

Universal Lengths in Microbial Genomes and Implication for Early Genome Growth

Chang-Heng Chang¹, Li-Ching Hsieh¹, Liaofu Luo³ and Hoong-Chien Lee^{1,2,4,5*}

¹Department of Physics and ²Department of Life Sciences, National Central University, Chungli, Taiwan 320;

³Physics Department, Inner Mongolia University, Hohhot, China; ⁴Center for Complex Systems, National Central University, Chungli, Taiwan 320; ⁵National Center for Theoretical Sciences, Shingchu, Taiwan.

We report the discovery of a set of universal lengths that characterize all microbial complete genomes. The Shannon information [Shannon 1948] of 108 complete microbial genomes relative to those of their respective randomized counterparts are computed and the results are summarized in a two-parameter exponential relation: $L_r(k) = (42 \pm 21) \times 2.64^k$, $2 \leq k \leq 10$, where L_r is a "root-sequence length" in nucleotides (nt) and k is the length of the oligonucleotides with which the genome is analyzed. The method of the analysis does not attend to $k \geq 11$ owing to the finite lengths - 0.4 to 7 Mb - of the microbial genomes. The relation means that for a given k the Shannon information of any microbial complete genome, irrespective of its length, is the same as that of a random sequence of length $L_r(k)$ having the same base composition of the genome. We show that artificial sequences generated by maximally stochastic short segmental duplications from random root-sequences of about 200 nt long have Shannon information and root-sequence lengths very similar to those of microbial genomes. This implies that the ancestors of microbial genomes mainly grew by segmental duplication shortly after they acquired a length of around 200 nt. The inference is that the ancestral genomes lived in an RNA world and replicated and duplicated with small ribozymes or their precursors.

It has long been noted that genomes have significantly, even extremely, overrepresented oligonucleotides (oros).^{*} A well-studied feature described as "the sheer dominance of a few members over the overall population" is the approximate power-law relation, $N = aF^b$, obeyed by oligonucleotides six to ten nt long in essentially all genomes, where F is the occurrence frequency, N is the number of oligonucleotides and a and b are constants [Li 1992, Mantegna et al. 1994, Luscombe et al. 2002]. Just as there are oros, there are also underrepresented oligonucleotides (uros). Broadly speaking the terms uro and uro may be used in two senses. In the more common usage they are used with respect to the actual distribution of the occurrence frequency in a genome, and by definition such oros and uros will form a relatively small set. Many such uros [Karlin et al. 1992] and oros [Karlin et al. 1992, Karlin and Burge 1995, Smith et al. 1995, Colbert et al. 1998] are known to have biological functions and methods for finding control or signal sequences in the genome based on extreme representation in this sense have been devised [van Helden et al. 1998, Bussemaker et al. 2000]. In a less common usage of the terms, extreme representation in a genome is an attribute of an occurrence frequency relative to the distribution of the occurrence frequency in a random sequence having the same length and base composition as the genome.

This second, less common usage is adopted in this paper. As soon as extreme representation is considered from this point of view, one is quickly confronted with the fact that very large numbers of oros and uros is a common feature of all complete microbial and eukaryotic genomes. Consider for instance the occurrence frequencies of hexanucleotides, or 6-mers, in an arbitrary one million

nucleotide (M nt) stretch of (a single strand of) the genome of *E. coli*. The mean frequency is 244. If the genome were a random sequence then the standard deviation (std) of the frequency distribution would be 15.6 and the chances that any 6-mer would occur more than 400 times or less than 100 times would be practically zero. In reality, the std of the frequencies for *E. coli* is 140, about nine times greater than that for a random sequence, and about 500 6-mers do occur more than 400 times and about 510 6-mers do occur less than 100 times. This pattern is neither particular to 6-mers nor to *E. coli*. In fact wide spectral width in the distribution of occurrence frequency for word lengths of at least up to 9 letters is a universal feature of known genomes.

Very few of the oros and uros in *E. coli* are known to have biological function and we shall assume most do not. If this is the case then we may ask what is the root cause of the existence of so many oros and uros or, alternatively, the root cause for the distribution of occurrence frequencies of genomic sequences having spectral widths so much greater than those of random sequences. If (most) of the oros and uros do not have biological function then whatever the cause may be, it would not be something that acts on individual oligonucleotides.

The purpose of this paper is the following: (i) to evince the phenomenon that wide spectral width in the distribution of occurrence frequency for short oligonucleotides is indeed a universal feature of complete bacterial genomes; (ii) to explore the nature of the universality of the phenomenon; (iii) to demonstrate that a very simple biological motivated model for the growth of genomes generates sequences that exhibits the phenomenon.

In this paper we report on our study of frequency distributions of k -mers, called k -distributions, for $k = 2$ to 10, of all microbial complete genomes. Extension of the study to cover all eukaryotic complete genome is under way. This

^{*}Corresponding author.

E-MAIL hcllee@phy.ncu.edu.tw; FAX (886) 3-426-5318.

study does not go beyond $k = 10$ because the canonical length of microbial genomes in nt, 10^6 , is about the same as the number of distinct 10-mers (4^{10}), so that spectral widths of k -distributions for $k > 10$ are essentially just the genome length itself and no longer supply the information we seek in this work (we are not saying that the statistics of longer oligonucleotides do not contain possibly interesting information). In our discussion it will be assumed that 4^k is not greater than the genome length.

A primary driving force of this study is our view that a wide spectral width in the distribution of occurrence frequency of oligonucleotides in a genome is a signature that the genome carries large amounts of information. In this context the ratio of the std to the mean of a distribution, hereafter called the relative spectral width and denoted by σ , will be used throughout our discussion. The connection between a large value for the relative spectral width and large information capacity may be first understood in the obverse: a genome with all oligonucleotides evenly represented, in which case $\sigma = 0$, can carry very little or practically no information. The connection is formalized in a not so well-known relation: an appropriately defined Shannon information [Shannon 1948] of a bell-shaped distribution is proportional to σ^2 of the distribution when $\sigma \ll 1$, with a proportional coefficient that is close to 0.5 (see Methods).

We take as benchmark for relative spectral width of a k -distribution from a genome to be that of the corresponding k -distribution from a reference random sequence having the same length and base composition as the genome. This choice is made because a random sequence is in practice the customary standard of reference for giving an “expected” value of an occurrence frequency, because there is in fact a high degree of randomness in genomic sequences and because there is a simple expression for the expected value of σ for a random sequence. We define *relative spectral information* $\mathcal{M}_\sigma$ of a k -distribution to be the ratio of the square of σ for the distribution from a genome to that expected from a reference random sequence, and similarly define *relative Shannon information* $\mathcal{M}_R$ (see Methods). These quantities turn out to be very useful, for through their usage an unexpected universality of microbial genomes is revealed.

Some simple properties of random sequences and “ m -replicas”, concatenates of m replicates of the sequences are useful for gaining insight in the meaning of the results to be presented. Suppose Q is a random sequence of length L and S_k is its k -distribution. For simplicity let $1 \ll 4^k \equiv \tau \ll L$. Then S_k is a Poisson distribution with mean frequency $\bar{f} = L/\tau$, std $\bar{f}^{1/2}$ and relative spectral width $\sigma = \bar{f}^{-1/2}$. Consider now two ways of increasing m times the length of Q . One way is replicating Q m times to get an m -replica of Q . Then the frequency of every k -mer in the m -replica will be m times that in S_k , so that for the m -replica the mean frequency and std will both grow by a factor of m but the relative spectral width will remain unchanged. The other way is to let Q grow into a random sequence of length mL . Then for the new random sequence the mean frequency will grow by a factor of m but std will only grow by a factor of $m^{1/2}$ so that the relative spectral width will *decrease* by a factor of $m^{-1/2}$. The difference between the m -replica and the (long) random sequence is revealed in their respective values for the relative spectral information and relative Shannon information (of

their k -distributions); for m -replica both will be m whereas for the random sequence they will be unity. Thus $\mathcal{M}_R$ (and $\mathcal{M}_\sigma$) may be understood as an effective replication number. Note that by definition $\mathcal{M}_R$ is not supposed to be dependent on k for a random sequence or its replica, otherwise this k -independence is not expected.

The m -replica has larger values for $\mathcal{M}_\sigma$ and $\mathcal{M}_R$ because it has a “root-sequence” that is a random sequence whose length is $1/m$ of its own length. We are therefore motivated to have the following: if the value for the relative Shannon information of the k -distribution of a sequence (not necessarily a replica of a random sequence) of length L is $\mathcal{M}_R(k)$, then for k -mers the *root-sequence length* L_r of the sequence is defined as $L/\mathcal{M}_R(k)$. There would have been little practical difference if $\mathcal{M}_R$ were replaced by $\mathcal{M}_\sigma$ in this definition.

Our results for the 108 microbial complete genomes are briefly summarized as follows. We find that the relative Shannon information of a genome is proportional to the genome length and independent of base composition. Furthermore, the proportionality constant is the *same* (to within 50%) for all microbial genomes and grows exponentially with decreasing k . These results are compactly expressed by the relation $L_r(k) \approx 42 \times 2.6^k$ nt where $L_r(k)$ is the k -mer universal root-sequence length of the genomes. The root-sequence length is of the order of the genome length for $k = 10$ and decrease rapidly with decreasing k ; $L_r(2)$ is only about 300 b. This implies that as far as 2-mers are concerned a 2 M nt microbial genome looks like a 6600-replica of a 300 nt random sequence. However the genomes cannot be simple replicas because L_r has a strong k -dependence.

The universality (*i.e.*, same for all genomes) and the strong k -dependence of the root-sequence lengths suggests a universal but non-trivial mechanism for genome growth that involves much duplication, with a universal initial condition. We show that artificial sequences generated by *quasireplication*, that is maximally stochastic duplications of short segments (about 20 nt), from random root-sequences of about 200 nt long to the full length of microbial genomes indeed have k -distributions and Shannon information similar to those seen in the real genomes.

RESULTS

Relative Shannon Information for Random Sequences and Microbial Genomes

Fig. 1 shows log-log plots of relative spectral information ($\mathcal{M}_\sigma$) and relative Shannon information ($\mathcal{M}_R$) of k -distributions, $k = 2$ to 10, versus sequence length L computed for a “genome” set of the 108 complete microbial genomes ([GenBank 2003] as of April 1, 2003) and four control sets of sequences the 108 members in each of which have lengths and base compositions matching those in the genome set: a “random” set of random sequences and three “replica” sets of m -replicas of random root-sequences, $m = 10, 100$ and 1000. An m -replica is a sequence obtained by replicating a root-sequence m times. That is, if a 100-replica is 1 M nt long then it is the concatenation of 100 copies of a root-sequence 10 k nt long. We may think of a random sequence as the 1-replica of it self. The genomes in the genome set are heterogeneous in length (0.4 to 7 M nt) - and base composition 28 to 78% A+T) and sequences in the control sets are made to match the genome set in sequence length and base composition.

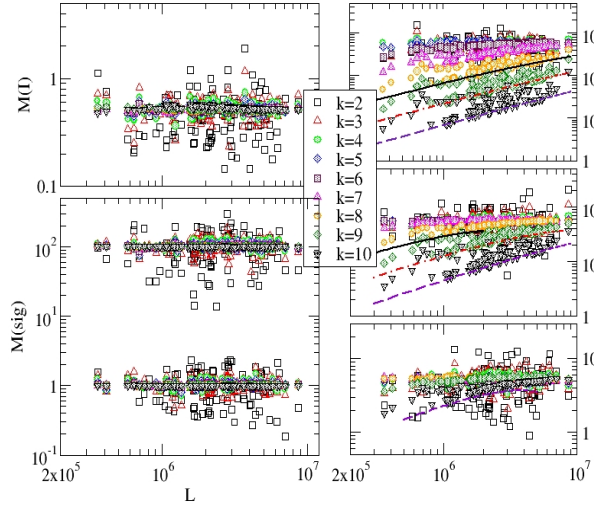


Figure 1: Relative Shannon information $\mathcal{M}_R$ for: (A) the random set; (B) the set of 100-replicas. For the replica set, if $4^k \lesssim L/100$, then $\mathcal{M}_R \approx 50$ is expected. Otherwise it is expected to be less; the three lines give the theoretical lowerbound for $k = 8, 9$ and 10 .

The random and replica sets are used to test our formulation and the accuracy of the computation. The values for $\mathcal{M}_\sigma$ and $\mathcal{M}_R$ (of a k -distribution) for a random sequence are expected to be 1 and a constant C approximately equal to 0.5, respectively, independent of k , sequence length and base composition. For an m -replica of length L , $\mathcal{M}_\sigma$ is expected to be m and $\mathcal{M}_R$ is expected to be mC for those k 's satisfying $4^k \lesssim L/m$. When $4^k > L/m$, $\mathcal{M}_R$ has a lowerbound $\mathcal{M}_R \approx (L/4^k) \log(4^k m/L)$ (see Methods).

In Fig. 1 each plot has 972 pieces of data and we see that the expectations described above are broadly met. In panels (A) and (B) about 50% of the error come from data for $k=2$ ("□" in the figure) and 25% from $k=3$ ("△" in the figure). This arises from the way $\mathcal{M}_\sigma$ and $\mathcal{M}_R$ are defined; both involve a theoretical expression as denominator whose value is very small when $k = 2$ and 3 , so that fluctuations in the computed value for the numerator are magnified yielding the large errors observed (see Methods). From panel (B) the value 0.514 ± 0.062 is extracted for the constant C . In panels (C), (D) and (E) the lines give the lowerbounds for the higher values of k . These results for the control sets assure us that we understand the relative Shannon information for random sequence and their replicas and may be summarized as follows: if the value of $\mathcal{M}_R$ of a sequence of length L is mC for k 's satisfying $4^k \lesssim L/m$ and is less than mC for k 's that are greater, then the sequence is likely an m -replica of a random root-sequence of length L/m .

The normalized relative Shannon information $\mathcal{M}_R/C_2$ for the genome set is shown in Fig. 2, where each piece of datum was multiplied by a factor of 2^{10-k} to partition data into different k groups for better viewing. In $\mathcal{M}_R/C_2$ is plotted against L ; in (B) $\mathcal{M}_R/C_2/L$ is plotted against the combined A+T probability (p_{AT}) in the sequence. The data display properties that are entirely different from those of the control sets: (i) They have a strong k -dependence. (ii) For given k the mixed p_{AT} data form a band (std is about 50% of mean) that depends linearly on

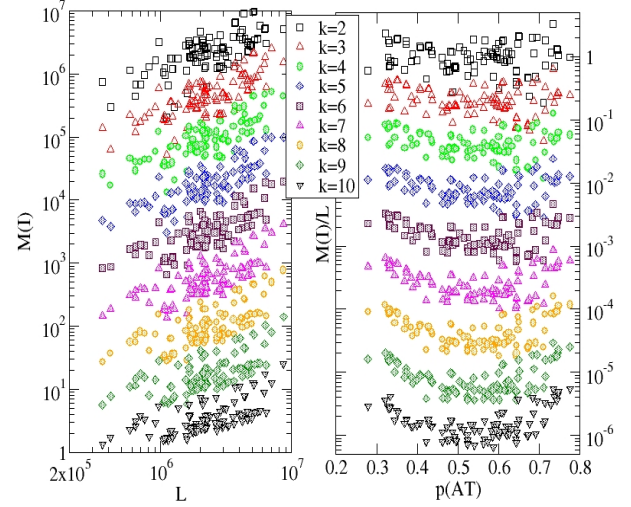


Figure 2: Relative Shannon information $\mathcal{M}_R$, multiplied by the factor $2^{10-k}/C$, for the genome set.

L . (iii) For given L , the mean $\mathcal{M}_R$ has an exponential dependence on k , $\langle \mathcal{M}_R/C_2 \rangle \propto 2.64^{-k}$. (iv) $\mathcal{M}_R/C_2/L$ has a weak dependence on p_{AT} that is approximately symmetric with respect to the point $p_{AT} = 0.5$.

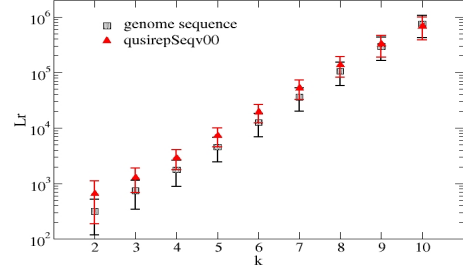


Figure 3: L_r versus k extracted from the relative Shannon information $\mathcal{M}_\sigma$ from the genome set and (•) and a set of model sequences (△) whose lengths and base compositions match those of the genome set.

Universal Root-Sequences Length of Genomes

The results in Fig. 2 shows that, especially for the smaller k 's, $m \equiv \mathcal{M}_R/C_2$ is a very large number. In such cases, as far as k -distributions are concerned, a genome of length L is like the m -replica of a random root-sequence of length $L_r = L/m$. But because $\mathcal{M}_R$ is (approximately) linear in L , L_r is independent of L . This, together with the fact that $\mathcal{M}_R$ has only a weak dependence on base composition, implies that for each k L_r is, within a factor of two, a *universal* root-sequence length, *i.e.*, the same for all microbial genomes. Property (iii) above implies that these universal lengths (squares in Fig. 3) satisfy the exponential relation (straight line is Fig. 3)

$$L_r(k) = \Lambda \times t^k; \quad 2 \leq k \leq 10 \quad (1)$$

where $\Lambda = 42 \pm 21$ nt and $t = 2.64 \pm 0.20$. For a specific genome of length L , the maximum value for k for

which the above equation holds is approximately given by $4^{k_{max}} \lesssim L$. Eq. (1) reduces the 972 pieces of data of the relative Shannon information of microbial genomics to two constants Λ and t and their errors. This strong dependence of L_r on k means that statistically, relative to random sequences (Fig. 1), the genome has a high sensitivity to the length scale with which it is examined; for 2-mers a 2 M nt genome is like a 6600-replica of a 300 nt random sequence, while for 10-mers it acts like a 3-replica of a 650 k nt random sequence. Eq. (1) is different from Zipf's law [Zipf 1949] and its variations that have been observed for k -mer frequencies ($k \geq 6$) in biological sequences [Mantegna et al. 1994, Luscombe et al. 2002].

DISCUSSION

Results from Model Sequences

The universality of the L_r 's to suggest the existence of a universal mechanism for genome growth from proto-genomes of a universal initial length. The small size of L_r for the smaller k 's implies a mechanism that involves a great deal of replication. However, simple whole-genome replication as discussed in [Ohno 1970, Skrabanek and Wolfe 1998, Hughes et al. 2001] at the early stages is ruled out because such a mode of growth yields, contrary to data, non-existent or too-weak k -dependence of $\mathcal{M}_R$ (panels (B-D), Fig. (1)). The observed strong k -dependence of L_r suggests a more complex duplication process.

We are motivated by the genomic data to consider a simple and biologically plausible growth model, through which "universal genomes" are generated *in silico* by a process we call *quasireplication* - growth by maximally stochastic short segmental duplication - from an short initial random sequence (see Methods). A set of model sequences with final lengths and base compositions matching those in the genome set were generated using an average length of $\bar{l} = 20 \pm 12$ (nt) for the duplicated segments and $L_0 = 200$ (nt) for the initial sequence. The relative Shannon information and root-sequence lengths (traingles in Fig. 3) computed for the set are shown in Fig. 4 and Fig. 3 (traingles), respectively. The agreement between these and the genomic data is quite good. The model is sensitive to its parameters $\bar{l}$ and L_0 . A very short initial genome is crucial for obtaining good results and is demanded by the shortness of $L_r(2)$ - about 300 b; L_0 much greater than 200 cannot generate sequences whose L_r has small enough values for the smaller k 's and has a strong enough k -dependence to match genomic data. Fig. 3 shows two sets of L_r 's given by two less good 2 M nt model sequences, one generated using $L_0 = 200$ and $\bar{l} = 25 \pm 12$ and the other using $L_0 = 500$ and $\bar{l} = 20 \pm 12$.

Frequency of Longer Words

We have not dealt with k -mers with k greater than 10. This is because the number of 10-mers, about 1 million, is just about the canonical size of microbial genomes. Distribution of longer k -mers in microbial genomes will be sparse and the study of such cases calls for a strategy different from the one used here. Such studies are being conducted.

Microbial Genomes are Quasireplicables

Our study shows that at least for up to 10-mers microbial genomes are statistically what we call quasireplicables. Like replicas of very high replication number (up to several

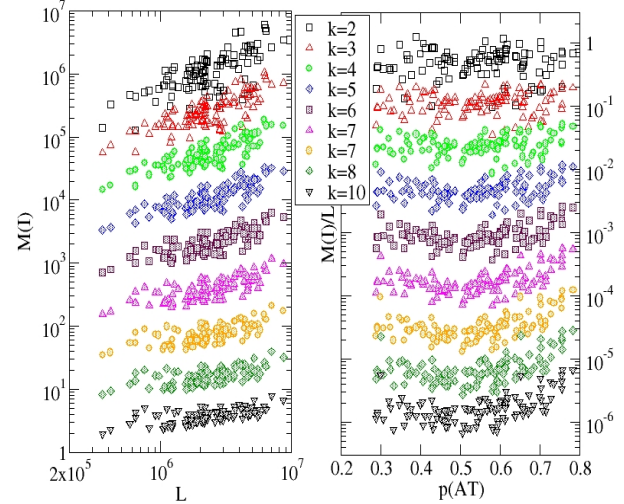


Figure 4: Relative Shannon information $\mathcal{M}_R$, multiplied by the factor $2^{10-k}/C$, for the quasireplica set.

thousand), these quasireplicables have very large $\mathcal{M}_R$ values for the shorter k -mers. Unlike m -replicas, the $\mathcal{M}_R$ values of these quasireplicables have a strong k dependence. Characteristically, quasireplicables are globally aperiodic, partially ordered, highly complex and evidently can carry large amounts of information. Recall that for a quasireplica of length L , $k_{max} \approx \ln L / \ln 4$ is the maximum k for which Eq. (1) applies. For $k \leq k_{max}$ a quasireplica acts as an m -replica, where m is approximately equal to the length of the sequence divided by $L_r(k)$, while for $k \gg k_{max}$ it appears essentially as a random sequence (this is not to say that microbial genomes are random sequences for long k -mers). Quasireplicables are robust structures that may be classified according to the two parameters Λ and t in Eq. (1). We have verified that, provided the typical duplicated segment length is significantly greater than k_{max} , quasireplication (including simple replication) of a quasireplica begets a longer quasireplica of the same class. We are currently analyzing eukaryotic genomes in the context of quasireplication.

A New Major Transition

We infer from results of this study the following conjecture for the early evolution of microbial genomes. A population of primitive genomes, shortly after its members had reached a length of about 200 b and had acquired a rudimentary duplication machinery, discovered and adopted quasireplication as a strategy for growth so that their progenies, at least the microbial genomes, are self-organized quasireplicables (belonging to the class given by Eq. (1)). If this description is substantially true then the onset to quasireplication must be another, hitherto unknown, major transition in evolution [Maynard Smith et al. 1997]. Prior to that transition the ancestral genomes evolved by unknown means and after the transition the genomes evolved via quasireplication.

RNA World

Although we cannot infer from the present study how primitive genomes got to the stage just before the transition, we do have two important inferences: (i) The primi-

tive genomes were short, about 200 nt long, possibly even shorter. This is less than a quarter of the size of a present-day gene coding for a typical enzyme and implies that the transition must have taken place in an “RNA world” devoid of proteins [Gilbert 1986, Joyce 2002]. Although the spontaneous synthesis of life requires a number of preconditions to be met [Szostak et al. 2001], a crucial one has been. Large - up to about 100 nt, highly active and complex RNA ligases have been isolated from a very limited pool of random RNA sequences [Ekland et al. 1995]. The 200 b length of the primitive genomes seems to be of an optimum size for the product of such a synthesis. (ii) The population of primitive genomes already had a spread in base composition similar to what is represented in present day microbial genomes. In a related issue, our model explains why the composition of a genome is essentially uniform over its entire length on scales of, say, greater than 10 k nt. The ribozymes that made up the duplication machinery at the time of transition probably were not much bigger than the smallest ribozymes now extant, about 31 to 50 nt [Forster and Symons 1987]. If this was the case, then the average duplicated segment length of about 20 nt was likely long enough to copy coded information which, presumably, could later be locally varied under the combined forces of mutation and natural selection for adaptation to new purposes. The short segment length is also consistent with the belief that the primitive small ribozymes were unlikely capable of faithfully duplicating much longer segment.

Rise of Proteins

It is not necessary that microbial genomes did grow to their full current lengths exclusively by short segment quasireplication. It is sufficient that they grew this way to become a quasireplica of a minimum length that belonged to the class defined by Eq. (1) (this length is about 250 k nt if $k = 10$ is excluded). This period of quasireplication, if it did occur, must have happened in the relatively short period between about 4.2 billion and 3.6 billion years ago, when the earth cooled down and protein must have appeared, respectively. After that the genomes could have continued to grow via quasireplication but, with the presence of proteins and enzymes, likely by duplicating much longer, including gene-length, segments [Yanai et al. 2000, Qian et al. 2001]. While the robustness of quasireplicates guarantees that such long-segment quasireplication would not take the genomes out of its original class of quasireplicates, the proliferation of pseudo-, duplicated and homologous genes in the genomes of larger organisms [Lynch and Conery 2000, Friedman and Hughes 2001, Bailey et al. 2002, Meyer 2003, Gu et al. 2003] could be understood as consequences of growth by such quasireplication.

METHODS

Complete Microbial Genomes

The 108 complete microbial genomes currently in the GenBank [GenBank 2003] are heterogeneous in length - 0.4 to 7 M nt - and base composition - 25 to 78% A+T. In most cases the numbers of A's and T's (and of C's and G's) in (one strand of) a genome are very similar. We therefore characterize the base composition of a genome by a single

Table 1: Shannon entropy and information (Eq. (2)) in units of $\log 2$ for random sequences and genomes. Genomic results are averaged over 108 genomes and random results are averaged over 108 random sequence with lengths and base compositions matching those in the genome set. Error in genomic information is less than 50% in all cases and much smaller for random sequences.

k	Random sequence		Genome average	
	Entropy	Inform'n	Entropy	Inform'n
2	3.9999941	5.90 E-6	3.97337	2.66 E-2
3	5.9999628	3.72 E-5	5.93353	6.65 E-2
4	7.9998282	1.72 E-4	7.88182	1.18 E-1
5	9.9992744	7.26 E-4	9.82101	1.79 E-1
6	11.997061	2.94 E-3	11.7525	2.74 E-1
7	13.988176	1.18 E-3	13.6653	3.35 E-1
8	15.952193	4.78 E-2	15.5313	4.69 E-1
9	17.797844	2.02 E-1	17.2668	7.33 E-1

parameter p , the combined probability of A and T, or of C and G, whichever is greater.

Relative Shannon Information

For the set of occurrence frequencies, or k -distribution, $\{f_i|L\}$ with $\sum_{i=1}^{\tau} f_i = L$ in a (circular) DNA of length L , where $\tau = 4^k$ is the number of types of k -mers, we define the relative spectral width for the distribution as $\sigma \equiv \Delta/\bar{f}$, where $\bar{f} = L/\tau$ is the mean frequency and Δ is the standard deviation (std). The Shannon uncertainty, often called the Shannon entropy, for the set $\{f_i|L\}$ is $H = -\sum_i p_i \log p_i$ where $p_i = f_i/L$ [Shannon 1948]. In the case of highest uncertainty, the maximum possible number of event types have nonzero and equal frequency and H has maximum value $\log \tau$ or $\log L$, whichever is less. In the opposite case of lowest uncertainty or highest certainty, which is of no interest to the present discussion, only a single type of event occurs and H has minimum value zero. The k -distributions of genomes are of a situation that is *maximally complex*: close to the maximum possible number of event types have nonzero frequencies. Shannon expressed the information in a system in terms of *decrease* in uncertainty, hence for maximally complex distributions we define the Shannon information as

$$R \equiv H_{max} - H = \min(\log L, \log \tau) + \sum_i p_i \log p_i \quad (2)$$

For simplicity we consider the case $L \gtrsim \tau$, that is, $k \lesssim \log L / \log 4$. For a maximally complex distribution H_{max} may be viewed as a large background in the distribution from which we want to extract the information-bearing small quantity $R \ll H_{max}$. This situation is illustrated in Table 1 where it is shown that the Shannon information is much greater in genomic sequences than in random sequences, especially for the smaller k -mers.

For a general unimodal distribution with relative spectral width σ it can be shown that

$$R = C_2 \sigma^2 + \mathcal{O}(\sigma^3) \quad (3)$$

where $C_2 \approx 0.5$; it is exactly 0.5 for Gaussian and square distributions. This reveals the close relation between Shannon information and relative spectral width σ . The relation is applicable even for distributions that are not unimodal or otherwise appear complicated, so long as it is

known that they have an underlying multimodal structure (see below).

We shall measure the R of distribution $\{f_i|L\}$ against that expected of a random distribution. The square of the std for the latter is (expected to be) $b\bar{f}$ so that $\sigma^2 = b/\bar{f}$, where $b < 1$ is a binomial factor depending on k and approaches unity when τ is large. We therefore define the relative Shannon information of $\{f_i|L\}$ to be

$$\mathcal{M}_R \equiv 2R\bar{f}/b \approx R/R_{ran} \quad (4)$$

where R_{ran} is the Shannon information of the distribution from a random sequence and the value 0.5 is used for C_2 . The expected result $\mathcal{M}_R \approx 1$ (hence $C_2 \approx 0.5$) for random distributions was borne out in Fig. 1.

Case when Distribution is not Bell-Shaped

In Fig. 5 the 6-distributions (cyan/gray curves) of random sequences 1M nt long with p equal to 0.5 (A), 0.6 (B) and 0.7 (C), respectively, are shown together with the per 1M nt 6-distributions (black) of three genomes with matching p values: *E. coli* [Blattner et al. 1997] (A), *C. muridarum* [Read 2000] (B) and *M. jannaschii* [Bult et al. 1996] (C). Only the two distributions in (A) for the $p = 0.5$ sequences satisfy the unimodal requirement to make Eq.(4) directly applicable. In this case the random distribution is indeed Poissonian and genomic distribution is much wider, $\sigma_{E.coli.}/\sigma_{ran} \sim 9$ so that, as far as 6-mers are concerned, a 1 M nt stretch of the *E. coli* genome is statistically like the 81-replica of a 12 k nt random sequence.

When $p \neq 0.5$ a k -distribution is a composite of $k+1$ sub-distributions, one for each of the subsets of k -mers with m AT's (called m -sets). This is clearly seen in the distributions for the random sequences in panels (B) and (C) of Fig. 5. In this case the sub-distributions are Poisson distributions whose respective means are $\bar{f}_m(p) = \bar{f} 2^k p^m (1-p)^{k-m}$, $m=0$ to k . For the genomes these sub-distributions are sufficiently broadened such that no individual peak is discernable in the summed distribution. We thus generalize the definition for $\mathcal{M}_R$ given in Eq.(4) to be the weighted average over the relative Shannon information of the m -sets:

$$\mathcal{M}_R \equiv 2 \sum_{m=0}^k L^{-1} \left(2^k(k, m) \bar{f}_m \right) R_{k,m} \bar{f}_m / b \quad (5)$$

where (k, m) is a binomial, $\sum_m 2^k(k, m) \bar{f}_m = L$. The factor b would normally be $1-\tau^{-1}$, but because A and T (and C and G) are counted together and the number of each monomer in the sequence is fixed, it is taken to be $b=1-2^{1-k}$. In practice, to circumvent large fluctuations in $R_{k,m}$ 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 is divided by a factor $(2^k/p^m(1-p)^{k-m}) \prod_s p_s^{m_s}$, where s runs over the four bases and $\sum_s m_s = k$.

For checking purposes we also define a "relative spectral information" $\mathcal{M}_\sigma$ where $R_{k,m}$ in Eq. (5) is replaced by $\sigma_{k,m}^2$.

Replicas of Random Sequences

Take a sequence Q' and replicate it m times to obtain its m -replica Q . Assume m is much less than the length L of Q . Then every k -mer appearing in Q' will appear m times in Q (we ignore small deviations caused by replication), and k -mers not appearing in Q' will also not

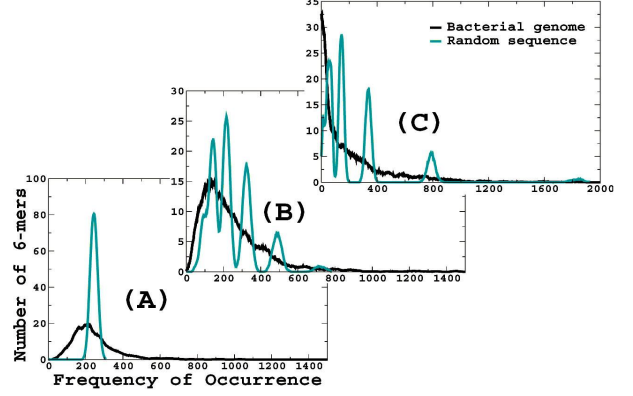


Figure 5: Comparison of 6-distributions of genomes (black) and random sequences (green/gray). (A), (B) and (C) are *T. pal.*, *C. mur.* and *M. jan.*, respectively, and random sequences with $p=0.5, 0.6$ and 0.7 , respectively

appear in Q . Because Q' is a random sequence, the k -distribution $S'_k = \{f'_i|L/m\}$ of Q' is a random distribution (provided $k \lesssim \log(mL)/\log 4$), whereas the k -distribution $S_k = \{f_i = m f'_i|L\}$ of Q is an m -multiple of S'_k . Therefore the relative spectral width of S_k will be $m^{1/2}$ times that of S'_k and $\mathcal{M}_R(S_k) \approx m \mathcal{M}_R(S'_k) \approx m C_2$. The difference is not caused by Q' being shorter than Q (the k -distribution of a random sequence of length L still has $\mathcal{M}_R \approx C_2$), but is caused by Q being an m -replica of Q' .

Root-Sequence Length

The above is a demonstration that a large value for $\mathcal{M}_R$ suggests the presence of replication or duplication. This motivates the following: If Q is a sequence of length L and has $\mathcal{M}_R(S_k)=m$, then we define $L_r=L/m$ as the root-sequence length of Q for k -mers. For as far as k -mers are concerned, Q (not necessarily an m -replica) has the order and information of the m -replica of a random "root-sequence" of length L_r . L_r expected to be independent of k when a sequence is actually the replica of a random sequence, but generally not so otherwise.

Quasireplication - Growth by Stochastic Short Segmental Duplication

The quasireplicas whose properties were shown in Fig. 4 were obtained from a two-phase growth model for generating nucleotide sequences [Hsieh et al. 2003, Hsieh et al. 2003b]. In the first phase of the model an initial "genome" (*i.e.*, a nucleotide sequence) is assumed to have grown by unknown means to a length L_0 . In practice we simply take this initial genome to be a random sequence of length L_0 and with a given p . In the second phase the genome grows by random short segmental duplications possibly modulated by random single mutations. The duplication process, or quasireplication, is maximally stochastic: a segment of length l , chosen according to some probability density function $f(l)$, is copied from one site and inserted into another site, both randomly selected. This model shares some features with models used to explain the power-law behavior of the occurrence frequency of genes in genomes [Yanai et al. 2000, Qian et al. 2001], except that there the units of duplication are genes, not short oligonucleotides, and those models were not aimed to describe the early growth of genomes.

A product of our model, called a quasireplica, of a certain length and p value is said to be universal in the sense that it is used to represent any microbial genome having the same length and p . Two quasireplicas of the same length and p are not the same sequences, nor would either be the same as any of the microbial genomes they represent. But all of them would have similar root-sequence lengths.

For the quasireplicas whose properties were shown in Fig. 4, an initial length of $L_0 = 200$ b, the Erlang density function $f(l) = 1/(an!)(l/a)^n e^{-l/a}$, with integer $n=2$ and the length scale $a=6.7$ (nt) were used after a coarse-grained search of parameter space to find optimum values. The chosen parameters imply the average length $\bar{l}$ of the duplicated segments is 20 ± 12 nt. An exhaustive, fine-grained search to find the absolute best set of parameters for the model was not conducted, nor do we think it necessary. However, the value $L_r \approx 300$ for $k = 2$ sets an upper limit for L_0 and, with $L_0 < 300$, $\bar{l}$ cannot much exceed 20.

The p value for a quasireplica is chosen at the outset. However the subsequent quasireplication may cause p to drift slightly and make the numbers of complementary nucleotides unequal. This last effect is dealt with using the method described in an earlier section for treating the same effect in real genomes.

ACKNOWLEDGEMENT

This work is supported in part by the grant 91-2119-M-008-012 from the National Science Council, ROC.

REFERENCES

- Bailey JA et al. (2002) Recent Segmental Duplications in the Human Genome. *Science* 297:1003-1007.
- Blattner FR, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453-1474.
- Bult CJ, et al. (1996) Complete genome sequence of the methanogenic Archaeon, *Methanococcus jannaschii*. *Science* 273:1058-1073.
- Bussemaker HJ, Li H and Siggia (ED 2000) Building A Dictionary for Genomes: Identification of Presumptive Regulatory Sites by Statistical Analysis. *Proc. Natl. Acad. Sci.* 97:10096-10100.
- Colbert T, Taylor AF and Smith GR (1998) Genomics, Chi sites and codons: 'islands of preferred DNA pairing' are oceans of ORFs. *Trends in Genetics* 14:485-488.
- Eklund EH, Szostak JW and Bartel DP (1995) Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* 269:364-370.
- Forster AC and Symons RH (1987) Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. *Cell* 49:211-220.
- Friedman R and Hughes AL (2001) Gene Duplication and the Structure of Eukaryotic Genomes. *Genome Res.* 11:373-381.
- GenBank: www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html.
- Gilbert W (1986), The RNA world. *Nature* 319:618.
- Gu Z, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421:63-66.
- Hsieh LC, Luo LF, Ji FM and Lee HC (2003) Minimal model for genome evolution and growth. *Phys. Rev. Lett.* 90:018101-018104.
- Hsieh LC, Luo LF and Lee HC, (preprint archive: <http://arXiv.org/abs/physics/0302031>).
- Hughes AL, da Silva J and Friedman R (2001) Ancient genome duplications did not structure the Human Hox-bearing chromosomes. *Genome Res.*, 11:771-780.
- Joyce GF (2002) The antiquity of RNA-based evolution. *Nature* 418:214-221.
- Karlin S, et al. (1992) Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucl. Acids Res.* 20:1363-1370.
- Karlin S and Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics* 11:283-290.
- Li W (1992) Generating non-trivial long-range correlations and 1/f spectra by replication and mutation. *Int. J. Bifurcation and Chaos* 2:137-154.
- Luscombe NM, et al. (2002) The dominance of the population by behaviour applies to a wide variety of genomic properties. *Gen. Biol.* 3:res.0040.1-0040.7
- Lynch M and Conery LC (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151-1155.
- Mantegna RN, et al. (1994) Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett.* 73:3169-3172.
- Maynard Smith J and Szathmari E, *The Major Transitions in Evolution*, (Oxford Univ. Press, London, 1997).
- Meyer A (2003) Duplication, duplication. *Nature* 421:31-32.
- Ohno S (1970) Evolution by gene duplication. (Springer Verlag, New York).
- Qian J, et al. (2001) Protein family and fold *J. Mol. Biol.* 313 673-681.
- Read TD, et al. (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucl. Acids Res.* 28:1397-1406.
- Shannon CE (1948) A mathematical theory of communication. *Bell Sys. Techn. J.* 27:379-423; 623-656.
- Skrabanek L and Wolfe KH (1998) Eukaryote genome duplication - where's the evidence? *Cur. Op. Gen. and Dev.* 8:694-700.
- Smith HO, et al. (1995) Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* 269:538-540.
- Szostak JW, Bartel, DP and Luisi, L (2001) Synthesizing life. *Nature* 409:387-39.
- van Helden J, Andre B and Collado-Vides J (1998) *J. Mol. Biol.* 281:827-842.
- Yanai I, et al. (2000) Predictions of Gene Family Distributions in Microbial Genomes: Evolution by Gene Duplication and Modification. *Phys. Rev. Lett.* 85:2641-2644.
- Zipf GK (1949) Human Behavior and the Principle of Least Effort (Addison-Wesley, Boston).