ are determined and divided into four parts shown in Fig. 2 due to the transfer of l-nucleotide strings in/between coding and/or non-coding regions of the genome. Fig. 2(a) and (d) display the transfer of l-nucleotide strings in the coding and the non-coding regions, respectively. One half is excluded due to the mirror symmetry about the diagonal line $j = i$. It means that the point (i_a, j_a) in Fig. 2(a)/(d) is the same with the point (j_a, i_a) in the excluded part of the figure. Fig. 2(b)/(c) displays the transfer of l-nucleotide strings from the coding/non-coding region

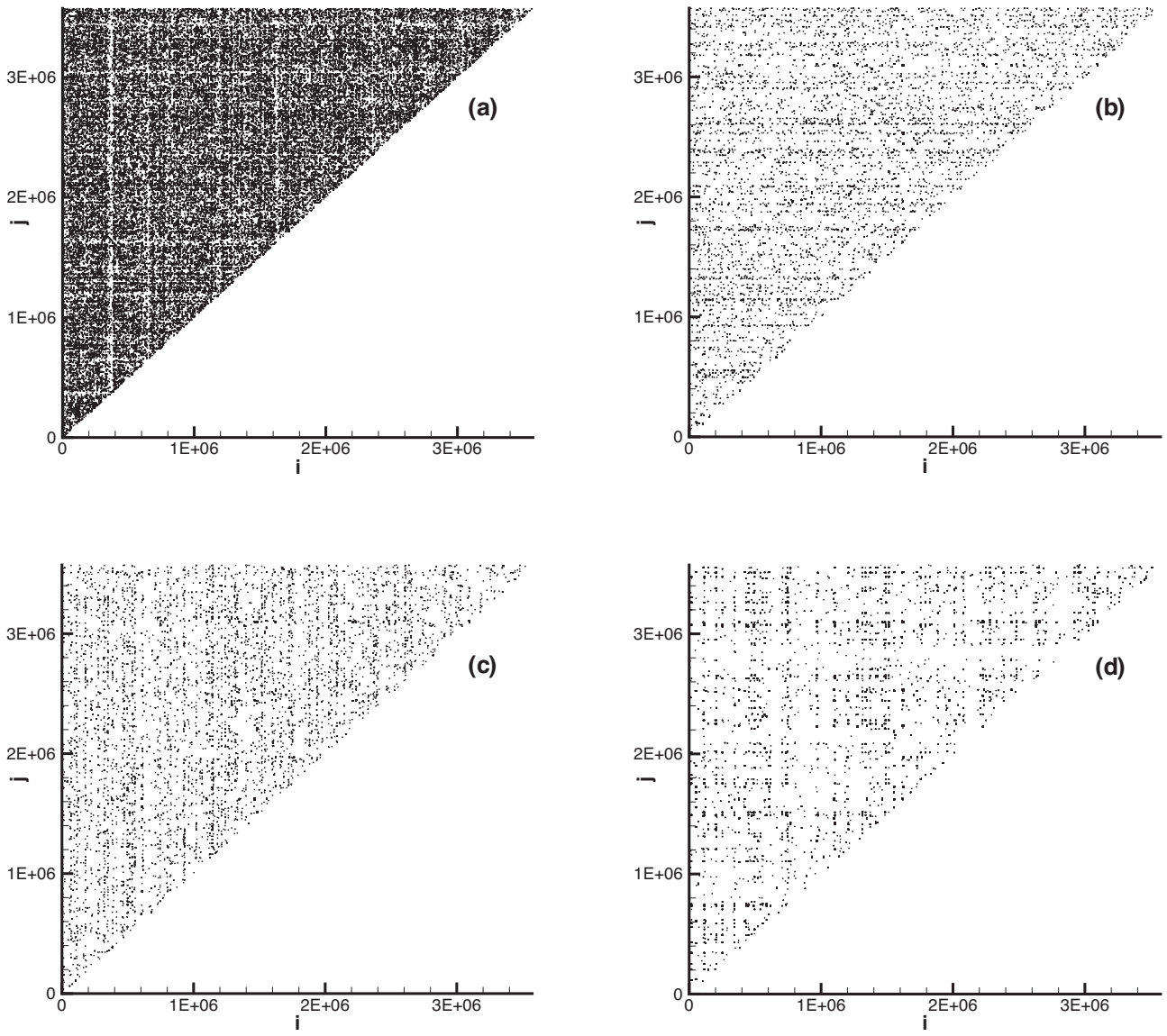

Fig. 2. Recurrence plots (a) in the coding region; (b) from the coding region to non-coding one; (c) from non-coding region to coding one; and (d) in the non-coding region of the *synecho* genome for 15-nucleotide strings.

to the non-coding/coding one. Only one half is kept due to the complementary symmetry about the diagonal line $j = i$. It means that the point (j_a, i_a) in the excluded part of Fig. 2(b) is the same with the mirror symmetrical point (j_a, i_a) of the point (i_a, j_a) in Fig. 2(c). So the point (i_a, j_a) in Fig. 2(b) and the mirror symmetrical point (j_a, i_a) in Fig. 2(c) can be combined into a figure, which displays the transfer of l-nucleotide strings from the coding region to the non-coding one. In the same way, a figure displaying the transfer of l-nucleotide strings from the non-coding region to the coding one can be generated by combining Fig. 2(c) and the mirror symmetrical part of Fig. 2(b). There exists the rotational symmetry about the diagonal line between two figures. It is evident that the transfer of l-nucleotide strings most and least frequently appears in the coding and non-coding regions, respectively. In Fig. 2(a) and (d), there appear some horizontal and vertical dense/sparse bands. In general, a vertical dense/sparse band $[i_{init}, i_{term}]$ means that l-nucleotide strings placed in the region $[i_{init}, i_{term}]$ of the genome are transferred to the more/less positions $[j, j + i_{term} - i_{init}]$ of the genome, where $j > i_{init}$. The more/less points $(i \in [i_{init}, i_{term}], j > i_{init})$ in the recurrence plot form the vertical dense/sparse band. A horizontal dense/sparse band $[j_{init}, j_{term}]$ means that l-nucleotide strings in the more/less positions $[i - j_{term} + j_{init}, i]$ of the genome are transferred into the region $[j_{init}, j_{term}]$ of the genome, where $i < j_{term}$. The more/less points $(i < j_{term}, j \in [j_{init}, j_{term}])$ in the recurrence plot form the horizontal dense/sparse band. Using the mirror symmetry, the horizontal dense/sparse bands in one figure can be reflected by the vertical ones in the excluded part of the same figure. In Fig. 2(b) and (c), there also exist some horizontal and vertical dense/sparse bands, respectively. In the same way, using the complementary symmetry, the horizontal dense/sparse bands in the excluded part of Fig. 2(b) can be reflected by the vertical dense/sparse ones in Fig. 2(c). The vertical dense/sparse bands in Fig. 2(c) imply that the l-nucleotide strings in several local positions of the non-coding region are transferred to more/less positions in the coding one. The horizontal dense/sparse bands in Fig. 2(b) imply that the l-nucleotide strings in more/less positions of the coding region are transferred into several local positions in the non-coding one.

To depict the above transfer behaviors of the nucleotide strings in one or several vertical bands in Fig. 2(a)–(d) the correlation

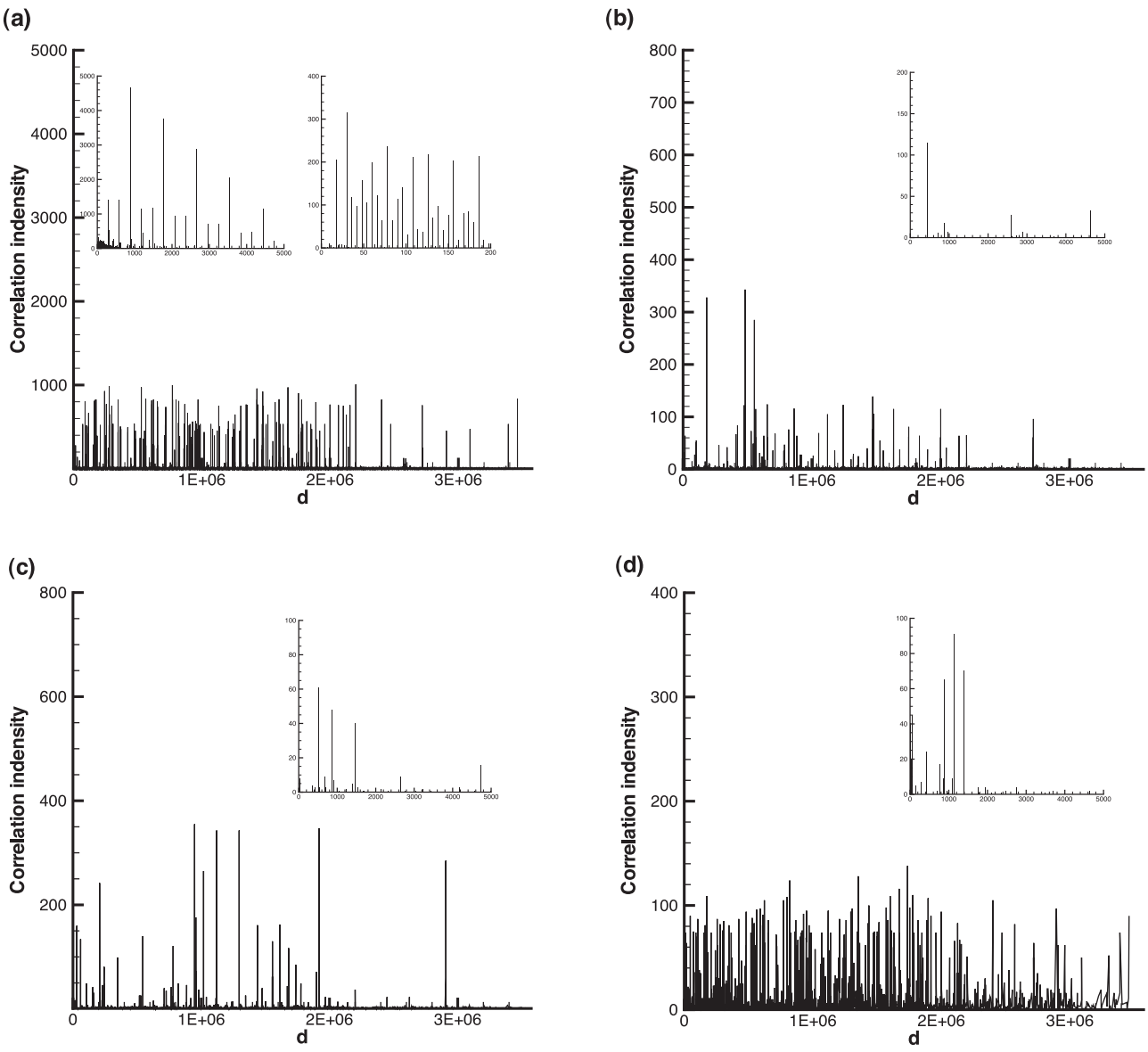

Fig. 3. Correlation intensity $\Xi(d)$ plots versus transfer distance d (a) in the coding region; (b) from the coding region to non-coding one; (c) from non-coding region to coding one; and (d) in the non-coding region of the *synecho* genome. Two local blow-up regions near $d = 0$ are redrawn in (a).

intensity is calculated by using Eq. (3) and drawn in Fig. 3. It describes the length of transferred nucleotide strings with the same correlation distance in one band or the sum of lengths of them in several bands. In the whole range of transfer distance, many discrete values are distributed and denoted as non-periodic correlation structures. To depict correlation structures in local regions, the region $d \in [0, 5000]$ is magnified in Fig. 3. Some equidistant parallel lines with a basic transferring length $d_{b_2} = 888$ appear in Fig. 3(a). Another basic transferring length $d_{b_1} = 6$ can be also determined when the region $d \in [0, 200]$ is further magnified in Fig. 3(a). They form periodic correlation structures. In the following subsections, we will analyze the periodic and non-periodic correlation structures.

3.2. Correlation analysis of nucleotide strings

3.2.1. Periodic correlation structures

By using the above method in Section 2, it is found that the periodic transfer of nucleotide strings with lengths ($L \geq 20$) is confined in local regions of the recurrence plot for $l = 15$. Fig. 4(a)–(d) displays several parallel lines in the local regions, where basic transferring lengths are determined as 6, 306, 18 and 888, respectively. In general, parallel lines in a local region of the recurrence plot reflect periodic transfer of several substrings in a long nucleotide string. The periodic correlation structures are classified into three kinds in terms of the relation of the substrings to the long nucleotide strings.

(1) The long nucleotide string is a periodic one composed of one basic string, which is divided into several substrings by gaps. The long nucleotide strings in Fig. 4(c) and (d) appear in the local regions (2,082,442–2,082,614) and (2,354,010–2,358,767), respectively. The local region in Fig. 4(c) is embedded in the coding region (2,080,885–2,082,720) of the gene *slr*0422. It is evident that there appear some equidistant parallel lines with the basic transferring length $3d_{b1} = 18$. For each transfer distance, the total nucleotide string is divided into several substrings with different lengths for transfer. The gap between two neighboring substrings consists of one base only. There exist 8 independent gaps, which are “t” at 2,082,471, “c”

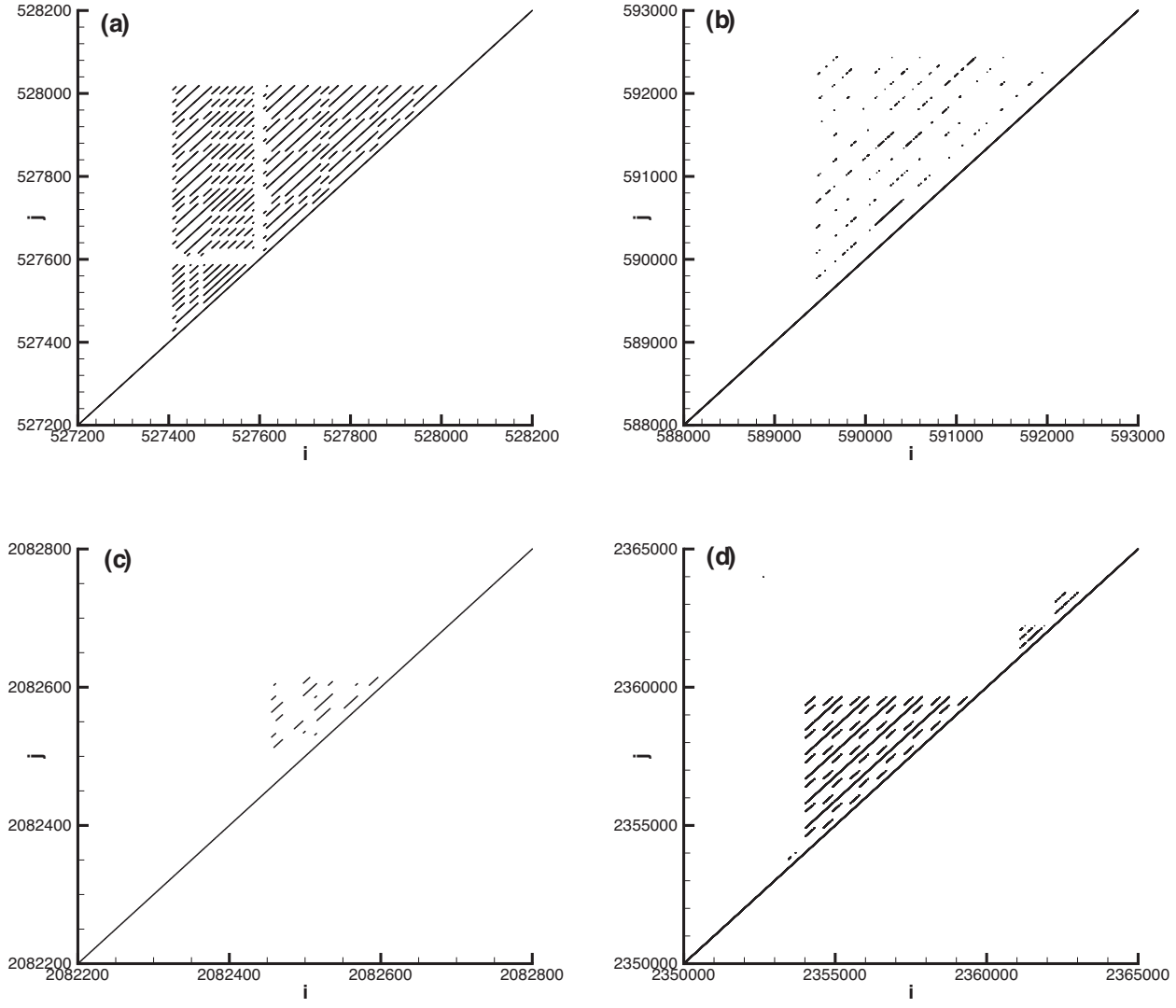

Fig. 4. Four local regions in the recurrence plot for the *synecho* genome, where periodic correlation structures are exhibited. The diagonal line ($i = j$) is also plotted.

at 2,082,480, "g" at 2,082,483, "c" at 2,082,498, "c" at 2,082,534, "g" at 2,082,537, "c" at 2,082,588 and "c" at 2,082,606. The substrings are transferred with integer times of the basic transferring length to form periodic correlation structures as shown in Fig. 3(a). Once the 8 replaced bases "c", "t", "a", "t", "t", "a", "c" and "c" in the gaps are restored, respectively, the divided substrings will combine to form a continuous periodic nucleotide string and make the periodic transfer of the basic string.

Similarly, the local region in Fig. 4(d) is embedded in the coding region (2,351,323–2,360,412) of the gene *slr*0364. It is evident that there exist 6 equidistant parallel lines, whose lengths decrease as the transfer distance increases. For the transfer of nucleotide strings in the local region, we reorder the transferred nucleotide strings in terms of the increasing periods in Table 1. The transferred nucleotide strings have the same starting position 2,354,010 and the basic transferring length $d_{b2} = 888$. For each transfer distance, the total nucleotide string is divided into several substrings with different lengths. The gap between two neighboring substrings consists of one base only. Since the gap may appear at either the original position 1 or transferred one 2, we present all gaps and their restored bases at the position 1 or 2 in Table 1. It is found that there appear 5 independent gaps in the long nucleotide string, which are denoted by square brackets in Table 1. Other gaps are just reappearance of them. The 5 independent gaps are "a" at 2,355,834, "t" at 2,356,965, "a" at 2,359,386, "t" at 2,358,741 in the first transfer and "t" at 2,357,853 in the second transfer. Once the replaced bases "t", "c", "t", "c" and "c" in the gaps are restored, respectively, the divided substrings will combine to form a continuous periodic nucleotide string and make the periodic transfer of the basic string. Different from Fig. 4(c), Fig. 4(d) also displays two groups of some short discrete parallel lines between two long ones with the distance d_{b2} in the local region. It is found that two identical nucleotide strings with the length 294 are placed at the phases 3–296 and 594–887 in the basic string with the length d_{b2}. It means that the transfer of the nucleotide string in a period from the first position to the second one has the correlation distance 591. In the same way, the transfer of the nucleotide string between two periods from the second/first position to the first one has the correlation distance 297/888. The transfer of the nucleotide string between two periods from the first position to the first/second one has the correlation distance 888/(591 + 888) and so on so forth. Of course, the 5 gaps also divide the nucleotide string into several substrings for transfer to form the periodic correlation structures in Fig. 3(a). So, once the replaced bases in the gaps are restored, the divided substrings will combine and make the periodic transfer of the nucleotide string.

Table 1
Periodic transfer of nucleotide strings with lengths $L(\geq 20)$ in the local region $(2.354–2.36 \times 10^6)$ for the *synecho* genome. $d_T(d_{b2})$ is the (basic) transfer distance. L_T is the total lengths of transferred nucleotide strings. $s_1 \rightleftharpoons s_2$ denotes the restoration of bases in gaps.

k	$d_T(L_T)$	Position 1	L	Position 2	$s_1 \rightleftharpoons s_2$
1	d_{b2}(4753)	2,354,010–2,354,945	936	2,354,898–2,355,833	$t \rightarrow [a]$
		2,354,947–2,355,833	887	2,355,835–2,356,721	$a \leftarrow t$
		2,555,835–2,356,076	242	2,356,723–2,356,964	$c \rightarrow [t]$
		2,356,078–2,358,497	2420	2,356,966–2,359,385	$t \rightarrow [a]$
		2,358,499–2,358,740	242	2,359,387–2,359,628	$[t] \leftarrow c$
		2,358,742–2,358,767	26	2,359,639–2,359,655	
2	$2d_{b2}$(3864)	2,354,010–2,354,057	48	2,355,786–2,355,833	$t \rightarrow a$
		2,354,059–2,355,188	1130	2,355,835–2,356,964	$c \rightarrow t$
		2,355,190–2,355,833	644	2,356,966–2,357,609	$a \leftarrow t$
		2,355,835–2,356,076	242	2,357,611–2,357,852	$c \rightarrow [t]$
		2,356,078–2,357,609	1532	2,357,854–2,359,385	$t \rightarrow a$
		2,357,611–2,357,852	242	2,359,387–2,359,628	$t \leftarrow c$
		2,357,854–2,357,879	26	2,359,630–2,359,655	
3	$3d_{b2}$(2976)	2,354,010–2,354,300	291	2,356,674–2,356,964	$c \rightarrow t$
		2,354,302–2,355,188	887	2,356,966–2,357,852	$c \rightarrow t$
		2,355,190–2,355,833	644	2,357,854–2,358,497	$a \leftarrow t$
		2,355,835–2,356,076	242	2,358,499–2,358,740	$c \rightarrow t$
		2,356,078–2,356,721	644	2,358,742–2,359,385	$t \rightarrow a$
		2,356,723–2,356,964	242	2,359,387–2,359,628	$t \leftarrow c$
		2,356,966–2,356,991	26	2,359,630–2,359,655	
4	$4d_{b2}$(2092)	2,354,010–2,354,300	291	2,357,562–2,357,852	$c \rightarrow t$
		2,354,302–2,355,188	887	2,357,854–2,358,740	$c \rightarrow t$
		2,355,290–2,356,103	914	2,358,742–2,359,655	
5	$5d_{b2}$(1204)	2,354,010–2,354,300	291	2,358,450–2,358,740	$c \rightarrow t$
		2,354,302–2,354,945	644	2,358,742–2,359,385	$t \rightarrow a$
		2,354,947–2,355,215	269	2,359,387–2,359,655	
6	$6d_{b2}$(317)	2,354,010–2,354,057	48	2,359,338–2,359,385	$t \rightarrow a$
		2,354,059–2,354,327	269	2,359,387–2,359,655	

(2) The long nucleotide string is a periodic one composed of two basic strings, which are divided into several substrings by gaps. The long nucleotide string in Fig. 4(a) appears in the local region (527,395–528,016), which is embedded in the coding region (524,346–529,595) of the gene *slr*1753. It is evident that there appear some equidistant parallel lines with the basic transferring length $d_{b1} = 6$. Basically, the long nucleotide string is composed by two basic strings with lengths 12 and 6. However, the second one disappears at some positions in the long nucleotide string. There also exist 2 independent gaps with "t" at 527,588 and "t" at 527,594, which divides the long nucleotide string into several substrings. The substrings are transferred with the integer times of the basic transferring length to generate periodic correlation structures as shown in Fig. 3(a). Once the 2 replaced bases "c" and "g" in the gaps are restored, respectively, the divided nucleotide strings will combine to form a continuous periodic nucleotide string and make the periodic transfer of the basic strings.

(3) The long nucleotide string is not a periodic one but is composed of several substrings for the transfer and non-transfer. The long nucleotide string in Fig. 4(b) appears in the local region (589,463–592,418), which is embedded in the coding region (587,228–593,125) of the gene *slr*2046. It is evident that there appear some equidistant parallel lines with the basic transferring length 306. However, there does not exist the basic string with length 306 in the genome. In the long nucleotide string, several substrings can be transferred to generate the periodic coherence structures, but others cannot. So the transferred nucleotide strings cannot combine to form a continuous periodic nucleotide string, but can be still transferred periodically in the long nucleotide string.

3.2.2. Non-periodic correlation structures

Fig. 5 displays local regions in the recurrence plot for the non-periodic correlation structures, where exist several non-equidistant parallel lines. Consider the local region (379,000–379,942) denoted by zone 2 in Fig. 5(b), which covers the coding region (379,065–379,913) of the gene *slr*1075, as an example. It is evident that the lengths of the non-equidistant parallel lines are almost identical in the increasing of transfer distance. For the non-periodic transfer of nucleotide strings, we reorder the transferred nucleotide strings in terms of the increasing of transfer distance in Table 2. For the first transfer distance 719,258, the original nucleotide string with length 943 located at 379,000–379,942 is divided into 12 substrings by 11 gaps and transferred to the position 1,098,258–1,099,200 covering the coding region (1,098,323–1,099,171) of the gene *slr*1357. The 12 nucleotide substrings have different lengths. Each gap has just one base. Since the change of base in the gap cannot be predicted to happen at the original position 1 or the transferred one 2 in Table 2, the replacement of the base at the position 2 is taken as the gap in the transfer process. And the gaps appearing first in the process are defined as independent and are denoted by square brackets in Table 2. Other gaps in the following process are just reappearance of them. Once the replaced bases in the gaps are restored, as given in Table 2, the continuous transfer of the original nucleotide string would happen. For the seventh, eighth and tenth transfer distances 2,064,934, 2,155,041 and 2,718,370, the transfer of the original nucleotide string is similar to that for the first one in covering the coding regions (2,443,999–2,444,847), (2,534,106–2,534,954) and (3,097,436–3,098,284) of the genes *slr*0352, *slr*0230 and *slr*0704, respectively. For the second transfer distance 967,132, the original nucleotide string is divided into 16 substrings by 15 gaps and transferred to the position 1,346,132–1,347,073 covering the coding region (1,346,197–1,346,556) of the gene *slr*0856 and the coding region (1,346,550–1,347,044) of the gene *slr*0857. Most of the gaps are just the replacement of one base. In particular, the base "a" at the original position 379,367 is not replaced but just removed in the transferred position 1,346,499 for the gap. Since the nucleotide string in the position 2 has

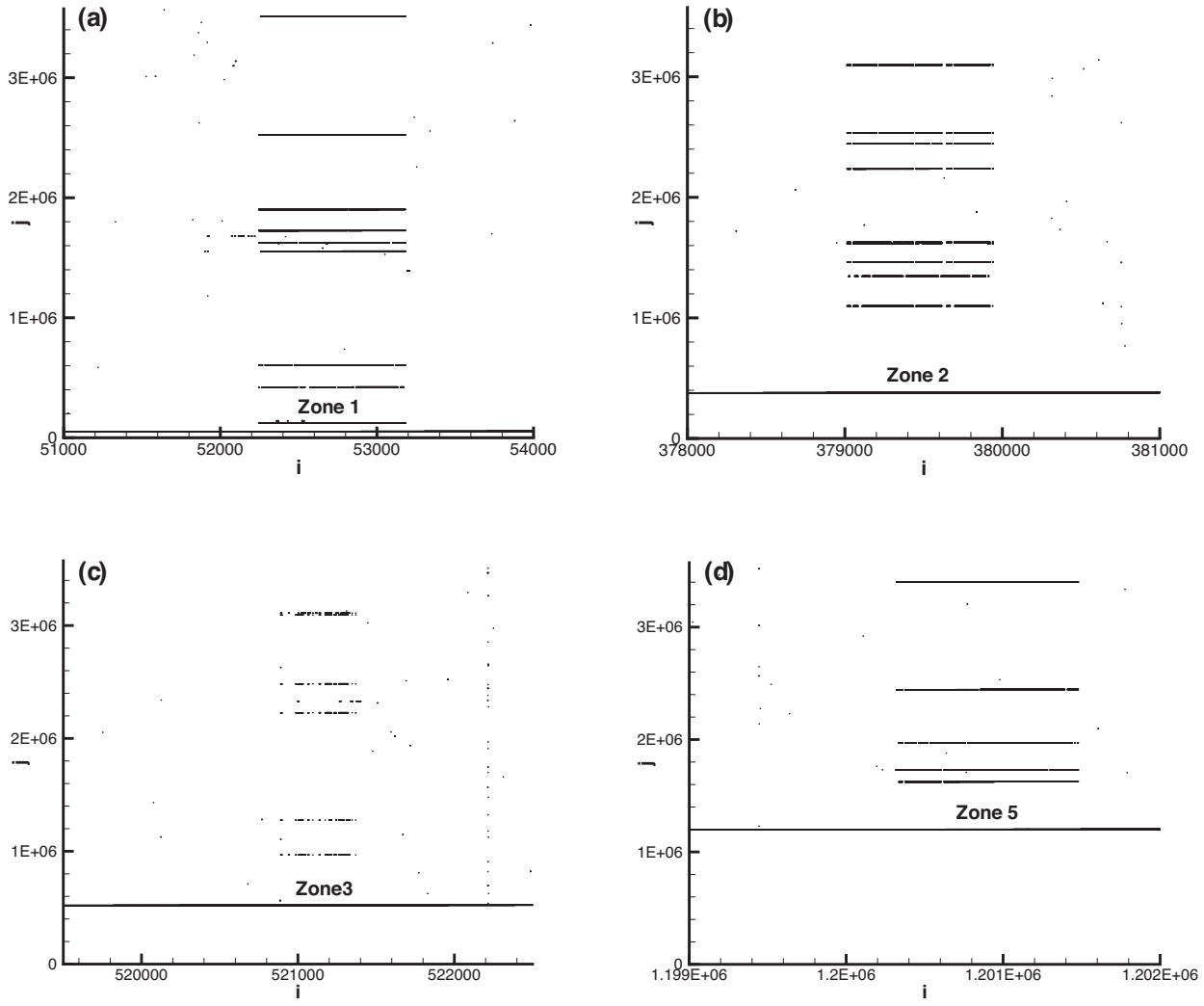

Fig. 5. Four local regions (52,217–53,173), (379,000–379,942), (520,876–521,321) and (1,200,307–1,201,479) denoted by zones 1, 2, 3 and 5 in the recurrence plot for the *synecho* genome, where non-periodic correlation structures are exhibited. The diagonal line ($i = j$) is also plotted.

the global movement with one-base, the transfer distance is shrunk to 967,131. For the third, fifth and sixth transfer distances 1,084,390, 1,247,100 and 1,856,496, the transfer of the original nucleotide string is similar to that for the second one on covering three groups of the coding regions of two genes. They are the coding regions (1,463,455–1,463,814), (1,463,808–1,464,302), (1,626,165–1,626,524), (1,626,518–1,627,012), (2,235,561–2,235,902) and (2,235,914–2,236,408) of the genes *slr*1715, *slr*1716, *slr*2112, *slr*2113, *slr*1936 and *slr*1937, respectively. For the fourth transfer distance 1,237,839, the original nucleotide string with length 626 is divided into 8 substrings by 7 gaps and transferred to the position 1,616,839–1,617,473 covering the coding region (1,616,904–1,617,419) of the gene *slr*1524. Most of the gaps are just the replacement of one base. In particular, the bases "a" and "g" are inserted in the positions 1,617,312 and 1,617,316 for the gaps, respectively, so the nucleotide strings in the position 2 have global movements with one-base and two-bases. The transfer distance is expanded into 1,237,840 and 1,237,841, respectively. For the ninth transfer distance 2,716,982, the original nucleotide string with length 336 is divided into 4 substrings by 3 gaps and transferred to the position 3,095,982–3,096,322 covering the coding region (3,096,048–3,096,332) of the gene *ssr*1175. Each gap is just the replacement of one base. There also exist some short nucleotide strings with the similar transfer to the above one, which are not presented in Table 2. Therefore, the non-periodic correlation structure is generated by the transfer of non-periodic nucleotide strings divided by several gaps.

In the *synecho* genome, mobile elements with different lengths, which referred to as selfish repetitive DNA sequences, are provided. Due to the comparison of the relative positions and lengths, 11 mobile elements located near the transferred nucleotide strings in the zone 2 are given in Table 2. It is evident that the mobile elements are almost identical to the nucleotide strings at the transfer positions. The slight differences between them are due to the different choices of their starting and ending positions. From the above correlation analysis, the mobile elements are almost the same or their substrings and are transferred with the non-periodic distance. In the transfer process, the mobile elements are divided into the substrings by several gaps. Most of gaps are either the replacement of one base or the insertion/reduction of one base.

The similar transfer of nucleotide strings with different lengths for the non-periodic correlation structures appears in other local regions. Six independent cases for the non-periodic transfer of the nucleotide strings with the lengths larger than 445 bases are distributed in the local regions (52,217–53,173)

Table 2
Non-periodic transfer of nucleotide strings in the local region ($3.79–3.8 \times 10^5$) for the *synecho* genome. Notation as in Table 1. The mobile elements (with lengths) located near the transferred nucleotide strings are given in the last line of each transfer step.

k	$d_T(L_T)$	Position 1	L	Position 2	$s_1 \to s_2$
1	719,258(932)	379,000–379,037	38	1,098,258–1,098,295	$g \to [a]$
		379,039–379,084	46	1,098,297–1,098,342	$a \to [g]$
		379,086–379,093	8	1,098,344–1,098,351	$g \to [a]$
		379,095–379,435	341	1,098,353–1,098,693	$c \to [t]$
		379,437–379,615	179	1,098,695–1,098,873	$a \to [g]$
		379,817–379,630	14	1,098,875–1,098,888	$t \to [c]$
		379,632–379,669	38	1,098,890–1,098,927	$a \to [t]$
		379,671–379,678	8	1,098,929–1,098,936	$a \to [g]$
		379,680–379,682	3	1,098,938–1,098,940	$c \to [t]$
		379,684–379,922	239	1,098,942–1,099,180	$c \to [t]$
		379,924–379,939	16	1,099,182–1,099,197	$t \to [c]$
		379,941–379,942	2	1,099,199–1,099,200	
	ME(947)	378,993–379,939		1,098,251–1,099,197	
2	967,132(361)	379,000–379,007	8	1,346,132–1,346,140	$a \to [g]$
		379,009–379,032	24	1,346,141–1,346,164	$t \to [c]$
		379,034–379,046	13	1,346,166–1,346,178	$a \to [g]$
		379,048–379,084	37	1,346,180–1,346,216	$a \to g$
		379,086–379,093	8	1,346,218–1,346,225	$g \to a$
		379,095–379,161	67	1,346,227–1,346,293	$c \to [t]$
		379,163–379,366	204	1,346,295–1,346,498	$a \to [\,]$
	967,131(567)	379,368–379,373	6	1,346,499–1,346,504	$c \to [t]$
		379,375–379,594	220	1,346,506–1,346,725	$a \to [g]$
		379,596–379,792	197	1,346,727–1,346,923	$g \to [a]$
		379,594–379,795	2	1,346,925–1,346,926	$a \to [g]$
		379,797–379,894	98	1,346,928–1,347,025	$t \to [c]$
		379,896–379,914	19	1,347,027–1,347,045	$g \to [a]$
		379,916–379,921	6	1,347,047–1,347,052	$t \to [c]$
		379,923–379,932	10	1,347,054–1,347,063	$a \to [t]$
		379,934–379,942	9	1,347,065–1,347,073	
	ME(946)			1,346,125–1,347,070	
3	1,084,390(367)	379,000–379,037	38	1,463,390–1,463,427	$g \to a$
		379,039–379,366	328	1,463,429–1,463,756	$a \to [\,]$
	1,084,389(579)	379,368–379,435	68	1,463,757–1,463,824	$c \to t$
		379,437–379,615	179	1,463,826–1,464,004	$a \to g$
		379,617–379,630	14	1,464,006–1,464,019	$t \to c$
		379,632–379,678	47	1,464,021–1,464,067	$a \to g$
		379,680–379,922	243	1,464,069–1,464,311	$c \to t$
		379,924–379,942	19	1,464,313–1,464,331	
	ME(946)			1,463,383–1,464,328	
4	1,237,839(470)	379,000–379,037	38	1,616,839–1,616,876	$g \to a$
		379,039–379,435	397	1,616,878–1,617,274	$c \to t$
		379,437–379,469	33	1,617,276–1,617,308	$g \to [a]$
		379,471–379,472	2	1,617,310–1,617,311	$g \to [ac]$
	1,237,840(2)	379,474–379,475	2	1,617,314–1,617,315	$t \to [gc]$
	1,237,841(154)	379,477–379,615	139	1,617,318–1,617,456	$a \to g$
		379,617–379,625	9	1,617,458–1,617,466	$t \to [c]$
		379,627–379,632	6	1,617,468–1,617,473	
	ME(678)			1,616,832–1,617,509	
5	1,247,100(365)	379,000–379,037	38	1,626,100–1,626,137	$g \to a$
		379,039–379,293	255	1,626,139–1,626,393	$c \to [t]$
		379,295–379,366	72	1,626,395–1,626,466	$a \to$
	1,247,099(569)	379,368–379,435	68	1,626,467–1,626,534	$c \to t$
		379,437–379,615	179	1,626,536–1,626,714	$a \to g$
		379,417–379,630	14	1,626,716–1,626,729	$t \to c$
		379,632–379,678	47	1,626,731–1,626,777	$a \to g$
		379,680–379,922	243	1,626,779–1,627,021	$c \to t$
		379,624–379,942	19	1,627,023–1,627,041	
	ME(946)			1,626,093–1,627,038	
6	1,856,496(366)	379,000–379,037	38	2,235,496–2,235,533	$g \to a$
		379,039–379,366	328	2,235,535–2,235,862	$a \to$
	1,856,495(569)	379,368–379,435	68	2,235,863–2,235,930	$c \to t$
		379,437–379,615	179	2,235,932–2,236,110	$a \to g$
		379,617–379,630	14	2,236,112–2,236,125	$t \to c$
		379,632–379,678	47	2,236,127–2,236,173	$a \to g$
		379,680–379,922	243	2,236,175–2,236,417	$c \to t$
		379,924–379,941	18	2,236,419–2,236,436	$t \to [\,a]$
	ME (946)			2,235,496–2,236,436	

Table 2 (*Continued*)

k	$d_T(L_T)$	Position 1	L	Position 2	$s_1 \to s_2$
7	2,064,934(936)	379,000–379,037	38	2,443,934–2,443,971	$g \to a$
		379,039–379,435	397	2,443,973–2,444,369	$c \to t$
		379,437–379,539	103	2,444,371–2,444,473	$g \to [a]$
		379,541–379,615	75	2,444,475–2,444,549	$a \to g$
		379,517–379,630	14	2,444,551–2,444,564	$t \to c$
		379,632–379,678	47	2,444,566–2,444,612	$a \to g$
		379,680–379,922	243	2,444,614–2,444,856	$c \to t$
		379,924–379,942	19	2,444,858–2,444,876	
	ME (947)			2,443,927–2,444,873	
8	2,155,041(935)	379,000–379,037	38	2,534,041–2,534,078	$g \to a$
		379,039–379,203	165	2,534,080–2,534,244	$c \to [t]$
		379,205–379,435	231	2,534,246–2,534,476	$c \to t$
		379,437–379,615	179	2,534,478–2,534,656	$a \to g$
		379,417–379,630	14	2,534,458–2,534,671	$t \to c$
		379,632–379,678	47	2,534,673–2,534,719	$a \to g$
		379,680–379,922	243	2,534,721–2,534,963	$c \to t$
		379,924–379,941	18	2,534,965–2,534,982	$t \to [c]$
	ME (947)			2,534,034–2,534,980	
9	2,716,982(338)	379,000–379,037	38	3,095,982–3,096,019	$g \to a$
		379,039–379,203	165	3,096,201–3,096,185	$c \to t$
		379,205–379,337	133	3,096,187–3,096,319	$g \to [a]$
		379,339–379,340	2	3,096,121–3,096,322	
	ME (345)			3095975–3,096,319	
10	2,718,370(936)	379,000–379,037	38	3,097,370–3,097,407	$g \to a$
		379,039–379,203	165	3,097,409–3,097,573	$c \to t$
		379,205–379,435	231	3,097,575–3,097,805	$c \to t$
		379,437–379,615	179	3,097,807–3,097,985	$a \to g$
		379,617–379,630	14	3,097,987–3,098,000	$t \to c$
		379,632–379,678	47	3,098,002–3,098,048	$a \to g$
		379,680–379,922	243	3,098,050–3,098,292	$c \to t$
		379,924–379,942	19	3,098,294–3,098,312	
	ME (947)			3,097,363–3,098,309	

in Fig. 5(a), (520,876–521,321) in Fig. 5(c), (573,394–574,580), (1,200,307–1,201,479) in Fig. 5(d), (1,483,390–1,484,062) and (1,614,619–1,615,832), which are denoted by zones 1, 3, 4, 5, 6 and 7, respectively. On the one hand, the zones 1, 4, 5, and 6 cover the coding regions (52,260–53,108), (573,394–574,580), (1,200,307–1,201,479) and (1,483,390–1,484,062) of the genes *sll*1397, *slr*2036, *sll*1780 and *ssr*2898, respectively. The zone 3 covers the coding region (520,918–521,169) of the gene *ssl*1922 and overlaps with the coding region (521,076–521,357) of the gene *ssl*1920. The zone 7 overlaps with the coding region (1,614,572–1,615,648) of the gene *slr*522. On the other hand, the nucleotide strings in the zones are also transferred to cover or overlap with coding regions of several genes in the genome. In the same way, the nucleotide strings at the transfer positions in the 6 zones are almost identical the mobile elements.

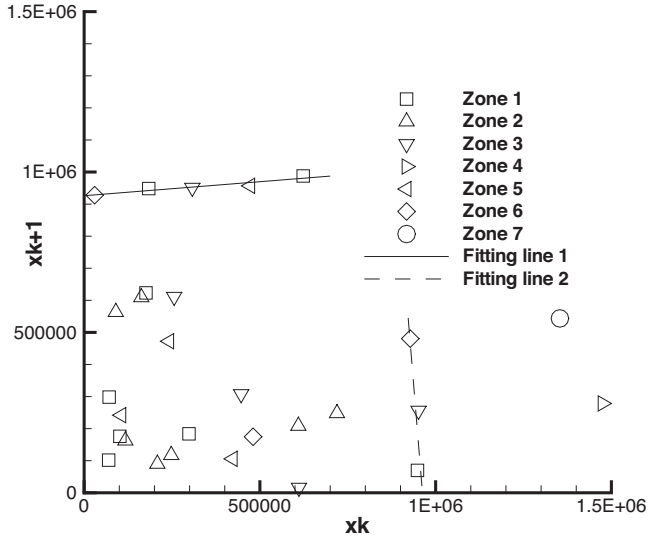

Fig. 6. Two-dimensional vectors in the phase space reconstructed from the iterative transfer distances in the seven zones.

For the non-periodic transfer in each zone, firstly, iterative transfer distance x_k is determined by using Eq. (4). Two-dimensional reconstructed vectors $\mathbf{y}_k$ generated by using Eq. (5) are then drawn in Fig. 6. In the genome, the maximal distance in the non-periodic transfer is about 1.4×10^6 in the zones 4 and 7. Any transfer with a smaller distance disappears before it. The second maximal distance in the non-periodic transfer is about 9.5×10^5 in the zones 1, 2, 3 and 6. It follows the transfer with a smaller distance and then goes on or stops due to its position in the genome. It is evident that some points in the reconstructed phase space are situated along a line, demonstrating linear dependence. By using the least square method, the fitting lines 1 and 2 drawn in Fig. 6 are determined as $x_{k+1} = 954,438 + 0.0874(x_k - 323,449)$ and $x_{k+1} = 268,827 + 13.47(x_k - 942,503)$, respectively. Line 1 reflects two steps for the continuous iterative transfer until the second maximum. Line 2 describes the transfer for a departure from the second maximum. The two fitting lines probably imply an intrinsic dynamics in the transfer of nucleotide strings.

4. Conclusion and discussions

In summary, by using the recurrence plot method and the phase space reconstruction technique, we have investigated the transfer properties of nucleotide strings in the *synecho* genome and demonstrated the presence of periodic and non-periodic correlation structures. The periodic correlation structures are generated by periodic transfer of several substrings in long periodic or

non-periodic nucleotide strings embedded in the coding regions of genes.

The non-periodic correlation structures are generated by non-periodic transfer of several substrings covering or overlapping with the coding regions of genes. In the periodic and non-periodic transfer, some gaps divide the long nucleotide strings into the substrings and prevent their global transfer. Most of the gaps are either the replacement of one base or the insertion/reduction of one base. In the reconstructed phase space, the points generated from two or three steps for the continuous iterative transfer via the second maximal distance can be fitted by two lines. It partly reveals an intrinsic dynamics in the transfer of nucleotide strings. Due to the comparison of the relative positions and lengths, the substrings concerned with the non-periodic correlation structures are almost identical to the mobile elements annotated in the genome. The mobile elements have been thus endowed with the basic results on the correlation structures.

Although the repeats of nucleotide strings in the genome may be determined by the general *k*-mer method, the correlation analysis can reflect the relative positions among the repeated nucleotide strings in the genome and their internal structures generated by gaps. The periodic and non-periodic structures in the cording and non-cording regions of the genome revealed by the correlation analysis may relate to the heredity and variance of the cells: the transfer of continuous/interrupted nucleotide strings in the genome keeps/changes the nucleotide composition for the heredity/variance. Moreover, the junk DNA of a genome includes of many transferable elements in non-cording regions. Its unknown biological functions in cells are covered by the mystery of transfer of the elements in the whole genome. The proposed periodic and non-periodic correlation structures may have fundamental importance for the biological functions of the junk DNA.

Acknowledgements

We would like to thank the National Science Foundation for partial support through the Grant No. 11172310 and the IMECH/SCCAS SHENTENG 1800/7000 research computing facilities for assisting in the computation.

References

Bennetzen, J.L., 2000. Transposable element contributions to plant gene and genome evolution. Plant Mol. Biol. 42, 251–269.
Bergman, C.M., Quesneville, H., 2007. Discovering and detecting tranposabe elements in genome sequences. Brief. Bioinform. 8, 382–392.
Bhaya, D., Vaulot, D., et al., 2000. Isolation of Regulated Genes of the Cyanobacterium *Synechocystis* sp. Strain PCC 6803 by Differential Display. J. Bacteriol. 182, 5692–5699.
Cao, Y., Tung, W.-W., Gao, J.B., 2005. Recurrence time statistics: versatile tools for genomic DNA sequence analysis. J. Bioinform. Comput. Biol. 3, 677–696.
Conte, E., Conte, S., Giuliani, A., 2012. Identification of possible differences in coding and non coding fragments of DNA sequences by using the method of the recurrence quantification analysis. Int. J. Res. Rev. Appl. Sci. 13, 370–397.
Delihas, N., 2011. Impact of small repeat sequences on bacterial genome evolution. Genome Biol. Evol. 3, 959–973.
Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B., 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. 16, 1391–1399.
Eckmann, J.-P., Kamphorst, S.O., Ruelle, D., 1987. Recurrence plots of dynamical systems. Europhys. Lett. A 4, 973–977.
Feschotte, C., Jiang, N., Wessler, S.R., 2002. Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. 3, 329–341.
Frahm, K.M., Shepelyansky, D.L., 2012. Poincaré recurrences of DNA sequences. Phys. Rev. E 85, 016214.
Fu, P., 2009. Genome-scale modeling of *Synechocystis* sp. PCC 6803 and prediction of pathway insertion. J. Chem. Technol. Biotechnol. 84, 473–483.
Garte, S., 2004. Fractal properties of the human genome. J. Theor. Biol. 230, 251–260.
Hao, B.-L., 2000. Fractals from genomes – exact solutions of a biology-inspired problem. Physica A 282, 225–246.
Herzel, H., Gloße, I., 1995. Measuring correlations in symbol sequences. Physica A 216, 518–542.
Herzel, H., Weiss, O., Trifonov, E.N., 1999. 10–11 bp periodicities in complete genomes refelct protein structure and DNA folding. Bioinformatics 15, 187–193.
Holste, D., Grosse, I., Beirer, S., Schieg, P., Herzel, H., 2003. Repeats and correlations in human DNA sequences. Phys. Rev. E 67, 061913.
Hong, S.-J., Lee, C.-G., 2007. Evaluation of central metabolism based on a genomic database of *Synechocystis* PCC6803. Biotechnol. Bioprocess Eng. 12, 165–173.
Jeffrey, H.J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18, 2163–2170.
Kandiah, V., Shepelyansky, D.L., 2013. Google matrix analysis of DNA sequences. PLOS ONE 8, e61519.
Kaneko, T., Sato, S., et al., 1996. Sequence analysis of the genome of the unicellular Cyanobacterium *Synechocystis* sp. Strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3, 109–136.
Karlin, S., Mrázek, J., Campbell, A.M., 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. 179, 3899–3913.
Katani, H., Tabata, S., 1999. Lessons from sequencing of the genome of a unicellular Cyanobacterium, *Synechocystis* sp. PCC6803. Annu. Rev. Plant Physiol. Plant Mol. Biol. 49, 151–171.
Katsaloulis, P., Theoharis, T., Zheng, W.-M., Hao, B.-L., Bountis, A., Almirantis, Y., Provata, A., 2006. Long-range correlations of RNA polymerase II promoter sequences across organisms. Physica A 366, 308–322.
Kazazian, H.H., 2004. Mobile elements: drivers of genome evolution. Science 303, 1626–1632.
Knoop, H., Zilliges, Y., Lockau, W.R., 2010. Steuer The metabolic network of *Synechocystis* sp. PCC 6803: systemic properties of autotrophic growth. Plant Physiol. 154, 410–422.
Kucho, K., Okamoto, K., et al., 2005. Global analysis of circadian expression in the Cyanobacterium *Synechocystis* sp. strain PCC 6803. J. Bacteriol. 187, 2190–2199.
Lönnig, W.-E., Saedler, H., 2002. Chromosome rearrangements and transposable elements. Annu. Rev. Genom. 36, 389–410.
Li, W., 1990. Mutual information functions versus correlation functions. J. Stat. Phys. 60, 823–837.
Li, W., Kaneko, K., 1992. Long-range correlation and partial $1/f$ spectrum in a non-coding DNA sequence. Europhys. Lett. 17, 655–660.
Messer, P.W., Arndt, P.F., Lässig, M., 2005. Solvable sequence evolution models and genomic correlations. Phys. Rev. Lett. 94, 138103.
Mitschke, J., Georg, J., et al., 2011. An experimentally anchored map of transcriptional start sites in the model Cyanobacterium *Synechocystis* sp. PCC6803. Proc. Natl. Acad. Sci. U. S. A. 108, 2124–2129.
Mount, D., 2004. Alignment of pairs of sequences. In: Bioinformatics: Sequence and Genome Analysis, 2nd ed. Cold Spring Harbor Lab. Press, New York (Chapter 3).
Packard, N.H., Crutchfield, J.P., Farmer, J.P., Shaw, R.S., 1980. Geometry from a time series. Phys. Rev. Lett. 45, 712–716.
Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., Stanley, H.E., 1992. Long-range correlations in nucleotide sequence. Nature 356, 168–170.
Peyrard, M., 2004. Nonlinear dynamics and statistical physics of DNA. Nonlinearity 17, R1–R40.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.
Tajima, N., Sato, S., Maruyama, F., et al., 2011. Genomic structure of the Cyanobacterium *Synechocystis* sp. PCC 6803 Strain GT-S. DNA Res. 19, 393–399.
Thiel, T., 1994. Genetic analysis of cyanobacteria. In: Bryant, D.A. (Ed.), The Molecular Biology of the Cyanobacteria. Kluwer Academic Press, Dordrecht, The Netherlands, pp. 581–611.
Wu, Z.-B., 2000. Metric representation of DNA sequences. Electrophoresis 21, 2321–2326.
Wu, Z.-B., 2004. Recurrence plot analysis of DNA sequences. Phys. Lett. A 232, 250–255.
Wu, Z.-B., 2013. Periodic correlation structures in bacterial and archaeal complete genomes. Curr. Bioinform. 8, 267–274.