ABSTRACT
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 - thousands of times in the case of short words - than that in matching random sequences. Furthermore, the Shannon information in all available complete genomes belong to two universality classes given by an extremely simple formula. *Plasmodium* alone belongs to one class and all the other genomes belong to the other class. The data are consistent with a model for genome growth composed of two main ingredients: random homologous 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 is that the large-scale and coarse-grain growth of genomes was selectively neutral and this suggests an independent corroboration of Kimura’s neutral theory of evolution.

Contact: hclee@phy.ncu.edu.tw

INTRODUCTION
Shannon information [1] has been widely used in many diverse fields related to information, including the study of information in DNA sequences, in particular in sequence alignment [2]. 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 non-coding 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.

MATHEMATICAL BACKGROUND
Shannon entropy and Shannon information
Consider a set of occurrence frequencies, \( \mathcal{F} = \{f_i|\sum_{i=1}^{\tau} f_i = N\} \equiv \{f_i/N\} \), for \( \tau \) types of events. Shannon’s uncertainty [1], or entropy, for the set is

\[
H(\mathcal{F}) = - \sum_i (f_i/N) \log(f_i/N)
\]

(1)

This quantity has maximum value \( H_{\text{max}} = \log \tau \) when all the occurrence frequencies are equal: \( f_i = \bar{f} = \tau^{-1}N \). Shannon expressed the information in a system in terms of decrease in uncertainty and there are many ways
to apply this notion of information. 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 [3]) in \( F \) as

\[
R(F) \equiv H_{max} - H(F) = \log \tau - H(F)
\]  

(2)

**Relation to relative spectral width**

From a set of occurrence frequencies \( F \) we can construct a distribution \( S = \{n_f|L\} \) where \( n_f \) is the number of events with frequency \( f \). The sum-rules \( \sum n_f = \tau \) and \( \sum f n_f = L \) are satisfied. If \( f \) is considered as light frequency - discrete in this case - and \( n_f \) as light intensity, i.e., number of photons, then \( S \) can be considered analogously to a standard optical spectrum. We henceforth call \( S \) a spectrum. In terms of \( n_f \) the Shannon entropy is

\[
H(F) = H(S) = -\sum_f (n_f/N) \log(f/N)
\]

There is a close relation between our definition of the Shannon information and the relative spectral width of \( S \) when the latter is a unimodal distribution. The relative spectral width \( \sigma \) of \( S \) is its half-width \( \Delta \) (or standard deviation) divided by its mean \( \bar{f} \) : \( \sigma = \Delta / \bar{f} \). The relation is

\[
R(F) = R(S) = \frac{\sigma^2}{2} + O(\sigma^3) \quad \text{(unimodal \( S \))}
\]

(3)

which is particularly useful when \( \sigma \) is small. We give two explicit examples. A histogram approximation of a unimodal distribution is to have a fraction \( x \) each of the events respectively have frequencies \( f = \bar{f} \pm (2x)^{-1/2} \), and the rest (fraction 1-2\( x \)) has mean frequency. Then

\[
R = \frac{\sigma^2}{2} - \frac{\sigma^4}{8} + O(\sigma^6) \quad \text{(histogram)}
\]

A Gaussian distribution with relative spectral width \( \sigma \) yields

\[
R = \frac{\sigma^2}{2} + \frac{\sigma^4}{4} + O(\sigma^6) \quad \text{(Guassian)}
\]

In general, an order \( \sigma^3 \) term will occur when the distribution is not symmetric around the mean frequency. Eq. (3) gives one a heuristic understanding of the Shannon information in a spectrum: there is no information when the spectrum is extremely narrow, or when all types of events occur with 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.

**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 \( F_k \) the set of occurrence frequencies \( \{f_i|L\}_k \), where \( f_i \) is the occurrence frequency of the \( i^{th} \) \( k \)-letter word, or \( k \)-mer. 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 [4, 5]. Given \( F_k \) we can construct a \( k \)-spectrum, \( S_k = \{n_f\}_k \), where \( n_f \) is the number of \( k \)-mers occurring with frequency \( f \). The number of event types is now \( \tau = 4^k \), so \( F_k \) and \( S_k \) satisfy the sum rules

\[ \sum_i 1 = \sum_j n_f = 4^k \]

and

\[ \sum_i f_i = \sum_j f n_f = L \]

and the mean frequency is \( \bar{f} = 4^{-k} L \). To simplify language we will refer to \( F_k \) 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.

**Shannon information in random sequence**

The \( k \)-spectrum \( F_k \) obtained from a random sequence \( Q \) 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 \( F_k \) (more properly, \( S_k \)) will be nearly a Poisson distribution with half-width \( \Delta_{ran} = (bf)^{1/2}, \)
where \( b = 1 - \tau^{-1} \). 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. (3), \( R(F_k) \approx b\tau/2L \). 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.

### \( n \)-replica and root-sequence

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

### 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.

### A FIRST LOOK AT GENOMES

#### Length and base composition of genomes

Genomes vary greatly in their 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.

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

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

#### 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* [6], with the occurrence frequencies of the 6-mers normalized to correspond to a 1 Mb sequence. The green curve in Fig. 1 shows the 6-spectrum of a random sequence, called a random copy of the genome, obtained by thoroughly scrambling the *P. aerophilum* genome. A random copy can of course also be generated using a random number generator. When this is done a totally different sequence would obtain but it would
have a 6-spectrum practically identical to the green 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.)

**Shannon information in a p=0.5 genome**

Given a k-spectrum $F_k$ we have from Eqs. (1) and (2) $H_{\text{max}}(F_k)=2k \ln 2$. The Shannon entropy and information in the k-spectra, $k=2$ to 10, of $P. aerophilum$ genome and its random copy 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:

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

(4)

Here $b'_k$ is used instead of the value $b=1-\tau^{-1}$ 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.

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, miniscule 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. (2) is a tiny signal buried in a huge background. (iii) The ratio of the genomic Shannon information to its random copy 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. (3), 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 qualitatively. 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.

**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 copy. We can therefore enhance the genomic signal by dividing it by the random signal. Let Q be a genome sequence with $p \approx 0.5$, $F_k$ be its k-spectrum and $F'_k$ be the k-spectrum of the random copy of 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 $F_k$ as the ratio of the Shannon information in $F_k$ to that expected in $F'_k$:

$$M_R^{(0)}(F_k) \equiv R(F_k) / R_{ex}(F'_k) = 2R(F_k) \tilde{f} / b'_k$$

(5)

Obviously, if Q is itself a random sequence, then $M_R$ is expected to be unity in any of its k-spectra.

**Case when genome is compositionally biased**

The situation is slightly more complicated for genomes with $p$ deviating significantly from 0.5. Fig. 2 shows the 6-spectra from the genome of *Chlamydia muridarum* [7] (black) and its random copy (green). Both have $p \approx 0.6$. Whereas the genomic spectrum is still unimodal, the random spectrum is composed of several sharp

<table>
<thead>
<tr>
<th>$k$</th>
<th>$H/\ln 2$</th>
<th>$R/\ln 2$</th>
<th>$R_{ex}/\ln 2$</th>
<th>$P. aerophilum$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.9999</td>
<td>5.90 E-6</td>
<td>5.77 E-6</td>
<td>3.973</td>
</tr>
<tr>
<td>3</td>
<td>5.9999</td>
<td>3.72 E-5</td>
<td>3.46 E-5</td>
<td>5.933</td>
</tr>
<tr>
<td>4</td>
<td>7.9999</td>
<td>1.72 E-4</td>
<td>1.62 E-4</td>
<td>7.881</td>
</tr>
<tr>
<td>5</td>
<td>9.9993</td>
<td>7.26 E-4</td>
<td>7.53 E-4</td>
<td>9.821</td>
</tr>
<tr>
<td>6</td>
<td>11.999</td>
<td>2.94 E-3</td>
<td>2.90 E-3</td>
<td>11.75</td>
</tr>
<tr>
<td>7</td>
<td>13.988</td>
<td>1.18 E-3</td>
<td>1.17 E-3</td>
<td>13.66</td>
</tr>
<tr>
<td>8</td>
<td>15.955</td>
<td>4.78 E-2</td>
<td>4.71 E-2</td>
<td>15.53</td>
</tr>
<tr>
<td>9</td>
<td>17.798</td>
<td>2.02 E-1</td>
<td>1.88 E-1</td>
<td>17.26</td>
</tr>
<tr>
<td>10</td>
<td>19.408</td>
<td>5.92 E-1</td>
<td>5.24 E-1</td>
<td>18.59</td>
</tr>
</tbody>
</table>

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 copy. $R_{ex}$ is the expected information in the random copy.
peaks. These are caused by the biased composition in the sequence. To see this, we denote by \( m \)-set the subsets \( F_{k,m} \) of \( k \)-mers with \( m \) (A+T)’s, \( m=1 \) to \( k \). Owing to the biased composition, the mean occurrence frequencies of the subsets \( F_{k,m} \) are spread out: 

\[
\bar{f}_m(p) = \bar{f} 2^k p^m (1 - p)^{k-m},
\]

where \( \bar{f} \) is the overall mean. (Notice that \( \bar{f}_m(p) \) approaches \( \bar{f} \) when \( p \) approaches 0.5.) The narrowness of the corresponding subspectra causes the \( k \)-spectrum of the random copy to appear as the superposition of \( k+1 \) non-overlapping sharp peaks as shown in the green spectrum in Fig. 2. Apparently, for the genome the subspectra are sufficiently broad and overlapping such that no individual peak is discernable in its \( k \)-spectrum.

Table 2: Shannon information in the \( m \)-set of \( k \)-mers, \( F_{k,m} \), from the genome \( C. muridarum \) and its random copy. Frequencies are normalized to that of a 1 Mb sequence. Eq. (8) is a universal formula given later in the text.

<table>
<thead>
<tr>
<th>( k, m )</th>
<th>( f_m )</th>
<th>( R_{C_{mur}} ) measured</th>
<th>Eq. (8)</th>
<th>( R_{random} ) measured</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>2, 1</td>
<td>60,000</td>
<td>1.96 E-2</td>
<td>2.00 E-2</td>
<td>2.88 E-6</td>
<td>4.17 E-6</td>
</tr>
<tr>
<td>3, 2</td>
<td>18,000</td>
<td>4.36 E-2</td>
<td>2.93 E-2</td>
<td>2.22 E-5</td>
<td>2.08 E-5</td>
</tr>
<tr>
<td>4, 2</td>
<td>3,600</td>
<td>8.18 E-2</td>
<td>7.18 E-2</td>
<td>1.94 E-4</td>
<td>1.21 E-4</td>
</tr>
<tr>
<td>5, 3</td>
<td>1,080</td>
<td>1.10 E-1</td>
<td>0.92 E-1</td>
<td>5.19 E-4</td>
<td>4.34 E-4</td>
</tr>
<tr>
<td>6, 3</td>
<td>216</td>
<td>1.53 E-1</td>
<td>1.84 E-1</td>
<td>2.98 E-3</td>
<td>2.24 E-3</td>
</tr>
<tr>
<td>7, 4</td>
<td>64.8</td>
<td>1.95 E-1</td>
<td>2.42 E-1</td>
<td>9.98 E-3</td>
<td>7.65 E-3</td>
</tr>
<tr>
<td>8, 4</td>
<td>13.0</td>
<td>2.84 E-1</td>
<td>4.77 E-1</td>
<td>5.82 E-2</td>
<td>3.83 E-2</td>
</tr>
<tr>
<td>9, 5</td>
<td>3.89</td>
<td>4.53 E-1</td>
<td>6.17 E-1</td>
<td>1.82 E-1</td>
<td>1.28 E-1</td>
</tr>
<tr>
<td>9, 7</td>
<td>8.75</td>
<td>3.91 E-1</td>
<td>2.74 E-1</td>
<td>7.97 E-2</td>
<td>5.70 E-2</td>
</tr>
<tr>
<td>10, 6</td>
<td>64.8</td>
<td>0.93 E 0</td>
<td>0.80 E 0</td>
<td>6.66 E-1</td>
<td>5.15 E-1</td>
</tr>
<tr>
<td>10, 8</td>
<td>2.62</td>
<td>6.87 E-1</td>
<td>3.55 E-1</td>
<td>2.87 E-1</td>
<td>1.98 E-1</td>
</tr>
</tbody>
</table>

The overall width 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 approximately given by

\[
\Delta_k(p) = \bar{f} \left( 2^k (p^2 + (1 - p)^2)^k - 1 \right)^{1/2}
\]

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 copy 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 copy. The measured Shannon informations (column 5) in the \( m \)-sets of the random copy are close to their expected values \( b_k'f_m \) (column 6). The values of the Shannon information in the genomic subspectra, in absolute magnitudes and relative to their respective random counterparts, are both similar to those seen in Table 1. Therefore we generalize the definition for \( M_R \) given in Eq. (5) to be the weighted average over the reduced Shannon information in the \( m \)-sets:

\[
M_R(F_k) \equiv \sum_{m=0}^{k} L^{-1} \left( 2^k (k,m)\bar{f}_m \right) M_R^{(0)}(F_{k,m})
\]
where \( M_R^{(0)}(F_{k,m}) \) is as defined in Eq. (5), but with \( F_k \) replaced by \( F_{k,m} \) and \( \bar{f} \) replaced by \( \bar{f}_m \), and \((k,m)\) is a binomial satisfying \( \sum_m 2^k (k,m) f_m = L \). The Shannon information in an \( m \)-set is given by Eq. (2) except that now \( \tau = 2^k (k,m) \). In practice, to circumvent large fluctuations in \( R(F_{k,m}) \) induced by small unevenness in the A/T (or C/G) contents - this can occur when \( f_m \) is very large at \( k=2 \) and 3 - each frequency was divided by a factor \( (2^k / p_m (1-p)^{k-m}) \prod_s p_m^s \), where \( m_s \) is the number of the \( s^{th} \) type of base in the \( k \)-mer and \( \sum_s m_s = k \).

**Tests with control sequences**

The reduced Shannon information (Eq. (6)) 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 meant to be “copies” of sequences in the genomes set, composed of 135 prokaryotic complete genomes (the prokaryotes) and 127 complete chromosome sequences of 10 eukaryotes (the eukaryotes). Each copy has the same length and approximately the same \( p \) of its genomic target. The targets of the random and century sets are the 127 prokaryotes. The targets of the common-root set are the combined 262 prokaryotes and eukaryotes. Sequences in the random set are just random copies of the genome sequences. Sequences in the century set are 100-replicas of random root-sequences. Sequences in the common-root set are replicated from 300 b random root-sequences. That is, corresponding to a complete genome sequence of length \( L \), there is \( L/300 \)-replica in the common-root set.

![Figure 3](image_url)

Fig. 3 shows log-log plots of reduced Shannon information from the \( k \)-spectra, \( k=2 \) to 10, of sequences in the three control sets whose composition are explained in the text. (A) The random set (135 sequences); \( M_{R\text{ave}}=1.03\pm0.12 \). (B) The century set (135 sequences); \( M_{R\text{ave}}=101\pm12 \). (C) The common-root set (square symbols; 262 sequences); \( (300/L)M_{R\text{ave}}=1.02\pm0.13 \). These results gives us confidence in the normalization used in equations Eq. (5) and Eq. (6) for defining the reduced Shannon information.

**INFORMATION IN WHOLE GENOMES**

**Complete genome sequences**

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 inverse is the case for the eukaryotes where length ranges from 0.2 Mb (smaller chromosomes of E. cuniculi) to 268
Figure 4: Reduced Shannon information, $M_R$, from 135 complete microbial genomes and 127 eukaryotes. Each symbol is the $M_R$ value of one $k$-spectrum from one complete sequence. Left panel, $M_R$ color-coded by organism; right panel, $M_R$ 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 $M_R \approx 1$ regardless of sequence content, have been discarded. Straight red lines in the plots are $M_R \propto L$ lines.

Mb ($R. norvegicus$ Chromosome I) and $p$ ranges from 53% to 64%. The exception is $Plasmodium$ whose $p$ is $81\pm 1\%$ [9].

Shannon information in complete genomes

The reduced Shannon information in the $k$-spectra of the 135 prokaryotes and 127 chromosomes of eukaryotes are color-coded by organism and shown in Fig. 4(A), where each piece of datum gives the $M_R$ in one $k$-spectrum of a sequence. The values of $M_R$ 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 not color-separated. Instead, they are all shown as black crosses. 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 $M_R$. The data from human (24 chromosomes), mouse (21 chromosome) 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 [8], 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 multiplication 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 red $M_R \propto L$ 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.

Universality classes of genomes

The linear relation between $M_R$ and $L$ implies that the effective root-sequence length $L_r(k)$ given by $L_r(k) \equiv L/M_R$ approximates a $k$-dependent but genome-independent constant. In Fig. 5, the black symbols are values for $L_r(k)$ obtained by averaging over subsets of genome data: ▲ from prokaryotes, ■ from eukaryotes ($Plasmodium$ excluded) and ▼ from sequences formed by concatenating the non-coding segments in prokaryote genomes. These results are well summarized by the simple formula ($L_r(k)$ in units of bases):

$$\log L_r(k) = ak + B; \quad 2 \leq k \leq 10$$

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

We refer to Eq. (7) as a universality class, whose mean is given by the straight line in Fig. 5. (Orange symbols in Fig. 5 are results obtained from model sequences, to be discussed later.) There is a number of ways to understand the universality of meaning of $L_r(k)$. One way is to see that, for given $k$, the Shannon information
in a genome 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 decreases as 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 Eq. (7) we have \( L_r(2) \), \( L_r(6) \) and \( L_r(10) \) being approximately 250 b, 11 kb and 480 kb, respectively. Hence the Shannon information in the 2-, 6- and 10-spectra of a genome approximately 2 Mb long is about 8,000, 1,820 and 4.2 times that of a 2 Mb random sequence having the same base composition as the genome.

The universality class expressed by Eq. (7) includes all the genomes/chromosomes 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 \( a=0.146\pm0.012 \) and \( B=2.14\pm0.05 \). Coincidently, the two classes essentially have a common \( L_r(2) \): 240±160 for the main class and 270±120 for Plasmodium.

**Universal formula for Shannon information**

From Eq. (7) we derive a formula for the Shannon information in an \( m \)-set \( F_{k,m} \) of a genome sequence of composition \( p \) in the main class:

\[
R(F_{k,m}) \approx 0.012(1 - 2^{1-k})e^{0.44k}(2^k p^m (1-p)^{k-m})^{-1}
\]  

(8)

When \( p \) approaches 0.5 the formula collapses to

\[
R(F_k) \approx 0.012 \ (1 - 2^{1-k}) \ e^{0.44k}
\]  

(9)

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. (8) is independent of \( L \) and Eq. (9) is independent of both \( L \) and \( p \). Eq. (8) was used to produce the numbers given in column 4 of Table 2.

From the above and Eq. (3) we also obtain a formula for the relative spectral width for \( F_{k,m} \): \( \sigma(F_{k,m}) \approx (2R(F_{k,m}))^{1/2} \) when the genome has \( p \neq 0.5 \), and \( \sigma(F_k)\approx(2R(F_k))^{1/2} \) for the whole \( k \)-spectrum when \( p\approx0.5 \). Note that \( \sigma(F_k) \) 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.

**Coding and non-coding 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. In Fig. 5 the \( L_r(k) \) for sequences obtained by concatenating the non-coding segments in prokaryotes are shown as ▼. Both these and the eukaryote data (■) show a slight leveling-off beginning at \( k=9 \). Overall, from the data shown in Fig. 5 one may infer that no essential difference in \( M_R \) between coding and noncoding regions obtains.
This is not to say that statistical sequence similarity between coding and non-coding 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 $M_R$. 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 non-coding 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 non-coding 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, the excess (if there is any) in A's in genes in one orientation will approximately be equal to the excess in, say, T's in genes in the opposite orientation.

**INTERPRETATION OF RESULTS**

**Duplications increases $M_R$ 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 “+’s” (red for eukaryotes and yellow for prokaryotes) in Fig. 3(C). It is extremely similar to the band of data (squares) 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 $L_n(2)\approx 250$ b common to the two universality classes 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 $M_R$ from $n$-replicas are $k$-independent. Hence the presence of a strong $k$-dependence in the genome data rules out the possibility that genomes are simple $n$-replicas. Some other mechanism or mechanisms must have been at work to generate the observed $k$-dependence in $M_R$.

**Point mutations decreases $M_R$ 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, $d > > 10$ (10 is the maximum $k$ in the present study), the effect of the mutations on the $k$-spectrum will be negligible to give $M_R \approx n$. Conversely, when the number of mutations is very large, $d < < 1$ and all traces of replication in the $n$-replica will be obliterated reducing the $n$-replica to a random sequence yielding $M_R \approx 1$. In between, when $d$ is of the order of $k$, the mutation will affect the $k$-spectra in such a way that the $M_R$ 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 $M_R$ similar to that observed in Fig. 4.

**MODEL FOR GENOME GROWTH**

**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 homologous duplications to create large values for $M_R$; a suitable number of random point replacements to generate the correct $k$-dependence in $M_R$. 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 [11]. After full growth the sequence is subjected to random point replacements at a frequency 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 for its simplicity in order to limit the complexity of 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 \( \chi^2 \) procedure based on comparing empirical values of \( L_r(k) \) and those computed from model sequences was used to determine optimal values for the parameters \( L_0 \), \( r \) and \( l_x \). The \( \chi^2 \) 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 \( r=0.95 \) (detail 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 in our model \( p \) and \( 1-p \) sequences 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 \( l_x >> L_0 \), 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.

**Results from model**

Using the optimal parameters (\( L_0=8 \), \( l_x=250 \) and \( r=0.95 \)) we generated 248 model sequences whose lengths and base compositions more or less match those of the genomes/chromosomes in the main universality class and computed \( M_R \) and \( L_r(k) \) for the model sequences. Red triangles 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 \( M_R \) of a \( k \)-spectrum guarantees a correct shape for that spectrum. The plotted 5-spectra in Fig. 6, where the the spectra from the model sequences are given in orange and those from three genome sequences in black (green curves are from the random copies) indicate the typical agreement between model and genome spectra. 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.

![Figure 6](image)

Figure 6: Comparison of 5-distributions of genome (black), random (green) and model (orange) 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).

The existence of the *Plasmodium* chromosomes as a separate universality class is a blessing in disguise, for it shows that there is nothing inevitable about the main universality class. 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 red circles 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 length, or both, than genomes in the main class. The real cause for the distinctiveness of *Plasmodium* may be far more complex. Among the eu karyotes studied *Arabidopsis*, which belongs to the main class, is phylogenetically the least remote from *Plasmodium* [9, 10]. It will be interesting to see how closer taxonomic relatives of *Plasmodium* [10] are classified by \( M_R \).

**DISCUSSION**

Universality in diversity

Our study of 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; (ii) the Shannon information in complete genomes un equivocally 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 replicative transpositions 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. It will need to be refined when it is confronted with finer textual details in the genome.

More universality classes?

The existence of the *Plasmodium* class raises several questions: (i) What caused the formation of different classes? (ii) Are there other classes than the two reported here in existence? (iii) Are there other organisms in the *Plasmodium* class? (iv) What are the biological signatures that characterize the different classes? This study may have provided a plausible partial answer to the first question. The answers to the next two questions presumably will become known in due course as more complete genomes are sequenced. It would be surprising if there were only two classes or if *Plasmodium* were to be the sole member of the second class. The last question can only be answered by detailed comparative studies of *Plasmodium* with genomes in the main classes.

Neutral theory of evolution

Whereas the complete genomes studied varied greatly in coding regions as a percentage of the whole genome (from 85% in microbes to less than 2% in the mammals), the universal genome property reported here seems not to depend on that percentage. If we assume that coded words other than genes such as binding sites, regulatory signals, and microRNA’s [12] collectively do not occupy a dominant portion of the noncoding regions in eukaryotes, then our findings appear to imply that the majority of the individual fixed duplications and replacements during genome growth were selectively neutral. 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 [13, 14], a theory that was based on the investigation of polymorphisms of genes.

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 other evidence of duplications in genomes: the existence of many transposable elements; the large amounts of repeats in both prokaryotes [15] and eukaryotes [16, 17]; the preponderance of paralogs (genes) and pseudogenes in all life forms [18, 19]; chromosome segment exchanges that seem to characterize mammalian [8] and plant [20] radiations. Our proposed growth model may at least be taken as a starting point for an explanation of all these phenomena.

Did random duplication speed up evolution?

An important lesson we have learned from this study is that when a genome adds homologous sequence to itself its \( M_R \) will increase. Hence stochastic duplication is a highly efficient process for a sequence to increase its reduced Shannon information in a non-directed fashion. Homologous duplication, even when
made randomly, also makes good evolutionary sense. For sometimes such duplications will copy a segment in which is embedded a coded sequence, say a gene, which can later evolve 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. Thus having random segmental duplication as a major mode of genome growth makes the rapid evolution of life easier to understand. Furthermore, because the process is random, it can be argued that on some appropriate length scale - the length \( l_e \) in our model - the entropy of the genome is being maximized in the process. In other words, growth by random segmental duplication enables the genome to have the best of two worlds: to gain (the capacity for biological) information while not fighting against the second law of thermodynamics. At the same time, because the difference in Shannon information between coding and non-coding regions is small, the fine-tuning by natural selection of potential information into real potential will not be met with strong entropic resistance. There is some independent evidence suggesting that a growth strategy with a reliance on duplication may have the effect of enhancing the rate of evolution \([21, 22]\) and increasing the robustness of organisms \([23]\).

ACKNOWLEDGMENTS

This work is supported in part by the grants 92-2119-M-008-012 and 92-2112-M-008-041 from the National Science Council, ROC.

REFERENCES