Palindrome-like Patterns in Genomes

Arthur Chun-Chieh Shih[1], D. T. Lee [1,2] , Chi-Fang Chin[1], Hong-Yuan Mark Liao[1], and Wen-Hsiung Li[2,3]

[1]Institute of Information Science and [2]Genome Research Center, Academia Sinica, Taiwan
[3]Departmnet of Ecology and Evolution, University of Chicago, Chicago, IL, 60637 USA

ABSTRACT

Motivation: It has been of great interest to know whether some simple nucleotide-level patterns or permutation structures can give rise to genome-wide general properties. In this study, we wish to find such patterns or structures. For this purpose, we develop a representation method different from those used to detect long range correlations (LRCs) in genomes. Specifically, we propose a set of crystalline-like walking spaces in which DNA sequences can be coded into a set of 3D curves by following a set of walking rules.

Results: Our representation method reveals that genomic DNA sequences in both prokaryotes and eukaryotes have linear trends and discernable orientations in the walking spaces. Some palindrome-like sequences are then found to have the trends and orientations similar to those found in genomic DNA sequences. After a series of verifications and simulations, we speculate that these patterns may be derived from a very old ancestral genome and some of the patterns might still possess certain functions in present-day genomes. These patterns may explain the existence of LRCs in genomes and why Chargaff's second parity rule holds approximately in many genomic regions.

Availability: The supplementary material can be found at http://www.iis.sinica.edu.tw/~arthur/INPR_in_Genomes.

INTRODUCTION

Although the genomes of human and many other organisms have been sequenced, we still cannot explain many genome-wide properties such as Chargaff's 2nd parity rule (Chargaff, 1951; Rudner et al., 1968) and long range correlations (Peng et al., 1992a; Buldyrev et al., 1993; Karlin and Brendel, 1993; Peng et al., 1994; Buldyrev et al., 1995; Li, 1997; Herzel *et al*., 1999).

Chargaff and his colleagues proposed two parity rules for DNA sequences (Chargaff, 1951; Rudner et al., 1968). The first rule stated that the frequencies of guanine (G) and adenine (A) are equal to those of cytosine (C) and thymine (T), respectively, in double-stranded DNA sequences. In 1953, Waston and Crick proposed the double-helix DNA model, providing an elegant explanation for the first parity rule. The second parity rule postulated that the first parity rule also applies to single-strand DNA (Chargaff, 1951; Rudner et al., 1968). Three decades later, Prabhu (1993) and Qi and Cuitcchia (2001) found that the frequencies of certain oligonucleotides closely approximate to those of their complementary ones in single-strand DNA. Although Forsdyke (1995) and Forsdyke and Mortimer (2000) suggested that the existence of the second parity rule for single stranded DNA may arise from intra-strand base pairing, concrete evidence for their hypothesis is still lacking.

The discovery of long range correlations (LRCs) in genomes (Peng et al., 1992a; Buldyrev et al., 1993; Karlin and Brendel, 1993; Peng et al., 1994; Buldyrev et al., 1995; Li, 1997; Herzel et al., 1999) has received much attention in the past decade. Studies based on various 1-D or 2-D mapping rules (Hamori and Ruskin, 1983; Gates, 1985; Hamori, 1985; Berthelsen et al., 1992; Buldyrev et al., 1995; Lobry, 1996b) revealed certain LRCs between nucleotides in the genomic sequences of various species and in different genomic regions. Although some speculations have been proposed to explain the observation in biological sense, the cause of LRCs remains controversial (Li and Kaneko, 1992; Peng et al., 1992b; Karlin and Brendel, 1993; Ohno, 1993; Peng et al., 1994; Li, 2002; Beirer, 2003).

In this paper, we are interested in whether some simple nucleotide-level patterns or permutation structures exist in genomes that give rise to genome-wide general properties. In particular, we

speculate that some palindrome-like patterns exist or ever existed in genomes. We develop a method that is different from those used to detect LRCs in genomes (Hamori and Ruskin, 1983; Gates, 1985; Hamori, 1985; Li and Kaneko, 1992; Peng et al., 1992b). In this method, we propose a set of crystalline-like walking spaces in which DNA sequences can be coded into a set of 3D curves by following a set of walking rules. Our method reveals that genomic DNA sequences have linear trends and discernable orientations in the walking spaces. We then find some palindrome-like sequences which display the trends and orientations very similar to those found in genomic DNA sequences. We have conducted a series of verifications and simulations, and speculate that these patterns may be derived from a very old ancestor and some of the patterns might still possess a certain function in present-day genomes. Furthermore, these patterns may explain Chargaff's second parity rule and LRCs in genomes.

METHODS

Construction of Crystalline-like Walking Spaces

For the four elements A, C, G, and T, there are four basic structures with one of the four elements as the central node (Fig. 1a). Let $\boldsymbol{S}$ be a DNA sequence, L be the length of $\boldsymbol{S}$, and S[i] be the ith element of $\boldsymbol{S}$. For an element S[i] in $\boldsymbol{S}$, the next element S[i+1] can be A, C, G, or T. The four potential relations between S[i] and S[i+1] define the four basic connection structures as shown in Fig. 1a; in each of them, the central node represents S[i], and the three neighboring nodes and the central node itself denote the possible elements for S[i+1]. Each of these basic connection structures is called a repeating unit and a tessellation of repeating units in the plane is called a walking plane. A walking space is defined by a walking plane, i.e., the (x,y)-plane, and an index axis, the z axis, which represents element positions in the DNA sequence. In view of its regular and uniform

structure similar to a crystal structure, a walking space is referred to as a *crystalline-like walking space* (CWS). Fig. 1b shows three CWSs that are correlated but cannot be obtained from one another by any linear transformation. There are other possible crystalline walking spaces, but all of them are the linear transformations of these three CWSs. We also define step-move vectors to characterize the walking track of a DNA sequence in a CWS. Let $\vec{p}(i) = (x(i), y(i), i)$, where $i \geq 1$, be the coordinates of node i in a walking space and $\Delta\vec{p}(i-1)$ be the step-move vector from node $\vec{p}(i-1)$ to node $\vec{p}(i)$ in the walking space, so that $\vec{p}(i) = \vec{p}(i-1) + \Delta\vec{p}(i-1)$. For example, if S[$i$-1]=A and S[$i$]=T, then in $CWS_{GC\text{-}AT}$ the step-move vector that moves from A to T is (0,1).

Characterization of a Walking Space

Consider a DNA sequence ***S*** of length L, and let S[i], $1 \leq i \leq L$, denote the i^{th} character in ***S***. If the starting point of ***S*** in the walking track is located at the origin (x(0), y(0), 0), then a 3-D ordered walking track, or trajectory, for ***S*** can be defined as $\{ \vec{p}(i) = (x(i), y(i), i) \mid 1 \leq i \leq L\}$. The pattern obtained by projecting the walking track of ***S*** onto the walking plane (the x-y plane) is called the "*configuration*" of ***S***.

Let $\overline{S}$ denote the complementary strand of ***S***. We call the configurations of $\overline{S}$ the complementary configurations of ***S***. Although the shapes of the primary and complementary configurations are the same, their orientations are different. In $CWS_{GC\text{-}AT}$, the orientations of the configuration of ***S*** and its complementary configuration differ by 180^o, while in $CWS_{CT\text{-}AG}$ and $CWS_{TG\text{-}CA}$, the orientations of the configurations of ***S*** and its complementary configurations differ by not only 180^o but also a mirror or reflection transformation. Note that if the trends and orientations of both configurations point in the right-up or left-down diagonal directions, we need

not use two strands of DNA sequences to do the analysis because the results will be the same. For example, if *S*=5'ACTGCAG 3', its complementary strand is $\overline{S}$ =5'CTGCAGT3'. As shown in Fig. 2, when the configuration of $\overline{S}$ is rotated by 180^o, its shape is exactly the same as that of S; the only difference is that the orientation is reversed.

In a walking space, a discernable trend and orientation of the configurations may imply that the dinucleotide orderings of the DNA sequence follow some specific permutations. Because the step-move vectors in each CWS can be divided into vertical, horizontal and diagonal moves, the walking track can be extended with a discernable trend and orientation that is governed by permutations of the vertical/horizontal step-moves. Similar to analyzing a function of multiple variables, we can directly observe the permutations of four of the 16 possible dinucleotides in a long sequence by mapping it onto a CWS. The ordering distributions of these dinucleotides will also be reflected in the configurations in CWSs.

Furthermore, if one character is changed or deleted, the neighboring transitional status may also be changed and the configurations in all CWSs can also be changed. This high correlation of the configurations in different CWSs can be used to detect some regular dinucleotide permutations hidden in a sequence if the obtained configurations have strong linear trends and discernable orientations.

RESULTS

In what follows, all contigs, chromosomes, and genomes used were downloaded from GenBank (http://www.ncbi.nim.nih.giv) as of May 2004.

Walking Tracks of Genomic Sequences in Human, *Arabidopsis*, and *E. coli*

First, as an example of application, NT_011362, a contig of human chromosome 20 with 24,982,240 bp (24.98 Mb), was selected. Fig. 3 shows the walking tracks of the sequence in $CWS_{GC\text{-}AT}$ at different scales (window sizes) and their corresponding two-dimensional configurations. Note that at the full-length scale the configurations in $CWS_{GC\text{-}AT}$, $CWS_{CT\text{-}AG}$, and

$CWS_{TG\text{-}CA}$ share similar orientation tendencies (Fig. 4). Second, other human DNA contigs were also considered and their configurations can be found in our Website (http://www.iis.sinica.edu.tw/~arthur/INPR_in_Genomes). The configurations for these contigs all possess similar orientation tendencies.

Since interspersed repetitive sequences account for over 40% of the human genome, we ask whether the regularities could be due to these repeats. As a test, we masked the repeats and low-complexity regions in the contig used in Fig. 4 using RepeatMasker (Smit and Green, 1997), but found that the resulting configurations also possess similar tendencies (Fig. 5). Therefore, interspersed repeats do not seem to have a strong effect on the configurations.

Note further that the configurations of a randomly generated sequence (2 Mb) in $CWS_{GC\text{-}AT}$, $CWS_{CT\text{-}AG}$, and $CWS_{TG\text{-}CA}$ look like random walks, showing no discernable tendency (Fig. 6).

For characterizing the differences between a discernable orientation and a "kinky" one, we define three measures to quantify the configurations in different CWSs: the consistency angle $\theta(i,w)$, the persistence length $r(i,w)$, and the persistence angle $\phi(i,w)$, where w is a given window size. For a specific $CWS_{xx\text{-}xx}$, where *xx-xx* represents GC-AT, CT-AG, or TG-CA, let $\vec{p}(i)$ be the point of the trajectory corresponding to S[i] and w be a given window size (w>0). For all $i \geq w$, $\vec{p}(i-w)$ and $\vec{p}(i+w)$ are two points on the trajectory w elements before and after S[i], respectively. Let $\vec{p}_{i-w,i}$, $\vec{p}_{i,i+w}$, and $\vec{p}_{i-w,i+w}$ be the vectors from $\vec{p}(i-w)$ to $\vec{p}(i)$, from $\vec{p}(i)$ to $\vec{p}(i+w)$, and from $\vec{p}(i-w)$ to $\vec{p}(i+w)$, respectively. We define $\theta(i,w)$, $r(i,w)$, and $\varphi(i,w)$ as follows:

$$\theta(i,w) = \cos^{-1}\left(\frac{\vec{p}_{i-w,i} \cdot \vec{p}_{i,i+w}}{\| \vec{p}_{i-w,i} \| \| \vec{p}_{i,i+w} \|}\right), \quad (1)$$

$$r(i,w) = \frac{1}{2w+1} \| \vec{p}_{i-w,i+w} \|, \quad (2)$$

and

$$\phi(i,w) = \cos^{-1}\left(\frac{\vec{p}_{i-w,i+w} \cdot \vec{e}_X}{\| \vec{p}_{i-w,i+w} \|}\right), \quad (3)$$

where $\vec{e}_X = (1,0)$ and **A•B** represents the inner product of vectors **A** and **B.** Note that $\theta(i,w)$ denotes the angle between vectors $\vec{p}_{i-w,i}$ and $\vec{p}_{i,i+w}$. If $\vec{p}(i-w)$, $\vec{p}(i)$ and $\vec{p}(i+w)$ share a discernable

orientation tendency, the value of $\theta(i,w)$ will be close to 0^o, whereas if $\vec{p}(i-w)$, $\vec{p}(i)$ and $\vec{p}(i+w)$ are randomly distributed as a coiling or kinky track, the value of $\theta(i,w)$ will, on average, be close to 90^o, because the possible values of $\theta(i,w)$ range from 0^o to 180^o. In addition, $r(i,w)$ and $\phi(i,w)$ are used to estimate, respectively, the persistence length and the direction of the orientation tendency with respect to the *x*-axis. If the orientation tendency is close to a straight line, the value of $r(i,w)$ will be close to unity and the value of $\phi(i,w)$ indicates the direction of $\vec{p}_{i-w,i+w}$. However, if the orientation is coiling or kinky, the estimated $r(i,w)$ will be close to a small nonzero value and its $\phi(i,w)$ is meaningless. The averages of $\theta(i,w)$, $r(i,w)$, and $\phi(i,w)$ of a configuration in $CWS_{xx\text{-}xx}$ are defined as follows,

$$\bar{\theta}_{xx-xx}(w) = \frac{1}{L-2w}\sum_{w\le i\le L-w}\theta(i,w), \tag{4}$$

$$\bar{r}_{xx-xx}(w) = \frac{1}{L-2w}\sum_{w\le i\le L-w} r(i,w), \tag{5}$$

$$\text{and} \quad \bar{\phi}_{xx-xx}(w) = \frac{1}{L-2w}\sum_{w\le i\le L-w}\phi(i,w), \tag{6}$$

where L is the length of the sequence. $\bar{\theta}_{xx-xx}(w)$, $\bar{r}_{xx-xx}(w)$, and $\bar{\phi}_{xx-xx}(w)$ are called the *average consistency angle*, the *average persistence length*, and the *average persistence angle*, respectively.

Fig. 7 shows the features of the configurations of NT_011362. The average angles $\bar{\theta}_{GC-AT}(w)$ and $\bar{\theta}_{TG-CA}(w)$ drop quickly to below 10^o when the window size w increases to 2,000 bp, whereas $\bar{\theta}_{CT-AG}(w)$ decreases to the same degree only when w increases to ~10,000 bp. This difference reflects the fact that the configurations of human sequences in $CWS_{CT\text{-}AG}$ are twisted, whereas those in $CWS_{GC\text{-}AT}$ and $CWS_{TG\text{-}CA}$ are not. On average, the values of $\bar{\theta}_{CT-AG}(w)$ are twice as large as the values of either $\bar{\theta}_{GC-AT}(w)$ or $\bar{\theta}_{TG-CA}(w)$. For all CWSs, the values of $\bar{r}_{GC-AT}(w)$, $\bar{r}_{TG-CA}(w)$, and $\bar{r}_{CT-AG}(w)$ approach a constant as w increases beyond 2,000 bp. Interestingly, the values of $\bar{r}_{GC-AT}(w)$ and $\bar{r}_{TG-CA}(w)$ are very close to each other, whereas those of $\bar{r}_{CT-AG}(w)$ are smaller by

a factor of 2 for the same *w*. For the average persistence angles with *w* larger than 5,000 bp, the values of both $\bar{\phi}_{CT-AG}(w)$ and $\bar{\phi}_{TG-CA}(w)$ are estimated to be 45^{o}, while the values of $\bar{\phi}_{GC-AT}(w)$ are estimated to be only ~ 20^{o}.

We downloaded additional 352 contigs of the human genome whose sizes are larger than 2 Mb; the total length of these contigs is over 2.2 billion bp. Using 1,000 bp as the window size, we calculated $\bar{r}_{xx-xx}$ and $\bar{\phi}_{xx-xx}$ of the configurations of each contig in $CWS_{GC\text{-}AT}$, $CWS_{CT\text{-}AG}$, and $CWS_{TG\text{-}CA}$. Then, each pair of $\bar{r}_{xx-xx}$ and $\bar{\phi}_{xx-xx}$ was transformed into a Cartesian coordinate system $(\bar{x}_{xx-xx}, \bar{y}_{xx-xx})$ (Fig. 8). Note that both the distributions of $(\bar{x}_{\text{TG-CA}}, \bar{y}_{\text{TG-CA}})$s and $(\bar{x}_{\text{CT-AG}}, \bar{y}_{\text{CT-AG}})$s spread out along a sloping line, while that of $(\bar{x}_{\text{GC-AT}}, \bar{y}_{\text{GC-AT}})$s spreads out in the horizontal direction. This means that in $CWS_{TG\text{-}CA}$ and $CWS_{CT\text{-}AG}$, the variations in the configurations of human contigs depend only on the persistence length, whereas the variations in $CWS_{GC\text{-}AT}$ depend on both the persistence length and the angle.

For comparison, the genomes of *E. coil K12* and *Arabidopsis* were used to see whether any specific orientation tendencies exist in these distantly related organisms. Fig. 9a and b show that the configurations of *E. coil K12* and *A. thaliana* chromosome I in different CWSs also have discernable orientation tendencies. Interestingly, in $CWS_{CT\text{-}AG}$ the orientations of the configurations in both genomes point toward the lower left direction, while those of the human genome all point toward the upper right direction. Moreover, in $CWS_{GC\text{-}AT}$, the slopes of the orientation tendencies in the genomes of *E. coil K12* and *A. thaliana* are different from those in the human genome.

Fig. 10a and b show some quantitative analyses of the configurations in Fig. 9a and b. Compared with the characteristic curves in the human genome, all average persistence lengths ($\bar{r}_{GC-AT}(w)$, $\bar{r}_{TG-CA}(w)$, and $\bar{r}_{CT-AG}(w)$) of both *E. coil K12* and *A. thaliana* chromosome I are less than those of human DNA sequences by a factor of 2. Moreover, the two persistence angles $\bar{\phi}_{CT-AG}(w)$ and $\bar{\phi}_{TG-CA}(w)$ of *E. coil K12* and *A. thaliana* are close to 45^{o} when the window

size reaches 20,000 bp. Actually, the $\bar{\phi}_{CT-AG}(w)$ of *E. coil K12* and *A. thaliana* should be close to 225^o because the orientation of the configurations points toward the lower left direction. Note also that the $\bar{\phi}_{GC-AT}(w)$s of these two species are close to 70^o, while those of human sequences are close to 20^o. Thus, the values of $\bar{\phi}_{GC-AT}(w)$ can be similar or very different between species. Table 1 shows the summaries of the directions of the orientation tendencies for these genomes in the three CWSs considered.

Coding and noncoding regions in genomes are supposed to evolve at different rates and by different evolutionary mechanisms (Li, 1997). For *E. coli,* a large proportion of the genome is coding. We can consider that the linear trends and orientations are independent of coding and noncoding regions in *E. coli*. However, for other eukaryotic species, most regions are noncoding and we are interested in seeing whether coding regions in the *human* and *A. thaliana* genomes also possess the linear properties.

Because a single coding sequence is usually less than 1,000 bp long, it is difficult to estimate the global trend and orientation. We therefore concatenate the coding sequences together to gain enough sequence lengths for analysis. Since the orientations of coding regions in each strand are usually inconsistent, we generate two sets of sequences: (1) sequences that are concatenated from the coding sequences in the same orientations and (2) sequences that are concatenated from the coding sequences in the same strand (Table 2). Interestingly, for the first set of sequences the persistent angles, $\phi_{TG\text{-}CA}$s and $\phi_{CT\text{-}AG}$s, for the human, *A. thaliana* and *E. coli* genomes are quite different from those in Table 1, whereas for the second set, the angles are similar to those in Table 1. Thus, the coding sequences in the same strand also show the linear trend and orientations similar to those in the noncoding regions for the human and *A. thaliana* genomes.

The observations up to now can be summarized as follows:

(a) *Linear orientation tendency*

The configurations of the DNA sequences we examined possess linear orientation tendencies in

$CWS_{\mathbf{CT\text{-}AG}}$, $CWS_{\mathbf{TG\text{-}CA}}$, and $CWS_{\mathbf{GC\text{-}AT}}$, which are independent of interspersed repeats and strands. Moreover, since the examined species belong to three different kingdoms (bacteria, plants, and animals), the linear tendencies are also independent of nucleosomal structure (Audit et al., 2001; Li, 2002) and compositional isochoric structure.

(b) *Direction of tendency*

The orientation tendencies for all sequences in $CWS_{TG\text{-}CA}$ are along the upper-right diagonal direction; that is, their persistence angles are all close to 45^o. But in $CWS_{CT\text{-}AG}$, the orientation tendencies for the *E. coli* and *A. thaliana* sequences point in the opposite direction of those for human sequences. Moreover, the orientation tendencies for human sequences in $CWS_{GC\text{-}AT}$ approach 20^o, while those for the *E. coli* and *A. thaliana* sequences are approximately 70^o.

(c) *Persistence ratio of the tendency in* CWSs

Table 1 shows also the average persistence lengths of the configurations for human, *E. coli K12*, and *A. thaliana* genomes in different CWSs; the window size is 50,000 bp. For all configurations of the genomes in CWSs, the average persistence lengths for human contigs are the longest. Those for *E. coli K12* are shorter than those for human but longer than those for *A. thaliana*. However, the ratio among $\bar{r}_{GC-AT}$, $\bar{r}_{TG-CA}$, and $\bar{r}_{CT-AG}$ is approximately constant and is independent of the species. For the genomes of human, *A. thaliana*, and *E. coli*, the ratios of their $\bar{r}_{GC-AT}$, $\bar{r}_{TG-CA}$, and $\bar{r}_{CT-AG}$ are 2.22**:**2.08**:**1.00, 2.47**:**2.28**:**1.00, and 2.42**:**2.25**:**1.00, respectively. The fact that these ratios were very similar for different species prompted us to study them further.

We examined further DNA sequences from other species, including some bacteria, worm, and fruit fly genomes. The related average persistence angles and the average persistence lengths for these sequences are shown in Table 3. Generally speaking, the configurations of these species all possess fixed linear trends and discernable orientations in CWSs. These regular properties widely exist among various species but whether this phenomenon is coincidental remains to be seen.

Relationship between Strand Asymmetry and Dinucleotides

As seen above, the configurations of *E. coli* in different CWSs all possess linear trends and discernible orientations and the complementary configurations also possess the same properties because all of their orientations point in the diagonal directions. Theses trends and orientations seem to be strand-independent. However, in the *E. coli* genome the numbers of Gs and Cs within a fixed window (≈ 50 kb) are known to be strand asymmetric (Lobry, 1996a; Francino and Ochman, 1997; Freeman *et al.*, 1998b). In the above analyses, the dinucleotides AA, TT, GG, and CC were not considered because these dimers do not influence the configurations in CWSs. In what follows we study what role these dinucleotides play and whether they are related to strand asymmetry.

In Karlin and his colleagues' studies, no signature differences were observed between pairs of complementary dinucleotide relative abundances across the genomes (Karlin, 1999; Karlin et al., 2002). They concluded that the constancy of dinucleotide abundances relative to the two strands is consistent with the constancy of the genome signature. The same observation was made by Shioiri and Takahata (2001). They also found that no strand bias exists in prokaryote genomes and that the trend holds true even for tri- or tetra-nucleotides. In their analyses, however, only the magnitudes of the dinucleotide relative abundances were considered; the accumulation and permutations of the dinucleotides were not considered. In what follows, we shall explore whether the dinucleotides in bacterial genomes are indeed strand-independent.

Considering the numerator in the formula for GC skew (Blattner et al., 1997), we have

$$S_G - S_C = S_{GG}+S_{GC}+S_{GA}+S_{GT} - (S_{CC}+S_{GC}+S_{TC}+S_{AC})$$

$$= (S_{GG} - S_{CC}) + (S_{GA} - S_{TC}) + (S_{GT} - S_{AC}),$$

(7)

where S_{XX} represents the number of dinucleotides XX within a fixed window size. A cumulative GC skew curve $H_{G\text{-}C\ skew}(i)$ of a DNA sequence up to $\mathbf{S}(i)$ is defined as

$$H_{G-C\ skew}(i) = \begin{cases} H_{G-C\ skew}(i-1)+1, & \text{if } S(i) \text{ is a G}, \\ H_{G-C\ skew}(i-1)-1, & \text{if } S(i) \text{ is a C}, \\ H_{G-C\ skew}(i-1), & \text{otherwise}, \end{cases} \quad (8)$$

where $H_{G\text{-}C\ skew}(0) = 0$ (Freeman *et al.*, 1998b). We also define three cumulative dinucleotides skew functions $H_{GG\text{-}CC\ skew}(i)$, $H_{GA\text{-}TC\ skew}(i)$, and $H_{GT\text{-}AC\ skew}(i)$ as follows:

$$H_{PQ-XY\ skew}(i) = \begin{cases} H_{PQ-XY\ skew}(i-1)+1, & \text{if } S(i) = P \text{ and } S(i+1) = Q \\ H_{PQ-XY\ skew}(i-1)-1, & \text{if } S(i) = X \text{ and } S(i+1) = Y \\ H_{PQ-XY\ skew}(i-1), & \text{otherwise}, \end{cases} \quad (9)$$

where $\{PQ, XY\}$ is $\{GG,CC\}$, $\{GA,TC\}$, or $\{GT,AC\}$, and $H_{PQ\text{-}XY\ skew}(0) = 0$. From Eq. (7), $H_{G\text{-}C\ skew}(i)$ is a combination of the cumulative dinucleotides curves $H_{GG\text{-}CC\ skew}(i)$, $H_{GA\text{-}TC\ skew}(i)$, and $H_{GT\text{-}AC\ skew}(i)$:

$$H_{G\text{-}C\ skew}(i) \approx H_{GG\text{-}CC\ skew}(i) + H_{GA\text{-}TC\ skew}(i) + H_{GT\text{-}AC\ skew}(i).$$

If $H_{G\text{-}C\ skew}(i)$ is nonzero, we can have the following equation:

$$\frac{H_{GG-CC\,skew}(i)}{H_{G-C\,skew}(i)} + \frac{H_{GA-TC\,skew}(i)}{H_{G-C\,skew}(i)} + \frac{H_{GT-AC\,skew}(i)}{H_{G-C\,skew}(i)} = k_{GG\text{-}CC}(i) + k_{GA\text{-}TC}(i) + k_{GT\text{-}AC}(i) \approx 1, \quad (10)$$

where $k_{PQ\text{-}XY}(i)=H_{PQ\text{-}XY\ skew}(i)/H_{G\text{-}C\ skew}(i)$, where $\{PQ, XY\}$ is $\{GG,CC\}$, $\{GA,TC\}$, or $\{GT,AC\}$, respectively. If $H_{GG\text{-}CC\ skew}(i)$, $H_{GA\text{-}TC\ skew}(i)$, and $H_{GT\text{-}AC\ skew}(i)$ are linearly correlated with $H_{G\text{-}C\ skew}(i)$, then $k_{GG\text{-}CC}(i)$, $k_{GA\text{-}TC}(i)$ and $k_{GT\text{-}AC}(i)$ will be close to constants for all i's. Thus, to estimate $k_{GG\text{-}CC}(i)$, $k_{GA\text{-}TC}(i)$, and $k_{GT\text{-}AC}(i)$ and calculate their variations we can check whether the cumulative dinucleotides skew functions are correlated with the GC skew function or not.

For calculating these factors, we define the averages of $k_{GC\text{-}AT}(i)$, $k_{GA\text{-}TC}(i)$ and $k_{GT\text{-}AC}(i)$ within a fixed window w ($w > 0$) as follows:

$$\bar{k}_{XX-XX}(i) = \frac{1}{w+1} \sum_{j=i-\frac{1}{2}w}^{j=i+\frac{1}{2}w} k_{XX-XX}(i+j),$$

where $(1/2)\cdot w \le i \le$ **L**$-(1/2)\cdot w$ and *xx-xx* represent GC-AT, GA-TC, and GT-AC, respectively. Using a non-overlapping shifting step (such as 1,000 bp in this paper), we can have three sets of the

factors: $\{\bar{k}_{GC-AT}(i)\text{s}\}$, $\{\bar{k}_{GA-TC}(i)\text{s}\}$, and $\{\bar{k}_{GT-AC}(i)\text{s}\}$. Then, the averages and variances of each set can be calculated. In what follows, we will examine two well-known bacterial genomes with distinguished GC skew properties to see whether the cumulative dinucleotides skew functions are also correlated with the GC skew functions.

Fig. 11a and b show the cumulative skew curves of related dinucleotides in *E. coli K12* and *Bacillus subtilis* (Kunst et al., 1997). Interestingly, for the *E. coli* genome the shapes of $H_{GG\text{-}CC\ skew}$, $H_{GA\text{-}TC\ skew}$, and $H_{GT\text{-}AC\ skew}$ are the same as the purine-excess curve $H_{G\text{-}C\ skew}(i)$ (Lobry, 1996a; Freeman *et al.*, 1998a); only the compositional factor ratios are different. Also, the minima and maxima of $H_{GG\text{-}CC\ skew}$, $H_{GA\text{-}TC\ skew}$, and $H_{GT\text{-}AC\ skew}$ correspond to the origins and the termini of DNA replication, respectively. Fig. 11c and d show the distributions of these compositional factors in these two genomes. In *E. coli*, the factors $k_{GG\text{-}CC}$, $k_{GA\text{-}TC}$, and $k_{GT\text{-}AC}$ are close to 0.5, 0.25, and 0.25, respectively, while in *B. subtilis*, the factors $k_{GG\text{-}CC}$, $k_{GA\text{-}TC}$, and $k_{GT\text{-}AC}$ approach 0.5, 0.5, and 0.0, respectively. In contrast to $H_{GG\text{-}CC\ skew}$, $H_{GA\text{-}TC\ skew}$, and $H_{GT\text{-}AC\ skew}$, the other cumulative dinucleotide curves, such as $H_{GA\text{-}AC\ skew}$, $H_{GG\text{-}TC\ skew}$, $H_{GT\text{-}TC\ skew}$, and $H_{AA\text{-}TT\ skew}$, do not show an entirely linear relationship with $H_{G\text{-}C\ skew}(i)$ (Fig. 11c). Thus, the dinucleotide cumulative curves are correlated with the strand compositional asymmetry and related to the leading and lagging strand of DNA. According to Lobry's hypothesis, a mutational bias is responsible for the strand asymmetry and the two DNA strands undergo unequal patterns of replication error (Lobry, 1996a). Thus, the mechanisms that generate the asymmetry between the two DNA strands in bacterial genomes should also operate at the dinucleotide level. This result is consistent with the observation that mutation is generally context-dependent (Kunkel, 1992; Karlin, 1999). However, since the result of strand asymmetry does not influence the linear trends and discernable orientations of the configurations, we will not discuss this issue further.

Palindrome-like Patterns in Genomes

From the above observations, we conjectured that an intrinsic nucleotide permutation structure exists in each genome, so that the configurations of these DNA sequences possess the linear trends and the specific orientation tendencies in the CWSs.

It is known that duplication, recombination, deletion/insertion, and point mutation are the main mechanisms that produce changes in genomic sequences. Changes caused by these mechanisms are subject to random drift or natural selection or both, but it is inconceivable that the above observed regularities were caused by random drift alone. One possible explanation for the observed phenomena is that parts of the examined genomes were composed of some simple, regular nucleotide permutation structures. Although these structures are subject to the evolutionary forces mentioned above, they are undistorted macroscopically and the genomes retain the linear trend and discernable orientations in different CWSs. In what follows, we will use a computational method to find the speculated patterns and analyze these patterns in different genomes to support our conjecture.

As shown in the previous sections, all of the DNA configurations possess linear trends and some specific angles in the CWSs. We could generate exhaustively all possible sequences of nucleotides of a *fixed* length and find which ones would produce linear trends and angles consistent with those in real DNA sequences. However, how to select a fixed length is very difficult. Based on some observations we found two sequences of 10 nucleotide long (10-mers) that could provide the similar trends and orientations as those for real DNA sequences (Fig. 12). Thus, 10-mers were chosen to be the fixed length for the query sequences in this paper. To exhaustively search for all possible sequences with the same properties, we generated all candidate sequences of 10 nucleotides and calculated the persistence lengths and angles of their configurations in different CWSs. The persistence angles for those possible patterns should satisfy either $\{15^o<\phi_{GC\text{-}AT}<25^o$, $40^o<\phi_{TG\text{-}CA}<50^o$, and $40^o<\phi_{CT\text{-}AG}<50^o\}$ or $\{65^o<\phi_{GC\text{-}AT}<75^o$, $40^o<\phi_{TG\text{-}CA}<50^o$ and $-50^o<\phi_{CT\text{-}AG}<-40^o\}$. In addition, the persistence lengths, $r_{GC\text{-}AT}$, for these patterns should be larger than $r_{CT\text{-}AG}$ and $r_{TG\text{-}CA}$ and should be as close to one as possible. Although the number of all possible candidates is 4^{10},

only 40 sequences have persistence lengths and angles that satisfy either of the above two criteria. All persistence lengths of these sequences are 0.949, 0.849, and 0.424 in the $CWS_{GC\text{-}AT}$, $CWS_{TG\text{-}CA}$, and $CWS_{CT\text{-}AG}$, respectively, with the ratio equal to 2.24:2.00:1.00. This ratio is very close to those for the human, *E. coli*, and *A. thaliana* genomes. Moreover, based on their persistence angles, these sequences can be classified into two groups: $\mathbf{\Gamma_H}$ and $\mathbf{\Gamma_{AE}}$, where H stands for human and AE for *A. thaliana* and *E. coli*. For sequences in $\mathbf{\Gamma_H}$, their persistence angles are 18.44^o, 45^o, and 45^o in the $CWS_{GC\text{-}AT}$, $CWS_{TG\text{-}CA}$, and $CWS_{CT\text{-}AG}$, respectively, while for sequences in $\mathbf{\Gamma_{AE}}$ their persistence angles are 71.57^o, 45^o, and 45^o. Table 4 lists all sequences in $\mathbf{\Gamma_H}$ and $\mathbf{\Gamma_{AE}}$.

For the sequences that include partial repeats or duplicates, such as 5..CATGTGTGCAGCTG..3' or 5..CAAAATTTTGCAGCTG..3', the orientation tendencies are also the same as those in either $\mathbf{\Gamma_H}$ or $\mathbf{\Gamma_{AE}}$, and therefore we propose a 10-node-cyclic-ring structure as a general model to generate the sequences with similar behaviors to those sequences in $\mathbf{\Gamma_H}$ and $\mathbf{\Gamma_{AE}}$. We call this ring the *Intrinsic Nucleotide Permutation Ring* (INPR). In this ring, each node represents a nucleotide (A, T, G or C) and there are three transitional states at each node: going forward (the clockwise direction), going backward (the anti-clockwise direction), and staying put. To generate a sequence, we start with any node in the INPR, record it, and then select a transitional state to go to the next node. However, if we want to generate sequences with the same orientations as the sequences in $\mathbf{\Gamma_H}$ and $\mathbf{\Gamma_{AE}}$, their effective transitional directions should be in forward direction; that is, the total number of forward transitions must exceed that of backward transitions. All possible sequences thus generated are *members* of the INPR, if their effective transitional directions are in forward direction in the INPR. Fig. 13 shows two pairs of the INPRs, where $INPR_H=\{R_{H1},R_{H2}\}$ and $INPR_{AE}=\{R_{AE1},R_{AE2}\}$, from which we can generate the sequences with the same trends and orientations as those in $\mathbf{\Gamma_H}$ and $\mathbf{\Gamma_{AE}}$, respectively. In either $INPR_H$ or $INPR_{AE}$, we find that only two nodes are different in each group and GCAT is the common pattern in all INPRs. Since the persistence lengths for the sequences in either $\mathbf{\Gamma_H}$ or $\mathbf{\Gamma_{AE}}$ are the longest of those generated by the INPRs, we call the sequences in $\mathbf{\Gamma_H}$ and $\mathbf{\Gamma_{AE}}$ the most *Compact Intrinsic Permutation Patterns* (CIPPs).

We calculated the ratios of the average persistence lengths of the human, *A. thaliana*, and *E. coli* genomes over those in the related CIPPs. The ratios indicate the diversities of the three genomes with respect to their corresponding CIPPs. We call these ratios the *pattern appearance ratios* $PAR_{xx\text{-}xx}$, where *xx-xx* represents the walking spaces $CWS_{GC\text{-}AT}$, $CWS_{CT\text{-}AG}$, and $CWS_{TG\text{-}CA}$. $PAR_{CT\text{-}AG}$, $PAR_{TG\text{-}CA}$, and $PAR_{GC\text{-}AT}$ are consistent within each species (Fig. 14). This seems to indicate that the conjectured CIPP for each genome may indeed be responsible for the linear trends and discernable orientations of the configurations in CWSs. Moreover, the $PAR_{xx\text{-}xx}$s also represent the percentages of the CIPP fragments hidden within the examined DNA sequences. Computing the averages of $PAR_{GC\text{-}AT}$, $PAR_{CT\text{-}AG}$, and $PAR_{TG\text{-}CA}$ for each genome, we obtain the average pattern appearance ratios for human, *E. coli*, and *A. thaliana* as 0.1075, 0.07068, and 0.05366, respectively. The *A. thaliana* genome has the smallest ratio for these species, while the ratio for human genome is the largest.

When analyzing $INPR_H$ and $INPR_{AE}$, we found four palindromic sequences in each group: TGCAGCTGCA, TGCTGCAGCA, CTGCATGCAG, and CAGCATGCTG in $\mathbf{\Gamma_H}$ and CATCATGATG, CATGATCATG, TGATGCATCA, and TCATGCATGA in $\mathbf{\Gamma_{AE}}$. Palindromes are very important structural patterns in DNA sequences because they can easily form a secondary structure that might be recognized and bound by some proteins or enzymes. The next question of interest is where in the genomes these *INPR*s are located and whether they have any biological functions.

Location of the *INPR*s in genomes

In each INPR, if we do not limit the lengths, many sequences such as 5'..CACATGTGC..3' and 5'..CAAAATTGGC..3' have the same linear trends and orientations in the CWSs as the sequence 5'..CATGC..3'. To simplify the problem, we consider only the sequences obtained by forward or stationary steps in the INPRs, ignoring those derived from backward moves, followed by forward moves. We used the following two criteria to search for members of each INPR in the human

genome and the *E. coli* and *A. thaliana* genomes: the number of exactly matched nodes is at least N_v and the number of mismatched nodes is at most N_a, while allowing only forward and stationary moves in the INPR. In the following experiments, N_v is set to 11 and N_a is set to either 0 or 1 in different cases. For instance, for INPR $\boldsymbol{R_{H1}}$ (i.e., TGCAGCTGCA) the sequence 5'..$T^+G^+C^+A^+G^+C^+T^+$ $G^+C^+A^+T^+$..3' is a member of $\boldsymbol{R_{H1}}$ satisfying the first criterion (exact match), where X^+ denotes a repeated sequence of X's of length at least 1. Note that the last character must be a repeat of the first in the INPR, when N_v is set to 11. The total length of this sequence is recorded as such. The search results for the human genome and the *E. coli* and *A. thaliana* genomes are shown in Table 5.

In the human genome, the members of $\boldsymbol{R_{H1}}$ and $\boldsymbol{R_{H2}}$ are found to be uniformly distributed among the sequences and the total length is 0.0653% of the genome size when no mismatch is allowed and 0.8415% when one mismatch is allowed. In comparison, the ratios in the *E. coli* and *A. thaliana* are 0.0404% and 0.0661%, respectively, if no mismatch is allowed, and are 0.6761% and 0.8780%, if one mismatch is allowed. Nevertheless, the ratios of the detected *INPR*s in the genomes are all much smaller than the values of their related *PAR*s, ~10% in the humane genome, ~7% in *E. coli*, and ~5% in A. thaliana. So, we wanted to know whether these detected *INPR*s dominate the trends and orientations of the configurations. To answer this question, we removed these detected patterns from the sequences and checked whether the remaining sequences still possess the linear trend and orientation properties. Using human contig NT_011362 as a test sample, we removed the detected sequences of $\boldsymbol{R_{H1}}$ and $\boldsymbol{R_{H2}}$ from the sequence and then searched the sequence for the members of $\boldsymbol{R_{H1}}$ and $\boldsymbol{R_{H2}}$ again. After two iterations of the remove-and-search process, no *INPR*s (length ≥ 11 bp) of the remaining sequence were found. We calculated the persistence lengths and angles of the configurations for the "purified" sequence. Surprisingly, its configurations in each CWS still possess the linear trends and the same orientations. The average persistence angles are very close to those for the original sequence but the average persistent lengths are slightly shorter! This means that most of the *INPR*s are fragmented in the genomes and these fragments still

constitute and dominate the linear trends and orientations.

For estimating how divergent the members of the *INPR*s in the genomes are, we generated a synthetic sequence $\boldsymbol{S_s}$ with 26 Mb containing 2,600,00 consecutive copies of a sequence in $\boldsymbol{\Gamma_H}$, 5'-TGCAGCTGCA-3', and introduced different percentages of mutated sites, from 10% to 100%, to mutate the sequence $\boldsymbol{S_s}$ randomly. Fig. 15a and b show the persistence lengths and angles for the sequences with different percentages of mutated sites in the CWSs. Before the percentage reaching 70%, the persistence angles retain almost the same values in each CWS, whereas the persistence lengths decay in an exponential manner. When the percentage of mutated sites is 67%, the persistence lengths for the sequence, called $S_S^{0.67}$, are 0.100, 0.093, and 0.046 in $CWS_{GC\text{-}AT}$, $CWS_{CT\text{-}AG}$, and $CWS_{TG\text{-}CA}$, respectively, and the related persistence angles are 18.6^o, 45^o, and 45^o. Interestingly, these values are almost the same as those for the human genome. Then for comparison, we took both the human contig NT_001934 and the $S_S^{0.67}$ sequence and introduced the same proportions of the mutated sites, from 10% to 60%, simultaneously. As shown in Fig. 16, all persistence lengths and angles for $S_S^{0.67}$ and NT_001934 at different percentages of the mutated sites have the same decreasing trends!

These results imply that the members of the $\boldsymbol{R_{H1}}$ and $\boldsymbol{R_{H2}}$ are divergent, fragmented and interspersed in the human genome. Using another synthetic sequence that was constituted by a member in $\boldsymbol{\Gamma_{AE}}$, we applied the same method to the *E. coli* and *A. thaliana* genomes. When the percentages of the mutated sites of the sequence are 73% and 75%, the persistent lengths and angles correspond, respectively, to those for the *E. coli* and *A. thaliana* genomes. Thus, the number of patterns of the $\boldsymbol{R_{AE1}}$ and $\boldsymbol{R_{AE2}}$ found in the *E. coli* genome is more than those found in *A. thaliana*.

When we checked how many of the members of $\boldsymbol{R_{H1}}$ and $\boldsymbol{R_{H2}}$ are present in $S_S^{0.67}$, only less than 0.0002% (no mismatch allowed) and 0.005% (one mismatch allowed) were detected. Actually, the other members of $\boldsymbol{R_{AE1}}$ and $\boldsymbol{R_{AE2}}$ were also found in the $S_S^{0.67}$. Because the original sequence

was composed of only the members of $\boldsymbol{R_{H1}}$ and $\boldsymbol{R_{H2}}$, these patterns were generated by chance and they could not constitute the regular configurations. By contrast, the fragments belonging to $\boldsymbol{R_{H1}}$ and $\boldsymbol{R_{H2}}$ still dominate and form the linear trend and discernible orientations for these synthetic and mutated sequences. The result suggests that the permutations of the members of the *INPR*s in genomes dominate the linear trend and discernable orientation. These members are very short and fragmented so that we can hardly detect them individually by using current approaches, such as alignments or other comparative approaches.

DISCUSSION

INPRs and recognition sites of restriction enzymes

Palindromic and quasi-palindromic (imperfect palindrome) sequences have unique structural properties and often occur near regulatory sites in genomic DNA to serve as recognition sequences for restriction enzymes (Varani, 1995; Glucksmann-Kuis et al., 1996; Kaushik et al., 2003). The term *restriction* comes from the fact that these enzymes were discovered in *E. coli* strains that appeared to be restricting the infection by certain bacteriophages. Most of the enzymes are believed to have a mechanism evolved by bacteria to resist viral attack and to help remove viral sequences (Weaver, 2002). Interestingly, we found several members of the INPRs that have subsequences that are the recognition sites of some of restriction enzymes. For examples, the recognition sites of the enzymes *Alu*I, *Pou*II, and *Pst*I are, respectively, $\begin{matrix}5'\text{–}AG\downarrow CT-3'\\ 3'\text{–}TC\uparrow GA-5'\end{matrix}$, $\begin{matrix}5'\text{–}CAG\downarrow CTG-3'\\ 3'\text{–}GTC\uparrow GAC-5'\end{matrix}$, and $\begin{matrix}5'\text{–}C\quad TGCA\downarrow G-3'\\ 3'\text{–}G\uparrow ACGT\quad C-5'\end{matrix}$ where ↑ and ↓ are the cutting points. The recognized sites 5'-AGCT-3' and 5'-CAGCTG-3' are both partial patterns of $\boldsymbol{R_{H1}}$ while 5'-CTGCAG-3' is a member of $\boldsymbol{R_{H2}}$. Moreover, 5'-↓GATC-3', 5'-GAT↓ATC-3', and 5'-G↓AATTC-3', are the recognition sites of the enzymes *Mbo*I, *Eco*RV, and *Eco*RI, respectively, and they are all members of $\boldsymbol{R_{AE2}}$. In addition, the recombinational hotspot *Chi*, 5'-GCTGGTGG-3', is a recognition site in the major recombination pathway, the RecBCD pathway, in *E. coli*. We found that the *Chi* sequence is a member of $\boldsymbol{R_{H2}}$. Besides, the other

sequence that may display *Chi* activity, GCAGGGCG, is also a member of both $\boldsymbol{R_{H1}}$ and $\boldsymbol{R_{H2}}$. Thus, it may imply that INPRs were once or are now certain recognition sequences and involved in DNA recombination process or other DNA splitting and merging processes, although the INPRs in genomes are found to be divergent and fragmented at the present time.

INPRs and Chargaff's second parity rule

Chargaff's second parity rule stated that the frequencies of guanine (G) and adenine (A) are equal to those of cytosine (C) and thymine (T), respectively, in single-strand DNA (Chargaff, 1951; Rudner *et al.*, 1968). This parity rule indicates that single-strand DNAs are symmetric. The same symmetry results were also observed by Prabhu (1993) and Qi and Cuitcchia (2001). By extrapolation, Forsdyke presented evidence for an evolutionary selection pressure to distribute stem loops generally through genomes (Forsdyke, 1995; Forsdyke and Mortimer, 2000). He also indicated the existence of an evolutionary pressure on DNA base order, rather than on base composition. In our results, we discovered the palindrome-like patterns, CIPPs, in the genomes. Because in the palindromic sequences the frequencies of guanine (G) and adenine (A) are equal to those of cytosine (C) and thymine (T), respectively, the compositional nucleotides of the CIPPs would contribute symmetry to single-strand DNAs. However, the *E. coli* and *Bacillus subtilis* genomes have been found to be strand asymmetric and our analysis also identified the cumulative dinucleotides curves of these genomes to be correlated with the GC skew property. The asymmtry conclusion of these genomes seems in conflict with the symmetry result from other studies. This contradiction has been explained by Frank and Lobry (1999). Chargaff's second parity rule can be formally derived from the first parity rule to give the base frequencies within each strand at equilibrium: #A=#T and #G=#C, where #X denotes the number of X's. Any deviation from the second parity rule implies asymmetric substitution: the result of different mutation rates, different pressures, or both, between the two strands of DNA (Frank and Lobry, 1999). Without a significant mutational bias, the DNA sequences should be strand approximately symmetric.

INPRs and long range correlations in genomes

Many studies revealed certain long range correlations (LRCs) between nucleotides in the genomic sequences of various species and in different genomic regions, but the cause of LRCs remains controversial (Li and Kaneko, 1992; Peng *et al.*, 1992b; Karlin and Brendel, 1993; Ohno, 1993; Peng *et al.*, 1994; Li, 2002; Beirer, 2003). Therefore, Zhou and Mishra (2003) suggested that at the nucleotide level, genomes have evolved independently to possibly share a common scale-free global structure. In our study, the INPRs are found to exist in the genomes. Although these patterns are distorted in the genomes, they can still give rise to LRCs. However, the CIPPs are not able to explain why different sequence lengths of LRCs were found in genomes. It may imply that the other unknown evolutionary mechanisms and scale-free structures in genomes were still hidden and not uncovered yet.

INPRs from the RNA world?

The RNA world hypothesis postulates that in the very beginning RNA molecules functioned both as genetic materials and as enzyme-like catalysts (Orgel, 1998). Although no physical evidence of RNA-based organisms was ever found, because the RNA world would have been extinct for almost four billion years, molecular archaeologists have uncovered artifacts of this ancestral era (Joyce, 2002). These discoveries suggested “molecular fossils” of the RNA World in modern organisms (McGinness and Joyce, 2003).

The RNA world hypothesis postulates that there ever existed particular polymer sequences with some special properties such as adherence to a mineral surface and unusual resistance to degradation (Gilbert, 1986; Joyce, 2002) in the pre-RNA world. The polymers could self-replicate and their replication rate was significantly higher than the decay rate so that they could survive for a while in the extreme prebiotic world (Levy and Ellington, 2001; Joyce, 2002). Several contemporary experimental results have demonstrated the possibility of the hypothesis *in vitro*

(Zielinski and Orgel, 1987; Sievers and von Kiedrowski, 1994; Johnston et al., 2001; Li and Chmielewski, 2003). The design of self replicating molecules had its genesis in the ground breaking work of Sievers and von Kiedrowski (1994) and Orgel and Zielinski (1987). Using palindromic oligonucleotides as a starting point, these researchers designed a means by which the product of a reaction between two smaller oligonucleotides could act as the template to promote this reaction. Hence the product could be formed in an autocatalytic or self replicating fashion (Li and Chmielewski, 2003).

The INPRs are divergent and uniformly distributed in the whole genome. These patterns are palindromic and parts of the patterns are also the recognition sites of enzymes. Thus, we wonder whether these patterns are related to any particular polymers in the RNA world. Although we have no direct evidence that Γ_{H} and Γ_{AE} can play a role like ribozymes, it may be interesting to study their biochemical properties.

ACKNOWLEDGMENTS

We thank Dr. F.-C. Chen and Dr. H.-K. Tsai for valuable discussions. This work was supported by Academia Sinica, and by the National Science Council, Taiwan, under the Grants NSC-92-3112-B-001-018-Y, NSC-92-3112-B-001-021-Y, and NSC-93-2752-E-002-005-PAE.

REFERENCES

Audit, B., Thermes, C., Vaillant, C., d'Aubenton-Carafa, Y., Muzy, J.F., and Arneodo, A. (2001) Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Phys Rev Lett*, **86**, 2471-2474.

Beirer, S. 2003. Modelling long-range Correlations in Genomic DNA Sequences. In *Institute of Theoretical Physics, Institute of Theoretical Biology*. Humboldt University, Berlin.

Berthelsen, C.L., Glazier, J.A., and Skolnick, M.H. (1992) Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Physical Review. A*, **45**, 8902-8913.

Blattner, F.R., Plunkett, G.I., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F. *et al.* (1997) The complete genome sequence of Escherichia coli K-12. *Science*, **277**, 1453-1474.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.K., Simons, M., and Stanley, H.E. (1995) Long-range correlation properties of coding and noncoding DNA

sequences: GenBank analysis. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, **51**, 5084-5091.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C., Simons, M., Sciortino, F., and Stanley, H.E. (1993) Long-range fractal correlations in DNA. *Physical Review Letters*, **71**, 1776.

Chargaff, E. (1951) Structure and function of nucleic acids as cell constituents. *Fed Proc*, **10**, 654-659.

Forsdyke, D.R. (1995) A stem-loop "kissing" model for the initiation of recombination and the origin of introns. *Mol Biol Evol*, **12**, 949-958.

Forsdyke, D.R. and Mortimer, J.R. (2000) Chargaff's legacy. *Gene*, **261**, 127-137.

Francino, M.P. and Ochman, H. (1997) Strand asymmetries in DNA evolution. *Trends Genet*, **13**, 240-245.

Frank, A.C. and Lobry, J.R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, **238**, 65-77.

Freeman, J.M., Plasterer, T.N., Smith, T.F., and Mohr, S.C. (1998a) Patterns of Genome Organization in Bacteria. *Science*, **279**, 1827-1828.

Freeman, J.M., Plasterer, T.N., Smith, T.F., and Mohr, S.C. (1998b) Patterns of Genome Organization in Bacteria. *Science*, **279**, 1827a-.

Gates, M.A. (1985) Simpler DNA sequence representations. *Nature*, **316**, 219.

Gilbert, W. (1986) The RNA world. *Nature*, **319**, 618.

Glucksmann-Kuis, M.A., Dai, X., Markiewicz, P., and Rothman-Denes, L.B. (1996) E. coli SSB activates N4 virion RNA polymerase promoters by stabilizing a DNA hairpin required for promoter recognition. *Cell*, **84**, 147-154.

Graur, D. and Li, W.-H. 2000. *Fundamental of Molecular Evolution*. Sinauer Press, Sunderland, MA.

Hamori, E. (1985) Novel DNA sequence representations. *Nature*, **314**, 585-586.

Hamori, E. and Ruskin, J. (1983) H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J Biol Chem*, **258**, 1318-1327.

Herzel, H. and Grosse, I. (1997) Correlations in DNA sequences: The role of protein coding segments. *Phys Rev E*, **55**, 800-810.

Herzel, H., Weiss, O., and Trifonov, E.N. (1999) 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics*, **15**, 187-193.

Johnston, W.K., Unrau, P.J., Lawrence, M.S., Glasner, M.E., and Bartel, D.P. (2001) RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science*, **292**, 1319-1325.

Joyce, G.F. (2002) The antiquity of RNA-based evolution. *Nature*, **418**, 214-221.

Karlin, S. (1999) Bacterial DNA strand compositional asymmetry. *Trends Microbiol*, **7**, 305-308.

Karlin, S. and Brendel, V. (1993) Patchiness and correlations in DNA sequences. *Science*, **259**, 677-680.

Karlin, S., Brocchieri, L., Trent, J., Blaisdell, B.E., and Mrazek, J. (2002) Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes. *Theor Popul Biol*, **61**, 367-390.

Kaushik, M., Kukreti, R., Grover, D., Brahmachari, S.K., and Kukreti, S. (2003) Hairpin-duplex equilibrium reflected in the A-->B transition in an undecamer quasi-palindrome present in the locus control region of the human beta-globin gene cluster. *Nucleic Acids Res*, **31**, 6904-6915.

Kunkel, T.A. (1992) Biological asymmetries and the fidelity of eukaryotic DNA replication. *Bioessays*, **14**, 303-308.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S. *et al.* (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature*, **390**, 249-256.

Lander, E.S. Linton, L.M. Birren, B. Nusbaum, C. Zody, M.C. Baldwin, J. Devon, K. Dewar, K. Doyle, M. FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

Levy, M. and Ellington, A.D. (2001) The descent of polymerization. *Nat Struct Biol*, **8**, 580-582.

Li, W. (1997) The study of correlation structures of DNA sequences: a critical review. *Comput Chem*, **21**, 257-271.

Li, W. (2002) Are isochore sequences homogeneous? *Gene*, **300**, 129-139.

Li, W. and Kaneko, K. (1992) DNA correlations. *Nature*, **360**, 635-636.

Li, X. and Chmielewski, J. (2003) Challenges in the design of self replicating peptides. *Org Biomol Chem*, **1**, 901-904.

Lobry, J.R. (1996a) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol*, **13**, 660-665.

Lobry, J.R. (1996b) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, **78**, 323-326.

McGinness, K.E. and Joyce, G.F. (2003) In search of an RNA replicase ribozyme. *Chem Biol*, **10**, 5-14.

Ohno, S. (1993) Patterns in genome evolution. *Curr Opin Genet Dev*, **3**, 911-914.

Orgel, L.E. (1998) The origin of life--a review of facts and speculations. *Trends Biochem Sci*, **23**, 491-495.

Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H.E. (1992a) Fractal landscape analysis of DNA walks. *Physica A*, **191**, 25-29.

Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H.E. (1992b) Long-range correlations in nucleotide sequences. *Nature*, **356**, 168-170.

Peng, C.K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., and Goldberger, A.L. (1994) Mosaic organization of DNA nucleotides. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, **49**, 1685-1689.

Prabhu, V.V. (1993) Symmetry observations in long nucleotide sequences. *Nucleic Acids Res*, **21**,

2797-2800.

Qi, D. and Cuticchia, A.J. (2001) Compositional symmetries in complete genomes. *Bioinformatics*, **17**, 557-559.

Rudner, R., Karkas, J.D., and Chargaff, E. (1968) Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proc Natl Acad Sci U S A*, **60**, 921-922.

Shioiri, C. and Takahata, N. (2001) Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *J Mol Evol*, **53**, 364-376.

Sievers, D. and von Kiedrowski, G. (1994) Self-replication of complementary nucleotide-based oligomers. *Nature*, **369**, 221-224.

Smit, A.F. and Green, P. 1997. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html.

THE ARABIDOPSIS GENOME INITIATIVE. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796-815.

Varani, G. (1995) Exceptionally stable nucleic acid hairpins. *Annu Rev Biophys Biomol Struct*, **24**, 379-404.

Weaver, R. 2002. *Molecular Biology*. McGraw-Hill Companies, Inc.

Weiss, O. and Herzel, H. (1998) Correlations in protein sequences and property codes. *J Theor Biol*, **190**, 341-353.

Zhou, Y. and Mishra, B. (2003) Models of Genome Evloution. *Modeling in Molecular Biology, Lecture Notes in Computer Science*.

Zielinski, W.S. and Orgel, L.E. (1987) Autocatalytic synthesis of a tetranucleotide analogue. *Nature*, **327**, 346-347.

TABLES

Table 1. The average persistence angles and the average persistence lengths for three examined genomes: human, *E. coli K12,* and *A. thaliana* in different **CWS**s. Because these sequenced genomes were very long, the counting window sizes were set to be 50,000 bp to let the estimated values have more statistical significance. (+,+) and (–,–) represent the orientation directions of the configurations in the cartesian coordinate system. We also calculated the variances of the average persistence angles. The range of variances is marked by ***: the variance is less than 10^o; **: the variance is within 10^o and 20^o; *: the variances are between 20^o and 30^o; and no *: the variance is larger than 30^o and the estimated angles may be considered not statistically significant.

Species	$CWS_{GC\text{-}AT}$		$CWS_{TG\text{-}CA}$		$CWS_{CT\text{-}AG}$	
	$r_{GC\text{-}AT}$	$\phi_{GC\text{-}AT}$	$r_{TG\text{-}CA}$	$\phi_{TG\text{-}CA}$	$r_{CT\text{-}AG}$	$\phi_{CT\text{-}AG}$
Human	0.100	$19.8(+,+)^{***}$	0.094	$45.0(+,+)^{***}$	0.045	$44.6(+,+)^{***}$
A. thaliana	0.052	$66.9(+,+)^{**}$	0.048	$44.6(+,+)^{***}$	0.021	$43.8(–,–)^{**}$
E. coli	0.068	$69.2(+,+)^{***}$	0.063	$45.8(+,+)^{***}$	0.028	$42.5(–,–)^{**}$

Table 2. The average persistence angles and the average persistence lengths for the coding regions in the human, *E. coli* and *A. thaliana* genomes; we generated two sets of sequences: (1) sequences that are concatenated from the coding sequences in the same orientations and (2) sequences that are concatenated from the coding sequences in the same strand. The counting window sizes were 50,000 bp. (a) All coding sequences in the human genome were used. (b) The coding regions in chromosomes 12 and Y were neglected because these regions are too short compared to their chromosome sizes.

Species	$CWS_{GC\text{-}AT}$		$CWS_{TG\text{-}CA}$		$CWS_{CT\text{-}AG}$	
	$r_{GC\text{-}AT}$	$\phi_{GC\text{-}AT}$	$r_{TG\text{-}CA}$	$\phi_{TG\text{-}CA}$	$r_{CT\text{-}AG}$	$\phi_{CT\text{-}AG}$
(1) The sequences by concatenating in the same orientations						
Human[a]	0.118	24.0(+,+)***	0.111	38.1(+,+)***	0.041	23.9(+,+)***
Human[b]	0.119	23.6(+,+)***	0.111	38.0(+,+)***	0.044	25.6(+,+)***
A. thaliana	0.073	66.9(+,+)***	0.069	35.9(+,+)***	0.028	63.8(–,–)**
E. coli	0.070	69.7(+,+)***	0.074	13.7(+,+)***	0.047	-78.2(+,–)**
(2) The examined sequences are concatenated by the coding regions in the same strands						
Human[a]	0.118	24.1(+,+)***	0.110	45.2(+,+)***	0.039	40.0(+,+)***
Human[b]	0.119	23.6(+,+)***	0.110	45.2(+,+)***	0.042	43.1(+,+)***
A. thaliana	0.073	66.8(+,+)***	0.068	45.4(+,+)***	0.026	41.9(–,–)***
E. coli	0.069	70.0(+,+)***	0.063	45.1(+,+)***	0.029	43.3(–,–)**

Table 3. The average persistence angles and the average persistence lengths for some bacteria and eukaryote genomes. The superscript notations are the same as in Table 1. The counting window sizes were 50,000 bp.

Species	$CWS_{GC\text{-}AT}$		$CWS_{TG\text{-}CA}$		$CWS_{CT\text{-}AG}$	
	$r_{GC\text{-}AT}$	$\phi_{GC\text{-}AT}$	$r_{TG\text{-}CA}$	$\phi_{TG\text{-}CA}$	$r_{CT\text{-}AG}$	$\phi_{CT\text{-}AG}$
Bacteria						
Aeropyrum pernix	0.07399	-30.74(+,–)***	0.01889	44.22(+,+)**	0.06755	41.93(+,+)***
Archaeolglobus_fulgidus	0.06479	49.06(+,+)***	0.06482	45.63(+,+)***	0.00717	-3.82(–,–)
Bacillus subtilis	0.09584	69.1(+,+)***	0.08822	45.1(+,+)***	0.0406	43.45(–,–)**
Haemophilus influenzae	0.06879	56.56(+,+)***	0.06741	45.09(+,+)***	0.0138	35.2(–,–)*
Mycopasma pneumoniae	0.04255	-24.24(+,–)***	0.01864	27.97(+,+)	0.03604	35.18(+,+)**
Mycoplasma genitalium	0.06205	3.98(+,+)**	0.04782	40.86(+,+)**	0.03504	37.58(+,+)**
Thermoplasma volcanium	0.05268	48.58(+,+)***	0.05217	44.34(+,+)***	0.00658	20.38(–,–)
E. coli O175	0.06822	68.95(+,+)***	0.06315	44.89(+,+)***	0.02888	44.46(–,–)**
Eukaryotes						

S. pombe	0.04594	60.32(+,+)***	0.04435	45.73(+,+)	0.01255	39.82(–,–)*
Drosophila	0.06618	50.24(+,+)***	0.06574	44.66(+,+)***	0.00771	24.15(–,–)
C. elegan	0.07547	58.54(+,+)	0.05795	44.67(+,+)***	0.04791	44.89(–,–)***

Table 4. A list of all CIPPs (*Compact Intrinsic Permutation Patterns*) in two groups. The sequences in $\mathbf{\Gamma_H}$ and $\mathbf{\Gamma_{AE}}$ have the same trends and orientations as those for human and those for *E. coli* and *A. thaliana*, respectively.

$\mathbf{\Gamma_H}$	TGCAGCTGCA, GCAGCTGCAT, CAGCTGCATG, AGCTGCATGC, GCTGCATGCA, CTGCATGCAG, TGCATGCAGC, GCATGCAGCT, CATGCAGCTG, ATGCAGCTGC, TGCTGCAGCA, GCTGCAGCAT, CTGCAGCATG, TGCAGCATGC, GCAGCATGCT, CAGCATGCTG, AGCATGCTGC, GCATGCTGCA, CATGCTGCAG, ATGCTGCAGC
$\mathbf{\Gamma_{AE}}$	CATCATGATG, ATCATGATGC, TCATGATGCA, CATGATGCAT, ATGATGCATC, TGATGCATCA, GATGCATCAT, ATGCATCATG, TGCATCATGA, GCATCATGAT, CATGATCATG, ATGATCATGC, TGATCATGCA, GATCATGCAT, ATCATGCATG, TCATGCATGA, CATGCATGAT, ATGCATGATC, TGCATGATCA, GCATGATCAT

Table 5. The search results for human, *E. coli*, and *A. thaliana*. The INPRs used for searching human are $\boldsymbol{R_{H1}}$ and $\boldsymbol{R_{H2}}$. To search E. coli and A. thaliana, the INPRs are $\boldsymbol{R_{AE1}}$ and $\boldsymbol{R_{AE2}}$. The number of minimum visiting nodes, N_v, in the INPRs is set to be 11 and the maximum number of the allowed missing visiting nodes, N_a, is set to either be 0 or 1 in this table. [a] The percentage is the detected number of INPRs in inter-genetic regions divided by the total number detected in the whole genome. [b] The percentage shows that the detected number of INPRs in intra-genetic regions divided by the total number detected in the whole genome. [c] The percentage shows the detected number of INPRs in coding regions divided by the total number detected in the whole genome. [d] The percentage shows the total lengths of the detected INPRs divided by the total lengths of the whole genome.

	Human ($\boldsymbol{R_{H1}}$ & $\boldsymbol{R_{H2}}$)		*E. coli* ($\boldsymbol{R_{AE1}}$ & $\boldsymbol{R_{AE2}}$)		*A. thaliana* ($\boldsymbol{R_{AE1}}$ & $\boldsymbol{R_{AE2}}$)	
$N_v = 11$ and $N_a = 0$						
No. in inter-genetic region	70895	57.8%[a]	15	12.3%[a]	2213	45.0%[a]
No. in intra-genetic region	51863	42.2%[b]	107	87.7%[b]	2706	55.0%[b]
No. in coding region	4021	3.3% [c]	N.A.	N.A.	1117	22.7% [c]
Total length(bp)	$\boldsymbol{R_{H1}}$: 951617		$\boldsymbol{R_{AE1}}$: 1085		$\boldsymbol{R_{AE1}}$: 39216	

	$\boldsymbol{R_{H2}}$: 1019176		$\boldsymbol{R_{AE2}}$: 791		$\boldsymbol{R_{AE2}}$: 39603	
% in the whole genome	0.0653%[d]		0.0404%[d]		0.0661%[d]	
N_v = 11 and N_a = 1						
No. in inter-genetic region	928288	57.5%[a]	298	14.5%[a]	30438	46.5%[a]
No. in intra-genetic region	686452	42.5%[b]	1755	85.5%[b]	35032	53.5%[b]
No. in coding region	44618	2.8%[c]	N.A.	N.A.	14978	22.9%[c]
Total length	$\boldsymbol{R_{H1}}$: 11469386 $\boldsymbol{R_{H2}}$: 13924293		$\boldsymbol{R_{AE1}}$: 18175 $\boldsymbol{R_{AE2}}$: 13191		$\boldsymbol{R_{AE1}}$: 546073 $\boldsymbol{R_{AE2}}$: 500379	
% in the whole genome	0.8415%[d]		0.6761%[d]		0.8780%[d]	

FIGURE LEGENDS

Fig. 1. Construction of the three Crystalline-like Walking Spaces (CWSs). (a) There are four basic structures with each of the four elements A, C, T, and G as the central node. (b) Three crystalline-like walking spaces. These CWSs cannot be obtained from one another by any linear transformation.

Fig. 2. Given the double-stranded DNA sequence $\begin{matrix}5'\ \text{ACTGCAG}\ 3'\\ 3'\ \text{TGACGTC}\ 5'\end{matrix}$, the two constituent strands are S=5'ACTGCAG 3' and $\bar{S}$ =5'CTGCAGT3'. The configurations of S and $\bar{S}$ are, respectively, represented by the red and yellow vectors. If the configuration of $\bar{S}$ is rotated by 180^o, its shape is exactly the same as that of S; the only difference is that the orientation is reversed.

Fig. 3. (a) The 3-D walking tracks of NT_011362 in $CWS_{GC\text{-}AT}$ for different window sizes. (b) The two-dimensional configurations of NT_011362 in $CWS_{GC\text{-}AT}$ for different window sizes.

Fig. 4. The configurations of NT_011362 in $CWS_{GC\text{-}AT}$, $CWS_{CT\text{-}AG}$, and $CWS_{TG\text{-}CA}$ at the full-length scale.

Fig. 5. The configurations of NT_011362 with repeats masked in $CWS_{GC\text{-}AT}$, $CWS_{CT\text{-}AG}$, and $CWS_{TG\text{-}CA}$ at the full-length scale.

Fig. 6. The configurations of a random sequence (20 Mb) in $CWS_{GC\text{-}AT}$, $CWS_{CT\text{-}AG}$, and $CWS_{TG\text{-}CA}$.

Fig. 7. The estimates of $\bar{\theta}_{xx-xx}(w)$, $\bar{r}_{xx-xx}(w)$, and $\bar{\phi}_{xx-xx}(w)$ for the sequence NT_011362. The vertical axes in the three diagrams represent $\bar{\theta}_{xx-xx}(w)$, $\bar{r}_{xx-xx}(w)$, and $\bar{\phi}_{xx-xx}(w)$, respectively, and the horizontal axis in each diagram represents different window sizes *w* from 10 to 50,000 bp.

Fig. 8. The distributions of $(\bar{x}_{\text{TG-CA}}, \bar{y}_{\text{TG-CA}})$s, $(\bar{x}_{\text{CT-AG}}, \bar{y}_{\text{CT-AG}})$s, and $(\bar{x}_{\text{GC-AT}}, \bar{y}_{\text{GC-AT}})$s of 352 human contigs. (X,Y) is a Cartesian coordinate system of $(\bar{r}_{TG-CA,}\bar{\phi}_{TG-CA})$, $(\bar{r}_{CT-AG,}\bar{\phi}_{CT-AG})$, and $(\bar{r}_{GC-AT,}\bar{\phi}_{GC-AT})$.

Fig. 9. The configurations of two genomes in $\text{CWS}_{\text{GC-AT}}$, $\text{CWS}_{\text{CT-AG}}$, and $\text{CWS}_{\text{TG-CA}}$. (a) The *E. coli* genome. (b) Chromosome I of the *A. thaliana* genome.

Fig. 10. Estimates of $\bar{\theta}_{xx-xx}(w)$, $\bar{r}_{xx-xx}(w)$, and $\bar{\phi}_{xx-xx}(w)$ for the *E. coli* genome and *A. thaliana* chromosome I with different window sizes.

Fig. 11. The cumulative dinucleotide curves, $H_{GG\text{-}CC\ skew}$, $H_{GA\text{-}TC\ skew}$, and $H_{GT\text{-}AC\ skew}$, of *E. coli* K12 and *Bacillus subtilis* genomes. (a) *E. coli* K12: The origin and the terminus of the DNA replication are located at 3,923,500bp and 1,588,800bp (Blattner et al., 1997), respectively. (b) *Bacillus subtilis*: The origin and the terminus of the DNA replication are located at 1bp and 2,017,000bp (Kunst et al., 1997), respectively. (c) and (d) The distributions of the compositional factor ratios, $k_{GG\text{-}CC}$, $k_{GA\text{-}TC}$, and $k_{GT\text{-}AC}$, in *E. coli* and *B. subtilis*, respectively. (c) The other cumulative dinucleotide curves of *E. coli* K12 genomes: $H_{GA\text{-}AC}$, $H_{GG\text{-}TC}$, $H_{GT\text{-}TC}$, and $H_{AA\text{-}TT}$.

Fig. 12. Two repetitive sequences, $\boldsymbol{S_H}$ and $\boldsymbol{S_{AE}}$, mapped onto the different walking spaces where the arrows in purple color represent the tracks for one elementary subsequence of the sequences and the bold-face arrows in red color represent the total displacement vector for each elementary subsequence. (a) The configuration of $\boldsymbol{S_H}$ = 5'..$(\text{CATGCAGCTG})_n$..3'. (b) The configuration of $\boldsymbol{S_{AE}}$ = 5'..$(\text{CATGATCATG})_n$..3'.

Fig. 13. Two groups of **INPR**s. (a) For the human genome. (b) For the *E. coli* and *A. thaliana* genomes.

Fig. 14. The relative percentage diversity ratios. For the human genome, $\boldsymbol{PDR_{GC\text{-}AT}}$, $\boldsymbol{PDR_{CT\text{-}AG}}$, and

$\boldsymbol{PDR_{TG\text{-}CA}}$ are 0.1054, 0.111, and 0.1061, respectively. The corresponding ratios are 0.0717, 0.0754, and 0.0754 for the *E. coli* genome, and 0.0559, 0.0556, and 0.0472 for the *A. thaliana* genome.

Fig. 15: A synthetic sequence $\mathbf{S_s}$ composed of 2,600,00 consecutive copies of a sequence in $\boldsymbol{\Gamma_H}$: TGCAGCTGCA. (a) The persistence lengths for different percentages of mutated sites. (b) The persistence angle for different percentages of mutated sites.

Fig. 16: Comparison of decay ratios between a synthetic sequence $S_S^{0.67}$ and a real human contig, NT_001934, where $S_S^{0.67}$ is composed of 2,600,00 consecutive copies of a sequence in $\boldsymbol{\Gamma_H}$ but 67% of the sites have been mutated at random. (a) The persistence length versus the percentage of mutated sites. (b) The persistence angle versus the percentage of mutated sites.

FIGURES

Fig. 1

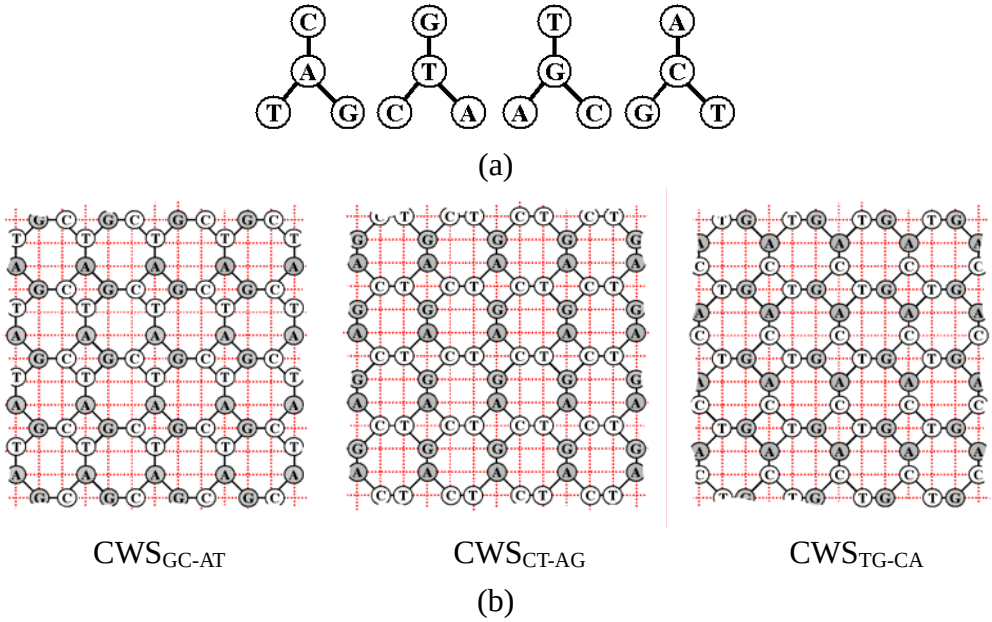


Fig. 2

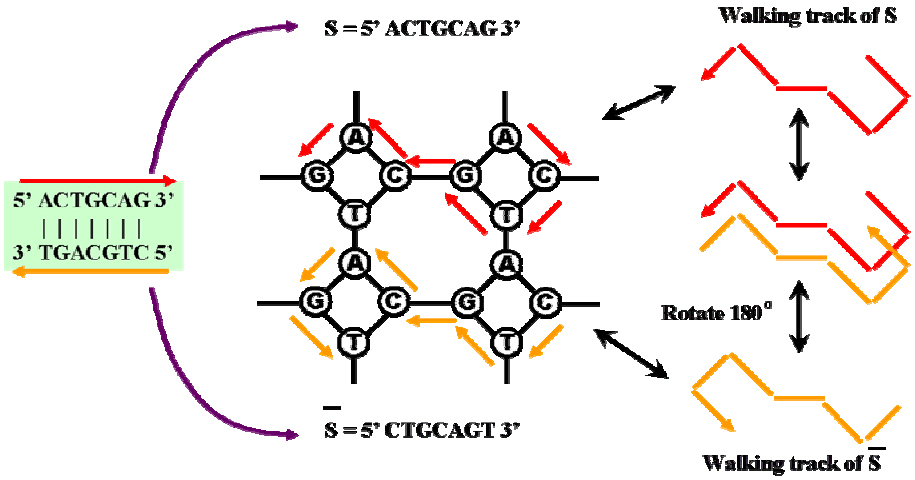
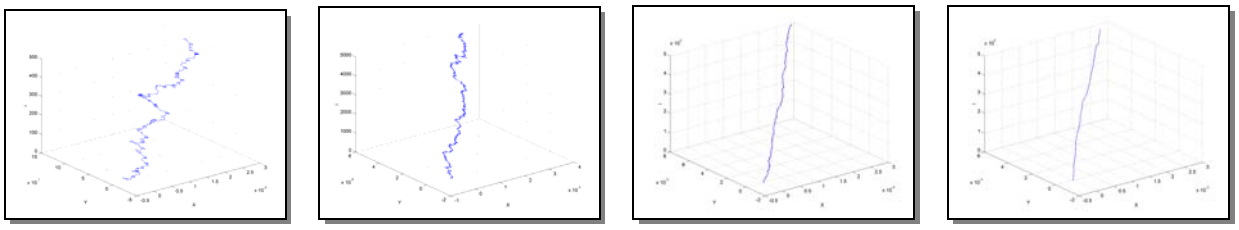
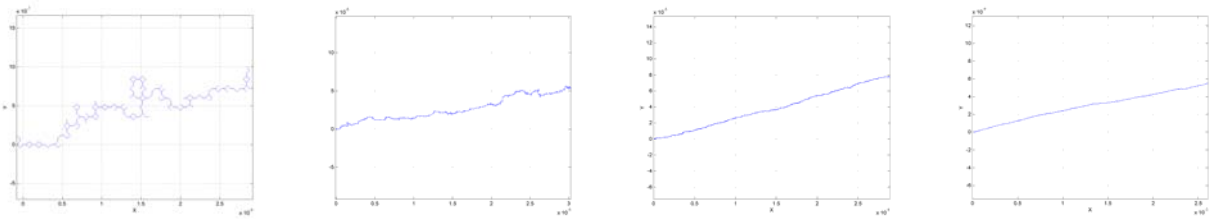


Fig. 3



500 bp 5,000bp 50,000bp 500,000bp

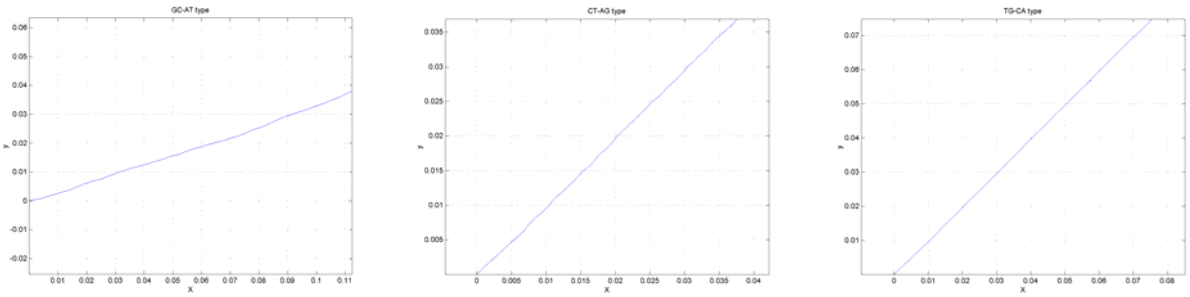
(a)



500 bp 5,000bp 50,000bp 500,000bp

(b)

Fig. 4



$CSW_{GC\text{-}AT}$ $CWS_{CT\text{-}AG}$ $CWS_{TG\text{-}CA}$

Fig. 5

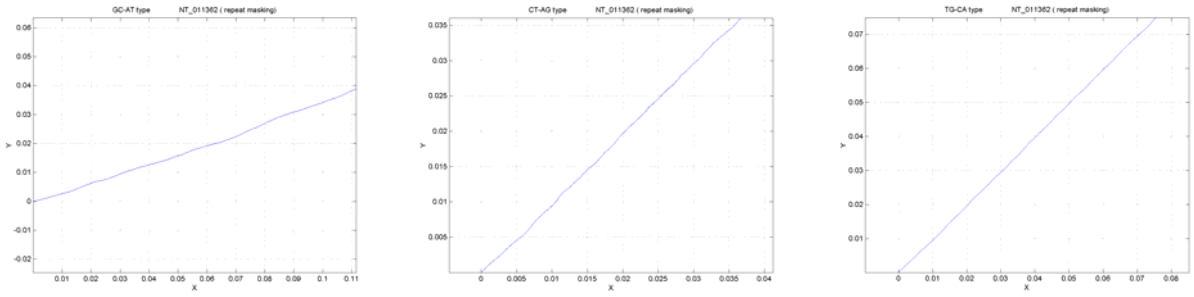
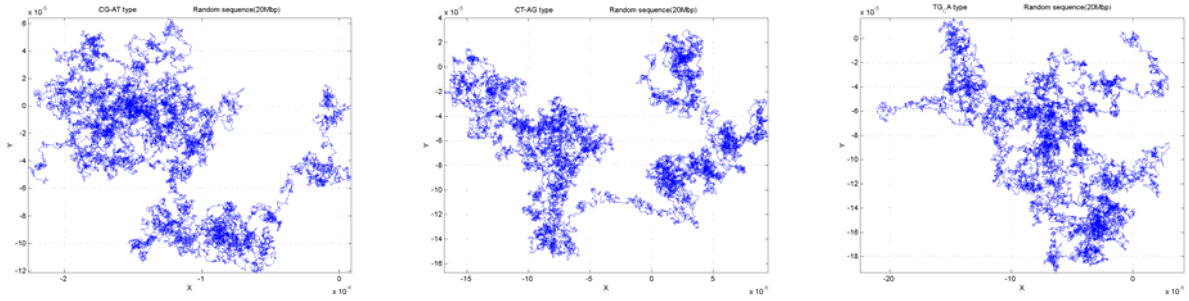


Fig. 6



$CSW_{GC\text{-}AT}$. $CWS_{CT\text{-}AG}$. $CWS_{TG\text{-}CA}$.

Fig. 7

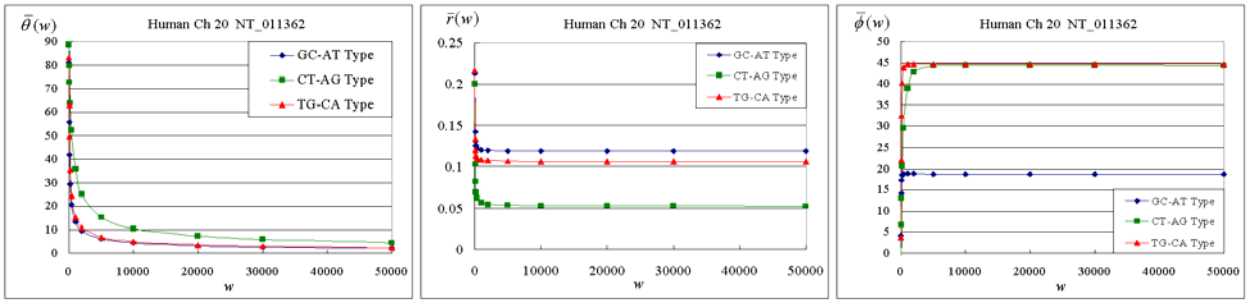


Fig. 8

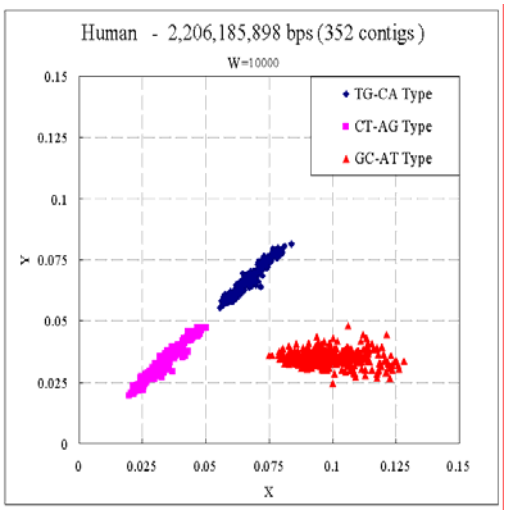
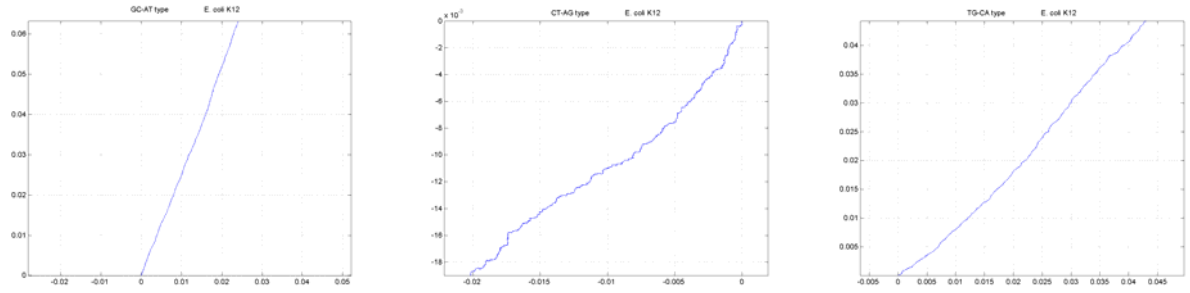
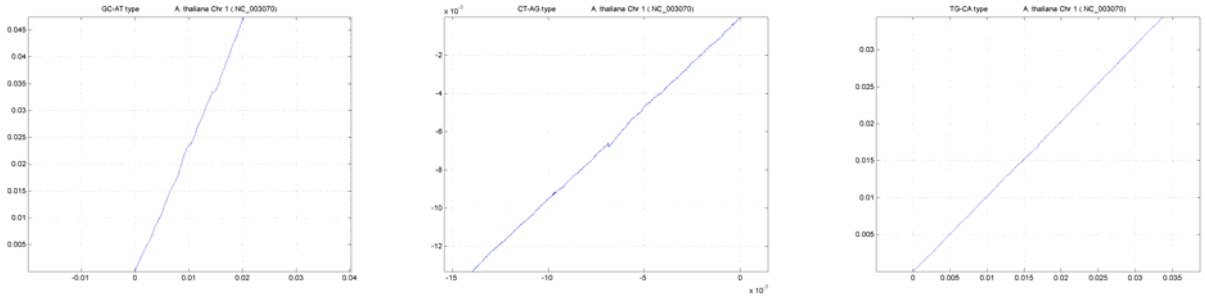


Fig. 9



CSW_{GC-AT}. CWS_{CT-AG}. CWS_{TG-CA}.

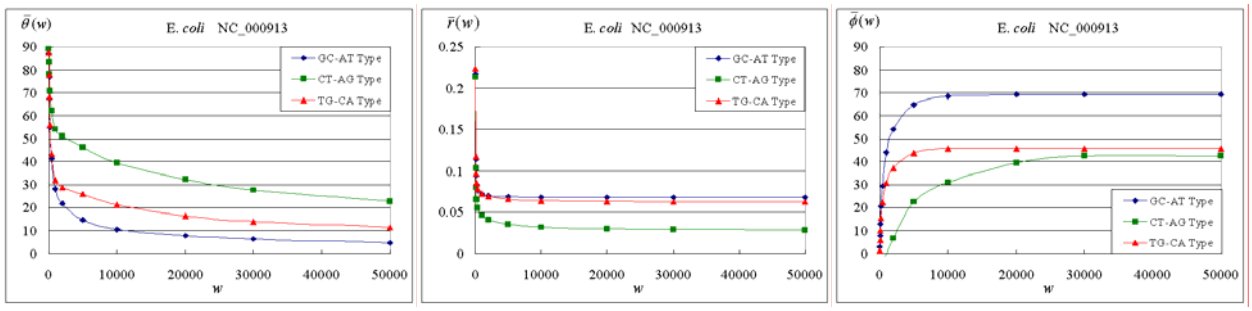
(a)



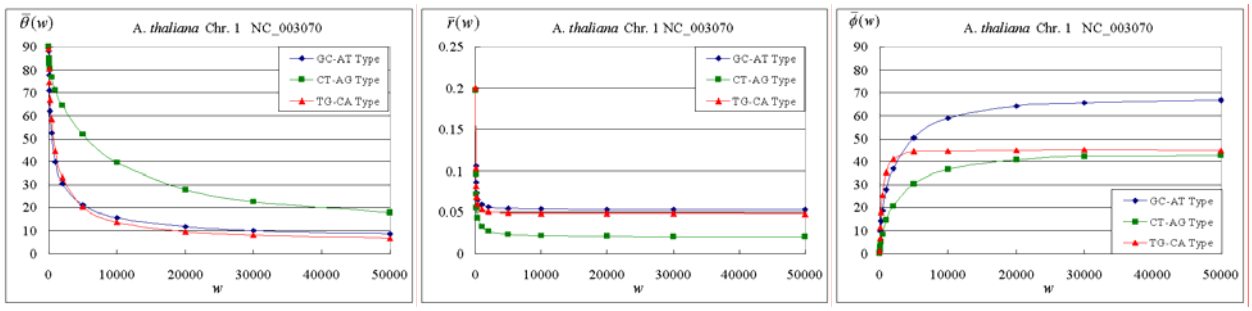
CSW_{GC-AT}. CWS_{CT-AG}. CWS_{TG-CA}.

(b)

Fig. 10

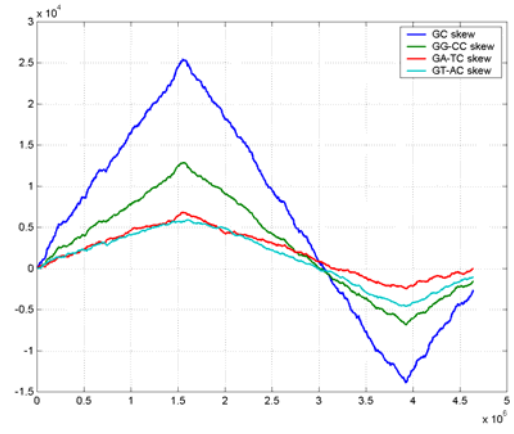


(a) *E. coli* whole genome.

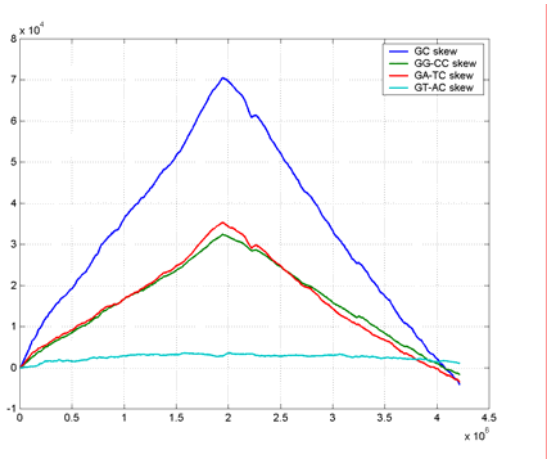


(b) *A. thaliana* chromosome I.

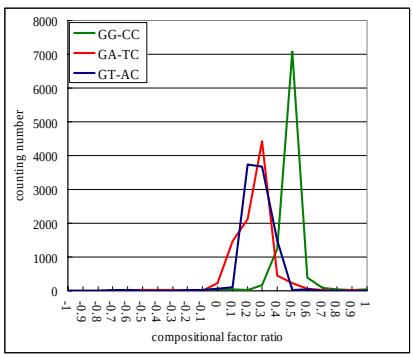
Fig. 11

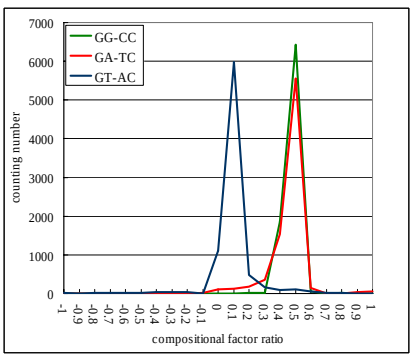


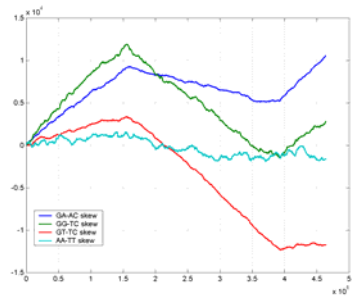
(a) *E. coli* K12.



(b) *Bacillus subtilis*.

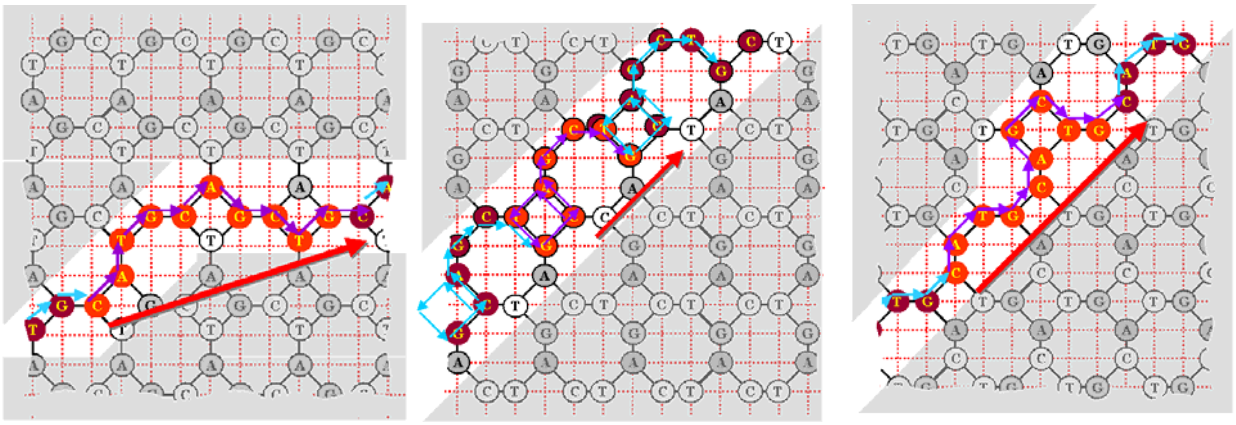






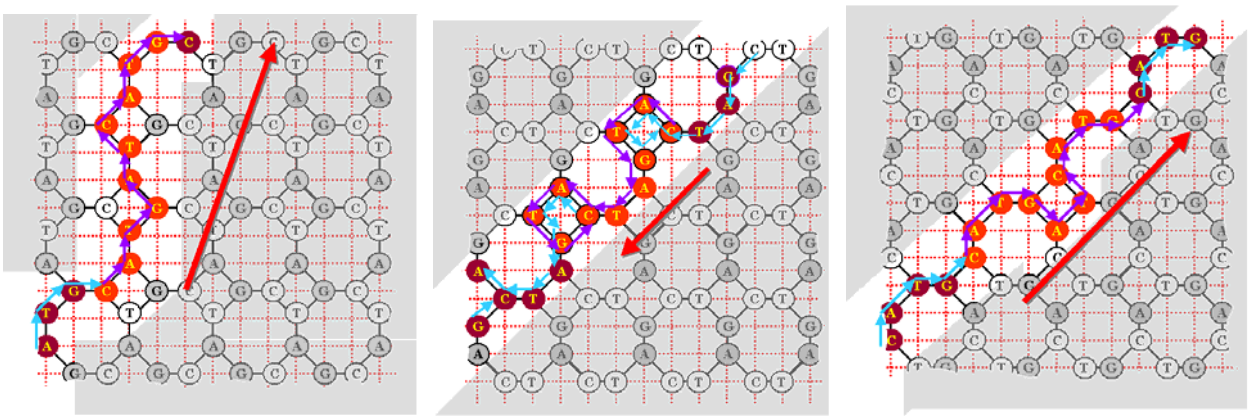
(c) *E. coli* K12. (d) *Bacillus subtilis*. (e) *E. coli* K12.

Fig. 12



$CWS_{GC\text{-}AT}$ $CWS_{CT\text{-}AG}$ $CWS_{TG\text{-}CA}$

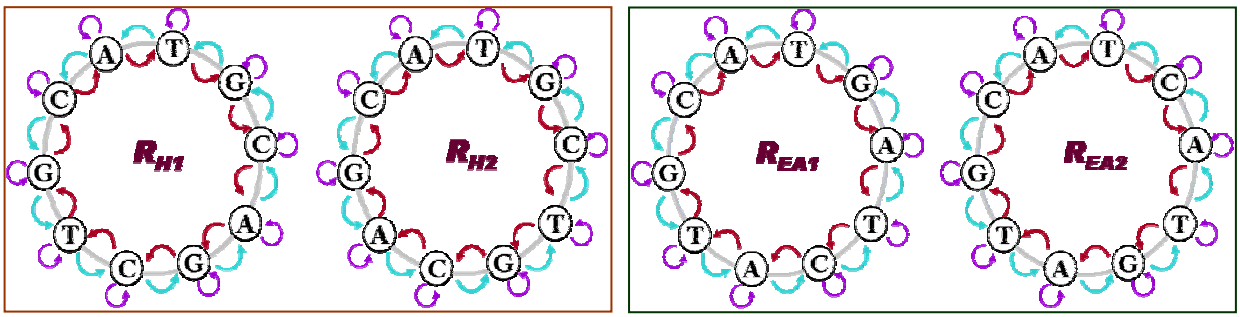
(a)



$CWS_{GC\text{-}AT}$ $CWS_{CT\text{-}AG}$ $CWS_{TG\text{-}CA}$

(b)

Fig. 13



(a) $\mathbf{INPR_H}=\{\boldsymbol{R_{H1}},\boldsymbol{R_{H2}}\}$. (b) $\mathbf{INPR_{EA}}=\{\boldsymbol{R_{AE1}},\boldsymbol{R_{AE2}}\}$.

Fig. 14

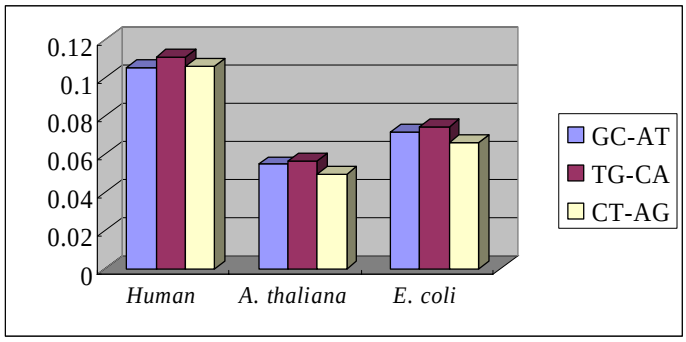
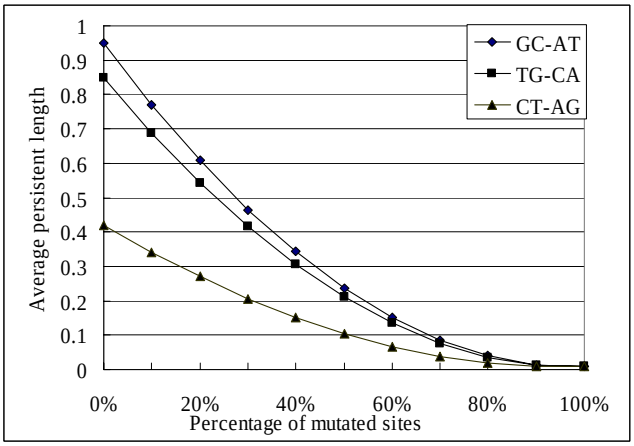
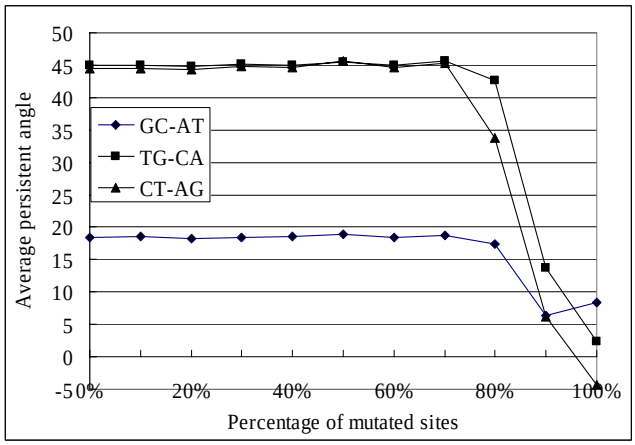


Fig. 15

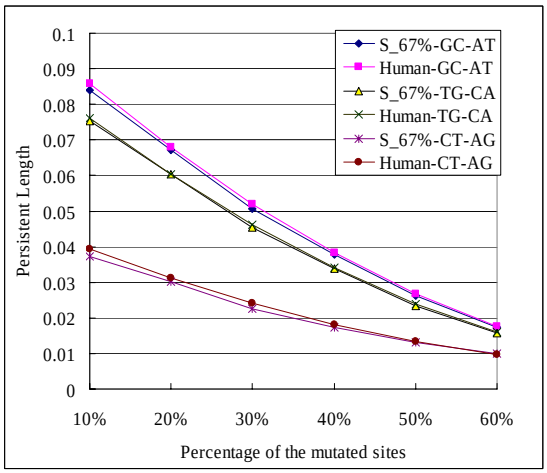


(a)

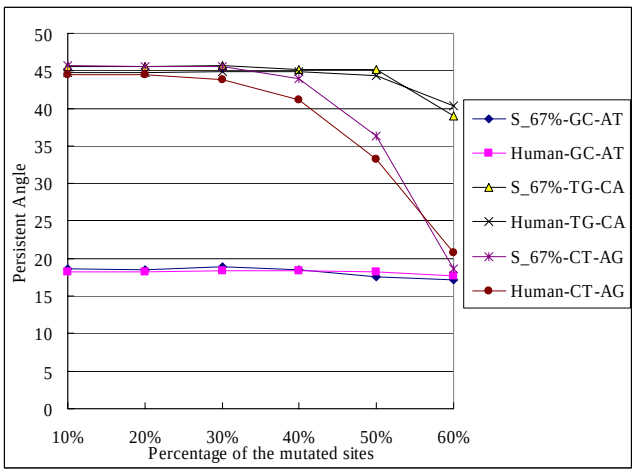


(b)

Fig. 16



(a)



(b)