Genome Size, Self-Organization and DNA’s Dark Matter

Chromosomes exhibit several features indicating that its spatiotemporal dynamics is self-organized. It has been recently suggested that a negative correlation between genome size and mean chromosome number would also be a fingerprint of self-organization, related to how human language is organized at the level of words and syllables. However, the vast dominance of non-coding DNA in eukaryotic genomes should prevent an interpretation of genome/chromosome size based on functional trade-offs related to information storage and transmission. Moreover, the reported negative correlation is shown to be an inevitable consequence of the definitions of chromosome and genome length and it is thus unrelated to any type of special generative process. © 2010 Wiley Periodicals, Inc. Complexity 000: 000–000, 2010

Key Words: genome; chromosomes; junk DNA; self-organization; language

Few examples in biology are more illustrative of the mixed nature of evolutionary paths than genomes and chromosomes in high multicellular organisms [1, 2]. On the one hand, chromosomes are the physical carriers of genes, the basic units of genetic information. They are also an example of the sophisticated mechanisms of molecular packing exhibited by eukaryotic cells. They are nonrandomly organized in space and mounting evidence indicates the presence of self-organized traits pervading their large-scale organization [3–6]. On the other hand, genome and chromosome evolution is marked by widespread, random events occurring at multiple scales. The analysis of genome composition reveals that most genomes are composed by a vast majority of noncoding DNA. This DNA defines a dark matter of genome organization: it contains great amounts of retroviral or retrotansposon sequences, pseudogenes and millions of repeated sequences [2]. Additionally, chromosomes evolve in time through a number of mechanisms of recombination, duplication, and loss that are constantly reshaping their number and size. The multiscale nature of genomic changes has made difficult the reconstruction of a natural history of eukaryotic chromosomes until recently [1].

In a recent paper [7] it has been suggested that a statistical trend displayed by genome size (described as the number of chromosomes \( L_g \)) versus chromosome size (computed in terms of millions of base pairs, \( \langle L_c \rangle \)) in eukaryotic organisms might be a signal of self-organizing behavior. More interestingly, the authors suggest that this trend is related to human language. Specifically, they plot \( \langle L_c \rangle \) against \( L_g \) for all available species from three public databases. These include animals, plants, and fungi. The correlation is rather rough for Fungi and Insects.
FIGURE 1

A negative correlation between the average length of chromosomes \( L_c \) and the genome size measured as the number of chromosomes naturally decays following a hyperbolic law. Here the open circles correspond to available data for mammals, obtained from http://www.genomesize.com. The continuous line is a power law fit, i.e., \( L_c \sim L_g^{-\alpha} \) which gives \( \alpha = 1.04 \pm 0.06 \).

It is also highly variable for plants. In mammals (a much more coherent group to establish comparisons) the correlation seems rather clean, with an almost hyperbolic decay (Figure 1).

Using this (statistically significant) decay, the authors go a step further in making a comparison with another correlation reported from the analysis of human language: the so called Menzerath-Altmann law. This empirical law indicates that the length of words (measured in number of syllables) is negatively correlated with the length of syllables (measured in phonemes). Roughly speaking, the longer a word, the shorter the syllables composing it [8]. In a more general context, it could be stated that an increase of a linguistic construct results in a decrease of its constituents. The seemingly general negative correlation observed at the genome level supports, in their view, a general principle of self-organization shared with human language. They conclude in their article that “our results suggest that human language and genomes share similar principles of self-organization.”

Using metaphors from a given field to extract universal patterns for another is a common strategy in complexity sciences [9]. Sometimes, apparently weak analogies turn into powerful correlations. However, the power of the metaphor needs to be properly established, even at the qualitative level. Words in language can be related to genes in genomes: well defined units that capture meaningful information. Their internal organization is also constrained by phonological and other constraints. What about whole genomes? Genes are certainly information carriers and they play well-defined roles in the orchestration of cellular order. Such roles are the result of gene-gene interactions and other molecular processes. Some features of genes have been recognized as related to formal aspects of language [10] but extrapolations to a larger scale of genome organization should be taken cautiously.

The problem considered by Ferrer and Forns, particularly the evolution of chromosomes in terms of their size, has been previously addressed from theoretical models (see for example [15] and references therein). Such models have been successful in accounting for the distribution of chromosome lengths by considering their dynamics as the result of random events of splitting, duplication, translocation, and mutation. The good fit obtained by these theoretical approaches and the current knowledge on the fluid nature of chromosomal rearrangements through time rule against any special multiscale link between genome-level and chromosome-level patterns. In fact, there is a serious flaw in [7] related to their choice of the correlated quantities that makes their results artificial. They choose genome complexity in terms of the actual number of chromosomes (more correctly, chromosome pairs). However, in this way what we have
Decay of the average chromosome size with the actual number of chromosomes $N_c$ (used as a measure of genome size in [7]). Here we have used a statistical model where a string of $G$ sites, representing genome size, is broken into $N_c$ pieces, representing chromosomes. The average chromosome size is thus obtained from $\frac{G}{N_c}$. We have used $N^M = 10$, $N^m = 200$ as the minimal and maximal number of chromosomes (partitions of the string). Moreover, genome size $G$ is also varied in a range $G_0 < G < G_M$ where $G = 10^4$ and $G_0$ is variable. The difference $G_M - G_0$ gives the range of genome size variation used. We made 500 runs for each set of parameters, choosing $N_c$ and $G$ at random within their interval of definition using a uniform distribution. In (a) we represent the result for a narrow range ($G_0 = 9000$).

in fact is that $L_c = \frac{|G|}{L_g}$, where $|G|$ is the total genome size (in base pairs), i.e., if $L_c(k)$ is the length (in base pairs) of the $k$-th chromosome, we have:

$$|G| = \sum_{k=1}^{L_g} L_c(k)$$

consistently with the definition of mean value. In other words, they simply say that the average number of chromosomes will fall-off as the inverse of the number of chromosomes, i.e., in their own notation,

$$L_c \sim L_g^{-a}$$

with $a = 1$. This is confirmed by estimating the scaling exponent from the mammalian data set, where we obtain $a = 1.04 \pm 0.06$ (continuous line, Figure 1). A similar value is obtained when using the more noisy data set for plants, where we get $a = 1.07 \pm 0.09$. Actually, the dispersion of points shown by the plant group can be easily explained in terms of a higher range of genome sizes found in plants.

A toy model illustrates our point. Let us consider a string of $|G|$ sites (here we use a maximum of $G_M = 10^4$ in our examples) which represents the genome, to be split into $N_c$ pieces by randomly cutting the genome in $N_c - 1$ sites. The number of cuts is chosen also at random within the interval $N^m < N_c < N^M$. Here we use $N^m = 10$ and $N^M = 200$. The range of genome sizes is controlled by an additional parameter $G_0$, which gives the lower bound. Small values of $G_0$ allow a broader range of genomes to appear, whereas for $G_0 \to G_M$ genome sizes will be more similar. In Figure 2(a) we show an example of the later case. Here we allow little variation for genome size but wide variation in the number of cuts, and as a consequence different partitions give an average chromosome value that falls perfectly in the hyperbolic shape. When the range of genome sizes increases, so do the fluctuations in average chromosome size, as illustrated in Figure 2(b) for two cases. As we can clearly appreciate, the result is very similar to the one shown in Figure 1 in [7]. This consistency clearly rules against the claims linking genome size properties and self-organization or informational constraints.

ACKNOWLEDGMENTS

The author thanks Sergi Valverde, Bernat Corominas-Murtra, Carlos Rodriguez-Caso and Jordi Fortuny for useful discussions. This work has been supported by the James McDonnell Foundation and the Santa Fe Institute. The author also thank the European Center for Living Technology for its hospitality.
REFERENCES