Model for the distributions of *k*-mers in DNA sequences

Yaw-Hwang Chen
Department of Electronics Engineering, Kun Shan University of Technology, Yung-Kang, Tainan Hsien, Taiwan 710, Republic Of China

Su-Long Nyeo and Chiung-Yuh Yeh
Department of Physics, National Cheng Kung University, Tainan, Taiwan 701, Republic Of China

(Received 21 January 2005; published 18 July 2005)

The evolutionary features based on the distributions of *k*-mers in the DNA sequences of various organisms are studied. The organisms are classified into three groups based on their evolutionary periods: (a) *E. coli* and *T. pallidum* (b) yeast, zebrafish, *A. thaliana*, and fruit fly, (c) mouse, chicken, and human. The distributions of 6-mers of these three groups are shown to be, respectively, (a) unimodal, (b) unimodal with peaks generally shifted to smaller frequencies of occurrence, (c) bimodal. To describe the bimodal feature of the *k*-mer distributions of group (c), a model based on the cytosine-guanine "*CG*" content of the DNA sequences is introduced and shown to provide reasonably good agreements.

DOI: 10.1103/PhysRevE.72.011908 PACS number(s): 87.14.Gg, 87.15.Aa, 02.50.−r

I. INTRODUCTION

A DNA sequence is a long string of four types of bases and is generally tightly packed inside a cell with a complicated three-dimensional structure. It carries the heredity information of an organism with the genetic content of the coding regions given by the local sequential structure based on the universal triplets of nucleotides or codons. A sequence of codons may then be translated to a sequence of amino acids for protein. What is less well known is the possible long-distance statistical structures of DNA sequences. Such structures can vary for different organisms and chromosomes.

The short- and long-distance structures are correlated in many ways and may arise from certain evolutionary mechanisms. In the textual analysis of microbial genomes [1] for their distributions of the frequency of occurrence of various *k*-mers, it was found that they possess the statistical characteristics of a much smaller *random* systems. Here a *k*-mer is an oligonucleotide of length *k* and frequency of short *k*-mers has been considered in studies of molecular evolution [2]. It was suggested [1] that the growth of microbial genomes follows a two-stage process. An initial stage of random growth of a genome to a length of approximately 1000 nucleotides, followed by random duplications of segments averaging around 25 nucleotides.

In this paper, we shall consider applying the analysis method used for the microbial genomes [1] to analyze various organisms including vertebrates. In particular, we consider the *k*-mer distributions for the chromosomes of various organisms including vertebrates. In this study, we shall consider a simple time-independent model for describing a typical feature of the *k*-mer distributions for human chromosomes.

II. MATERIALS AND METHODS

A. Data sources

The DNA sequences of the organisms we shall analyze are listed in Table I and are available from Ref. [3]. In particular, the DNA sequences of human chromosomes will be taken from Ref. [4]. We shall consider DNA sequences given in FASTA sequence format. In this format, the human DNA sequences contain undetermined bases denoted by the letter "*N*."

In Table I, we group the organisms into three evolutionary periods [5]. The species *Arabidopsis thaliana*, which is a small flowering plant, is included for comparison.

B. *k*-mer distribution

In this paper, we shall study how *k*-mers are distributed in DNA sequences. Let us first define the distribution of *k*-mers by considering, as an example, tossing of a die 15 000 times. Suppose the outcomes (numbers of occurrence) for the values on the faces of the die are given in Table II.

Table II provides a simple statistical property that the numbers of occurrence are approximately the same. A more useful information about the die may be described by considering the standard deviation and higher moments of the distribution. But, here we shall consider a different quantitative description of the statistical property. We round the numbers of occurrence to the nearest hundred and obtain the resulted numbers which are also given in Table II.

TABLE I. List of organisms considered in this study.

Taxonomic name	Name used in text
Escherichia coli K 12	*E. coli*
Treponema pallidum	*T. pal*
Drosophila melanogaster	Fruit fly
Saccharomyces cerevisiae	Yeast
Arabidopsis thaliana	*A. thaliana*
Danio rerio	Zebrafish
Mus musculus	Mouse
Gallus gallus	Chicken
Homo sapiens	Human

TABLE II. The numbers of occurrence of the values on the faces of a die.

Value	1	2	3	4	5	6
Number of occurrence	2350	2510	2540	2510	2600	2490
Round-off number	2400	2500	2500	2500	2600	2500

From Table II, we add up the same number of the frequency of occurrence to obtain Table III, which are then depicted in Fig. 1.

In this example, the number of tosses 15 000 may be identified as the total length of a DNA sequence, and the values on the faces of the die are then identified as the types of bases. The number distribution describes the "monomers" ($k=1$) with die values $1\sim 6$. Here we denote N_k as the number of k-mers with frequency of occurrence f.

With the analysis procedure, one may consider a sequence of k-mers of any length. In the die-tossing example, we see that when $k=2$, the number of 2-mers or dimers to be considered is $6^2=36$. Thus, it becomes very complicated to evaluate the distribution of the frequencies of occurrence for the cases when k is large.

In the following sections, we shall study the k-mer distributions for the DNA sequences of the organisms listed in Table I. We shall consider the cases with $k=2$, 4, 6, 8, in which the numbers of k-mers are 4^k ($k=2$, 4, 6, 8). In this paper, all distributions will be plotted with sequences normalized to one megabase (Mb) for comparison purpose; that is, all distributions are normalized to 1 Mb of sequence length.

C. Human chromosome 21 and random DNA sequence

As an example, we shall compare the distributive properties of k-mers in human chromosome 21 and in a random DNA sequence. The results of our analysis for the cases $k=2$, 4, 6, 8 are represented in Figs. 2–5. From these figures, we see that, as have been pointed out long ago [1,6], the distributions of k-mers in human chromosome 21 are very different from those in the random DNA sequence. We note that all human chromosomes have very similar k-mer distributions with a bimodal structure for large k values (Figs. 4 and 5). The fact that all human chromosomes have very similar k-mer distributions has been pointed out in [7], in which all the eukaryotic genomes examined were shown to possess very similar distributions.

In Figs. 2–5, the k-mer distributions for the random DNA sequence of length L ($L \gg 4^k$) are known to obey the Poisson's formula with the peak values

$$\bar{f}_m(p)=\bar{f}2^k p^m(1-p)^{k-m}, \quad 0 \leqslant m \leqslant k, \tag{1}$$

where p is the fraction of (adenine + thymine) ($A+T$) in the random sequence (there is a general symmetry of genomes that the numbers of the bases A (C) and T (G) are almost the same), $\bar{f}$ is the average frequency of occurrence $\bar{f}=L/4^k$. Thus, for the 6-mer distribution for the random sequence given in Fig. 4, there are seven peaks given by Eq. (1) with $m=0$, 1, …, 6.

TABLE III. The frequencies and their numbers of occurrence in the die tossing.

Frequency	Number
2400	1
2500	4
2600	1

D. k-mer distributions of organisms of different groups

From the above analysis, we see that the k-mer distributions for human chromosomes and a random sequence are very different. Here we shall consider and evaluate the distributions of k-mers ($k=2\sim 8$) of the organisms in Table I. The distributions for the case $k=6$ are plotted according to the three evolutionary groups. We observe that the k-mer distributions for group (a) organisms (Fig. 6) are unimodal. The k-mer distributions for group (b) organisms are also unimodal (Fig. 7), but compared to group (a) organisms, their peaks are generally slightly shifted to smaller frequencies of occurrence. The k-mer distributions for group (c) organisms (Fig. 8) are bimodal. Thus, the complexity of a k-mer distribution increases from group (a) to group (c) organisms. Here we also note that *Arabidopsis thaliana*, a small flowering plant, behaves like the bacteria in group (a).

We should mention that a particular characteristic property of a k-mer distribution for an organism does not necessarily reflect that the organism belongs to a particular evolutionary group, since we have analyzed only a few organisms in each evolutionary group. In particular, many microbial genomes have highly biased base compositions, and their distributions [1,7] can have a larger skewness than those of the genomes in group (b).

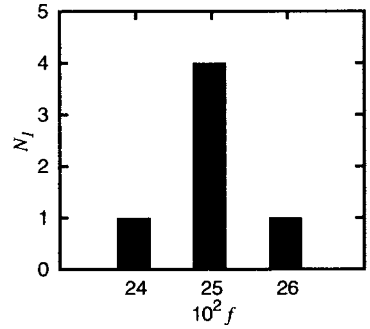

FIG. 1. The plots of the number distribution of the frequencies of occurrence of the die tossing.

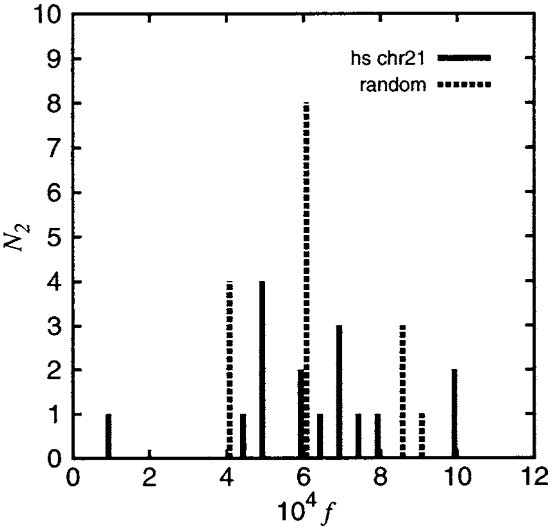

FIG. 2. The distributions of 2-mers in human chromosome 21 and in a random DNA sequence.

E. Minimal microbial genome growth model

In order to explain the distributive features of *k*-mers in microbial genomes in terms of an evolutionary model, a minimal model which employs two types of evolutionary events, mutation and DNA duplication, was proposed [1]. In this model, mutation events are caused by single base replacements (SBR), and DNA duplication events are given by occasional random duplication (RD) of a stretch of oligonucleotide with a characteristic length scale σ.

A model DNA sequence is generated with an initial simple random sequence of length L_0 of a prescribed base composition. This initial sequence is then allowed to evolve and grow by SBR and RD events until its final length just exceeds 1 Mb. In RD event, a length ℓ of the copied sequence is randomly selected, then a site s sufficiently far from the end of the genome is randomly chosen and the sequence from s to $s+\ell-1$ is copied and inserted into the genome behind a second randomly selected site. This model has three adjustable parameters: an initial length L_0, a ratio η of the chances of having an SBR or an RD event, and a length scale σ. For example, to fit the 6-mer distribution of *E. coli*, the parameters $L_0=1000, \eta=500$, and $\sigma=15\,000$ were used.

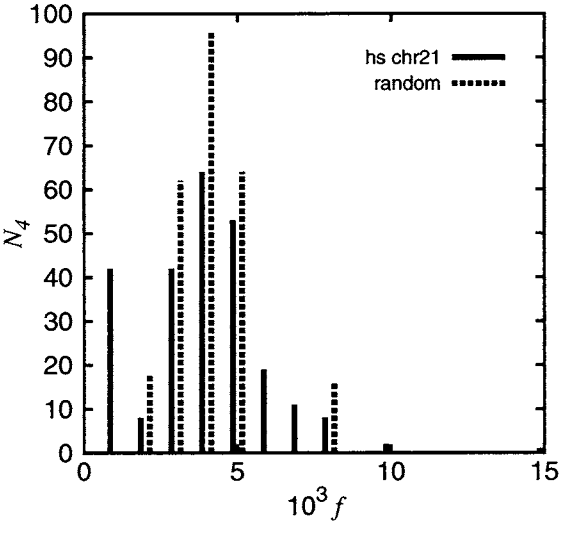

FIG. 3. The distributions of 4-mers in human chromosome 21 and in a random DNA sequence.

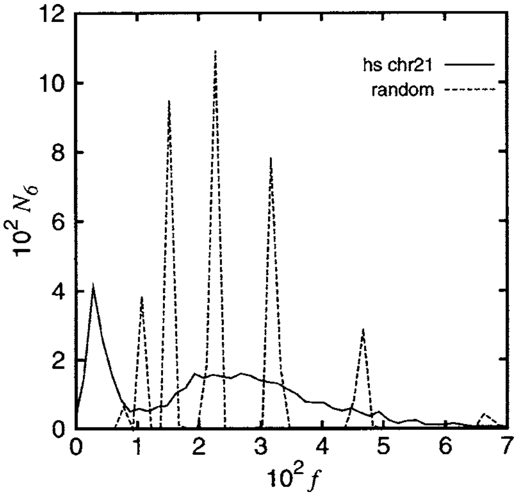

FIG. 4. The distributions of 6-mers in human chromosome 21 and in a random DNA sequence.

In a computer simulation [1], a random length ℓ is chosen as according to the Erlang probability function [8]

$$F_m(\ell)=\frac{1}{\sigma m!}\left(\frac{\ell}{\sigma}\right)^m e^{-\ell/\sigma}, \quad m=0,1,2,\ldots, \tag{2}$$

where σ is a chosen length scale (in bases) and m is a shape parameter. The probability function (2) has an average value $(m+1)\sigma$ and a standard deviation $(m+1)^{1/2}\sigma$. We let L_c be the current length of an artificial genome. Then the probability per unit length of a selected segment of length x is given $w(x)=\sigma^{-1}e^{-x/\sigma}(1-e^{-L_c/\sigma})^{-1}$ with normalization $\int_0^{L_c}w(x)dx=1$. To select a segment ℓ using a random number generator, we construct a function G such that $G(\sigma,y)=\ell$ and $y\in[0,1)$ is

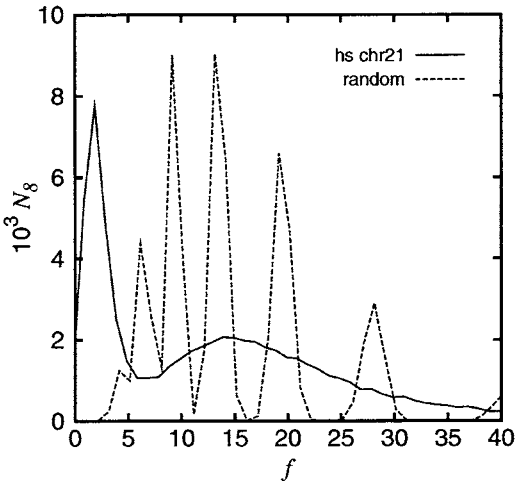

FIG. 5. The distributions of 8-mers in human chromosome 21 and in a random DNA sequence.

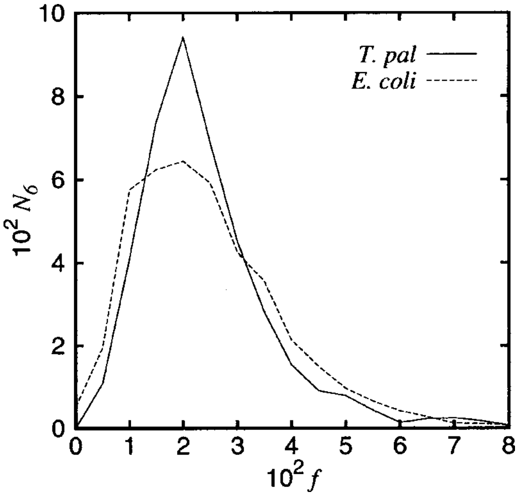

FIG. 6. The 6-mer distributions for group (a) organisms.

a computer generated random number. The inverse function $G^{-1} \equiv y = \int_0^{\ell} w(x)dx$ yields the required random length $\ell = G(\sigma, y) = -\sigma \ln[1 - y(1 - e^{-L_c/\sigma})]$ [1].

Clearly, the final genome length L is much greater than L_0 and σ. Also, the total number of RD events is greater than L/σ and the total number of SBR events is greater than $\eta L/\sigma$.

This model generates k-mer distributions that agree with the distributions for the corresponding real microbial genomes with chosen parameters [1]. In addition, it was shown [7] that the minimal model with one set of three universal parameters can be used to account for the primary feature of the k-mer distributions of essentially all microbial and eukaryotic genomes. The primary feature of a k-mer distribution is specified by its effective random-sequence length, which provides a reference parameter for generating a model sequence.

However, the minimal model does not give bimodal k-mer distributions as seen in some eukaryotes such as the third group of organisms in Table I. To generate a k-mer distribution with a bimodal structure, which is a secondary feature, one or more additional conditions are needed and we shall describe a simple condition as follows.

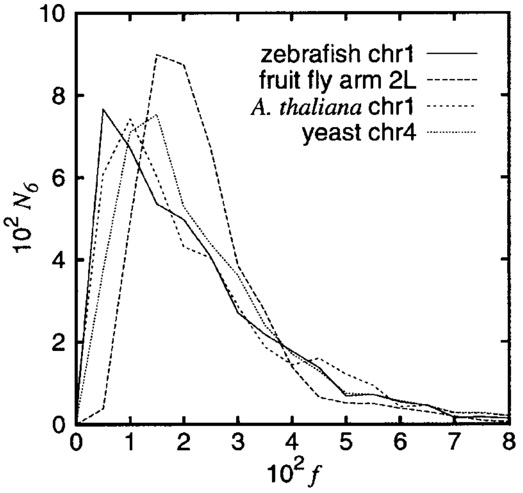

FIG. 7. The 6-mer distributions for group (b) organisms.

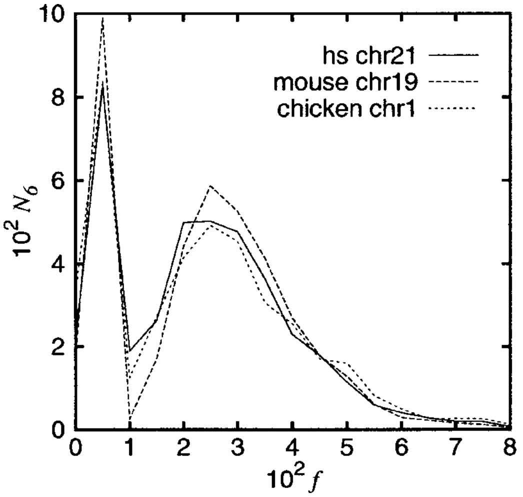

FIG. 8. The 6-mer distributions for group (c) organisms.

F. Model for bimodal k-mer distributions

To account for the bimodal structure in the k-mer distributions for the human genome, we now propose a random duplication mechanism, which is based on the minimal model. As in the minimal model, we need to estimate the initial length L_0 by calculating the effective length from the average and standard deviation. For a random DNA sequence which has a Poisson type of k-mer distributions, we define the average

$$\overline{N_k} = \frac{\sum_{n=1}^{4^k} N_k(n)}{4^k}, \tag{3}$$

and the standard deviation

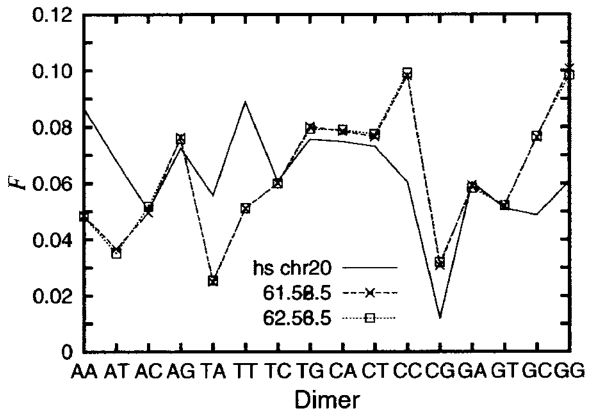

FIG. 9. The fractions of the dimers in the complete human chromosome 20 and in sequences in the chromosome of 1 Mb length.

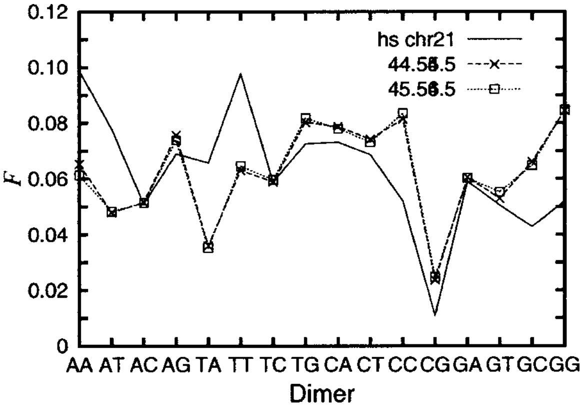

FIG. 10. The fractions of the dimers in the complete human chromosome 21 and in sequences in the chromosome of 1 Mb length.

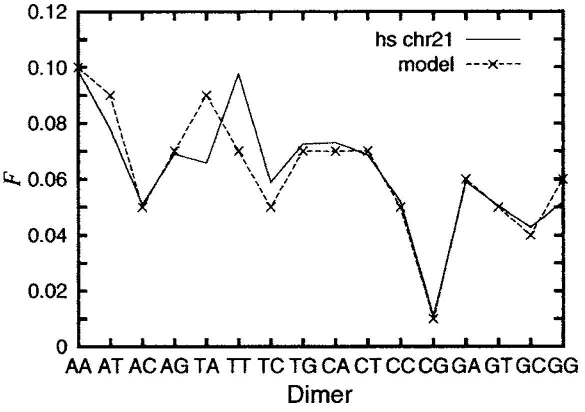

FIG. 12. The fractions of the dimers in the complete human chromosome 21 and in the model sequence.

$$\sigma_k = \sqrt{\frac{\sum_{n=1}^{4^k} (N_k(n) - \overline{N_k})^2}{(4^k - 1)}}, \qquad (4)$$

where $N_k(n)$ denotes the number of a k-mer in a sequence of length normalized to 1 Mb. The effective random-sequence length $L_{\rm eff}$ and the ratio $r = \overline{N_k}/\sigma_k$ are related by $L_{\rm eff} = 4^k r^2$ [1]. For human chromosomes, we calculate the values of $L_{\rm eff}$ for various k values in order to determine L_0. We note that $L_{\rm eff}$ ranges from the average value of 149 base pairs (bp) for $k = 2$ to about $L_{\rm eff} = 3.9 \times 10^4$ bp for $k = 8$. To obtain the optimal value for L_0, we perform a χ^2 procedure based on comparing the empirical values of $L_{\rm eff}$ with those from a set of model sequences [7,9]. For human chromosomes, we use an initial length $L_0 = 300$ bp.

In order to understand the bimodal structure of the k-mer distributions for group (c) organisms, in particular, for human chromosomes, we shall analyze the content of the dimers in human chromosomes to determine how the dimers are distributed. In particular, the fraction of the dimer "*CG*" that plays an important role in DNA methylation, which varies widely among different organisms [10]. Thus, suppose the ratio of the fraction of the number of "*CG*"s in the total length is p_0. Then for a chosen length ℓ, we assign a suppressing factor for the evolution of a "*CG*" given by e^{-p/p_0}, where $p = n_\ell(CG)/\ell$ with $n_\ell(CG)$ denoting the number of "*CG*"s in the length ℓ. Here we note that the definition of p is not to be considered as the fraction of "CG"s with respect to the total number of dimers in the chosen length ℓ. To define such a fraction, we would have to evaluate the total number of dimers in the chosen length ℓ, but ℓ can have an odd number of bases. A more appropriate suppressing factor may be $e^{-\alpha p/p_0}$ with some constant α.

To implement a duplication suppressing factor into our evolutionary model, we consider the Erlang probability function (2) for a DNA segment of length ℓ (in bases). We take an average duplication length $\overline{\ell}$ bp, corresponding to a value pair (m, σ), and the ratio η of SBR to RD. Then we impose the additional condition such that a duplication is allowed if the randomly computer generated value between 0 and 1 is smaller than e^{-p/p_0}, and it is forbidden, otherwise.

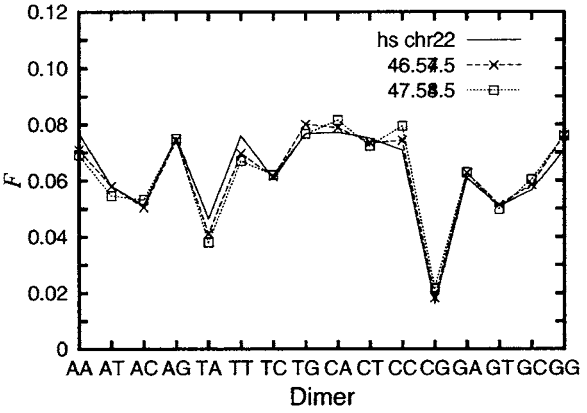

FIG. 11. The fractions of the dimers in the complete human chromosome 22 and in sequences in the chromosome of 1 Mb length.

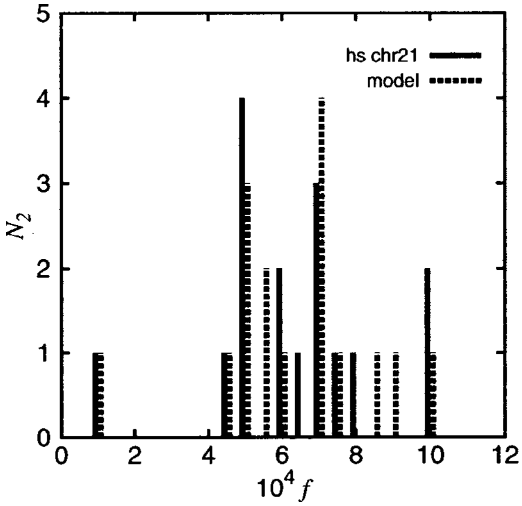

FIG. 13. The distributions of 2-mers in human chromosome 21 and in the model sequence.

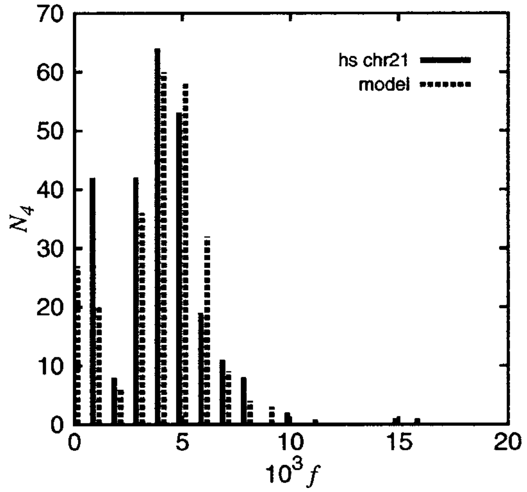

FIG. 14. The distributions of 4-mers in human chromosome 21 and in the model sequence.

G. Fractions of dimers in human chromosomes

As an application, we evaluate the numbers of the dimers in human chromosomes 20, 21, and 22. Their dimer distributions for DNA sequences of length of 1 Mb are given in Figs. 9–11. We see that for the fractions of the dimers on the left of the dimer *TA* in these figures are generally smaller than the values of the dimers in the whole sequences. Other chromosomes exhibit similar features.

From the distributions of the fractions of dimers, we shall see that the dimer "*CG*" occupies about 0.01 in the total base length of a chromosome; i.e., the ratio of the number of " *CG*"s to the total number of dimers in the complete chromosome is 0.01, which is much smaller than the average value $1/4^2=0.0625$. Therefore, we shall take the duplication suppressing factor for the the evolution of the "*CG*" to be $e^{-p/0.01}$. Note that the fractions of the dimer "*CG*" in the chosen sequences of 1 Mb length in the chromosome are about 0.01.

III. RESULTS AND CONCLUSION

To generate model sequences with our model, we choose the average duplication length $\bar{\ell}=20$ bp $(m=3,\sigma=5)$ and the ratio of SBR to RD to be $\eta=0.6$. We note that when $\bar{\ell}=20$, the number of duplications reaches about 1.5×10^5 times. In our model, we first generate ℓ and evaluate $n_\ell(CG)$ to obtain $p=n_\ell(CG)/\ell$, then a DNA duplication is allowed if the randomly computer generated value between 0 and 1 is smaller than $e^{-p/0.01}$, and it is forbidden, otherwise. Thus, using this generating procedure, we obtain a model sequence for most of human chromosomes whose fractions of the dimer "*CG*" are about 0.01.

TABLE IV. The average fractions of dimers with weak-H bond bases (*W*) and strong-H bond bases (*S*) in human chromosome 21 and in model sequence

Dimer	Fraction	
	Chromosome 21	Model
$\overline{WW}$	0.340	0.350
$\overline{WS}+\overline{SW}$	0.502	0.490
$\overline{SS}$	0.158	0.160
$\overline{SS}-CG$	0.147	0.150

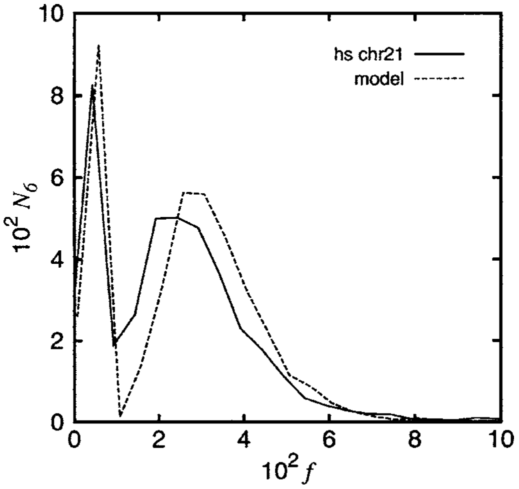

FIG. 15. The distributions of 6-mers in human chromosome 21 and in the model sequence.

A. Fractions of dimers in the model sequence

From the model generated sequence, it is easy to evaluate the fractions of the dimers, which are given in Fig. 12. The fractions of the dimers in the model sequence are seen to be in reasonably good agreements with those in the real human chromosome 21, in particular, for the dimer "*CG*" with fraction 0.01. We see that there are reasonably good agreements in the fractions of the dimers associated with the strong-H bond bases $S=(C,G)$. However, there are large differences in the fractions of the dimers associated with the weak-H bond bases $W=(A,T)$, such as "*AT*", "*TA*", and "*TT*". We note that

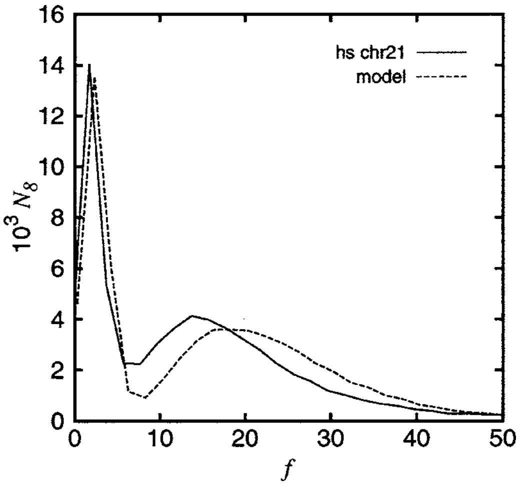

FIG. 16. The distributions of 8-mers in human chromosome 21 and in the model sequence.

all other chromosomes have small “*CG*” fraction of about 0.01 and exhibit similar feature. Such good agreements may be fortuitous, since we have used one condition on only the dimer “*CG*” to fit all 16 dimers. Thus, it is more appropriate to compare the average fractions of dimers of the types WW, WS, SW, and SS with four dimers each. From these, we have evaluated the average values $\overline{WW}, \overline{WS}+\overline{SW}, \overline{SS}$, and $\overline{SS-CG}$ for human chromosome 21 and model sequence. The values given in Table IV agree reasonably well.

B. The *k*-mer distributions in the model sequence

Also, from the model sequence, we can evaluate the *k*-mer distributions. The results for the cases $k=2,4,6,8$ are plotted in Figs. 13–16 with those for human chromosome 21. There are reasonably good agreements in the *k*-mer distributions for the model sequence and the real human chromosome 21. In particular, the model is able to produce a sequence with bimodal *k*-mer distributions.

In conclusion, we note that the reason for using a duplication suppressing factor in our model may be seen as follows. In human chromosomes, the fractions of the bases A, T, C, and G are about 30% for A and T, and 20% for C and G. In these, about 5% of C are methylated and composed of about 1% of the total bases. This value is about the fraction of “*CG*” in human chromosomes. Thus, we see that most of the “*CG*”s are methylated. By suppressing the duplications of the dimer “*CG*,” we have effectively adjusted the distributions of the dimers related with the bases C and G. The small content of “*CG*” in human chromosomes may be caused by the fact that methylated “*CpG*”s in the double helix DNA structure lead to tighter helix structure and hence to fewer duplications. Here “*CpG*” simply indicates that “C” and “G” are connected by a phosphodiester bond. This in turn leads to a suppression of “*CG*”s in the genomes of higher organisms such as those in group (c).

Finally, we note that larger discrepancies exist in the fractions of the dimers associated with the weak-H bond bases like “*AT*”, “*TA*”, and “*TT*”. One reason is that there are 16 different dimers in a sequence and we have one duplication suppressing factor on only the dimer “*CG*.” Also, other factors may be taken into account. For example, during the evolution of a DNA sequence, the number of “*CG*”s in the sequence is not necessarily a constant.

ACKNOWLEDGMENTS

This work was supported by the National Science Council of the Republic of China under Contract No. NSC93-2112-M006-012 and by the National Center for Theoretical Sciences of the Republic of China.

[1] L. C. Hsieh and H. C. Lee, Mod. Phys. Lett. B **16**, 821 (2002); L. C. Hsieh, L. Luo, F. Ji, and H. C. Lee, Phys. Rev. Lett. **90**, 018101-1 (2003); L. C. Hsieh, L. Luo, and H. C. Lee, *Evidence for Growth of Microbial Genomes by Short Segmental Duplications*, in Proceedings of the 2nd IEEE Computer Society Bioinformatics Conference (CSB 2003), Stanford University, Stanford, CA (2003), p. 474.

[2] C. Burge, A. M. Campbell, and S. Karlin, Proc. Natl. Acad. Sci. U.S.A. **89**, 1358 (1992); S. Karlin, J. Mrázek, and A. M. Campbell, J. Bacteriol. **179**, 3899 (1997).

[3] ftp://ftp.ncbi.nih.gov/genomes/

[4] ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/April_14_2003/

[5] A. Sidow, Curr. Opin. Genet. Dev. **6**, 715 (1996); A. Sidow, Cell **111**, 13 (2002).

[6] H. Xie and B. Hao, *Visualization of K-tuple distribution in procaryote complete genomes and their randomized counterparts*, in Proceedings of the 2002 IEEE Computer Society Bioinformatics Conference (CSB 2002), Stanford University, Palo Alto, CA (2002), p. 31.

[7] C. H. Chang, L. C. Hsieh, T. Y. Chen, H. D. Chen, L. Luo, and H. C. Lee, *Shannon Information in Complete Genomes*, in Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004), Stanford, CA (2004), p. 20.

[8] E. Brockmeyer, H. L. Halstrøm, and Arne Jensen, *The Life and Works of A.K. Erlang*, (Trans. Dan. Acad. Techn. Sci. No. 2, Copenhagen, 1984). [A free copy is available from http://www.com.dtu.dk/teletraffic/Erlang.html.]

[9] T. Y. Chen, L. C. Hsieh, C. H. Chang, L. Luo, F. M. Ji, and H. C. Lee, Int. J. Mod. Phys. B **18**, 2448 (2004).

[10] B. Hendrich and S. Tweedie, Trends Genet. **19**, 269 (2003).