Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2014, Article ID 176943, 9 pages
http://dx.doi.org/10.1155/2014/176943

Research Article

An Exploration of the Triplet Periodicity in Nucleotide Sequences with a Mature Self-Adaptive Spectral Rotation Approach

Bo Chen[1,2] and Ping Ji[3]

[1] *College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China*
[2] *Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, Fuzhou 350116, China*
[3] *Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong*

Correspondence should be addressed to Bo Chen; bo.chen@fzu.edu.cn

Received 19 April 2014; Revised 20 July 2014; Accepted 25 July 2014; Published 12 August 2014

Academic Editor: Ning Hu

Previously, for predicting coding regions in nucleotide sequences, a self-adaptive spectral rotation (SASR) method has been developed, based on a universal statistical feature of the coding regions, named triplet periodicity (TP). It outputs a random walk, that is, TP walk, in the complex plane for the query sequence. Each step in the walk is corresponding to a position in the sequence and generated from a long-term statistic of the TP in the sequence. The coding regions (TP intensive) are then visually discriminated from the noncoding ones (without TP), in the TP walk. In this paper, the behaviors of the walks for random nucleotide sequences are further investigated qualitatively. A slightly leftward trend (a negative noise) in such walks is observed, which is not reported in the previous SASR literatures. An improved SASR, named the mature SASR, is proposed, in order to eliminate the noise and correct the TP walks. Furthermore, a potential sequence pattern opposite to the TP persistent pattern, that is, the TP antipersistent pattern, is explored. The applications of the algorithms on simulated datasets show their capabilities in detecting such a potential sequence pattern.

1. Introduction

Coding region prediction for nucleotide sequences is an active issue in the field of computational biology [1–10]. Techniques, including the dynamic programming (DP) and the Hidden Markov Model (HMM), have been adopted to process information collected from *ab initio* experiments and predict potential coding regions. Besides, researchers suggest that the usages of codons are highly nonrandom in coding regions [11], and the biased appearance of codons raises a universal property in coding regions, called the "triplet periodicity (TP)." Investigating the TP property can be a subject of interest for developing the coding regions detection algorithm [12, 13], as well as some other significant gene related issues.

The TP property was first presented by Fickett [14]. It is said to be a simple and universal difference between coding and noncoding regions. After Fickett's work, the TP property was analyzed with various theoretical tools, such as the hidden Markov chains [15, 16], the time series [17, 18], the information theory [11, 12], and the Fourier transform [19–25]. Studies on the TP property are with the aim of predicting coding regions [26] and, especially, detecting frame shift points in nucleotide sequences [27, 28]. Among such methods, the self-adaptive spectral rotation (SASR) provides a visualization of the TP property hidden in nucleotide sequences and can be employed for training-free coding region prediction [24, 25]. This method takes only the query sequence as its input and outputs a random walk in the complex plane, called the TP walk, which conveniently presents the locations of coding (TP intensive) regions as well as frame shifts. Here, a "frame shift" Δ is related to the length of the interregion gap g (the non-TP region between two TP intensive regions), and it is defined as $\Delta = g \bmod 3$.

In Chen and Ji's work [24], they claimed that, for simple random sequences, the TP walks should be random

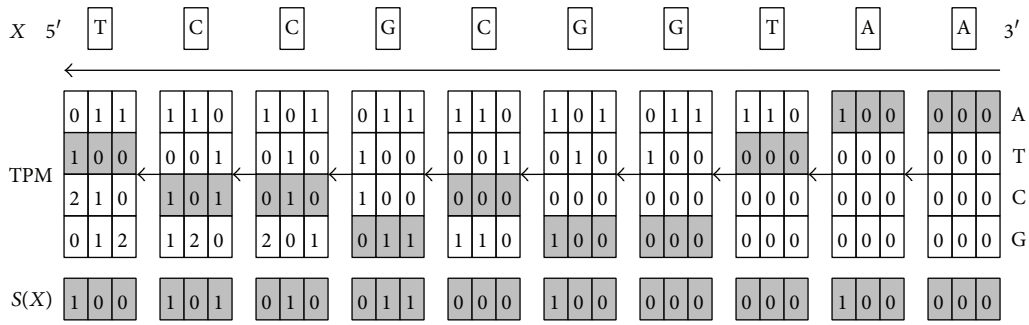

Figure 1: An example of generating the TP sequence for a nucleotide sequence.

around zero point, and for simple TP intensive sequences, the walks should obviously move rightward. A measure, named rightward rate (RR), was used to verify such a point and qualitatively discriminate TP intensive sequences from random ones. However, the RR measure is defined in the positive domain and cannot indicate the potential walk trend that moves leftward. In this work, we proposed a new measure, that is, the symmetrical rightward rate (SRR), to qualitatively investigate the behavior of the TP walks for sequences considering both the leftward trend and rightward trend. After that, a slightly leftward trend (a negative noise) in TP walks is observed, which is not reported in the previous SASR literatures. The cause of such an abnormity is discussed with the probability theory, and a modification of the original method, named mature SASR, is given to correct it. Furthermore, a potential sequence pattern opposite to the TP persistent pattern, that is, the TP anti-persistent pattern, is explored. The applications of the algorithms on simulated datasets show their capabilities in detecting such a potential sequence pattern.

All the involved methods in this work are introduced in Section 2, as well as the preparation of the experimental data. Section 3 demonstrates our experiments, findings, and discussions. A conclusion is reached at the end of the paper.

2. Methods and Materials

2.1. Previous Work: The Original SASR. For a certain base sequence $X = \{x_1, x_2, \ldots, x_N\}$, there is a TP profile, describing the preferred usages of the codons. And the TP profile was presented, in Frenkel and Korotkov's work [11, 12], using a triplet periodicity matrix (TPM). The TPM is a 4×3 matrix, each row i ($i = 1, 2, 3, 4$) stands for a nucleotide base Λ ($\Lambda = A, T, C,$ or G), each column stands for a "period position" j ($j = 1, 2, 3$), and the entry m_{ij} (or $m_{\Lambda j}$) is the count by which the base Λ appears at the period position j. As a previous work [24], in the SASR, the TPM of the posterior subsequence at each position t is calculated recursively from $t = N$ to 1, with the recurrence formula and initial value

$$M_\Lambda(P_X(t)) = \begin{cases} M_\Lambda(P_X(t+1)) \gg 1 & x_{t+1} \neq \Lambda, \\ M_\Lambda(P_X(t+1)) \gg 1 + \{1, 0, 0\} & x_{t+1} = \Lambda, \end{cases}$$
$$M_\Lambda(P_X(N)) = \{0, 0, 0\}. \tag{1}$$

Here, $P_X(t)$ stands for the posterior subsequence of the complete sequence X at position t (excluding position t). $M_\Lambda(P_X(t))$ is the row vector in the TPM of this posterior subsequence for each base Λ ($\Lambda = A, T, C,$ or G). The operation "$V \gg n$" means n times right cyclic shift (RCS) on the triplet row vector V:

$$\{m_1, m_2, m_3\} \xrightarrow{\text{RCS}} \{m_3, m_1, m_2\}. \tag{2}$$

Then, for each position t, a triplet vector s_t, called TP vector, is selected from the TPM of the posterior subsequence, according to the base at the position, that is, x_t. It follows that $s_t = M_{x_t}(P_X(t))$. A sequence of TP vectors is generated as $S(X) = \{s_1, s_2, \ldots, s_N\}$, called the TP sequence. Figure 1 gives an example of generating the TP sequence for a given nucleotide sequence.

The TP walk is then defined as a random walk in the complex plane, generating a moving trace according to the TP sequence. The trace is a sequence $W = \{w_0, w_1, w_2, \ldots, w_N\}$ with the initial value $w_0 = 0$, and for each step $t > 0$,

$$w_t = \begin{cases} w_{t-1} + \dfrac{u(s_t)}{|u(s_t)|} & |u(s_t)| \neq 0, \\ w_{t-1} & |u(s_t)| = 0. \end{cases} \tag{3}$$

Here, the function $u(s_t)$ maps the triplet vector $s_t = \{m_1, m_2, m_3\}$ into a complex number by

$$u(s_t) = m_1 e^{-i2\pi/3} + m_2 e^{-i4\pi/3} + m_3. \tag{4}$$

The above process, that generates a TP walk from the query sequence, is called a SASR process. The TP walk generated from (3) can provide a good visualization of the TP property: for TP intensive regions, the TP walk shows obvious moving trends, while the walk in non-TP regions moves much slower or randomly around stable points. These walk patterns are clues to the discrimination between TP intensive and non-TP regions [24]. Moreover, the walk shifts in direction from a TP intensive region to a neighboring one and the angle of the "corner" indicate the frame shift Δ between the two regions, following a "corner rule" [24].

2.2. The Symmetrical Rightward Rate. According to Chen and Ji [24], the TP walks for simple TP intensive sequences

have an obvious trend to move rightward and those for random sequences move randomly around the zero point. To quantitatively verify this principle in practice, a rightward rate (RR) measure has been presented in Chen and Ji's work. For a given nucleotide sequence, an RR measure is calculated from its TP walk $W = \{w_0, w_1, w_2, \ldots, w_N\}$:

$$\mathrm{RR} = \frac{1}{N} \max \{\mathrm{Re}(w_t) \mid t = 0, 1, 2, \ldots, N\}. \tag{5}$$

Here, $\mathrm{Re}(w)$ stands for the real part of the complex number w. This measure is used to reveal the average speed at which the walk moves rightward in the complex plane.

According to the above definition, an RR measure should not be less than 0 and does not allow revealing the walk trend that moves leftward. However, in some cases, a leftward trend should also be considered. So a symmetrical rightward rate (SRR) is further presented here:

$$\begin{aligned}\mathrm{SRR} = \frac{1}{N} \big[&\max \{\mathrm{Re}(w_t) \mid t = 0, 1, 2, \ldots, N\} \\ &+ \min \{\mathrm{Re}(w_t) \mid t = 0, 1, 2, \ldots, N\}\big].\end{aligned} \tag{6}$$

If a walk has an obvious trend to move rightward, its SRR measure tends to be positive, while a walk to move leftward provides a negative SRR measure. And a walk to move randomly around the zero point has an SRR measure close to 0. The SRR considers both the leftward and rightward trends and is employed to reveal the true behavior of the TP walks in this work.

2.3. Improvement: The Mature SASR. A modification of the original SASR is proposed here, called the mature SASR. In the original SASR, at each position t, the TPM of the posterior subsequence is calculated and the TP vector s_t is selected directly from this TPM, as mentioned previously. In this modification, s_t is selected from a "mature" TPM, instead of from the original matrix. Here, "mature" means that the TPM satisfies

$$\sum_{\Lambda} m_{\Lambda 1} = \sum_{\Lambda} m_{\Lambda 2} = \sum_{\Lambda} m_{\Lambda 3}. \tag{7}$$

A mature TPM $matM$ is maintained with a simple recurrence formula only involving a RCS: $matM_{\Lambda}(P_X(t)) = matM_{\Lambda}(P_X(t+1)) \gg 1$. Besides, the original TPM is still maintained as mentioned before, so that the mature TPM can be updated by copying it, when the original TPM becomes "mature," in every three steps. Figure 2 shows a simple example of generating a TP sequence with this new algorithm.

With this improved method obtaining a TP sequence, the complete algorithm in generating a TP walk is described as shown in Pseudocode 1. And its usage and advantages are shown in Section 3.

2.4. Simulating Random Sequences. In this work, a random sequence dataset is generated, containing 2,000 nucleotide sequences with lengths of 300 bp ~ 5,000 bp. These sequences are unbiasedly random without any periodicity, which are obtained by simply assigning each site in the sequences as nucleotide base Λ ($\Lambda = A, T, C$, or G) with the probability $p_{\Lambda} = 1/4$.

2.5. Simulating TP Antipersistent Sequences. Besides the random sequence dataset, another sequence dataset is generated, containing 2,000 simulated TP antipersistent sequences with lengths of 300 bp~5,000 bp (see the elaboration about TP antipersistent in Section 3). To generate a simulated TP antipersistent DNA sequence with a length of N, the flow chart in Figure 3 is followed. Firstly, a short subsequence at the end (the "seed"), that is, $\{x_{N-L+1}, x_{N-L+2}, \ldots, x_N\}$, is randomly generated. Here, we use the seed length $L = 9$. The TPM of the complete sequence is calculated as follows:

$$\begin{aligned}m_{\Lambda j} = \mathrm{count}\big\{&t \mid 1 \le t \le N, t =_3 j, \\ &x_t \text{ has been assigned as } \Lambda\big\}.\end{aligned} \tag{8}$$

Here, "count" means get the number of the elements in the following set, and "$t =_3 j$" denotes "$t \bmod 3 = j \bmod 3$."

Then, the bases in the anterior part are assigned recursively from position $N - L$ to 1. For each given position t, $1 \le t \le N - L$, x_t is assigned to be base Λ with a probability:

$$\Pr\{x_t = \Lambda\} = \frac{\sum_{j \neq_3 t} m_{\Lambda j}}{\sum_{j \neq_3 t} m_{Aj} + m_{Tj} + m_{Cj} + m_{Gj}}. \tag{9}$$

Here, "$j \neq_3 t$" denotes "$j \bmod 3 \neq t \bmod 3$."

After assigning the base at each position, the TPM of the complete sequence is immediately updated following equation (8), with the newly assigned x_t.

3. Results and Discussions

3.1. Application of the Original SASR to Random Sequences. The original SASR is applied to the simulated random sequences. The distribution of the SRR values of the TP walks is plotted in Figure 4 in the form of its probability density function (PDF). It shows that the distribution is close to the normal distribution with a slight shift to the negative.

The sample mean $\overline{X}$ of the SRR values is -7.95×10^{-3} and the sample standard deviation S is 2.06×10^{-2}. A one-sample t-test with the hypothetical mean $\mu_0 = 0$ obtains a P value of 0. Here, the P value in a one-sample t-test is a statistical term indicating the likelihood to get the observed sample if the population is with the hypothetical mean μ_0. In practice, a t statistic is first calculated:

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}. \tag{10}$$

The sample size $n = 2{,}000$ as mentioned before. Once the t statistic is determined, a P value can be found using a table of values from "Student's t-distribution". A P value of 0 indicates that the distribution is significantly different from the unbiased (with the expectation of 0) normal distribution. So the TP walks for the random sequences slightly move

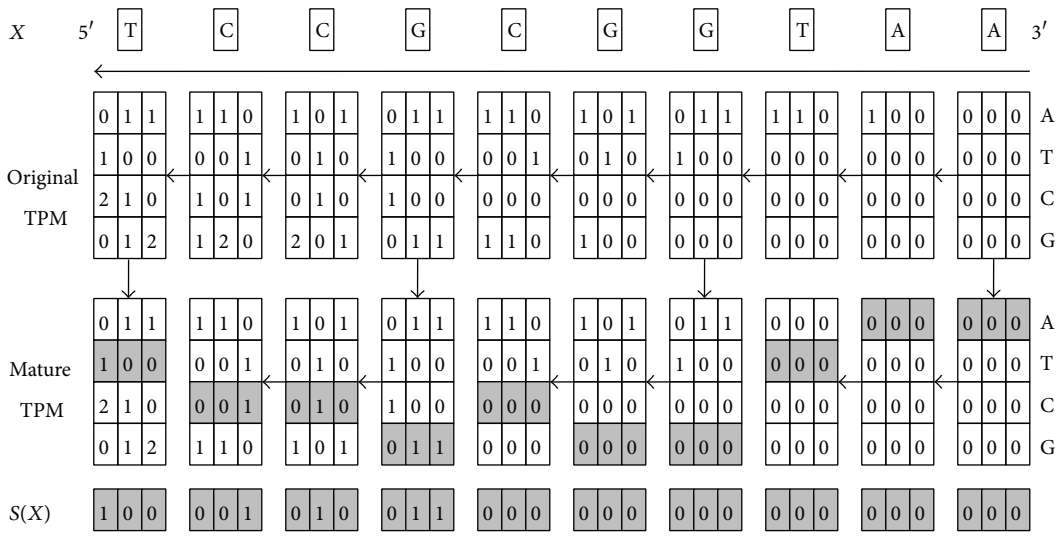

Figure 2: A simple example of generating a TP sequence with the new algorithm.

```
Input: nucleotide sequence x[1,...,N]
Output: TP walk as a complex sequence w[0,...,N]
(1) for each Λ do matM[Λ] = M[Λ] = {0, 0, 0};
(2) for t from N to 1 do {
(3)     if ((N − t) mod 3 == 0) matM[Λ] = M[Λ];
(4)     s[t] = matM[x[t]];
(5)     for each Λ do {
(6)         M[Λ] = M[Λ] ≫ 1;
(7)         if (x[t] == Λ)M[Λ] += {1, 0, 0};
(8)     }
(9) }
(10) w[0].re = 0; w[0].im = 0;
(11) for t from 1 to N do {
(12)         u.re = −0.5 ∗ s[t][0] − 0.5 ∗ s[t][1] + s[t][2];
(13)         u.im = sqrt(3)/2 ∗ s[t][0] − sqrt(3)/2 ∗ s[t][1];
(14)         r = sqrt(u.re ∗ u.re + u.im ∗ u.im);
(15)         w[t] = w[t − 1];
(16)         if (r! = 0) {
(17)                 w[t].re += u.re/r;
(18)                 w[t].im += u.im/r;
(19)         }
(20) }
```

Pseudocode 1

leftward, rather than unbiased random as expected in Chen and Ji's work [24].

The reason for the slightly leftward trend is discussed below. Consider a random sequence $X = \{x_1, x_2, \ldots, x_N\}$. At any position, a certain base Λ ($\Lambda = A, T, C$, or G) appears with a fixed probability p_Λ and $p_A + p_T + p_C + p_G = 1$. Suppose a base Λ appears at position t_0; according to Chen and Ji [24] (also find the original SASR in Section 2), we have the step t_0:

$$s_{t_0} = M_\Lambda\left(P_X\left(t_0\right)\right) = \left\{m_{\Lambda 1}, m_{\Lambda 2}, m_{\Lambda 3}\right\}$$
$$\text{where } m_{\Lambda j} = \text{count}\left\{t \mid x_t = \Lambda, \left(t - t_0\right) =_3 j, t > t_0\right\}. \tag{11}$$

It is easy to find that the random variable $m_{\Lambda j}$ follows the Binomial distribution:

$$m_{\Lambda j} \sim B\left(n_j, p_\Lambda\right). \tag{12}$$

Here, n_j is the count of the positions t that satisfy $t > t_0$ and $(t - t_0) =_3 j$. And the expected value $E(m_{\Lambda j}) = n_j p_\Lambda$. So the expected value of the step is

$$E\left(s_{t_0}\right) = p_\Lambda \cdot \left\{n_1, n_2, n_3\right\}. \tag{13}$$

According to the definition of n_j, although the differences among n_1, n_2, and n_3 are no more than 1, n_3 is always the

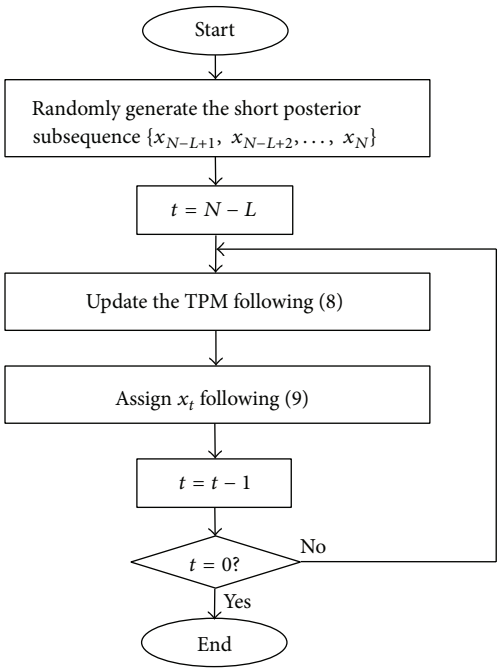

Figure 3: The flow chart to generate a simulated TP antipersistent sequence.

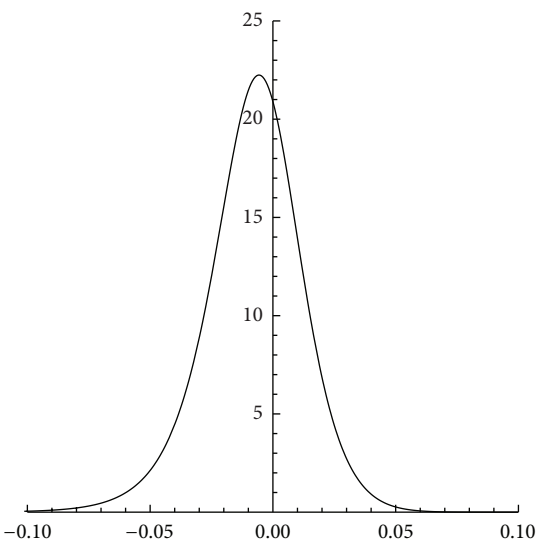

Figure 4: The PDF of the distribution of the SRR values when applying the original SASR to the random sequences.

minimum in the three. According to the mapping from the triplet to the complex number (4), it causes the walk to move leftward slightly for each step and further produces a slightly negative SRR value.

As discussed above, the slightly leftward trend is caused by a negative noise raised by the original SASR method itself. The noise may comprehensively exist in all TP walks. It needs an improved method to eliminate it.

3.2. Application of the Mature SASR to Random Sequences. As mentioned in Section 2, the mature SASR uses a mature TPM

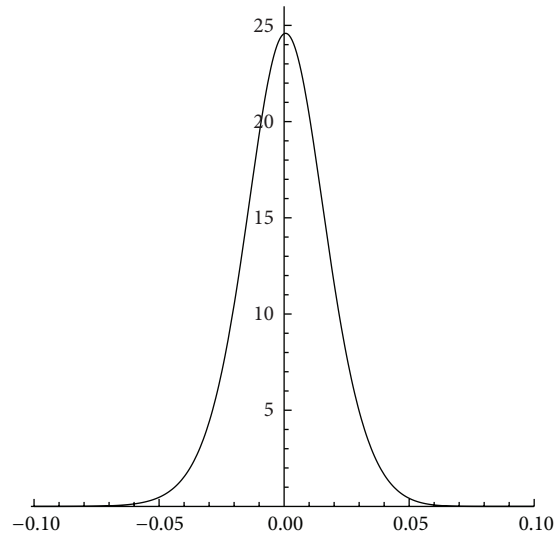

Figure 5: The PDF of the distribution of the SRR values when applying the mature SASR to the random sequences.

instead of the original matrix. The mature TPM is always derived when $n_1 = n_2 = n_3$, and it is supposed to eliminate the noise.

The mature SASR is applied to the random sequences and the distribution of the SRR values is plotted in Figure 5. It shows that it is close to the normal distribution with an expected value of 0. The sample mean of the SRR values is 4.87×10^{-4} and the sample standard deviation is 1.79×10^{-2}. The one-sample t-test obtains a P value of 22.5% (two-tailed), which shows no significant difference from the unbiased distribution. So it is verified that, by using the mature SASR, the TP walks for random sequences are unbiasedly random around the zero point in the complex plane. The negative noise is eliminated by this modification.

It should be pointed out that, compared with the original SASR method, the mature SASR equally eliminates the negative noise that originally exists in all TP walks for both non-TP and TP intensive sequences. Therefore, this modification of the method does not impact the capability of the method in detecting the TP intensive pattern.

3.3. The TP Antipersistent Sequences. The TP profile was presented in Frenkel and Korotkov's work [11, 12] using a triplet periodicity matrix (TPM) as mentioned in Section 2. The TP profiles in the parts of a non-TP sequence have no correlation with each other. It shows a "Brownian pattern" in the sequence. On the other hand, in a simple TP intensive sequence $X = \{x_1, x_2, \ldots, x_N\}$, a certain base Λ appears at position j in the 3 bp period with a probability:

$$\Pr\{x_t = \Lambda, t =_3 j\} = \frac{m_{\Lambda j}}{N}. \tag{14}$$

Kotlar and Lavner's finding [23] suggests that, in coding regions of a given organism, the TP profile, by which nucleotide bases appear in the triplet period, tends to keep

unchanged. It can be considered as a "persistent pattern" in the sequence. That is, for any position t_0,

$$\begin{aligned}&\Pr\{t =_3 j_0 \mid t \le t_0, x_t = \Lambda\}\\&\quad = \Pr\{t =_3 j_0 \mid t > t_0, x_t = \Lambda\}.\end{aligned} \tag{15}$$

Besides the "Brownian pattern" and the "persistent pattern" mentioned above, a theoretically potential pattern is considered, namely, the "antipersistent pattern." For the antipersistent pattern, any part of the sequence has the TP profile opposite to the rest parts. In other words, a certain base Λ avoids appearing at the position j in the 3 bp period, which is preferred in other parts of the sequence. An ideal probability model is built here as a simple case of the TP antipersistency. That is, at any given position t_0 in the sequence, a certain base Λ appears with a probability:

$$\begin{aligned}&\Pr\left\{x_{t_0} = \Lambda \mid t_0 =_3 j_0\right\}\\&\quad = \Pr\{x_t = \Lambda \mid t > t_0, t \ne_3 j_0\}.\end{aligned} \tag{16}$$

So that, for any position t_0 presenting Λ in the sequence,

$$\begin{aligned}&\Pr\left\{t_0 =_3 j_0 \mid x_{t_0} = \Lambda\right\}\\&\quad = \frac{\Pr\{t_0 =_3 j_0\} \cdot \Pr\left\{x_{t_0} = \Lambda \mid t_0 =_3 j_0\right\}}{\sum_j \Pr\{t_0 =_3 j\} \cdot \Pr\left\{x_{t_0} = \Lambda \mid t_0 =_3 j\right\}}\\&\quad \approx \frac{\Pr\{x_t = \Lambda \mid t > t_0, t \ne_3 j_0\}}{\sum_j \Pr\{x_t = \Lambda \mid t > t_0, t \ne_3 j\}}\\&\qquad \left(\text{for } j = 1,2,3 \ \Pr\{t_0 =_3 j\} \approx \frac{1}{3}\right)\\&\quad \approx \frac{\Pr\{t \ne_3 j_0 \mid t > t_0\} \cdot \Pr\{x_t = \Lambda \mid t > t_0, t \ne_3 j_0\}}{\sum_j \Pr\{t \ne_3 j \mid t > t_0\} \cdot \Pr\{x_t = \Lambda \mid t > t_0, t \ne_3 j\}}\\&\qquad \left(\text{for } j = 1,2,3 \ \Pr\{t \ne_3 j \mid t > t_0\} \approx \frac{2}{3}\right)\\&\quad = \frac{1}{2}\Pr\{t \ne_3 j_0 \mid t > t_0, x_t = \Lambda\}.\end{aligned} \tag{17}$$

Therefore, this model is found to be opposite to the "persistent pattern" of (15). In practice, we simulate such TP antipersistent sequences by the method mentioned in Section 2.

3.4. Applications of the Algorithms to TP Antipersistent Sequences. The original SASR is first applied to the simulated TP antipersistent sequences (see Section 2). The distribution of the SRR values of the TP walks is plotted in Figure 6, compared with that for the random sequences. It shows an obvious difference between these two distributions.

The simulation above reveals a leftward moving trend of the TP walks for TP antipersistent sequences. The reason of such a behavior is discussed as below. Consider any short section containing three sequential positions $t_0 - 2$, $t_0 - 1$, and t_0 (t_0 is a multiple of 3; i.e., $t_0 \bmod 3 = 0$) in a sequence with the TP antipersistent probability model mentioned previously. The posterior subsequences at these three positions share a similar TPM with a shift:

$$M_\Lambda\left(P_X\left(t_0 - i\right)\right) \approx M_\Lambda\left(P_X\left(t_0\right)\right) \gg i, \quad (i = 0,1,2). \tag{18}$$

Meanwhile, according to (9), base Λ appears at these positions with a probability:

$$\Pr\left\{x_{t_0-i} = \Lambda\right\} \approx \frac{\sum_{j \ne 3-i} m_{\Lambda j}}{\sum_{j \ne 3-i} m_{Aj} + m_{Tj} + m_{Cj} + m_{Gj}}. \tag{19}$$

Here, $m_{\Lambda j}$ stands for the entry in the TPM of the posterior subsequence at position t_0; that is, $M_\Lambda(P_X(t_0)) = \{m_{\Lambda 1}, m_{\Lambda 2}, m_{\Lambda 3}\}$. Meanwhile, we have

$$\sum_\Lambda m_{\Lambda 1} \approx \sum_\Lambda m_{\Lambda 2} \approx \sum_\Lambda m_{\Lambda 3} \approx \frac{N - t_0}{3}. \tag{20}$$

Hence, these three steps in the walk move to

$$\begin{aligned}&E\left(\sum_i \frac{s_{t_0-i}}{\left|u\left(s_{t_0-i}\right)\right|}\right)\\&\approx E\left(\sum_i \frac{M_{x_{t_0-i}}\left(P_X\left(t_0\right)\right) \gg i}{\left|u\left(M_{x_{t_0-i}}\left(P_X\left(t_0\right)\right)\right)\right|}\right)\\&\qquad \text{(according to (11) and (18))}\\&= \sum_i \sum_{\Lambda=A,T,C,G} \left(\frac{\sum_{j \ne 3-i} m_{\Lambda j}}{\sum_{j \ne 3-i} m_{Aj} + m_{Tj} + m_{Cj} + m_{Gj}} \cdot \frac{M_\Lambda\left(P_X\left(t_0\right)\right) \gg i}{\left|u\left(M_\Lambda\left(P_X\left(t_0\right)\right)\right)\right|}\right)\\&\qquad \text{(according to (19))}\\&\approx \frac{3}{2\left(N - t_0\right)} \sum_{\Lambda=A,T,C,G} \frac{\sum_i \left[\left(\sum_{j \ne 3-i} m_{\Lambda j}\right) \cdot M_\Lambda\left(P_X\left(t_0\right)\right) \gg i\right]}{\left|u\left(M_\Lambda\left(P_X\left(t_0\right)\right)\right)\right|}\\&\qquad \text{(according to (20))}\\&= \frac{3}{2\left(N - t_0\right)} \sum_{\Lambda=A,T,C,G} \frac{\{\alpha_1, \alpha_2, \alpha_3\}}{\left|u\left(M_\Lambda\left(P_X\left(t_0\right)\right)\right)\right|},\end{aligned} \tag{21}$$

where

$$\begin{aligned}\alpha_1 = \alpha_2 &= m_{\Lambda 1} m_{\Lambda 2} + m_{\Lambda 2} m_{\Lambda 3} + m_{\Lambda 3} m_{\Lambda 1}\\&\quad + m_{\Lambda 1}^2 + m_{\Lambda 2}^2 + m_{\Lambda 3}^2,\\\alpha_3 &= 2 m_{\Lambda 1} m_{\Lambda 2} + 2 m_{\Lambda 2} m_{\Lambda 3} + 2 m_{\Lambda 3} m_{\Lambda 1}.\end{aligned} \tag{22}$$

Obviously, in this case, we have $\alpha_1 = \alpha_2 \ge \alpha_3$. Therefore, in (21), the first two elements of the expected vector dominate the third one. According to (4), it causes the TP walk to move leftward in the complex plane.

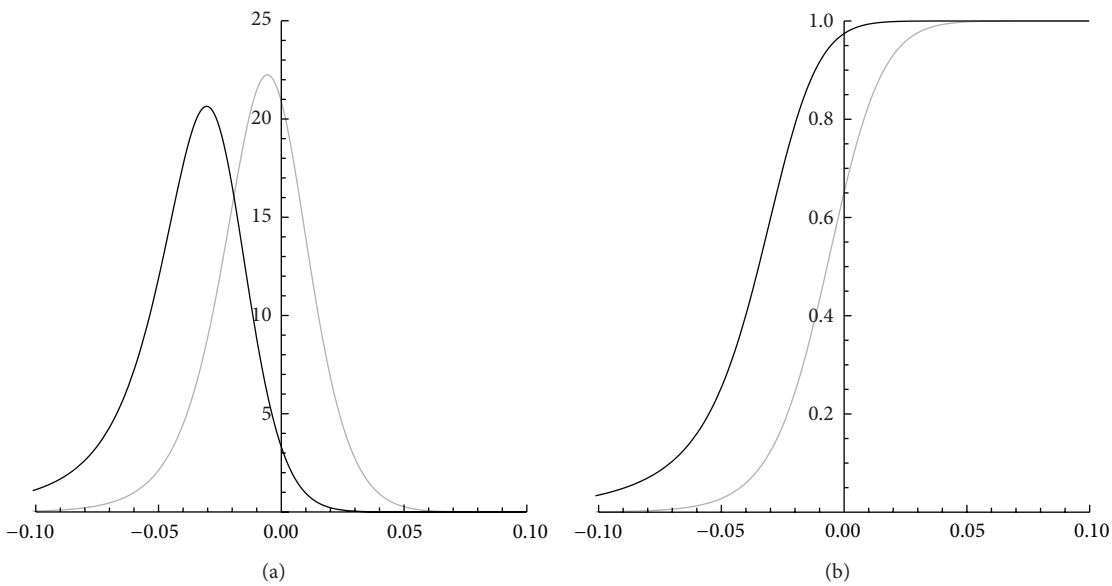

Figure 6: The distribution of the SRR values when the original SASR is applied to the simulated TP antipersistent sequences (black) compared with those for the random sequences (gray). (a) The probability density function (PDF). (b) The cumulative distribution function (CDF).

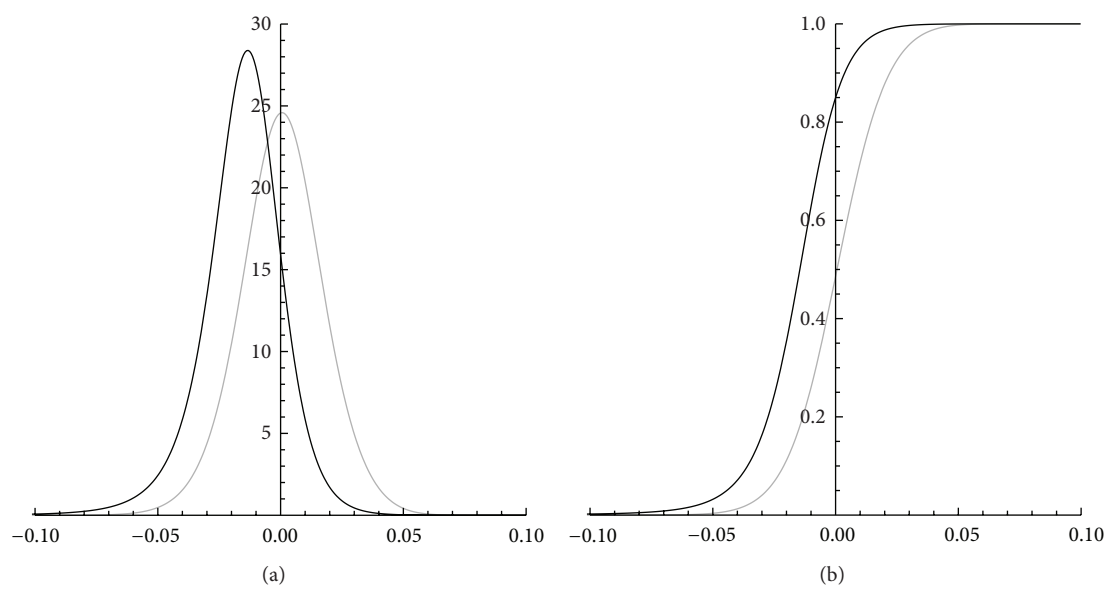

Figure 7: The distribution of the SRR values when the mature SASR is applied to the simulated TP antipersistent sequences (black) compared with those for the random sequences (gray). (a) The probability density function (PDF). (b) The cumulative distribution function (CDF).

However, since the TP walks, from the original SASR, comprehensively contain a negative noise as mentioned before, it is difficult to determine to what extent the noise has impacted the gap between the two distributions in Figure 6. Therefore, to visualize the real gap between the two patterns, it needs the mature SASR, in which the noise has been eliminated.

The mature SASR is then applied to the simulated TP antipersistent sequences. The distribution of the SRR values of the TP walks is plotted in Figure 7, compared with that for the random sequences. The PDF curve for the simulated TP antipersistent sequences is on the left side to that for the random sequences, and the cumulative distribution function (CDF) curves indicate that there are 85% simulated sequences with negative SRR values, while the SRR values of the random sequences distribute fifty-fifty in negative and positive areas. It is found that the sample mean $\overline{X}$ and the sample deviation S of the 2,000 SRR values for the simulated anti-TP dataset are -1.57×10^{-2} and 1.73×10^{-2}, respectively. A P value of 0 indicates the significant difference between this distribution and that for random sequences. It must be noticed that, although the gap between the two distributions

is less than that in Figure 6, such a gap is completely due to the difference between the sequence patterns, without any noise. So the mature SASR is more suitable in visualizing the TP antipersistency than the original SASR.

The results from the simulations and the discussions above indicate that the mature SASR is able to discriminate TP antipersistent sequences from random sequences. The antipersistent pattern can be identified according to a leftward moving trend in the TP walk.

4. Conclusions

In this work, a new measure, that is, SRR, is presented to qualitatively investigate the behavior of the original SASR's outputs, that is, the TP walks, for sequences considering both the leftward trend and rightward trend. After that, for random sequences, an abnormal behavior of the walks from the original SASR is revealed: the TP walks for the random sequences slightly move leftward, rather than unbiased random as expected in Chen and Ji's work [24]. This abnormity is caused by a negative noise raised by the original SASR method itself. And the noise comprehensively exists in all TP walks.

A modification of the original SASR, that is, the mature SASR, is then given in order to eliminate the noise and correct the behavior of the TP walks, without impacting the capability of the method in detecting the TP intensive pattern. The application to the simulated random sequences verifies that, by using the mature SASR, the TP walks for random sequences are unbiasedly random around the zero point in the complex plane.

Furthermore, a potential sequence pattern opposite to the TP persistent pattern, that is, the TP antipersistent pattern, is explored. The applications of the algorithms on simulated datasets show their capabilities in detecting such a potential sequence pattern. The mature SASR is said to be an effective tool for the visualization of TP-related features, including non-TP, TP persistency, and TP antipersistency.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The project was supported by the Natural Science Foundation of Fujian Province, China (Grant no. 2012J05114), and Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing (Fuzhou University).

References

[1] K. Song, Z. Zhang, T. Tong, and F. Wu, "Classifier assessment and feature selection for recognizing short coding sequences of human genes," *Journal of Computational Biology*, vol. 19, no. 3, pp. 251–260, 2012.

[2] P. K. Sree and I. R. Babu, "AIS-INMACA: a novel integrated MACA based clonal classifier for protein coding and promoter region prediction," *Journal of Bioinformatics and Comparative Genomics*, vol. 1, pp. 1–7, 2014.

[3] J. Khatun, Y. Yu, J. A. Wrobel et al., "Whole human genome proteogenomic mapping for ENCODE cell line data: Identifying protein-coding regions," *BMC Genomics*, vol. 14, no. 1, article 141, 2013.

[4] J. Mena-Chalco, H. Carrer, Y. Zana, and R. M. Cesar Jr., "Identification of protein coding regions using the modified Gabor-wavelet transform," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 2, pp. 198–207, 2008.

[5] A. D. Haimovich, B. Byrne, R. Ramaswamy, and W. J. Welsh, "Wavelet analysis of DNA walks," *Journal of Computational Biology*, vol. 13, no. 7, pp. 1289–1298, 2006.

[6] Y. L. Orlov, R. Te Boekhorst, and I. I. Abnizova, "Statistical measures of the structure of genomic sequences: entropy, complexity, and position information," *Journal of Bioinformatics and Computational Biology*, vol. 4, no. 2, pp. 523–536, 2006.

[7] J. H. Do and D. K. Choi, "Computational approaches to gene prediction," *The Journal of Microbiology*, vol. 44, no. 2, pp. 137–144, 2006.

[8] M. Stanke and S. Waack, "Gene prediction with a hidden Markov model and a new intron submodel," *Bioinformatics*, vol. 19, no. 2, pp. ii215–ii225, 2003.

[9] C. T. Zhang and J. Wang, "Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve," *Nucleic Acids Research*, vol. 28, no. 14, pp. 2804–2814, 2000.

[10] W. Li, "The complexity of DNA," *Complexity*, vol. 3, pp. 33–37, 1997.

[11] F. E. Frenkel and E. V. Korotkov, "Classification analysis of triplet periodicity in protein-coding regions of genes," *Gene*, vol. 421, no. 1-2, pp. 52–60, 2008.

[12] F. E. Frenkel and E. V. Korotkov, "Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes," *DNA Research*, vol. 16, no. 2, pp. 105–114, 2009.

[13] J. W. Fickett, "The gene identification problem: an overview for developers," *Computers and Chemistry*, vol. 20, no. 1, pp. 103–118, 1996.

[14] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Research*, vol. 10, no. 17, pp. 5303–5318, 1982.

[15] R. K. Azad and M. Borodovsky, "Probabilistic methods of identifying genes in prokaryotic genomes: connections to the HMM theory," *Briefings in Bioinformatics*, vol. 5, no. 2, pp. 118–130, 2004.

[16] J. Henderson, "Finding genes in DNA with a Hidden Markov model," *Journal of Computational Biology*, vol. 4, no. 2, pp. 127–141, 1997.

[17] Y. H. Cao, W. W. Tung, J. B. Gao, and Y. Qi, "Recurrence time statistics: versatile tools for genomic DNA sequence analysis," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 3, pp. 677–696, 2005.

[18] J. B. Gao, Y. Qi, Y. H. Cao, and W. Tung, "Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences," *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, pp. 139–146, 2005.

[19] X. Y. Jiang, D. Lavenier, and S. S. Yau, "Coding region prediction based on a universal DNA sequence representation method," *Journal of Computational Biology*, vol. 15, no. 10, pp. 1237–1256, 2008.

[20] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, and L. Marcourt, "Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences," *Journal of Theoretical Biology*, vol. 206, no. 3, pp. 323–326, 2000.

[21] D. Anastassiou, "Frequency-domain analysis of biomolecular sequences," *Bioinformatics*, vol. 16, no. 12, pp. 1073–1081, 2000.

[22] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 8–20, 2001.

[23] D. Kotlar and Y. Lavner, "Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions," *Genome Research*, vol. 13, no. 8, pp. 1930–1937, 2003.

[24] B. Chen and P. Ji, "Visualization of the protein-coding regions with a self adaptive spectral rotation approach," *Nucleic Acids Research*, vol. 39, no. 1, article e3, 2011.

[25] B. Chen and P. Ji, "Numericalization of the self adaptive spectral rotation method for coding region prediction," *Journal of Theoretical Biology*, vol. 296, pp. 95–102, 2012.

[26] M. Bellani, J. Epps, and G. A. Huttley, "A comparison of periodicity profile methods for sequence analysis," in *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS '12)*, pp. 78–81, Washington, DC, USA, December 2012.

[27] Y. M. Suvorova, V. M. Rudenko, and E. V. Korotkov, "Detection change points of triplet periodicity of gene," *Gene*, vol. 491, no. 1, pp. 58–64, 2012.

[28] A. M. Michel, K. R. Choudhury, A. E. Firth, N. T. Ingolia, J. F. Atkins, and P. V. Baranov, "Observation of dually decoded regions of the human genome using ribosome profiling data," *Genome Research*, vol. 22, no. 11, pp. 2219–2229, 2012.