Growth of microbial genomes by short segmental duplications

Li-Ching Hsieh*, Liaofu Luo‖ and H.C. Lee*†§

*Department of Physics and †Department of Life Science, National Central University, Chungli, Taiwan 320

‖Department of Physics, Inner Mongolia University, Hohot, China

§Institute for Theoretical Physics, Chinese Academy of Science, Beijing 100080, China

A DNA sequence can be analyzed as a text of four letters by counting the times each word in the set of k-letter words occurs in the text. If the text is random and long enough, then the frequencies of word occurrence are expected to obey a Poisson distribution. Examination of complete microbial genomes shows that for k less than 9, the distribution has a width many times the width of a Poisson distribution - 42, 24, 9 and 3.2 times, for k being 2, 4, 6 and 8, respectively. The cause of this phenomenon is not known. Here we propose a simple biologically plausible model for the growth of genomes to explain it: the genome first grows randomly to a length much shorter than its final length, thereafter mainly grows by random segmental duplication. We show that using an initial length of 1000 bases (1 kb) and duplicated segments with lengths averaging 25 b, one can generate a model sequence the size of microbial genomes - of the order of 1 Mb - that exhibits genomic statistical characteristics.

It is a general rule of statistics that very large systems have sharply defined average properties. If one million apples were dropped at random into sixty-four barrels, the number of apples falling into the barrels is governed by the Poisson distribution. In 95 of 100 cases, the number of apples in a barrel would deviate by less than 250 apples from the mean of 15,625 apples. This probability declines rapidly as the number of apples departs from the mean: there is a less than one in 10^{830} (10^{980}, respectively) chance that one barrel would get as many (few) as 24,000 (8,000) apples.

Microbial genomes are large and seemingly random systems when viewed as texts of the four nucleotides represented by A, C, G and T. They are of the order of 1 million bases long, yet disobey the large-system rule in an astonishing manner. To count the number of times each of the sixty-four trinucleotides, or 3-mers, occur in a genome-as-text is similar to counting apples in barrels. The genome of the bacterium *Treponema pallidum*, the causative agent of syphilis [1], shows how enormously a genome differs from a random sequence. This 1.14 Mb long, almost evenly composed genome has six 3-mers (CGC, GCG, AAA, TTT, GCA, TGC) occurring more than 24,000 times per 1 Mb and two (CTA, TAG) less than 8,000 times. Scrambling the genome sequence thoroughly reduces it to a random sequence obeying Poisson distribution.

T. pallidum is not exceptional in this respect. For the fourteen complete microbial genome sequences with approximately even base composition (see Methods), the observed standard deviation of the distribution of the frequency of occurrence (hereafter, simply distribution) of 3-mers per 1 Mb is 4,080±630 around the mean of 15,625. This deviation is about 32 times that of a Poisson distribution with the same mean.

Table 1: Standard deviation of k-mer distributions: for the genome of *T. pallidum*; averaged over 14 microbial genomes with unbiased base composition; of a random sequence with Poisson distribution; of the model genome described in text. Last column is the size of a random sequence with the observed ratio of mean count to standard deviation.

k	*T. pal*	Genomic average	Poisson	Present model	Eff. size (in kb)
2	8227	10580±2040	250	8207	0.56
3	3977	4080±630	125	3415	0.94
4	1384	1490±210	62.5	1202	1.8
5	434	469±66	31.2	402	4.4
6	129	141±21	15.6	134	12
7	37.5	41.9±6.7	7.8	45.3	35
8	11.0	12.4±2.3	3.9	15.9	100
9	3.4	3.84±0.84	1.9	5.9	260
10	1.3	1.33±0.34	1.0	2.3	540

Nor is the 3-mer exceptional in the k-mer-statistics of genomic sequences. In Table 1, the standard deviations of the distributions of k-mers per 1 Mb, $k = 2$ to 10, averaged over the fourteen genomic sequences and for a Poisson distribution are given in the third and fourth columns respectively. Statistically the genomic sequences begin to resemble a random sequence only when k becomes greater than 9. This implies that in genomic sequences there are many grossly over- and under-represented short oligomers, and that the shorter the oligomer, the more plentiful it is, and extreme its over- or under-representation. This phenomenon also occurs with progressively less frequency for k-mers where $k > 9$. Here we focus our attention on the $k \leq 9$ cases.

That a genome has deviations much greater than the Poisson deviations suggests that it has the statistical properties of a system much smaller than itself. This smaller size, or effective random-sequence length, given in the last column of Table 1, is estimated to be the size of a random sequence that has the ratio of the mean count to standard deviation equal to that observed in the average genomic distribution. One notices that the effective random-sequence length is very short for the smaller k's and grows with k. When k=10, it is essentially the same length as the real genome.

There are many known examples of individual oligonucleotides that exhibit extreme relative abundance. For dinucleotides this was noted to be common and has genome-wide consistency [2]; tetrapalindromes and hexapalindromes are almost always under-represented in bacteriophages and are systematically under-represented in bacteria where 4-cutting and/or 6-cutting restriction enzymes are common [3]; an 8-mer that appears as Chi sites, hotspots of homologous recombination, is highly over-represented in *E. coli* [4]; in the human pathogens *Haemophilus influenzae* [5, 6] and *Neisseria* [7] there are 9- and 10-mers functioning as uptake signal sequences that are vastly over-represented. The causes for these extreme cases are generally not known and, with the exception of the dinucleotides, these individual cases do not much affect the statistical properties of the genome.

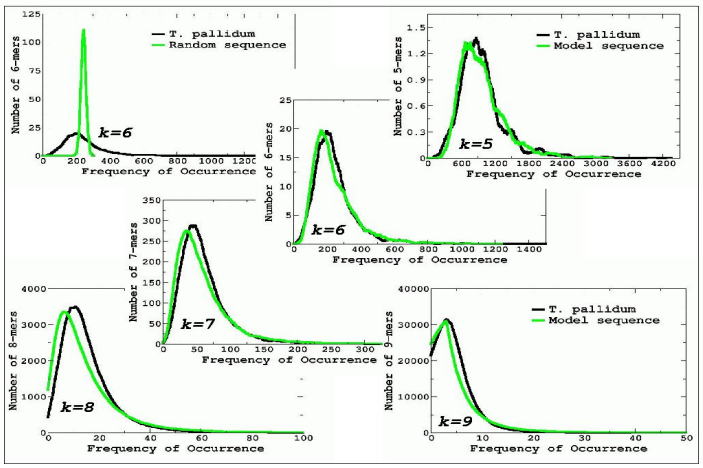

Figure 1: Comparison of k-mer distributions, k= 5 to 9. Black: *T. pallidum*; Gray: simulated sequence. Top-left panel: *T. pal.* and random sequence, k=6. Other panels: *T. pal.* and model sequence.

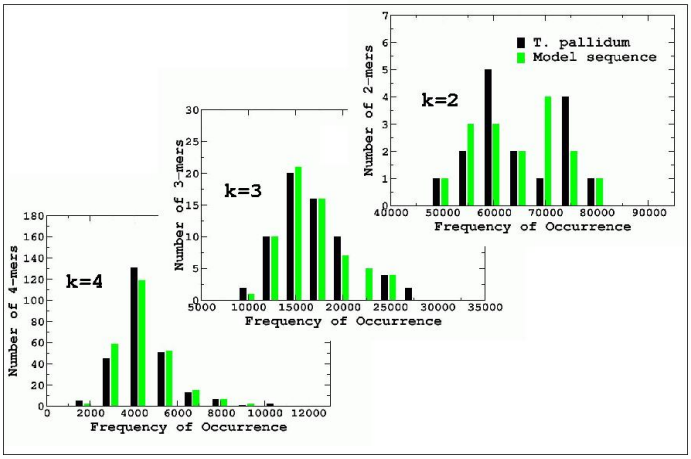

Figure 2: Histograms of k-mer distributions of genome of *T. pal.* (hatched) and model sequence (solid gray), k=2 to 4.

What caused a genome to have statistical characteristics so starkly distinct from those of a random sequence? Natural selection suggests itself as a prime explanatory candidate. For instance, the 64 frequencies of codons, 3-mers used by the genome to code proteins in genes, exhibit very wide distributions. But natural selection by itself does not directly cause any change in a genome. Such changes are caused by mutation and other mechanisms, all believed to occur at random. Natural selection may account for what changes come to pass; if, however, such changes always tend to promote or retain a randomness that exhibits Poisson distribution, then the ability of natural selection to push the genome very far in a non-Poisson direction seem to have its limits.

Here we propose a biologically plausible model for the growth and evolution of a genome that can generate the observed statistical characteristics of genomic sequences without the benefit of natural selection. The model is very simple and 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 acquires a length of 1 Mb.

Growth by self-copying or segmental duplication is one way in which a genome could gain size while retaining its small-system statistical characteristics. In one extreme, if the sequence grew by whole-sequence duplication, then the final product would wholly preserve the statistical characteristics of the original small sequence. In the other extreme, if the genome were to grow by duplicating very short - one or two bases long - segments, then the resultant product would be a random sequence with a Poisson distribution. Neither will produce a sequence that behaves like a microbial genome.

We found it comparatively easy to generate a sequence that could faithfully reproduce the genomic k-mer distribution of a particular k but not those of other k's. Typically such a sequence had an excessively rigid effective random-sequence length and, consequently, a distribution too narrow (broad) for smaller (greater) k's. Several such examples are given in the Methods. Generating a sequence that would emulate a real genome was a much more exacting task.

After extensive experimentation, it was found that sequences exhibiting the statistical characteristics sought after could be generated from an initial random sequence approximately 1 kb long (L_0) which was then grown to 1 Mb by random duplication of segments of length ($\bar{l}$) averaging 25 b with a spread (Δ_l) of approximately 11 b (see Methods for detail).

The standard deviations of the k-mer distribution of a good model sequence are given in column five of Table 1. They agree quite well with the observed genomic values in columns two and three although their k-dependence is slightly too strong. Fig. 1 shows comparisons between the k-mer distributions for k=5 to 9 of the genome of *T. pallidum* and those of the model sequence. The panel at the top-left corner compares the 6-mer distribution from *T. pallidum* with that of a random sequence obtained by scrambling the *T. pallidum* genome. The strong agreement between the microbial genome and the model sequence contrasts sharply with the glaring differences between the genome and the random sequence. Histograms in Fig. 2 show comparisons for k=2, 3 and 4. In all three cases, the histogram for a random sequence would be represented by a single tower located at the mean frequency. For k=2 and to a lesser extent k=3, the histograms for both genomic and model sequences display large fluctuations. The model sequence is not expected to exactly reproduce the counts of the genomic sequence. Indeed, generated stochastically, another (good) model sequence would give distributions indistinguishable from those shown in Fig. 1 but something rather different than those shown in the k=2 and 3 panels of Fig. 2. In any case, all model sequences would show patterns of fluctuation similar to those exhibited by the genomic sequence and have standard deviations similar to those given in column 5 of Table 1.

The model sequence is parameter-sensitive: If L_0 was 10 kb or longer no good model sequence was found; if either $\bar{l}$ or Δ_l was was changed by more than 10% from their

optimal values of 25 b and 11 b respectively the agreement between the genomic and model sequences would worsen noticeably (see Methods). No mutations were imposed on the model sequence whose properties are shown here; twenty thousand mutation fixations would reduce the standard deviations of the k-mer distributions of the model sequence by 4% (for k=2) to 10% (k=10) but under casual inspection the model sequence - with or without mutation - has the appearance of a random sequence. Results showing the model reproducing the k-mer distributions of microbial genomes with highly biased compositions will be presented elsewhere.

In bacterial genomes, typically about 12% of genes represent recent duplication events - 12% in *T. pallidum* [1], 11.2% in *H. influenzae* [8] and 12.8% in *V. cholerae* [9]. Our model sequence as presented here does not yet fully explain the pattern of all such duplications, many of which would involve segments up to several kb long. Work is under way to extend the model to account for the genomic pattern of repeat sequences of all lengths.

We mention some biological and evolutionary implications assuming our model does capture the essence of the growth mechanism of microbial genomes and, by extension, perhaps of all genomes. Of primary significance, the model greatly lessens the burden on natural selection as the only force that could have driven genomes uniformly so far in a non-Poisson direction. This in turn suggests that codon usage may not be the primary cause of the very broad distribution of the 3-mer counts seen in genomes. Indeed, since our ancestral genome might well have acquired the machinery for recombination when life was dominated by RNA and before codon was invented, the very biased codon usage seen today might well be an adaptation to the already-wide 3-mer distribution resulted from growth by duplication. The average duplication length of 25 b is about the length above which repeats in microbial genomes are rare, or the length of an oligonucleotide that can almost uniquely determine an organism. It is also the length that can encode a reasonable length of secondary structure motif in proteins. Growth by duplication is in itself a brilliant strategy as it would have increased the rate of evolution enormously; the preponderance of intra-genomic and inter-genomic homologous genes across all life forms [10, 11, 12] bears witness to its success.

Methods

The fourteen microbial genome sequences (length (L) in Mb and G+C probability (p) in brackets) *E. coli* K12 (4.64, .50), *E. coli* 0157 (5.52, .50), *M. thermoautotrophicum* (1.75, .50), *A. fulgidus* (2.18, .49), *T. pallidum* (1.14, .53), *X. fastidiosa* (2.67, 0.53), *V. cholerae* chromosomes I (2.96, .48) and II (1.07, .47), *Synechococcus sp.* (3.57, .48), *N. meningitidis* serogroup B strain MC58 (1.57, .52), *Y. pestis* (4.65, .48), *S. typhimurium* (4.86, .52), *S. enterica* (4.81, .52) and *P. aerophilum* (2.22, .51) are obtained from the GenBank [13]. Counting of k-mers is done by reading through a k-base wide window that is slid around the (circular) genome once. Counts are normalized to per 1 Mb and bias in base composition is corrected for by dividing the actual counts by the factor $L2^k p^n (1-p)^{k-n}$, where n is the total number of G's and C's in each k-mer.

Generation of model sequence. 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. The function gives a mean duplicated segment length $\bar{l} = (m+1)\sigma$ with standard deviation $\Delta_l = (m+1)^{1/2}\sigma$. The values $m = 0$ to 8 and selected values for σ from 3 to 15,000 were used. The model sequence compared with genomic sequences in the Figures 1 and 2 and in Table 1 was generated with $L_0 = 1000$, $m = 4$, $\sigma = 5$ and without mutation events. Fine-tuning to find the best parameters was not attempted. The following are some examples that gave very good distributions for specific k-mers but not generally; all were generated with $L_0 = 1000$ and $m = 0$: for 6-mer, $\sigma = 13,000 \pm 2,000$ and on average 0.04σ mutations per duplication (these parameters also work for genomes with biased base compositions) [14]; for 2-mer, $\sigma = 50$, no mutation; for 5-mer, $\sigma = 30$, no mutation; for 9-mer, $\sigma = 15$, no mutation.

Presentation of data. In Fig. 1 the curves shown are the result of a small amount of forward and backward averaging - to remove excessive fluctuations. In Fig. 2 data bunching were use to produce the towers shown.

1. C.M. Fraser et al., *Complete genome sequence of Treponema pallidum, the syphilis spirochete*. Science **281** (1998) 375-388.
2. S. Karlin and C. Burge, *Dinucleotide relative abundance extremes: a genomic signature*. Trends in Genetics **11** (1995) 283-290.
3. S. Karlin et al., *Statistical analyses of counts and distributions of restriction sites in DNA sequences*. Nucl. Acids Res. **20** (1992) 1363-1370.
4. T. Colbert, A.F. Taylor and G.R. Smith, *Genomics, Chi sites and codons: 'islands of preferred DNA pairing' are oceans of ORFs*. Trends in Genetics **14** (1998) 485-488.
5. H.O. Smith et al., *Frequency and distribution of DNA uptake signal sequences in the Haemophilus influenzae Rd genome*. Science **269** (1995) 538-540.
6. S. Karlin, J. Mrazek and M. Campbell, *Frequent oligonucleotides and peptides of the Haemophilus influenzae genome*. Nucl. Acid Res. **24** (1996) 42634272.
7. H.O. Smith et al., *DNA uptake signal sequence in naturally transformable bacteria*. Res. Microbiol. **150** (1999) 603-616.
8. Arabidopsis Genome Initiative, *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana*. Nature **408** (2000) 796-815.
9. J.F. Heidelberg, et al., *DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae*. Nature **406** (2000) 477-483.
10. W. H. Li, *Molecular Evolution*. (Sinauer Associates, 1997).
11. J.M. Smith, *Evolution Genetics*, (Oxford University Press, 1998).
12. S. Otto and P. Yong, *The evolution of gene duplicates*. Adn. Genetics **46** (2001) 451-483.
13. GenBank: www.ncbi.nlm.nih.gov/PMGifs/ Genomes/micr.html.
14. L.C. Hsieh, Liaofu Luo, Fengmin Ji and H.C. Lee, *Minimal model for genome evolution and growth*, to be published.

Acknowledgment

HCL thanks the National Science Council (ROC) for the grant NSC 91-2119-M-008-012 and members of the Redfield Lab and the Otto Lab, Department of Zoology, University of British Columbia, for discussion and Centre de Recherches Mathématiques, Université de Montréal, and Center for Theoretical Biology, Beijing University, for hosting visits.

Correspondence and requests for material should be addressed to HCL at Physics, NCU {e-mail: hclee@phy.ncu.edu.tw}.