Genomes: at the edge of chaos with maximum information capacity

Sing-Guan Kong1,2, Hong-Da Chen1,2, Wen-Lang Fan1,2, Jan Wigger4, Andrew Torda4, and H.C. Lee2,3,5

Department of Physics, 2Graduate Institute of Biophysics, and 3Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan 32001, ROC
4Center for Bioinformatics, University of Hamburg, Hamburg, Germany
5National Center for Theoretical Sciences, Hsinchu, Taiwan 30043, ROC

(Dated: February 1, 2008)

We propose an order index, \( \phi \), 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. The 786 complete genomic sequences in GenBank were found to have \( \phi \) 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.

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

The 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, 2, 3]. The notion has since been used to describe biological states, and life in general, on the assumption that life necessarily involves self-similarity [10, 11, 12, 13], and distinctive Shannon complexity, including long-range correlations and scale redundancies [14, 15, 16]. However, these have not been adapted to the wider biological context, even for the simplest of organisms. But the transition from non-chaotic to chaotic states [1, 3] has been used to describe biological states, and life in general, on the assumption that life necessarily involves self-similarity [10, 11, 12, 13], and distinctive Shannon complexity, including long-range correlations and scale redundancies [14, 15, 16]. However, these have not been adapted to the wider biological context, even for the simplest of organisms. But the transition from non-chaotic to chaotic states [1, 3]

\[ L_m \approx L_m^{\infty} \approx L^{1/2} \]

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, the AT to GC ratio may differ significantly. This partition is needed for preventing biased base composition from masking crucial statistical information in genomes [2, 3]. For \( k \geq 2 \), the \( k \)th order index for a sequence of length \( L \) (in bases) is

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

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 \( \phi \) 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 \( \phi = 1 \) 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 \( \phi \) to be proportional to \( L^{-1/2} \) on average. The log-log plots in Fig. (a) and (b) show \( \phi \) 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 \( \phi \) scales very well as \( L^{-1/2} \) (with sizable fluctuations), and is only weakly dependent on \( k \) and \( p \). These results can be summarized for all \( k \) and \( p \) by an empirical 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 \( \phi \)-value sequence.
A nearly universal value—the average over all sequences is \( \phi_g = 0.037 \pm 0.027 \) (this defines the symbol \( \phi_g \)). We have verified that, as a general rule, within a genome the variation in segmental \( \phi \) decreases with segmental length and the average \( \phi \) reaches its whole-genome value when the size of the segment exceeds 50 kb. In Fig. 2(a) the spread...
in $\phi$ 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 (owing to imperfect annotation, the sum of the parts sometimes differ slightly from the whole), then concatenated the parts into two separate sequences and computed their order indexes, $\phi_{cd}$ and $\phi_{ncd}$, respectively. A summary of the ratio $\phi_{cd}/\phi_{ncd}$ 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. Notable exceptions are the three bacteria with exceptionally large genomes ($L \leq 10$ Mb) with ratios ranging from 5 to 7: *S. avermitilis*, *S. coelicolor*, and *Mycobacterium sp. MCS*.

For the eukaryotic chromosomes longer than 10 Mb the ratios do not significantly deviate from unity. Mustard, whose coding and noncoding parts have nearly equal lengths ($\sim 10$–12 Mb), is the only exception in this category with $\phi_{cd}/\phi_{ncd} \approx 1$ (these ratios are beyond the 90th percentile and therefore are not included in Fig. 2(d)). In this case $\phi_{cd} \approx 0.055$ is similar to other genomes while $\phi_{ncd} \approx 0.0075$ is about seven times less than the norm. Rice, the only other plant included in this study, with $\phi_{cd}/\phi_{ncd} \approx 0.35$ is unlike mustard but more like the other eukaryotes. 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 $\phi$’s computed separately. Averaged over sets of organisms, $\phi_{mRNA}/\phi_{nmRNA}$ is of the order of 1, with the ratio being $\sim 0.5$ for insects and $\sim 2$ for plants (Fig. 2(e)). For the latter, the ratio is $\sim 1$ for the five chromosomes of mustard and $\sim 2$ for the twelve chromosomes of rice. In summary, the differences in $\phi$ between coding and noncoding parts, and between mRNA and non-mRNA parts are much smaller than the difference between genomes and random sequences.

The ratio $\mu_{eq}(\phi)/\mu_\varepsilon$ is an indication of how close a sequence is to being random. Fig. 2(f) shows that the shorter ($L \geq 10$ 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 $\phi$, hence $\mu_{eq}(\phi)$, is approximately constant, whereas $\mu_\varepsilon$ is proportional to $\ln L$. The overall average of the ratio is $0.45 \pm 0.11$.

We summarize our results by considering the function

$$I(z) = -z \ln z - (1 - z) \ln(1 - z)$$

(4)

where $z = \phi^\lambda$ and $\lambda = 0.21$. The value of the exponent $\lambda$ is determined by requiring that $z = 0.5$ at $\phi = 0$. $I(z)$ 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: $\phi_1 \log_{10} L_{eq}(\phi)$, where $L_{eq}(\phi)$ is the equivalent length (Eq. (2)); and $\mu_{eq}(\phi)$, the equivalent mutation rate. A scale linear in $z$, relative to one in $\phi$, 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 and far removed from the random ($z \sim 0$) and ordered ($z \sim 1$) sequences.

The genomic equivalent lengths, occupying a small neighborhood around at $L_{eq}(\phi_g) = 730$ b, are far shorter than the actual lengths of complete sequences. Among the many possible mechanisms that may cause long sequences to have short equivalent lengths, by far the simplest is replication. This is because a long sequence of length $L$ composed of multiple replications of a random sequence $l$ bases long will have $L_{eq} \sim L$, independent of $L$. Similarly, if genome growth is dominated by random segmental duplication, then the genomic $L_{eq}$ will be much shorter genome length.

![FIG. 3: The function $I(z)$ (Eq. (4)) plotted as a function of: $z = \phi^\lambda$ ($\lambda = 0.21$); $\phi_1 \log_{10} L_{eq}(\phi)$ (Eq. (3)); $\mu_{eq}(\phi)$ (in units of bp$^{-1}$; Eq. (3)). Data from prokaryotic (gray) and eukaryotic (black) genomes occur near the peak of $I$ and have $\phi \sim \phi_g$, $L_{eq} \sim 25$–10 kb, and $\mu_{eq} \sim 1.8 \pm 0.5$ bp$^{-1}$.](image-url)
occurrence frequencies very close to the theoretical mean frequency of the set, $\bar{f}_{m}^{(\infty)} = \overline{L_{m}^{(\infty)}} / \tau_{m}$. However, this also implies minimum word-usage variation, which prevents a random sequence from being information-rich. 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 $\phi$’s near $\phi_{g}$ may be interpreted as the signature of a certain robust characteristic in the genomic evolution processes. The near equality of $\phi$’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 [16, 28]. 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 $\phi_{g}$ 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 (non-deleterious) random segmental duplications [23, 24, 25], occurring singly [16, 28] and in tandem [24]. We consider random segmental duplication to be an infrastructure-building process because it does not necessarily produce information directly. Instead, it causes genomic $\phi$ to be close to $\phi_{g}$, 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 $\phi_{g}$ to have a very large configuration space, hence relatively low free energy, thus rendering $\phi_{g}$-valued states widely accessible. In contrast, non-neutral, 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]. We may assume that during the long history of the genome’s growth and evolution, the twin-processes acted in a ratchet-like, complementary manner, driving the genome, in successive stages, to a state of maximum information capacity, and helping it to acquire, at each stage, near-maximum information content.

This work is supported in part by grant nos. 95-2311-B-008-001 and 95-29111-I-008-004 from the National Science Council (ROC).

[1] C. G. Langton. Physica D 42, 12-37 (1990).
[2] J. P. Crutchfield and K. Young. In W. H. Zurek, editor, Complexity, entropy, and the physics of information, 223-269 (Addison-Wesley, Redwood City, CA, 1990).
[3] M. Mitchell, P. T. Hraber, and J. P. Crutchfield. Complex Systems, 7, 89-130 (1993).
[4] S. A. Kauffman. (Oxford Univ. Press, London, 1993).
[5] S. P. Davies, et al. Biochem. J. 351, 95-105 (2000).
[6] W. Li and K. Kaneko. Europhys. Lett. 17, 655-660 (1992).
[7] C. K. Peng, et al. Nature 356, 168-170 (1992); Phys. Rev. E 47, 3730-3733 (1993).
[8] R. F. Voss. Phys. Rev. Lett. 68, 3805-3808 (1992).
[9] N. E. Israeloff, et al. Phys. Rev. Lett. 76, p1976 (1996); S. Bonhoeffer, et al. loc. cit. p1977; R. F. Voss. loc. cit. p1978; N. Mantegna, et al. loc. cit. p1979. Also: C. A. Chatzidimitriou-Dreismann, et al. NAR 24, 1676-1682 (1996).
[10] K. W. Church, J. I. Helfman, J. Comp. Graph. Stat. 2, 153-174 (1993).
[11] X. Lu, Z. Sun, H. Chen, Y. Li. Phys. Rev. E 58, 3578-3584 (1998).
[12] N. Nagai. Jpn J Physiol. 51, 159-68 (2001).
[13] T. Y. Chen, L. C. Hsieh and H. C. Lee. Comp. Phys. Comm. 169, 218-221 (2005).
[14] R.N. Mantegna, et al. Phys. Rev. Lett. 73, 3169-3172 (1994); Phys. Rev. E 52, 2939-2950 (1995).
[15] H. E. Stanley, et al. Physica A 273, 1-18 (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 Solitons Fract. 11, 825-836 (2000).
[18] V. V. Prabhu Nucl. Acids Res. 21, 2797-2800 (1993).
[19] J. Mrazek and S. Karlin. Proc. Nat. Acad. Sci. (USA) 95, 3720-3725 (1998).
[20] S. J. Bell and D. R. Forsdyke J. theor. Biol. 197, 51-61 (1999).
[21] Rice Annotation Project Database; http://rapdb.lab.nig.ac.jp/ (Nov. 2006).
[22] National Center for Biotechnology Information genome database; http://www.ncbi.nlm.nih.gov/ (Oct. 2006).
[23] M. Lynch. Science 297 945-947 (2002).
[24] J. A. Bailey, et al. Science 297, 1003-1007 (2002).
[25] L. Zhang, et al. Mol. Bio. Evol. 22 135-141 (2005).
[26] M. Kimura. J. Mol. Evol. 16 111-120 (1980).
[27] Y. X. Fu and W. H. Li. Genetics 133 693-709 (1993).
[28] L. S. Hsieh, L. F. Luo, F. M. Ji and H. C. Lee. 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 1151-1155 (2000).
[31] J. Zhang. Trends Eco. Evol. 18, 292-298 (2003).