Statistical correlation of nucleotides in a DNA sequence

Liaofu Luo*
Department of Physics, Inner Mongolia University, Hohhot 010021, China

Weijiang Lee†
CCAST (World Laboratory), P.O. Box 8730, Beijing 100080, China
and Department of Physics, Inner Mongolia University, Hohhot 010021, China

Lijun Jia, Fengmin Ji, and Lu Tsai
Department of Physics, Inner Mongolia University, Hohhot 010021, China

(Received 26 November 1997; revised manuscript received 17 February 1998)

We review methods in the study of nucleotide correlation in DNA sequence, and demonstrate two basic properties of the correlation through statistical analysis, namely, the short-range dominance of nucleotide correlation in most DNA sequences and the coarse-grained evolutionary dependence of the short-range correlation in coding sequences. A corresponding evolutionary mechanism is suggested. By the use of spectral analysis a large inhomogeneity in long-range base correlations for different sequences is indicated. Some results on three-dimensional DNA walks are reported. The linguistic differences between coding and noncoding sequences are also indicated. [S1063-651X(98)01107-6]

PACS number(s): 87.10.+e, 02.50.−r, 05.40.+j, 87.15.−v

I. INTRODUCTION

The nucleotide sequence data stored in GenBank have exceeded hundreds of millions of bases and they increase by ten times each five years. A great deal of information, such as the origin of life, the evolution of species, the development of individuals, and the expression and regulation of genes, exist in these sequences. On the other hand, the nucleotide sequences are generally close to random sequences. For example, the information parameter analysis shows that the first-order informational redundancies of most coding sequences are lower than 0.05 [1]. Where is the information stored then? A key point is the base correlation existing in the DNA sequences. In fact, base correlation is the basis for the grammatical construction of genetic language.

Thus, investigation into nucleotide correlation is of special importance. In recent years many authors have discussed the correlation properties of nucleotides in DNA sequences. To our knowledge, there are at least six methods that have been proposed to study the correlation property of nucleotides in DNA sequences:

(1) The method of informational parameters [1–4]: The authors defined Markovian entropy with lag or mutual information to describe the nucleotide correlation between adjacent or nonadjacent sites in the sequence.

(2) DNA walk and fractal analysis [5–11]: The sequence has been mapped onto a one-dimensional walk [7,8] or more completely, onto a two or three-dimensional (2D or 3D) walk [5,6] and a corresponding fractal analysis is given.

(3) Correlation spectrum method: Some authors characterized the symbolic sequence by decomposing it into binary sequences and quantified the base correlation, and then a spectral analysis has been done [12–15].

(4) The method of subsequence or inhomogeneity analysis [15–19]: This emphasizes the inhomogeneity of the three positions in a codon and provides an approach to finding the reading frame in DNA sequence.

(5) Preferential mode analysis [20–24]: This method examines the preferred modes and poor modes in DNA sequences of a variety of species. The preferred modes may be related to specific codes of nucleotide sequences. The method is of great importance in the linguistic analysis of hereditary information.

(6) A method of evolutionary or dynamical model [25,26]: The nucleotide correlations are investigated under an assumed evolutionary model or other dynamical model, which can account for both random mutation and natural selection in the formation of DNA sequences.

The problems that have been extensively discussed in the physical literature as follows are: What is the fundamental characteristic of the nucleotide correlation? Is there any unified picture on fractal-like organization about the long-range correlation? How do we estimate the correlation length of a sequence in the presence of fluctuations? What important differences in base correlation are there between coding and noncoding sequences? Does the correlation show any evolutionary dependence and what is the mechanism responsible for its evolution? It seems there is no generally accepted conclusion on each of the above questions. In this article we shall discuss some aspects of these problems. The organization of the article is as follows. In Sec. II we shall report the short-range dominance of base correlation in coding and many noncoding sequences and analyze the linguistic implication of this result. By the use of a spectral analysis method we shall also indicate the large inhomogeneity for the spectrum exponent from sequence to sequence. So, in contrast to

*Author to whom correspondence should be addressed. Electronic address: lfluo@nmg2.imu.edu.cn

†Permanent address: Department of Physics, Inner Mongolia University, Hohhot 010021, China.

TABLE I. The statistical distribution of D_{k+1} $(k=1,2,...)$ in coding sequences.

	Pri	Rod	Mam	Vrt	Inv	Pln	Bct	Vrl	Phg	Org
Correlation strength	9.0	7.3	7.0	9.8	6.9	4.0	3.2	5.5	2.2	2.4
Main max. at $k=1,2$(%)	89.5	93.8	92.7	81.5	63.3	74.4	66.0	84.8	40.9	47.6
Short tail (%)	23.3	33.2	28.2	24.1	14.9	24.2	35.5	50.9	45.1	42.8

short-range correlations, it seems that there is no unified picture on the long-range component of nucleotide correlation. In Sec. III we shall indicate, in a coarse-graining sense the evolutionary dependence of the short-range correlation on coding sequences and propose a formulation, maximum information principle, to describe the mechanism of sequence evolution. Since many works on DNA sequences in the literature are based on the DNA walk, in Sec. IV we shall generalize a 1D walk to a 3D walk, which is the most complete one, considering four bases equivalently in base space, and report some results on base correlation with this approach. In the last section we shall summarize the statistical peculiarities of genetic language that should be considered seriously in the attempt to establish a reasonable model of genetic language.

II. SHORT-RANGE DOMINANCE OF BASE CORRELATION

The nucleic acid sequence as a genetic language can be investigated by an information-theoretic method. We introduce informational entropy H and its related redundancy D_1, which describes the divergence from equiprobability of four bases (the vocabulary composition of genetic language):

$$H=-\sum_a p_a \log_2 p_a,$$
$$D_1=H_{\max}-H=2-H, \tag{1}$$

where p_a is the probability of base a occurring in the sequence, and the summation runs over all four bases, adenine (A), cytosine (C), guanine (G), and thymine (T). $H_{\max}$ is the maximum value of H, which is taken when the four bases occur equiprobably in the sequence.

We introduce Markovian entropy H^M (averaged conditional entropy) and its related second-order information redundancy D_2, which describes the (neighboring) base correlation in the DNA sequence.

$$H^M=-\sum_a p_a \sum_b p_{b|a} \log_2 p_{b|a},$$
$$D_2=H-H^M=2H+\sum_{a,b} p_{ab} \log_2 p_{ab}, \tag{2}$$

where p_{ab} is the joint probability of base pair ab occurring in the sequence, and $p_{b|a}=p_{ab}/p_a$ is the conditional probability.

To describe the nonadjacent correlation we generalize D_2 to

$$D_{k+2}=2H+\sum_{a,b} p_{a(k)b} \log_2 p_{a(k)b}, \quad k=1,2,3,..., \tag{3}$$

where $p_{a(k)b}$ means the joint probability of base b occurring after base a at a distance k along the sequence. $D_2, D_3, \ldots$ describe the divergence from independence of the sequence (the grammatical construction of genetic language). It is easily shown that D_{k+1} is the mutual information $I(k)$ exactly [4] since

$$I(k)=\sum_{i,j} p_{i(k-1)j} \log_2 \frac{p_{i(k-1)j}}{p_i p_j}=-2\sum_j p_j \log_2 p_j + \sum_{i,j} p_{i(k-1)j} \log_2 p_{i(k-1)j}.$$

For a random sequence with infinite length, $D_1=D_2=\cdots=0$. However, for a real sequence with finite length, the fluctuation makes them nonvanishing. To draw out the meaning of D_k calculated from the nucleotide sequence, one should subtract the effect of stochastic fluctuation. The latter can be seen as the error bars for D_k. Through expansion of logarithm in D_k and by use of Pearson's theorem one can prove that, for a random sequence of length N, $(2N \ln 2)D_1$ obeys a χ^2 distribution with 3 degrees of freedom,

TABLE II. The statistical distribution of D_{k+1} $(k=1,2,...)$ in complete sequences.

	Pri	Rod	Mam	Vrt	Inv	Pln	Bct	Vrl	Phg	Org
Correlation strength	18.9	19.8	14.8	15.8	8.7	4.5	7.2	11.8	6.2	5.9
Main max. at $k=1,2$(%)	95.5	96.4	96.3	88.6	59.2	81.7	86.9	87.7	85.9	94.0
Short tail (%)	7.5	14.2	11.2	16.2	1.6	21.5	6.1	14.7	11.0	22.0

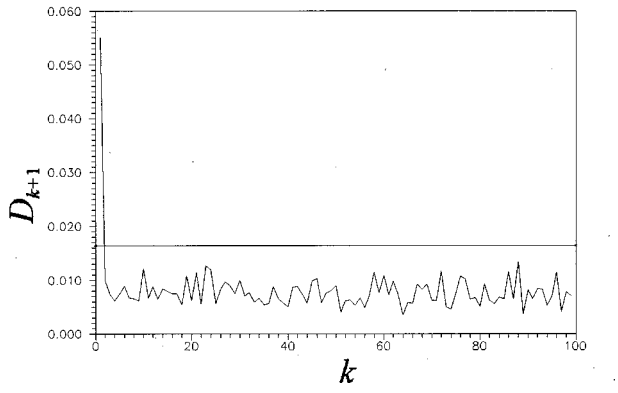

FIG. 1. D_{k+1} vs k ($k=1,2,3,...$) for a typical DNA sequence, HUMDYZ1 (length=3564), showing short-range dominance of nucleotide correlations. The straight line indicates $1.5D_{k>1}$ (f.b.), see Eq. (4). To save space, only $k<100$ are plotted. For $k>100$ D_k's are always below the bound.

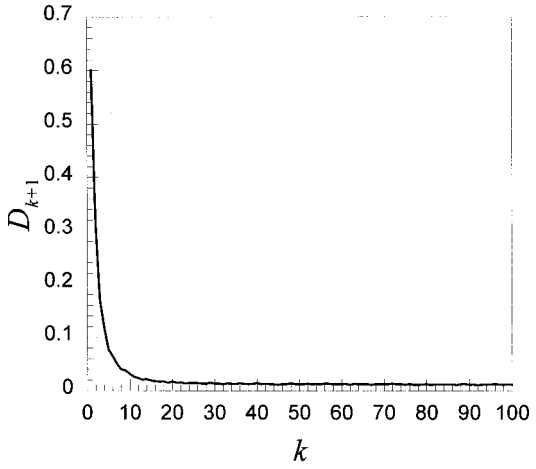

FIG. 2. D_{k+1} vs k ($k=1,2,3,...$) for English language. The English passages (about 30 000 letters) are taken from technical reference of Windows 95. It shows clearly the short-range dominance of letter correlation. The curve decreases rapidly to a value near the average of an independent sequence. To save space, only $k<100$ are plotted. For $k>100$ D_k's always fluctuate around the average value.

$(2N \ln 2)D_{k>1}$ obeys a χ^2 distribution with 9 degrees of freedom as $N\to\infty$. Thus, the fluctuation bounds (f.b.) for random sequences

$$D_1(\text{f.b.}) = 8.15/N,$$

$$D_{k>1}(\text{f.b.}) = 15.65/N \tag{4}$$

(99% confidence level). (The second equation also holds for independent sequence.) This means that for a random sequence with a length of 1000 only 1% of D_k's, namely, about 10 of them, may exceed the bound. Since most DNA sequences analyzed by us have length 1000 to 4000, to guarantee information carrying to each D_k higher than some defined bound we shall multiply a factor 1.5 to the $D_{k>1}(\text{f.b.})(15.65/N)$, which will be defined as the actual fluctuation bound.

Based on the above results we can discuss the k dependence of D_k and split off the fluctuation effect due to finite length. The results calculated from 3709 coding sequences and 2883 complete sequences (each complete sequence including coding regions and introns, 5′-caps and 3′-tails) from 1991 GenBank are shown in Tables I and II. The first lines in both tables refer to the category average of correlation strength of main maximum in D_{k+1} [in unit of $D_{k+1}(\text{f.b.})$]. The second lines give the percentage of sequences with main maximum located on the nearest-neighboring ($k=1$) and next-to-nearest neighboring ($k=2$) sites. We find that for most sequences (50% to more than 90%) the main maxima are located on $k=1$ and 2 neighbors. So, the nucleotide correlations in the nearest-neighboring and next-to-nearest neighboring sites are dominant ones [1]. The third lines give the percentage of sequences with a short-tail base correlation. The short tail means $D_{k+1} < 1.5D_{k+1}(\text{f.b.})$ as $k\geqslant 3$. From the tables we see that the sequences with short-tail correlations amount to 25–50% for coding regions but only 5–20% for complete sequences. Figure 1 gives a typical example of D_{k+1} versus k for sequence HUMDYZ1 (GenBank sequence identification codes). The above results prove the short-range dominance of base correlations in coding sequences. For complete sequences the short-range dominance of correlations exists, too, but most parts of them have long tails.

Since the nucleic acid sequence, as a whole, plays its role in biological activity, it is generally anticipated that the sequence will show long-range correlations. However, we have found short-range dominance of base correlations in coding and many complete sequences, which is unexpected. Part of the above results was obtained by us in Ref. [3]. The universal existence of strong nucleotide correlation in adjacent sites of DNA sequence is an important characteristic of genetic language. The feature is more notable for coding sequences since the correlation with a short tail occurs more frequently in these regions. Therefore, in spite of many long-range in-

TABLE III. The statistical distribution of $F_{T(k)A}$ in coding sequences.

	Pri	Rod	Mam	Vrt	Inv	Pln	Bct	Vrl	Phg	Org
Correlation strength	17.5	6.6	7.0	7.8	6.9	5.9	2.9	4.8	2.2	1.8
Main max. at $k=1,2$(%)	82.0	87.4	81.8	86.1	79.4	85.1	41.8	50.0	29.9	11.9
Short tail (%)	13.7	18.4	9.1	20.4	10.9	13.2	21.4	26.5	30.6	34.5

TABLE IV. The statistical distribution of $F_{T(k)A}$ in complete sequences.

	Pri	Rod	Mam	Vrt	Inv	Pln	Bct	Vrl	Phg	Org
Correlation strength	12.3	10.9	12.0	11.8	9.6	7.3	6.4	8.3	5.8	3.6
Main max. at $k=1,2$(%)	97.0	97.5	96.3	94.3	84.5	92.9	78.1	71.9	67.2	39.0
Short tail (%)	1.0	2.3	2.8	1.0	2.8	5.1	2.3	2.1	3.2	8.0

teractions between nucleotides, there exists a definite simplicity against the complex background—the strong short-range correlation of adjacent bases.

How can we understand the short-range dominance of base correlation in DNA sequences? It is related to some frequently occurring words with 2–3 bases in genetic language. We have found these words through preferential-mode analysis [24]. To understand this point we compare genetic language with the English language. In English there also exist many frequently used words and syllables. We can also calculate D_k versus k and investigate the correlation of letters. Figure 2 gives an example of D_k versus k in English, which shows the short-range dominance clearly.

The informational redundancy D_{k+1} can be rewritten as

$$D_{k+1} \approx \frac{1}{\ln 2} \sum_{a,b} \frac{F_{a(k)b}}{p_a p_b}, \tag{5}$$

where

$$F_{a(k)b} = (p_{a(k-1)b} - p_a p_b)^2 \tag{6}$$

is defined to describe the particular base correlation ($k = 1,2,3,\ldots, p_{a(0)b} = p_{ab}$). It can be proved that, for an independent sequence,

$$\frac{NF_{a(k)b}}{p_a(1-p_a)p_b(1-p_b)}$$

obeys a χ^2 distribution with 1 degree of freedom. So, one has

$$F_{a(k)b}(\text{f.b.}) = \frac{6.63}{N} p_a(1-p_a)p_b(1-p_b) \tag{7}$$

(99% confidence level).

There are 16 kinds of base correlations included in D_k. We have studied each base correlation for 3709 coding sequences and 2883 complete sequences and found that the modes CG and TA are the strongest for most sequences. As an example, Tables III and IV give the statistical distribution of $F_{T(k)A}$ versus k. We find that for these main modes of base correlation the short-range dominance still holds, but the percentage of sequences with short tails is largely lowered as compared with D_k. The reason in part lies in that the overall factor 1.5 has been used in the fluctuation bound. Actually, due to different degrees of freedom for $D_{k>1}$ and $F_{a(k)b}$ a factor larger than 2 should be multiplied in the $F_{a(k)b}$ case.

From previous discussion we know that the difference in correlation D_k between coding and noncoding sequences does exist. The former has more short-tail correlations than the latter. But both the coding and complete sequences show strong correlations between adjacent bases. A large part of the long tails of base correlation in complete sequences comes from noncoding regions, but it seems not mainly from introns. The statistical data on base correlation in 924 intron sequences are given in Table V.

Spectral Analysis and long-range correlation. Spectral analysis of a nucleotide sequence after it has been changed into a numeric sequence, or more directly, of a correlation function D_k may give more information on nucleotide correlation. An important feature is the many peaks in the spectrum. Especially for coding sequence there is a universal resonance peak at $k/N = 1/3$ (N = sequence length). Evidently, the phenomenon is related to codon triplets in the sequence. Reference [15] gave a thorough analysis of its origin and demonstrated that the 1/3 peak occurs in the spectrum if and only if the base composition is not uniformly distributed in three codon positions. The theorem can be generalized to explain other kinds of peaks in the spectrum. So, a spectrum line means the inhomogeneous distribution of base composition in a given range. It is reasonable to infer that the abundant spectrum lines existed in the high-frequency range are related to short-range dominance of base correlations.

TABLE V. The statistical distribution of correlations in introns.

	D_k				$F_{C(k)G}$				$F_{T(k)A}$			
	Pri	Rod	Mam	Vrt	Pri	Rod	Mam	Vrt	Pri	Rod	Mam	Vrt
Correlation strength	6.0	6.0	4.8	3.4	11.7	9.5	8.0	5.1	3.0	2.3	2.6	2.2
Main max. at $k=1,2$ (%)	98.2	95.1	95.2	88.4	96.3	97.4	93.5	82.7	56.1	43.8	59.7	42.1
Short tail (%)	66.3	58.4	59.7	78.5	36.5	39.2	29.0	52.1	38.3	44.1	30.6	56.2

We have pointed out the short-range dominance of base correlation and its universal property. But what about the long-range component of the correlation? Is there any peculiarity about long-range correlation in DNA sequences? The problem can be approached through investigating the low frequency behavior of the power spectrum of sequences. Applying a FFT (fast Fourier analysis) method to each sequence, many long-period components in the spectrum of base correlation can be found. We calculate the percentage of sequences that have marked components of correlation with a long period larger than 100 in the spectrum of D_k and the average spectrum of $\langle F_{a(k)b}\rangle_{ab}$. The data include 3720 coding sequences and 2895 complete sequences. The result shows that only 1% (or less) of coding sequences and 5% (or less) of complete sequences have such long-period correlations.

We define the power spectrum as

$$P_\nu = \left| \sum_{k=1}^{N} D_{k+1} \exp \frac{2\pi i k \nu}{N} \right|^2 . \tag{8}$$

We then investigate the relation between power spectrum P_ν and frequency ν (or inverse period) for each sequence and check if the power spectrum can be put in the form

$$P_\nu = c\,\nu^{-\alpha} \tag{9}$$

for low ν. By linear regression of $\ln P_\nu$ versus $\ln \nu$ ($\ln P_\nu = -\alpha \ln \nu + \beta$) we find for sequences with long-period correlations (period$>$100) that the low frequency behavior can be classified into three categories: (1) For 60% or more of the sequences the linear relation holds (and $\alpha \neq 0$) in the low frequency region with deviation $\sigma < 1$ ($\sigma = [\Sigma_1^n \{\ln P_\nu - (-\alpha \ln \nu + \beta)\}^2]^{1/2}/n$) [see Fig. 3(a)]; (2) The linear relation holds only approximately with $1 < \sigma < 2$ [see Fig. 3(b)]; (3) No linear relation but strong oscillation can be found between $\ln P_\nu$ and $\ln \nu$ [see Fig. 3(c)]. The average spectrum exponents α for various species are shown in Table VI. For most (about 60%) sequences with long-period correlations they take values between -1 and -2. However, it is demonstrated that if the sequences are stochastically chosen then these exponents take values near -0.5. Many works proposed the power spectrum of $\nu^{-\alpha}$ type but their results are scattered due to different samplings and statistical methods [6,8,12]. They all neglected the difference between sequences with and without long-period correlations. We emphasize the large inhomogeneity in spectrum exponents from sequence to sequence. Only for a part of the sequences that have long-period components in base correlation the low frequency behavior of the power spectrum takes the form of $1/\nu^\alpha$ with $\alpha = 1-2$.

III. EVOLUTIONARY DEPENDENCE OF SHORT-RANGE CORRELATION IN CODING SEQUENCES

Dobzhansky said that everything in biology will be nonsense if it is not viewed from evolution. In this section we will study the statistical correlation between nucleotides from the point of evolution. We shall demonstrate the adjacent base correlation increasing with evolution. This is the very reason why base correlation is important in biology.

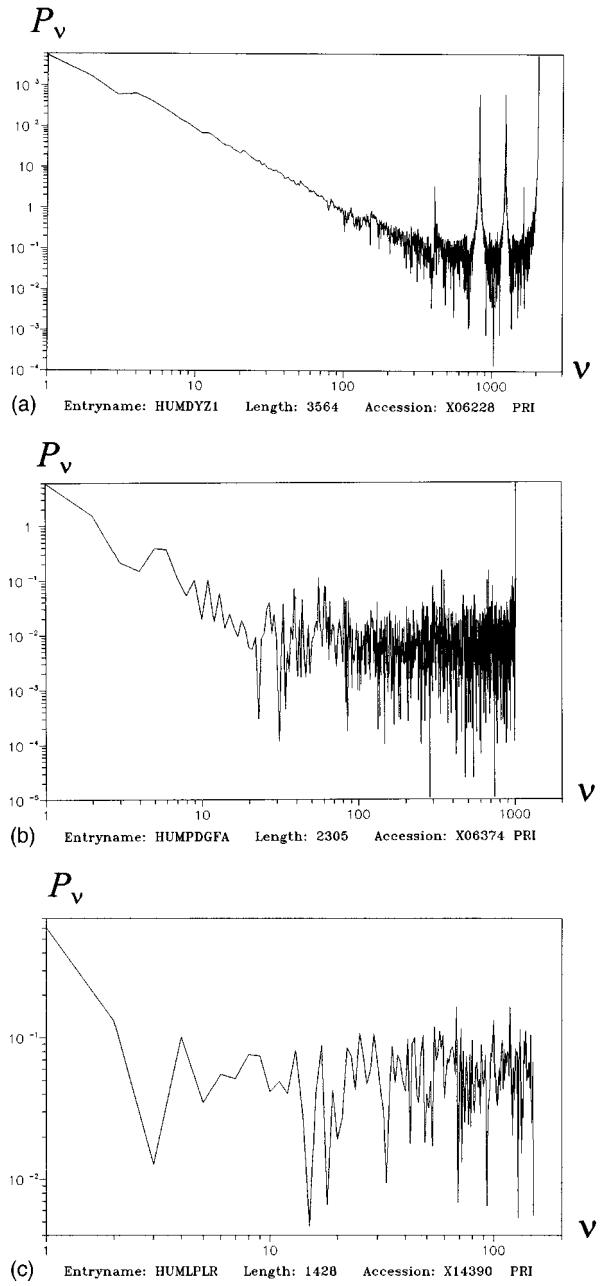

FIG. 3. Log-log diagram for power spectrum $P(\nu)$ vs frequency ν. (a) for sequence HUMDYZ1 (length$=$3564), showing a good linear relation; (b) for sequence HUMPDGFA (length $=$2305), showing an approximate linear relation; and (c) for sequence HUMLPLR (length$=$1428), showing no linear relation existed.

Since adjacent base correlation is the main part of correlations between nucleotides, the correlation property in coding sequences can be described through D_2 approximately. D_2 describes the grammatical construction of genetic lan-

TABLE VI. The averaged spectrum exponent α in sequences with long-periodicity-correlation for given deviation σ.

	Pri	Rod	Mam	Vrt	Inv	Pln	Bct	Vrl
$0<\sigma<1$	1.29	1.16	1.51	1.25	1.24	1.32	1.40	1.10
$1<\sigma<2$	0.93	1.60	1.27	1.63	1.21	1.31	1.16	0.85

TABLE VII. The informational parameters of coding sequences.

	Pri	Rod	Mam	Vrt	Inv	Pln	Bct	Vrl	Phg	Org
$\langle D_1 \rangle$	0.029	0.016	0.031	0.024	0.047	0.033	0.037	0.036	0.040	0.067
$\langle D_2 \rangle$	0.059	0.059	0.055	0.056	0.037	0.030	0.032	0.037	0.021	0.022
$\langle X \rangle$	0.71	0.80	0.68	0.73	0.54	0.52	0.56	0.56	0.45	0.36
$\langle F \rangle$	0.54	0.53	0.56	0.51	0.48	0.46	0.52	0.46	0.43	0.40

guage, while D_1 describes its vocabulary composition. And one can define $X=D_2/(D_1+D_2)$, which describes both aspects. On the other hand, the $G+C$ content (denoted by F) is also introduced to describe the variation of strong-bond components of DNA sequence in the course of evolution. We calculate the informational parameters D_1, D_2, X, and F for each sequence, average them in each category of species, and then study their evolutionary dependence. The first study was done based on 1985 GenBank data with some complement where 732 coding sequences and 0.8 million base pairs have been statistically analyzed [1]. Next, we used 1987 GenBank data and studied 1944 coding sequences with 2.5 million base pairs. Recently, we utilized 1991 GenBank data and their catalog on 8226 coding sequences with 15 million base pairs. The results of three statistical investigations are consistent with each other. The results recently obtained are listed in Table VII. From Table VII we see that $\langle D_1 \rangle$ is a small quantity, ~ 0.03, which decreases roughly with evolution but the trend is not clear; $\langle D_2 \rangle$ increases gradually from lower organisms to higher species, which shows the base correlation strengthened in the course of evolution. The main trends of $\langle X \rangle$ and $\langle F \rangle$ changing with evolution are the same as $\langle D_2 \rangle$. Of course, the formation and bifurcation of species in evolution is a complicated process. The informational parameters change with time. However, as soon as a new species has formed, its phenotype and gene structure begin to be stabilized (the punctuated evolution as stated by some biologists). So, the result obtained by use of the present data can reflect the evolutionary history. On the other hand, ''evolution is a mender'' as said by Monod. The complexity of evolution and the individual difference in biology make many exceptions to any general law. The correlation of informational parameters with evolution could only be manifested through a coarse-grained average. The experience in our statistical investigation shows that the coarse-grained average is not an expedient measure but an appropriate form to express the biological complexity.

In order to show the difference in three positions of a codon we break the data of coding regions into three constituent subsequences. We define informational parameters of subsequences. For example,

$$D_1^{(n)} = 2 + \sum_a p_a^{(n)} \log_2 p_a^{(n)}. \tag{10}$$

$$D_2^{(n)} = -\sum_a p_a^{(n)} \log_2 p_a^{(n)} - \sum_a p_a^{(n-1)} \log_2 p_a^{(n-1)} + \sum_{ab} p_{ab}^{(n-1,n)} \log_2 p_{ab}^{(n-1,n)} \tag{11}$$

($n=1$, 2, and 3; the case of $n-1=0$ means the third subsequence), in which $p_a^{(n)}$ denotes the probability of base a in the nth subsequence, etc. In 1990 we made the subsequence analysis on 1024 coding sequences with length 1200 base pairs [17]. Recently, a similar analysis was made for enlarged data—8226 coding sequences with length >1000. Both analyses give the same result. The results are tabulated in Table VIII. Combining Tables VII and VIII we find that $\langle D_2 \rangle$ and $\langle D_2^{(1)} \rangle$ for the four higher species—Pri (primate), Rod (rodent), Mam (other mammalian), and Vrt (other vertebrate)—are higher than Inv (invertebrate) and Vrl (viral), the parameters of Inv and Vrl are higher than Pln (plant) and Bct (bacterial), and the latter are, in turn, higher than Phg (phage) and Org (organelle). $\langle D_2^{(1)} \rangle$ changes in a wider range than $\langle D_2 \rangle$. However, there is no evolutionary correla-

TABLE VIII. The informational parameters of subsequences.

	n	$\langle D_1^{(n)} \rangle$	$\langle D_2^{(n)} \rangle$	$\langle X^{(n)} \rangle$	$\langle F^{(n)} \rangle$
Pri	1	0.066	0.098	0.62	0.57
	2	0.051	0.065	0.59	0.44
	3	0.135	0.083	0.49	0.62
Rod	1	0.055	0.102	0.67	0.55
	2	0.049	0.060	0.57	0.42
	3	0.091	0.087	0.57	0.61
Mam	1	0.071	0.086	0.60	0.57
	2	0.060	0.063	0.54	0.42
	3	0.20	0.076	0.42	0.68
Vrt	1	0.074	0.091	0.58	0.55
	2	0.064	0.073	0.55	0.42
	3	0.110	0.087	0.60	0.57
Inv	1	0.089	0.050	0.39	0.53
	2	0.084	0.089	0.54	0.41
	3	0.186	0.091	0.40	0.51
Pln	1	0.080	0.038	0.34	0.51
	2	0.070	0.077	0.55	0.40
	3	0.125	0.076	0.44	0.48
Bct	1	0.097	0.035	0.28	0.58
	2	0.052	0.067	0.59	0.41
	3	0.164	0.094	0.48	0.56
Vrl	1	0.064	0.054	0.47	0.50
	2	0.040	0.057	0.60	0.43
	3	0.092	0.055	0.52	0.46
Phg	1	0.075	0.030	0.31	0.52
	2	0.064	0.068	0.55	0.39
	3	0.140	0.055	0.35	0.39
Org	1	0.078	0.032	0.37	0.48
	2	0.077	0.082	0.60	0.40
	3	0.199	0.073	0.30	0.31

TABLE IX. The informational D_1 of noncoding sequences as compared with coding regions.

	Imunoglob	Mammal	Eukaryot	E. coli	Prokar	Virus
5′-cap	0.037	0.033	0.111	0.036	0.094	0.089
3′-tail	0.034	0.058	0.106	0.023	0.056	0.058
Intron	0.061	0.064	0.124			
Coding	0.027	0.026	0.037	0.017	0.045	0.026

tion for $\langle D_2^{(2)}\rangle$ and $\langle D_2^{(3)}\rangle$. This shows that the neighboring base correlation between two codons has the strongest evolutionary dependence. For parameter $\langle X\rangle$, Pri, Rod, Mam, Vrt take higher values than Inv, Pln, Bct, Vrl, and the latter higher than Phg and Org. The same is true for $\langle X^{(1)}\rangle$. Next, we find that these parameters for viruses take larger values which are near their hosts—invertebrates. This supports the view that viruses could not be primitive life but retrograde type in evolution. Organelles include mitochondria and chloroplasts. They occupy a very low level, which supports the symbiosis theory about the origin of these organelles. Another interesting result obtained by subsequence analysis is $D_1^{(3)}$ much larger than $D_1^{(1)}$ and $D_1^{(2)}$, and varying from one species to another [17].

The base composition and base correlation are inhomogeneous for different segments of DNA sequence. The above discussion is done for protein-coding regions. We have studied the noncoding sequences too and found their informational parameters largely different from the coding region. An important law is the following: D_1 in all noncoding regions are larger than the corresponding coding region (see Table IX, which is taken from Ref. [1]). The reason may be due to the regulating and controlling signals existing in 5′-caps, 3′-tails, introns, and other noncoding regions, which decreases their informational entropies [1]. The result is also consistent with linguistic analysis in Ref. [27].

The evolutionary dependence of information parameters exists not only in the statistical average of a large amount of unrelated genes, but also in some particular gene or protein if the latter has enough sequence data and occurs in a wide range of species. We have studied 37 MHC (myosin heavy chain) genes and the results for coding and complete sequences are shown in Table X. The results are consistent with Ref. [28], which uses the DNA walk method but the evolutionary dependence is more obvious in our information-theoretic approach.

TABLE X. The informational parameters of MHC genes.

		$\langle D_1\rangle$	$\langle D_2\rangle$	$\langle X\rangle$	$\langle F\rangle$
Pri	coding	0.051	0.086	0.64	0.55
	complete	0.028	0.089	0.78	0.53
Rod	coding	0.048	0.077	0.65	0.56
	compete	0.031	0.067	0.75	0.54
Mam	coding	0.065	0.088	0.58	0.58
	complete	0.061	0.085	0.59	0.57
Vrt	coding	0.034	0.081	0.73	0.48
	complete	0.029	0.069	0.74	0.46
Inv	coding	0.049	0.053	0.58	0.49
	complete	0.047	0.036	0.50	0.44
Pln	coding	0.076	0.042	0.34	0.36
	complete	0.080	0.040	0.32	0.36

The most direct and obvious representation of the evolutionary relationship is preferential mode analysis done recently by us [24]. The nucleic acid sequence is reduced to a two-letter sequence written as S ($=C$ or G) and W ($=A$ or T) or by R ($=A$ or G) and Y ($=C$ or T). The preferential modes of dinucleotides and trinucleotides are emphasized since the nearest-neighboring and next-to-nearest neighboring correlation of bases are the most important ones. We define the ordered fragment $(Y_l,\ldots,Y_k)_n$, $Y_i\in(S,W)$ or (R,Y) and n—repetition times of $(Y_l,\ldots,Y_k)$. From sequence data the frequency of mode $m=(Y_l,\ldots,Y_k)$, denoted as N_m, is calculated first. We then define and calculate the relative mode content (RMC)

$$W_m=\frac{N_m-\langle N_m\rangle}{\sigma_m}. \tag{12}$$

$\langle N_m\rangle$ is the expectation value of mode frequency in the independent sequence and σ_m its deviation. Likewise, one can calculate the relative ordered-fragment content (ROFC) through the numeration of the ordered fragment $(Y_l,\ldots,Y_k)_n$ $(n=2,3,\ldots)$

$$W_{mn}=\frac{N_{mn}-\langle N_{mn}\rangle}{\sigma_{mn}}. \tag{13}$$

$W_m(W_{mn})>1$ means preferred mode, $W_m(W_{mn})<1$ means poor mode. Further, to describe the deviation of sequence from independent sequence one may sum up the modes for given k and reading frame i $(i=1,2,\ldots,k)$. Define deviation

$$U_{ki}=2^{-k}\sum_m (W_m)^2. \tag{14}$$

It can be shown that U_{ki} is related to informational redundancy D_2 (as $k=2$), D_3 (as $k=3$), etc. but with a given reading frame.

We have studied 6.8×10^6 bp sequence data and found all preferred (and poor) modes of $k=2$ and 3. Several examples for exons are shown in Table XI. The RMC and ROFC (with number in brackets showing the repetition times of mode) are tabulated. The data indicate clearly the evolutionary dependence of these modes. The deviations $U_{3i}(S\text{-}W)$, $U_{2i}(S\text{-}W)$, and $U_{2i}(R\text{-}Y)$ for exons and $U_{2i}(R\text{-}Y)$, $U_{3i}(R\text{-}Y)$ for introns, 5′-caps, and 3′-tails are correlated with evolution, too, which can be found in Ref. [24].

TABLE XI. Relative ordered-fragment content W_m and W_{mm} of preferred modes.

	Pri	Rod	Mam	Vrt	Inv	Pln	Bct	Vrl	Phg	Org
Exon										
SWS	4.41	4.69	5.57	4.20	1.96	0.82	2.28	1.67	0.47	−1.86
SWS(2)	2.23	2.54	2.65	2.36	1.33	0.64	1.19	0.79	0.61	−0.48
SWS(3)	2.32	2.11	3.20	2.23	1.42	0.64	1.10	0.96	0.58	−0.12
SWS(4)	2.39	1.94	2.52	1.92	1.50	0.76	0.61	0.70	0.09	0.19
WWS	1.85	2.20	3.02	2.73	1.52	1.85	0.81	0.18	−0.12	−1.42
WWS(2)	1.29	1.47	2.02	1.82	0.89	1.18	0.39	0.33	−0.12	−0.47
WWS(3)	1.10	1.28	1.85	1.71	1.07	1.08	0.38	0.39	−0.05	−0.28

Through the above analyses we have found many preferred modes in S-W language for exons and many preferred modes in R-Y language for introns, 5′-caps and 3′-tails [24]. It seems that exons preferentially use S-W language and these noncoding regions preferentially use R-Y language. The different preference of genetic languages in coding and noncoding sequences can be explained by their functional difference. The S-W representation is more convenient for the coding region in its translation and proofreading since two strands of DNA are precisely the same in S-W language. The R-Y reprsentation is more useful for noncoding regions in the regulation and control of gene expression since the sites of large local deviation in DNA (which are recognized by repressors and enzymes) are presumably determined by an R-Y rather than an S-W sequence.

The evolutionary dependence of short-range correlation in coding sequences can be shown not only in informational parameters D_2, etc. and preferential mode analysis, but also in other statistical properties of the sequence. For example, the eigenvalue and limit-approaching length of the probability matrix [17], the fractal dimension of the 2D DNA walk [5], etc. On the other hand, since the short-range correlations of nucleotides are correlated with evolution, we can reconstruct an evolutionary relation with this knowledge. Based on base probabilities p_a and conditional probabilities $p_{b|a}$ and $p_{b|a*}$ ($*$ means any base) we have succeeded in deducing the evolutionary tree by classifying 16S rRNA sequence data [29].

Evolutionary mechanism. Nucleic acid sequences evolve under two factors, namely, random mutations (including insertions, deletions, and recombinations), which cause the entropy H to increase, and natural selection, which cause the Markovian entropy to decrease (the increase of base correlation). The latter can be viewed as some constraints imposed on the random drift of bases. By using the maximum entropy principle (MIP)—a general principle for the nonequilibrium system suggested by Haken [33]—we can express the evolutionary mechanism of joint action of random mutation and natural selection successfully [25]. Following MIP, the entropy H is maximized under constraints

$$N=\Sigma p_a=1$$

and fixed H^M

$$H^M=\sum_a p_a \log_2 p_a-\sum_{ab} p_{ab} \log_2 p_{ab}$$

$$=-\sum_{ab} p_b p_{a|b} \log_2 p_{a|b} \tag{15}$$

and H^{M*}

$$H^{M*}=\sum_a p_a \log_2 p_a-\sum_{ab} p_{a(1)b} \log_2 p_{a(1)b}$$

$$=-\sum_{ab} p_b p_{a|b*} \log_2 p_{a|b*} \tag{16}$$

or $C+G$ content C

$$C=p_C+p_G, \tag{17}$$

namely,

$$\delta H-\lambda_0\delta N-\lambda_1\delta H^M-\lambda_2\delta H^{M*}=0 \tag{18}$$

or

$$\delta H-\lambda_0\delta N-\lambda_1\delta H^M-\lambda_2\delta C=0. \tag{19}$$

From Eqs. (18) and (19) the probabilities p_a for each sequence can be found. Through calculation of p_a for 1469 protein-coding DNA sequences of various species we find the deviations of the theoretical probabilities from the experimental values are generally lower than 10%. Moreover, the Lagrangian multipliers averaged over each category are correlated with evolution. The agreement of MIP analysis with observational data is satisfactory. So the above assumed evolutionary mechanism of nucleotide sequences is reasonable and their formation can be explained under the MIP principle basically in spite of the complexity inherent in the evolutionary process.

IV. DNA WALK

So far we have discussed the main results obtained by us on base correlations in DNA sequences—the finding of short-range dominance of the correlation and its evolutionary dependence. In these discussions different methods have been synthetically used but the main approach is based on information-theoretic consideration. Recently, 1D DNA walk as a method to investigate nucleotide correlation appeared widely in physical literatures [7,8,11,28]. However, following Silverman and Linsker, each symbol in s-symbol sequence can be represented by a vertex in the $(s-1)$ simplex.

TABLE XII. The averaged fractal dimensions in 3D, 2D, and 1D DNA walk (primate).

		Code	Intron	Cap	Tail	Complete			Code	Intron	Cap	Tail	Complete
	D_a	1.69	1.56	1.47	1.53	1.56		D_a	1.80	1.60	1.53	1.65	1.67
3D	D_b	1.38	1.40	1.26	1.33	1.33	1D	D_b	1.59	1.57	1.48	1.49	1.67
	D_f	1.38	1.38	1.31	1.37	1.38	$(R\text{-}Y)$	D_f	1.50	1.52	1.48	1.50	1.58
	D_a	1.65	1.51	1.41	1.47	1.50		D_a	1.51	1.48	1.41	1.41	1.40
2D	D_b	1.35	1.39	1.22	1.30	1.37	1D	D_b	1.27	1.28	1.14	1.26	1.22
	D_f	1.36	1.38	1.26	1.34	1.36	$(S\text{-}W)$	D_f	1.29	1.30	1.17	1.30	1.30

So, for the nucleotide sequence the mapping space is three dimensional [6,10,30]. The 3D DNA walk is the most complete one that considers four bases equivalently in base space. In the following we shall generalize DNA walk into three-dimensional space. By using S-L mapping, put A, C, G, and T on the four vertices of a tetrahedron. We use the following representation of nucleotides in which A corresponds to a walk in the direction $(\mathbf{i}+\mathbf{j}+\mathbf{k})$, C corresponds to $(-\mathbf{i}+\mathbf{j}-\mathbf{k})$, G corresponds to $(\mathbf{i}-\mathbf{j}-\mathbf{k})$, and T corresponds to $(-\mathbf{i}-\mathbf{j}+\mathbf{k})$, respectively. So the projection of DNA walk on the X axis is purine along $+X$ and pyrimidine along the $-X$ direction; the projection of the DNA walk on the Z axis is A and T (weak bond) along $+Z$ and C and G (strong bond) along the $-Z$ direction; the projection of DNA walk on the Y axis is A and C along $+Y$, and G and T along the $-Y$ direction. The mean square separation of the end points in a sequence (length N) containing n bases is denoted by $\langle R_n^2\rangle_N$. The local fractal dimension (FD) is defined by [5]

$$d_N(n)=\ln\frac{n+1}{n}\Big/\ln\{\langle R_{n+1}^2\rangle_N/\ln\langle R_n^2\rangle_N\}^{1/2}. \tag{20}$$

We found that $d_N(n)$ changes smoothly only for $n\leqslant N/2$ as in polymer chains in configuration space [31]. So an averaged FD is defined, namely,

$$D_f=\langle d_N(n)\rangle_{N/2}. \tag{21}$$

Here the average is taken over different $n(n\leqslant N/2)$. Moreover, from the log-log diagram of end-point separation versus base pair number in 3D DNA walk we found a good linearity remaining between $\ln\langle R_n^2\rangle_N^{1/2}$ and $\ln n$ in a range of 20 bases and a comparatively good linearity in 400 bases. So, we can define two other quantities,

$$D_a=\langle d_N(n)\rangle_a \tag{22}$$

and

$$D_b=\langle d_N(n)\rangle_b, \tag{23}$$

where $\langle\cdots\rangle_a$ means the average over first 21 bases (n from 1 to 21, or $\ln n$ from 0 to 3) and $\langle\cdots\rangle_b$ means the average from 22nd to 403rd bases ($\ln n$ from 3 to 6). The above analyses show that the FD can be defined rigorously only in regions of 20 bases but it can be generalized to about 400 bases or half the length of the sequence approximately [5,9].

In 3D walk it is easily shown that the square separation of end points R_n^2 obeys

$$3R_n^2=4(n_A^2+n_G^2+n_c^2+n_T^2)-n^2, \tag{24}$$

$$n=n_A+n_G+n_c+n_T.$$

So, for $n_A\sim n_G\sim n_c\sim n_T$, we have

$$3R_n^2=4n^2\sum_a\,(p_a^2-\tfrac{1}{4})\approx n^2\,\ln 2\;D_1(n),$$

$$(p_a=n_a/n,\;\; a=A,G,C,T),$$

$$\langle R_n^2\rangle\approx\frac{\ln 2}{3}\,n^2\langle D_1(n)\rangle, \tag{25}$$

where $D_1(n)$ means the first-order information redundancy for a window with length n. The average in Eq. (25) is taken over n-bases fragments along the sequence. Since the dependence of $\langle D_1(n)\rangle$ on n is not in a form of power function for a large variation of n, the linear relation between $\ln\langle R_n^2\rangle_N$ and $\ln n$ holds only in region a, i.e., for n small as stated before. Set $\langle D_1(n)\rangle\sim n^{-\beta}$ for small n. For a random sequence, $\beta=1$. If a sequence has a strong bias of bases, then the change of $D_1(n)$ with n will be weaker than that in random sequence. This leads to $\beta<1$ and a relatively small D_a [$D_a\sim 2/(2-\beta)$]. On the other hand, when $n>N/2$, the window with n bases can be shifted along the sequence only for a few times and one has a large fluctuation about the average $\langle R_n^2\rangle$. This makes the fractal description through end-point separation impossible. Equations (20)–(25) give a relation between fractal description and information-theoretic description.

We calculate fractal dimensions D_a, D_b, and D_f in 3D DNA walk, 1D purine-pyrimidine walk, 1D strong-weak bond walk and 2D walk for each sequence. The 2D DNA walk is defined by map of the sequence onto 2D plane in which A, C, T and G correspond to $+X$, $-X$, $+Y$, and $-Y$, respectively. More than 1000 sequences are calculated and a part of results (for primate only) are listed in Table XII. (The details of the results can be found in Ref. [34].)

From Table XII we find the following.

(1) $D_a>D_b\sim D_f$. Since D_a is the rigorous fractal dimension defined in a range of 20 bases, the result shows the long-range (larger than 20 bases) correlation lowering the average FD. The point can also be seen from a plot of $\ln\langle R_n^2\rangle_N$ versus $\ln n$ in which the curve goes slightly up from a straight line for n larger than 20.

(2) The FD in the 3D walk is slightly larger than that in 2D and they are both smaller than the 1D purine-pyrimidine $(R\text{-}Y)$ walk and larger than the 1D weak-strong bond $(S\text{-}W)$ walk. The 3D walk includes three kinds of 1D walks as its projections. The purine-pyrimidine $(R\text{-}Y)$ classification of nucleotides is related to the hydrophobic-hydrophilic charac-

teristics of encoded amino acids. The classification of weak and strong hydrogen bonds is also important for the biological function of nucleic acids. These two walks are more important than the third one. The 2D walk defined above has its projections in the direction of $\mathbf{i}+\mathbf{j}$ and $\mathbf{i}-\mathbf{j}$ corresponding to S-W walk and R-Y walk, respectively. So the fractal dimension in the 2D walk is approximately equal to that in 3D. However, the two 1D walks—R-Y type and S-W type—take much different values of fractal dimensions. The former is larger than the latter and the fractal dimensions of 3D and 2D walks lie in the middle of them. The 1D S-W walk taking a smaller FD than the R-Y walk can be explained by calculation of D_1 in reduced languages. We have proved that D_1 in S-W language is larger than D_1 in R-Y language.

(3) D_f of coding sequences takes value between 1.1 and 1.4 for 2D and 3D walks and between 1.2 and 1.5 for 1D R-Y walks. The deviation of these dimensions from 2 shows the existence of base correlation even in coding region. The result is different from Ref. [7], where $\alpha=0.50$ is reported for some cDNA sequences which corresponds to our $D_f=2$.

(4) D_a of coding sequences is found to be larger than D_a of corresponding introns, 5′-caps and 3′-tails. This implies possibly that some regulation signals (biased in base composition) with length smaller than 20 bases occurred in these noncoding regions [1,27].

The above results are consistent with that obtained through the information-theoretic method. However, the fractal dimension and the end-point separation in the DNA walk are determined by first-order informational redundancy $D_1(n)$ versus n. Its relation to base correlation is not clear. So, the property of short-range dominance of base correlation which has been found in the information-theoretic method could not be demonstrated obviously in this approach.

V. CONCLUSION: THE STATISTICAL CHARACTERISTICS OF GENETIC LANGUAGE

The formal language theory has been developing conspicuously since the classic work of Chomsky [35]. However, it is not clear if the genetic language can be put in the framework of Chomsky hierarchy. To establish a reasonable mathematical model on genetic language one should first study the statistical characteristics of the language, especially from the point of evolution. The following points should be noticed seriously [32].

(1) Strong noise background and small informational redundancy of DNA sequences due to the pressure of neutral or nearly neutral mutations of bases in the evolution.

(2) Short-range dominance of base correlations and its evolutionary relationship. The short-range dominance is nearly universal for all sequences. The evolutionary dependence exists only for coding sequences and in the coarse-grain average. These points have been indicated and discussed thoroughly in Secs. II and III in this article.

(3) Division of the language into two kinds—coding region and noncoding region—and the existence of the reading frame (inhomogeneity of base composition) in coding sequences.

(4) Different vocabularies in coding and noncoding sequences. The coding sequences preferentially use S-W language and many noncoding sequences preferentially use R-Y language. The particular vocabularies in some noncoding region lead to its relatively large D_1 (see Sec. III).

(5) Inhomogeneity of long-range correlation property. Several percents of sequences in single-copied genes show the long-period of base correlation which obey $v^{-\alpha}$ law with $\alpha=1-2$. The law seems related to fragment repetitions [6] (see Sec. II).

(6) $G+C$ content. In addition to adjacent base correlation the most strong evolutionary relationship in base composition is $G+C$ content, which provides a constraint to random mutation (see Table VI).

(7) Maximum information principle that describes the evolutionary mechanism basically. The evolution of nucleotide sequences is dominated by two major factors—random mutation that maximizes the entropy and the natural selection which strengthens the statistical correlation between bases, especially the correlation between neighboring bases. The MIP, as a formulation of molecular evolution, can summarize the above mechanism and it achieves success in the study of coding sequences (see Sec. III).

ACKNOWLEDGMENTS

The work (Project No. 39670188) was supported by the National Science Foundation of China. The authors are also grateful to Drs. Hong Li, Guoyi Bai, and Qianzhong Li for their helpful discussions.

[1] L. F. Luo, L. Tsai, and Y. M. Zhou, J. Theor. Biol. **130**, 351 (1988).
[2] L. Gatlin, *Information Theory and Living System* (Columbia University Press, New York, 1972).
[3] L. F. Luo and H. Li, Bull. Math. Biol. **53**, 345 (1991).
[4] H. Herzel and I. Grosse, Physica A **216**, 518 (1995).
[5] L. F. Luo and L. Tsai, Chin. Phys. Lett. **5**, 421 (1988).
[6] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[7] C. K. Peng *et al.*, Nature (London) **356**, 168 (1992).
[8] S. V. Buldyrev *et al.*, Phys. Rev. E **51**, 5084 (1995).
[9] Y. Xiao *et al.*, J. Theor. Biol. **175**, 23 (1995).
[10] C. T. Zhang and R. Zhang, Nucl. Acids Res. **19**, 6313 (1991).
[11] A. Arneodo *et al.*, Phys. Rev. Lett. **74**, 3293 (1995).
[12] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[13] V. R. Chechetkin and A. Yu. Turygin, J. Phys. A **27**, 4875 (1994).
[14] V. R. Chechetkin and V. V. Lobzin, Phys. Lett. A **222**, 354 (1996).
[15] W. J. Lee and L. F. Luo, Phys. Rev. E **56**, 848 (1997).
[16] S. Tavare and B. Song, Bull. Math. Biol. **51**, 95 (1989).
[17] L. F. Luo and Shengli, Acta Scientiarum Naturalium Universitatis Intramongolicae **21**, 229 (1990); see also L. F. Luo, Evolutionary Theory **10**, 75 (1991); and *Collected Works on*

Theoretical Biophysics (Inner Mongolia University Press, Hohhot, 1997).
[18] J. Xu *et al.*, Comput. Biol. Med. **23**, 333 (1993).
[19] W. J. Lee and L. F. Luo, in *Collected Works on Theoretical Biophysics* (Inner Mongolia University Press, Hohhot, 1997).
[20] E. N. Trifonov and V. Brendel, *Gnomic—A Dictionary of Genetic Codes* (Balaban, Philadelphia, 1986).
[21] G. Lennon and R. Nussinov, J. Theor. Biol. **22**, 427 (1985).
[22] P. Lio *et al.*, J. Theor. Biol. **167**, 413 (1994).
[23] A. O. Schmitt, W. Ebeling, and H. Herzel, BioSystems **37**, 199 (1996).
[24] L. F. Luo and F. M. Ji, J. Theor. Biol. **188**, 343 (1997).
[25] L. F. Luo and G. Y. Bai, J. Theor. Biol. **174**, 131 (1995).
[26] P. Allegrini, P. Grigolini, and B. J. West, Phys. Lett. A **211**, 217 (1996).
[27] R. N. Mantegna *et al.*, Phys. Rev. Lett. **73**, 3169 (1994).
[28] S. V. Buldyrev *et al.*, Biophys. J. **65**, 2673 (1993).
[29] L. F. Luo, F. M. Ji, and H. Li, Bull. Math. Biol. **57**, 527 (1995).
[30] R. Silverman and R. Linsker, J. Theor. Biol. **118**, 295 (1986).
[31] S. Havlin and D. Ben-Avraham, J. Phys. A **15**, L311 (1982).
[32] L. F. Luo *et al.*, in *Proceedings of the International Symposium on Theoretical Biomathematics* (Inner Mongolia University Press, Hohhot, 1997); L. F. Luo, Prog. Phys. (in Chinese) **17**, 320 (1997).
[33] H. Haken, *Information and Self organization* (Springer-Verlag, Berlin, 1988).
[34] L. F. Luo and L. Tsai, Acta Scientiarum Naturalium Universitatis Intramongolicae **27**, 781 (1996).
[35] N. Chomsky, IRE Trans. Inf. Theory **2**, 113 (1956).