

Genomes: At the edge of chaos with maximum information capacity*

Sing-Guan Kong[†], Hong-Da Chen[‡], Andrew Torda[§] and H. C. Lee^{¶,¶}

[†]*Department of Mechatronics and Biomedical Engineering,
Universiti Tunku Abdul Rahman, 43000 Kajang, Selangor, Malaysia*

[‡]*Department of Physics, Chung Yuan Christian University,
Zhongli, Taoyuan City 32023, R. O. C. Taiwan*

[§]*Center for Bioinformatics, University of Hamburg, Hamburg, Germany*

[¶]*Department of Biomedical Sciences and Engineering, National Central University,
Zhongli, Taoyuan City 32001, R. O. C. Taiwan*

^{¶¶}*hclee12345@gmail.com*

Received 29 September 2015

Accepted 19 October 2015

Published 9 December 2015

We propose an order index, ϕ , which quantifies the notion of “life at the edge of chaos” when applied to genome sequences. It maps genomes to a number from 0 (random and of infinite length) to 1 (fully ordered) and applies regardless of sequence length and base composition. The 786 complete genomic sequences in GenBank were found to have ϕ values in a very narrow range, 0.037 ± 0.027 . We show this implies that genomes are halfway towards being completely random, namely, at the edge of chaos. We argue that this narrow range represents the neighborhood of a fixed-point in the space of sequences, and genomes are driven there by the dynamics of a robust, predominantly neutral evolution process.

Keywords: Edge of chaos; randomness; order; order index; equivalent length; equivalent mutation rate; complete genome sequence; dynamic fixed point; genome evolution; information.

PACS numbers: 87.14.Gg, 87.15.Cc, 02.50.-r, 05.45.-a, 89.70.+c, 87.23.Kg

Edge of chaos originally refers to the state of a computational system, such as cellular automata, when it is close to a transition to chaos, and gains the ability for complex information processing.¹⁻³ The notion has since been used to describe biological states, and life in general, on the assumption that life necessarily involves

*This article will also appear in “80th Birthday of Professor Hao Bailin”, edited by: Phua Kok Khoo and Ge Molin (World Scientific, 2016).

complex computation.⁴ In model systems such as cellular automata, there are well-defined procedures for recognizing the change in computational capability during the transition from nonchaotic to chaotic states.^{1,3} However, these have not been adapted to the wider biological context, even for the simplest of organisms. But if we represent a living organism by its genome, view evolution as a dynamical process that drives genomes in the space of sequences, and consider chaos as a state of genome complexity, then we have a framework within which the meaning of “life occurs at the edge of chaos” may be investigated.

When genomes are seen as texts of the four chemical letters or bases, A (adenine), C (cytosine), G (guanine) and T (thymine), they have several key properties reflecting their complexity, including long-range correlations and scale invariance^{5–9} self-similarity^{10–13} and distinctive Shannon redundancy.^{14–16} However, these properties do not give a measure of the proximity of a genome to chaos or randomness. Before the edge of chaos notion can be explored, one needs to have a quantity that measures the randomness of genomes as texts.

Here, we analyze genomes in terms of the frequency of occurrence of k -letter words, called k -mers, where k is a small integer.¹⁷ For a given k , the 4^k types of k -mers are partitioned into $k + 1$ “ m -sets”, $m = 0 - k$. An m -set is composed of all the k -mers containing m and only m A or T’s. There are $\tau_m = 2^k \binom{k}{m}$ types of k -mers in an m -set. The reason for partitioning the k -mers according to AT-content for statistical purposes is that although the A:T and C:G ratios are invariably close to 1,^{18–20} the AT to GC ratio may differ significantly. This partition is needed for preventing biased base composition from masking crucial statistical information in genomes.^{8,9,16} For $k \geq 2$, the k th order index for a sequence of length L (in bases) is

$$\phi \equiv \frac{1}{(2 - 2(p^k + q^k))} \sum_m \frac{1}{L} |L_m - L_m^{\{\infty\}}|, \quad (1)$$

where $0 < p < 1$ is the fractional AT-content in the sequence; $q = 1 - p$; L_m is the total number of k -mers in the m -set; and $L_m^{\{\infty\}}$ is the expected value for L_m in a p -valued random sequence of infinite length: $L_m^{\{\infty\}} = L 2^{-k} \tau_m p^m q^{k-m}$. The definition of ϕ is based on the observation that distribution-averages are useful indicators of the randomness of a sequence. The denominator on the right-hand side of Eq. (1) is a normalization factor which ensures $\phi \approx 1$ for an ordered sequence (in which all AT’s are on, say, the 5’ end and all CG’s are on the 3’ end). The singularities at $p = 0$ and 1 are not a practical problem since no genome has such extreme base composition.

From the central limit theory we expect, for random sequences, $|L_m - L_m^{\{\infty\}}|$ to scale as $L_m^{-1/2}$. We therefore expect ϕ to be proportional to $L^{-1/2}$ on average. The log-log plots in Fig. 1(a) and 1(b) show ϕ as a function of sequence length for different k ’s and p ’s. Each datum is averaged over 500 random sequences. It is seen that ϕ scales very well as $L^{-1/2}$ (with sizable fluctuations), and is only weakly dependent on k and p . These results are summarized for all k and p by an empirical

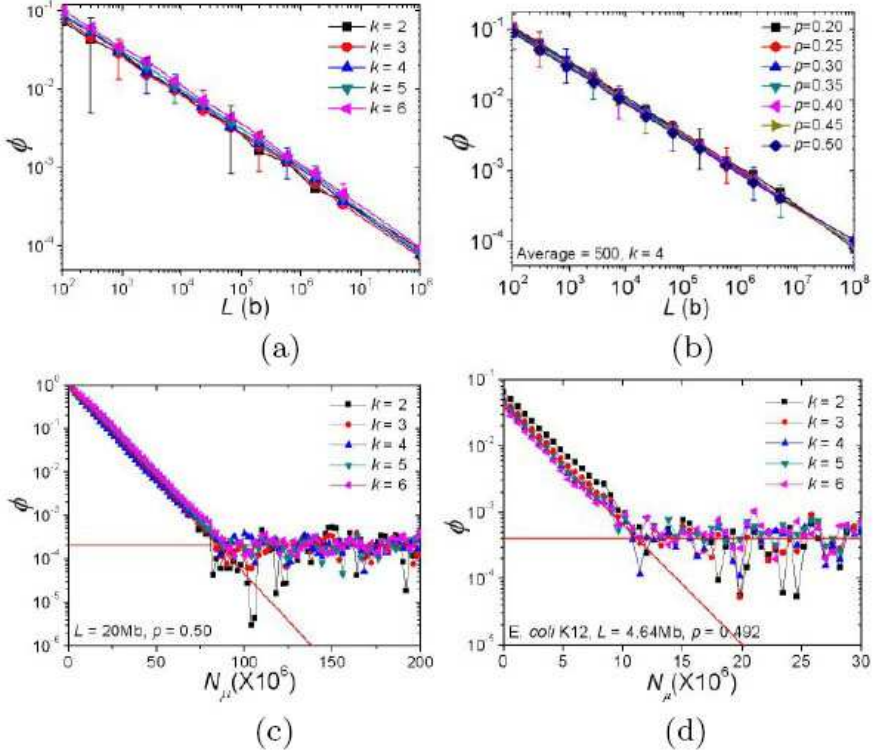


Fig. 1. (Color online) (a) Log-log plot of order index, ϕ , versus length of random sequence for $p = 0.5$ and $k = 2 - 6$. (b) Same as (a); for $k = 4$ and $p = 0.20 - 0.50$. (c) Semi-log plot of ϕ versus N_μ , number of random point mutations, for an initially ordered 20 Mb, $p = 0.5$ sequence. The intersection of the red lines is the critical point where sequence becomes random. (d) Same as (c); initial sequence is genome of *E. coli*.

relation:

$$\phi^{\{\text{ran}\}} = c_\phi L^{-\gamma_\phi} \quad (2)$$

with $\gamma_\phi = 0.50 \pm 0.01$ and $c_\phi = 1.0 \pm 0.2$, or, to a good approximation, $\phi^{\{\text{ran}\}} \approx L^{-1/2}$. This leads to the convenient concept of an *equivalent length* for a ϕ -value sequence, $L_{\text{eq}}(\phi) \equiv \phi^{-2}$, the nominal length of a random sequence whose order index is ϕ .

Random point mutation acting on a nonrandom sequence decreases its order, and hence its ϕ . Figure 1(c) shows that the ϕ of a $p = 0.5$, 20 Mb, ordered sequence decreases exponentially with the number of mutations N_μ , until N_μ reaches a critical number $N_{\mu c}$. The critical value reflects the fact that a random sequence does not become more random with further mutations. If one thinks of random point mutation as a dynamical action taking a sequence from one point to another in the sequence space projected onto ϕ , then a randomized sequence is a fixed-point of

the action. Our numerous simulations yield,

$$\phi = \begin{cases} \exp(-2N_\mu/L), & N_\mu \lesssim N_{\mu c}, \\ \phi_c \approx L^{-1/2}, & N_\mu > N_{\mu c}, \end{cases} \quad (3)$$

where the $N_{\mu c} \approx (1/4)L \ln L$, and the critical mutation rate is $\mu_c \equiv N_{\mu c}/L \approx (1/4) \ln L$. In the case of Fig. 1(c), the coordinates of the simulation ($k = 4$) critical point are $(\phi_c, N_{\mu c}) = (2.2 \times 10^{-4}, 8.5 \times 10^7)$, as compared to the “theoretical” values $(2.2 \times 10^{-4}, 8.4 \times 10^7)$. For typical sequences of genomic length ($L \sim 10^{1 \pm 1}$ Mb), $\mu_c = 4.0 \pm 0.6$ mutations per base (b^{-1}). We use Eq. (3) to assign to a ϕ -valued sequence an *equivalent mutation rate*, $\mu_{\text{eq}}(\phi) \equiv \ln \phi^{-1/2}$, the nominal number of random point mutations per base required to bring the index of an ordered sequence to ϕ .

Equation (3) can be adapted for application to sequences not initially ordered. For example, the equivalent mutation rate for the 4.6 Mb genome of *E. coli* ($\phi = 0.049$) is $1.5 b^{-1}$. Since for a 4.6 Mb sequence $\mu_c = 3.8 b^{-1}$, one expects an additional $2.3 \times 4.6 \times 10^6 = 1.1 \times 10^7$ mutations are needed to randomize it. In the simulation shown in Fig. 1(d), the actual number needed is found to be $(1.1 \pm 0.1) \times 10^7$.

We computed ϕ for 384 complete prokaryotic genomes (28 archaeobacteria and 356 eubacteria) and 402 complete chromosomes from 28 eukaryotes of lengths ranging from 200 kb to 230 Mb. The rice genome was downloaded from the Rice Annotation Project Database²¹ and all other sequences from the National Center for Biotechnology Information genome database,²² during the period 26 Feb–27 Nov, 2006. The 28 eukaryotes include the unicellular *P. falciparum*, 11 fungi, 2 plants, 5 insects and 9 vertebrates.

Results summarized in Fig. 2 (details are available upon request) indicate that genomic ϕ systematically vary neither with sequence length (a) and (b), nor with base composition (c). Instead they have a nearly universal value — the average over all sequences is $\phi_g \equiv 0.037 \pm 0.027$ (this defines the symbol ϕ_g). We have verified that, as a general rule, within a genome the variation in segmental ϕ decreases with segmental length and the average ϕ reaches its whole-genome value when the size of the segment exceeds 50 kb. In Fig. 2(a), the spread in ϕ of the genomic data shows a tendency to decrease with sequence length. Part of this effect may be purely statistical: smaller sample sizes (i.e., sequence lengths) tend to have larger statistical fluctuations. Part of it may also be because sequences longer than 10 Mb are all from chromosomes of multicellular eukaryotes that are phylogenetically close. In any case, Fig. 2(a) clearly puts the genomes in a category apart from random sequences.

From each complete sequence, we extracted the coding and noncoding parts, then concatenated the parts into two separate sequences and computed their order indexes, ϕ_{cd} and ϕ_{ncd} , respectively. A summary of the ratio $\phi_{\text{cd}}/\phi_{\text{ncd}}$ for sets of genomes grouped by length is given in Fig. 2(d). For prokaryotes the ratio ranges (10th to 90th percentile) from 0.15 to 3 with a median of about 0.5. For the eukaryotic chromosomes longer than 10 Mb the ratios do not significantly deviate from

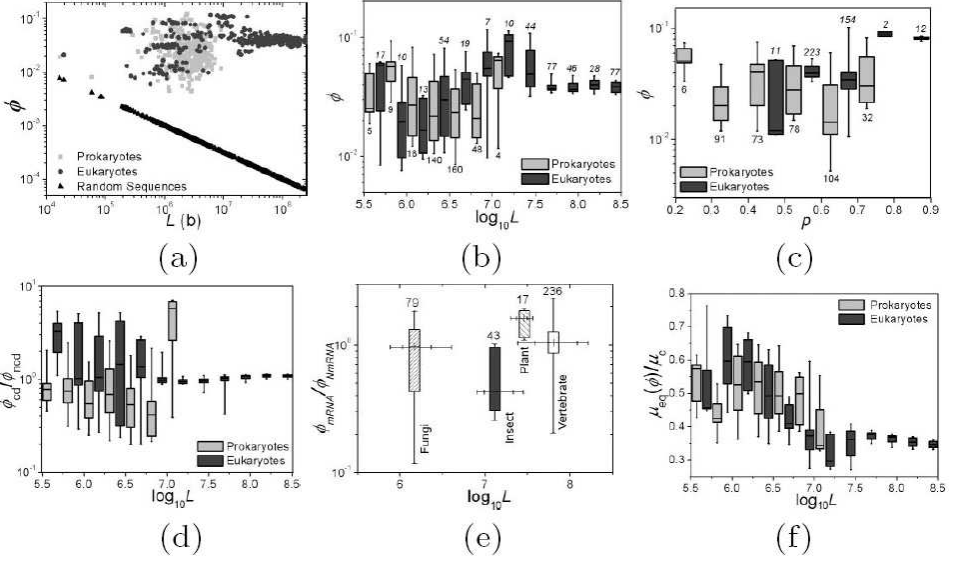


Fig. 2. (a) Order index, ϕ , versus sequence length, L , for 384 prokaryotic genomes (gray ■'s), 402 eukaryotic chromosomes (black •'s) and random sequences (line composed of ▽'s). In (b)–(f): Box (gray for prokaryotes; black for eukaryotes) height is given by 25% to 50% values and the range represents 10%–90% values; numbers above boxes are numbers of sequences in group; all ϕ 's are averaged over $k = 2$ to 6. (b) ϕ versus $\log L$. (c) ϕ versus fractional AT-content, p . (d) Ratio of ϕ_{cd} (for coding parts) to ϕ_{ncd} (noncoding) versus $\log L$. (e) Ratio of ϕ_{mRNA} (mRNA segments) to ϕ_{nmRNA} (non-mRNA), averaged over classes of eukaryotes. (f) Ratio of equivalent mutation rate, μ_{eq} , to critical mutation rate, μ_c , versus $\log L$.

unity. For the eukaryotic chromosomes shorter than 10 Mb the ratios average to about 2 but show greater variation.

The coding parts of eukaryotic genomes are further partitioned into mRNA and non-mRNA parts, and their ϕ 's computed separately. Averaged over sets of organisms, ϕ_{mRNA}/ϕ_{nmRNA} is of the order of 1, with the ratio being ~ 0.5 for insects and ~ 2 for plants [Fig. 2(e)]. For the latter, the ratio is ~ 1 for the five chromosomes of mustard and ~ 2 for the 12 chromosomes of rice. In summary, the differences in ϕ between coding and non-coding parts, and between mRNA and non-mRNA parts are much smaller than the difference between genomes and random sequences.

The ratio $\mu_{eq}(\phi)/\mu_c$ is an indication of how close a sequence is to being random. Figure 2(f) shows that the shorter ($L \gtrsim 10$ Mb) sequences are roughly half-way and the longer sequences, one-third of the way, towards becoming random. The systematic but weak length-dependence of the ratio is explained by the fact that the genomic ϕ , hence $\mu_{eq}(\phi)$, is approximately constant, whereas μ_c is proportional to $\ln L$. The overall average of the ratio is 0.45 ± 0.11 .

We summarize our results by considering the function

$$I(z) = -z \ln z - (1 - z) \ln(1 - z), \quad (4)$$

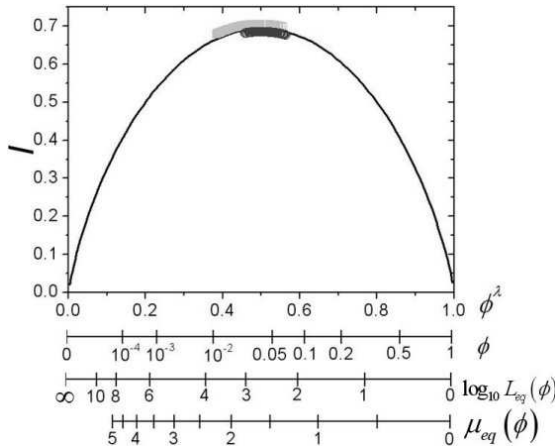


Fig. 3. The function $I(z)$ [Eq. (4)] plotted as a function of: $z = \phi^\lambda$ ($\lambda = 0.21$); ϕ ; $\log_{10} L_{\text{eq}}(\phi)$ [Eq. (2)]; $\mu_{\text{eq}}(\phi)$ [in units of b^{-1} ; Eq. (3)]. Data from prokaryotic (gray) and eukaryotic (black) genomes occur near the peak of I and have $\phi \sim \phi_g$, $L_{\text{eq}} \sim 0.25 - 10$ kb and $\mu_{\text{eq}} \sim 1.8 \pm 0.5 b^{-1}$.

where $z = \phi^\lambda$ and $\lambda = 0.21$. The value of the exponent λ is determined by requiring that $z = 0.5$ at $\phi = \phi_g$. $I(z)$ is the simplest function that maps the range $(0, 1)$ to a positive real value, has zeros at (and only at) $z = 0$ and 1 , has a maximum at $z = 0.5$ and is symmetric with respect to the point $z = 0.5$. In Fig. 3, the parabola-like curve shows $I(z)$ plotted against z . In addition, three other sets of abscissas are given: ϕ ; $\log_{10} L_{\text{eq}}(\phi)$, where $L_{\text{eq}}(\phi)$ is the equivalent length [Eq. (2)] and $\mu_{\text{eq}}(\phi)$, the equivalent mutation rate. A scale linear in z , relative to one in ϕ , is a better representation of the space of possible sequence lengths. It is seen in Fig. 3 that genomes are concentrated near the peak of the I -curve and equally far removed from both random ($z \sim 0$) and ordered ($z \sim 1$) sequences.

The genomic equivalent lengths, occupying a small neighborhood around at $L_{\text{eq}}(\phi_g) = 730 b$, are far shorter than the actual lengths of complete sequences. Replication is the basic mechanism for producing long sequences with short equivalent lengths: a long sequence of length L composed of multiple replications of a random sequence l bases long will have $L_{\text{eq}} \sim l$, independent of L . Similarly, if genome growth is dominated by random segmental duplication,²³⁻²⁵ then the genomic L_{eq} will be much shorter genome length.¹⁶

The genomic equivalent mutation rates span a small range around $1.8 b^{-1}$, or about 45% of the critical mutation rate of approximately $4 b^{-1}$ that would randomize the genomes. Thus, for example, a typical worm (*C. elegans*) chromosome, with an average length of 17 Mb and an equivalent mutation rate of $1.8 b^{-1}$, is as random as an initially ordered 17 Mb sequence after having undergone 31 million random mutations — as compared to the 68 million mutations which would randomize the sequence. In this sense genomes are quasi-random or “at the edge of chaos”. For a linear text, quasi-randomness satisfies two crucial necessary conditions for high

information content: high efficiency and large variation in word usage. A random sequence has maximum word-usage efficiency but minimum word-usage variation, because all its k -mers in an m -set have occurrence frequencies very close to the theoretical mean frequency — $\bar{f}_m^{\{\infty\}} = L_m^{\{\infty\}}/\tau_m$ — of the set. In a quasi-random sequence a compromise between high efficiency and large variation in word usage is obtained by suitably relaxing the equal-frequency condition,¹⁶ thus allowing a genome at the edge of chaos to have close to maximum information *capacity*.

The high concentration of genomic ϕ s near ϕ_g may be interpreted as the signature of a certain robust characteristics in the genomic evolution processes. The near equality of ϕ 's for coding and noncoding regions within a genome suggests that the underlying evolution processes are not dominated by codon selection, but are likely predominantly selectively neutral.^{26,27} We therefore propose the following conjecture: Just as randomness is a fixed-point of the action of random point mutations, the state of genomes defined by $\phi \sim \phi_g$ is a fixed-point of the action of a robust, predominantly neutral evolution process. The observed shortness of $L_{eq}(\phi_g)$ suggests that the neutral process is dominated by (nondeleterious) random segmental duplications.^{16,23–25,28,29} We consider random segmental duplication to be an infrastructure-building process because it does not necessarily produce information directly. Instead, it causes genomic ϕ to be close to ϕ_g , giving genomes maximum information capacity. Since this enhances genomic fitness indirectly, the neutral process may in itself be a product of natural selection. The near randomness of the neutral process guarantees the fixed-point associated with ϕ_g to have a very large configuration space, hence relatively low free energy, thus rendering ϕ_g -valued states widely accessible. In contrast, nonneutral, information-gathering processes dominated by selection (narrowly construed) are predominantly point mutations: they are poor mechanisms for inducing genomic states of maximum capacity, and do not lead to widely accessible states. Taken together these suggest that the evolution of the genome may have been driven by a two-stage process: one neutral, robust, infrastructure-building and universal, and the other selective, fine-tuning, information-gathering and diverse. An example of such a two-step process is found in the paradigm of accidental gene duplication followed by mutation driven subfunctionalization.^{30,31}

Acknowledgment

This work is supported in part by Grant No. MOST 104-2314-B-033-001 from the Ministry of Science and Technology (R.O.C).

References

1. C. G. Langton, *Physica D* **42**, 12 (1990).
2. J. P. Crutchfield and K. Young, Computation at the onset of chaos, in *Complexity, Entropy and the Physics of Information*, ed. W. H. Zurek (Addison-Wesley, Redwood City, CA, 1990), pp. 223–269.

3. M. Mitchell, P. T. Hraber and J. P. Crutchfield. *Complex Syst.* **7**, 89 (1993).
4. S. A. Kauffman, *The Origin of Order* (Oxford University Press, London, 1993).
5. W. Li and K. Kaneko, *Europhys. Lett.* **17**, 655 (1992).
6. C. K. Peng, et al. *Nature* **356**, 168 (1992); *Phys. Rev. E* **47**, 3730 (1993).
7. R. F. Voss, *Phys. Rev. Lett.* **68**, 3805 (1992).
8. N. E. Israeloff, et al., *Phys. Rev. Lett.* **76**, 1976 (1996); S. Bonhoeffer, et al., *ibid.* 1977; R. F. Voss. *ibid.* 1978; N. Mantegna, et al., *ibid.* 1979.
9. C. A. Chatzidimitriou-Dreismann, et al., *Nucl. Acid. Res.* **24**, 1676 (1996).
10. K. W. Church and J. I. Helfman, *J. Comput. Graph. Stat.* **2**, 153 (1993).
11. X. Lu et al., *Phys. Rev. E* **58**, 3578 (1998).
12. N. Nagai, *Jpn. J. Physiol.* **51**, 159 (2001).
13. T. Y. Chen, L. C. Hsieh and H. C. Lee, *Comput. Phys. Commun.* **169**, 218 (2005).
14. R. N. Mantegna, et al., *Phys. Rev. Lett.* **73**, 3169 (1994); *Phys. Rev. E* **52**, 2939 (1995).
15. H. E. Stanley, et al., *Physica A* **273**, 1 (1999).
16. H. D. Chen, et al., *Phys. Rev. Lett.* **94**, 178103 (2005).
17. B. L. Hao, H. C. Lee and S. Y. Zhang, *Chaos Solit. Fract.* **11**, 825 (2000).
18. V. V. Prabhu, *Nucl. Acids Res.* **21**, 2797 (1993).
19. J. Mrazek and S. Karlin, *Proc. Nat. Acad. Sci. (USA)* **95**, 3720 (1998).
20. S. J. Bell and D. R. Forsdyke, *J. Theor. Biol.* **197**, 51 (1999).
21. Rice Annotation Project Database; <http://rapdb.lab.nig.ac.jp/> (2006).
22. National Center for Biotechnology Information genome database; <http://www.ncbi.nlm.nih.gov/> (2006).
23. M. Lynch, *Science* **297**, 945 (2002).
24. J. A. Bailey, et al., *Science* **297**, 1003 (2002).
25. L. Zhang, et al., *Mol. Biol. Evol.* **22**, 135 (2005).
26. M. Kimura, *J. Mol. Evol.* **16**, 111 (1980).
27. Y. X. Fu and W. H. Li, *Genetics* **133**, 693 (1993).
28. L. S. Hsieh et al., *Phys. Rev. Lett.* **90**, 018101 (2003).
29. P. W. Messer, P. F. Arndt and M. Laessig, *Phys. Rev. Lett.* **94**, 138103 (2005).
30. M. Lynch and J. S. Conery, *Science* **290**, 1151 (2000).
31. J. Zhang, *Trends Ecol. Evol.* **18**, 292 (2003).