UNIVERSAL LENGTHS IN COMPLETE MICROBIAL GENOMES

T.Y. CHEN[1], L.C. HSIEH[1,3], C.H. CHANG[1], L.F. LUO[4], F.M. JI[5] and H.C. LEE[1,2,3]

[1]*Department of Physics*[2]*, Department of Life Sciences and* [3]*Center for Complex Systems, National Central University, Chungli, Taiwan 320*

[4]*Department of Physics, University of Inner Mongolia Hohhot 010021, China*

[5]*Department of Physics, Northern JiaoTong University, Beijing 100044, China*

Received 30 October 2004

Statistical analysis of frequency occurrence of short words in complete genomes reveals the existence of a set of universal lengths common to all extant complete microbial genomes. This phenomenon is consistent with a model for genome growth in which primitive genomes grew mainly by maximally stochastic duplications of short segments from an initial length of about 200 nucleotides. The relevance of these results to the so-called RNA world in which life began and evolved before the rise of proteins is discussed.

Introduction. Genomes are books of life for organisms and necessarily carry huge amounts of information. By and large bigger genomes carry more information than smaller ones (there are noted exceptions). Yet as far as we know genomes grew and evolved stochastically, modulated by natural selection [1]. This raises a puzzling question: how does genomes stochastically grow *and* accumulate information simultaneously? This paper uses the set of all 108 sequenced complete microbial genomes as data in exploring ideas on randomness, entropy, information and growth with the aim of finding an answer. What emerges is the discovery of a set of universal lengths governed by a simple recursion formula that fits microbial genomic data. Attempts to understand the data motivated a model for the growth and evolution of genomes that provides a possible answer to our question.

In what follows we first discuss some properties of the relative spectral width of a distribution of occurrence frequencies for a set of random events, and of "replicas" of such sets, then show a simple relation between the relative spectral width and a definition of Shannon information for such sets. We then compute the reduced spectral widths of "k-spectra" of complete microbial genomes, present the results and interpret them by way of a growth model for genomes.

Frequency sets, their multiples, relative spectral width and Shannon information. Consider a set $\mathcal{S}$ of occurrence frequencies for τ types of events $\{f_i|\sum_{i=1}^{\tau} f_i{=}N\}{\equiv}\{f_i|N\}$, with mean frequency $\bar{f}{=}\langle f\rangle{=}N/\tau$ and standard deviation (std) $\Delta{=}\langle(f-\bar{f})^2\rangle^{1/2}$. If each frequency is increased by a factor of m then the mean and std for the new set $\mathcal{S}'{\equiv}\{f_i' = mf_i|mN\}$, an *$m$-multiple* of $\mathcal{S}$, will both increase by a factor of m while the *relative spectral width* will not change:

$\sigma' \equiv \Delta'/\bar{f}' = \sigma \equiv \Delta/\bar{f}$.

If the frequencies in $\mathcal{S}$ are of random events of equal likelihood, then the event probability versus frequency will be nearly a Poisson distribution (provided $N >> \tau$) and std $\Delta_{ran} = (b\bar{f})^{1/2}$, where b=1-τ^{-1}. That σ scales in this case as $N^{-1/2}$ for large N is the basis for a well known effect in thermodynamics: the average of some measure of a random system gains sharpness as the system gains size, and achieves infinite sharpness in the thermodynamic (large N) limit.

As we shall see, genomes exhibit the essence of this property

The relative spectral width defined given above is closely connected to a version of information in $\mathcal{S}$. Shannon expressed the information in a system in terms of decrease in uncertainty [2]. Shannon's uncertainty, or entropy, for the system $\mathcal{S}$ is $H = -\sum_i (f_i/N)\log(f_i/N)$. We define the *Shannon information* of the system as $R \equiv H_{max} - H$, which accords with the thermodynamical notion that an increase in entropy results in a decrease in information. We are interested in cases when most of the f_i's are non-zero, then H acquires its maximum value $\log\tau$ when all $f_i = \bar{f}$ and one finds for a unimodal distribution

$$R \equiv \log\tau - H = 0.5\ \sigma^2 + \mathcal{O}\left(\sigma^3\right) \approx 0.5\ \sigma^2 \qquad (\sigma << 1) \tag{1}$$

Reduced spectral widths and effective root-sequence lengths. Consider now single strands of DNA, or nucleotide sequences and view them as linear texts written in the four bases, or letters, A, C, G, T [3,4]. For a sequence of L nucleotides (nt) we denote by $\mathcal{S}_k$, or a k-spectrum, the set of frequencies $\{f_i|L\}_k$, where f_i is the occurrence frequency of the i^{th} k-letter word, or k-mer, that may be obtained by moving a sliding window of width k across the genome; τ=4^k and $\bar{f}$=$4^{-k}L$. To measure the information of a real genomic sequence relative to those expected of a random sequence of the same length (and base composition, see below) we define *reduced spectral width* for unimodal distributions to be:

$$\mathcal{M}_\sigma \equiv \sigma^2\bar{f}/b \approx (\sigma/\sigma_{ran})^2 \tag{2}$$

A random sequence is expected to have $\mathcal{M}_\sigma \approx 1$.

Suppose $\mathcal{Q}'$, an *m-replica*, is obtained by replicating m times a sequence $\mathcal{Q}$, then every k-spectrum $\mathcal{S}'_k$ of $\mathcal{Q}'$ is an m-multiple of the corresponding k-spectrum $\mathcal{S}_k$ of $\mathcal{Q}$. In particular, if $\mathcal{Q}$ is a random sequence, then we expect (when $m << L$) $\mathcal{M}_\sigma(\mathcal{S}_k) \approx 1$ and $\mathcal{M}_\sigma(\mathcal{S}'_k) \approx m$ to a high degree of accuracy, independent of k. This motivates the following: Given sequence $\mathcal{Q}$ of length L and k-spectrum $\mathcal{S}_k$, $L_r \equiv L/\mathcal{M}_\sigma(\mathcal{S}_k)$ is defined as the *effective root-sequence length* of $\mathcal{Q}'$ for k-mers. For as far as k-mers are concerned, $\mathcal{Q}$ (not necessarily an m-replica) has the same reduced spectral width as that of a random root-sequence of length L_r. Only an m-replica of a random sequence is expected to have L_r's independent of k for $k < \log L/\log 4$.

Complete microbial genomes. The 108 complete microbial genomes currently in the GenBank [5] are heterogeneous in length - 0.4 to 7 million bases (Mb) - and base composition - 25 to 75% A+T. In most cases the numbers of A's and T's (and

of C's and G's) in a genome are very similar. We therefore characterize the base composition of a genomes by a single parameter p, the combined probability of A and T, or C and G, whichever is greater.

In Fig. 1 the 5-spectra, or occurrence frequency distributions of 5-mers, of random sequences 2 Mnt long (green-dotted; with sharp peaks) with p equal to 0.5 (A), 0.6 (B) and 0.7 (C), respectively, are shown together with the per 2 Mnt 5-spectra (black) of three genomes with matching p values: *A. fulgidus* (A), *S. pneumoniae* (B) and *C. acetobutylicum* (C) [5]. In the plots, fractional numbers of 5-mers appear as a result of averaging of the ordinates over a small range of abscissa to smooth out excessive fluctuation of the spectra for better viewing. Only the spectra in (A) for the $p = 0.5$ sequences satisfy the unimodal requirement to make Eq.(1) directly applicable. In this case the random spectrum is indeed Poissonian and the genomic spectrum is much wider, $\sigma_{A.ful.}/\sigma_{ran} \sim 20$ so that information-wise, as far as 5-mers are concerned, a 2 Mnt stretch of the *A. fulgidus* genome is like the 400-replica of a random sequence merely 5 knt long.

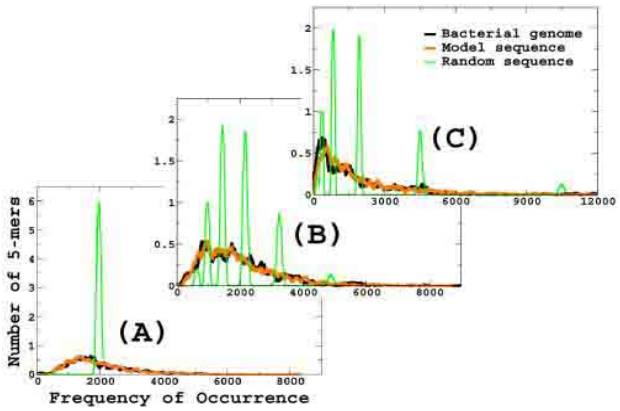

Fig. 1. Comparison of 5-spectra of genome (black), random (green-dotted; with sharp peaks) and model (orange/gray) sequences with p=0.5 (A), 0.6 (B) and 0.7 (C), respectively. The genomes are *A. fulgidus* (A), *S. pneumoniae* (B) and *C. acetobutylicum* (C). Fractional numbers of 5-mers appear in the ordinates as a result of curve-smoothing.

For random sequences with p>0.5 the single Poisson spectrum is split into $k+1$ smaller Poisson spectra (Fig. 1 (B) and (C)), one for each of the subsets of k-mers with m AT's (called m-sets), whose respective means are $\bar{f}_m(p) = \bar{f} 2^k p^m (1-p)^{k-m}$, m= 0 to k. We thus generalize the definition for $\mathcal{M}_\sigma$ given in Eq.(2) to be the weighted average over the reduced spectral width of the m-sets:

$$\mathcal{M}_\sigma \equiv \sum\nolimits_{m=0}^{k} L^{-1} \left(2^k (k,m) \bar{f}_m\right) \sigma_{k,m}^2 \bar{f}_m / b \tag{3}$$

where (k,m) is a binomial, $\sum_m 2^k (k,m) \bar{f}_m = L$. Since A and T (and C and G) are counted together and the number of each monomer in the sequence is fixed, the binomial factor is taken to be $b = 1 - 2^{1-k}$. To verify Eq. (1), we also define a

reduced Shannon information $\mathcal{M}_R$ where $\sigma^2_{k,m}$ in Eq. (3) is replaced by $R_{k,m}$, the Shannon information of the m-set.

Results. Fig. 2 shows log-log plots of $\mathcal{M}_\sigma$ and/or $\mathcal{M}_R$ versus sequence length L computed from a "genome" set composed of 108 complete microbial genomes [5] and two control sets with lengths and base compositions matching those in the genome set: a "random" set of random sequences and a "replica" set of 100-replicas of random sequences. The results for the control sets are shown in Fig. 2 (A) and (B). They are essentially independent of k, sequence length L and base composition and have the expected values: (A) $\mathcal{M}_R = 0.514\pm0.062$ - this verifies Eq. (1); (B) $\mathcal{M}_\sigma = 1.02\pm0.11$ for the random set and $\mathcal{M}_\sigma = 101\pm12$ for the replica set. Each set contains 972 pieces of data and in each plot about 50% of the error comes from data for k=2 ("□" in the figure) and 25% from k=3 ("△"). This is because $\bar{f}_{k,m}$ for these cases are very large and magnify fluctuations in $\sigma_{k,m}$ in Eq (3).

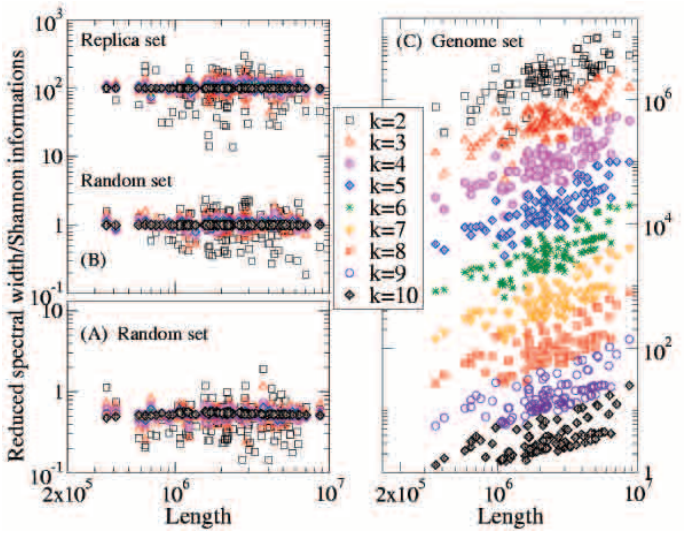

Fig. 2. (A) Reduced Shannon information ($\mathcal{M}_R$) for the random set; (B) reduced spectral information ($\mathcal{M}_\sigma$) for the random and replica sets; (C) $\mathcal{M}_\sigma$ for the genome set.

Fig. 2 (C) shows $\mathcal{M}_\sigma$ for the genome set, where each piece of data was multiplied by a factor of 2^{10-k} to delineate data into different k groups for better viewing. Still essentially p-independent, the data are otherwise entirely different from those of the control sets: (i) For given k they form a band (std is about 50% of mean) that depends linearly on L, implying that $L_r = L/\mathcal{M}_R$ is about the same for all genomes; namely L_r has a *universal* value. (ii) For given L, the mean $\log \mathcal{M}_R$ decreases approximately linearly with increasing k, such that the universal lengths (squares in Fig. 3) satisfy an approximate geometric recursion formula

$$L_r(k) = t\ L_r(k-1); \quad 3 \le k \le \min(10, \log L/\log 4) \tag{4}$$

with $L_r(2) = 300 \pm 180$ nt and $t = 2.64 \pm 0.20$.

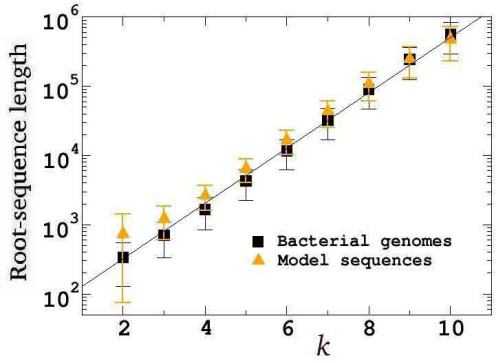

Fig. 3. L_r versus k extracted from reduced spectral width of the genome set (squares) and a set of 108 model sequences (triangles) whose lengths and base compositions match those in the genome set. Line shows mean of Eq. (4).

Model for genome growth and evolution. The universality of the L_r's suggests the existence of a universal mechanism for (microbial) genome growth from proto-genomes of a universal initial length. The very large values of $\mathcal{M}_\sigma$ (hence the shortness of the L_r's) for the smaller k's reminds us of the m-multiple system $\mathcal{S}'$ discussed earlier and suggests a mechanism for genome growth involving a great deal of replication or duplication. One obvious mechanism, growth simply by whole-genome replications [6,7] is ruled out because that would yield k-independent L_r's, contrary to data. The observed strong k-dependence of L_r suggests a more complex duplication process.

We now show that artificial genomes generated in a simple and biologically plausible growth model [8] possess properties similar to those of microbial genomes seen in Figs. 1, 2(C) and 3. In the model the initial condition of a genome is a random sequence of length L_0 with a base composition p. The condition $L_0 < L_r(2) \approx 300$ nt is necessary if the large values of $\mathcal{M}_\sigma$ for the small k's are to be attained. The genomes then grow by random short segmental duplications - or *quasireplication* - possibly modulated by random single mutations. The model shares some features with those used to explain the power-law behavior of the occurrence frequency of genes in genomes [9,10], except that there the units of duplication are genes, not the short oligonucleotides used here. The quasireplication process is maximally stochastic: a segment of length l, chosen according to the probability density function $f(l) = 1/(an!)(l/a)^n e^{-l/a}$, is copied from one site and inserted into another site, both randomly selected. The Erlang function $f(l)$ was chosen for convenience. The values $L_0 \approx$200 nt, n=2 and a=6.7 (nt) gave the best fit to data, implying a typical length of 20 ± 12 nt for the duplicated segments.

In Fig. 3 the L_r's (triangles) extracted from a set of 108 model sequences with length and base composition matching those in the genome set and generated *in*

silico by quasireplication are compared with the L_r's for the genome set (squares). The two sets of lengths essentially agree although those from the model sequences have a slightly weaker k-dependence. The k-spectra of 5-mers computed from three representative model sequences with p= 0.5, 0.6 and 0.7, respectively, are shown as orange/gray curves in Fig. 1. Our many trials to reproduce the data lead us to believe it is unlikely that any simple model not having the two main ingredients of our model, very short initial genome length (<300 nt) and random duplication of short segments, would be able to obtain the results of the quality shown in Fig. 3.

Discussion and conclusion. The answer to the question posed at the beginning of this articles may be this: It seems that by choosing the mode of quasireplication for growth and evolution, microbial genomes also adopted a superb strategy for information acquisition and accumulation. The choice, if it did happen, should be viewed as a result of natural selection, for it led to evolution at a rate much higher than would have obtained had genomes grew only by fitness-driven natural selection of random insertions. In choosing quasireplication the genomes left a clear evolutionary track: the universal root-sequence lengths. Unlike an m-replica, a quasireplica is globally aperiodic. For $k \leq k_{max} \approx \log L / \log 4$ a quasireplica acts as an m-replica where $m = L/L_r(k)$, while for $k >> k_{max}$ it appears essentially as a random sequence. Quasireplicas are partially ordered, highly complex and evidently capable of carrying large amounts of information. Our results show that all extant complete microbial genomes belong to a single universality class specified by Eq. (4). Quasireplicas are extremely robust. We have verified (but not shown here) that, segments as short as one-thousandth of a quasireplica belong to the same class as the parent quasireplica and, provided the typical duplicated segment length is significantly greater than k_{max}, quasireplication (including simple replication) upon a quasireplica begets a longer quasireplica of the same class. In this sense genomes seem to be genuine biological realizations of self-organized critical systems [11].

Acknowledgment. This work was supported in part by the grant 92-2119-M-008-012 from the National Science Council, ROC.

References

1. R. Dawkins, *The Blind Watchmaker*, (Penguin, 1988).
2. C. E. Shannon, Bell Sys. Techn. J. **27** (1948) 379.
3. R. N. Mantegna, *et al.*, Phys. Rev. Lett. **73** (1994) 3169.
4. S. Karlin and C. Burge, Trends in Gen. **11** (1995) 283.
5. The GenBank, http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html (Version January 26, 2003).
6. S. Ohno, *Evolution by Gene Duplication*, (Springer Verlag, New York, 1970).
7. A. L. Hughes, *et al.*, Genome Res. **11** (2001) 771.
8. L. C. Hsieh, *et al.*, Phys. Rev. Lett. **90** (2003) 018101.
9. I. Yanai, *et al.*, Phys. Rev. Lett. **85** (2000) 2641.
10. J. Qian, *et al.*, J. Mol. Biol. **313** (2001) 673.
11. P. Bak, *et al.*, Phys. Rev. Lett. **59** (1987) 381.