Characterizing self-similarity in bacteria DNA sequences

Xin Lu,[1,*] Zhirong Sun,[2] Huimin Chen,[1] and Yanda Li[1]
[1]*Department of Automation, Tsinghua University, 100084 Beijing, People's Republic of China*
[2]*Department of Biology, Tsinghua University, 100084 Beijing, People's Republic of China*
(Received 24 November 1997; revised manuscript received 27 April 1998)

In this paper some parametric methods are introduced to characterize the self-similarity of DNA sequences. Compared with Fourier analysis, these methods perform statistically more stably and yield more reliable results. Using these methods, eight whole genomes of bacteria provided by NCBI are analyzed. Long-range correlation properties in the nucleotide density distribution along these DNA sequences are explored. Estimation results show that the long-range correlation structure prevails through the entire molecule of DNA. Higher order statistics through coarse graining reveal that rather than multifractal, there are only monofractal phenomena presented in the sequences. Hence, the nucleotide density distribution can be modeled asymptotically as fractional Gaussian noise. This result points to a new direction for analyzing and understanding the intrinsic structures of DNA sequences. [S1063-651X(98)01009-5]

PACS number(s): 87.10.+e

I. INTRODUCTION

Many attempts have been made to find informational content in DNA sequences, especially with the progress of the human genome project. The coding sequences have been studied extensively, and in recent years, the study of correlation structure in DNA sequences has evolved from local consideration to global awareness. As commented by Román-Roldán *et al.*: ''(1987 to present), . . . The most outstanding result was the finding of long-range correlation in DNA sequences.'' [1].

During the 1990s, several statistical methods originating from the field of signal processing and pattern recognition were introduced into the study of long-range correlation in DNA sequences. More details about the studies in this field, their methods and results, can be found in [2–11] and the review of Román-Roldán *et al.* [1] and Li [12].

In previous works, Fourier analysis was used as one of the major techniques to detect and analyze long-range correlation properties. A power law in the lower-frequency part of the DNA spectrum is considered to be evidence of long-range correlation structure. The long-range correlation structure is described as a $1/f^{\alpha}$ process, and the α value is used as the distinguishing feature among different classifications as well as between the coding and noncoding sequences [3,5,9,11]. However, in terms of statistical accuracy, Fourier analysis is not the best way to estimate the fractional exponent, for it lacks statistical robustness in parameter estimation.

Before further investigating the long-range correlation properties in DNA sequences, we must first be aware of some technical notions about the length scale as well as the model assumption of the correlation structures.

First, what is the length scale meant by the term ''long-range'' in the terminology ''long-range (base-base, statistical) correlation''? In some early papers, it is meant to be longer than (a) 3–6 bases [13] or (b) 800 bases [5] or (c) 1–10kb [8]. The progress in sequencing has enabled researchers to analyze longer DNA sequences and gather more information. The human genome project is scheduled to finish in about 2005; at that time, we will have DNA sequences of several hundred million bases, or even several billion bases in length. This makes it possible to study the asymptotic long-range correlation structure of DNA sequences. At the same time, the study can reveal much about correlation structure in the lower-frequency part. Therefore the method to estimate the correlation structure must also be considered for longer sequences. In other words, although the DNA sequence lengths are always bounded, the statistical method used must still be able to describe the asymptotic properties of the correlation structure when the data are approaching infinity.

The second is about the potential mathematical model when we use various methods to describe the correlation structures. The DNA sequences must be assumed to be ergodic and stationary before being analyzed by Fourier transformation. DNA sequences all have finite lengths and can be modeled as self-similar processes embedded in white background noise. Thus, under this condition, parametric methods will be more promising than nonparametric methods such as Fourier analysis, as far as modeling accuracy is concerned, and the result will be more robust. This is the reason that we use the Fractional Gaussian Noise (FGN) model and Hurst index to describe the long-range correlation properties in this paper.

Another statistical issue we should be aware of is that the parametric methods are model dependent. When we model the DNA sequence as a FGN process, we must check the valid scale and model accuracy of the correlation structure. As discussed in [5], DNA sequences exhibit a ''partial $1/f^{\alpha}$ spectrum,'' so that the length scale that possesses these properties should be checked carefully while using these methods.

In the field of traffic modeling in computer networks, long-range correlation phenomena have been investigated

*Electronic address: luxin@jerry.au.tsinghua.edu.cn

extensively [14–17]. Many statistical algorithms to verify long-range correlation have also been implemented successfully. Using these methods, we analyzed in great detail the long-range correlation properties in the DNA sequences. The goal of this paper is to explore the asymptotic long-range correlation properties of DNA sequences, therefore we selected whole genomes of bacteria, which usually have the length of more than 1 000 000 base pairs (BP).

In this paper whole genomes of seven eubacteria: Escherichia coli (4 639 221 bp, G+C: 50.8%), Haemophilus influenzae (1 830 140 BP, G+C: 38.1%), Helicobacter pylori (1 167 867 BP, G+C: 38.9%), Mycoplasma genitalium (580 073 BP, G+C: 31.7%), Mycoplasma pneumoniae (816 394 BP, G+C: 40.0%), Rhizobium sp. NGR234 (536 165 BP, G+C: 58.5%), Synechocystis PCC6803 (3 573 470 BP, G+C: 47.7%), and 1 archaea: Methanococcus jannaschii (1 664 970 BP, G+C: 31.4%) are investigated. (These data are provided by NCBI ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria G+C stands for guanine+cytosine). Using the average spectrum and other statistical methods, it was found that there exist long-range correlation properties in the DNA sequences of bacteria, within scales longer than 200 BP. This property does not vanish when the length is extended to the whole genome. Thus the correlation structure under the length scale between 200 BP and the whole sequence can be modeled by the FGN process asymptotically.

The paper is organized as follows. In Sec. II we give a brief introduction to the self-similar stochastic process (or long-range dependent process), as well as the statistical methods of Hurst index estimation. In the study of long-range correlation structures of DNA sequences, the parametric methods introduced in Sec. II will be more promising and the results will be more robust than nonparametric methods such as Fourier analysis. In Sec. III the long-range correlation properties in the DNA sequences of bacteria are studied. Their Hurst indexes are estimated by various methods, and the results are presented. The consistency among these results further supports the existence of long-range correlation properties in DNA sequences of bacteria and reinforces the reliability of these methods. Finally, biological meaning and origin of long-range correlation properties are discussed in Sec. IV.

II. MATHEMATICAL DEFINITION OF SELF-SIMILAR STOCHASTIC PROCESS AND ESTIMATION OF THE HURST INDEX

In this section we briefly introduce the self-similar stochastic process and the methods of Hurst index estimation. Details can be found in [15,18,19] and relative topics listed therein.

Suppose we have a covariance stationary stochastic process $X=(X_1,X_2,\ldots,X_n,\ldots)$, which has the following autocorrelation structure:

$$r(k)=E[(X_t-\mu)(X_{t+k}-\mu)]\sim k^{-\beta}L_1(k),\quad k\to\infty \tag{1}$$

where $0<\beta<1$, and L_1 is a slowly varying function, that is, $\lim_{k\to\infty}L_1(kx)/L_1(k)=1$ for all $x>0$. Then we call X a self-similar stochastic process in the sense of Eq. (1) [14]. In frequency domain, it can be expressed equivalently as

$$f(\lambda)\sim\lambda^{-1+\beta}L_2(\lambda),\quad \lambda\to 0 \tag{2}$$

where L_2 is a slowly varying function and $f(\lambda)$ is the spectral density function.

Definition 1. Let $X^{(m)}=(X_1^{(m)},X_2^{(m)},\ldots)(m=1,2,3,\ldots)$ be the m block process of X, defined as summing the original series X over nonoverlapping blocks of size m: $X^{(m)}(k)=(1/m)\Sigma_{i=(k-1)m+1}^{km}X(i)$, $k=1,2,\ldots$, and $r^{(m)}$ is the autocorrelation function of $X^{(m)}$.

Definition 2. X is called strictly (second order) self-similar with the Hurst index $H=1-\beta/2$, if the m block process $X^{(m)}$ has the same correlation structure with the original process X, that is,

$$r^{(m)}(k)=r(k)\quad \text{for all } (m=1,2,\ldots,k=1,2,\ldots). \tag{3}$$

Definition 3. If Eq. (3) can only hold when $m\to\infty$, then X is called asymptotically self-similar, with the Hurst index $H=1-\beta/2$. That is, when $m\to\infty$,

$$r^{(m)}(1)\to 2^{1-\beta}-1, r^{(m)}(k)\to\frac{1}{2}\delta^2(k^{2-\beta})(k=2,3,\ldots), \tag{4}$$

where $\delta^2(f(k))=f(k+1)-2f(k)+f(k-1)$. Definition 2 is often used in modeling the self-similar stochastic process, for it can express the properties of the stochastic process under large m (time scale), which is what we are interested in.

The main feature of a strictly and asymptotically self-similar stochastic process is the autocorrelation function $r^{(m)}(k)\neq 0$ when $m\to\infty$. But in the short-range correlation process, the $X^{(m)}$ asymptotically approaches white noise, or $r^{(m)}(k)\to 0$ when $m\to\infty$. Therefore stochastic processes that satisfy definition 3 are also called long-range dependent processes. Its characteristic in the frequency domain is that $f(\lambda)\sim\lambda^{-1+\beta}L_2(\lambda)$ $(0<\beta<1)$ when $\lambda\to 0$. $L_2(*)$ is also a slowly varying function and $f(\lambda)=\Sigma_k r(k)e^{-ik\lambda}$ is the spectrum density function. The FGN process, which is implemented in this paper for description of the correlation properties in the lower-frequency part of the DNA spectrum, is one of the simplest and most well established models of stationary self-similar processes.

One of the attractive features of using self-similar models, when appropriate, is that the degree of self-similarity can be expressed by only a single parameter. This parameter expresses the speed of decay of the autocorrelation function. The parameter used is the Hurst index $H=1-\beta/2$. Thus, for self-similar series with long-range dependence, $1/2<H<1$. As $H\to 1$, the degree of both self-similarity and long-range dependence increases. Otherwise, if the Hurst index estimated $\hat{H}$ is approximately equal to 0.5 when $m\to\infty$, it means that only short-range correlation is presented, or that $X^{(m)}$ is asymptotic white noise [20].

It has been discussed in [3,5,9,11] that the DNA sequences act as self-similar processes. Therefore, in the estimation of self-similar exponents of DNA sequences, the implementation of fractal-based methods, which possess the nature of self-similarity, is natural and convenient, and the

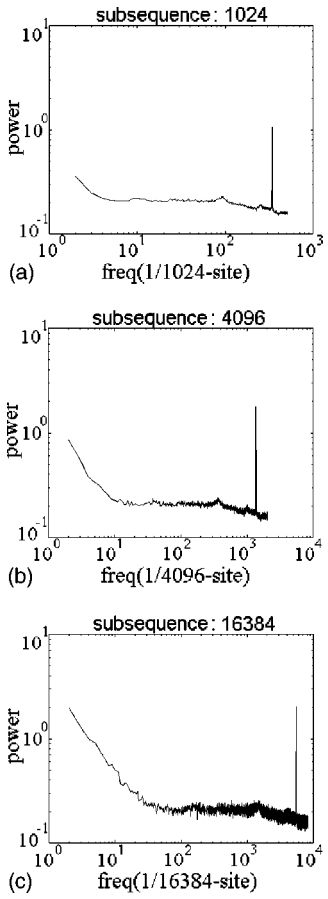

FIG. 1. Average spectrum of adenine from E. coli with subsequences (a) 1024, (b) 4096, and (c) 16 384.

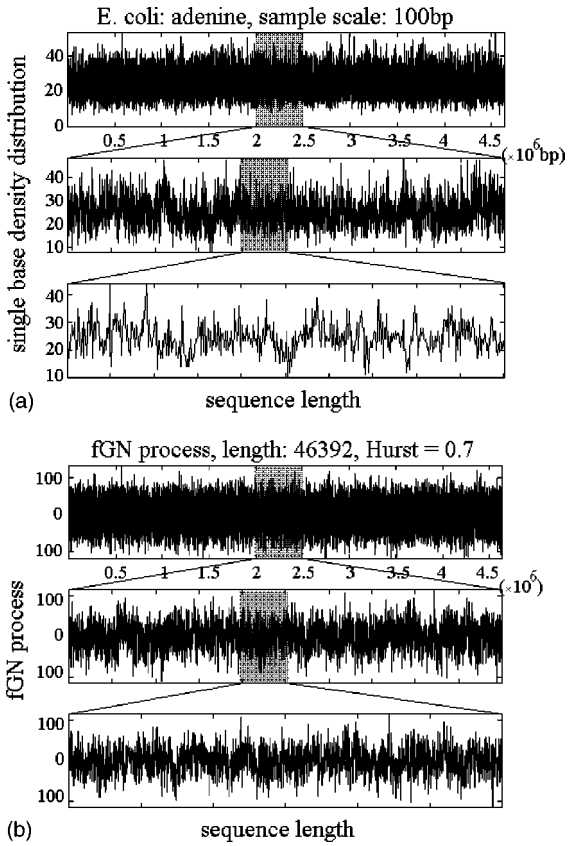

FIG. 2. (a) 100 BP density distribution of adenine from E. coli and (b) a FGN process with the same length and a Hurst index of 0.7.

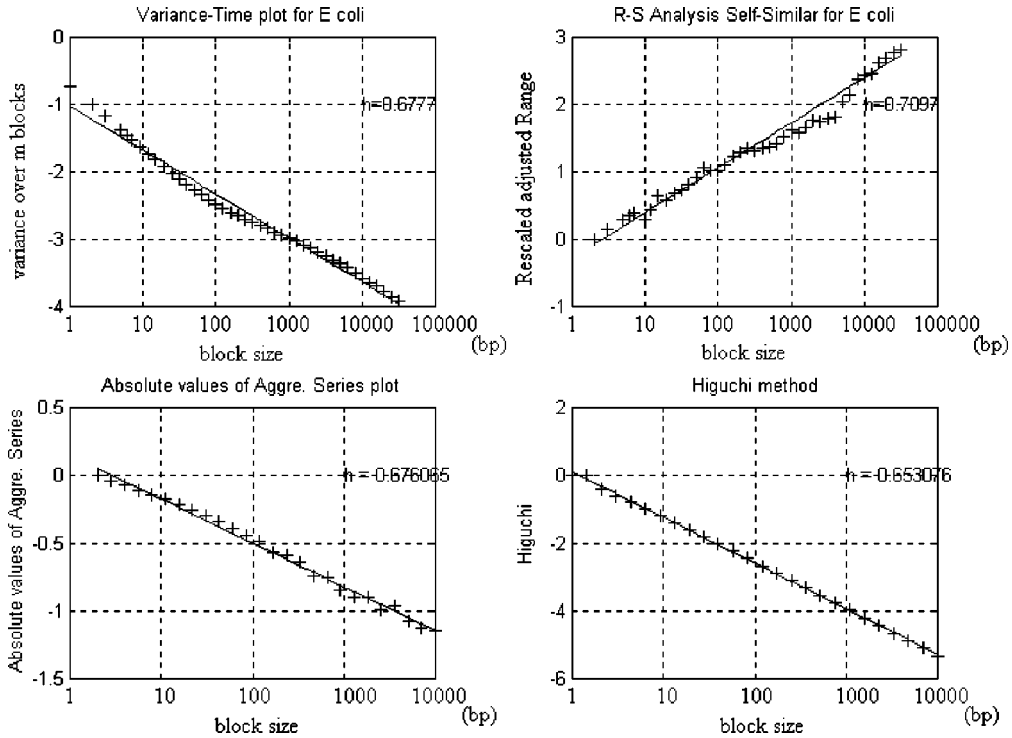

FIG. 3. Linear fitting figure for Hurst index estimation of adenine from E. coli. The X axis is the log plot of block size, which reflects the length scale in which the correlation structure exists. The Y axis is the statistics calculated by various methods. Only the slopes in the figures are analyzed, which are directly related to the Hurst index.

TABLE I. Estimation results of various methods: E. coli.

	A	T	C	G
RS	0.7097	0.7014	0.7284	0.7414
Variance time	0.6777	0.6774	0.7819	0.8061
Scale 2	0.7301	0.7330	0.7259	0.7246
Scale 4	0.7310	0.7337	0.7266	0.7250
Scale 6	0.7309	0.7340	0.7259	0.7241
Scale 8	0.7284	0.7319	0.7207	0.7176
Scale 10	0.7096	0.7171	0.6980	0.6918
Higuchi	0.6531	0.6615	0.7770	0.7934
AVAS	0.6760	0.6770	0.7940	0.8222
$q=1.5$	0.6732	0.6726	0.7828	0.8088
$q=2$	0.6692	0.6685	0.7732	0.7977
$q=2.5$	0.6644	0.6646	0.7648	0.7883
$q=3$	0.6559	0.6608	0.7571	0.7800
$q=3.5$	0.6536	0.6570	0.7501	0.7501
$q=4$	0.6481	0.6532	0.7437	0.7437

results yielded will be more accurate. The statistical tools used in this paper for the estimation of Hurst index include time domain methods such as RS (rescaled adjusted range statistic) analysis [21], variance-time analysis [18], the AVAS method (absolute values of the aggregated series), the Higuchi method [19], and the wavelet-domain-based EM analysis method [22].

The main idea of time domain methods can be summarized as two steps: First, to calculate some statistics in the m block process; second, to investigate the power-law relationship between the block size m and the calculated statistics. The wavelet domain EM method can be used for the estimation of fractal parameter of signals under additive white noise. This method is an iterative algorithm, including an E step and an M step. Using parameters under various wavelet scales, the E step gets the estimation for variance of signal and noise. The M step iterates the estimation results until they converge in a MLE (maximum likelihood estimate) sense. This method can be used to examine the consistency of estimation results under various wavelet scales.

TABLE II. Estimation results of various methods: H. influenzae.

	A	T	C	G
RS	0.6779	0.7234	0.7809	0.7916
Variance time	0.6533	0.6807	0.7570	0.7688
Scale 2	0.6954	0.7053	0.7075	0.7083
Scale 4	0.6963	0.7035	0.7072	0.7066
Scale 6	0.6992	0.7044	0.7049	0.7014
Scale 8	0.6940	0.6989	0.6815	0.6780
Scale 10	0.6648	0.6595	0.6441	0.6336
Higuchi	0.6344	0.6741	0.7517	0.7498
AVAS	0.6545	0.6649	0.7677	0.7678
$q=1.5$	0.6506	0.6684	0.7582	0.7654
$q=2$	0.6466	0.6725	0.7492	0.7611
$q=2.5$	0.6424	0.6760	0.7406	0.7560
$q=3$	0.6382	0.6784	0.7321	0.7497
$q=3.5$	0.6338	0.6795	0.7237	0.7237
$q=4$	0.6294	0.6795	0.7155	0.7155

TABLE III. Estimation results of various methods: H. pylori.

	A	T	C	G
RS	0.7419	0.7460	0.7844	0.7697
Variance time	0.6716	0.5931	0.8051	0.7700
Scale 2	0.7412	0.7477	0.7054	0.7023
Scale 4	0.7427	0.7492	0.7050	0.7026
Scale 6	0.7430	0.7483	0.6987	0.6988
Scale 8	0.7353	0.7420	0.6817	0.6771
Scale 10	0.7035	0.7029	0.6506	0.6368
Higuchi	0.6830	0.6186	0.8032	0.7686
AVAS	0.6729	0.5920	0.8059	0.7738
$q=1.5$	0.6676	0.5887	0.8007	0.7666
$q=2$	0.6624	0.5847	0.7959	0.7609
$q=2.5$	0.6657	0.5803	0.7915	0.7560
$q=3$	0.6519	0.5754	0.7874	0.7516
$q=3.5$	0.6465	0.5700	0.7835	0.7835
$q=4$	0.6411	0.5641	0.7798	0.7798

III. LONG-RANGE CORRELATION PROPERTIES IN THE DNA SEQUENCES OF BACTERIA

To answer the question of whether or not there exist long-range correlation properties in the DNA sequences of bacteria, the average spectrum is studied first. In Buldyrev *et al.* [3], the subsequence length is 512, but in this paper, the subsequence length varies from 1024 to 16 384. Figure 1 shows the average spectrum of adenine from E. coli; other base density distributions along each bacteria show the same phenomenon.

Figure 1 shows that by increasing the subsequence length, the long-range correlation property is gradually more apparent in the lower-frequency part of the DNA spectrum. The middle part of the frequency is approximately white noise, but the lower-frequency part displays the long-range correlation property. The turning point between the middle and the lower part is about 200–400 BP. The long-range correlation property is gradually more clearly displayed with the in-

TABLE IV. Estimation results of various methods: M. genitalium.

	A	T	C	G
RS	0.7433	0.7813	0.7858	0.7632
Variance time	0.6512	0.8509	0.7850	0.8626
Scale 2	0.7167	0.7377	0.6461	0.6788
Scale 4	0.7190	0.7302	0.6343	0.6727
Scale 6	0.7166	0.7074	0.5896	0.6564
Scale 8	0.6936	0.6984	0.5760	0.6338
Scale 10	0.6371	0.6211	0.5042	0.6416
Higuchi	0.6432	0.8416	0.7477	0.8492
AVAS	0.6348	0.8522	0.7757	0.8657
$q=1.5$	0.6387	0.8457	0.7760	0.8597
$q=2$	0.6401	0.8406	0.7740	0.8527
$q=2.5$	0.6399	0.8363	0.7705	0.8448
$q=3$	0.6387	0.8324	0.6470	0.8362
$q=3.5$	0.6369	0.8288	0.7609	0.7609
$q=4$	0.6346	0.8255	0.7552	0.7552

TABLE V. Estimation results of various methods: M. pneumoniae.

	A	T	C	G
RS	0.8026	0.7897	0.7884	0.8277
Variance time	0.7529	0.7397	0.7303	0.7001
Scale 2	0.7526	0.7260	0.7233	0.7366
Scale 4	0.7541	0.7269	0.7233	0.7354
Scale 6	0.7385	0.7141	0.6931	0.7013
Scale 8	0.6881	0.6491	0.6300	0.6296
Scale 10	0.6590	0.5931	0.6041	0.5511
Higuchi	0.7338	0.7249	0.7170	0.7062
AVAS	0.7526	0.7194	0.7336	0.7005
$q=1.5$	0.7502	0.7277	0.7292	0.6971
$q=2$	0.7460	0.7324	0.7232	0.6929
$q=2.5$	0.7412	0.7346	0.7264	0.6883
$q=3$	0.7361	0.7352	0.6648	0.6834
$q=3.5$	0.7309	0.7346	0.7023	0.7023
$q=4$	0.7257	0.7333	0.6952	0.6952

TABLE VI. Estimation results of various methods: R. NGR234.

	A	T	C	G
RS	0.7542	0.7709	0.7765	0.7597
Variance time	0.6145	0.5905	0.6568	0.6147
Scale 2	0.7049	0.6773	0.6746	0.7030
Scale 4	0.7063	0.6774	0.6751	0.7018
Scale 6	0.6792	0.6451	0.6367	0.6795
Scale 8	0.6578	0.6167	0.6088	0.6266
Scale 10	0.5891	0.5212	0.5721	0.5773
Higuchi	0.6261	0.6250	0.6656	0.6076
AVAS	0.6120	0.5868	0.6574	0.6007
$q=1.5$	0.6078	0.5843	0.6508	0.6024
$q=2$	0.6025	0.5828	0.6449	0.6030
$q=2.5$	0.5968	0.5814	0.6393	0.6025
$q=3$	0.5918	0.5798	0.5217	0.6012
$q=3.5$	0.5852	0.5778	0.6288	0.6288
$q=4$	0.5793	0.5756	0.6239	0.6239

crease of subsequence length. Similar results can be derived for other bacteria genomes.

As discussed in Sec. I, Fourier analysis is a nonparametric method. The stationary property of the spectrum is hard to guarantee under a larger scale. In previous works, the DNA sequences are segmented into subsequences with identical lengths, and the average spectrum is calculated over these subsequences [3,5,11]. With this approach, the correlation properties revealed are limited by the subsequence length. The number of subsequences for the calculation of average spectrum is limited, and the scale range of asymptotic long-range correlation is hard to determine. In Buldyrev *et al.* [3], the power-law relationship in the low-frequency range of coding sequences is explained as the distortion by artifacts of the Fourier transformation method. But in this section, with Hurst index estimation, we will show that there does exist long-range correlation in the DNA sequences of bacteria, but in a scale longer than several hundred BP.

In order to investigate the long-range correlation properties in the DNA sequence more extensively, the statistical tools introduced in Sec. II are implemented. Hurst indexes of the seven eubacteria and one archaea are calculated, and their long-range correlation properties are discussed. When mapping the DNA sequences to numerical sequences, an *equal-symbol multiplication* method, which is recommended by Voss [11] and commented on by Maddox [6], is implemented. This method decomposes the nucleotide sequence into four 0-1 sequences, and calculates their Hurst indexes separately. So what we have investigated are only the long-range correlation properties of four single bases along the DNA molecule. This also points to some hints for the construction of the DNA molecule. Referring to Li [12], only three of them are independent, and when the approximate strand symmetry is considered, only one is independent, but for the sake of statistical accuracy, all of them are calculated in this paper.

For statistical convenience, we compressed the sequence by sampling it with a range of 100 BP. Therefore our focus is narrowed down to the long-range correlation property within the density distribution of the four single bases, as well as the difference between this distribution and an asymptotically second order self-similar process (such as a FGN process).

Figure 2 shows the 100 BP density distribution of adenine from E. coli, as well as a FGN process with the same length and a Hurst index equal to 0.7. From Fig. 2 it can be seen that the density distribution of nucleotide and the FGN process all perform as self-similar processes. The algorithms introduced in Sec. II are implemented in this paper. The least squares linear fitting in the log-log plots shows that the results are reliable. The corresponding calculated Hurst indexes are all larger than 0.5, which testifies to the existence of the long-range correlation structure within the nucleotide density distribution.

Figure 3 is the least squares fitting figure for the log-log plot of adenine from E. coli, with the Hurst indexes estimated by RS, variance-time, AVAS, and Higuchi method, respectively.

TABLE VII. Estimation results of various methods: Synechocystis.

	A	T	C	G
RS	0.7198	0.7158	0.7250	0.7141
Variance time	0.6458	0.6053	0.6050	0.6085
Scale 2	0.7269	0.7185	0.6946	0.6958
Scale 4	0.7278	0.7193	0.6956	0.6969
Scale 6	0.7230	0.7160	0.6870	0.6911
Scale 8	0.7142	0.7104	0.6767	0.6830
Scale 10	0.6749	0.6761	0.6307	0.6440
Higuchi	0.6503	0.6301	0.6113	0.6224
AVAS	0.6429	0.6097	0.6005	0.6050
$q=1.5$	0.6418	0.6051	0.5999	0.6041
$q=2$	0.6402	0.5990	0.5994	0.6030
$q=2.5$	0.6377	0.5940	0.5986	0.6016
$q=3$	0.6343	0.5879	0.5975	0.5999
$q=3.5$	0.6303	0.5816	0.5959	0.5959
$q=4$	0.6357	0.5752	0.5939	0.5939

TABLE VIII. Estimation results of various methods: M. jannaschii.

	A	T	C	G
RS	0.7642	0.7697	0.7875	0.7851
Variance time	0.6939	0.6066	0.7180	0.6309
Scale 2	0.7855	0.7936	0.8463	0.8496
Scale 4	0.7871	0.7956	0.8479	0.8515
Scale 6	0.7883	0.7963	0.8495	0.8534
Scale 8	0.7799	0.7834	0.8503	0.8492
Scale 10	0.7278	0.7361	0.7826	0.7899
Higuchi	0.6832	0.6166	0.7103	0.6385
AVAS	0.6818	0.5962	0.7010	0.6089
$q=1.5$	0.6842	0.5972	0.7063	0.6166
$q=2$	0.6847	0.5976	0.7088	0.6218
$q=2.5$	0.6837	0.5972	0.7094	0.6250
$q=3$	0.6815	0.5959	0.7082	0.6264
$q=3.5$	0.6782	0.5937	0.7056	0.7056
$q=4$	0.6741	0.5908	0.7015	0.7015

In Tables I–VIII, estimation results of the Hurst index for all eight bacteria sequences by various methods are presented. Because the DNA sequences are not strictly self-similar processes and the lengths of sequences are finite, the estimation results of various methods may slightly differ from each other. But so long as the Hurst indexes estimated are all greater than 0.5 and consistent, the long-range correlation property is verified.

Scale 2, scale 4, scale 6, scale 8, and scale 10 are the estimation results of the wavelet-based EM method under various scales. When the scale of wavelet changes from two to ten and the frequency range becomes narrower, a local (or high-frequency) feature can be detected. Under coarse scales of wavelet (scales 2, 4, and 6), which correspond to the lower part of the frequency, the estimation results remain relatively constant, which means that the density distribution of the nucleotide can be described asymptotically by the FGN process in the lower part of the frequency. Under fine scales of wavelet, which correspond to the higher part of the frequency, the decrease of the estimated Hurst index can be interpreted as in the smaller scale of the DNA sequence; it cannot be well modeled by the FGN process any more; this is also due to the local features of DNA sequences.

Apart from the estimation of the Hurst index, the AVAS method can also serve as the estimation of higher order statistics by means of the central moment of the m block process. This can serve to testify whether the long-range correlation in the DNA sequence is constructed by a multifractal or a monofractal process [23]. Under the condition that the estimation results by coarse grain are all linear with the growth of the order of statistics, if the estimation results remain relatively constant, then the signal is only a self-similar process; otherwise, the sequence is considered to present a multifractal phenomenon. In Tables I–VIII, the line labeled AVAS is the estimation result of the AVAS algorithm. $q=1.5$, $q=2$, $q=2.5$, $q=3$, $q=3.5$, $q=4$ are the estimation results of various orders of statistics. It can be seen from Tables I–VIII that the estimation results of the various orders of statistics are consistent, which means that the nucleotide density distribution along the DNA sequence can be asymptotically modeled by the FGN process.

In Tables I–VIII, the results of scale 8 and scale 10 from the wavelet-based EM method are decreased, therefore only the results from the RS, variance-time, Higuchi, and AVAS methods and wavelet scales 2–6 are used to calculate the mean and standard deviation of the Hurst index estimation. From Tables I–VIII, it can be seen that the Hurst index estimated by various methods for the four single base density distributions are all larger than 0.5. This means that there do exist long-range correlation properties along the DNA sequences. The mean and standard deviation of the estimation results are listed in Table IX. It can be seen that the mean values are between 0.65 and 0.8, with reason-

TABLE IX. Mean and standard deviation of estimation results from various methods.

Name of bacteria		A	T	C	G
E. coli	Mean($\hat{H}$)	0.7012	0.7026	0.7514	0.7624
	Std($\hat{H}$)	0.0321	0.0312	0.0312	0.0432
H. influenzae	Mean($\hat{H}$)	0.6730	0.6938	0.7396	0.7420
	Std($\hat{H}$)	0.0257	0.0209	0.0322	0.0364
H. pylori	Mean($\hat{H}$)	0.7138	0.6850	0.7582	0.7408
	Std($\hat{H}$)	0.0357	0.0788	0.0522	0.0371
M. genitalium	Mean($\hat{H}$)	0.6964	0.7859	0.7092	0.7498
	Std($\hat{H}$)	0.0395	0.0624	0.0831	0.0876
M. pneumoniae	Mean($\hat{H}$)	0.7553	0.7344	0.7299	0.7297
	Std($\hat{H}$)	0.0224	0.0256	0.0290	0.0461
R. NGR234	Mean($\hat{H}$)	0.6710	0.6533	0.6775	0.6667
	Std($\hat{H}$)	0.0549	0.0635	0.0456	0.0604
Synechocystis	Mean($\hat{H}$)	0.6909	0.6735	0.6599	0.6620
	Std($\hat{H}$)	0.0419	0.0553	0.0522	0.0476
M. jannaschii	Mean($\hat{H}$)	0.7406	0.7107	0.7801	0.7454
	Std($\hat{H}$)	0.0515	0.0981	0.0693	0.1144

ably small deviation, which further supports the reliability of these results.

IV. CONCLUSION

In this paper the long-range correlation properties in the whole genomes of bacteria are verified. The genomes of bacteria always have a length of more than 1 000 000 BP. The results of this paper suggest that the asymptotic long-range correlation property is one of the natural properties that DNA sequences possess, and is directly related to the structure and function of the whole DNA molecule.

There are still many discussions concerning the biological meaning and the origin of the long-range correlation properties in the DNA sequences. In [24–26], these properties are considered to be related to the construction of the higher order structure of the DNA molecule. In [11], Voss points out that the scale-independent correlations seem to offer the best compromise between efficient information transfer and immunity to errors on all scales. In [10,27], it is argued that neither the patchiness, the alternation of coding and noncoding regions, nor the repetitive sequences is able to fully explain the long-range correlation properties of DNA. The correlation structure of DNA sequences can also be considered to be a result of the evolution process [12]. A model called the expansion-modification model, which is introduced in [28,29], can emulate this process, and was shown to exhibit the same properties.

In this paper some parametric methods to estimate the Hurst index from the FGN point of view are introduced. Compared with nonparametric methods such as Fourier analysis, the performance of these methods is more stationary when applied to sequences with finite length, and the results yielded are more robust. With these methods, eight whole bacteria genomes are analyzed, and the long-range correlation properties in these sequences are verified. The lower limit of the length scale for this property is about 200–400 BP, and it can extend through the whole molecule of DNA. In longer scales, the results from higher order statistics are consistent, so the lower-frequency part can be well described asymptotically by a FGN process. This study can motivate the analysis of the DNA sequence structures under larger scales. Furthermore, the biological meaning behind the long-range correlation properties still calls for further investigation.

[1] R. Román-Roldán, P. B. Galván, and J. L. Oliver, Pattern Recogn. **29**, 1187 (1996).
[2] I. Amato, Science **257**, 74 (1992).
[3] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **51**, 5084 (1995).
[4] H. Herzel and I. Große, Phys. Rev. E **55**, 800 (1997).
[5] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[6] J. Maddox, Nature (London) **358**, 103 (1992).
[7] Sean Nee, Nature (London) **357**, 450 (1992).
[8] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).
[9] V. V. Prabhu and J.-M. Claverie, Nature (London) **359**, 782 (1992).
[10] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, and H. E. Stanley, Biophys. J. **72**, 866 (1997).
[11] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[12] W. Li, Computers Chem. **21**, 257 (1997).
[13] W. Li, Int. J. Bifurcation Chaos Appl. Sci. Eng. **2**, 137 (1992).
[14] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, IEEE Trans. Commun. **43(2/3/4)**, 1566 (1995).
[15] H. Cai, H. Chen, and Y. Li, J. China Inst. Commun. **18**, 59 (1997).
[16] W. E. Leland and D. V. Wilson (unpublished).
[17] N. Likhanov, B. Tsybakov, and N. D. Georganas (unpublished).
[18] D. R. Cox, in *Statistics; An Appraisal*, edited by H. A. David and H. T. David (The Iowa State University Press, Ames, IA, 1984), Vol. 55.
[19] M. S. Taqqu and V. Teverovsky, Fractals **3**, 785 (1995).
[20] M. E. Crovella and A. Bestavros, IEEE/ACM Trans. Networking **5**, 835 (1997).
[21] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Vilson, IEEE/ACM Trans. Networking **2**, 1 (1994).
[22] G. W. Wornell and A. V. Oppenheim, IEEE Trans. Signal Process. **40**, 611 (1992).
[23] R. H. Riedi, J. L. Vehel, ''Multifractal properties of TCP traffic: A numerical study,'' INRIA report, 1997.
[24] A. Grosberg, Y. Rabin, S. Havlin, and A. Neer, Europhys. Lett. **23**, 373 (1993).
[25] J. Widom, Proc. Natl. Acad. Sci. USA **89**, 1095 (1992).
[26] J. Yao, P. T. Lowary, and J. Widom, Proc. Natl. Acad. Sci. USA **90**, 9364 (1993).
[27] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).
[28] W. Li, Europhys. Lett. **10**, 395 (1989).
[29] W. Li, Phys. Rev. A **43**, 5240 (1991).