Hierarchical structure analysis describing abnormal base composition of genomes

Zhengqing Ouyang,[1,2] Jian-Kun Liu,[1] and Zhen-Su She[1,3,*]

[1]*State Key Lab for Turbulence and Complex Systems and Center for Theoretical Biology, Peking University, Beijing 100871, People's Republic of China*

[2]*School of Physics, Peking University, Beijing 100871, People's Republic of China*

[3]*Department of Mathematics, University of California, Los Angeles, Los Angeles, California 90095, USA*

(Received 8 December 2004; published 14 October 2005)

Abnormal base compositional patterns of genomic DNA sequences are studied in the framework of a hierarchical structure (HS) model originally proposed for the study of fully developed turbulence [She and Lévêque, Phys. Rev. Lett. **72**, 336 (1994)]. The HS similarity law is verified over scales between 10^3bp and 10^5bp, and the HS parameter β is proposed to describe the degree of heterogeneity in the base composition patterns. More than one hundred bacteria, archaea, virus, yeast, and human genome sequences have been analyzed and the results show that the HS analysis efficiently captures abnormal base composition patterns, and the parameter β is a characteristic measure of the genome. Detailed examination of the values of β reveals an intriguing link to the evolutionary events of genetic material transfer. Finally, a sequence complexity (S) measure is proposed to characterize gradual increase of organizational complexity of the genome during the evolution. The present study raises several interesting issues in the evolutionary history of genomes.

DOI: 10.1103/PhysRevE.72.041915 PACS number(s): 87.14.Gg, 87.15.Aa, 87.15.Cc

I. INTRODUCTION

The DNA sequence of a complete genome of an organism contains the information not only for making all the proteins (genes) necessary for the organism, but also for assembling them to form the organism in a specific time order with specific three-dimensional patterns. While small-scale (from several to hundreds base pairs) patterns of the nucleotide arrangement are certainly important for determining its coding or noncoding nature and some regulatory biological functions (e.g., binding site or splicing site signal) [1], more large-scale variation across several thousands base pairs or longer may be related to higher level biological functions such as controlling networks of genes which are likely important indices in evolution [2]. It is important to develop tools for analyzing the structures of these patterns for a large set of available sequence data and to use it as a laboratory for quantitative tests of biological laws governing the evolution [3].

There has been considerable efforts in studying the statistical property of nucleotide composition pattern [4]. The algorithms for DNA sequence alignment and similarity search have been developed for the study of phylogeny and evolution of many biological species [5]. Methods developed in nonlinear analysis and information theory were introduced to characterize coding and noncoding DNA sequences [6,7]. Other methods derived from statistical physics [8] including spectrum analysis [9–12], wavelet analysis [13], etc., have also been proposed to measure the correlation between nucleotides over long distances along one-dimensional DNA chain, although the so-called long-range correlations in DNA sequences remain a subject of debate [14,15].

Among the previous studies of scaling behaviors of DNA sequence, the most well-known result is the $1/f$-like power law at moderate length scales (typically 10–1000 bp) [9,10]. Less effort has been made to examine large-scale correlations, partially due to the lack of very long sequences in the past decades. Recently, large-scale structure of DNA sequence at the complete genomic level has been studied [16]. These structures are believed to be related to the evolution of the organism, and such global information is difficult to analyze with traditional methods. For example,"base-base" correlation at long distances does not reveal clear biological meaning [14].

Complex heterogeneity of the base composition was studied in the context of multi-scale correlations by Bernaola *et al.* [17] who have proposed the concept of "domains within domains." The method uses the Jensen-Shannon entropic divergence based segmentation to characterize the hierarchical organization of heterogeneous base composition. The local structure description has been successfully applied to the analysis of the complexity of sequence composition [18], detection of borders between coding and non-coding regions [19,20], partition of homogeneous isochore structures in eukaryotic genomes [21], etc.

The present study is also a multi-scale approach to the problems of base composition heterogeneity. But we focus on the study of the global structure of the genome and on a comparative genomic study. Earlier, we have reported a hierarchical structure (HS) description of multi-scale structures of DNA sequences [22,23], which has been inspired by an earlier HS model for hydrodynamic turbulence [24]. The analysis begins by constructing a nucleotide composition fluctuation field and by calculating the scale variation of the probability density functions (PDF) in the HS scaling model framework. The range of scale in the study is between 10^3 bp and 10^5 bp, and the entire genome forms the ensemble for calculate the average. Our analysis reveals that the base composition variations along genomes are far from random, but present complex self-organized, intermittent structures, for which the HS model gives an excellent description. In the

*Electronic address: she@pku.edu.cn

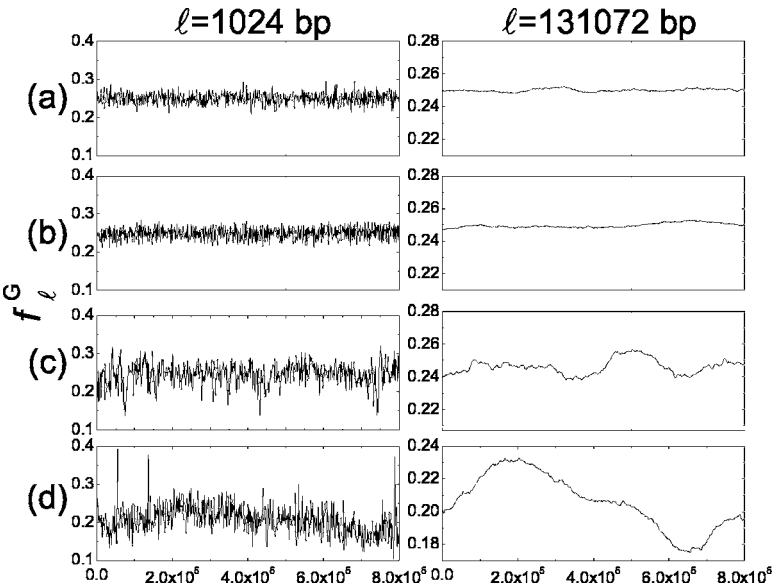

FIG. 1. Nucleotide guanine (G) composition variation f_ℓ^G of (a) random, (b) simulation, (c) ecoli, and (d) Hsap4 (for abbreviations, see the text). Representative local densities are calculated over a window of range of $\ell_{\min}=2^{10}$ bp (the left) and $\ell_{\max}=2^{17}$ bp (the right), respectively. The sliding window moves at a step of length $\Delta=1024$ bp. Note the intensive fluctuations of natural sequences away from artificial ones.

present work, we report detailed measurement of a HS parameter β for over one hundred sequences of three kingdoms (eukaryotes with *Homo sapiens* and *Saccharomyces cerevisiae*; prokaryotes with archaea and bacteria; viruses). In particular, we use the HS parameter to analyze abnormal base composition patterns and sequence organizational complexity of the genomes.

The paper is organized as follows: A multi-scale variable f_ℓ concerning base composition fluctuations of genome sequences is introduced in Sec. II. We present briefly measurements of scaling property with special emphasis on the HS model in Sec. III. Section IV is devoted to a detailed HS analysis of various genomes. Section V offers a summary and some additional discussion.

II. BASE COMPOSITION FLUCTUATIONS

A single-stranded DNA chain is a symbolic sequence $\{n_i\}(i=1,2,\ldots,L)$ of length L comprised of four base letters (A, C, G and T). Many transformation methods have been proposed to represent numerical properties of DNA sequences; for instance, "DNA walk" [8] constructs a numerical sequence $\{u_i^G\}$ with a rule (e.g., if $n_i=G$ then $u_i^G=1$, and otherwise $u_i^G=0$), and then uses a running sum $y^G(n)=\Sigma_{i=1}^n[u_i^G-(1-u_i^G)]$ to present graphically a walk in the one-dimensional landscape of the original DNA sequence. Below, we consider the very simple window-averaged composition defined by

$$f_\ell^G(i)=\frac{1}{\ell}\sum_{k=i}^{i+\ell-1}u_k^G, \tag{2.1}$$

where i is the location of the first base within the window. The definition can be extended to any other single nucleotide (A, C, or T) or their degeneracy (R, Y, etc.), or to any dinucleotide molecules (AT, AG, etc.), even to any specific word. By sliding the window along the DNA sequence with a moving step Δ and by changing the window size ℓ, we obtain a series of fluctuation fields $f_\ell(i)$ which form a multiscale description of the base composition fluctuation structures. This multi-scale moving-window analysis helps to capture base composition heterogeneity at varies detail, and to reveal the correlations between different scales. It avoids the drawback of using a single subjective sliding window which often reaches cursory conclusions [25]. When ℓ is taken to be the length of the entire genome, f_ℓ becomes simply the overall mean base composition. The set of the multi-scale variable f_ℓ are similar to the set of locally averaged energy dissipation rate ϵ_ℓ in a turbulence field, which has played an important role in characterizing turbulent structures. We propose to use it to characterize genomic fluctuation structures.

It is instructive to visualize f_ℓ for various sequences. Figure 1 displays a segment of 0.8 million bp of base composition of guanine (G), f_ℓ^G, at two scales $2^{10}(\approx 10^3\text{bps})$ and $2^{17}(\approx 10^5$ bps) for four different sequences: an independent identical distributed (*i.i.d.*) random sequence with 50% A+T content (Random), a simulated genome sequence by a so-called minimal model (Simulation) [26], the *E. coli* genome sequence (Ecoli) and the *H. sapiens* sequence of chromosome 4 contig 8 (Hsap4), respectively. Both the random sequence and simulated sequence by the minimal model present essentially a white noise signal at the small scale and a flat plateau at the large scale with little variation. On the other hand, the Ecoli and Hsap4 sequences clearly present many intermittent bursts with local abnormal composition. Even at large scale of 10^5 bps, base composition variations are still remarkable, which is more pronounced for Hsap4 than for Ecoli. We believe that this is the most distinctive features of true genomes from random sequences and it deserves detailed study by theorists of the evolution. The present work focuses on one specific aspect of the structure, namely a relative scaling that characterizes how the fluctuation varies from small to large scales.

To quantify the statistical properties of the fluctuations across multiple scales, it is customary to compute the probability density functions (PDF) of f_ℓ, $P(f_\ell)$. At small window size, the local base composition is a discrete random variable. For easy comparison and with neglectable difference,

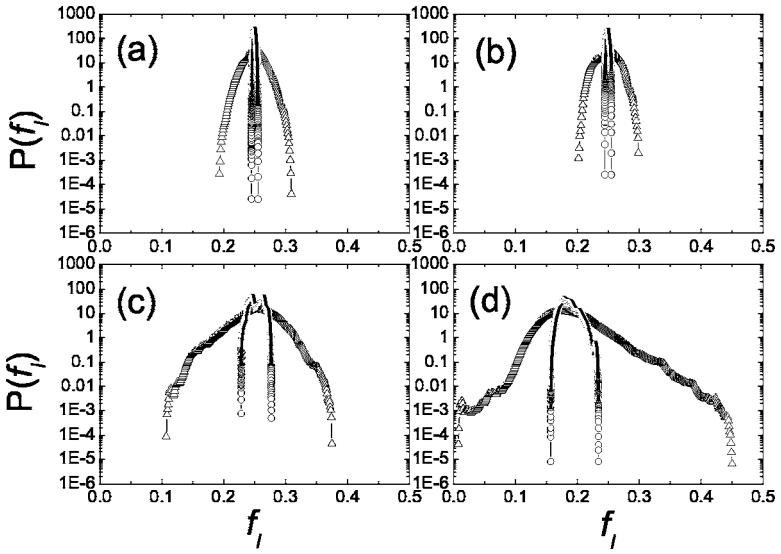

FIG. 2. Typical PDFs of guanine density (G) f_ℓ of (a) random, (b) simulation, (c) ecoli, and (d) Hsap4 at two scales $\ell_{\min}=2^{10}$ (triangle) and $\ell_{\max}=2^{17}$ (circle). Note that with ℓ decreasing, the right wings of PDFs progress further, indicating the appearance of high intensity fluctuations in f_ℓ which can be captured by the HS analysis.

we choose to treat it like a continuous random variable using linear interpolation. PDFs at two scales for the four sequences in Fig. 1 are shown in Fig. 2 where the two boundary scales of our scaling studies below, namely $\ell_{\min}=2^{10}$ and $\ell_{\max}=2^{17}$, are chosen. As expected, all PDFs exhibit a much narrower range of variations at large scales (inner lines) compared to that at smaller scales. It is quite clear that both the random sequence and the simulated sequence are close to Gaussian distribution with a mean of 0.25, as expected from the law of large numbers, which is asymptotic to the binomial distribution (also noticed by [27]). On the other hand, the natural DNA sequences show clear departure from a binomial distribution for small window size, and present several abnormal features compared to Gaussian distribution at large window size, as can be seen from Fig. 2(c) and 2(d). The variation of $P(f_\ell)$ with scale ℓ can be seen from the propagation of the tails of PDFs: as ℓ decreases, highly intense fluctuation events become more probable and the tails extend longer. A noticeable longer tail is observed from the right wings of Hsap4 PDFs and an opposite situation occurs for the Ecoli PDFs. We believe that these are remarkable features to be understood from an evolutionary point of view. Some aspects of this variation are described by our quantitative HS analysis below.

III. METHODS

The HS model was originally proposed by She and Lévêque [24] to describe inertial-range scaling for fluctuations in a turbulent fluid. It postulates a new similarity relation between structures of increasing intensities of successive moment-orders p as a generalization of the Kolmogorov's complete-scale-similarity [28]. The new similarity hypothesis has been successfully tested in various systems including the Couette-Taylor flow [29], flows in rapidly rotating disk [30], turbulent climate variations [31], magnetohydrodynamic turbulence in astrophysical environment [32], and in many other complex systems such as the diffusion-limited aggregates [33], the luminosity fields of natural image [34] and chemical reaction patterns [35]. Preliminary analysis of the base density fluctuations at moderate length scales along genomes [22,23] has given also an encouraging sign that has motivated the present work.

Denote by $S_p(\ell)$ the pth order moment of the fluctuation f_ℓ:

$$S_p(\ell)=\langle f_\ell^p\rangle=\int f_\ell^p P(f_\ell)df_\ell, \tag{3.1}$$

the HS model introduces a hierarchy of functions for successive fluctuation intensities:

$$\mu_p(\ell)=\frac{S_{p+1}(\ell)}{S_p(\ell)}=\frac{\int f_\ell^{p+1}P(f_\ell)df_\ell}{\int f_\ell^p P(f_\ell)df_\ell}=\int f_\ell Q_p(f_\ell)df_\ell, \tag{3.2}$$

where $Q_p(f_\ell)=f_\ell^p P(f_\ell)/\int f_\ell^p P(f_\ell)df_\ell$ is a weighted PDF for which $\mu_p(\ell)$ is the mathematical expectation. Such a hierarchy, $\mu_p(\ell)(p=0,\ldots,\infty)$, covers the mean fluctuation intensity μ_0, and a series of increasing hierarchical intensities with increasing order p, and finally approaches to the highest intensity in the whole ensemble, i.e., $\mu_\infty(\ell)=\lim_{p\to\infty}\mu_p(\ell)$. Therefore, one can associate each $\mu_p(\ell)(p=0,\ldots,\infty)$ with a structure of a certain intensity which varies from the mean intensity to the extreme intensity. The ensemble of the scaling for all p constitutes a description which is related to an overall self-organized state of the fluctuation structures that is much richer than a simple scaling of the power spectrum or other correlation measures. This is the most distinctive feature of the present framework.

The so-called HS similarity law among various intensities reads

$$\frac{\mu_{p+1}(\ell)}{\mu_1(\ell)}=\frac{\alpha_p}{\alpha_0}\left(\frac{\mu_p(\ell)}{\mu_0(\ell)}\right)^\beta, \tag{3.3}$$

where the exponent β is a constant and α_p is a coefficient independent of ℓ. The validation of the HS similarity relation

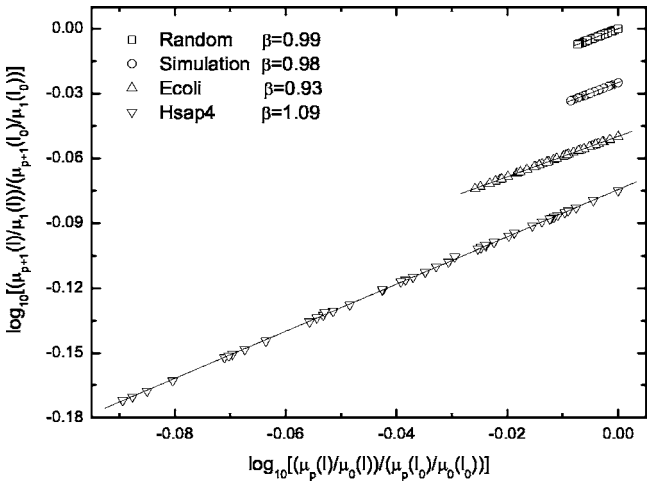

FIG. 3. The β test of guanine density fluctuation for random, simulation, ecoli, and Hsap4 at the range of $2^{10} \leqslant \ell \leqslant 2^{17}$ and $0 \leqslant p \leqslant 8$. A straight line indicates the validity of the HS similarity. The slope β is estimated by a least square fitting. For clarity, the second, third, and fourth set of data points are displaced vertically up by a suitable amount.

Eq. (3.3) can be directly tested by a log-log plot of $\mu_{p+1}(\ell)/\mu_1(\ell)$ vs $\mu_p(\ell)/\mu_0(\ell)$ (which is called a β-test [29,36]). The HS similarity is verified with the linearity in the plot, and the HS parameter β can be obtained by measuring the slope using a least square fitting. Technically speaking, this completes the HS analysis of a given set of fluctuation data.

As an example, the results of the β-test for the four kinds of sequences (the random, the simulation, Ecoli and Hsap4) are shown in Fig. 3, where the scale range is between 10^3 bp and 10^5 bp and $\ell_0 = 1024$. Remarkable linearity is observed, so the HS similarity is accurately verified for the multi-scale base composition fluctuation data. The values of β_G obtained are very close to 1 (0.99 ± 0.000 and 0.98 ± 0.000) for the first two kinds of sequences, and are noticeably different from 1 (0.93 ± 0.001 for Ecoli and 1.09 ± 0.002 for Hsap4) for the two true genomic sequences, respectively. It is believed [28] that the HS similarity is an indication of the self-organization of the ensemble of the fluctuation events. When $\beta = 1$, it is the situation of homogeneity (or with no intermittency). The random sequence and the simulated sequence are indeed homogeneous, and their HS analysis is consistent with this prediction. The situation where $\beta \neq 1$ reveals the presence of heterogeneous correlations in the fluctuation structures. All our studies below focus on the quantitative study of the heterogeneity described by the value of β.

IV. HS ANALYSIS FOR GENOMIC DATA

We have conducted the HS analysis for over one hundred organisms spreading over three kingdoms of species: eukaryote with *Homo sapiens* (24 chromosomes) and *Saccharomyces cerevisiae* (16 chromosomes); prokaryote with 16 archaea complete genomes and 124 bacteria complete genomes/chromosomes; and 67 viruses complete genomes. Details of these genomes data can be found in [37]. The 67 viruses sequences studied are long enough (beyond 2×2^{17} bp) for statistic analysis. Scales for analyzing the nucleotide fluctuations is consistent with the above, from 10^3 to 10^5bp. For each sequence, four kinds of bases (adenine (A), cytosine (C), guanine (G) and thymine (T) as four fundamental "words" in DNA sequences are analyzed independently. All sequences pass the β-test (most of which have a correlation coefficient above 0.9995) for four different bases, thus the HS parameter β can be measured with a high accuracy. The measured β of four base composition fields of viruses, bacteria, archaea, *S. cerevisiae* (yeast) and *H. sapiens* (human) are documented in [37].

A. Characterizing genome's heterogeneity with β

According to our calculation [37], many β values significantly deviate from unity, indicating the existence of abnormal fluctuation structures in base compositional patterns compared to a random sequence. We believe that this heterogeneity has biological interest, and hence β carries biological information. In this section, we report a few typical cases where the meaning of β can be clarified.

A first example to be shown here is the bacterium *Xylella fastidiosa* (*B-Xfa*) that causes a range of economically important plant diseases [38]. The length of the genome is 2,679,305 bp and the overall G+C content is 52.7%. It comprises 2904 predicted coding regions and at least 83 genes are bacteriophage-derived and include virulence-associated genes from other bacteria, providing direct evidence of phage-mediated horizontal gene transfer (HGT), which will be discussed in more detail in the next subsection. Figure 4 shows the local base compositional pattern of the four nucleotides of *B-Xfa* (top) and the results of the β test (bottom). The (top) graphs indicate that A has a similar large-scale variation pattern with C, and T with G. However, both A and T contain sharp downward sparks corresponding to small-scale structures of low A/T segments. On the other hand, both C and G contain sharp upward sparks corresponding to small-scale structures of rich C/G segments. The β test clearly indicates a symmetry between A and T with $\beta_A \approx \beta_T < 1$, and a symmetry between C and G with $\beta_C \approx \beta_G > 1$. This suggests that a β less than unity is associated with the appearance of downward sparks in the composition fluctuation field.

In Fig. 5, we show the guanine composition fluctuation patterns and the results of the β-test for three typical cyanobacteria which is a general term for the bacteria capable of oxgenic photosynthesis: *Thermosynechococcus elongatus* BP-1 (BP-1), *Synechocystis* PCC6803 (PCC6803) and *Synechococcus* sp. WH8102 (WH8102). BP-1 and PCC6803 are freshwater cyanobacteria, and BP-1 has notable thermophilic character and is branched very close to the original cyanobacteria [39]. On the other hand, WH8102 is a kind of marine cyanobacterium [40]. Figure 5 clearly shows that the three organism have different fluctuating composition patterns and that WH8102 has a notable small β and extensive low-guanine regions comparing to the other two. The values of β are directly related to the amount of sharp sparks, which

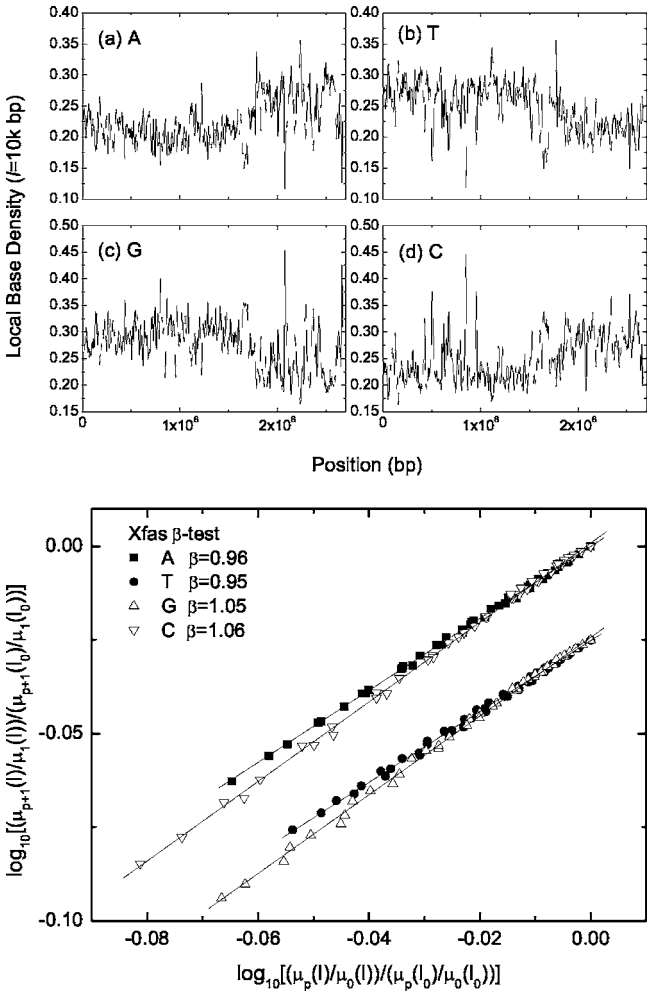

FIG. 4. (Top) Local density fluctuations with a scaling window of $\ell = 10^4$ bp of four bases: (a) A, (b) T, (c) G, and (d) C for *B-Xfa*. The sliding window moves at a step of length $\Delta = 10^3$ bp. (Bottom) β test of base density fluctuation for *B-Xfa* at the range of $2^{10} \leqslant \ell \leqslant 2^{17}$ and $0 \leqslant p \leqslant 8$. A straight line indicates the validity of the HS similarity. Note that the parity rule $\beta_A \approx \beta_T$ and $\beta_C \approx \beta_G$ is obeyed. For clarity, the second, third, and fourth set of data points are displaced vertically up by a suitable amount.

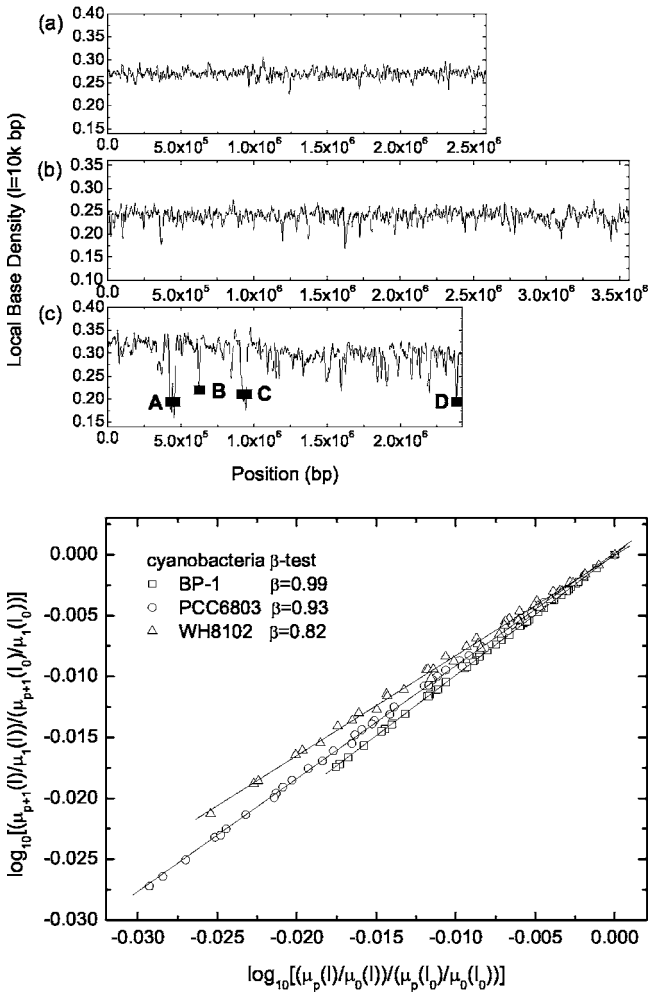

FIG. 5. (Top) Local density fluctuations with a scaling window of $\ell = 10^4$ bp of G for (a) BP-1, (b) PCC6803, and (c) WH8102. The sliding window moves at a step of length $\Delta = 10^3$bp. (Bottom) β test for BP-1, PCC6803, and WH8102. The test range is $2^{10} \leqslant \ell \leqslant 2^{17}$ and $0 \leqslant p \leqslant 8$. A straight line indicates the validity of the HS similarity. Note the smaller value of β for WH8102 corresponds to many low-concentration regions of G, some of which (like those marked by A, B, C, D) have been identified to be related to the acquisition of new functions for WH8102 (see the text).

are believed to be related to the HGT as an important evolutionary event [41]. Indeed, the extensive low G+C segments of the genome of WH8102 have been comprehensively identified as obtained by HGT [40], contributing to its functionalization of marine living, e.g., A and B of Fig. 5 of WH8102 (Top) correspond to the acquisition of modification of cell envelope, and C and D correspond to motility and cyanate usage acquisition, respectively.

In summary, we present here two examples indicating that the HS analysis is effective in revealing these abnormal compositional patterns in a genome, which seem to be the result of the evolution. This study deserves to be carried out further and to be applied to other organisms.

B. Anticorrelation with the mean composition

The results above suggest that $\beta > 1$ (or $\beta < 1$) implies more upward (or downward) sparks in the composition fluctuations. It is natural to ask what the nature of those sparks is. Imaging the following scenario of the genome evolution: a number of species carrying different composition sets of DNA exchange pieces of their DNA among themselves during the evolution. Assume there are two organisms O_1 and O_2 for which there is a flux of DNA from the genome of O_1 to O_2 (the process does not have to be symmetric; see discussion below). Assume also that O_1 is C/G rich and O_2 is A/T rich in their genome. The transfer of pieces of nucleotides from O_1 to O_2 will introduce a strong upward spark in the C/G composition field of O_2 and a strong downward spark in the A/T composition field of O_2. As the evolution proceeds, the accumulation of more downward (or upward) sparks in the A/T (or C/G) composition fields of O_2 will yield $\beta^{A/T} < 1$ and $\beta^{C/G} > 1$ for O_2 which is originally A/T rich, according to the results reported in the last subsection.

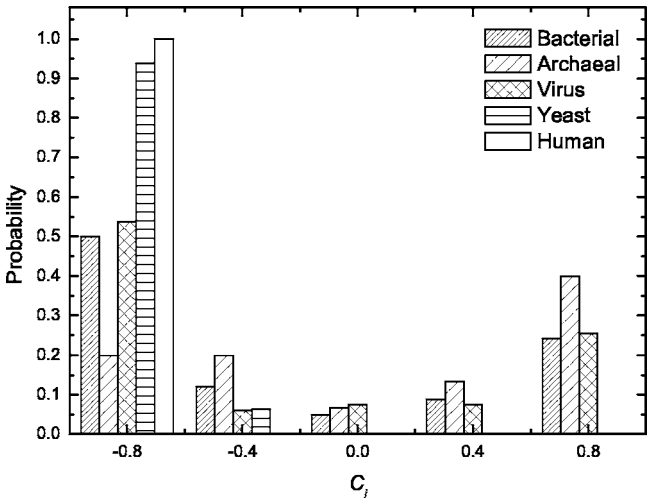

FIG. 6. Probability plot of the correlation coefficient C_i of measured β and mean base composition for sequences of various categories.

In other words, the mean composition of O_2 is anticorrelated with its β value among the four nucleotides.

A more general conclusion holds: if the mean composition of a nucleotide α of a genome is above (or below) the average of the ensemble with which the genome exchanges this nucleotide material, with higher probability the genome will get segments of DNA with higher (or lower) density of α, then the value of its β^α is typically above (or below) unity, which anticorrelates with the mean composition of α of this genome.

We have attempted to test the above hypothesis with the computed β_i^α and mean composition f_i^α for many species and chromosomes. Let us define a correlation coefficient between measured β_i and the genomewise mean composition f_i for each sequence (organism or chromosome) i as

$$C_i = \frac{\sum_\alpha (\beta_i^\alpha - \beta_i^m)(f_i^\alpha - f_i^m)}{\sqrt{\sum_\alpha (\beta_i^\alpha - \beta_i^m)^2 \sum_\alpha (f_i^\alpha - f_i^m)^2}}, \tag{4.1}$$

where α denotes nucleotides, and β_i^m and f_i^m are averages over the four nucleotides: $\beta_i^m = \langle \beta_i^\alpha \rangle$, $\mu_i^m = \langle \mu_i^\alpha \rangle$, $(\alpha = \{A, T, G, C\})$. In an ideal situation where $\beta^A \approx \beta^T = 1 - \epsilon$ and $\beta^C \approx \beta^G = 1 + \epsilon$, and $f^A \approx f^T = 1/4 + \delta$ and $f^C \approx f^G = 1/4 - \delta$, the correlation coefficient $C_i = -\epsilon\delta/|\epsilon\delta|$, which is -1 if ϵ and δ have the same sign.

We have calculated C_i for all genome sequences of five different categories and the results are summarized in a probability plot in Fig. 6 [37], where the probability to find C_i selected from fives categories are calculated separately. It is clear that human genome (with 24 chromosomes) and yeast genome (with 16 chromosomes) present an uniform ensemble with a correlation coefficient close to the ideal situation (-1). We consider this as an indication that a process of genetic material transfer (GMT) like what we proposed above has taken place between two heterogeneous DNA and this process is responsible for abnormal composition patterns as well as for deviations of β from unity. On the other hand, archaea seems to present certain correlation in the opposite direction, the nature of which deserves further studies. The bacteria and virus genomes display complex variations with a weak overall anticorrelation, but a substantial portion of them (over 50%) exhibits strong anticorrelations. A comparative study of this subgroup versus others may reveal the mechanism behind the DNA transfer events postulated above.

In summary, the result above indicates that β reflects the presence of abnormal base composition patterns (sharp up and down sparks) which, when combined with other measures such as the mean composition, may lead to hypotheses concerning the evolutionary history of the genomes. We do not intend to go into more details further, but leave the biological aspects of the problem to future studies.

C. Sequence organizational complexity

One of the most intriguing results discovered in [23] is the systematic change of β with evolutionary categories. When β_A and β_T are plotted for several dozens of prokaryotic and eukaryotic genomes, a tendency of decreasing β from unity was observed, which is interpreted as a sign of increasing organizational complexity. The results reported in the present paper enable us to understand this finding and to develop a systematic measure of the sequence complexity (S) as we describe below.

It is generally believed that Darwinian evolution leads to increasing organizational complexity of organisms and hence of genomes. One has not yet found a quantitative measure for such a global organizational complexity at the whole genome level. Several evolutionary forces may be identified. First, random point mutations and other insertion/deletion events are primary mechanisms acting to increase the complexity of the noncoding regions of the genome [42]. Reference [11] reported that noncoding sequences have long correlation structures while coding sequences are absent, and concluded that the complexity of noncoding sequences increases with evolution. Second, duplication of genes with post-evolutionary variations may be responsible for gene clusters and hence for coherent variations over large scales. This seems to be a promising mechanism for the formation of large-scale structures in base compositional patterns. Third, genetic material transfer (GMT) is a mechanism by which species acquire new genotypes and thus become more complex. Horizontal gene transfer (HGT) is identified as one of the major forces in prokaryotic genome evolution [45]. There is a growing body of evidence that demonstrates that transposition has been major players in eukaryotic genome evolution [43]. Other similar mechanisms causing eukaryotic genome heterogeneity include sexual process and viral infection, etc. Since the HS analysis seems to be effective for capturing abnormal composition patterns which are the potential signature of HGT, and more generally of GMT between genomes during the evolution, it is natural to develop a complexity measure based on the present HS analysis.

The key basis for a HS-based complexity measure is the fact that the deviation of β from unity may be related to

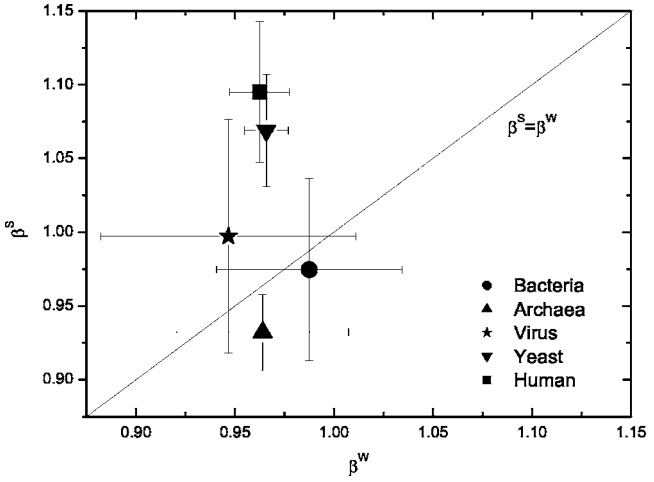

FIG. 7. The mean HS measure β^s versus β^w for 124 bacterial genomes, 16 archaeal genomes, 16 yeast chromosomes, 23 human chromosomes, and 67 virus genomes. Note the cluster property of the three kingdoms and the diversity of viruses.

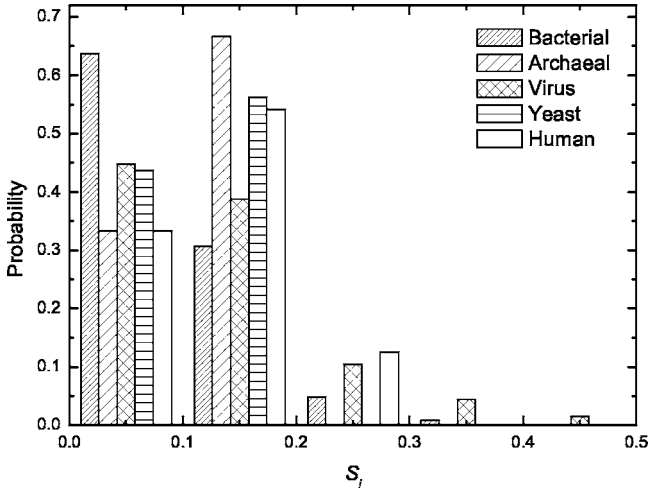

FIG. 8. Histogram plot of the HS-based S measure for organisms of various species.

certain evolutionary events which introduce heterogeneity in the base composition pattern. We then postulate that the degree of the heterogeneity, and hence the degree of the sequence complexity of the genome, is proportional to the departure of β from unity. This would be qualitatively correct if the abnormal composition events accumulates during the evolution for the set of organisms living in similar environment, and so does the organism's global complexity by the Law of Increasing Complexity [44]. Here, we propose to use the strand symmetry to reduce the number of the HS parameter β for each organism to two: $\beta^S=(\beta^A+\beta^T)/2$ and $\beta^W=(\beta^C+\beta^G)/2$. A plot of the mean β^S and β^W is shown in Fig. 7 for various organisms. It is interesting to observe that the deviation property of β^S of the organisms of various categories differs from that of β^W, indicating that they carry complementary information about the evolution. Thus, we define a combined measure of sequence complexity (S) for each sequence (organism or chromosome) i as

$$S_i = \sum_{\alpha=\{S,W\}} |\beta_i^\alpha - 1|. \tag{4.2}$$

It is interesting to note that, mathematically speaking, the deviation of β from unity is related to the so-called "singularity" index (γ) of the most intense fluctuations [$\gamma=1/(1-\beta)$] in the HS model. For a random sequence, the S measure is nearly zero; so the random sequence set up the background. The deviation of β from unity is considered as a measure of the amount of sequence complexity related to large-scale organization and to the evolution. We have calculated the S for all kinds of genome sequences that have passed the β-test and the results are summarized by a probability plot in Fig. 8 [37]. We have also calculated the mean S measure for various species, the result is presented in Table I.

Table I shows that the minimal model proposed earlier generates minimal sequence complexity related to large-scale organization, because of the lack of the mechanism for introducing abnormal composition patterns, an essential feature of the true genomes. As shown by the S values in Fig. 8 and in Table I, human has the highest S. Interestingly, we find that virus has a very high mean value and a substantial deviation of S; we interpret this as the evidence of diversity of virus genomes: some exchanging DNA dynamically with the host genomes, others are at stasis. We note also that on average, archaea is more complex than bacteria. Yeast has a S higher than bacteria while lower than archaea. We speculate that yeast is a kind of unicellular eukaryotic organism undergoing relatively high pressure of natural selection which has reduced the rate of growth of genome complexity. On the other hand, recent studies [46] suggest that eukaryotic genomes originate from a fusion of archaea and becteria genomes. This is an interesting phenomenon which needs further study.

Previous studies using spectrum exponents to characterize sequence complexity have led to controversial results [10,11]. While Voss [10] found the spectrum exponents decrease with evolution, Buldyrev *et al.* [11] found the opposite. These incompatible findings are due to the equivocal meaning of spectrum exponents. The present notion of S and its quantitative measure in terms of the HS parameter β gives a biological more meaningful solution. It establishes the ground for our previous finding in [23].

TABLE I. Mean and rms deviation of the correlation coefficient C_i and the sequence complexity measure S_i for different kinds of sequences.

Sequences	$\langle C_i \rangle$	C_i^{rms}	$\langle S_i \rangle$	S_i^{rms}
Random			0.014	
Simulation			0.026	
Bacterial	−0.219	0.778	0.093	0.057
Archaeal	0.133	0.742	0.117	0.034
Virus	−0.221	0.762	0.129	0.085
Yeast	−0.775	0.158	0.106	0.036
Human	−0.800	0.176	0.129	0.052

V. CONCLUSIONS

We have explored the application of the genomic HS model [23] to the study of abnormal base composition structure of over one hundred genome sequences in three kingdoms, and solidly established the evidence that the HS analysis effectively characterizes abnormal base composition patterns of genome. Here, we have reported a number of interesting facts. First, we show that high quality ESS scaling is found for most organisms within the range of scale 10^3 bp $\leqslant \ell \leqslant 10^5$ bp, and the whole sets of scaling exponents are reliably described by the HS model. Second, we demonstrate that the *i.i.d.* random sequence and a simulated genome by short segmental duplications have a HS parameter β close to unity, and the departure from unity of β reveals the existence of abnormal base composition patterns (e.g., upward or downward sharp sparks in base composition fluctuation field). Thirdly, we argue that extensive GMT should occur during genome evolution. The anticorrelation of β and average base composition is an evidence of this. The human and yeast genomes have uniform and strongest anticorrelation, indicating a continuing process of GMT in their evolutionary history. The majority of GMTs in these organisms may be transposition, which are to be confirmed by the future studies. On the other hand, virus, bacteria and archaea show a bimodal distributions for the β-average-composition correlation: some members display strong anticorrelations and some others are on the reverse. It reveals the across-species inhomogeneity for these kinds of organisms. Finally, we define a measure of sequence complexity (S) based on the departure of β from unity to characterize the heterogeneity of genome base composition. S is a measure derived from the scaling, and is shown to have a comprehensive interpretation when applied to various species. We show that human has the highest S, consistent with its multicellular nature; virus has large fluctuations on S, owing to their diversity. A notable conclusion is that a simulation model of genome sequence based on segmental duplication [26] is unable to exhibit abnormal base fluctuations, so it needs to be enriched.

The proposed HS analysis is basically a scaling analysis. The success of its application to analyzing abnormal base composition patterns of genomic is encouraging to the basic idea that scaling a kind of global measure of the system. On the other hand, the HS analysis is a sophisticated scaling analysis, because it involves a complete set of scaling instead of one or two exponents. It is therefore free of obscure meaning of earlier scaling analysis of long-range correlations. In different words, the HS analysis avoids to discuss long-range correlation in general, but is able to reveal a special kind of long-range correlations by abnormal base fluctuation structures, which is believed to be evolutionary relevant. More interestingly, β_i^α for individual base α and genome (chromosome) i can become a characteristic measure of each genome (chromosome).

A number of directions can be pursued further. It is interesting to test computationally the mechanism suggested in this paper for the formation of abnormal base composition during the evolution. The computational study can even be carried out for a community of microbial organisms [47]. Another interesting issue is the origin and evolution of eukaryotes. Genome fusion is probably the largest type of GMT and is speculated to be the origin of eukaryotes [46]. When such fusion takes place, abnormal base composition patterns will arise, which can now be quantitatively characterized in the HS framework. The measure of sequence complexity can also be pursued. Although the C_i of yeast is consistent with its eukaryotic nature, it is surprising that its S is lower than that of archaea, a distinct pattern away from human. Several evolutionary mechanisms may be responsible for this. One possibility is that many GMTs in yeast have undergone strong negative selection and lost during evolution. Another possibility is the recent HGT of prokaryotes after the genome fusion which increase the complexity of prokaryotes. All these studies need to incorporate the HS analysis as the basic quantitative tool.

ACKNOWLEDGMENTS

We have benefited from useful discussions with many people at the LTCS and CTB of Peking University, especially Dr. Huaiqiu Zhu. This work was supported by NSFC No. 10225210 and by 973 Project founded by MOST of China.

[1] D. Boffelli, M. A. Nobrega, and E. M. Rubin, Nat. Rev. Genet. **5**, 456 (2004).
[2] N. Banerjee and M. Q. Zhang, Curr. Opin. Microbiol. **5**, 313 (2002).
[3] B. Dujon *et al.*, Nature (London) **430**, 35 (2004).
[4] B. V. Reddy and M. W. Pandit, J. Biomol. Struct. Dyn. **12**, 785 (1995).
[5] A. M. Sugden, B. R. Jasny, E. Culotta, and E. Pennisi, Science **300**, 1691 (2003), and related articles in this special issue: Tree of Life.
[6] F. Flam, Science **266**, 1320 (1994); P. Gtaziano, A. Marcella, and S. Cecilia, Methods Enzymol. **266**, 281 (1996).
[7] Z. Ouyang, H.-Q. Zhu, J. Wang, and Z.-S. She, J. Bioinform. Comput. Biol. **2**, 353 (2004); H.-Q. Zhu, G.-Q. Hu, Z. Ouyang, J. Wang, and Z.-S. She, Bioinformatics **10**, 1093 (2004).
[8] C.-K. Peng *et al.*, Nature (London) **356**, 168 (1992).
[9] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
[10] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[11] 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).
[12] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).
[13] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995); A. Arneodo *et al.*, Physica A **249**, 439 (1998).

[14] W. Li, Comput. Chem. (Oxford) **21**, 257 (1997).
[15] S. Nee, Nature (London) **357**, 450 (1992); J. Maddox, Nature (London) **358**, 103 (1992); S. Karlin and V. Brendel, Science **259**, 677 (1993).
[16] M. de Sousa Vieira, Phys. Rev. E **60**, 5932 (1999).
[17] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).
[18] R. Román-Roldán, P. Bernaola-Galván, and J. L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).
[19] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).
[20] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H. E. Stanley, Phys. Rev. E **65**, 041905 (2002).
[21] J. L. Oliver, P. Bernaola-Galván, P. Carpena, and R. R. Román-Roldán, Gene **276**, 47 (2001).
[22] J. Wang, Q. Zhang, K. Ren, and Z.-S. She, Chin. Sci. Bull. **46**, 1988 (2001).
[23] Z. Ouyang, C. Wang, and Z.-S. She, Phys. Rev. Lett. **93**, 078103 (2004).
[24] Z.-S. She and E. Leveque, Phys. Rev. Lett. **72**, 336 (1994).
[25] W. Li, P. Bernaola-Galván, P. Carpena, and J. L. Oliver, Neurobiol. Aging **27**, 5 (2003).
[26] L.-C. Hsieh, L. Luo, F. Ji, and H. C. Lee, Phys. Rev. Lett. **90**, 018101 (2003).
[27] W. Li, G. Stolovitzky, P. Bernaola-Galván, and J. L. Oliver, Genome Res. **8**, 916 (1998).
[28] Z.-S. She, Prog. Theor. Phys. Suppl. **130**, 87 (1998).
[29] Z.-S. She, K. Ren, G. S. Lewis, and H. L. Swinney, Phys. Rev. E **64**, 016308 (2001).
[30] C. Baroud, B. Plapp, H. Swinney, and Z.-S. She, Phys. Fluids **15**, 2091 (2003).
[31] Z.-S. She *et al.*, Prog. Nat. Sci. **12**, 747 (2002).
[32] P. Padoan, S. Boldyrev, W. Langer, and A. Nordlund, Astrophys. J. **583**, 308 (2003)
[33] D. Queiros-Conde, Phys. Rev. Lett. **78**, 4426 (1997).
[34] A. Turiel, G. Mato, N. Parga, and J.-P. Nadal, Phys. Rev. Lett. **80**, 1098 (1998).
[35] J. Liu, Z.-S. She, H. Guo, L. Li, and Q. Ouyang, Phys. Rev. E **70**, 036215 (2004); J. Liu, Z.-S. She, Q. Ouyang, and X. T. He, Int. J. Mod. Phys. B **17**, 4139 (2003); H. Y. Guo, L. Li, Q. Ouyang, J. Liu, and Z.-S. She, J. Chem. Phys. **118**, 5038 (2003).
[36] Z.-S. She and L. Liu, Acta Mech. Sin. **19**, 453 (2003); L. Liu and Z.-S. She, Fluid Dyn. Res. **33**, 261 (2003).
[37] Supplementary data, including details of genomic data, tables of measured β, C_i, and S_i, can be found in our website (http://ctb.pku.edu.cn/main/SheGroup/Index.htm) for interested readers.
[38] A. L. G. Simpson *et al.*, Nature (London) **406**, 151 (2000).
[39] Y. Nakamura *et al.*, DNA Res. **9**, 123 (2002); D. Honda, A. Yokota, and J. Sugiyama, J. Mol. Evol. **48**, 723 (1999).
[40] B. Palenik *et al.*, Nature (London) **424**, 1037 (2003).
[41] J. G. Lawrence and H. Ochman, Proc. Natl. Acad. Sci. U.S.A. **95**, 9413 (1998); W. Martin, BioEssays **21**, 99 (1999); A. M. Campbell, Theor Popul. Biol. **57**, 71 (2000); H. Ochman, J. G. Lawrence, and E. A. Groisman, Nature (London) **405**, 299 (2000).
[42] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M. H. R. Stanley, and M. Simons, Biophys. J. **65**, 2673 (1993).
[43] N. J. Bowen and I. K. Jordan, Curr. Issues Mol. Biol. **4**, 65 (2002).
[44] C. Adami, C. Ofria, and T. C. Collier, Proc. Natl. Acad. Sci. U.S.A. **97**, 4463 (2000).
[45] E. V. Koonin, K. S. Makarova, and L. Aravind, Annu. Rev. Microbiol. **55**, 709 (2001).
[46] M. C. Rivera and J. A. Lacke, Nature (London) **431**, 152 (2004).
[47] E. F. DeLong, Curr. Opin. Microbiol. **5**, 520 (2002).