# Short Segmental Duplication: Parsimony in the Growth of Microbial Genomes

**Li-Ching Hsieh[1], Liaofu Luo[4] and Hoong-Chien Lee[1−3,5]**

[1]Department of Physics and [2]Department of Life Sciences, National Central University, Chungli, Taiwan 320
[3]Center for Complex Systems, National Central University, Chungli, Taiwan 320
[4]Department of Physics, Inner Mongolia University, Hohot, China
[5]National Center for Theoretical Sciences, Shinchu, Taiwan

Received: May 11, 2003/Accepted:

**Abstract**.* We show that textual analysis of microbial complete genomes reveals telling footprints of their early evolution. If a DNA sequence considered as a text in its four bases is sufficiently random, the distribution of frequencies of words of a fixed length from the text should be Poissonian. We point out that in reality, for words less than nine letters complete microbial genomes universally have distributions that are uniformly many times wider than those of corresponding Poisson distributions. We interpret this phenomenon as follows: the genome is a large system that possesses the statistical characteristics of a much smaller random system, and certain textual statistical properties of genomes observable now are remnants of those of their ancestral genomes, which were much shorter than genomes today. This interpretation motivates a simple biologically plausible model for the growth of genomes: the genome first grew randomly to an initial length of not more than one thousand bases (1 kb), thereafter mainly grew by random short segmental duplications. Setting the lengths of duplicated segments to average around 25b, we have generated model sequences *in silico* whose statistical properties emulate those of present day genomes. The small size of the initial random sequence and the shortness of the lengths the duplicated segments both dictate an RNA world at the time growth by duplication began. Growth by duplication allowed the genome repetitive use of hard-to-come-by codes increasing thereby the rates of evolution and species diversion enormously.

*Corresponding author: HCL, hclee@phy.ncu.edu.tw.

**Key words:** Complete microbial genomes - Oligonucleotide frequency - Statistical analysis - Genome growth model - Short segmental duplication - Evolution - RNA world

## Introduction

It is a general rule of statistics that the larger the system the more sharply defined its average properties. When apples are randomly dropped into barrels, the distribution of apples in the barrels is governed by the Poisson distribution. If 1,024 apples were dropped into sixty-four barrels, there is a 5% chance that one of the barrels would have less than 8 or more than 24 apples. If 1 million apples were dropped into the barrels the chances that the number of apples received by any barrel falling outside the range of 14,600 to 16,600 would be exceedingly small, and there is a less than one in $10^{980}$ ($10^{830}$, respectively) chance that one barrel would get as few (many) as 8,000 (24,000) apples.

Microbial genomes are seemingly random systems when viewed as texts of the four bases represented by A, C, G and T. 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 the number of apples after they have been dropped into barrels. The genome of the bacterium *Treponema pallidum*, the causative agent of syphilis, is about 1M base pairs long and has almost even base composition [Fraser et al. 1998]. In an astonishing departure from what is expected of a sys-

Table 1: For given $k$'s, standard deviation of $k$-distributions from $p \approx 0.5$ sequences: for the genome *T. pallidum*; averaged over 25 Class A genomes; for a $p = 0.5$ random sequence; for a $p = 0.5$ model sequence (see text). All stds are normalized to correspond to a sequence with $p = 0.5$ (see Methods).

| $k$ | *T. pallidum* | Class A genomes | Random | Model |
|---|---|---|---|---|
| 2 | 8227 | 10610±2107 | 250 | 8207 |
| 3 | 3977 | 4379±707 | 125 | 3415 |
| 4 | 1384 | 1490±232 | 62.5 | 1202 |
| 5 | 434 | 468±72.5 | 31.2 | 402 |
| 6 | 129 | 141±22.3 | 15.6 | 134 |
| 7 | 37.5 | 41.6±7.0 | 7.8 | 45.3 |
| 8 | 11.0 | 12.3±2.3 | 3.9 | 15.9 |
| 9 | 3.4 | 3.76±0.85 | 1.9 | 5.9 |
| 10 | 1.3 | 1.29±0.34 | 1.0 | 2.3 |

tem of its size, the 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 restores it to a random sequence obeying Poisson distribution and the large-system rule.

*T. pallidum* is not exceptional in disobeying the large-system rule. For the twenty-five complete microbial "Class A" genomes whose combine probability $p$ for AT or CG content is 0.46 to 0.55, the observed standard deviation (std) of the distribution of the frequency of occurrence of 3-mers per 1 Mb (hereafter called 3-distribution) is 4,080±630 around the mean of 15,625. This is about 32 times the std of a Poisson distribution of the same mean that a random sequence would yield.

Nor is the statistics of 3-mers special in genomes. In Table 1, column 3 gives the std of the $k$-distribution, $k = 2$ to 10, averaged over the twenty-five Class A genomic sequences and column 4 gives the std for a Poisson distribution with mean value $10^6/4^k$. The genomic stds approach those of a random sequence when $k$ increases beyond 10. For $k$ less than 10, the Poisson std increases as $2^{-k}$ with decreasing $k$ whereas the genomic std increases at a much higher rate, such that already at $k$=8 the genomic std is many times greater than the Poisson std. Because the variance in the genomic std is typically much smaller than the difference between the genomic and Poisson stds, the genomic $k$-distribution differs from the Poisson distribution in a universal fashion. Hence we shall speak of a

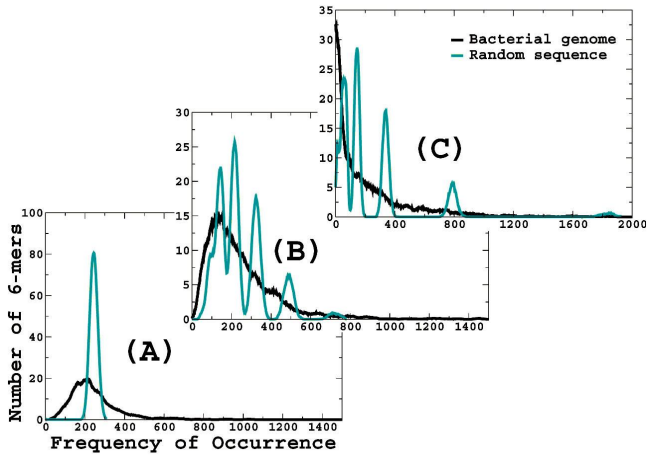


Figure 1: Comparison of 6-distributions of genomes (black) and random sequences (green/gray), with abscissa giving the frequency of occurences of 6-mers and ordinates showing the number of 6-mers having a given frequency: (A) *T. pallidum* genome and a random sequence with $p = 0.5$; (B) *C. muridarum* genome and a random sequence with $p = 0.6$; (C) *M. jannaschii* genome and a random sequence with $p = 0.7$

universal (Class A) genome.

The base composition of a genome has a conspicuous effect on its $k$-distributions. The black curve in Figure 1 shows 6-distributions of three representative genomes: (A) *T. pallidum* with $p \approx 0.5$; (B) *Chlamydia muridarum* [Read 2000] with $p \approx 0.6$; (C) *Methanococcus jannaschii* [Bult et al. 1996] with $p \approx 0.7$. For comparison the green/gray curves in the figure show 6-distributions of random sequences with $p = 0.5$, 0.6 and 0.7, respectively. The $k$-distribution of a random sequence is composed of $k$+1 Poisson distributions with mean frequencies $10^6\ 2^{-k} p^m (1\text{-}p)^{k-m}$, $m$=0 to $k$, which coalesce into a single Poisson distribution when $p$ is close to 0.5. In contrast, narrow sharp spikes are completely absent in the 6-distributions for the microbial genomes.

**Microbial genomes are large systems with small-system statistics**. The universal (Class A) genome has the statistical property of a random sequence much shorter in length than itself. To see this, we define the "statistical length" $L_{stat}$ of the universal genome as the length of a random sequence that has a $k$-distribution with a mean to std ratio equal to that of the observed genomic ratio $r$. Then $L_{stat}$=$4^k r^2$, and its values for the various $k$'s are given in column 2 of Table 2. Column 3 gives almost identical results extracted from "Class C", or $p \approx 0.7$, genomes (see Methods). $L_{stat}$ has a strong $k$ dependence: it is very short for the smaller $k$'s - of the order of 1 kb for $k \leq 3$ - and grows rapidly with $k$.

Table 2: Universal statistical lengths of microbial genomes. Genomes in Class A have $p \approx 0.5$ and those in Class C have $p \approx 0.7$.

| | $L_{stat}$ (kb) | |
|---|---|---|
| $k$ | Class A genomes | Class C genomes |
| 2 | $0.65 \pm 0.35$ | 0.53±0.30 |
| 3 | $1.0 \pm 0.3$ | 1.1±0.6 |
| 4 | $1.9 \pm 0.5$ | 2.1±1.1 |
| 5 | $4.7 \pm 1.3$ | 5.2±2.5 |
| 6 | $13 \pm 4$ | 14±6 |
| 7 | $37 \pm 12$ | 36±17 |
| 8 | $110 \pm 40$ | 93±44 |
| 9 | $300 \pm 130$ | 230±110 |
| 10 | $640 \pm 300$ | 600±240 |

When $k$=10 it is about half the length (normalized to 1 Mb) of the real genome.

A signature of the universal genome, by comparison to a random sequence, lies in its very large numbers of both overrepresented and underrepresented oligonucleotides. As a typical representative of the universal genome, the genome of *E. coli* [Blattner et al. 1997] has 500 and 510 6-mers whose frequencies of occurrence are greater than 400 and less than 100 per 1 Mb, respectively, while a 1 Mb random sequence has none in either category. On the other hand, a 1 Mb sequence obtained by replicating a 13 kb ($L_{stat}$ for $k$=6) random sequence 7.7 times would have overrepresented and underrepresented 6-mers as numerous as those in *E. coli*. There are many known examples of individual oligonucleotide that exhibit extreme relative abundance. For 2-mers this was noted to be common and has genome-wide consistency [Karlin and Burge 1995]; 4- and 6-palindromes are almost always underrepresented in bacteriophages and are systematically underrepresented in bacteria where 4-cutting and/or 6-cutting restriction enzymes are common [Karlin et al. 1992]; an 8-mer that appears as Chi sites, hotspots of homologous recombination, is highly overrepresented in *E. coli* [Colbert et al. 1998]; in the human pathogens *Haemophilus influenzae* [Smith et al. 1995, Karlin et al. 1996] and *Neisseria* [Smith et al. 1999] there are 9- and 10-mers functioning as uptake signal sequences that are vastly overrepresented. The causes for these extreme cases are generally not known and, with the exception of the 2-mers, such individual cases do not decisively determine the statistical properties of a genome.

What caused a genome to have $k$-distributions so much wider than 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 random mutations and other stochastic mechanisms. 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-Poissonian direction would seem to have its limits.

**Model for early genome growth**. Here 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. 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 segmental duplications possibly modulated by random single mutations. In this work a snapshot is taken of the model sequence shortly after it reaches a length of 1 Mb. The key aspect of the model is growth by segmental duplication, the most straightforward and biologically viable way for the universal genome to become what it appears to be - a large system that exhibits small-system statistical characteristics.

Growth by whole-genome duplication [Ohno 1970, Skrabanek and Wolfe 1998, Hughes et al. 2001] coupled with mutation is ruled out because such a mode of growth yields genomes whose $k$-distributions have the incorrect $k$-dependence - their $L_{stat}$ vary with $k$ too weakly. Indeed we found it comparatively easy to generate a sequence that could faithfully reproduce the genomic $k$-distribution for any given $k$, say $k = k'$, but not simultaneously those of other $k$'s. Typically such a sequence had an $L_{stat}$ that has a $k$-dependence far too weak than required to fit genomic data and, consequently, $k$-distributions that are too narrow when $k < k'$ and too broad when $k > k'$. Ways to generate several such examples are given in the Methods.

Table 3: Name, GenBank code, length and base composition of microbial complete genomes analyzed in the paper; $p_{AT}$ is the combined probability of A and T in the genome, and $p$ is the greater of $p_{AT}$ and $1-p_{AT}$. Top-half entries are the 25 Class A genomes with $0.46 \geq p \geq 0.55$ and the bottom-half are the 28 Class C genomes with $0.66 \geq p \geq 0.75$.

| Code | Name | length (M bp) | $p_{AT}$ | $p$ |
|---|---|---|---|---|
| NC_000853 | *Thermotoga maritima* | 1.86 | 0.54 | 0.54 |
| NC_000868 | *Pyrococcus abyssi* 1.76 | | 0.55 | 0.55 |
| NC_000911 | *Synechococcus sp. PCC 6803* | 3.57 | 0.52 | 0.52 |
| NC_000913 | *Escherichia coli K12* | 4.64 | 0.49 | 0.51 |
| NC_000916 | *Methanobacterium thermoautotrophicum* | 1.75 | 0.50 | 0.50 |
| NC_000917 | *Archaeoglobus fulgidus* | 2.18 | 0.51 | 0.51 |
| NC_000919 | *Treponema pallidum* | 1.14 | 0.47 | 0.53 |
| NC_002488 | *Xylella fastidiosa 9a5c* | 2.67 | 0.47 | 0.53 |
| NC_002505 | *Vibrio cholerae* chromosome 1 | 2.96 | 0.52 | 0.52 |
| NC_002506 | *Vibrio cholerae* chromosome 2 | 1.07 | 0.53 | 0.53 |
| NC_002578 | *Thermoplasma acidophilum* | 1.56 | 0.54 | 0.54 |
| NC_002655 | *Escherichia coli O157:H7* EDL933 | 5.52 | 0.50 | 0.50 |
| NC_002695 | *Escherichia coli O157:H7* | 5.49 | 0.49 | 0.51 |
| NC_003112 | *Neisseria meningitidis serogroup B strain MC58* | 2.27 | 0.48 | 0.52 |
| NC_003116 | *Neisseria meningitidis serogroup A strain Z2491* | 2.18 | 0.48 | 0.52 |
| NC_003143 | *Yersinia pestis strain CO92* | 4.65 | 0.52 | 0.52 |
| NC_003197 | *Salmonella typhimurium LT2 LT2* | 4.86 | 0.48 | 0.52 |
| NC_003198 | *Salmonella enterica subsp. enterica serovar Typhi* | 4.80 | 0.48 | 0.52 |
| NC_003364 | *Pyrobaculum aerophilum* | 2.22 | 0.49 | 0.51 |
| NC_003450 | *Corynebacterium glutamicum ATCC* 13032 | 3.31 | 0.46 | 0.54 |
| NC_004088 | *Yersinia pestis KIM* | 4.60 | 0.52 | 0.52 |
| NC_004113 | *Thermosynechococcus elongatus BP*-1 | 2.59 | 0.46 | 0.54 |
| NC_004337 | *Shigella flexneri 2a str. 30*1 | 4.60 | 0.49 | 0.51 |
| NC_004347 | *Shewanella oneidensis MR-1* | 4.96 | 0.54 | 0.54 |
| NC_004431 | *Escherichia coli CFT073* | 5.23 | 0.50 | 0.50 |
| NC_000908 | *Mycoplasma genitalium* | 0.580 | 0.68 | 0.68 |
| NC_000909 | *Methanococcus jannaschii* | 1.66 | 0.69 | 0.69 |
| NC_000962 | *Mycobacterium tuberculosis* | 4.41 | 0.34 | 0.66 |
| NC_000963 | *Rickettsia prowazekii strain Madrid E* | 1.11 | 0.71 | 0.71 |
| NC_001263 | *Deinococcus radiopugans R1* chromosome 1 | 2.64 | 0.33 | 0.67 |
| NC_001264 | *Deinococcus radiopugans R1* chromosome 2 | 0.412 | 0.33 | 0.67 |
| NC_001318 | *Borrelia burgdorferi* | 0.910 | 0.71 | 0.71 |
| NC_002162 | *Ureaplasma urealyticum* | 0.751 | 0.75 | 0.75 |
| NC_002163 | *Campylobacter jejuni* | 1.64 | 0.69 | 0.69 |
| NC_002516 | *Pseudomonas aeruginosa* | 6.26 | 0.33 | 0.67 |
| NC_002528 | *Buchnera sp. APS APS* | 0.640 | 0.74 | 0.74 |
| NC_002607 | *Halobacterium sp. NRC-1* | 2.01 | 0.32 | 0.68 |
| NC_002696 | *Caulobacter crescentus* | 4.01 | 0.33 | 0.67 |
| NC_002745 | *Staphylococcus aureus subsp. aureus N315* | 2.81 | 0.67 | 0.67 |
| NC_002755 | *Mycobacterium tuberculosis CDC1551* | 4.40 | 0.34 | 0.66 |
| NC_002758 | *Staphylococcus aureus strain Mu50* | 2.87 | 0.67 | 0.67 |
| NC_002771 | *Mycoplasma pulmonis* | 0.963 | 0.73 | 0.73 |
| NC_003030 | *Clostridium acetobutylicum ATCC824* | 3.94 | 0.69 | 0.69 |
| NC_003103 | *Rickettsia conorii Malish 7* | 1.27 | 0.68 | 0.68 |
| NC_003106 | *Sulfolobus tokodaii* | 2.69 | 0.67 | 0.67 |
| NC_003295 | *Ralstonia solanacearum* | 3.71 | 0.33 | 0.67 |
| NC_003296 | *Ralstonia solanacearum* | 2.09 | 0.33 | 0.67 |
| NC_003366 | *Clostridium perfringens* | 3.03 | 0.71 | 0.71 |
| NC_003454 | *Fusobacterium nucleatum subsp. nucleatum ATCC 25586* | 2.17 | 0.73 | 0.73 |
| NC_003888 | *Streptomyces coelicolor A3(2)* | 8.66 | 0.28 | 0.72 |
| NC_003923 | *Staphylococcus aureus subsp. aureus MW2* | 2.82 | 0.67 | 0.67 |
| NC_004061 | *Buchnera aphidicola str. Sg* | 0.641 | 0.75 | 0.75 |
| NC_004432 | *Mycoplasma penetrans* | 1.35 | 0.74 | 0.74 |

Generating a sequence that would emulate a real genome was a much more exacting task.

## Materials and Methods

**Complete microbial genome sequences** are obtained from GenBank [GenBank 2003]. The names, GenBank codes, lengths and base compositions of the genomes are listed in Table 3; $p_{AT}$ is the combined probability of A and T in the genome, and $p$ is the greater of $p_{AT}$ and $1-p_{AT}$. The 25 Class A genomes are given in the top half of the table and the 28 Class C genomes in the bottom half. Counting of $k$-mers is done by reading through a $k$-base wide overlapping sliding window. Counts are normalized to per 1 Mb and variation in genomic base composition is compensated for by dividing the actual genomic counts by the factor $L(p/\bar{p})^m((1-p)/(1-\bar{p}))^{k-m}$, where $p$ is the greater of the joint AT or CG probability, $\bar{p}=0.5$ ($\bar{p}=0.7$) for Class A (Calss C) genomes and $m$ is the total number of AT bases (or CG) in each $k$-mer.

**Generation of model sequence**. A random sequence of length $L_0$ with a given base composition 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 $f(l)=1/(\sigma n!)(l/\sigma)^n e^{-l/\sigma}$, is copied from one site and pasted onto another site, both randomly selected. In the above, $n$ is an integer and $\sigma$ is a length scale in bases. The function gives a mean duplicated segment length $\bar{l}=(n+1)\sigma$ with std $\delta_l=(n+1)^{1/2}\sigma$. The values $n=0$ to 8 and selected values for $\sigma$ from 3 to 15,000 were used. In the text, the model sequences used to compare with genomic sequences were generated with $L_0=1000$, $n=4$, $\sigma=5$ and without mutation events. This model has $\bar{l}=25$ and $\delta_l=11.2$. When $f(l)$ is replaced by a Gaussian distribution with the same values for $\bar{l}$ and $\delta_l$, respectively, less satisfactory results are obtained. Fine-tuning to find the best parameters was not attempted. The following are some examples that gave very good $k$-distributions for specific $k$-mers but not generally; all were generated with $L_0=1000$ and $n=0$: for 6-mer, $\sigma=13,000\pm2,000$ and on average $0.04\sigma$ mutations per duplication (these parameters also work for genomes with biased base compositions) [Hsieh et al. 2003]; for 2-mer, $\sigma=50$, no mutation; for 5-mer, $\sigma=30$, no mutation; for 9-mer, $\sigma=15$, no mutation.

**Sequences with highly biased compositions**. The standard deviation (std) in the $k$-distribution of a sequence with a significantly biased base composition is essentially determined by the value of $p$. This is because when $p$ is large (i.e., significantly greater than 0.5), the $k$-distribution of even a random sequence is spread out, as is seen in the green/gray curves of the (B) and (C) panels in Fig. 1. There the sharp peaks occur at the mean frequencies $\bar{f}_m$ of subsets of $k$-mers with $m$ AT's (called $m$-sets), $m$= 0 to $k$:

$$\bar{f}_m=\bar{f}2^k p^m(1-p)^{k-m},\quad m=0,1,\cdots,k \tag{1}$$

Table 4: Standard deviations of $k$-distributions from from $p\approx0.7$ sequences: column 2, for the genome of *M. jannaschii*; column 3, averaged over 28 $p\approx0.7$ Class C genomes; column 4, for a $p=0.7$ random sequence; column 5, for a $p=0.7$ model sequence. All genomic stds are normalized to correspond to a $p=0.7$ sequence (see Methods).

| $k$ | *M. jan.* | Class C Genomes | Random | Model |
|---|---|---|---|---|
| 2 | 38,700 | 37,300±5,900 | 36,700 | 36,700 |
| 3 | 12,600 | 12,200±2,060 | 11,700 | 11,800 |
| 4 | 3,930 | 3,830±680 | 3,520 | 3,580 |
| 5 | 1,180 | 1,150±220 | 1,020 | 1,060 |
| 6 | 347 | 339±69 | 292 | 311 |
| 7 | 101 | 100±22 | 82.8 | 91.8 |
| 8 | 29.5 | 29.3±6.8 | 23.3 | 27.6 |
| 9 | 8.62 | 8.58±2.13 | 6.68 | 8.74 |
| 10 | 2.60 | 2.60±0.67 | 2.01 | 2.99 |

where $\bar{f}=10^6 4^{-k}$ is the overall mean. When $p$ is away from 0.5 it is possible for $\bar{f}_m$ to be much greater or much less than $\bar{f}$. For a random sequence the distribution wthin a $m$-set is a Poisson distribution with mean $\bar{f}_m$. If the widths of such distributions are ignored then the std for the entire $k$-distribution is

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

$\Delta_k(p)$ is zero at $p$=0.5 but grows rapidly when $2p$ deviates from 1. This means that for a random sequence the std is given by the Poissonian value $\bar{f}^{1/2}$ when $2p\approx1$ but is given by $\Delta_k$ when $\max\{2p,1/2p\}$ is significantly greater than 1. This is verified by data shown in Table 4, where the stds for the genome *M. jannaschii*, those averaged over 28 Class C genomes, for a $p$=0.7 random sequence, and for a $p$=0.7 model sequence are given. The genomic results are normalized - by dividing the frequency of every $k$-mer by a factor $(L/10^6)(p/0.7)^m((1-p)/0.3)^{k-m}$ - to those for a sequence $10^6$ bases long with exactly $p$=0.7; $L$ is the actual length of the genome. The entries in the table are very accurately given by Eq. (2): $\Delta_k(0.7)$= 36,700, 11,700, 3,520, 1,020, 293, 82.5, 23, 6.4 and 1.8 for $k$= 2 to 10, respectively. Thus for sequences with highly biased base compositions the std of the entire $k$-distribution is not sensitive to details of the sequence that we are interested in.

***k*-distributions of *m*-sets**. Such details show up in widths of the distributions of $m$-sets. In a random sequence the distribution of such a set is still approximately Poissonian, with its std approximately equal to the mean frequency $\bar{f}_m$. The total number of $k$-mers in an $m$-set is $N_m=2^k(k,m)\bar{f}_m$ giving $\sum_m N_m=10^6$, where $(k,m)$ is the binomial. Table 5 shows for some $m$-sets: the mean frequency (column 2); std averaged over 28 Class C genomes (± its own std) (column 3); std averaged over 50 random sequences (column 4); std for a $p=0.7$ model sequence. In computing the genomic stds, the genomic $k$-mer frequencies per 1 Mb are first normalized to correspond to a sequence with $p$=0.7, then another overall normalization is made to set the total number of $k$-mers in

Table 5: Mean frequency and standard deviation of $k$-distribution for $m$-sets from $p \approx 0.7$ sequences: column 2, mean frequency $\bar{f}_m$ for the set; column 3, std (and its own std) averaged over 50 $p = 0.7$ random sequences; column 4, std averaged over 28 Class C genomes; column 5, std of $p = 0.7$ model sequence (see Methods).

| | | Standard deviation | | |
|---|---|---|---|---|
| $k, m$ | $\bar{f}_m$ | Class C genomes | Random | Model |
| 2,1 | 52,500 | 6,381±1,674 | 168±58 | 2,555 |
| 3,2 | 18,375 | 3,736±1,038 | 120±20 | 1,851 |
| 4,2 | 2,756 | 1,142±246 | 51.6±4.6 | 688 |
| 5,3 | 964 | 452±92.2 | 30.8±1.5 | 296 |
| 6,3 | 145 | 92.6±19.9 | 12.0±0.28 | 92.3 |
| 7,4 | 50.6 | 36.6±7.7 | 7.12±0.10 | 50.6 |
| 8,4 | 7.60 | 7.44±1.57 | 2.76±0.02 | 9.93 |
| 9,5 | 2.65 | 3.05±0.61 | 1.63±0.01 | 4.69 |
| 9,7 | 14.5 | 11.3±2.87 | 3.80±0.03 | 12.1 |
| 10,6 | 0.93 | 1.33±0.28 | 0.97±0.00 | 2.28 |
| 10,8 | 5.06 | 4.66±1.18 | 2.25±0.01 | 5.79 |

the $m$-set to $N_m$. This guarantees the genomic mean frequency of the $k$-mers in the $m$-set to be $\bar{f}_m$. In Table 5 it is noticed that for $k$= 2 and 3 the std for the random sequence is less than $\bar{f}_m^{1/2}$ as would be expected of a Poisson distribution. This reflects the fact that we have ignored a binomial factor that would reduce the std. Nevertheless in each case the std of average genomic $k$-distribution is either greater or much greater (for the smaller $k$'s) than the std of a random sequence. The relation between the stds of genomic and random sequences is now similar to that seen in Table 1 for the Class A genomes. The stds for the model sequence are in general agreement with the genomic values but has a $k$-dependence that is slightly too weak. Model sequences that fit the data better can be found but that is not the primary purpose of the present study.

**Computation of $L_{stat}$.** Denote the average genomic std for an $m$-set by $\Delta_{k,m}$ and that for random sequences by $\Delta'_{k,m}$, then the statistical length for the "universal" $p = 0.7$ genome is defined as the weighted average

$$L_{stat} = L \sum_{m=0}^{k} \frac{N_m}{L} \left( \frac{\Delta'_{k,m}}{\Delta_{k,m}} \right)^2 \qquad (3)$$

Errors for $L_{stat}$ are similarly computed. The results are given in column 3 of Table 2.

**Presentation of data**. In Figs. 3 and 5 the curves shown are the result of a small amount of forward and backward averaging - to remove excessive fluctuations. In Figs. 2 and 4 data bunching was used to produce the towers shown.

## Result

After extensive experimentation, it was found that sequences having the statistical characteristics sought after could be generated by choosing: (i) the length ($L_0$) of the initial random sequence to be approximately 1 kb; (ii) the average length ($\bar{l}$) of the (randomly chosen) duplicated segments to be 25b with a spread ($\delta_l$) of approximately 11b. It is emphasized that every step in the growth procedure is taken stochastically.

The stds of the $k$-distributions of a $p = 0.5$ model sequence thus generated are given in column five of Table 1. They agree quite well with the observed genomic values in columns 2 and 3 although their $k$-dependence is still slightly too weak. Histograms of $k$-distributions of *T. pallidum* (black) and the model sequence (orange/dark gray), $k$=2, 3 and 4, are compared in Fig. 2. In all three cases, the histogram for a random sequence would be represented by a single narrow 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 model sequence (generated using the same parameters) will yield histograms that differ in detail from those shown in the $k$=2 and 3 panels of Fig. 2 but show patterns of fluctuation still similar to those exhibited by the genomic sequence and have stds very close to those given in column 5 of Table 1. Fig. 3 shows comparisons for

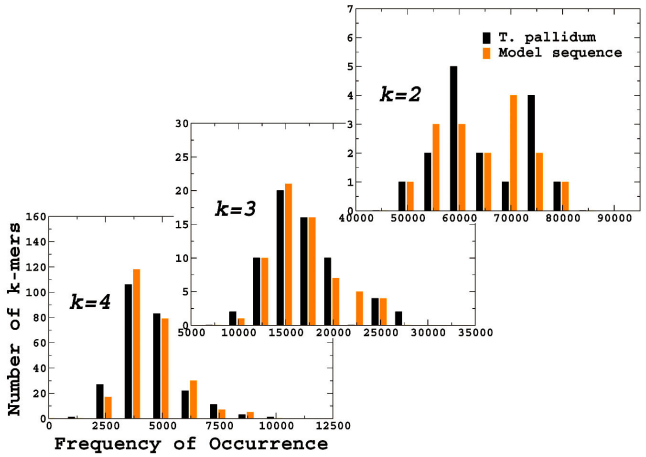


Figure 2: Histograms of $k$-distributions for genome of *T. pallidum* (black) and a $p = 0.5$ model sequence (orange/dark gray), $k$=2 to 4, with abscissa indicating intervals of frequency of occurrence of $k$-mers and ordinates giving the number of $k$-mers falling within a given interval of frequency of occurrence. In each case the histogram of the $k$-distributions for a random sequence would be represented by a single tower located at the mean frequency $10^6 4^{-k}$.

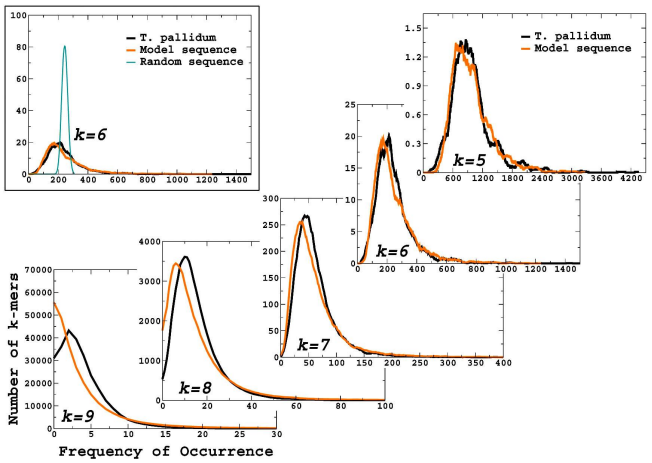


Figure 3: $k$-distributions for genome of *T. pallidum* (black) and a $p = 0.5$ model sequence (orange/dark gray), $k$= 5 to 9. See legend of Fig. 1 for further detail. The green/gray curve in the top-left panel is the 6-distribution for a $p = 0.5$ random sequence.

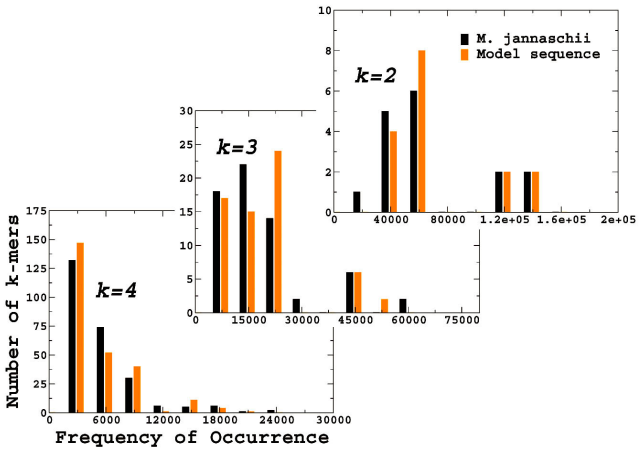


Figure 4: Histograms of $k$-distributions for genome of *M. jannaschii* (black) and a $p = 0.7$ model sequence (orange/dark gray), $k$=2 to 4. See legend of Fig. 2 for further detail. The histograms for a random sequence would be given by $k + 1$ narrow towers.

$k$=5 to 9. The panel in the top-left corner gives the 6-distributions from *T. pallidum*, a random sequence (green/gray) and the model sequence. In every case the model sequence succeeds in broadening out the narrow peaks that come with a random sequence and has $k$-distributions very similar to those obtained from *T. pallidum*.

The same growth model also can account for the $k$-distributions of genomes with $p$ significantly different from 0.5. We demonstrate this by comparing the $k$-distributions for *M. jannaschii* (black), which has $p \approx 0.7$, with those of a $p = 0.7$ model sequence (orange/dark gray) in Fig. 4 (histograms for $k$=2, 3 and 4) and Fig. 5 (distributions for $k$=5 to 9). The model sequence was generated using exactly the same procedure and parameters that generated the $p = 0.5$ model sequence (see Methods), except that the initial 1 kb random sequence has $p = 0.7$ rather than 0.5. The top-left panel of Fig. 5 shows that a $k$-distribution from a $p \neq 0.5$ random sequence (green/gray), with its $k$+1 narrow peaks, is entirely distinct from a genomic distribution.

Because the $L_{stat}$'s depend only on $k$ but not on the length and base composition of the genomes (Table 2), they are universal - same for all microbial genomes - lengths. In other words, being a large system with small-system statistics is a universal characteristic of microbial genomes. To summarize, the genomes: (a) have essentially identical sets of $L_{stat}$'s and (b) have $k$-distributions that are emulated by model sequences generated using identical parameters (but with predetermined $p$ values).

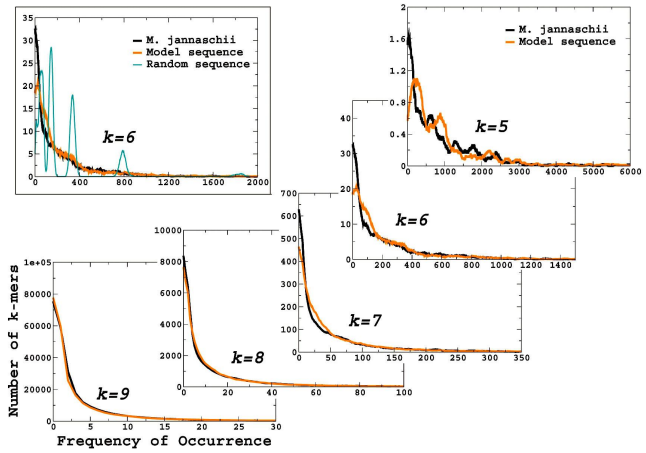


Figure 5: $k$-distributions for genome of *M. jannaschii* (black) and a $p = 0.7$ model sequence (orange/dark gray), $k$= 5 to 9. See legend of Fig. 1 for further detail. The green/gray curve in the top-left panel is the 6-distribution for a $p = 0.7$ random sequence.

The model sequences are parameter-sensitive. If $L_0$ was significantly longer than 1 kb then no good model sequence could be found. This is unsurprising because $L_0$ cannot be much longer than the shortest $L_{stat}$ in Table 2. If, with $L_0$=1 kb, either $\bar{l}$ or $\delta_l$ was changed by more than 10% from their optimal values of 25b and 11b respectively then the agreement between the genomic and model sequences would worsen noticeably.

## Discussion

In the spirit of simplicity no (point) mutations were imposed on the model sequences whose properties

are shown here. This is because mutations randomize a sequence whereas we are looking for a way to stochastically generate model sequences with less randomness than that of a random sequence. If the model sequence were allowed to have too many mutations then a compensating mechanism would have to be found to de-randomize the sequence. In reality the randomizing effect of mutation is partially compensated for by the stabilizing force of natural selection, whose effect is also not explicitly considered in our model. In any case the model sequences as given can tolerate about one mutation per two duplication events: twenty thousand mutation fixations reduces the std of the $k$-distributions of the Class A model sequence by 4% (for $k$=2) to 10% (for $k$=10).

In bacterial genomes, typically about 12% of genes represent recent duplication events - 12% in *T. pallidum* [Fraser et al. 1998], 11.2% in *H. influenzae* [Arabidopsis 2000] and 12.8% in *V. cholerae* [Heidelberg et al. 2000]. The model sequences presented here do not explain the pattern of all such duplications, many of which would involve segments up to several kb long. Work is underway to extend the model to account for the genomic pattern of repeat sequences of all lengths.

The generic statistical textual properties of eukaryotic genomes have also been examined and findings will be reported elsewhere. So much now is believed to obtain: when the great difference in length between microbial and eukaryotic genomes is accounted for, what is said here of the statistical textual properties of microbial genomes should hold true *mutatis mutandis* for eukaryotic genomes.

To be sure there will be many textual aspects of the generic microbial genome that the growth model proposed here will not be able to account for in detail. Nevertheless we believe the evidence presented here is sufficiently strong to support the following proposition: the ancestors of microbial genomes underwent a fundamental transition in their growth and evolution shortly after they had reached a length of not more than 1 kbp and had acquired a rudimentary duplication machinery, thereafter grew (and diverged) mainly by stochastic duplication of short segments whose lengths averaged to about 25b. Assuming this model to be substantially correct we mention some of its implications for biology and evolution.

Our results suggest that the base composition of a genome was essentially inherited from an ancestor whose own composition had been determined either randomly or by some unknown cause by the time of the transition to growth-by-duplication and that subsequent compositional changes caused by natural selection have been minor. If so, the base composition of a present-day genome should be close to being uniform over the entire genome because of the relatively small size of the ancestor genome - regardless of how it came into being - by comparison to present-day genomes. Indeed, this essentially holds true for genomes [Karlin and Burge 1995] (although coding regions often have a relatively richer GC content [Bult et al. 1996, Arabidopsis 2000, Lander et al. 2001, Venter et al. 2001]) and this phenomenon is of unknown biological significance.

A genome of the order of 1 kbp long is far too short to encode enough proteins for DNA duplication. Setting the initial length of our model universal genome to not greater than 1 kbp at the point of transition to growth-by-duplication thus necessarily implies that the universal genome began its life in an RNA world, when there were no proteins and when RNAs had the dual roles of genotype and phenotype [Gilbert 1986]. This view of the origin of life was advocated [Woese 1967, Crick 1968, Orgel 1968] even before RNA was discovered to exhibit self-splicing and enzymatic activities [Kruger et al. 1982, Guerrier-Takada et al. 1983]. Some RNA enzymes, or ribozymes, are very small; the hammerhead ribozyme is only 31 to 42 nucleotides (nt) long [Forster and Symons 1987] and the hairpin ribozyme is only 50 nt long [Hampel and Tritz 1989]. Hence we can infer with reasonable certainty that the 1 kbp initial universal genome was of sufficient size to encode the machinery necessary for sustained evolution and duplication (for many other issues of the RNA world see [Joyce 2002] for a review). Our model does not address the origin of this initial genome. The likelihood that it evolved from something arising spontaneously beforehand is enhanced by its short length and supported by the successful isolation of artificial ribozymes from pools of random RNA sequences *in vitro* [Ekland et al. 1995]. The average duplicated segment length of 25b likely represents a good portion of the length of a typical ribozyme encoded in the early universal genome even if it is very short

compared to a present-day gene that codes for an enzyme.

Natural selection has not been explicitly included in our model not because it is unimportant in evolution, but rather because we believe it could not have played a lead role in generating the non-Poissonian statistical characteristics of genomes discussed here. Suppose the genome grew not by segmental duplication but via "events" in which oligonucleotides were inserted at random into the genome (event specifics are unimportant to our argument) and at every such event the oligonucleotide was accepted or rejected through natural selection according to some preference causing a broadening of the $k$-distributions of the genome. If the oligonucleotides had been random, then there would have been many rejections between fixations and the number of events needed to generate a genome with $k$-distributions as wide as those of real genomes would have been orders of magnitude greater than the number of segmental duplications required to achieve the same effect. Against a scenario in which change was driven by natural selection, the principle of parsimony would dictate that segmental duplication was the overwhelming dominant force generating the wide $k$-distributions we now observe.

On the other hand, the effect of natural selection is implicitly included in our model, so far as it is a model for evolution. Natural selection determines whether a change brought upon by segmental duplication becomes a fixation. Furthermore, natural selection fine-tunes the fixations for adaptation.

As a corollary of the above reasoning, we can consider that uneven codon usage may not have been the primary cause of the very broad distribution of the 3-mer counts now seen in genomes. It is much more likely that codons were evolutionary "spandrels" [Gould and Lewontin 1979], that is to say, their rise as codes for proteins came as a consequence of an opportunistic evolutionary adaptation to the already-wide 3-distribution that had resulted from growth by duplication. Similarly, many of the highly under- or overrepresented oligonucleotides that now have biological functions might have originated as spandrels.

Our analysis presented here supports the view that statistical characteristics of present day genomes were already substantially determined by the characteristics of their ancestors by the time of their transition to growth by duplication; conversely, that statistical characteristics of genomes today can be regarded just in this manner as a basis from which to explore the nature and properties of early ancestral genomes; further analysis made along this line of reasoning may bring us a step nearer in understanding the universal ancestor [Woese 1998].

Growth by duplication is in itself a brilliant strategy because it allowed the genome to utilize hard-to-come-by codes repeatedly, thereby increasing the rates of evolution and species diversion enormously. For this strategy to have worked, the length of the duplicated segments used and the typical length of coding sequences must match. This condition is likely met by our model because most ribozymes in the early universal genome must have been small. Was this strategy continued after the "early life" of the universal genome - after the rise of codons and proteins? If so, then, with proteins/enzymes much larger than the small ribozymes, the duplicated segments in the post-protein era must have been much longer than 25b for the strategy to have been effective. In higher organisms many repeat sequences with lengths ranging from 1 base to many kilobases are believed to have resulted from at least five modes of duplication, and about 50% - perhaps even more - of the human genome is composed of such duplications [Lander et al. 2001, Venter et al. 2001]. Furthermore, as already mentioned, typically about 12% of genes in bacterial genomes represent recent duplication events. So certainly, the continuity of this strategy into the protein era is abundantly in evidence [Friedman and Hughes 2001, Bailey et al. 2002]. In eventual answers to questions such as why genes have been duplicated [Meyer 2003] at the high rate of about 1% per gene per million years [Lynch and Conery 2000], and why in all life forms so many duplicate genes are found [Maynard Smith 1998, Otto and Yong 2001, Gu et al. 2003], this growth strategy, if adopted universally in genomes, may be a simple and crucial, indeed a parsimonious part.

*Acknowledgments.* HCL thanks the National Science Council (ROC) for the grant NSC 91-2119-M-008-012, and Sally Otto, Rosie Redfield and Ceaga Lee for comments and discussions.

## References


Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796-815.

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.

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.

Crick FHC (1968) The origin of the genetic code. J. Mol. Bio. 38:367-379.

Ekland 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.

Fraser CM, et al. (1998) Complete genome sequence of Treponema pallidum, the syphilis spirochete. Science 281:375-388.

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.

Gould SJ and Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc. R. Soc. Lond. B 205:581-598.

Gu Z, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421:63-66.

Guerrier-Takada C, et al. (1983) The RNA moiety of RNAase P is the catalytic subunit of the enzyme. Cell 35:849-857.

Hampel A and Tritz RR (1989) RNA catalytic properties of the minimum (-)sTRSV sequences. Biochemistry 28:4929-4933.

Heidelberg JF, et al. (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406:477-483.

Hsieh LC, Luo LF, Ji FM and Lee HC (2003) Minimal model for genome evolution and growth. Phys. Rev. Lett. 90:018101-018104.

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.

Karlin S, Mrazek J and Campbell A (1996) Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. Nucl. Acid Res. 24:4263-4272.

Kruger K, et al. (1982) Self splicing RNA: autoexcision and autocyclization of the ribosomal intervening sequences of *Tetrahymena*. Cell 31:147-157.

Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lynch M and Conery LC (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.

Maynard Smith J (1998) Evolution Genetics. (Oxford University Press).

Meyer A (2003) Duplication, duplication. Nature 421:31-32.

Ohno S (1970) Evolution by gene duplication. (Springer Verlag, New York).

Orgel LE (1968) Evolution of the genetic apparatus. J. Mol. Bio. 38:381-393.

Otto S and Yong P (2001) The evolution of gene duplicates. Adn. Genetics 46:451-483.

Read TD, et al. (2000) Genome sequences of Chlamydia trachomatis MoPn and Chlamydia pneumoniae AR39. Nucl. Acids Res. 28:1397-1406.

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.

Smith HO, et al. (1999) DNA uptake signal sequence in naturally transformable bacteria. Res. Microbiol. 150:603-616.

Venter JC, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.

Woese C (1967) The Genetic Code. (Harper & Row, New York) 179-195.

Woese C (1998) The universal ancestor. Proc. Natl. Acad. Sci. USA 95:6854-6859.