Journal of Bioinformatics and Computational Biology
Vol. 3, No. 3 (2005) 587–608

Imperial College Press
www.icpress.co.uk

SHANNON INFORMATION IN COMPLETE GENOMES

CHANG-HENG CHANG*, LI-CHING HSIEH*,#, TA-YUAN CHEN*, HONG-DA CHEN*
LIAOFU LUO‡ and HOONG-CHIEN LEE*,†,§,¶

*Department of Physics and †Department of Life Sciences,
National Central University, Chungli, Taiwan, ROC

‡Physics Department, Inner Mongolia University, Hohot, China

§National Center for Theoretical Sciences, Hsinchu, Taiwan, ROC
¶hclee@phy.ncu.edu.tw

#Department of Ecology and Evolution, University of Chicago,
Chicago, ILL 60637, USA

Received 3 September 2004
Revised 11 November 2004
Accepted 12 November 2004

Shannon information in the genomes of all completely sequenced prokaryotes and eukaryotes are measured in word lengths of two to ten letters. It is found that in a scale-dependent way, the Shannon information in complete genomes are much greater than that in matching random sequences — thousands of times greater in the case of short words. Furthermore, with the exception of the 14 chromosomes of *Plasmodium falciparum*, the Shannon information in all available complete genomes belong to a universality class given by an extremely simple formula. The data are consistent with a model for genome growth composed of two main ingredients: random segmental duplications that increase the Shannon information in a scale-independent way, and random point mutations that preferentially reduces the larger-scale Shannon information. The inference drawn from the present study is that the large-scale and coarse-grained growth of genomes was selectively neutral and this suggests an independent corroboration of Kimura's neutral theory of evolution.

Keywords: Genomics; Shannon information; statistical analysis; molecular evolution; genome growth.

1. Introduction

Shannon information[1] has been widely used in many diverse fields related to information. In the study of information in DNA sequences, it has been applied, for instance, to sequence alignment[2] and to the discovery of the DNA motif.[3] But it seems not to have been applied to the field of comparative genomics. This could be for a number of reasons. The availability of a large number of completely sequenced genomes is a relatively recent phenomenon. The high heterogeneity of complete genomes may make comparison difficult. For instance, how is the 0.58 million bases (Mb) genome of *Mycoplasma genitalium* to be compared with the 3000 Mb genome

of *Homo sapiens*? Within a genome different sections such as coding and noncoding regions are thought to have varying amounts of information. What section should be used to represent the genome? There is also the question of Shannon information itself, which as a broadly defined concept may be applied in many different ways and a definitive way to use it for comparative genomics has not been established.

In this paper, we devise a method to measure the Shannon information in a complete genome relative to that in a matching random sequence and apply it to all extant prokaryotic and eukaryotic complete genomes. The method is scale-dependent and highly sensitive to the amount of repeats in the sequence. The results are surprisingly unequivocal. We find that in spite of the wide diversity of the genomes in length, base composition and internal structure, the Shannon information in complete genomes (relative to random sequences) is uniformly very large for shorter words, in a way so regular that all the studied genomes except one — that of the malaria causing protozoan *Plasmodium falciparum* — can be put into a single universality class defined by an exceedingly simple formula; the fourteen chromosomes of *Plasmodium* belong to a related but distinct small class. By inquiring into how these results could have possibly come about we arrive at a simple model for genome growth and discuss its implications.

2. Mathematical Background

2.1. *Shannon entropy and information*

Consider a set $\mathcal{F}$ of occurrence frequencies for τ types of events,

$$\mathcal{F} = \left\{ f_i \in \mathbb{N} \left| \sum_{i=1}^{\tau} f_i = L \right. \right\} \equiv \{f_i | L\}. \tag{1}$$

The Shannon's uncertainty,[1] or entropy, for the set is

$$H(\mathcal{F}) = -\sum_i (f_i/L) \log(f_i/L) \tag{2}$$

This quantity has maximum value $H_{\max} = \log \tau$ when all the occurrence frequencies are equal: $f_i = \bar{f} = L/\tau$. Shannon suggested the notion of information as a measure of decrease in uncertainty and there are many ways this notion may be applied. Here we are interested in cases when most of the f_i's are non-zero and for such cases we define a Shannon information (called *Divergence* in Gatlin[4]) in $\mathcal{F}$ as

$$R(\mathcal{F}) \equiv H_{\max} - H(\mathcal{F}) = \log \tau - H(\mathcal{F}). \tag{3}$$

2.2. *Relation to relative spectral width*

From a set of occurrence frequencies $\mathcal{F}$, we can construct a distribution $\mathcal{S} = \left\{ n_f \in \mathbb{N} | \sum_f f n_f = L \right\}$, where n_f satisfying $\sum_f n_f = \tau$ is the number of events with frequency f. If f is considered as light frequency — discrete in this case — and n_f as light intensity, i.e., number of photons, then $\mathcal{S}$ can be considered analogously

to a standard optical spectrum. We shall consider $\mathcal{S}$ and $\mathcal{F}$ as interchangeable and shall refer to either as a spectrum. In terms of n_f, the Shannon entropy is

$$H(\mathcal{F}) = H(\mathcal{S}) = -\sum_f (n_f f/L)\log(f/L).$$

Using the relations $\sum_i (n_f f/L) = 1$ and $\log\tau = -\log(\bar{f}/L)$ we can rewrite Eq. (3) compactly as

$$R(\mathcal{F}) = R(\mathcal{S}) = \sum_f (n_f f/L)\log(f/\bar{f}) \tag{4}$$

where, as before, $\bar{f}$ is the mean frequency. This form of $R(\mathcal{F})$ lends itself to be expressed in terms of important spectral properties of $\mathcal{F}$, especially when $\mathcal{F}$ is a well-defined unimodal spectrum. In that case, we write $f = \bar{f}(1+x)$ and expand $\log(1+x)$ in a power series in x to obtain

$$R(\mathcal{F}) = \sum_{n=2}^{\infty} (-1)^n \frac{1}{(n-1)n}\langle x^n\rangle, \tag{5}$$

where $\langle x^n\rangle$ is the quantity x^n averaged over the spectrum, or the nth moment of $\mathcal{F}$. The leading term on the right-hand-side of Eq. (5), $\langle x^2\rangle$, is just the square of the *relative spectral width*, σ, of $\mathcal{F}$, namely, the ratio of the standard deviation Δ of the occurrence frequency to its mean $\bar{f}$: $\sigma \equiv \Delta/\bar{f}$. Equation (5) is particularly useful when σ is small, and is further simplified when $\mathcal{F}$ is symmetric with respect to it mean. Then odd moments vanish and we have

$$R(\mathcal{F}) \approx \frac{\sigma^2}{2} + \frac{\sigma^4}{12} + \mathcal{O}(\sigma^6), \quad \text{(sym. unimodal)} \tag{6}$$

where the approximation $\langle x^4\rangle \approx \sigma^4$ was used. This expression gives one a heuristic understanding of the Shannon information in a unimodal spectrum: there is no information when the spectrum is extremely narrow, that is, when all types of events occur with almost the same frequency. Conversely, so long as $\sigma < 1$, the broader the spectrum the higher the Shannon information. We remark that our definition of Shannon information is not intuitively useful for cases when the occurrences concentrate in a few types of events. Such situations do not arise in the systems — complete genomes — we are here interested in.

2.3. *k-spectrum from a DNA sequence*

Consider now a single strand of DNA and view it as a linear text written in the four bases, or chemical letters, A, C, G, T. For a sequence of L nucleotides (nt) we denote by $\mathcal{F}_k$ the set of occurrence frequencies $\{f_i|L\}_k$; the notation is the same as used in Eq. (1) except that here, the extra subscript k signifies that f_i is the occurrence frequency of the ith k-letter word, or (overlapping) k-mer, in the sequence. In

this study we are not interested in the order of the k-mers. The frequencies are obtained by sliding a window of width k across the genome, one letter at a time, and recording the number of times each k-mer is seen through the window.[5,6] (For simplicity, we treat the genome as being circular, which is not always true but true for many microbial genomes. Otherwise f_i should sum to $L - k + 1$ instead of L. The difference is negligible in any case because L is of the order of 1 million and in many cases much greater whereas $k \leq 10$.) Given $\mathcal{F}_k$ we can construct a *k-spectrum* giving n_f, the number of k-mers occurring with frequency f. The number of event types is now $\tau = 4^k$, so f_i and n_f satisfy the sum rules $\sum_i 1 = \sum_f n_f = 4^k$ and $\sum_i f_i = \sum_f f n_f = L$, and the mean frequency is $\bar{f} = 4^{-k}L$. To simplify language we will refer to $\mathcal{F}_k$ also as a k-spectrum. To insure good statistics we do not want k to be so large that $\bar{f}$ is less than one. Since the canonical size of microbial complete genomes is 2 Mb and 4^{10} is just over 10^6, the maximum k we consider in this study is 10.

2.4. *Shannon information in random sequence*

The k-spectrum $\mathcal{F}_k$ obtained from a random sequence $\mathcal{Q}$ with even base composition is a set of frequencies of random events of equal likelihood. If the mean frequency $\bar{f}$ is a very large number, which we assume to be the case, then $\mathcal{F}_k$ (more properly, $\mathcal{S}_k$) will be nearly a Poisson distribution with half-width $\Delta_{\text{ran}} = (b\bar{f})^{1/2}$, where $b = 1 - \tau^{-1}$ is a binomial factor. Thus the relative spectral width $\sigma_{\text{ran}} = (b\tau/L)^{1/2}$ falls off as $L^{-1/2}$ with increasing L and, from Eq. (6), $R(\mathcal{F}_k) \approx b\tau/2L$. That is, the Shannon information in a random sequence diminishes as $1/L$ with increasing L. This is but a simple manifestation of a well-known effect in statistics: the average of some measure of a random system gains sharpness as the system gains size, and achieves infinite sharpness in the large-system limit.

2.5. *n-replica and root-sequence*

There is a simple way for $\mathcal{Q}$ to grow and escape the large-system rule. Suppose we replicate $\mathcal{Q}$ n times to generate a sequence $\mathcal{Q}'$. We call $\mathcal{Q}'$ an *n-replica* of $\mathcal{Q}$ and $\mathcal{Q}$ a *root-sequence* of $\mathcal{Q}'$. If n is much less than L, then to a high degree of accuracy, the set of occurrence frequencies for k-mers in $\mathcal{Q}'$ is $\mathcal{F}'_k s = \{nf_i | nL\}_k$. Then $\bar{f}$ and Δ for the k-spectrum of $\mathcal{F}'_k s$ will both increase by a factor of n, hence its relative spectral width will remain unchanged. Thus, although $\mathcal{Q}'$ is n times longer than $\mathcal{Q}$, the Shannon information in $\mathcal{F}'_k$ for any k will be the same as that in $\mathcal{F}_k$, instead of being n times smaller. Conversely, the Shannon information in $\mathcal{Q}'$ is n times greater than that in a random sequence having the same length as $\mathcal{Q}'$.

2.6. *Random mutation and homologous insertion*

We thus have the notion of replication as an undesigned way for a sequence to gain length *and* "gain" Shannon information. Here, gaining means not losing in

absolute magnitude, as compared to the change in a random sequence when it gains length. Replication is a special case of a general way of gaining length by insertions of homologous segments. The latter is the last step in a common mode of mutation known as replicative transposition, where a segment of the genome is first copied and then inserted back into the genome at another site. Whereas a random mutation would generally decrease the Shannon information in a sequence, replicative transposition is an exception.

3. A First Look at Genomes

3.1. *Length and base composition of genomes*

Genomes vary greatly in their "profiles" — lengths and base compositions. An empirical fact is that genomes are almost always compositionally self-complementary, meaning that on a single strand the numbers of A's and T's are approximately equal, as are the numbers of C's and G's. Therefore, for simplicity, we characterize the base composition of a genome by a single number, p, the percentage content of (A+T). In the complete genomes or chromosomes of genomes studied in this work, the length spans a range of about 0.2 to 300 million base pairs and p spans a range of about 0.25 to 0.82 in complete genomes. We say two sequences match if they have the same profile.

3.2. *A view of genomic and random k-spectra*

The black curve in Fig. 1 is the 6-spectrum of the genome of the $p \approx 0.5$ hyperthermophile *Pyrobaculum aerophilum*,[7] with the occurrence frequencies of the 6-mers

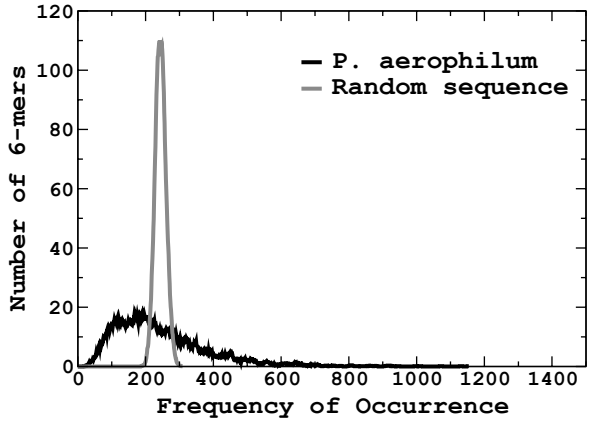

Fig. 1. 6-spectra of the genome of *P. aerophilum* (black) ($p \approx 0.5$) and its random match (gray). The frequencies have been normalized to that of a 1 Mb sequence. For better viewing only, the large fluctuation in the actual spectra have been smoothed out by forward and backward averaging, hence ordinates n_f need not be integers.

normalized to correspond to a 1 Mb sequence. The spectrum as shown has been smoothed out by forward and backward averaging (over twenty-one frequencies). Without this averaging, the spectrum has large fluctuations that would obscure the nature of its overall shape. In what follows, all computations are based on the real, not the smoothened, spectrum. The gray curve in Fig. 1 shows the 6-spectrum of the random match of the genome, obtained by thoroughly scrambling the genome of *P. aerophilum*. A random match can of course also be generated using a random number generator. When this is done a totally different sequence would be obtained which nevertheless would have a 6-spectrum practically identical to the gray curve in Fig. 1. (This is because a k-spectrum does not specify which k-mer has a certain occurrence frequency; it only specifies how many k-mers have frequency f.)

3.3. *Shannon information in a $p = 0.5$ genome*

Given a k-spectrum $\mathcal{F}_k$ we have from Eqs. (2) and (3) $H_{\max}(\mathcal{F}_k) = 2k \ln 2$. The Shannon entropy and information in the k-spectra, $k = 2$ to 10, of the genome of *P. aerophilum* and its random match are given in Table 1. The column under the heading R_{ex} gives the expected Shannon information in the k-spectrum of a random sequence:

$$R_{ex} = b'_k 4^k/2L, \quad b'_k = 1 - 1/2^{k-1}. \tag{7}$$

Here b'_k is used instead of the binomial factor $b = 1 - \tau^{-1}$ given previously. This is a semi-empirical value used to partly compensate for the fact that the random sequence is not completely random because (i) it is made to be approximately compositionally self-complementary (as most genomes are), and (ii) its percentage (A+T) content, or p, is fixed to be 0.5. Table 1 shows that R_{ex} is in excellent agreement with the actual Shannon information computed from a $p = 0.5$ random sequence.

Table 1. Shannon entropy H and information R in units of $\ln 2$ in the k-spectra of the genome sequence of *P. aerophilum* and its random match. R_{ex} is the expected information in the random match.

	Random match			*P. aerophilum*	
k	$H/\ln 2$	$R/\ln 2$	$R_{ex}/\ln 2$	$H/\ln 2$	$R/\ln 2$
2	3.9999	5.90 E−6	5.77 E−6	3.973	2.66 E−2
3	5.9999	3.72 E−5	3.46 E−5	5.933	6.65 E−2
4	7.9999	1.72 E−4	1.62 E−4	7.881	1.18 E−1
5	9.9993	7.26 E−4	7.53 E−4	9.821	1.79 E−1
6	11.999	2.94 E−3	2.90 E−3	11.75	2.74 E−1
7	13.988	1.18 E−3	1.17 E−3	13.66	3.35 E−1
8	15.955	4.78 E−2	4.71 E−2	15.53	4.69 E−1
9	17.798	2.02 E−1	1.88 E−1	17.26	7.33 E−1
10	19.408	5.92 E−1	5.24 E−1	18.59	1.41 E−0

We make several remarks concerning Table 1. (i) For both sequences the Shannon entropy is in every case very close to its maximum value, $2k \ln 2$. (ii) The Shannon information is very small, minuscule in the case of the smallest k's, compared with the Shannon entropy. That is, in most cases the Shannon information as defined in Eq. (3) is a tiny signal buried in a huge background. (iii) The ratio of the genomic Shannon information to its random match is very large for the small k's and decreases rapidly with increasing k. For instance, the ratio is about 4600, 100 and 2, respectively, at $k = 2$, 6 and 10. This, according to Eq. (6), implies that the spectral widths of the genomic k-spectra are about 68, 10 (see Fig. 1) and 1.4 times their random counterparts. We have tested this phenomenon on many $p \approx 0.5$ genomes and in every case the remarks made above apply substantially. We thus conclude that in so far as such sequences are concerned, our definition of Shannon information seems to be well suited for delineating genomes from random sequences.

3.4. *Reduced Shannon information*

We have seen that the Shannon information in genome and random sequences alike is a very small signal compared to Shannon entropy, but the Shannon information in a genome tends to be much larger than that in its random match. A better sense of the magnitude of the Shannon information in a sequence is obtained by measuring it relative to the Shannon information in the random match. Let $\mathcal{Q}$ be a genome sequence with $p \approx 0.5$, $\mathcal{F}_k$ be its k-spectrum and $\mathcal{F}'_k$ be the k-spectrum of the random match of $\mathcal{Q}$. From our discussion above we expect its k-spectrum to be unimodal, similar to the black curve in Fig. 1. We define a *reduced Shannon information* in $\mathcal{F}_k$ as the ratio of the Shannon information in $\mathcal{F}_k$ to that expected in $\mathcal{F}'_k$:

$$\mathcal{M}_R^{(0)}(\mathcal{F}_k) \equiv R(\mathcal{F}_k)/R_{ex}(\mathcal{F}'_k) = 2R(\mathcal{F}_k)\bar{f}/b'_k. \tag{8}$$

Obviously, if $\mathcal{Q}$ is itself a random sequence, then $\mathcal{M}_R$ is expected to be unity in any of its k-spectra.

3.5. *Case when genome is compositionally biased*

The situation is slightly more complicated for genomes with p deviating significantly from 0.5. Figure 2 shows the 6-spectra from the genome of *Chlamydia muridarum*[8] (black) and its random match (gray). Both have $p \approx 0.6$. Whereas the genomic spectrum is still unimodal, the random spectrum is composed of several sharp peaks. These are caused by the biased composition in the sequence. To see this, we denote by m-set the subsets $\mathcal{F}_{k,m}$ of k-mers with m (A+T)'s, $m = 0$ to k. Owing to the biased composition, the mean occurrence frequencies of the subsets $\mathcal{F}_{k,m}$ are spread out:

$$\bar{f}_m(p) = L\left(\frac{p}{2}\right)^m \left(\frac{q}{2}\right)^{k-m} = \bar{f}2^k p^m q^{k-m}, \tag{9}$$

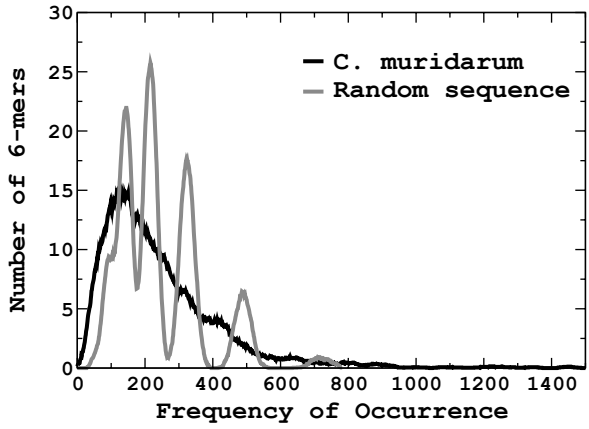

Fig. 2. 6-spectra of the genome of *C. muridarum* (black) and its random match (gray). The frequencies have been normalized to that of a 1 Mb sequence and, for better viewing the large fluctuation in the actual spectra have been smoothed out by forward and backward averaging.

where $q = 1 - p$ and $\bar{f} = L/4^k$. For $p = 0.6$, this gives $\bar{f}_m = 64$, 96, 144, 216, 324, 482 and 729, for $m = 0$ to k, respectively. Peaks at these positions, except the one at 64, are discerned in the spectrum for the random sequence shown in Fig. 2. (Notice that $\bar{f}_m(p)$ approaches $\bar{f}$ when p approaches 0.5.) The narrowness — because they are Poisson distributions with large means[9,10] — of the corresponding subspectra causes the k-spectrum of the random match to appear as the superposition of $k + 1$ separate sharp peaks as shown in the gray spectrum in Fig. 2. Apparently, for the genome the subspectra are sufficiently broad and overlapping such that no individual peak is discernible in its k-spectrum.

The overall variance (standard deviation squared) of the k-spectrum of a random sequence is determined by the spread of the subspectra which, when the widths of the individual subspectra are ignored, is given by

$$\begin{aligned}\Delta_k^2(p) &= \tau^{-1} \sum_m \tau_m \big(\bar{f}_m - \bar{f}\big)^2 \\ &= \bar{f}^{\,2} \big(2^k \left(p^2 + (1-p)^2\right)^k - 1\big). \end{aligned} \tag{10}$$

For $k = 6$, this gives 126 which is close to the width of 132 of the 6-spectrum of *C. muridarum* (normalized to 1 Mb). That is, the difference in Shannon information in the genome and its random match is no longer reflected in these widths. Rather, the difference lies in the widths of the subspectra of the m-sets. Table 2 gives the Shannon information in the subspectra of the m-sets in *C. muridarum* and in its random match. The measured Shannon informations (column 5) in the m-sets of the random match are close to their expected values $b'_k/\bar{f}_m$ (column 6). The values of the Shannon information in the genomic subspectra, in absolute magnitudes and

Table 2. Shannon information in the m-set of k-mers, $\mathcal{F}_{k,m}$, from the genome *C. muridarum* and its random match. Frequencies are normalized to that of a 1 Mb sequence. Eq. (15) is a universal formula given later in the text.

		R_{Cmur}		R_{random}	
$k,\ m$	$\bar{f}_m$	Measured	Eq. (15)	Measured	Expected
2, 1	60,000	1.96 E−2	2.00 E−2	2.88 E-6	4.17 E−6
3, 2	18,000	4.36 E−2	2.93 E−2	2.22 E-5	2.08 E−5
4, 2	3,600	8.18 E−2	7.18 E−2	1.94 E−4	1.21 E−4
5, 3	1,080	1.10 E−1	0.92 E−1	5.19 E−4	4.34 E−4
6, 3	216	1.53 E−1	1.84 E−1	2.98 E−3	2.24 E−3
7, 4	64.8	1.95 E−1	2.42 E−1	9.98 E−3	7.65 E−3
8, 4	13.0	2.84 E−1	4.77 E−1	5.82 E−2	3.83 E−2
9, 5	3.89	4.53 E−1	6.17 E−1	1.82 E−1	1.28 E−1
9, 7	8.75	3.91 E−1	2.74 E−1	7.97 E−2	5.70 E−2
10, 6	0.97	0.93 E−0	0.80 E−0	6.66 E−1	5.15 E−1
10, 8	2.62	6.87 E−1	3.55 E−1	2.87 E−1	1.98 E−1

relative to their respective random counterparts, are both similar to those seen in Table 1. Therefore we generalize the definition for $\mathcal{M}_R$ given in Eq. (8) to be the weighted average over the reduced Shannon information in the m-sets:

$$\mathcal{M}_R(\mathcal{F}_k) \equiv L^{-1} \sum_{m=0}^{k} L_m \mathcal{M}_R^{(0)}(\mathcal{F}_{k,m}), \tag{11}$$

where $\mathcal{M}_R^{(0)}(\mathcal{F}_{k,m})$ is as defined in Eq. (8), but with $\mathcal{F}_k$ replaced by $\mathcal{F}_{k,m}$ and $\bar{f}$ replaced by $\bar{f}_m$, and

$$L_m = 2^k(k,m)\bar{f}_m \tag{12}$$

is the number of k-mers in the m-set. Here (k,m) is the binomial satisfying $\sum_m (k,m)p^k(1-p)^{k-m} = 1$. The Shannon information in an m-set is given by Eq. (3) except that τ in the equation is replaced $\tau_m = 2^k(k,m)$, the number of types of k-mers in the m-set. [Note that $\sum_m L_m = L$, and $\bar{f}_m$ averaged over the m-sets gives $\bar{f}$:

$$\tau^{-1} \sum_m \tau_m \bar{f}_m = 4^{-k} \sum_m L_m = 4^{-k} L = \bar{f}, \tag{13}$$

which verifies that $\bar{f}$ is the mean frequency regardless of base composition.] In practice, to circumvent large fluctuations in $R(\mathcal{F}_{k,m})$ induced by small unevenness in the A/T (or C/G) contents — this can occur when $\bar{f}_m$ is very large at $k = 2$ and 3 — each frequency was divided by a factor $(2^k/p^m(1-p)^{k-m}) \prod_s p_s^{m_s}$, where m_s is the number of the sth type of base in the k-mer and $\sum_s m_s = k$.

3.6. *Tests with control sequences*

The reduced Shannon information [Eq. (11)] is defined such that its expected value for the k-spectrum of any random sequence is expected to be one, provided the length of the sequence is greater than 4^k. We test this with three sets of control sequences: a "random" set, a "century" set, and a "common-root" set. Sequences in the control sets are matches of sequences that form subsets — called targets — of genomes (see below) comprising 262 complete genomes: 135 prokaryotic complete genomes (the prokaryotes) and 127 complete chromosomes of 10 eukaryotes (the eukaryotes). The random set is comprised of 135 random matches of the prokaryotes. The century set is comprised of 135 100-replicas of random root-sequences that targets the 135 prokaryotes. In this case, for every target of length L and (base) composition p there is a sequence constructed by replicating one hundred times a random sequence of length $L/100$ and composition p. The common-root set is comprised of 262 replicas of 300 b random root-sequences that targets the combined 262 prokaryotes and eukaryotes. In this case, for every target of length L and composition p, there is a sequence — an $L/300$-replica — constructed by replicating $L/300$ times a random sequence of length 300 and composition p.

The diamond symbols in Fig. 3 give reduced Shannon information versus sequence length from the k-spectra, $k = 2$ to 10, of sequences in the three control sets. The figures in panels (A) and (B) have 1,215 data points each (135 sequences times nine k values). Panel (C) has about 2300 data points (262×9, excluding data for which genome length is less than 4^k). The $\mathcal{M}_R$ averaged over all sequences and all k's are as expected: 1.03 ± 0.12 and 101 ± 12 in panels (A) and (B), respectively. In (C), $\mathcal{M}_R$ is proportional to L as expected; the averaged value for $(300/L)\mathcal{M}_R$ is 1.02 ± 0.13. (The $\circ$ symbols in (C) are genome data; see below.) These results

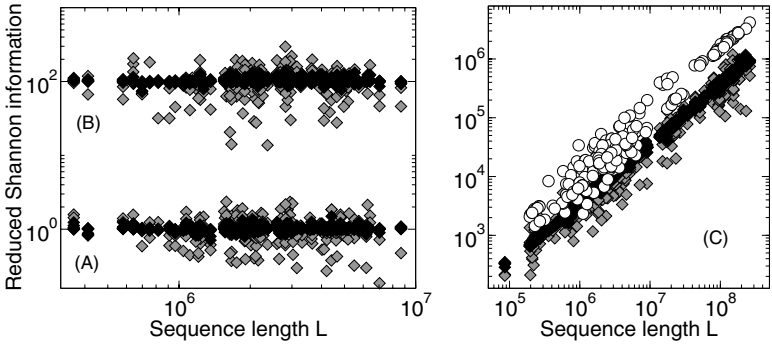

Fig. 3. Reduced Shannon information $\mathcal{M}_R$ in the k-spectra, $k = 2$ and 3 (gray diamonds) and 4 to 10 (black diamonds), of sequences in three control sets whose compositions are explained in the text. (A) The random set (135 sequences); $\mathcal{M}_{Rave} = 1.03 \pm 0.12$. (B) The century set (135 sequences); $\mathcal{M}_{Rave} = 101 \pm 12$. (C) The common-root set (262 sequences); $(300/L)\mathcal{M}_{Rave} = 1.02 \pm 0.13$. Also in (C) are the $\mathcal{M}_R$ (multiplied by a factor of 3) for $k = 2$ from the genomes (135 prokaryote and 127 eukaryotes).

gives us confidence in the normalization used in equations Eq. (8) and Eq. (11) for defining the reduced Shannon information.

4. Information in Whole Genomes

4.1. *Length and base composition of genomes*

Complete genome sequences used in the present study were downloaded from the genome FTP site of the (USA) National Center for Biotechnology Information. The 135 complete microbial genomes (the prokaryotes) were downloaded on October 9, 2003 from *ftp://ftp.ncbi.nih.gov/genomes/Bacteria/* and the 127 chromosome sequences of ten complete eukaryotic (the eukaryotes) were downloaded on July 15, 2003 from *ftp://ftp.ncbi.nih.gov/genomes/*. The ten eukaryotes (number of chromosomes in brackets) are *A. thaliana* (5), *C. elegans* (6), *D. melanogaster* (6), *E. cuniculi* (11), *H. sapiens* (24), *M. musculus* (21), *P. falciparum* (14), *R. norvegicus* (21; Chromosome Y missing), *S. cerevisiae* (16) and *S. pombe* (3). The prokaryotes are relatively homogeneous in length — 0.4 to 7 Mb — but highly heterogeneous in p — 26% to 0.75%. The reverse is the case for the eukaryotes where length ranges from 0.2 Mb (smaller chromosomes of *E. cuniculi*) to 268 Mb (*R. norvegicus* Chromosome I) and p ranges from 53% to 64%. The exception is *Plasmodium* whose p is $81 \pm 1\%$.[12]

4.2. *Shannon information in complete genomes*

The reduced Shannon information in the k-spectra of the 135 prokaryotes and 127 chromosomes of eukaryotes are color- (gray scale) and symbol-coded by organism and shown in Fig. 4(A), where each piece of datum gives the $\mathcal{M}_R$ in one

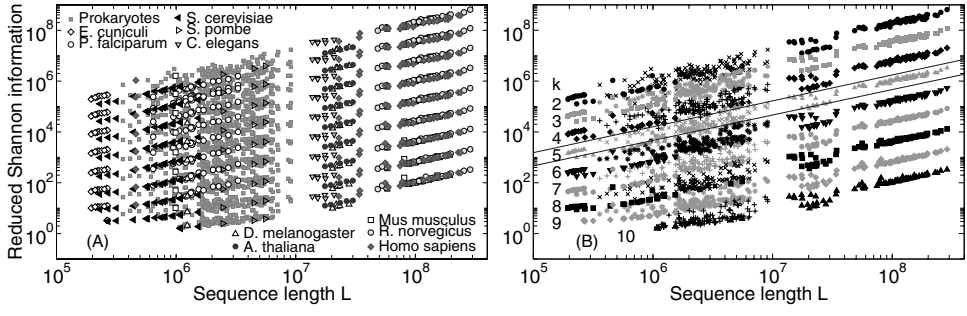

Fig. 4. Reduced Shannon information, $\mathcal{M}_R$, from 135 complete microbial genomes and 127 eukaryotes. Each symbol is the $\mathcal{M}_R$ value of one k-spectrum from one complete sequence. Left panel, $\mathcal{M}_R$ color-coded (gray scale) by organism; right panel, $\mathcal{M}_R$ color-coded by k, excluding data from 14 chromosomes of *P. falciparum*, where each "k-band" contains data from 248 complete sequences. Data have been multiplied by factor of 2^{10-k} to delineate the k-bands for better viewing. Data for which $4^k > L$, when $\mathcal{M}_R \approx 1$ regardless of sequence content, have been discarded. Straight lines in the plots are $\mathcal{M}_R \propto L$ lines.

k-spectrum of a sequence. The values of $\mathcal{M}_R$ in the figure have been multiplied by a factor of 2^{10-k} to partition data into different k groups for better viewing. The prokaryotic data are all shown as gray squares. Data for which sequence length is less than 4^k are deleted. For each organism, the data form separate k-dependent bands running diagonally across the figure, where bands for smaller k's give larger values of $\mathcal{M}_R$. The data from human (24 chromosomes), mouse (21 chromosomes) and rat (22 chromosomes) practically overlap when differences in sequence length is taken into account. Since relative to human chromosomal structure, there are large and numerous intra- and interchromosomal segment exchanges in the mouse and rat chromosome,[11] it is evident that Shannon information as applied in the present analysis is insensitive to whatever mutations that may have caused closely related organisms to diverge, from large chromosomal segment exchanges to gene-modifying point mutations. The data in Fig. 4(A) indicate the eukaryotes and the prokaryotes span a similar vertical range, about 2000 when the multiplicative factor of 2^{10-k} is removed. The only glaring exceptions to this similarity are the 14 chromosomes of the malaria causing parasite *Plasmodium falciparum*; they span a noticeably smaller vertical range of about 13. In Fig. 4(B) the data in (A) excluding those from *Plasmodium* are repeated and color-coded by k to highlight the well defined *k-bands*. Each band stretches over the full range of genome/chromosome length spanning three orders of magnitude. The two straight $\mathcal{M}_R \propto L$ lines, separated by a factor of 3.5 on the ordinate, are shown to give a sense of the linearity of a k-band and the vertical spread of the data within a band.

4.3. *Effective root-sequence length*

The linear relation between $\mathcal{M}_R$ and L implies that the *effective root-sequence length* $L_r(k)$, defined as $L_r(k) \equiv L/\mathcal{M}_R$, approximates a k-dependent but genome-independent constant. Table 3 gives the values for $L_r(k)$ extracted from $\mathcal{M}_R$ averaged over the prokaryotes (column 2), eukaryotes excluding *Plasmodium* (column 5),

Table 3. Effective root-sequence lengths L_r defined as length of sequence divided by reduced Shannon information.

	Prokaryotes				
k	Whole genome	Coding	Noncoding	Eukaryotes	*Plasmodium*
2	3.51 ± 2.17 E2	3.26 ± 1.91 E2	4.66 ± 3.74 E2	2.17 ± 1.26 E2	1.51 ± 0.35 E3
3	7.21 ± 3.91 E2	6.66 ± 3.37 E2	9.42 ± 6.36 E2	5.45 ± 2.98 E2	2.97 ± 0.22 E2
4	1.72 ± 0.84 E3	1.59 ± 0.73 E3	2.17 ± 1.25 E3	1.49 ± 0.74 E3	4.41 ± 0.30 E2
5	4.54 ± 2.10 E3	4.20 ± 1.87 E3	5.45 ± 2.85 E3	4.29 ± 2.00 E3	7.56 ± 0.56 E2
6	1.27 ± 0.57 E4	1.17 ± 0.51 E4	1.41 ± 0.70 E4	1.27 ± 0.55 E4	1.46 ± 0.12 E3
7	3.68 ± 1.67 E4	3.40 ± 1.52 E4	3.56 ± 1.76 E4	3.76 ± 1.51 E4	3.00 ± 0.28 E3
8	1.07 ± 0.49 E5	9.93 ± 4.48 E4	8.28 ± 4.01 E4	1.08 ± 0.40 E5	6.49 ± 0.68 E3
9	2.97 ± 1.34 E5	2.73 ± 1.22 E5	2.08 ± 0.61 E5	3.17 ± 1.05 E5	1.45 ± 0.16 E4
10	7.54 ± 2.95 E5	6.96 ± 2.66 E5	5.94 ± 0.37 E5	9.63 ± 2.97 E5	3.27 ± 0.40 E4

and the 14 chromosomes of *Plasmodium*. The prokaryotes and eukaryotes are very similar but the *Plasmodium* set is different. The meaning of $L_r(k)$ is this: if a genome has $L_r(k)$, then its reduced Shannon information (for k-mers) is the same as that in a random sequence of length $L_r(k)$, irrespective of the true length of the genome. This is to be compared with the Shannon information in a random sequence, which is proportional to the reciprocal of its length. In other words, if a genome of length L is x times $L_r(k)$, then the Shannon information in the genome is x times that in a random sequence of length L. From Table 3 we have $L_r(2)$, $L_r(6)$ and $L_r(10)$ being approximately 300 b, 13 kb and 800 kb, respectively. Hence the Shannon information in the 2-, 6- and 10-spectra of a genome approximately 2 Mb long is about 6700, 1500 and 2.5 times that of a 2 Mb random sequence matching the genome.

4.4. *Universality classes of genomes*

The data given in Table 3 are plotted as black symbols in Fig. 5: ▲ for prokaryotes, ■ for eukaryotes (*Plasmodium* excluded) and ▼ for sequences formed by concatenating the noncoding segments in prokaryotes. The relatively small standard deviation in $L_r(k)$ implies that there is a genome-independent, or universal, value for $L_r(k)$. These results are well summarized by the simple formula ($L_r(k)$ in units of bases):

$$\log L_r(k) = ak + B; \quad 2 \le k \le 10, \tag{14}$$

where $a = 0.410 \pm 0.030$ and $B = 1.58 \pm 0.19$.

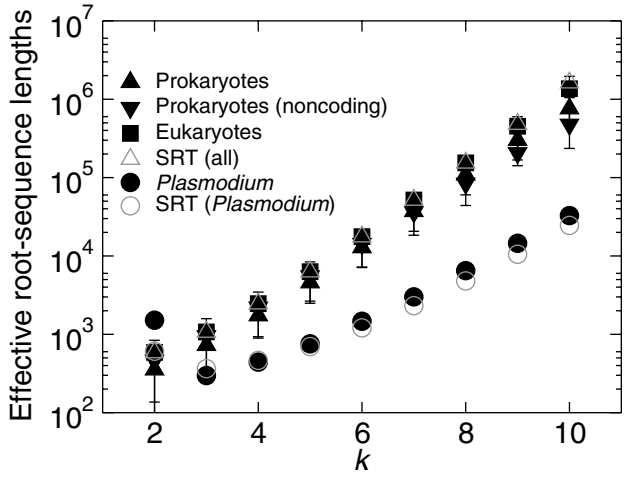

Fig. 5. Effective root-sequence lengths L_r. Each piece of data (with error flags) is obtained from averaging $L/\mathcal{M}_R$ over a k-band as seen in Fig. 4 (B). Black symbols are from genomic data: ▲, prokaryotes; ▼, noncoding regions in prokaryotes; ■, eukaryotes; ●, *Plasmodium*; Gray symbols show results obtained from model sequences: △, all model sequences except *Plasmodium*; ○, model sequence for Plasmodium.

We refer to Eq. (14) as a *universality class*, whose mean is given by the straight line in Fig. 5. (Gray symbols in Fig. 5 are results obtained from model sequences, to be discussed later.) The universality class expressed by Eq. (14) includes all the genomes studied except the fourteen chromosomes of *Plasmodium*, whose $L_r{}'s$ are shown as •'s in Fig. 5 (A). This small group forms a separate class given by the constants $a = 0.146 \pm 0.012$ and $B = 2.14 \pm 0.05$.

4.5. *A universal formula*

From Eq. (14), we extract a formula for the Shannon information in an m-set $\mathcal{F}_{k,m}$ of a genome sequence of composition p in the main class:

$$R(\mathcal{F}_{k,m}) \approx 0.012(1 - 2^{1-k})e^{0.44k}(2^k p^m (1-p)^{k-m})^{-1}. \tag{15}$$

When p approaches 0.5 the formula collapses to

$$R(\mathcal{F}_k) \approx 0.012(1 - 2^{1-k})e^{0.44k}. \tag{16}$$

This last formula gives not only the Shannon information in a genome sequence with $p \approx 0.5$, it also gives the weighted average (over the m-sets) of the Shannon information in any genome sequence in the main class. Note that Eq. (15) is independent of L and Eq. (16) is independent of both L and p. Equation (15) was used to produce the numbers given in column 4 of Table 2.

From the above and Eq. (6), we also obtain a formula for the relative spectral width for $\mathcal{F}_{k,m}$: $\sigma(\mathcal{F}_{k,m}) \approx (2R(\mathcal{F}_{k,m}))^{1/2}$ when the genome has $p \neq 0.5$, and $\sigma(\mathcal{F}_k) \approx (2R(\mathcal{F}_k))^{1/2}$ for the whole k-spectrum when $p \approx 0.5$. Note that $\sigma(\mathcal{F}_k)$ cannot be used as an estimate for the relative spectral width of the k-spectrum of a genome whose p deviates far from 0.5.

4.6. *Coding and noncoding regions*

About 85% of a prokaryote is comprised of coding regions, whereas coding regions typically occupy less than half of an eukaryotic chromosome. Generally, coding regions occupy a smaller the fraction the higher life form of the organism; coding regions make up less than 2% of the human genome. Columns 3 and 4 in Table 3 give the $L_r(k)$ for sequences obtained by concatenating the coding and noncoding segments, respectively, in prokaryotes. There is a small difference in the two sets of data but, on the level of accuracy maintained in the present discussion, on the whole one may infer that no essential difference in $\mathcal{M}_R$ between coding and noncoding regions obtains.

This is not to say that statistical sequence similarity between coding and noncoding sections is so great that no difference in Shannon information between them may be measured. Quite the contrary. But there are several reasons why such a difference tend not show in $\mathcal{M}_R$ for the *whole* genome. First, most genes are protein genes and they are coded in three-letter codons. This implies that the greatest difference between a coding and a noncoding segment will be detected when the

sliding window used to count word frequencies slides three letters at a time. Our sliding window slides one letter at a time. Second, differences between coding and noncoding regions tend to cancel when viewed over the whole genome. An example is the compositional self-complementarity on a *single* strand of a genome, in spite of the fact that, as a rule, the contents of complementary bases in coding regions are different. The reason that the difference cancels out over the entire strand is because coding regions are more or less uniformly distributed on *both* strands, such that on a single strand, there are as many positively oriented genes as there are negatively oriented genes. Consequently, on a single strand the excess (if there is any) in A's in genes in one orientation will approximately be equal to the excess in T's in genes in the opposite orientation.

5. Interpretation of Results

5.1. *Duplications increases $\mathcal{M}_R$ uniformly*

The existence of universality classes in reduced Shannon information implies that the latter is a signature in complete genomes undiminished by the enormous diversity in growth and evolution experienced by individual genomes. Since it is easy to show that most biologically plausible models for genome growth and evolution do not generate any class, even less so the observed universality classes, the existence of the universality classes and their precise form provide powerful constraints on models for genome growth and evolution. Our experience with robust signals in systems composed of highly diverse members suggests a growth process in which stochasticity plays a strong role.

The very large amount of reduced Shannon information in complete genomes, at least for the shorter k-mers, is consistent with the hypothesis that genomes contain very large amounts of duplications. The $k = 2$ band of genomic data in Fig. 4(B) is reproduced as $\circ$'s in Fig. 3(C). It is extremely similar to the band of data (black and gray $\diamond$'s) obtained from the common-root set of sequences composed of n-replicas made from replicating random root-sequences 300 b long. The fact that 300 b is close to the value of $L_r(2) \approx 300$ b of the main universality class hints at the possibility that genomes are to a large extent n-replicas with a common root-sequence length of about 300 b. However, the $\mathcal{M}_R$ from n-replicas lacks the clear k-dependence seen in the genome data and this rules out the possibility that genomes are simple n-replicas. Some other mechanism is needed to generate the observed k-dependence in $\mathcal{M}_R$.

5.2. *Point mutations decreases $\mathcal{M}_R$ differentially*

An obvious candidate that may generate the observed k-dependence are small mutations. For simplicity, we consider the effect of random point replacements on a k-spectrum of an n-replica. Suppose d is the average distance between two adjacent mutation sites. When the total number of mutations is very small, $d \gg 10$ (10 is the

maximum k in the present study), the effect of the mutations on the k-spectrum will be negligible to give $\mathcal{M}_R \approx n$. Conversely, when the number of mutations is very large, $d \ll 1$ and all traces of replication in the n-replica will be obliterated reducing the n-replica to a random sequence yielding $\mathcal{M}_R \approx 1$. In between, when d is of the order of k, the mutation will affect the k-spectra in such a way that the $\mathcal{M}_R$ in a k-spectrum of a larger k will suffer a higher degree of reduction. Presumably, given an n-replica, there may be an appropriate number of mutations whose effect is to generate a k-dependence in $\mathcal{M}_R$ similar to that observed in Fig. 4.

6. Model for Genome Growth

6.1. *A minimal model*

Based on the above considerations, we devised a number of simple growth models having the two main ingredients: a large number of random segmental duplications to create large values for $\mathcal{M}_R$; a suitable number of random point replacements to generate the observed k-dependence in $\mathcal{M}_R$. In addition, the model must have the flexibility allowing the growing genomes to diverge at any stage and the robustness to prevent the Shannon information from depending on the diverging events. Here we report the results obtained from a stochastic replicative transposition (SRT) model in which an initial random sequence of length L_0 is grown to full length via duplications of randomly selected segments (in the sequence) of random lengths that are then reinserted into the sequence at randomly selected sites.[10] After full growth, the sequence is subjected to random point replacements at a rate of r mutations per nucleotide. The replacements have the same compositional bias as the target sequence. Having the mutations all occur after the completion of growth does not necessarily reflect the actual workings of nature; indeed there is an infinite number of ways single mutations may be admixed with duplications. Rather the scheme is adopted in this paper simply to limit the number of parameters in the model.

The lengths l of the duplicated segments are given by a distribution on which the results have a weak dependence. Here we simply use a square distribution having the range $1 \leq l \leq l_x$. A χ^2 procedure based on comparing empirical values of $L_r(k)$ with those computed from a set of twenty model sequences that match twenty randomly selected prokaryotic genomes was used to determine optimal values for the parameters L_0, r and l_x. The χ^2 is observed to have a strong dependence on L_0 favoring very short initial sequence lengths and weaker dependence on l_x and r. We find that the best results for the prokaryotes are obtained when $L_0 = 8$, $l_x = 250$ and $r = 0.95$ (details of this search will be reported elsewhere). The initial sequences are compositionally self-complementary but otherwise random. Hence an $L_0 = 8$ sequence can only have $p = 0$, 0.25, 0.5, 0.75 or 1.0. Because p and 1-p sequences in our model are mathematically equivalent, the initial sequences are chosen to have $p = 0.25$ or 0.5. Two measures was taken to shorten computation time, neither of which is expect to qualitatively affect the presented results. Firstly, because

$l_x \gg L_0$, an initial sequence is first replicated to a length just greater than l_x before it is subjected to growth by stochastic segmental duplication. Secondly, for model eukaryote sequences, l_x is taken to be 10,000 once the sequence grows beyond 2 Mb.

6.2. *Results from model*

Using the optimal parameters ($L_0 = 8$, $l_x = 250$ and $r = 0.95$) we generated 248 model sequences whose profiles more or less match those of the genomes/chromosomes in the main universality class and computed $\mathcal{M}_R$ and $L_r(k)$ for the model sequences. The $\triangle$ in Fig. 5 summarize results for $L_r(k)$. Each symbol in the figure is obtained by averaging over 248 sequences; standard deviations from the mean are given by the error flags. It is fair to say that the extremely simple model accounts for the k-dependence and universality of the data very well. A general property of sequences generated by the model is that a correct value for $\mathcal{M}_R$ of a k-spectrum guarantees a correct shape for that spectrum.[10] The plotted 5-spectra in Fig. 6, where the spectra from the model sequences are given in light gray and those from three genome sequences in black (dark gray curves are from the random matches) indicate the typical agreement between model and genome spectra. All curves have been smoothed by forward and backward averaging for better viewing; the value shown at each frequency is the average over twenty-one

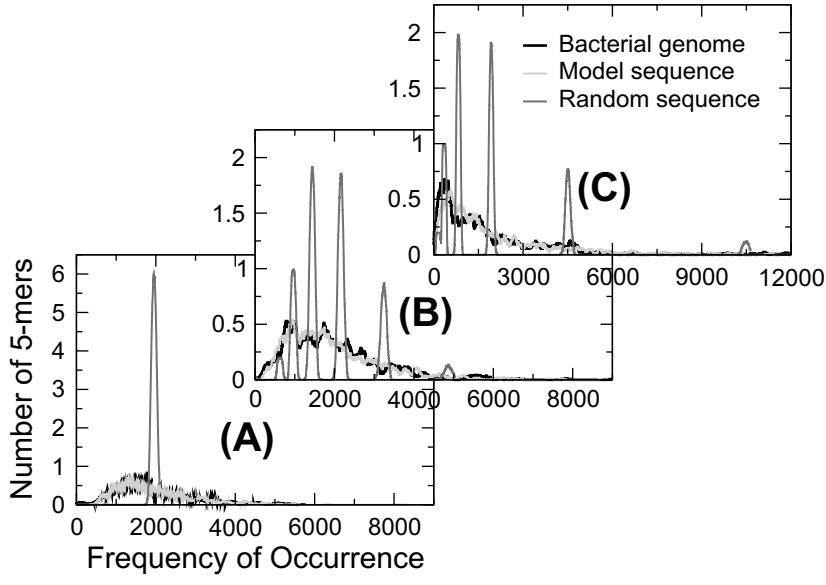

Fig. 6. Comparison of 5-distributions of genome (black), random (dark gray) and model (light gray) sequences with $p = 0.5$ (A), 0.6 (B) and 0.7 (C), respectively. The genomes are *A. fulgidus* (A), *S. pneumoniae* (B) and *C. acetobutylicum*(C). All curves have been smoothed by forward and backward averaging.

frequencies. Without averaging, the genomic spectra will show large fluctuations whereas the model will show much smaller fluctuations. This is an indication that textually, the genome is much rougher, or inhomogeneous, than what our simple model produces. We emphasize that it is not a trivial task to generate a sequence whose k-spectra are genome-like for all k's; it is far easier to generate sequences that do not have the observed properties of genomes than it is the opposite.

The 14 model *Plasmodium* chromosomes are similarly generated as the main group except that $L_0 = 80$ and $r = 0.20$. The results are shown as $\circ$ in Fig. 5. On the surface, the larger L_0 and smaller r for *Plasmodium* suggest that, compared to other organisms studied, this organism experienced either less duplication or significantly fewer point (or small) mutations per site, or both, than genomes in the main class. The real cause for the distinctiveness of *Plasmodium* may be far more complex. Among the eukaryotes studied, *Arabidopsis*, which belongs to the main class, is phylogenetically the least remote from *Plasmodium*.[12,13] It will be interesting to see how closer taxonomic relatives of *Plasmodium*[13] are classified by $\mathcal{M}_R$.

7. Discussion

7.1. *Universality in diversity*

Our main findings concerning Shannon information in complete genomes revealed two important facts: (i) for short k-mers Shannon information in complete genomes is uniformly very large, even enormous; and (ii) the Shannon information in complete genomes unequivocally exhibits a universality that coexists with the huge diversity of species. We have found a simple, coarse-grain model for genome growth and evolution that can account for both phenomena: very early on, when they were much less than 300 b long, genomes started to grow mainly by stochastic segmental duplication followed by (or admixed with) small mutations. The model allows a genome to diverge at any stage during its growth such that, in principle, all the genomes studies could have had a single common ancestor. The simplicity of the model and the maximally stochastic nature of the growth mechanisms may underlie the robustness of the results and explain the emergence of the universality classes in the presence of a huge diversity of species. As a computational device, the compositional bias and complementarity in the model sequences are generated by the bias in the replacement mutations. The proposed model should be viewed as a crude prototype for a realistic model for genome growth and evolution. In particular, it does not explain the origin of compositional bias. The model will need to be refined when it is confronted with finer textual details in the genome.

7.2. *Why is Plasmodium different?*

We need to examine the data and our model in greater detail to ascertain whether the genome *Plasmodium* is truly fundamentally different from all other genomes. In

particular, in view of the fact that the genome of *Plasmodium* has the most biased base composition among all completed genomes, we need to conduct a detailed study of the p-dependence of $\mathcal{M}_R$. The case of *Plasmodium* raises several questions: (i) Why is the $\mathcal{M}_R$ of *Plasmodium* different? (ii) (If *Plasmodium* is truly different then) Are there other organisms in the *Plasmodium* class? (iii) Are there more than the two universality classes reported here in existence? (iv) What are the biological causes of different classes?

[**Note added in revision**. We have noticed that *Plasmodium* becomes less anomalous and conspicuous if the normalization factor of $R_{ex}(\mathcal{F}'_k)$ is removed from the definition of $\mathcal{M}_R^{(0)}(\mathcal{F}_k)$ in Eq. (8). This considerably lowers the significance but does not eliminate the *Plasmodium* problem while not changing the observed universality and our proposed explanation of it. Details of this finding will be reported elsewhere.]

7.3. *Neutral theory of evolution*

Whereas the complete genomes studied vary greatly in coding regions as a percentage of the whole genome (from 85% in microbes to less than 2% in *H. sapiens*), the universal genome property reported here seems not to depend on that percentage. Indeed we have shown that in prokaryotes, there is no discernible difference between the reduced Shannon information of the coding and noncoding regions (Fig. 5). In the context of our growth model, our findings appear to imply that the majority of the individual fixed duplications and replacements during genome growth do not act differently in the two regions. If we assume that coded words other than genes such as binding sites, regulatory signals, and microRNA's[14] collectively do not occupy a dominant portion of the noncoding region in eukaryotes, then we may assume that the fixed events in the noncoding region were selectively neutral and hence, by inference, so were essentially all the fixed events. This notion of selective neutralism, based as it is on the present whole-genome analysis, seems to independently corroborate Kimura's neutral theory of molecular evolution,[15,16] a theory that was based on the investigation of polymorphisms of genes.

7.4. *Genomes are rich in duplications*

Independent from our contention that large Shannon information in a genome suggests a large amount of random duplications over the entire genome, there are many other evidence of duplications in genomes: the existence of many transposable elements; the large amounts of repeats in both prokaryotes[17] and eukaryotes[18–21]: the preponderance of paralogs (genes) and pseudogenes in all life forms;[22–24] chromosome segmental rearrangements that seem to characterize mammalian[11] and plant[25] radiations. Our proposed growth model may at least be taken as a starting point for an explanation of all these phenomena.

7.5. *Random segmental duplication as a result of natural selection*

We have learned from this study that the reduced Shannon information ($\mathcal{M}_R$) in a genome increases when it adds homologous sequence to itself. Hence stochastic duplication is a highly efficient process for a sequence to increase its $\mathcal{M}_R$ in a non-directed fashion. Lifeless random segmental duplication may have eased the path to the rise of life. A larger $\mathcal{M}_R$ implies a wider distribution of occurrence frequencies of oligonucleotides and the consequential concomitant rapid appearance of large numbers of over- and under-represented oligonucleotides, which would make easier — there will be less entropic resistance — the task of endowing some such oligonucleotides with biological meaning by natural selection at a later date. Random segmental duplication also makes good evolutionary sense after the rise of the earliest codes. For sometimes such duplications will copy a segment in which is embedded a coded sequence, say a proto-gene, which can later evolve by natural selection into a new gene in the host genome. This mode of generating new genes will be enormously faster than having a new gene evolved entirely from scratch, and may provide a basis for explaining why genes have been duplicated at such a high rate,[23] perhaps up to about 1% per gene per million years,[18] and evidently causing the human genome to expand by 15 to 20% in the last fifty million years.[21] Thus having random segmental duplication as a major mode of genome growth makes the rapid rise and evolution of life easier to understand, and may itself be a consequence of natural selection. This is consistent with the propositions that a growth strategy with a reliance on duplication may have the effect of enhancing the rate of evolution.[26,27]

Acknowledgments

This work is supported in part by the grants 92-2119-M-008-012 and 93-2311-B-008-006 from the National Science Council, ROC.

References

1. Shannon CE, A mathematical theory of communication, *Bell Sys Techn J* **27**:379–423; 623–656, 1948.
2. Clote P, Backofen R, *Computational Molecular Biology*, John Wiley & Sons, 2000.
3. Pesole G, Attimonelli M, Sacconne C, Linguistic approaches to the analysis of sequence information, *Trends Biotech* **12**:401–408, 1994.
4. Gatlin LL, Information theory and the living system, Columbia University Press, 1972.
5. Karlin S, Campbell AM, Mrazek J, Comparative DNA analysis across diverse genomes, *Annu Rev Genet* **32**:185–225, 1998.
6. Hao BL, Lee HC, Zhang SY, Fractal related to long DNA sequences and complete genomes, *Chaos, Solitons and Fractals* **11**:825–836, 2000.
7. Fitz–Gibbon ST *et al.*, Genome sequence of the hyperthermophilic crenarchaeon, *Pyrobaculum aerophilum*, *PNAS*, **99**:984–989, 2002.
8. Read TD *et al.*, Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae AR39*, *Nucl Acids Res* **28**:1397–1406, 2000.

9. Xie HM, Hao BL, Visualization of K-tuple distribution in prokaryote complete genomes and their randomized counterparts, *IEEE Proc Comp Sys Bioinf* 31–42, 2003.
10. Hsieh LC, Luo LF, Ji FM, Lee HC, Minimal model for genome evolution and growth, *Phys Rev Lett* **90**:018101–018104, 2003.
11. O'Brien SJ *et al.*, The promise of comparative genomics in mammals, *Science* **286**:458–481, 1999.
12. Gardner MJ *et al.*, Genome sequence of the human malaria parasite *Plasmodium falciparum*, *Nature* **419**:498–511, 2002.
13. Baldauf SL *et al.*, A kingdom-level phylogeny of eukaryotes based on combined protein data, *Science* **290**:972–977, 2000.
14. Ambros V, MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing, *Cell* **113**:673–676, 2003.
15. Kimura M, Evolutionary rate at the molecular level, *Nature* **217**:624–626, 1968.
16. Kimura M, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1983).
17. Jensen LJ *et al.*, Three views of microbial genomes, *Res Microbiol* **150**:773–777, 1999.
18. Lynch M, Conery LC, The evolutionary fate and consequences of duplicate genes, *Science* **290**:1151–1155, 2000.
19. Lander ES *et al.*, Initial sequencing and analysis of the human genome, *Nature* **409**:860–921, 2001.
20. Venter JC *et al.*, The sequence of the human genome, *Science* **291**:1304–1351, 2001.
21. Liu G *et al.*, Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome, *Gen Res* **13**:358–368, 2003.
22. Otto S, Yong P, The evolution of gene duplicates, *Adv Genetics* **46**:451–483, 2001.
23. Meyer A, Duplication, duplication, *Nature* **421**:31–32, 2003.
24. Gu Z *et al.*, Role of duplicate genes in genetic robustness against null mutations, *Nature* **421**:63–66, 2003.
25. Grant D, *et al.* Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soyabean and *Arabidopsis*, *Proc Natl Acad Sci USA* **97**:4168–4173, 2000.
26. Yanai I *et al.*, Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification, *Phys Rev Lett* **85**:2641–2644, 2000.
27. Zhang YX *et al.*, Genome shuffling leads to rapid phenotypic improvement in bacteria, *Nature* **415**:644–646, 2002.

Chang-Heng Chang was a member of the Computational Biology Lab and a Master's student at the Physics Department, National Central University, Taiwan ROC. He is now working in the information industry in Taiwan.

Hong-Da Chen is a member of the Computational Biology Lab and a PhD student at the Physics Department, National Central University.

Ta-Yuan Chen just obtained his PhD degree from the Physics Department, National Central University, Taiwan ROC. He is now a researcher with the cancer Research Center, Hoshin Cancer Hospital, Taipei, Taiwan, ROC.

Li-Ching Hsieh recently obtained his PhD degree from the Physics Department, National Central University, Taiwan. He is now a posdoctoral research fellow at the Department of Ecology and Evolution, University of Chicago.

Liaofu Luo is a professor with the Physics Department, Inner Mongolia University, Hohot, China. His main research interest is biological sequence analysis and theoretical biology.

Hoong-Chien Lee is a University Distinguished Professor jointly appointed with the Physics Department and the Life Sciences Department, National Central University, Taiwan. He works on a variety of topics in computational biology, including molecular evolution.