Bioinformatics, Proteomics and Immaging Analysis

OMMEGA Publishers
Answering Intelligence
www.ommegaonline.org

Short Communication Open Access

Mathematical Laws of Genomes

Vincenzo Manca*

Department of Computer Science, and Biomedical Computational Center (CBMC), University of Verona

***Corresponding author:** Manca, V. Department of Computer Science, and BioMedical Computational Center (CBMC), University of Verona, Italy; E-mail: vincenzo.manca@univr.it

Citation: Manca, V. Mathematical Laws of Genomes (2017) Bioinfo Proteom Img Anal 3(1): 172- 173.

DOI: 10.15436/2381-0793.16.1211

Keywords: Genomes; Texts; Evolution; Laws; Biobit

Received Date: November 10, 2016
Accepted Date: November 23, 2016
Published Date: November 30, 2016

Genomes follow laws resembling celestial orbits. In a recent publication on Nature Scientific Reports[1] two computer scientists (V. Manca and V. Bonnici) discovered strong mathematical regularities in the structure of genomes. These regularities can be expressed in terms of simple laws based on ellipses, hyperbolas and parabolas. But the geometrical forms do not cope with spatial movements, rather they refer to genomic indexes defined by means of information theory[2].

The research started in 2009 at Center of Biomedical Computing of the University of Verona (Italy) directed by one of two authors, with the aim of finding rigorous proofs supporting the intuition that genomes are texts, with a complex internal organization, written during the evolutionary process of living organisms. Along a line of thought, called Infogenomics[3-7], it was argued that these texts contain the deepest secrets that rule the universal mechanisms of life, in all the various and complex phenomena of biological processes.

The usual comparison between genomes and computer operating systems is surely significant, because genomes like programs direct procedures activating and regulating the cell processes that realize biological functions. However, this comparison misses a crucial and peculiar aspect of genomes. In fact, these texts are written by themselves during evolution, starting from small initial texts to transform in more complex texts, according to an internal mechanisms that explore enormous genomic spaces, but at same time by maintaining an internal coherence that guarantees the possibility of transmitting their information along species. This means that these texts need to be open to their future and, at same time, adequate to their present and conservative of their past. The equilibrium among these different temporal perspectives is probably the ultimate key of their success in living and evolving by always acquiring new and richer biological functionalities.

The publication on Scientific Reports shed light on the complex equilibrium between order and disorder, calculus and chaos, or according to the terminology of the paper, entropic and anti-entropic components. These components are defined in terms of Information Theory, by using the classical concepts introduced by Shannon[8], the notion of (empirical) entropy and entropic divergence, but even random genomes and genomic dictionaries of k-mers (sequences of length k), for suitable values of k.

A dedicated software, published on the present journal[8], was developed and enriched for analyzing 70 genomes from primitive genomes to the most complex genomes of plants and animals. Empirical entropies were computed and informational indexes based on these entropies were estimated. All these measures confirmed the invariance of some constraints over numerical ratios that all these genomes satisfy. It is surprising that the chaotic component of genomes is at least three times bigger than the ordered component measuring the degree of their structural richness, and related to the functions that genomes direct in the organisms. Genomes keep alive the fuel supporting their future in a measure that is preponderant with respect to the part

supporting their acquired structure. However, some strict constraints have to be respected, as expressed by a number of laws satisfied in all considered genomes.

Data were extracted from public genomic databases. No wet experiment was performed, only computations of informational quantities and verification of the formulae expressing the identified genomic laws. This is a new trend and something that can open the way to a great number of possible applications.

Current research on genomic laws shows that they can be reduced to three. The first one is a statement introducing three main informational indexes and some basilar relationships by an ellipse on which other derived indexes are defined for each genome.

The second law improves and collapses in one inequality four inequalities given in[1]. This law is a sort of genome well formedness law. A long experimental investigation of this law has to be developed for a better understanding of its meaning. Some preliminary results were reported in[1] (see supplementary Information). In our opinion this law, or some derivations of it, could be applied, with suitable modifications, in evaluating divergences of pathological genomes from "normal" genomes, by opening the way to global analysis of genomes that could complement the current clinical genomic approaches based on snips and local genetic variants.

The third law expresses a measure of genomic complexity, called biobit, that increases along the usual biological complexity of the corresponding organisms. Without no commitment with any biological information this measure gives a value expressing an informational complexity that seems completely coherent with its evolutive position. However, a detailed discussion, which we cannot develop here, should prevent some wrong interpretations confounding genomic complexity with the usual localization of the organisms in classical evolutive classifications. In fact, measures of our indexes are based on sequencing results of organisms as they are now, by ignoring the original genomes of the ancestors of these organisms as they appeared when the species arose.

The biobit-complexity function is based on a suitable equilibrium between entropic and anti-entropic components of genomes. What is surprising is that this function is related to a classical mathematical distribution based on the Beta Euler's function (applied also in physics to unify elementary particles, see Veneziano's formula[9]).

Moreover, it is mathematically apparent that the evolutionary trends of evolution is anti-entropic. This apparent paradox, due to the physical nature of genomes, can be explained by distinguishing between individual and species genomes, but this is a too technical discussion that we do not develop further.

In conclusion, mathematics and computer science allow us to consider genomes within a theoretical framework that could be disclose some of the deep principles on which life developed the evolutionary and functional strategies of its realization within species and biological individuals. If this perspective is essentially correct, surely it is destined to influence strongly the future research of computational genomics, and more generally, of the whole life sciences. Of course, it is too early for predicting the possible scenarios of this assertion. However, we want to remark that the scientific risks of such an approach are surely justified by the benefits that it could obtain along this way. Moreover, biology is not new to contaminations of its classical apparatus, which were crucial to its development. DNA structure was discovered by means of a significant contamination with physics, and was a famous physicist, Erwin Schrodinger who firstly, on the basis of mathematical and physical arguments, correctly predicted some essential features of DNA structure. And even the three-nucleotide structure of genetic code was proposed, in terms of simple mathematical arguments, by George Gamow (the founder of Big Bang theory)[10]. By paraphrasing one of Hilbert's famous sentence (referred to physics), Life is too important and too complex for being studied only by biologists.

References

[1] Bonnici, V., Manca, V. Informational laws of genome structures. (2016) Scientific Reports 6: 28840.

[2] Shannon, C. E. A mathematical theory of communication. (1948) Bell SysTech J 27: 623-656.

[3] Manca, V. Infogenomics: Genomes as information sources. (2016) Elsevier, Morgan Kauffman chap. 21: 317--324

[4] Manca, V. Research Lines in Infogenomics. (2015) Bioinfo Proteom Img Anal 1(1) 1-4.

[5] Manca, V. Infobiotics: information in biotic systems. (2013) Springer.

[6] Castellini, A. , Franco, G., Manca, V. A dictionary based in formational genome analysis. (2012) BMC genomics 13: 485.

[7] Kong, S. G. et al. Quantitative measure of randomness and order for complete genomes. (2009) Phys Rev E Stat Nonlin Soft Matter Phys 79(6): 061911.

[8] Bonnici,V., Manca, V. Infogenomics tools: A computational suite for informational analysis of genomes. (2015) J.Bio info Proteomics Rev 1: 8-14.

[9] Veneziano G. Construction of a crossing-symmetry, Regge-behaved amplitude for linearly rising trajectories. (1968) Nuovo Cimento A 57: 190-7.

[10] Gamow, G., Ycas. M. Mr Trompkins inside himself. (1967) MrTrompkinns Series.