Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Physics Communications 169 (2005) 218–221

Computer Physics Communications

www.elsevier.com/locate/cpc

Shannon information and self-similarity in whole genomes

Ta-Yuan Chen [a], Li-Ching Hsieh [b], Hoong-Chien Lee [a,*,1]

[a] *Department of Physics and Center for Complex Systems, National Central University, Chungli, Taiwan 320*
[b] *Institute of Information Science and Genomics Research Center, Academia Sinica, Taipei, Taiwan 115*

Available online 20 April 2005

Abstract

The Shannon information (SI) in distributions of occurrence frequency of short words in whole genomes is shown to exhibit universality. For given word length, the SI in genomes of all lengths is the same as that in random sequences of a universal lengths L_r. For the shorter words L_r is far shorter than the genome. For example, $L_r \sim 1000$ bases for three-letter words. We further show that whole genomes are highly self-similar in the sense that any segment of the genome down to a length of Λ_{sim}, about twice L_r, also shares the universal property. We devise a simple genome growth model in which genome-size sequences grown by maximally stochastic segmental duplication and random mutation possess the universal and self-similar properties of genomes.

PACS: 87.10.+e; 89.70.+c; 87.14.Gg; 87.23.Kg; 02.50.-r

Keywords: Complete genomes; Shannon information; Universality class; Self-similarity

1. Introduction

Recently is was shown that frequency distributions of overlapping k-mers, or k-spectra [1], of complete genomes exhibit universality [2,3]. For a given word length k, the Shannon information (SI) [4], or equivalently the relative spectral width (SW) [2] is the same for all complete genomes irrespective of their lengths and base compositions, or profiles. The length of the genomes range from 0.4 to 230 Mb (million bases) and their combined probability of (A+T) range from 0.25 to 0.75. In the papers cited, the SI and SW were calculated as *reduced Shannon information* (RSI) [2] and *reduced spectral width* (RSW) [3], which are quantities relative to those of a random sequence whose profile match that of the genome. These quantities are defined such that trivial dependence on base composition is to a large part eliminated and that RSI $\approx$ RSW ≈ 1 for any random sequence. The universality is inferred from the computed RSI [2] (and RSW [3]) being proportional to genome length (Fig. 1).

[*] Corresponding author.
E-mail address: hclee@phy.ncu.edu.tw (H.-C. Lee).

[1] Work supported by grants NSC (ROC) 92-2119-M-008-041 and 93-2311-B-008-006.

0010-4655/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cpc.2005.03.050

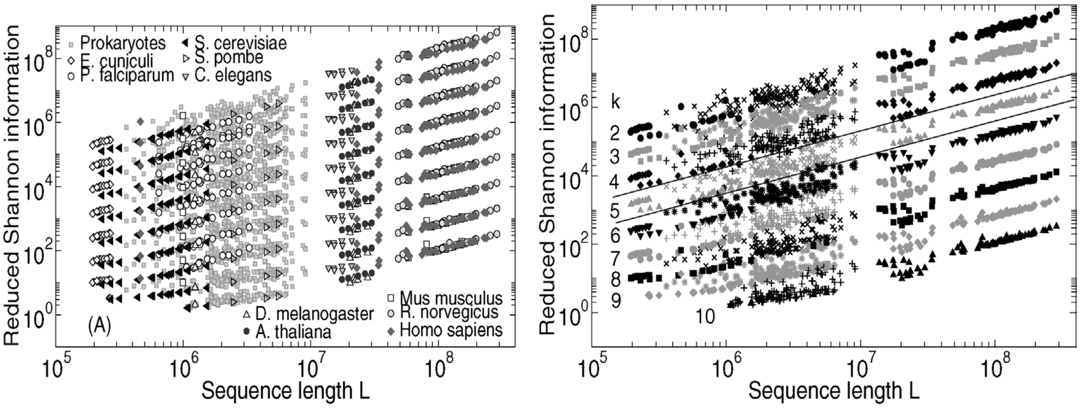

Fig. 1. RSI from 135 complete microbial genomes and 127 eukaryotes. Each symbol is the RSI value of one k-spectrum from one complete sequence. *Left panel*, RSI color-coded (gray scale) by organism; *right panel*, RSI color-coded by k. Data have been multiplied by factors of 2^{10-k} to delineate the k-bands for better viewing. Data for which $4k > L$, when RSI ≈ 1 for any sequence, have been discarded. Straight lines in the plots give RSI $\propto L$.

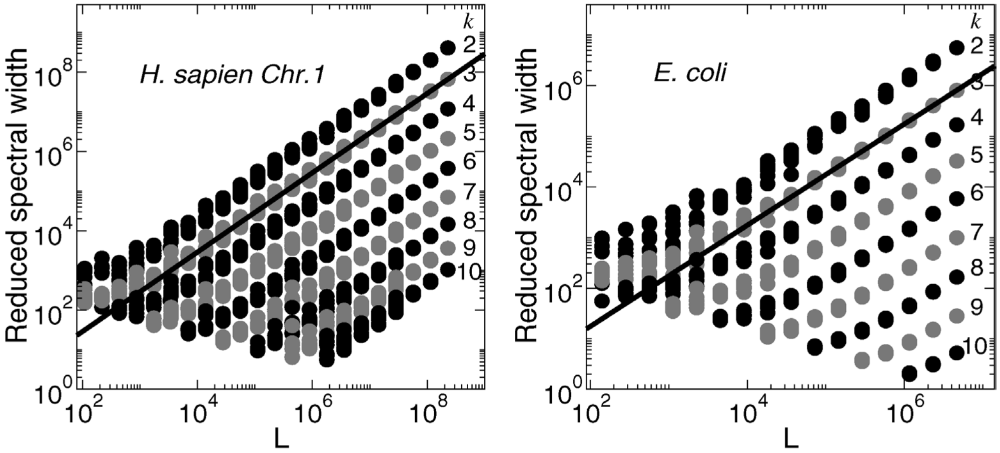

Fig. 2. *Left panel*: Average RSW of k-spectra, $k = 2$ to 10, of segments from the 246 Mb chromosome I of *H. sapiens*. Lengths of the segments are $1/2^n$ of full length, $n = 1$ to 21, and for each length eight segments are randomly selected. Data for which segment length is less than 4^k are not included. *Right panel*: Same for the and the 4.6 Mb genome of *E. coli*, except that lengths of the segments are $1/2^n$ of full length, $n = 1$ to 15.

The observed universality can be simply expressed in terms of equivalent random-sequence lengths (in bases), $L_r(k)$, one for each k, such that the universal genomic Shannon information is equal to that in a random sequence of length $L_r(k)$. The data in Fig. 1 summarized by the relation $L_r(k) = L_{r0}\exp(ak)$, $2 \leqslant k \leqslant 10$, where $a = 0.94 \pm 0.07$ and $L_{r0} = 42 \pm 17$. This relation defines a *universality class* of genomes. *Plasmodium falciparum* is the sole exception to the class [2].

2. Self-similarity

For the smaller k's, $L_r(k)$ is far shorter then the genomes. For instance, $L_r(2)$ is only about 300 and $L_r(6)$ is of the order of 10 000, as compared to the canonical length of 2 Mb for microbial genomes and the length of 100 Mb for some of the eukaryotic chromosomes. This being the case, one may ask whether individual segments from a genome in a (universality) class also belong to the class. Fig. 2 shows the result of testing this idea on two complete sequences, chromosome I of *Homo sapiens* (left panel) and the genome of *Escherichia coli* (right panel). In each case, for a given k, $k = 2$ to 10, eight segments of length one 2^nth, $n = 1, 2, \dots,$ of the full genome length, down to a length that is just less than $L_r(k)$, are randomly selected from the genome and their RSW computed. With $L_r(2) \approx 300$ b and the lengths of the *H. sapiens* chr. I and *E. coli* being 230 and 4.6 Mb, respectively, the maximum n used for the two genomes are 21 and 15, respectively. The computed RSW for the segments together with the RSW of the full sequence are plotted against segment length L. In the panels, data for the same k form clear *k-bands* suggesting a linear relation between RSW and L (lines in figure). If the data point from a segment sits on a parallel line that goes through the data point from the complete sequence, then the L_r of that segment is the same as that of the complete sequence and that segment is said to be similar to the complete sequence. Data shown in Fig. 2 suggest the *H. sapiens* chr. I and *E. coli* genomes to be self-similar down to a length not much greater than $L_r(k)$.

Complete genome sequences used in the present study were downloaded from the genome FTP site of the (USA) National Center for Biotechnology Information. The 163 complete microbial genomes (the prokaryotes) were downloaded on October 9, 2003 from ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ and the 127 chromosome sequences of ten complete eukaryotic (the eukaryotes) were downloaded on July 15, 2003 from ftp://ftp.ncbi.nih.gov/genomes/. For each complete sequence and each k we determine Λ_u, the length above which all segments are similar to the genome, and Λ_d, the length below which no segment is similar to the genome, by computing the RSW in sets of randomly selected segments of length Λ, where Λ is taken in steps decreasing from $10L_r(k)$ to $L_r(k)$. Λ is said to be a self-similar length if the RSW for each segment of length Λ is within a factor of two of Λ/L times the RSW for the genome.

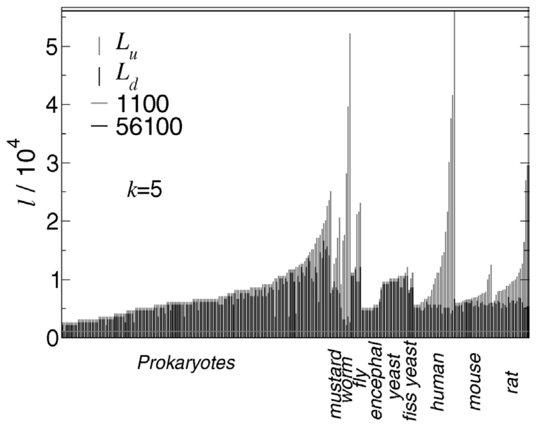

Fig. 3. Λ_u (gray bars) and Λ_d (black bars) for $k = 5$ for all complete sequences.

3. Results and discussion

Results for Λ_u and Λ_d for $k = 5$ are given in the Fig. 3. Results for Λ_u and Λ_d, $k = 2$ to 8 (not shown, see http://pooh.phy.ncu.edu.tw/cdy/genome/self_sim/main.html for details and more results) have the following general property: (i) Except for $k = 2$, in most cases $(\Lambda_u - \Lambda_d)/\Lambda_d \lesssim 1$ and decreases with increasing k. (ii) Even for $k = 2$ $(\Lambda_u - \Lambda_d)/\Lambda_d \lesssim 2$ for most cases. (iii) In a few cases Λ_u is up to an order of magnitude greater than Λ_d. (iv) For the prokaryotes, the difference between Λ_u and Λ_d is small, both being close to the eukaryotic Λ_d. The averages and standard deviations of Λ_u and Λ_d and given in Table 1 and Fig. 4(A). Denoting by Λ_{sim} the average of the prokaryotic Λ_u and Λ_d and the eukaryotic Λ_d we have $\Lambda_{\mathrm{sim}}(k) \approx \Lambda_{s0} \exp(a_s k)$, $2 \leqslant k \leqslant 8$, where $a_s \approx 0.780$ and $\Lambda_{s0} = 166 \pm 120$. The eukaryotic Λ_u is noticeably greater than Λ_{sim}. This is a reflection of the higher textual inhomogeneity in the eukaryotic genomes. The results satisfy the inequality $4^k < L_r(k) < \Lambda_{\mathrm{sim}}(k)$. In most cases $L_r(k)$ is not much greater than 4^k and, except for $k = 2$, $\Lambda_{\mathrm{sim}}(k)$ is less than twice $L_r(k)$. In this sense genomes are highly uniform and are very close to being maximally self-similar.

In [2], a minimal universal model for genome growth [5] in which growth is dominated by maximally stochastic segmental duplications was used to explain the rise of the universality class. With three universal parameters the model generates sequences that belong to the genomic universality class [2]. Here we compute Λ_u and Λ_d from the model sequences for the twenty-four *H. sapiens* chromosomes. The results are compared with the genomic Λ_{sim} in Fig. 4(B). The agreement between data and model is excellent. The relatively small standard deviations on the model results reflect the small sample size (24 sequences versus

Table 1
Average standard deviation Λ_u and Λ_d for complete eukaryotic and prokaryotic

k	Eukaryotes		Prokaryotes		Model sequence (*H. sapiens*)	
	Λ_d	Λ_u	Λ_d	Λ_u	Λ_d	Λ_u
2	4.20 ± 2.18 E2	1.63 ± 1.26 E3	5.43 ± 3.69 E2	1.12 ± 1.12 E3	6.45 ± 1.24 E2	9.14 ± 3.15 E2
3	9.62 ± 4.43 E2	2.08 ± 1.65 E3	1.20 ± 0.84 E3	1.70 ± 1.50 E3	1.07 ± 0.15 E3	1.22 ± 0.13 E3
4	2.35 ± 0.92 E3	4.88 ± 3.89 E3	2.66 ± 1.55 E3	3.47 ± 2.61 E3	2.20 ± 0.28 E3	2.45 ± 0.25 E3
5	6.39 ± 2.03 E3	1.18 ± 0.91 E4	6.15 ± 3.17 E3	7.54 ± 4.39 E3	4.97 ± 0.41 E3	5.47 ± 0.41 E3
6	1.63 ± 0.48 E4	3.11 ± 2.36 E4	1.53 ± 0.77 E4	1.78 ± 0.92 E4	1.23 ± 0.11 E4	1.34 ± 0.11 E4
7	4.12 ± 1.24 E4	6.82 ± 5.71 E4	3.99 ± 1.96 E4	4.60 ± 2.28 E4	3.16 ± 0.30 E4	3.66 ± 0.30 E4
8	9.77 ± 3.23 E4	1.76 ± 1.62 E5	1.10 ± 0.45 E5	1.20 ± 0.55 E5	9.14 ± 0.76 E3	1.01 ± 0.07 E5

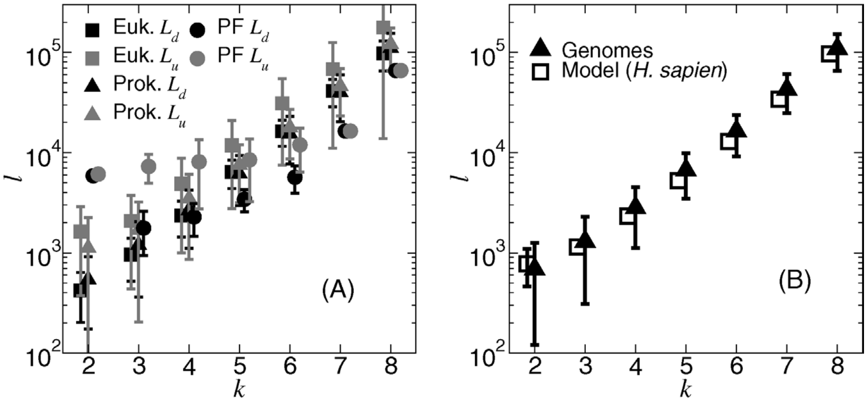

Fig. 4. (A) Average and standard deviation of L_u abd L_d (Λ_u and Λ_d in text) for complete prokaryotic eukaryotic genomes. (B) Genome data is the average of prokaryotic Λ_u and Λ_d and eukaryotic Λ_d, model data is average of Λ_u and Λ_d from 24 model sequences for the 24 human chromosomes.

276 sequences in the genome set) and the fact that textually sequences generated in the minimal model are significantly more homogeneous than genomes.

References

[1] B.L. Hao, H.C. Lee, S.Y. Zhang, Fractal related to long DNA sequences and complete genomes, Chaos, Solitons and Fractals 11 (2000) 825–836.

[2] C.H. Chang, et al., A universal signature in whole genomes, in: IEEE Proc. Computer Sys. Bioinformatics (CSB’04), 2004, pp. 20–30; J. Bioinf. Comp. Biology (2005), submitted for publication.

[3] T.Y. Chen, et al., Universal lengths in complete microbial genomes, Int. J. Mod. Phys. B 18 (2004) 2448–2454.

[4] C.E. Shannon, A mathematical theory of communication, Bell Sys. Techn. J. 27 (1948) 379–423; 623–656.

[5] L.S. Hsieh, et al., Minimal model for genome evolution and growth, Phys. Rev. Lett. 90 (2003) 018101–018104.