Minimal Model for Genome Evolution and Growth

Li-Ching Hsieh,[1] Liaofu Luo,[2] Fengmin Ji,[3] and H. C. Lee[1,4,5]

[1]*Department of Physics, National Central University, Chungli, Taiwan 320*
[2]*Department of Physics, University of Inner Mongolia, Hohhot 010021, China*
[3]*Department of Physics, Northern JiaoTong University, Beijing 100044, China*
[4]*Department of Life Science, National Central University, Chungli, Taiwan 320*
[5]*Centre de Recherches Mathématiques, Université de Montréal, Montréal, Quebec, Canada*

(Received 8 June 2002; published 3 January 2003)

Textual analysis of typical microbial genomes reveals that they have the statistical characteristics of a DNA sequence of a much shorter length. This peculiar property supports an evolutionary model in which a genome evolves by random mutation but primarily grows by random segmental duplication. That genomes grew mostly by duplication is consistent with the observation that repeat sequences in all genomes are widespread and intragenomic and intergenomic homologous genes are preponderant across all life forms.

DOI: 10.1103/PhysRevLett.90.018101 PACS numbers: 87.10.+e, 02.50.–r, 87.14.Gg, 87.23.Kg

When a genome is viewed as a linear text composed of the four "letters" A (adenine), C (cytosine), G (guanine), and T (thymine), it appears sufficiently random for the genomes to be described as being made by a "blind watchmaker" [1]. In spite of all that is known about genomes, it is still a challenge to delineate coding parts of a genome sequence from noncoding parts, especially when the effort cannot benefit from sequence similarity to other known coding sequences [2].

There is actually not a clear understanding of the randomness of a genomic sequence. Consider the distribution of the frequency (of occurrence) of an oligonucleotide of length k, hereafter called a k-mer distribution [3–5]. The frequency of a k-mer is the number of times it is seen through a window of width k made to slide once across the genome. When the sequence length L is much greater than 4^k, the k-mer distribution for a "simple random sequence" grown randomly one nucleotide at a time would have a Poisson distribution. Thoroughly scrambling any genome would yield a simple random sequence (of a given base composition), which is what is usually referred to as a random sequence in the literature.

Figure 1(a) shows the (density of the) 6-mer distribution in a simple random sequence of length 1×10^6 bases (1 Mb) with unbiased base composition. The mean frequency of 244 and root-mean deviation of 15.5 are characteristic of a Poisson distribution. We focus on $k = 6$ in the first instance because in this case the number of k-mers and the mean frequency are both sufficiently large for good statistical sampling, making it more meaningful to disregard the counts of individual k-mers. For small k, the abundance of the 16 dinucleotides in genomes is widely disparate [4], and counts of the 64 trinucleotides, or codons, are well known to be biased in coding regions of genomes. For $k > 9$, the average frequency is too small for the present statistical analysis. Figure 1(b) is the distribution obtained from the complete genome of *Escherichia coli* [6] with approximately 50% A + T content (genome sequences are taken from the GenBank [7]; unless explicitly mentioned otherwise genomic frequencies of k-mers are normalized to correspond to those of a 1 Mb long sequence). The large ratio of the root-mean deviation of the genomic 6-mer distribution to that of a Poisson distribution is typical of microbial genomes with an unbiased base composition. The difference persists for all short oligonucleotides. For example, the ratios of the root-mean deviations for k-mer distributions of the *E. coli* genome to those of a 4.6 Mb long random sequence are 78, 73, 50, 32, 19, 11, 6.4, and 3.8, for $k = 2$ to 9, respectively. We know of no previous explanation of this striking disparity between genomic and random sequences.

The disparity is equally striking when the base composition is biased. The distributions in Fig. 2 are those of (a) a simple random sequence and (b) the complete genome of *Methanococcus janaschii* [8], both with approximately 70% A + T content. The single narrow peak seen in Fig. 1(a) is now broken into seven smaller peaks in

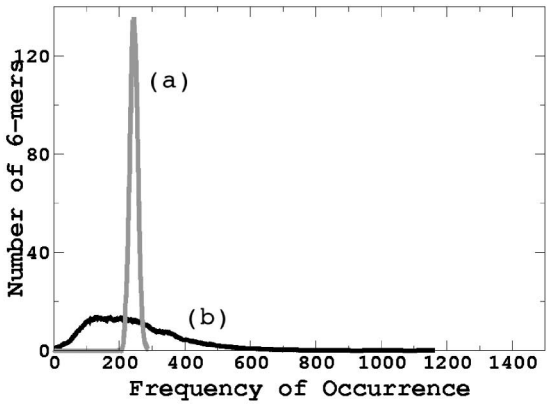

FIG. 1. Density of 6-mer distribution of (a) a simple random sequence and (b) the genome of *E. coli*, both with approximately 50% A + T content.

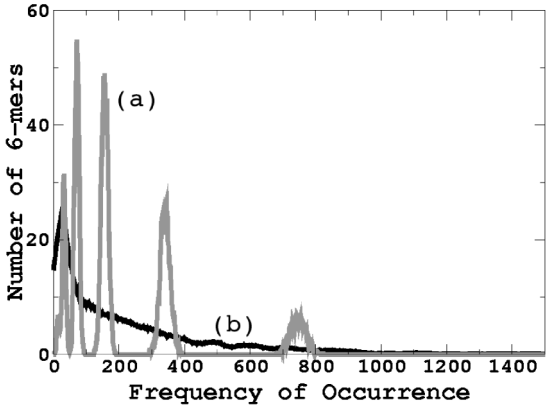

FIG. 2. Density of 6-mer distribution of (a) a simple random sequence and (b) the genome of *M. janaschii,* both with approximately 70% A + T content.

Fig. 2(a) whose appearance is caused by the bias in the base composition; the mean frequency of 6-mers with m A or T's is $244 \times (7/5)^m(3/5)^{6-m}$, giving the positions of the seven peaks to be 11.4, 26.6, 62.0, 144, 337, 787, 1837 (off scale), for $m = 0$ to 6, respectively.

Some of the known examples of oligonucleotides whose counts in genome sequences deviate with probable biological causes significantly from an expected random distribution are as follows: tetrapalindromes and hexapalindromes are almost always under-represented in bacteriophages and are systematically under-represented in bacterials where 4-cutting and/or 6-cutting restriction enzymes are common [9]; an 8-mer that appears as a chi site, hotspot of homologous recombination, is highly over-represented in *E. coli* [10]; in the human pathogens *Haemophilus influenzae* [11,12] and *Neisseria* [13] 9- and 10-mers that function as uptake sequence signals are vastly over-represented in their respective genomes. Frequencies of shorter oligomers are expected to, and do, influence those of longer oligomers. However, the correlation deviates significantly from random association, often in subtle ways [10]. In the case of *E. coli*, we have verified that random association of codons does not explain the majority of the 500 (510) 6-mers with counts greater than 400 (less than 100), a threshold more than 10 (9) standard deviations away from the expected value [see Fig. 1(b)]. In this work we therefore ignore the role of natural selection as it directly impinges on specific individual oligonucleotides.

Rather we choose to address the following question: Is there a simple, biologically motivated mechanism for stochastic genome growth that can reproduce the statistical characteristics of the observed k-mer distributions in bacterial genomes? This Letter identifies a potential mechanism and reports on some preliminary encouraging results.

Although bacterial genomes are of the order of 1 Mb long, the observed ratio of the mean of the 6-mer distribution to its root-mean deviation suggests the statistical property of a much shorter sequence, perhaps as short as 10 kb. The 6-mer distribution of a 10 kb simple random sequence would have about 3310 of the 6-mers occur 1 to 4 times, 3 to 4 occur 9 times, about one occurs 10 times, and about 350 not occurring altogether. Suppose we now duplicate this simple random sequence 100 times to produce a 1 Mb long sequence and let it undergo a number of single base mutations; then we can expect the long sequence to have a 6-mer distribution that begins to resemble Fig. 1(b). That is, it should have many 6-mers occurring more than 400 many times, some occurring close to 1000 times, and many occurring fewer than 100 times.

While it has been conjectured that great leaps in evolution had been the result of whole genome duplications [14], the present state of gene sequence information from vertebrates makes it difficult to either prove or disprove this hypothesis [15,16]. On the other hand, there certainly have been a very large number of events of gene duplications [17]. Indeed most genomes have repetitive sequences (or repeat sequences) with lengths ranging from 1 base to many kbs whose numbers of copies far exceed those that would be found in a simple random sequence. For example, in the human genome repeat sequences account for at least 50% and probably much more [18,19], and the fraction of genes that represent recent duplication events is 11.2% in *H. influenzae* [20].

Here, we propose a minimal model for microbial genome growth that incorporates duplication of DNA of all lengths and that exhibits the observed 6-mer (and to some extent other k-mers) distributions of real genomes. The model employs the two types of fixation events that drives genomic changes, mutation, and DNA duplication. For simplicity mutation fixations are represented by single base replacement (SBR). DNA duplication fixations are represented by occasional random duplication (RD) of a stretch of oligonucleotide with a characteristic length of σ bases.

In the model genomes are single stranded and the initial state of a genome is a simple random sequence of length L_0 with a given base composition. From the initial state the genome evolves and grows by SBR and RD fixations that preserve base composition until its length just exceeds 1 Mb. In an RD event, the length l of the duplicated segment is first randomly chosen (see below); then a site p at least l sites from the end of the genome is randomly chosen and the segment from site p to $p + l - 1$ is copied and inserted into the genome at a second randomly chosen site. The model has three parameters: the initial length L_0, the ratio η of the chances of having an SBR to an RD fixation, and the length scale σ. For the work reported here L_0 was held fixed at 1000 and only the two parameters η and σ were varied.

If the probability per unit length of selecting a segment of length l is $p(l)$ and the current length of the genome is L_c, then $\int_0^{L_c} p(l)dl = 1$. We assume $p(l) \propto e^{-l/\sigma}$ for simplicity. This has the property that segments of shorter lengths are more frequently duplicated than segments of

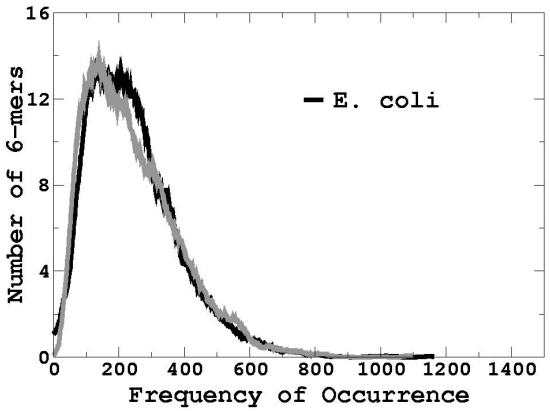

FIG. 3. Density of 6-mer distribution of the genome of *E. coli* (50% A + T content) (black) and a simple random sequence (50% A + T content) including segmental duplication mechanism with $\eta = 500$ and $\sigma = 15\,000$ (gray).

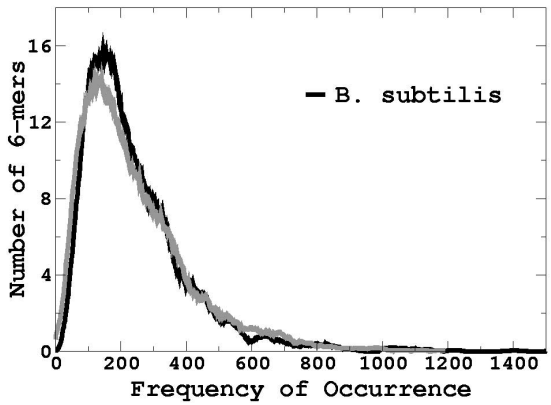

FIG. 4. Same as Fig. 3. Black: *B. subtilis* (60% A + T content); gray: model sequence with $\eta = 600$ and $\sigma = 15\,000$.

longer lengths and that σ is the average length of duplicated segments after the genome has grown to be significantly longer than σ. There are many other possible candidates for $p(l)$ and a survey of them will be reported elsewhere. With $p(l)$ given above and the relation $\int_0^l p(x)dx = y$ the segment length corresponding to a random number $0 \le y \le 1$ is $l = -\sigma \ln[1 - y(1 - e^{-L_c/\sigma})]$. Note that this model does not contain anything that directly acts on the distribution of any specific oligomer.

It turns out that if the model sequence is to have a 6-mer distribution similar to those of the representative real microbial genomes, the total number of mutations (for a sequence of canonical length 1 Mb) acting on the model sequence needs to be around 40 000. From the discussion in the previous section, this implies the relation $\eta \approx 0.04\sigma$ should hold. The best results are obtained when $\sigma \approx 15\,000$. In Fig. 3 the model genome with an unbiased base composition generated with the parameters $\eta = 500$ and $\sigma = 15\,000$ is seen to have a 6-mer distribution (gray) surprisingly similar to that of *E. coli* (black). No attempts were made to fine-tune the two parameters to get a "perfect" fit. The required 40 000 mutations in a genome of 1 Mb implies that the average interval between two neighboring mutation sites is about 25 bases long. Indeed, a model genome generated with $\eta = 0$ and $\sigma = 25$; that is, it has only duplication but no mutation events and has a 6-mer distribution similar to that shown in Fig. 3.

In Fig. 4 (5, respectively) the distributions for the model genome (gray) generated with $\eta = 600$ and $\sigma = 15\,000$ and for the genome (black) of *Bacillus subtilis* [21] (*M. janaschii,* respectively) are compared; both have approximately 60% (70%) A + T content. The peaks caused by biased base composition that one expects to see in a Poisson distribution [and seen in Fig. 2(a)] are no longer evident in the distributions from the model genomes, just as they do not show in the distributions from real genomes. In particular, the model seems to succeed with ease in accounting for the very large number of 6-mers that occur with exceptionally high and with exceptionally low frequencies seen in most microbial genomes.

The 6-mer distributions of microbial genomes are well represented by the two-parameter gamma distribution: $D(y) = y^{\alpha-1}\beta^{-\alpha}e^{-y/\beta}/\Gamma(\alpha)$. The distribution has mean $\langle y \rangle = \alpha\beta$ and root-mean deviation $\Delta = \alpha^{1/2}\beta$. In Table I the nth order deviations, defined as $\Delta^{(n)} = (\langle (y - \langle y \rangle)^n \rangle)^{1/n}$, n from 2 to 5, of 6-mer distributions of real genomes are compared with those of (a) the gamma distribution with the parameters α and β (in brackets) obtained from the real genome distribution, (b) the 6-mer distribution of a simple random sequence without duplication, and (c) the 6-mer distribution of the corresponding sequence given by the minimal model shown in Figs. 3–5. The data shown show that (i) the gamma distribution is a good representation of the distributions of the real genomes and (ii) distributions from the real and model genomes are similar to a high degree.

While the model genome sequences presented here reproduce the observed 6-mer distributions very well, they are less satisfactory in reproducing the distributions for other values of k, although in all cases (possibly

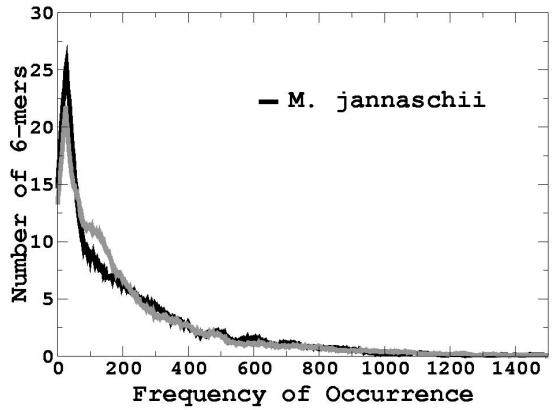

FIG. 5. Same as Fig. 3. Black: *M. janaschii* (70% A + T content); gray: model sequence with $\eta = 600$ and $\sigma = 15\,000$.

TABLE I. High order deviations

Sequence	$\Delta^{(2)}$	$\Delta^{(3)}$	$\Delta^{(4)}$	$\Delta^{(4)}$
E. coli (50% A + T content)	140	147	213	252
(a) ($\alpha = 3.05$; $\beta = 80.0$)	140	146	208	243
(b)	15.6	3.6	20.7	10
(c) ($\eta = 500$; $\sigma = 15$ K)	144	148	212	247
B. subtilis (60% A + T)	168	223	316	400
(a) ($\alpha = 2.12$; $\beta = 115$)	168	186	261	310
(b)	79	68	109	117
(c) ($\eta = 600$; $\sigma = 15$ K)	169	194	266	311
M. janaschii (70% A + T)	320	465	650	810
(a) ($\alpha = 0.58$; $\beta = 418$)	320	439	609	767
(b)	264	369	500	603
(c) ($\eta = 600$; $\sigma = 15$ K)	321	462	635	783

except for $k = 9$) the model sequence is much closer to the real genome than is a simple random sequence. Typically, they underestimate (overestimate) the spreading of the distribution for $k < 6$ ($k > 6$). The search for a possibly more elaborate model that gives a global account of the distributions of all short oligonucleotides is under way and results will be reported elsewhere.

Although the minimal model we have now is not exactly correct, we believe it captures the essential mechanism for genome growth: stochastic segmental duplication. That mechanism itself may be viewed as a force of natural selection. Because the probability that a random stretch of DNA would be a gene is so minuscule, a population of genomes that stumbled upon a duplication mechanism would have had an enormous evolutionary advantage over another population that did not. The preponderance of intragenomic and intergenomic homologous genes [17,22] across all life forms [6,8,18–21] is testament to the importance of this mechanism.

We have pointed out that the number of over- or under-represented short oligomers observed in all known microbial genomes is very large compared to what is expected of a simple random sequence, probably much larger than the number of such oligomers associated with specific biological functions, and we have proposed a model where genomes having this observed property are the result of a stochastic process. The implication is that many over- and under-represented oligomers might not have become so as a result of natural selection. Indeed, natural selection being an opportunistic process, some of the oligomers that do have specific biological functions might well have been recruited for such functions precisely because they were over- or under-represented in the first place. The model also has several interesting implications. It predicates both of the competing modes of evolution: by the gradual changes of classical Darwinism, and by stochastic spurts of various sizes, as in "punctuated equilibrium" [23,24]. It is resonant with the notion that likely a large fraction of genes within a genome are ultimately related by descent to a small number of genes that arose early in our evolutionary history [25]. That a present-day long genome may share a vital characteristic of its theoretical shorter earlier self implies one knows something about its ancestor, and by extension the common ancestor of its relatives. Perhaps, by pushing this notion harder and examining genomes closer, one may gain a deeper understanding of our universal ancestor [26].

H. C. L. thanks the National Science Council for Grant No. NSC 90-2119-M-008-019, members of the Redfield Laboratory and the Otto Laboratory, Department of Zoology, University of British Columbia, and members of the Center for Theoretical Biology, Beijing University for useful discussion, and Institute for Theoretical Physics, Chinese Academy of Science for hospitality to conclude this work.

[1] R. Dawkins, *The Blind Watchmaker* (Penguin, London, 1988).
[2] J. B. Hogenesch *et al.*, Cell **106**, No. 4, 413 (2001).
[3] C. Burge *et al.*, Proc. Natl. Acad. Sci. U.S.A. **89**, 1358 (1992).
[4] S. Karlin and C. Burge, Trends Genetics **11**, 283 (1995).
[5] L. F. Luo *et al.*, Bull. Math. Biol. **57**, 527 (1995).
[6] F. R. Blattner *et al.*, Science **277**, 1453 (1997).
[7] www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html.
[8] C. J. Bult *et al.*, Science **273**, 1058 (1996).
[9] S. Karlin *et al.*, Nucl. Acids Res. **20**, 1363 (1992).
[10] T. Colbert *et al.*, Trends Genetics **14**, 485 (1998).
[11] H. O. Smith *et al.*, Science **269**, 538 (1995).
[12] S. Karlin, J. Mrazek, and M. Campbell, Nucleic Acids Res. **24**, 4263 (1996).
[13] H. O. Smith *et al.*, Res. Microbiol. **150**, 603 (1999).
[14] S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, New York, 1970).
[15] L. Skrabanek and K. H. Wolfe, Curr. Opin. Genet. Dev. **8**, 694 (1998).
[16] A. L. Hughes *et al.*, Genome Res. **11**, 771 (2001).
[17] S. Otto and P. Yong, Adn. Genet. **46**, 451 (2001).
[18] E. S. Lander *et al.*, Nature (London) **409**, 860 (2001).
[19] J. C. Venter *et al.*, Science **291**, 1304 (2001).
[20] The Arabidopsis Initiative, Nature (London) **408**, 796 (2000).
[21] F. Kunst *et al.*, Nature (London) **390**, 249 (1997).
[22] W. H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
[23] N. Eldredge and S. J. Gould, in *Models in Paleobiology*, edited by T. J. M. Schopf (Freedman, San Francisco, 1972); S. J. Gould and N. Eldredge, Paleobiology **3**, 115 (1977).
[24] P. Bak, C. Tang, and K. Wiesenfeld, Phys. Rev. Lett. **59**, 381 (1987).
[25] J. Maynard Smith, *Evolution Genetics* (Oxford University Press, Qxford, 1998).
[26] C. R. Woese, Proc. Natl. Acad. Sci. U.S.A. **95**, 6854 (1997).