Periodicity of base correlation in nucleotide sequence

Weijiang Lee* and Liaofu Luo
CCAST (World Laboratory), P. O. Box 8730, Beijing 100080, China
and Department of Physics, Inner Mongolia University, Hohhot, 010021, China†
(Received 6 January 1997)

The correlation spectrum of a symbolic sequence is defined. The harmonic heights in the spectrum of a nucleotide sequence are calculated rigorously. A model of coding sequences is proposed and the origin of 3-periodicity is explained. The possible occurrence of a hidden periodicity of base correlation with no peak in the spectrum is discussed. [S1063-651X(97)13106-3]

PACS number(s): 87.10.+e, 02.90.+p

I. CORRELATION SPECTRUM

Many attempts have been made to find the informational content in DNA sequences. To investigate the correlation of a symbolic sequence, an essential problem is how to characterize the symbolic sequence by a numerical sequence [1–5]. We first delineate the problem in a slightly general manner.

Consider a symbolic sequence

$$x_0 x_1 \ldots x_{N-1} \tag{1}$$

consisting of N letters belonging to a given finite alphabet Λ. For DNA sequences, Λ={adenine (A), cytosine (C), guanine (G), and thymine (T)}. There are also other choices of the alphabet Λ for simplified representations of DNA sequences. For example, {R,Y} for purine (R)–pyrimidine (Y) representation; {S,W} for strong bond (G,C)–weak bond (A,T) representation; {A,A}, etc. For all $a,\ b \in \Lambda$, we introduce the inner product $S_{ab}=(a,b)$, which, with a meaning of the measure of similarity between a and b, can be evaluated according to their biochemical and physical properties, or according to practical purposes. For example,

$$S_{ab}=\delta(a,b), \quad \text{for all } a,b \in \Lambda \tag{2}$$

means the self-similarity of all bases and absolute dissimilarity between any two different bases, where $\delta(a,b)$ is the Kronecker symbol. In addition, the symmetry of the similarity between any two bases a and b is presumed: $S_{ab}=S_{ba}$.

Based on the definition of an inner product, we define the correlation function $c(\tau)$ as follows [2]:

$$c(\tau)=\frac{1}{N}\sum_{j=0}^{N-1}(x_j,x_{j+\tau}) \tag{3}$$

where, for the sake of convenience and simplicity, the periodic boundary condition

$$x_{N+j}=x_j \quad (j=0,1,2,\ldots,N-1) \tag{4}$$

is adopted. Easily shown that

$$c(\tau)=\frac{1}{N}\sum_{a,b\in\Lambda}\mu_{ab}(\tau)S_{ab}, \tag{5}$$

where $\mu_{ab}(\tau)$ is the number of base pair a and b at distance τ along the sequence. When N is large enough, the frequencies $\mu_{ab}(\tau)/N$ may be replaced by the joint probability $P_{ab}(\tau)$ (''the thermodynamic approximation''). Speaking rigorously, $P_{ab}(\tau)$ is the probability that base a appears at some site j and meanwhile base b appears at site $(j+\tau)$ of the same sequence. Under the thermodynamic approximation, Eq. (5) is written as

$$c(\tau)=\sum_{a,b\in\Lambda}P_{ab}(\tau)S_{ab}. \tag{6}$$

So the correlation function is a linear combination of the joint probabilities. In other words, the joint probabilities constitute bases of the correlation function. For a given pair of bases a and b, by setting $S_{ab}=S_{ba}=1$ and other inner products 0, we see that $[P_{ab}(\tau)+P_{ba}(\tau)]/2$ itself is also a correlation function. On the other hand, by setting $S_{aa}=1$ and other inner products 0, the obtained $c(\tau)$ is identical to that introduced in Ref. [1] where it is deduced from the sequence $\{U_j\}$ ($U_j=1$ if $x_j=a$, and $U_j=0$ otherwise) and the inner product replaced by conventional multiplication.

To investigate the periodicity of the correlation function we use the spectrum

$$V(k)=\frac{1}{N}\sum_{r=0}^{N-1}c(\tau)\exp\left(i\,\frac{2\pi k\tau}{N}\right). \tag{7}$$

Because of the periodic boundary condition and the symmetry of the similarity factors, one has

$$c(\tau)=c(N-\tau), \quad \forall\tau. \tag{8}$$

So $V(k)$ is real and symmetric

$$V(k)=V(N-k) \quad \forall k. \tag{9}$$

In Refs. [1–3] many resonances have been found in the correlation spectrum. Especially, for most coding sequences there exists a strong resonance at $k/N=1/3$. Though the sharp peak at 1/3 is anticipated to be related to the nucleotide triplet (codon) that specifies one of the 20 amino acids, but no thorough explanation has been given. The problem is

*Electronic address: wjlee@nmg2.imu.edu.cn
†Mailing address.

seemingly more difficult since the base correlation in most coding sequences is of short-range property [4,6]. On the other hand, apart from a sharp peak at 1/3 what is the meaning of the other resonances in the correlation spectrum? In this paper we shall investigate the origin and property of resonances in the spectrum of DNA sequence.

II. SOME RIGOROUS RESULTS ON HARMONIC HEIGHTS

Let $\mu_{ab}(\tau)$ be the number of base pair ab that is separated by distance τ along the sequence

$$\mu_{ab}(\tau)=\sum_{i=0}^{N-1}\delta(x_j,a)\,\delta(x_{i+\tau},b) \tag{10}$$

and define the correlation spectrum of base pair ab as

$$V_{ab}(k)=\frac{1}{N}\sum_{\tau=0}^{N-1}\frac{\mu_{ab}(\tau)+\mu_{ba}(\tau)}{2}\exp\left(i\,\frac{2\pi k\tau}{N}\right). \tag{11}$$

The overall correlation spectrum is then written as

$$V(k)=\sum_{a,b\in\Lambda}V_{ab}(k)S_{ab}.$$

If the sequence length is divisible by some small integer m, then the mth harmonic of the spectrum can be easily calculated through simple counting. Suppose $N=nm$. We first divide the sequence into m subsequences. The qth subsequence is

$$x_q x_{q+m} X_{q+2m}\ldots x_{q+(n-1)m},\quad q=0,1,2,\ldots,m-1.$$

Denote the number of base a in the kth subsequence by $N_k^{(m)}(a)$, $k=0,1,2,\ldots,m-1$

$$\begin{aligned}\sum_{r \bmod m=q}\mu_{ab}(\tau)&=\sum_{j=0}^{n-1}\sum_{i=0}^{N-1}\delta(x_i,a)\,\delta(x_{mj+q+i},b)\\&=\sum_{r=0}^{n-1}\sum_{k=0}^{m-1}\delta(x_{mr+k},a)\\&\quad\times\sum_{j=0}^{n-1}\delta(x_{m(j+r)+q+k},b)\\&=\sum_{k=0}^{m-1}N_{q+k}^{(m)}(b)\sum_{r=0}^{n-1}\delta(x_{mr+k},a)\\&=\sum_{k=0}^{m-1}N_k^{(m)}(a)N_{q+k}^{(m)}(b)\end{aligned} \tag{12}$$

in which for convenience, $N_k^{(m)}(a)=N_{k-m}^{(m)}(a)$ is assumed when $k\geqslant m$. According to the definition of correlation spectrum, we obtain

$$V_{ab}\left(\frac{N}{m}\right)=\frac{1}{2N}\sum_{q=0}^{m-1}\cos\frac{2\pi q}{m}\sum_{k=0}^{m-1}[N_k^{(m)}(a)N_{k+q}^{(m)}(b)+N_k^{(m)}(b)N_{k+q}^{(m)}(a)], \tag{13}$$

if $N \bmod m=0$. From Eq. (13), the following theorems can easily be deduced.

Theorem 1. The sufficient condition for period-m peak vanishing in the correlation spectrum $V_{ab}(N/m)$ of base pair a and b for a sequence with length N ($N \bmod m=0$) is the uniform distribution of base a or b in all the m subsequences, namely,

$$N_0^{(m)}(a)=N_1^{(m)}(a)=\cdots=N_{m-1}^{(m)}(a)$$

or

$$N_0^{(m)}(b)=N_1^{(m)}(b)=\cdots=N_{m-1}^{(m)}(b). \tag{14}$$

Theorem 2. The first three harmonic heights in the correlation spectrum $V_{ab}(N/m)$ ($m=2,3,4$) are

$$V_{ab}\left(\frac{N}{2}\right)=\frac{1}{N}[N_0^{(2)}(a)-N_1^{(2)}(a)][N_0^{(2)}(b)-N_1^{(2)}(b)],\quad \text{if } N \bmod 2=0; \tag{15}$$

$$V_{ab}\left(\frac{N}{3}\right)=\frac{1}{2N}\left[3\sum_{k=0}^{3}N_k^{(3)}(a)N_k^{(3)}(b)-N(a)N(b)\right],\quad \text{if } N \bmod 3=0; \tag{16}$$

[where $N(a)$ is the total number of base a in the whole sequence] and

$$V_{ab}\left(\frac{N}{4}\right)=\frac{1}{N}\sum_{k=0}^{1}[N_k^{(4)}(a)-N_{k+2}^{(4)}(a)][N_k^{(4)}(b)-N_{k+2}^{(4)}(b)],\quad \text{if } N \bmod 4=0. \tag{17}$$

Theorem 3. The necessary and sufficient condition for a period 2 resonance peak existing in the correlation spectrum $V_{ab}(N/2)$ is that the nucleotides are not uniformly distributed in the two subsequences.

In general cases of base pair ab, there are no similar necessary-sufficient condition for the period other than 2. However, if $a=b$, which is the case that was investigated in most literature [1,4,5], we have

Theorem 4. For the sequence $\{U_j\}$ ($U_j=0,1$) with length N ($N/\text{mod}m=0$), the correlation spectrum at $k=N/m$

$$V\left(\frac{N}{m}\right)=\frac{1}{N}\sum_{q=0}^{m-1}\cos\frac{2\pi q}{m}\sum_{k=0}^{m-1}N_k^{(m)}N_{k+q}^{(m)},\quad \text{if } N \bmod m=0, \tag{18}$$

where $N_k^{(m)}=$ the number of 1 occurring in the kth subsequence.

Theorem 5. If and only if $U_j=1$ is not uniformly distributed in the three subsequences, period 3 resonance peak exists in the correlation spectrum $V(k=N/3)$.

Theorem 6. The necessary and sufficient condition for $V(N/4)=0$ for sequence $\{U_j\}$ ($U_j=0,1$) is

$$N_0^{(4)}=N_2^{(4)}\quad\text{and}\quad N_1^{(4)}=N_3^{(4)}. \tag{19}$$

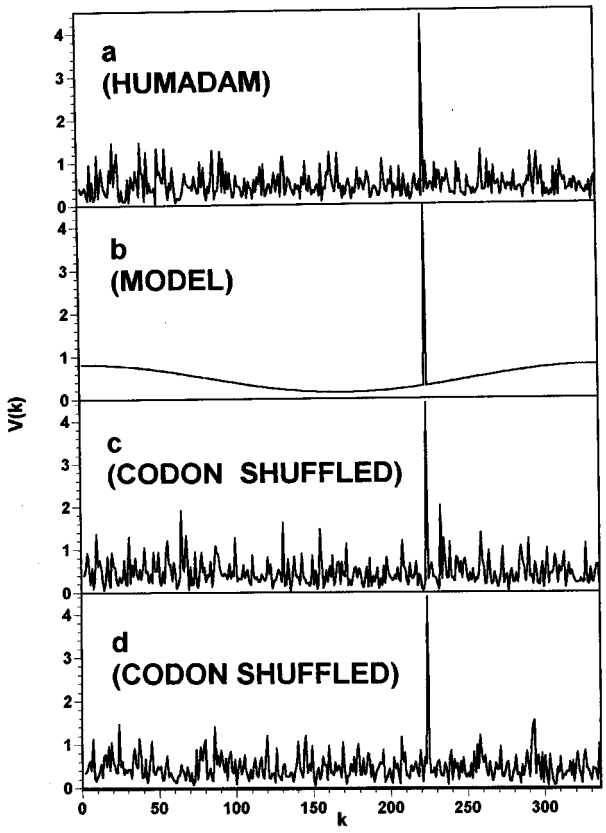

FIG. 1. (a) is a typical example for the correlation spectrum of DNA coding sequences. The sequence used here, HUMADAM, is taken from Entrez 8.0(1993). (b) is based on the theoretical calculation by use of the proposed model, namely, from Eqs. (22), (23), and (25). We see that (b) has a similar shape as (a). By random rearrangement of codons in the sequence, we obtained a number of codon-shuffled sequences. (c) and (d) are two examples of the correlation spectra of these sequences, which have the same height of the 1/3 resonance peaks as the original sequence (a).

III. A MODEL OF CODING SEQUENCES

Figure 1 gives a typical example of the spectrum of the coding sequence. In calculation the appropriate choice of the values of inner products have been made. In these calculations, the similarity factors

$$\{S_{ab}\}=\begin{array}{c} \begin{matrix} T & C & A & G \end{matrix} \\ \begin{bmatrix} 1 & \frac{1}{2} & 0 & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{4} & \frac{1}{2} \\ 0 & \frac{1}{4} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} \begin{matrix} T \\ C \\ A \\ G \end{matrix} \end{array}$$

have been taken [2]. We find that there is a sharp peak at $k/N=1/3$. The peak is typical to most of the coding sequences [1–3]. The peak height can be calculated by Eq. (15). To study the background of the peaks more carefully we introduce the following model.

For brevity, consider the sequence $\{U_j\}_{j=0}^{N-1}$ written by 1 and 0 at first. To answer the existence of the reading frame in the coding sequence, we divide the sequence with length $N=3n$ into n triplets. The probabilities of 1 occurring in the three positions of a triplet are supposed to P_{1**}, P_{*1*} and $P^{**}{}_1$, respectively, where * denotes an arbitrary base. Since $c(\tau)$ is related to the probability of letter 1 occurring in the sequence we have

$$c(0)=(P_{1*}+P_{*L*}+P_{**1})/3,$$

$$c(1)=(P_{11*}+P_{*11}+P_{**11**})/3,$$

$$c(2)=(P_{1*1}+P_{*1*1**}+P_{**1*1*})/3, \text{ etc.} \quad (20)$$

Note that in the expression of joint probability the reading frame has been indicated. Through statistical analyses of nucleic acid sequences it has been shown that the base correlations in most coding sequences are of short-range nature [4,6]. So we assume that the base correlation takes place only inside a triplet and the occurrence of a pair of bases in different triplets is independent. Then Eq. (20) can be simplified to

$$c(0)=(P_{1**}+P_{*1*}+P_{**1})/3,$$

$$c(1)=(P_{11*}+P_{*11}+P_{**1}P_{1**})/3=c(N-1),$$

$$c(2)=(P_{1*1}+P_{*1*}P_{1**}+P_{**1}P_{*1*})/3=c(N-2),$$

$$c(3)=c(6)=\cdots=c(N-3)$$
$$=(P_{1**}P_{1**}+P_{*1*}P_{*1*}+P_{**1}P_{**1})/3$$

$$c(4)=c(7)=\cdots=c(N-4)=c(5)=c(8)=\cdots=c(N-5)$$
$$=(P_{1**}P_{*1*}+P_{*1*}P_{**1}+P_{**1}P_{1**})/3. \quad (21)$$

We see that the periodicity does occur. By use of Eq. (7) we obtain the spectrum

$$V(k)=n\delta_{kn}[c(3)-c(4)]+R_k \quad (n=N/3,\ 1<k<N/2). \quad (22)$$

R_k comes from the breaking of periodicity of correlation function due to the short-range correlation in a codon

$$R_k=[c(0)-c(3)]+2[c(1)-c(4)]\cos(2\pi k/N)$$
$$+2[c(2)-c(5)]\cos(4\pi k/N) \quad (23)$$

Inserting Eq. (21) into Eq. (22) one has the peak height at $k/N=1/3$ proportional to $N/3$ and

$$c(3)-c(4)=\frac{3}{2}\left[\frac{P_{1**}P_{1**}+P_{*1*}P_{*1*}+P_{**1}P_{**1}}{3}\right.$$
$$\left.-\left(\frac{P_{1**}+P_{*1*}+P_{**1}}{3}\right)^2\right]=\frac{3}{2}\,\sigma_1 \quad (24)$$

(σ_1 means the deviation of letter 1 occurring in the three positions of a codon.) The first term on the right-hand side of Eq. (22) is identical to Eq. (18). The latter is an exact result of peak height. So the error brought in the approximation of no correlation between triplets in the model, cancels the R_k term in Eq. (22) at $k=n$ exactly. Equation (22) gives the correlation spectrum the profile which is consistent with the coding sequence data. (See Fig. 1.) The above discussion for sequence $\{U_j\}_{j=0}^{N-1}$ can be generalized to the symbolic se-

quence of several letters. The model can also be generalized to the case of the correlation occurring only in a range of $m(\neq 3)$ bases instead of codon triplet.

IV. THE HIDDEN PERIODICITY

As seen from Fig. 1 there are many peaks in the correlation spectrum of a nucleotide sequence. Each peak shows a kind of periodicity of base correlation. To find the periodicity is very important for understanding the meaning of the sequence. However, is there any periodicity which corresponds to no peak in the spectrum? If any, how do you find these hidden periodicities? For m-periodicity we can break the sequence into m subsequences and deduce the peak height at N/m. [See Eq. (13).] The periodicity means the inhomogenous distribution of base a or b in the m subsequences. But as stated in Theorem 1, the inhomogenous distribution of bases is only necessary, but not a sufficient condition, for $V_{ab}(N/m)\neq 0\ (m>2)$. To be definite, consider $m=4$. Suppose a kind of 4-periodicity with

$$N_0^{(4)}(a)=N_2^{(4)}(a)\neq N_1^{(4)}(a)=N_3^{(4)}(a) \tag{25a}$$

and/or

$$N_0^{(4)}(b)=N_2^{(4)}(b)\neq N_1^{(4)}(b)=N_3^{(4)}(b). \tag{25b}$$

From Eq. (17) we know that the peak is equal to zero. So the 4-periodicity satisfying Eq. (25) is a hidden periodicity which could not be found by spectral analysis. In fact, we find that the hidden 4-periodicity does exist in some intron sequences as an important symmetry [7]. In general, the hidden m-periodicity can be found through the inhomogeneous distribution of bases in the m subsequences but with Eq. $(13)=0$. The interference elimination of base-number components with different q ($q=0$ to $m-1$) is the origin of the spectrum peak vanishing at $k=N/m$.

[1] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[2] L. F. Luo and F. M. Ji, Acta Sci. Nat. Univ. Intramongolicae **26**, 419 (1995).
[3] V. R. Chechetkin and A. Yu. Turgin, J. Phys. A **27**, 4875 (1994).
[4] S. V. Buldyrev *et al.*, Phys. Rev. E **51**, 5084 (1995).
[5] A. Arneodo *et al.*, Phys. Rev. Lett. **74**, 3293 (1995).
[6] L. F. Luo and H. Li, Bull. Math. Biol. **53**, 345 (1991).
[7] F. M. Ji and J. D. Zhao (private communications).