Evidence for Growth of Microbial Genomes by Short Segmental Duplications

Li-Ching Hsieh[1], Liaofu Luo[3] and H.C. Lee[1,2,4]
[1]*Department of Physics and* [2]*Department of Life Science, National Central University, Taiwan*
[3]*Department of Physics, Inner Mongolia University, Hohot, China*
[4]*Center for Complex Systems, National Central University, Taiwan*
hclee@phy.ncu.edu.tw

Abstract

Textual analysis of microbial genomes reveals footprints of their early evolution of the genomes. It is shown that distributions frequency occurrence of words less than nine letters in genomes have widths that are many times those of Poisson distributions. This phenomenon suggests a simple biologically plausible model for the growth of genomes: the genome first grows randomly to an initial length of approximately one thousand nucleotides (1 kb), or about one thousandth of its final length, thereafter mainly grows by random short segmental duplication. We show that using duplicated segments averaging around 25 b, model sequences generated in this model possess statistical properties characteristic of present day genomes. Both the initial length and the duplicated segment length support an RNA world at the time duplication began.

1. Introduction

Microbial genomes are seemingly random and very large systems when viewed as texts of the four nucleotides represented by A, C, G and T [1]. It is a general rule of statistics that very large nearly random systems have sharply defined average properties. Yet when we examine word frequencies (k-distributions) of short words 2 to 10 letters long (k-mers), we find they are very uncharacteristic of large systems, their frequency distributions (k-distributions) are much wider than those expected of random sequences. We analyze this phenomenon in detail and arrive at a model for early genome growth: maximally stochastic short segmental duplication from a very short initial random sequence.

2. Spectral widths of k-distributions

In Table 1 gives the standard deviations (stds) of the distribution of k-mers per 1 Mb, $k = 2$ to 10, for the some sequences that has approximate even base composition: column 2, of *Treponema pallidum* [2], the causative agent of syphilis; column 3, the average of fourteen complete genomic sequences with approximate even base composition [3]; column 4, of a random sequence with even base composition whose whose distributions are Poissonian. The std's of the genomic and random sequence have about the same magnitude when k is equal to or greater than 10 (not shown in the Table). But with decreasing values of k the Poisson std increases as 2^{-k} whereas the genomic std increases at a much higher rate, such that for $k \leq 8$ the Poisson std is many times less than the genomic std Moreover, the uncertainty in the genomic std is typically much smaller than mean std and the difference between the genomic and Poisson std's, we shall speak of a universal genome as one having the mean spectral widths of the genomes.

Table 1: Standard deviation of k-mer distributions.

k	*T. pal*	Genomes	Random	Model	$L_r(kb)$
2	8227	10580±2040	250	8207	.65±.35
3	3977	4080±630	125	3415	1.0±0.3
4	1384	1490±210	62.5	1202	1.9±0.5
5	434	469±66	31.2	402	4.7±1.3
6	129	141±21	15.6	134	13±4
7	37.5	41.9±6.7	7.8	45.3	37±12
8	11.0	12.4±2.3	3.9	15.9	110±40
9	3.4	3.84±0.84	1.9	5.9	300±130
10	1.3	1.33±0.34	1.0	2.3	640±300

3. Effective random-sequence length

We define the effective random-sequence length L_r of the universal genome as the length of a random sequence that has a k-mer distribution with a mean to std ratio equal to that of the corresponding genomic ratio r. Then L_r=$4^k r^2$, and its values for the various k's are given in the last column of Table 1. One notices that the L_r of the universal genome is very short for the smaller k's - of the order of 1 kb for $k \leq 3$ - and grows with k. When k=10, it is essentially the same length as the real genome.

4. Model for early genome growth

We propose a biologically plausible model for the growth and evolution of a universal genome that can generate the observed statistical characteristics of genomic sequences [4]. The model consists of two phases. In the first phase the genome initially grows to a random sequence whose size is much smaller than the final size of the genome. In the second phase the genome grows by random duplications modulated by random single mutations. In this work a snapshot is taken of the model genome shortly after it reaches a length of 1 Mb.

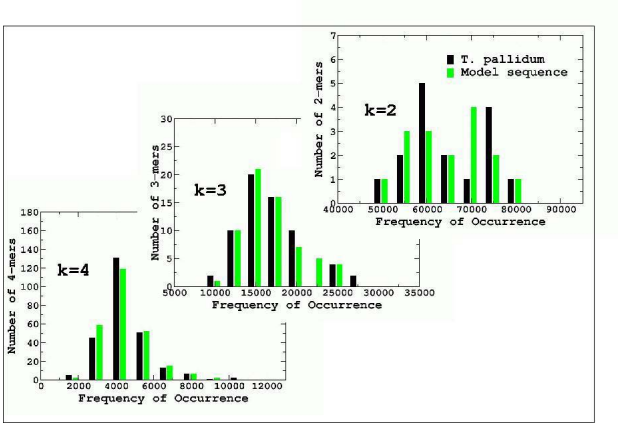

Figure 1: Histograms of k-distributions of genome of *T. pal.* (black) and model sequence (gray/green), k=2 to 4.

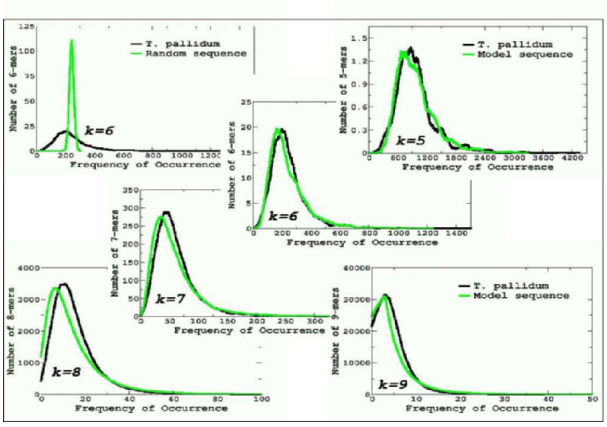

Figure 2: Comparison of k-distributions of genome of *T. pal.* (black) and model sequence (gray/green), k= 5 to 9. Top-left panel: *T. pal.* and random sequence, k=6.

In the model a random sequence of length L_0 is first generated. Thereafter the sequence is altered by single mutations (replacements only) and duplications, with a fixed average mutation to duplication event ratio. In duplication events, a segment of length l, chosen according to the Erlang probability density function $f(l) = 1/(\sigma m!)(l/\sigma)^m e^{-l/\sigma}$, is copied from one site and pasted onto another site, both randomly selected. In the above m is an integer and σ is a length scale in bases. After extensive experimentation, it was found that sequences having the statistical characteristics sought after could be generated by the model using the parameters $L_0 = 1000$, $m = 4$, $\sigma = 5$ and without mutation events. This implies an average length of 25 b width a std of 11 b for the duplicated segments. The resulting are given: as stds in column five of Table 1; as histograms of k-distributions for k=2, 3 and 4 in Fig. 1; as k-distributions for k=5 to 9 in Fig. 2. In general the model sequence seems to have the statistical property of the microbial genomes.

5. Discussion

Setting the initial length of our model universal genome before it began the growth by duplication process to about 1 kb but not much longer (as required by observed data) necessarily implies that the universal genome began its life in an RNA world [5, 6] in which there were no proteins and RNAs had the dual roles of genotype and phenotype. This implies the duplication most likely were carried out by ribzymes or their precursors, a likelihood consistent with the smallness of some of the ribozymes now extant; the hammerhead ribozyme is as small as 31 nt long [7]. Ribozymes this small enhances the code-copying efficiency of segments that are one average 25 b long. Our model does not address the origin of this initial genome but it could have arose spontaneously from pools of random RNA sequences, some of which happened to have encoded ribozyme [8]. In any case, being a natural way to repeatedly utilize hard-to-come-by codes, growth by duplication is in itself a brilliant strategy and must have increased the rates of evolution and species diversion enormously.

HCL thanks the National Science Council (ROC) for the grant NSC 91-2119-M-008-012.

6. References

[1] S. Karlin and C. Burge, *Dinucleotide relative abundance extremes: a genomic signature*. Trends in Genetics **11** (1995) 283-290.

[2] C.M. Fraser et al., *Complete genome sequence of Treponema pallidum, the syphilis spirochete*. Science **281** (1998) 375-388.

[3] GenBank: www.ncbi.nlm.nih.gov/PMGifs/Genomes /micr.html.

[4] L.C. Hsieh, Liaofu Luo, Fengmin Ji and H.C. Lee, *Minimal model for genome evolution and growth*, Phys. Rev. Lett. 90 (2003) 018101.

[5] W. Gilbert, *The RNA world*, Nature 319 (1986) 618.

[6] G. F. Joyce, *The antiquity of RNA-based evolution*, Nature 418 (2002) 214-221.

[7] *Forster AC, Symons RH. Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites*, Cell 49 (1987) 211-220.

[8] E.H. Ekland, J.W. Szostak and D.P. Bartel, *Structurally complex and highly active RNA ligases derived from random RNA sequences*, Science 269 (1995) 364-370.