Quantitative measure of randomness and order for complete genomes

Sing-Guan Kong,[1,2] Wen-Lang Fan,[1,2] Hong-Da Chen,[1,2] Jan Wigger,[4] Andrew E. Torda,[4] and H. C. Lee[2,3,5]
[1]*Department of Physics, National Central University, Chungli, Taiwan 32001, Republic of China*
[2]*Graduate Institute of Biophysics, National Central University, Chungli, Taiwan 32001, Republic of China*
[3]*Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan 32001, Republic of China*
[4]*Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, D-20146 Hamburg, Germany*
[5]*National Center for Theoretical Sciences, Hsinchu, Taiwan 30043, Republic of China*

(Received 7 July 2008; revised manuscript received 14 April 2009; published 9 June 2009)

We propose an order index, ϕ, which gives a quantitative measure of randomness and order of complete genomic sequences. It maps genomes to a number from 0 (random and of infinite length) to 1 (fully ordered) and applies regardless of sequence length. The 786 complete genomic sequences in GenBank were found to have ϕ values in a very narrow range, $\phi_g=0.031^{+0.028}_{-0.015}$. We show this implies that genomes are halfway toward being completely random, or, at the "edge of chaos." We further show that artificial "genomes" converted from literary classics have ϕ's that almost exactly coincide with ϕ_g, but sequences of low information content do not. We infer that ϕ_g represents a high information-capacity "fixed point" in sequence space, and that genomes are driven to it by the dynamics of a robust growth and evolution process. We show that a growth process characterized by random segmental duplication can robustly drive genomes to the fixed point.

DOI: 10.1103/PhysRevE.79.061911 PACS number(s): 87.14.G−, 02.50.−r, 05.45.−a, 87.15.Cc

I. INTRODUCTION

The *edge of chaos* originally refers to the state of a computational system, such as cellular automata, when it is close to a transition to chaos, and gains the ability for complex information processing [1–3]. The notion has since been used to describe biological states, and life in general, on the assumption that life necessarily involves complex computation [4]. In model systems such as cellular automata, there are well defined procedures for recognizing the change in computational capability during the transition from nonchaotic to chaotic states [1,3]. However, these have not been adapted to the wider biological context, even for the simplest of organisms. But if we represent a living organism by its genome, view evolution as a dynamical process that drives genomes in the space of sequences, and consider chaos as a state of genome randomness, then we have a framework within which the meaning of "life occurs at the edge of chaos" may be investigated. Genomes, linear sequences written in the four chemical letters, or bases, A (adenine), C (cytosine), G (guanine), and T (thymine) and often referred to as books of life, regulate the functioning of organisms through the many kinds of codes embedded in them (there are also nontextual posttranslational regulations; see, e.g., [5]). When genomes are seen as texts, they have several key properties reflecting their complexity, including long-range correlations and scale invariance [6–11], self-similarity [12–15], fractal property [16,17], and distinctive Shannon redundancy [18–20]. However, these properties do not give a measure of the proximity of a genome to chaos or randomness. Before the edge-of-chaos notion can be explored, one needs to have a quantity that measures the randomness of genomes as texts.

II. DEFINITION OF ORDER INDEX

Here we analyze a genomic sequence of length L (in bases) in terms of the frequency of occurrence of k-letter words, called k-mers, where k is a small integer, and denote the set of all $\tau\equiv 4^k$ types of k-mers by $\mathcal{S}$. Given a sequence, we count the frequency of occurrence (or frequency) f_u of each k-mer type u in $\mathcal{S}$ using an overlapping sliding window of width k and slide one [17]. The sum of the frequencies is $\Sigma_{u\in\mathcal{S}} f_u=L-k+1$, approximated by L. Let the fractional A/T and C/G content of a sequence be denoted by p and $q=1-p$, respectively. Whereas the A/T to C/G ratio, or p/q, varies widely from genome to genome, the well-verified Chargaff's second parity rule [21–23] states that in any long stretch of a single strand of genomic sequence the A:T and C:G ratios are both invariably close to 1. This property suggests a binary partition of $\mathcal{S}$ into subsets (m sets) $\mathcal{S}_m$, $m=0$ to k, where each of the $\tau_m=\binom{k}{m}2^k$ types of k-mers in $\mathcal{S}_m$ contain m and only m A/T's [note that $\Sigma_m\tau_m=\tau 2^{-k}\Sigma_m\binom{k}{m}=\tau$] [24]. For example, in the case of $k=2$, $\mathcal{S}_0$ is the set CC, CG, GC, and GG; $\mathcal{S}_1$ is the set CA, CT, GA, GT, AC, AG, TC, and TG; and $\mathcal{S}_2$ is the set AA AT, TA, and TT. Let $L_m=\Sigma_{u\in\mathcal{S}_m} f_u$ be the frequency sum of k-mers in $\mathcal{S}_m$, then $\Sigma_m L_m=L$. In a p-valued infinitely long random sequence that obeys Chargaff's second parity rule (always assumed unless otherwise stated), the relative frequency of a k-mer belonging to the m set is $p^m q^{k-m}$, hence $L_m^{\{\infty\}}/L=\lim_{L\to\infty} L_m/L=2^{-k}\tau_m p^m q^{k-m}$. For $k\geq 2$, we define the kth *order index* for the sequence as

$$\phi_k \equiv \frac{1}{[2-2(p^k+q^k)]}\sum_m \frac{1}{L}\left|L_m - L_m^{\{\infty\}}\right|. \qquad (1)$$

The definition of the index is based on the observation that as a sequence (or any sample) approaches randomness, its distribution approaches uniformity, here represented by L_m approaching $L_m^{\{\infty\}}$. The definition of ϕ is different, as it is based on distribution averages, as opposed to the more conventional distribution variances. We emphasize that our approach would not work without the m-set partition, nor if the nucleotide mapping were purine-pyrimidine or amino-keto instead of weak-strong (see below). By definition ϕ_k will be

TABLE I. The $k=2$ order index of artificial ordered, checkerboard long ($L \to \infty$) sequences and their concatenates.

				$L_m^{\{\infty\}}/L$			L_m/L			
Sequence	Description	p	N_ϕ [a]	$m=0$	$m=1$	$m=2$	$m=0$	$m=1$	$m=2$	$\phi_{k=2}$
X	"Ordered"[b]; 50%A/T, 50%C/G	0.5	1	0.25	0.5	0.25	0.5	0	0.5	1
Y	"Checkerboard"[c]	0.5	1.0	0.25	0.5	0.25	0	1	0	1
U	Ordered; 40%A/T, 60%C/G	0.4	0.96	0.36	0.48	0.16	0.6	0	0.4	1
V	80% checkerboard, 20% A/T	0.60	0.96	0.16	0.48	0.36	0	0.8	0.2	0.67
R	Random	0.5	1	0.25	0.5	0.25	0.25	0.5	0.25	0
XY [d]	X and Y set complement	0.5	1.0	0.25	0.5	0.25	0.25	0.5	0.25	0[e]
XU	X and U similar	0.45	0.99	0.303	0.495	0.203	0.55	0	0.45	1.00
XV	X and V nearly set complement	0.55	0.99	0.203	0.495	0.303	0.25	0.4	0.35	0.19
YU	Y and U nearly set complement	0.45	0.99	0.303	0.495	0.203	0.3	0.5	0.2	0.01
YV	Y and V similar	0.55	0.99	0.203	0.495	0.303	0	0.9	0.1	0.81
UV	U and V nearly set complement	0.5	1.0	0.25	0.5	0.25	0.3	0.4	0.3	0.20
RV	Mixing randomness and ordered	0.55	0.99	0.203	0.495	0.303	0.125	0.65	0.225	0.30

[a] Normalizing denominator in Eq. (1).

[b] All A and T (in equal portions) on the 5′ end and all C and G on the 3′ end.

[c] Odd sites are A/T (with equal probability) and even sites are C/G.

[d] XY is the concatenate of the X and Y.

[e] That XY is not random but has ϕ exactly equal to zero should be viewed as an artificial "accident" that does not occur in genomic sequences.

small for long random sequences and approaches zero asymptotically for any k with increasing random sequence length. In what follows, we will often suppress the subscript k from ϕ_k when we speak of generic properties of the index; later we will see that for small k's the index has only a mild k dependence. The p-dependent normalization factor on the right-hand side of Eq. (1) ensures that $\phi \approx 1$ for an "ordered" sequence (in which all A/T's are, say, on the 5′ end and all C/G's are on the 3′ end), and approaches 1 as L approaches ∞. The singularities at $p=0$ and 1 do not pose a practical problem since no genome has such extreme base compositions (ϕ has well-defined $p \to 1$ and $p \to 0$ limits). In defining ϕ we relied on an important property of $\Delta_m \equiv L_m - L_m^{\{\infty\}}$: as a sequence becomes more random, the frequency distribution of its k-mers (within an m set) will approach a random distribution and $|\Delta_m|$ will become smaller. Computing Δ_m separately for each m set also causes ϕ to be largely insensitive to p and makes it meaningful to define ϕ for a sequence that is compositionally heterogeneous, and to compare ϕ's of compositionally diverse genomes. There should be many other ways to quantify randomness/order but as we shall see below, ϕ has some remarkable properties that make it an especially useful metric.

III. GENERAL PROPERTIES OF ϕ

Because an L_m is measured separately for each m set, it is possible for sequences having quite different base compositions (different p's) or word contents, or both, to have very similar ϕ's. If the word contents of two sequences are similar, then they will have approximately equal ϕ's, and the concatenate of the two will have a similar ϕ. Generally, the concatenate of two (nonrandom) sequences will have a ϕ that is equal or less than the larger of the ϕ's of the components (this is analogous to entropy). When the word contents of two ϕ-similar sequences are set complement, then the ϕ of the concatenation of the two sequences will be drastically reduced. Here "complement" is employed in the sense of set theory, not in the sense of nucleotides. We use a pair of simple artificial sequences to illustrate this point. Consider the $k=2$ m sets in two very long ($L \to \infty$), equal-length, and even-composition sequences: the ordered sequence X and the "checkerboard" (A/T in odd sites are and C/G in even sites) sequence Y. In X, the eight 2-mers in $\mathcal{S}_0$ and $\mathcal{S}_2$ have frequencies equal to $L/8$ and the eight 2-mers in $\mathcal{S}_1$ have zero frequency. The opposite is true in Y; those in $\mathcal{S}_0$ and $\mathcal{S}_2$ have zero frequency and those in $\mathcal{S}_1$ have frequencies equal to $L/8$. The two sets of word contents in X and Y are said to be *set complement*. We have $\phi_X=1$ and $\phi_Y=1$, but the concatenate XY of the two sequences has $\phi_{XY}=0$; see Table I for details. There is huge range of possibilities between exactly the same and exact set complementarity. In Table I, XU and YV are examples of different degrees of (word-content) similarity and XV, YU, and UV are examples of set complementarity. In any case a significant reduction in the ϕ of a concatenate is sufficient evidence that the word contents of the two component sequences are more set complement than similar.

For random sequences, ϕ is approximately $L^{-1/2}$. We examine the properties of ϕ using random sequences. From the central limit theory we expect (for random sequences) $|\Delta_m|$ to scale as $L_m^{-1/2}$. We therefore expect ϕ to be proportional to $L^{-1/2}$ on average.

The log-log plots in Figs. 1(a) and 1(b) show ϕ as a function of sequence length for different k's and p's. Each datum is averaged over 500 random sequences. It is seen that ϕ scales very well as $L^{-1/2}$ (with sizable fluctuations) and is only weakly dependent on k and p. These results can be

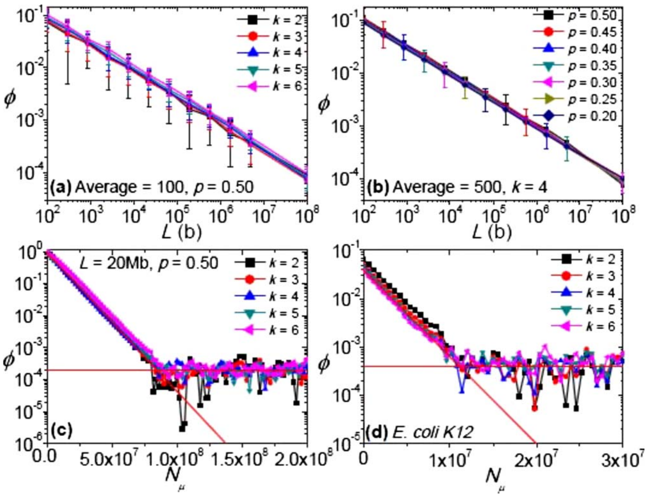

FIG. 1. (Color online) (a) Log-log plot of order index, ϕ, vs length of random sequence for $p=0.5$ and $k=2-6$. (b) Same as (a); for $k=4$ and $p=0.20-0.50$. (c) Semi-log plot of ϕ vs N_μ, number of random point mutations, for an initially ordered 20 Mb, $p=0.5$ sequence. The intersection of the lines is the critical point where sequence becomes random. (d) Same as (c); initial sequence is genome of *E. coli*.

summarized for all k and p by an empirical relation:

$$\phi^{\{ran\}} = c_\phi L^{-\gamma_\phi}, \tag{2}$$

with $\gamma_\phi^{\{ran\}}=0.50\pm0.01$ and $c_\phi^{\{ran\}}=1.0\pm0.2$ or, to a good approximation, $\phi^{\{ran\}}\approx L^{-1/2}$. This leads to the convenient concept of an (approximately k-independent) *equivalent length* for a ϕ-valued sequence,

$$L_{eq}(\phi) \equiv \phi^{-2}, \tag{3}$$

the nominal length of a random sequence whose order index is ϕ.

In Fig. 1(b), the weak dependence of ϕ on p is a consequence of the m-set partition of k-mers, designed specifically for the binary G/C versus A/T grouping, also known as strong-weak mapping, of nucleic letters. Two others, the purine-pyrimidine (A/G versus C/T) and amino-keto (A/C versus G/T) mappings, have also been used as binary reductions in DNA analysis [6–8]. The existence of Chargaff's second parity rule renders the strong-weak mapping special, and neither of the other two mappings will yield results similar to Figs. 1(a) and 1(b) and summarized in Eq. (2). To be specific we discuss the purine-pyrimidine mapping (the argument works similarly for the amino-keto mapping). In such a mapping all genomic sequences will have p' $(=p_{AG})$ very close the 0.5 (thanks to Chargaff's second parity rule), which would seem to remove the need to view p' as a variable for consideration. Suppose we carry out an m-set partition of the k-mers as before, by putting all k-mers with m AG's in an m set. However, unless $p(=p_{AT})=0.5$, the frequency distribution of the k-mers in an m set will be multimodal ($k+1$ modes to be precise) instead of unimodal as before [24]. In other words, the k-mers can have highly nonuniform frequencies even in a random sequence of infinite length, so that there is not a natural quantity corresponding to L_m. If we define $L_m^{\{\infty\}}=L\tau_m/\tau$ [because $p'=q'=0.5$, see Eq. (1)] regardless, then we will not have the result given in Eq. (2), namely, a $\phi^{\{ran\}}$ that always vanishes as $L\rightarrow\infty$, but instead a $\phi^{\{ran\}}$ that depends strongly with p and which, when p deviates significantly from 0.5, remains finite at all lengths. Neither will we get results for a genome that are easy to summarize (see below). In fact, the difference between the strong/weak mapping and the alternative mappings discussed above is the main reason why statistic properties of genomes are the least ambiguous when the strong/weak mapping is used [6–8,20,25].

A. Order index decays exponentially with rate of point mutation

Random events such as point mutations acting on a nonrandom sequence decreases its order, and hence its ϕ. Figure 1(c) shows that the ϕ of a $p=0.5$, 20 Mb ordered sequence, decreases exponentially with the number of mutations N_μ, until N_μ reaches a critical number $N_{\mu c}$. The critical value reflects the fact that the randomness of an already-random sequence cannot be increased. In other words, if one thinks of a random point mutation as a dynamical action taking a sequence from one point in the sequence space to another, then a randomized sequence is a fixed point of the action. Our studies of initially ordered sequences having a variety of lengths and base compositions yield

$$\phi = \begin{cases} \exp(-2N_\mu/L), & N_\mu \lesssim N_{\mu c} \\ \phi_c \approx L^{-1/2}, & N_\mu > N_{\mu c}, \end{cases} \tag{4}$$

where $N_{\mu c}\approx(1/4)L\ln L$ is the number of mutations after which the sequence becomes random, or "critical," hence we define the critical mutation rate as

$$\mu_c \equiv N_{\mu c}/L \approx (1/4)\ln L. \tag{5}$$

This formula for $N_{\mu c}$ compares well with simulation. In the case of Fig. 1(c), the coordinates of the simulation ($k=4$) critical point are $(\phi_c,N_{\mu c})=(2.2\times10^{-4},8.5\times10^7)$, as compared to the values $(2.2\times10^{-4},8.4\times10^7)$ given by Eqs. (4) and (5). For typical sequences of genomic length ($L\sim10^{7\pm1}$ Mb), $\mu_c=4.0\pm0.6$ mutations per base (b^{-1}). We use Eq. (4) to assign to a ϕ-valued sequence (before it becomes critical) an *equivalent mutation rate*,

$$\mu_{eq}(\phi) \equiv \ln \phi^{-1/2}, \tag{6}$$

the nominal number of random point mutations per base required to bring the index of an ordered sequence to ϕ.

Equation (4) can be adapted for application to sequences not initially ordered. For example, the equivalent mutation rate for the 4.6 Mb genome of *E. coli* ($\phi=0.049$) is 1.5 b^{-1}. Since for a 4.6 Mb sequence $\mu_c=3.8$ b^{-1}, one expects an additional $2.3\times4.6\times10^6=1.1\times10^7$ mutations are needed to randomize it. In the simulation shown in Fig. 1(d), the actual number needed is found to be $(1.1\pm0.1)\times10^7$.

B. Positions of *k*-mers in genomes are essentially uncorrelated

Although a casual glance at a genomic sequence is sufficient to see that it is very far from being ordered, here, before

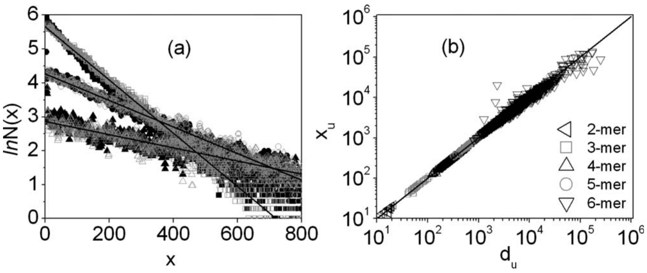

FIG. 2. (a) Interval distributions of three 4-mers with frequencies $f_u=\bar{f}=L/256$ (○), $2\bar{f}$ (□), and $\bar{f}/2$ (△), respectively, in *E. coli* (L=4.6 Mb, p=0.5; solid symbols) and in corresponding random sequence (hollow symbols); solid lines express Eq. (7). (b) Average interval x_u of k-mer u versus d_u, where $-d_u$ is the slope of the linear regression of the logarithm of the interval distribution of u; $x_u=d_u$ if the distribution is exponential. Data for all k-mers, k=2 to 6, in *E. coli* are shown; for each k the number of data and the mean value of x_u are both equal to 4^k.

computing the ϕ's for genomic sequences, we use a conventional method to show qualitatively that genomes are close to being random. Given a specific k-mer u, we consider the distribution $N(x)$ of intervals x of adjacent pairs of u's in a sequence. If the positions of the u's are uncorrelated, as they would be in a random sequence, then the distribution will fall along the exponential given by

$$N^{\{ran\}}(x) = N_0 e^{-x/x_u} \tag{7}$$

where $x_u=L/f_u$ is the average interval, f_u is the frequency of u, L is the sequence length, and $N_0=f_u^2/L$. Figure 2(a) shows that the interval distributions of three 4-mers in the genome of *E. coli* are quite well represented by Eq. (7). In each case the small scattering along the exponential indicates that the u sites are not entirely uncorrelated. Let $-d_u$ be the linear regression of the logarithm of an interval distribution. If the distribution is given exactly by $N^{\{ran\}}(x)$ then $d_u=x_u$. In Fig. 2(b) x_u is plotted against d_u for all k-mers in *E. coli*, k=2 to 6. It is seen the vast majority of data fall on the straight line $x_u=d_u$, which suggests that in *E. coli* the sites of at least most the shorter k-mers are substantially uncorrelated (this does not conflict with the known presence of long-range correlation in genomes [6–11]). This was shown to be a general property of 41 randomly complete bacterial genomes for k =2 to 6 [26] (because d_u cannot be statistically reliably extracted unless $f_u \gg 1$, the study was not extended to k-mers with k greater than 6). Note that even when this property holds for most k-mers, it need not hold for all k-mers. Known exceptions are some overrepresented k-mers, whose spatial distributions tend to be highly localized [27].

C. Genomic ϕ is length independent and nearly universal

We computed ϕ, for 384 complete prokaryotic genomes (28 archaebacteria and 356 eubacteria) and 402 complete chromosomes from 28 eukaryotes of lengths ranging from 200 kb to 230 Mb. When computing ϕ, "N-runs," or gaps in the chromosome, are spliced out from the sequence. The rice genome was downloaded from the Rice Annotation Project Database [28], and all other sequences from the National Center for Biotechnology Information genome database [29], during the period 26 February to 27 November, 2006. A list of the 786 genomes/chromosomes studied and basic properties—length and base composition—are given in the Order Index Database (OIDB) [30]. We present results for k=2 to 6 because, mainly due to statistical fluctuation, the k dependence of ϕ becomes noticeable for larger k's and increases with k, especially for the shorter chromosomes. Recall that the average frequency of a k-mer in an m set is $\bar{f}_m$ is approximately $L_m^{\{\infty\}}/\tau_m$, hence $\bar{f}_{\{m=0\}} \approx q^k L/2^k$ and $\bar{f}_{\{m=k\}} \approx p^k L/k^2$, and one of them can be very small when p or q is significantly less than 0.5. For instance, when p=0.3, for a chromosome (L=) 2 Mb long, $\bar{f}_{\{m=k\}} \sim 3.4$ when k=7 and ~0.51 when k=8, which are frequencies too small for reliable statistics. Statistical fluctuation for k up to 10 is not a concern for most of the vertebrate chromosomes, which are of the order of 100 Mb and have $p \sim 0.58$. Their order indexes for k=7 to 10 do not differ qualitatively from those reported here for the smaller k's. The k=2 to 6 order indexes for all individual chromosomes are given in tabulated form in OIDB [30]. Generally, intrachromosomal variation in ϕ with k, typically within 50%, is less than interchromosomal variation in ϕ. The situation is similar to that for the random and *E. coli* sequences shown in Fig. 1. While a weak k dependence in ϕ simplifies our narrative, a stronger k dependence would not alter our understanding of the matter. Henceforth we discuss the propertied of k-averaged ϕ. The results shown in Fig. 3 indicate that genomic ϕ's systematically vary neither with sequence length [(a) and (b)] nor with base composition [(c)]. The average and variance of the entire set of genomic data are

$$\ln \phi_g = - 3.49 \pm 0.65 \tag{8}$$

or, equivalently $\phi_g = 0.031^{+0.028}_{-0.015}$ (it is better to express deviation in logarithmic form because ϕ can span many decades). The range in L_{eq}, 0.29–4.4 kb, corresponding to the range of ϕ_g is very small when compared to the range of actual sequence lengths. For example, if the 4.6 Mb *E. coli* chromosome and the 226 Mb human chromosome 1 were random, then their ϕ's would be 4.7×10^{-4} and 6.7×10^{-5}, as opposed to the actual values of 0.049 and 0.038, respectively. Chromosomes from the same organism have highly similar word contents hence highly similar ϕ's. The variances seen in Fig. 3, where they are substantial, are caused by differences among organisms, not among chromosomes from the same organism. Relative to organisms in other categories, the vertebrates are phylogenetically extremely close and their genomes, other than differences arising from genome rearrangements [31] to which ϕ is insensitive, have a high degree of similarity. This explains why the ϕ's of their chromosomes [boxes with log L>7.3 in Fig. 3(b) and the "223 box" at $p \sim 0.6$ in panel (c)] are concentrated in an extraordinary narrow range 0.033–0.050. There are statistically significant differences among the boxes in Figs. 3(b) and 3(c). For instance, the P values (from two-sample t tests) for the ten vertebrates taken as a group paired with the seven boxes in Fig. 3(b) with 10 (log $L \sim 5.8$), 18, 13, 140, 160, 48, and 10 (log $L \sim 7.4$) chromosomes are 0.0022, 0.043, $\sim 5 \times 10^{-5}$,

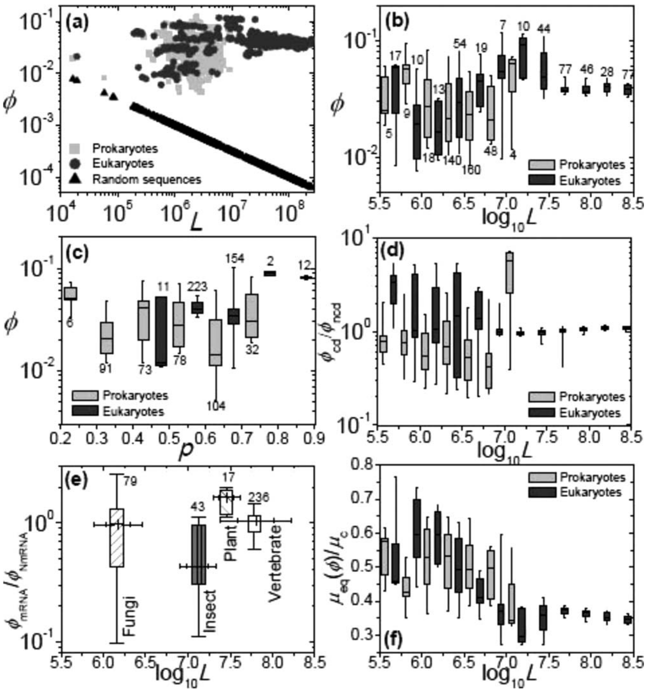

FIG. 3. (a) Order index, ϕ, vs sequence length, L, for 384 prokaryotic genomes (gray ■'s), 402 eukaryotic chromosomes (black ●'s), and random sequences (line composed of solid △'s). In (b)–(f): box (gray for prokaryotes; black for eukaryotes) height is given by 25% to 50% values and the range represents 10% to 90% values; numbers above boxes are numbers of sequences in group; all ϕ's are averaged over $k=2$ to 6. (b) ϕ vs log L. (c) ϕ vs fractional A/T content, p. (d) Ratio of ϕ_{cd} (for genic parts) to ϕ_{ncd} (nongenic) vs log L. (e) Ratio of ϕ_{mRNA} (mRNA segments) to ϕ_{nmRNA} (non-mRNA), averaged over classes of eukaryotes. (f) Ratio of equivalent mutation rate, μ_{eq}, to critical mutation rate, μ_c, vs log L.

$\sim 2\times 10^{-5}$, 0.00015, $\sim 1\times 10^{-5}$, and 0.030, respectively (but are significantly greater than 0.05 when paired with the other boxes). Nevertheless, as is evident from Fig. 3(a), taken together, the genomic ϕ's form a well-defined unimodal distribution without systematic p or L dependence. Figure 4 shows that P values of the boxes in Fig. 3(b) under the null hypothesis that each is part of a normal distribution defined by the right-hand side of Eq. (8). The 37 outlying chromosomes with $P<0.05$, including 14 with $0.04<P<0.05$, are listed in Table II. All except two of the outlying prokaryotic chromosomes have $\phi<0.01$, that is, they are more random than the norm. There are only two eukaryotes among the outliers; four of the 16 *S. cerevisiae* chromosomes have $\phi<0.01$ and eight of the 16 *A. mellifera* chromosomes have $\phi \gtrsim 0.1$, or more ordered than the norm. It would not be surprising had the genomic ϕ's spanned several decades, but it is so when genomes share the common feature of having equivalent lengths far shorter than their actual lengths, and have ϕ's congregate in a comparatively narrow range. For this reason we refer to ϕ_g, defined by Eq. (8), or equivalently, the range $0.016\le\phi\le 0.059$, as the *universal* ϕ of genomes.

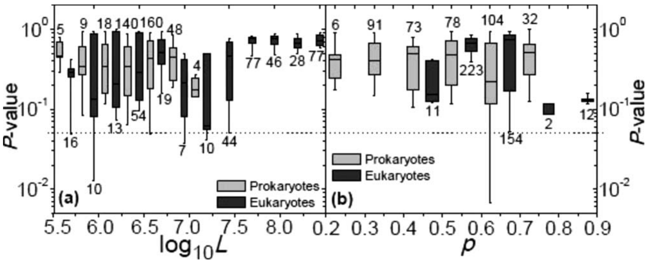

FIG. 4. P values of (a) data boxes from Fig. 3(b) and (b) from Fig. 3(c) being part a normal distribution defined by the right-hand side of Eq. (8). The horizontal dashed line indicates $P=0.05$.

D. Coding and noncoding regions have similar ϕ's

From each complete sequence, we extracted the genic (including introns in eukaryotes) and nongenic parts, then concatenated the parts into two separate sequences and computed their order indexes, ϕ_{cd} and ϕ_{ncd}, respectively. The effect on ϕ from splicing a chromosome to form the concatenates can be estimated. Of the four categories of concatenates—prokaryote/eukaryote and genic/nongenic—the prokaryote-nongenic concatenates are generally the shortest and are expected to be the most affected by the splicing. Consider a prokaryotic chromosome 3 Mb long, with 3500 genes occupying 88% (a typical value) of the chromosome. Then the nongenic concatenate will be 360 kb long with 3500 artificial connecting sites. This will generate $3500\times(k-1)$ k-mers not belonging to the chromosome, or an estimated $(k-1)$ percent error in ϕ_{ncd}. Possible errors on ϕ_{cd} and on the ϕ's of eukaryotic chromosomes are expected to be smaller. The coding concatenates are built from coding sequences from a single strand and includes genes in the positive and negative orientations. Another way to build the coding concatenate, not adopted here, is to have all the genes included, say, in the positive orientation. Because the computation of ϕ is in fact a binary scheme (strong versus weak nucleotides), the only difference the two concatenates have on ϕ will again come from the artificial connecting sties mentioned above. We therefore estimate the difference in ϕ that may arise from the two different ways of building the coding concatenate to be not greater than 10%.

Generally, ϕ_{cd}, ϕ_{ncd}, and the ϕ for the whole chromosome have similar magnitudes. A summary of the ratio ϕ_{cd}/ϕ_{ncd} for sets of genomes grouped by length is given in Fig. 3(d). The genic parts of eukaryotic genomes are further partitioned into mRNA (exon) and non-mRNA (intron) parts, and their ϕ's computed separately. Averaged over sets of organisms, ϕ_{mRNA}/ϕ_{nmRNA} is of the order of 1 [Fig. 3(e)]. In all cases the differences in ϕ between different parts of a genomes (single- or multiple-chromosome) is much smaller than that between genomes and random sequences.

Because noncoding parts (nongenic in prokaryotes and nongenic plus introns in eukaryotes), in spite of the fact that they contain regulation-related sequences, are expected to be more tolerant to random small mutations (point mutations and small indels—insertions and delations), they are expected to be noticeably more random than coding parts (genic in prokaryotes and exons in eukaryotes). That the two

TABLE II. Chromosomes belonging to the universal set defined by Eq. (8) with $P<0.05$.

Name	Accession no.[a]	$\bar{\phi}$ [b]	P value[b]	Name	Accession no.[a]	$\bar{\phi}$ [b]	P value[b]
S. aureus	9 strains[c]	~4.4(−3)	~3.0(−3)	*A. marginale*	4842	4.45(−3)	3.15(−3)
S. epidermidis	4461	4.87(−3)	4.89(−3)	*C. felis*	7899	5.20(−3)	6.66(−3)
L. johnsonii	5362	5.58(−3)	9.18(−3)	*S. hemolyticus*	7168	5.79(−3)	1.08(−2)
S. epidermidis	2976	6.49(−3)	1.77(−2)	*M. mobile 163 K*	6908	6.80(−3)	2.14(−2)
T. denitrificans	7404	7.12(−3)	2.58(−2)	*L. acidophilus*	6814	7.34(−3)	2.90(−2)
G. sulfurreducens	2939	7.40(−3)	2.99(−2)	*F. tularensis*	7880	7.50(−3)	3.15(−2)
W. succinogenes	5090	7.51(−3)	3.17(−2)	*C. hydrogenoformans*	7503	1.23(−1)	3.20(−2)
M. hungatei	7796	7.75(−3)	3.57(−2)	*F. tularensis*	6570	7.90(−3)	3.84(−2)
C. caviae	3361	7.94(−3)	3.91(−2)	*M. succiniciproducens*	6300	1.15(−1)	4.04(−2)
C. abortus	4552	8.06(−3)	4.14(−2)	*X. fastidiosa 9a5c*	2488	8.12(−3)	4.25(−2)
P. marinus	7335	8.19(−3)	4.37(−2)	*S. tokodaii*	3106	8.47(−3)	4.96(−2)
S. cerevisiae	Chr V	6.00(−3)	1.26(−2)	*S. cerevisiae*	Chrs XV, III	~7.7(−3)	~3.5(−2)
S. cerevisiae	Chr VI	8.43(−3)	4.87(−2)	*A. mellifera*	8 chrs.[d]	~1.1(−1)	~4.8(−2)

[a] 4842 indicates the accession no. NC_004842.

[b] The value 4.4(−3) means 4.4×10^{-3}.

[c] The nine strains, in order of increasing P value, are 3923, 2953, 7793, 7795, 2952, 7622, 2951, 2758, and 2745.

[d] The eight chromosomes, in order of increasing P value, are XV, X, XII, II, IV, V, I, and XI.

parts on average have similar ϕ's is therefore noteworthy. A closer examination of Fig. 3(d) shows there is not a fixed relation in randomness between the two parts. Whereas the nongenic parts in a vast majority of prokaryotes tend to be slightly more ordered than the genic parts ($\phi_{ncd}\gtrsim\phi_{cd}$), the genic parts in most nonvertebrate eukaryotes tend to be more ordered. Figure 3(e) shows the introns have a clear preference for being more ordered than exons ($\phi_{nmRNA}\gtrsim\phi_{mRNA}$) in insects, and the opposite in plants. In vertebrates the two parts, either genic and nongenic or exon and intron, have very close to the same randomness. These results suggest that selection-driven small mutations are not the main mechanisms deciding randomness and order on a genomic scale.

It is known that in many species a codon bias exists, and this may be expected to have an impact on ϕ, especially for $k=3$. This effect is not detected in ϕ. A major cause that weakens the codon effect is this: the sliding window used to count k-mer frequencies travels in one direction and has a slide of one, so that only one in three 3-mers is a codon. It is important to realize that ϕ sameness does not imply sequence similarity. However, when ϕ_{cd}, ϕ_{ncd}, and the ϕ for the whole chromosome all have similar magnitudes then it can be inferred that the word contents in the coding and noncoding part do not differ significantly.

E. Genomic scaling exponent $\gamma_\phi\approx 0$ when length exceeds 50 kb

The left panel of Fig. 5 shows for HS1 and *R. baltica* what is generally true for all cases studied: that $\bar{\phi}_l$, the average of segmental ϕ_l for segments of a fixed length l, tends a constant when l exceeds 50 kb, with a corresponding decrease in the variance in ϕ_l. In other words, the genomic ϕ_l is a scale-independent quantity, or

$$\gamma_\phi^{\{genome\}}\approx 0, \tag{9}$$

for scales greater than 50 kb, as compared to random sequences whose γ_ϕ is 1/2 [Eq. (2)]. This implies that word contents over two adjacent regions (at large scales) tend to be more similar than set complementary, for otherwise $\bar{\phi}_l$ would decrease with increasing l. On the other hand, that $\bar{\phi}_l$ does decrease noticeably with increasing l when l is less than 10 kb and the relative large variance in ϕ_l that persists in HS1 even beyond 50 kb both indicate significant intrachromosomal compositional heterogeneity at subchromosomal scales.

IV. INTERGENOMIC COMPOSITIONAL DIVERSITY

The universality of ϕ_g is not a result of intergenomic compositional homogeneity. To verify chromosome diversity we map each chromosome to a multidimensional unit $\vec{s}$ vec-

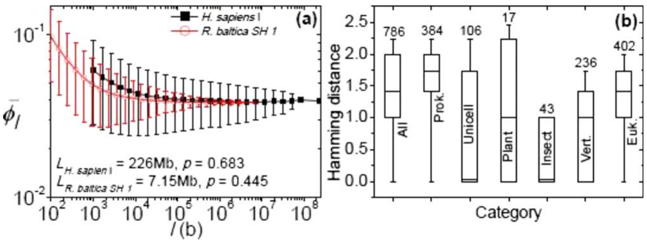

FIG. 5. (Color online) Left panel, average segmental index, $\bar{\phi}_l$ vs segment length l for the human chromosome I (solid square) and the *R. baltica* chromosome (circle). Right panel, category (number of chromosomes given above box) statistics of pairwise Hamming distances between $k=5$ $\vec{s}$ vectors; legends for boxes same as in Fig. 3.

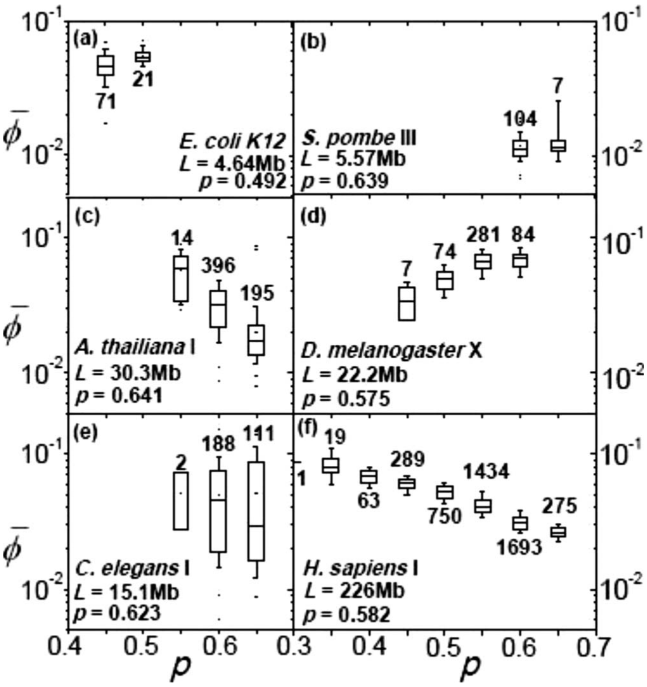

FIG. 6. Box plots of ϕ for 50 kb segments obtained from partition of a chromosome and grouped in 0.05 p intervals; numeral indicates number of segments represented by each box. (a) *E. coli*; (b) *S. pombe* chr. III; (c) *A. thaliana* chr. I; (d) *D. melanogaster* chr. X; (e) *C. elegans* chr. I; (f) *H. sapiens* chr. I.

tor as follows: For given k, the mth component, $m=0$ to k, of $\vec{s}$ is 0 or 1, respectively, when $L_m - L_m^{\{\infty\}}$ is negative or positive. The maximum Hamming distance between two $\vec{s}$ vectors is $\sqrt{k+1}$, and the maximum value for the rms pairwise distance, h_{rms}, for a sufficiently large set of vectors is $\sqrt{(k+1)/2}$. Figure 5, right panel, shows the $k=5$ results for h_{rms} for various categories of chromosomes; the median h_{rms} for eukaryotes, prokaryotes, and all chromosomes are 1.70, 1.41, and 1.60, respectively. Concatenating two chromosomes is yet another way to test their word-content similarity or the lack of it; a sharp drop in the ϕ of the concatenate indicates significant set complementarity in word content, hence interchromosomal heterogeneity. Many such cases can be found and an extreme example is the pair, the 1.23 Mb *Ch. pneumoniae* ($\phi=0.0204$) and the 1.49 Mb *T. whipperlei* ($\phi=0.0149$), whose concatenate has $\phi=0.00197$. Five other cases are given in Fig. 7.

V. DEPENDENCE OF ϕ ON INTRACHROMOSOMAL COMPOSITIONAL HETEROGENEITY

It is known that compositional heterogeneity is prevalent in genomes [11,32–35]. The extent of this is already suggested by the sizable variances in ϕ_l seen in Fig. 5(b). The box plots in Fig. 6, where each box gives the statistics of ϕ of 50 kb segments grouped according to p, explores this notion further. While $0.01 \leq \bar{\phi} \leq 0.1$ for almost all cases, the range of ϕ and its dependence on p are seen to vary significantly from organism to organism. In the case for HS1 [panel (f)], where the range of ϕ is consistent with what is shown in

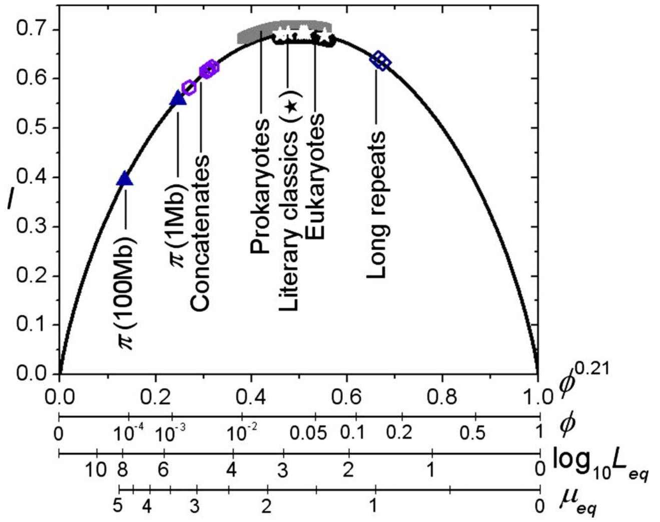

FIG. 7. (Color online) The function $I(z)$ [Eq. (10)] plotted as a function of $z=\phi^{0.21}$; ϕ; $\log_{10} L_{eq}(\phi)$; $\mu_{eq}(\phi)$ (in units of b^{-1}). Coordinates for 384 prokaryotic and 402 eukaryotic chromosomes are shown in gray and black, respectively, and those for the six classics (stars), two long repeats of short sentences (diamonds), five bichromosome concatenates (hexagons), and two π sequences (triangles).

Fig. 7 (for $l=50$ kb), there is a clear trend of $\bar{\phi}$ decreasing with increasing p. Because GC-rich regions are known to be richer in genes, this result may be interpreted as indicating gene-rich regions in HS1 are more ordered than gene-poor regions. However, this cannot be a general argument because in the case of *D. melanogaster* Chr. X [panel (d)] $\bar{\phi}$ trends with p in the other direction. Figure 6 also reveals a second aspect of intrachromosomal heterogeneity, a wide range of ϕ values for segments having the same p, as displayed by the boxes at $p=0.6$ and 0.65 in *C. elegans* [panel (e)] and the $p=0.6$ box in *A. thaliana* [panel (c)]. Notwithstanding their significant intrachromosomal variations, local p-specific ϕ's mostly fall within the range of ϕ_g.

A well-known type of intrachromosomal heterogeneity is cumulative base skews, or compositional asymmetry, where (in some genomes) there is a difference in base compositions of transcripts in the leading and lagging strands [36–39]. The most discussed compositional asymmetries are the AT skew [(A−T)/(A+T)] and GC skews, that is, deviations from Chargaff's second parity rule [21]. Because the relative magnitude of skews are small, of the order of a few percent [36,37], this asymmetry is not expected to have a significant effect on ϕ, which is a coarsed-grained measure. For instance, the eight genomes, given in the descending order of skew, *B. subtilis* (high level of skew), *E. coli*, *H. pylori*, *T. pallidum*, *H. influenzae*, *M. jannaschii*, *M. thermoautotrophicum*, *A. fulgidus*, and *Synechocystis* (almost no skew) [36,37], have ϕ values that are 0.036, 0.049, 0.084, 0.012, 0.051, 0.033, 0.026, and 0.069, respectively. Our study of skews in complete genome [40] indicates the lack of correlation between skew and ϕ to be a general trend.

VI. GENOME AT THE EDGE OF CHAOS

The ratio $\mu_{eq}(\phi)/\mu_c$ is an indication of how close a sequence is to being random. Figure 3(f) shows that the overall average of the ratio is 0.45 ± 0.11, meaning that the typical genomic equivalent mutation rate is about 1.8 b^{-1}, as compared to the critical mutation rate of approximately 4 b^{-1} (that would randomize the genome). Thus, for example, a typical worm (*C. elegans*) chromosome, with a length of about 17 Mb, is as random as an initially ordered 17 Mb sequence after having undergone 31 million random mutations. A mutation rate of 1.8 b^{-1} seems quite high, since one might well expect a sequence that has experienced an average of one mutation per site to be random. We think the term "at the edge of chaos" is a useful qualitative description of the state of randomness of genomes, namely, there are close to being but not quite random, and conveys the idea intended for the phrase "life at the edge of chaos." We emphasize that our inference is based on an observation of empirical data and does not have the force of a theoretical model with predictive power. For instance, we have not made the necessary study to say whether there is indeed a sharp edge—say, a first-order transition—beyond which the sequence is too unordered to code for life.

VII. ORDER INDEX OF CLASSICAL LITERARY TEXTS

In order to see if one may establish a deeper associate between ϕ_g with the fact that genomes are carriers of biological codes, we computed ϕ's for four sets of artificial pseudogenomes converted from nongenomic sequences: (1) six literary classics ranging in length from 0.08 to 6 million letters: *The Bible*, King James Versiov (3.22 million letters); *Sonnets*, Shakespeare (0.08); *Oliver Twist*, Dickens (0.69); *Remembrance of Things Past* (English translation), Proust (5.92); *Ulysses*, Joyce (1.19); *A Moveable Feast*, Hemingway (0.19). (2) long repeats (1 M times) of the two short sentences, "Though this be madness, yet there is method in't" (Hamlet) and "All the perfumes of Arabia will not sweeten this little hand" (Macbeth); (3) five bichromosome concatenates: 2491–4572, 2179–4551, 0912–7354, 4432–5072, and 0919–4605, where, e.g., the number 2491 designates the NCBI genome sequence access number NC_002491; (4) two sequences, 1 M [41] and 100 M [42] digits, respectively, of the irrational number π. The ϕ's of the pseudogenomes are given in Fig. 7.

In sets (1) and (2) the conversion of the literary texts to pseudogenomes is made by discarding all nonalphabetic symbols in a text and using an alphabet-to-base mapping (although a binary mapping is sufficient for our purpose) constructed according to two basic rules: (i) each category of alphabets is as much as possible evenly distributed to the two groups A/T and C/G, and to a lesser extent, to the four nucleotides; (ii) the resulting six pseudogenomes satisfy Chargaff's second parity rule [21] approximately. The categories of alphabets we used are as follows: (aeiouy), (bmpfvw), (cszdlnt), and (ghjkqxr). Rule (i) avoids mappings that cannot produce words, such as those that map all the vowels to A/T. The mapping used here, which yields nearly even-based pseudogenomes for the six classics, is (adjlsy) to "A," (chiopq) to "C," (efgnvxz) to "G," and (bkmrtuw) to "T." This maps each of the six classics to a pseudogenome with $p = 0.50 \pm 0.02$ ($p_A \sim p_C \sim p_G \sim p_T = 0.250 \pm 0.007$), and the Hamlet and Macbeth sentences to $p = 0.50$ and 0.56 sequences, respectively. The range of (the k-averaged) ϕ for the classics is 0.036–0.064, which lies within the range of ϕ_g. We have computed the ϕ's of many more long text and used alternative mappings obeying rules (i) and (ii) yielding pseudogenomes with $0.2 \lesssim p \lesssim 0.8$, and found that the ϕ's of the vast majority of the pseudogenomes lie with the range of 0.02–0.07.

Using the mapping given above to convert the two repeats in set (2) yields for the pseudogenomic sequences $\phi \approx 0.16 > \phi_g$ and $L_{eq} \approx 40$ b which, as expected, are properties close to those of the original short root sentences. The five bichromosome concatenates, whose ϕ's are of the order of 0.005, which is considerably less than ϕ_g. These examples show that phylogenetically distant chromosomes can have word contents that are closer to being set complement than similar, so that concatenating them can lead to sharply reduced ϕ. Conversely, concatenating long segments from the same chromosome, or from two chromosomes that are phylogenetically close, typically leaves ϕ little changed.

For the two series of π in set (4) the even and odd digits are mapped to A/T and C/G, respectively. The pseudogenomic π sequences have L_{eq}'s 0.63 and 280 Mb, respectively. Since these are close to their true sequence length, this implies that the π series are essentially random.

VIII. HYPOTHESIS

$\phi \sim \phi_g$ is a signature for high information capacity. We observe a general trend: long sequences with high information content—genomes and pseudogeneomes converted from literary classics—tend to have $\phi \sim \phi_g$, and L_{eq}'s of the order of 0.29–4.4 kb, several orders of magnitude shorter than their true lengths. In comparison, long sequences with low information content either (the long repeats) have ϕ significantly greater than ϕ_g because they are too ordered, in which case they have even shorter L_{eq}'s, or (the π series) have ϕ significantly less than ϕ_g because they are too random, with L_{eq}'s more closely tracking their true lengths. Series such as those of π are sometimes cited as examples of high complexity. Here we can make a distinction between high complexity and high information content. Since the meaning of π can be simply stated: "the ratio of circumference of a circle to its diameter," the information content of a long π series may indeed be said to be low, even though the textual structures of the series are extremely complex. The situation with the bichromosome concatenates is more complicated. In each case, the two component chromosomes each has high information content, but in the composite the information is admixed so as to reduce ϕ; just as a message written in two sets of codes will have some information garbled.

Our observation is consistent with the generally accepted understanding: if a text has close to the maximum information density then its textual structure cannot be either too random or too ordered. We therefore propose the following hypothesis: the criteria $\phi \sim \phi_g$ and $L_{eq}/L \leqslant 10^{-2}$ (L is the true length) are necessary conditions for high information-capacity sequences. We speak of information capacity instead of information content because ϕ only characterizes the textual structure of a sequence, but is not a direct measure of its information. To give the hypothesis mathematical form we construct a function $I(z)$, intended as a relative indicator of the information capacity of a sequence and satisfying the criteria: the variable z is a scaling function of ϕ with $z|_{\phi=0}=0$ and $z|_{\phi=1}=1$, and I has two minima at $I(0)=I(1)=0$ and a maximum at $z|_{\phi=\phi_g}=0.5$. The simplest solution is

$$I(z) = -z \ln z - (1-z)\ln(1-z); \quad z = \phi^{0.21}. \qquad (10)$$

Other than the rescaling of ϕ, which gives equal space (in z) to more ordered ($\phi > \phi_g$) and more random ($\phi < \phi_g$) sequences, I reminds one of entropy, except that instead of only one minimum at the random limit, it has two minima representing both ordered and random sequences. Figure 7 shows $I(z)$ plotted against z together with symbols indicating the coordinates of the sequences discussed in the text above. In addition, three other sets of abscissas are given: ϕ, $\log_{10} L_{eq}(\phi)$ [Eq. (3)], and $\mu_{eq}(\phi)$ [Eq. (6)]. As intended, Fig. 7 shows the high information-content sequences—genomes and literary texts—concentrated around the region, defined by $\phi_g \sim 0.015-0.059$, $\log L_{eq,g} \sim 2.5-3.7$ and $\mu_{eq,g} \sim 1.4-2.1$, near the peak of the I curve ($0.670 \lesssim I < I_{\max} = 0.693$) and equally far removed from the highly random ($z \sim 0$) and highly ordered ($z \sim 1$) regions. Roughly, one may set $\phi_{edge}=0.1$, corresponding to $\mu_{eq}(\phi)=1.2\ \mathrm{b}^{-1}$, as the beginning of the edge of chaos.

Sequence growth by segmental duplication has small equivalent lengths. Although genomes and literary texts share the property of having $\phi \sim \phi_g$, we cannot assume they were produced similarly. Whereas the classical texts were written with deliberation, at least large parts of genomes were "blindly" formed. Considering that genomes have very short equivalent lengths and that they have extreme scales invariance ($\gamma_\phi \sim 0$), segmental duplication likely is an important characteristic of the growth process. Segmental duplication [43–45] is known as an important mechanism in genome growth and evolution. From the physics perspective segmental duplication is the easiest way for genomes to acquire length while retaining a short equivalent length. From the biological perspective, duplication is the common explanation for proteins acquiring new functions without destroying old ones, and duplication events provide the only common explanation for the promiscuity of common protein domains. Figure 8 shows the ϕ's from model sequences generated in a minimal model of genome growth based on random segmental duplication [46]. Here the term random refers to the length of the duplicated segment and the sites on the sequence where it is selected and where the duplication is inserted. We choose a random process for the model because it is the simplest possible procedure; we know that (Fig. 2) at least the shorter k-mers in genomes are essentially randomly placed [26]; genes, most of which are homologs and therefore produced by duplications, are to a first order of approximation randomly placed over a chromosome (see, however, [47]); it is consistent with our finding that ϕ's for coding and noncoding regions within a genome are nearly equal, which is to say that to the extent segmental duplication affects ϕ, it has no preference in what region it acts on; it is consistent with the notion that mutation events in genomes are predominantly neutral [48,49]. The model has only three parameters: L_0, the length of a random sequence taken to be the initial state of the sequence; $\bar{d}$, the average length of the duplicated segments; and r_μ, the rate of point mutations per length administered after—this feature is chosen for simplicity—the completion of growth. Owing to the random features in the model, high diversity among sequences is guaranteed but each sequence is deficient in intrasequence compositional heterogeneity. Long-range correlation [6–11], closely related to long-range intragenomic compositional heterogeneity, can be accounted for when a suitable proportion of the duplication events are made tandem [11]. Here, for simplicity, tandem duplication is left out; it has only a small effect on a whole-genome ϕ.

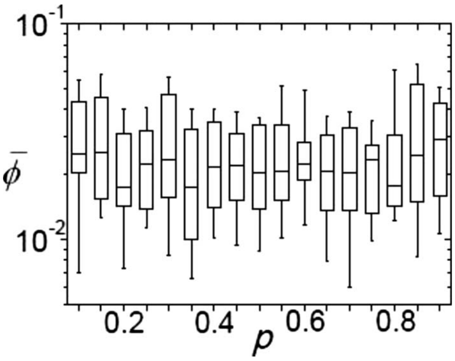

FIG. 8. Order index versus p of model sequences generated from a minimum growth model [46]. Lengths of the sequences are approximately 2 Mb and each p-interval bin gives the result of 20 sequences.

The results shown in Fig. 8 are those of sequences generated from a single set of parameters: $L_0=64$ b, $\bar{d}=1000$ b, and $r_\mu=0.73\ \mathrm{b}^{-1}$. These values were neither chosen nor adjusted for this study, but were previously set in a separate study of statistical properties of genomes—including the "Shannon information" [20]—not directly related to ϕ. The ϕ's of all the model sequences are seen to congregate around ϕ_g. Averaged over all model sequences, $\phi_{model}=0.022^{+0.023}_{-0.011}$ and $I_{model}=0.680 \pm 0.013$. The model ϕ's are essentially independent of sequence length and depend mildly on $\bar{d}$ and r_μ (within reasonable ranges) but sensitively on L_0. We find that it is not difficult to get sequences have ϕ_{model} congregating around *some* common ϕ' provided the dominant growth process is segmental duplication and the sequences have a common L_0 and similar $\bar{d}$ and r_μ. Hence we say that yielding $\phi_{model} \sim \phi'$ is a *robust* property of the growth model.

Is ϕ_g a fixed point of genome growth dynamics? Given that genome-length sequences can easily have $\phi \ll \phi_g$ and $L_{eq} \gg L_{eq,g}$, it is worthwhile asking how genomes were driven to have their ϕ's congregate in the small region near ϕ_g. One possibility is to view ϕ_g as a fixed point—more appropriately, a fixed basin, in ϕ space—of a robust dynamical process of genome growth and evolution. (We emphasize that for every sequence length this fixed point in ϕ space maps to many and possibly huge regions in sequence space which however sum to a minute portion of the entire sequence space.) As an analogy, we may view $\phi^{\{ran\}} \sim L^{-1/2}$ [Eq. (2)] as the fixed point of the processes of random point mutation and/or random base-by-base growth. The discussion in the last section suggests random segmental duplication, which can robustly drive genomic ϕ's to some fixed point ϕ', as a strong candidate for the dominant mechanism of genome growth. Other mechanisms, such as selection-driven point mutations and small indels, necessarily contributed to genome evolution, but these by themselves are not robust enough for driving genomic ϕ toward ϕ_g and cannot explain the ϕ's of the noncoding parts, nor can they yield as high a rate for growth and evolution as segmental duplication does [43–45]. All of this seems to point to a deeper level of insight on the dynamics of genome growth: the *dominance* of random segmental duplication as a dynamical process for genome growth may itself be a product of natural selection. Once this selection is made the matter of driving genomes to the correct fixed point becomes the relatively easy task of settling on the correct detailed characteristics—represented by parameters in the growth model with appropriate values—so that the fixed point is indeed ϕ_g. In its turn this fixed point is selected because it characterizes necessary structural characteristics of long sequences having maximum information capacity. Robustness aside, another important, perhaps even decisive, factor for natural selection to choose this path is its inherent speed; the process is very fast because of its random nature. Of course, the actual acquisition of information content, as opposed to the building of information capacity, still needs to be carried out by the much slower processes of selection-driven point mutations and indels.

ACKNOWLEDGMENTS

This work was partly supported by Grants No. 96-2112-M-008-025 and No. 97-2112-M-008-013 from the National Science Council (ROC) and the Cathy General Hospital-NCU Collaboration Grant No. 97-CGH-NCU-A1.

[1] C. G. Langton, Physica D **42**, 12 (1990).

[2] J. P. Crutchfield and K. Young, in *Complexity, Entropy, and the Physics of Information*, edited by W. H. Zurek (Addison-Wesley, Redwood City, CA, 1990).

[3] M. Mitchell, P. T. Hraber, and J. P. Crutchfeld, Complex Syst. **7**, 89 (1993).

[4] S. A. Kauffman, *The Origins of Order, Self-Organization, and Selection in Evolution* (Oxford University, London, 1993).

[5] S. P. Davies, H. Reddy, M. Caivano, and P. Cohen, Biochem. J. **351**, 95 (2000).

[6] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[7] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 3730 (1993).

[8] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[9] N. E. Israeloff, M. Kagalenko, and K. Chan, Phys. Rev. Lett. **76**, 1976 (1996).

[10] Z. G. Yu, V. Anh, and K. S. Lau, Phys. Rev. E **64**, 031903 (2001).

[11] P. W. Messer, P. F. Arndt, and M. Lassig, Phys. Rev. Lett. **94**, 138103 (2005).

[12] K. W. Church and J. I. Helfman, J. Comput. Graph. Statist. **2**, 153 (1993).

[13] X. Lu, Z. Sun, H. Chen, and Y. Li, Phys. Rev. E **58**, 3578 (1998).

[14] N. Nagai, K. Kuwata, T. Hayashi, and S. Era, Jpn. J. Physiol. **51**, 159 (2001).

[15] T. Y. Chen, L. C. Hsieh, and H. C. Lee, Comput. Phys. Commun. **169**, 218 (2005).

[16] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil, Mol. Biol. Evol. **16**, 1391 (1999).

[17] B. L. Hao, H. C. Lee, and S. Y. Zhang, Chaos, Solitons Fractals **11**, 825 (2000).

[18] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994).

[19] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Halvin, C. K. Peng, and M. Simons, Physica A **273**, 1 (1999).

[20] H. D. Chen, C. H. Chang, L. C. Hsieh, and H. C. Lee, Phys. Rev. Lett. **94**, 178103 (2005).

[21] R. Rudner, J. D. Karkas, and E. Chargaff, Proc. Natl. Acad. Sci. U.S.A. **60**, 921 (1968).

[22] V. V. Prabhu, Nucleic Acids Res. **21**, 2797 (1993).

[23] S. J. Bell and D. R. Forsdyke, J. Theor. Biol. **197**, 51 (1999).

[24] C. H. Chang, L. C. Hsieh, T. Chen, H. D. Chen, L. Luo, and H. C. Lee, J. Bioinf. Comput. Biol. **3**, 587 (2005).

[25] P. Bernaola-Galvan, J. L. Oliver, and R. Roman-Roldan, Phys. Rev. Lett. **83**, 3336 (1999).

[26] W.-L. Fan, Master's thesis, National Central University, 2004 (http://sansan.phy.ncu.edu.tw/~hclee/rpr/FanWL2004.pdf).

[27] S. Hampson, D. Kibler, and P. Baldi, Bioinformatics **18**, 513 (2002).

[28] Rice Annotation Project Database (http://rapdb.lab.nig.ac.jp/).

[29] National Center for Biotechnology Information Genome Database (http://www.ncbi.nlm.nih.gov/).

[30] Order Index Database (http://sansan.phy.ncu.edu.tw/~kensinro/OrdInd/order_index.htm).

[31] P. Pevzner and G. Tesler, Genet. Res. **13**, 37 (2003).

[32] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier, Science **228**, 953 (1985).

[33] M. P. Francino and H. Ochman, Nature (London) **400**, 30 (1999).

[34] A. Nekrutenko and W. H. Li, Genome Res. **10**, 1986 (2000).

[35] G. Bernardi, *Structural and Evolutionary Genomics: Natural Selection in Genome Evolution* (Elsevier, Amsterdam, 2005).

[36] J. R. Lobry, Mol. Biol. Evol. **13**, 660 (1996).

[37] J. Mrazek and S. Karlin, Proc. Natl. Acad. Sci. U.S.A. **95**, 3720 (1998).

[38] J. M. Freeman, T. Plasterer, T. F. Smith, and S. C. Mohr, Science **279**, 1827 (1998).

[39] S. L. Salzberg, A. J. Salzberg, A. R. Kerlavage, and J. F. Tomb, Gene **217**, 57 (1998).

[40] Inverse Symmetry Database (http://sansan.phy.ncu.edu.tw/kensinro/Index.htm).

[41] First 1 m digits of π (http://www.exploratorium.edu/pi/Pi10-6.html).

[42] First 100 m digits of π (http://crd.lbl.gov/dhbailey/expmath/software/).

[43] M. Lynch, Science **297**, 945 (2002).

[44] J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler, Science **297**, 1003 (2002).

[45] L. Zhang, H. H. Lu, W. Y. Chung, J. Yang, and W. H. Li, Mol. Biol. Evol. **22**, 135 (2005).

[46] L. C. Hsieh, L. F. Luo, and H. C. Lee, AAPPS Bull. **13**, 22 (2003).

[47] M. J. Lercher, A. O. Urrutia, and L. D. Hurst, Nat. Genet. **31**, 180 (2002).

[48] M. Kimura, J. Mol. Evol. **16**, 111 (1980).

[49] Y. X. Fu and W. H. Li, Genetics **133**, 693 (1993).