Correlation between three genome metrics and genome size in different phyla

Nuria de Frutos Andicoehcea[1], Wladimiro Diaz-Villanueva[1], Vicente Arnau[1], Andrés Moya[1,2,3]

1. Institute for Integrative Systems Biology(I2SysBio), Universitat de València(UV) and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain
2. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region(FISABIO), Valencia, Spain
3. Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública(CIBEResp), Madrid, Spain

Correo de contacto:nudefru@alumni.uv.es

INTRODUCTION

The tendency toward increasing complexity in biological evolution is a controversial issue. Having a complexity measure can help with its resolution. Biological complexity has been defined as the sum of its parts; as long as they interact with each other(1).
In a previous paper (2), three complexity metrics were applied in cyanobacteria(2), Sequence Compositional Complexity (SCC) (3), Genomic Signature (GS) (4), and Biobit (BB) (5). Its ability to measure complexity and its behavior along the tree of life is unknown.

The **genome size** measures an increment in parts of the sequence. Smaller genome sizes correspond to bacteria and archaea. The values increase until reaching a saturation point in multicellular eukaryotic genomes**.**

MATERIALS & METHODS

Selecction

Initially, 118 phylas were selected, 47% eukaryotes, 29% archaea, and 24% bacteria. Incomplete genomes were eliminated, and 28,336 genomes were obtained, 59% bacteria, 9% archaea, and 31% eukaryotes

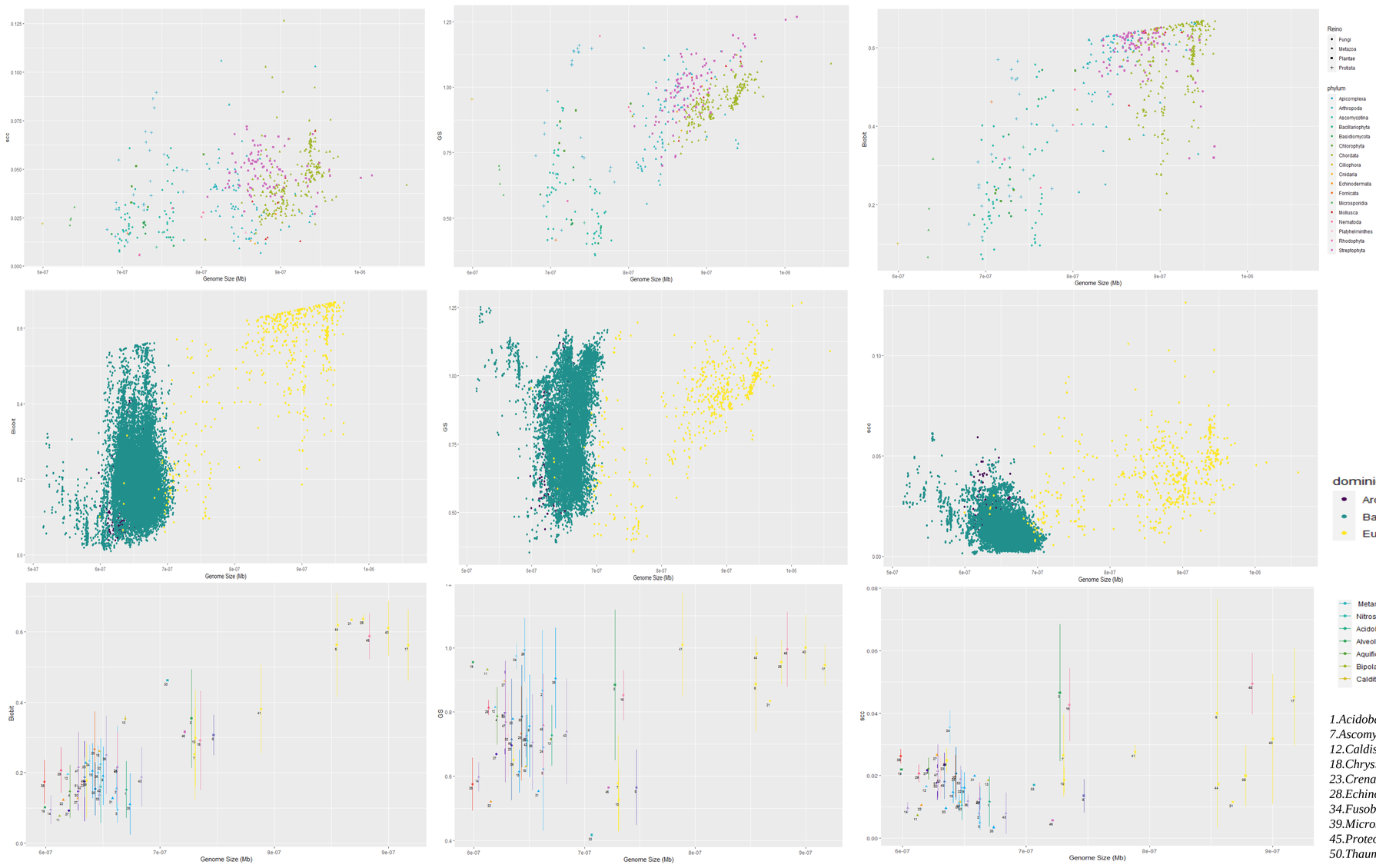
Biobit is a non-linear function that measures the balance between the entropic and anti-entropic parts of the genome.

The **Genomic Signature** is based on the Chaos Game Representation. It measures the relative frequency of K-mers, comparing it with the expected values of a random genome of the same length.

K=log4(S)-2 S is the size of the genome

Sequence Compositional Complexity measures the heterogeneity of the genome through non-overlapping compositional domains.

RESULTS



Analysis of the distribution obtains a log-normal distribution in SCC, GS, and BB. However, the p-value (p-value <0.05) obtained in the Anderson-Darling test rejects the normal distribution. In SCC, the distribution indicates no differentiation between the three domains. In the case of GS, the distribution of Archaea coincides with Bacteria. Meanwhile, BB has a difference between both groups. In GS and BB, eukaryotes have higher values.

1.Acidobacteria, 2.Actinobacteria, 3.Apicomplexa, 4.Aquificae, 5.Armatimonadetes, 6.Arthropoda, 7.Ascomycotina, 8.Bacillariophyta, 9.Bacteroidetes, 10.Basidiomycota, 11.Bipolaricaulota, 12.Caldiserica, 13. Calditrichaeota, 14.Chlamydiae, 15.Chlorobi, 16.Chlorophyta, 17.Chordata, 18.Chrysiogenetes, 19.Ciliophora, 20.Cloacimonetes, 21.Cnidaria, 22.Coprothermobacterota, 23.Crenarchaeaota, 24.Cyanobacteria, 25.Deferribacteres, 26.Deinococcus-Thermus, 27.Dictyoglomi, 28.Echinodermata, 29.Elusimicrobia, 30.Euryarchaeota, 31.Fibrobacteres, 32.Firmicutes, 33.Fornicata, 34.Fusobacteria, 35.Gemmatimonadetes, 36.Kiritimatiellaeota, 37.Korarchaeota, 38.Micrarchaeota, 39.Microsporidia, 40.Mollusca, 41.Nematoda, 42.Nitrospirae, 43.Planctomycetes, 44.Platyhelminthes, 45.Proteobacteria, 46.Rhodophyta, 47.Spirochaetes, 48.Streptophyta, 49.Synergistetes, 50.Thaumarchaeota, 51.Thermodesulfobacteria, 52.Thermotogae, 53.Verrucomicrobia/Clamydiae.

CONCLUSIONS

- **Biobit** seems to be a complexity metric.The phyles with closer ancestors are grouped. There is a distinction between bacteria and archaea, which share a range of values; and the eukaryotes, whose values present a positive correlation that corresponds to the evolutionary processes of appearance of the large phylogenetic groups. The highest values are Chordata and Streptophyta. The value of K depends on the size of the genome, and the metric has a limit of K = 32. The metric is affected by the size of the genome.There is an upper limit to the metric.
- **Genomic Signature** obtains an extensive intragroup dispersion in all the pylums. In eukaryotes, there is a positive correlation, whose highest values correspond to Streptophyta, in the kingdom Plantae, and the phyles of Metazoa are grouped.
- In **Sequence Compositional Complexity,** the highest values are the parasites. Deleting parts of its genome increment the differentiation of the compositional domains. There is a large dispersion in the values of Eukaryotes compared to prokaryotes. However, the means and the distribution show a minimal separation between the two groups.

REFERENCES

1.Moya, A. 2020 Progreso, Complejidad y Evolución. Éndoxa: Series Filosóficas. 46, pp. 427-440.
2.Moya, A. Olivera, J. L. Verdú, M. Delaye, L. Arnau, V. Bernaola-Galván, P. De la Fuente, R. Diaz, W. Gómez-Martín, et al. 2020 Driven progressive evolution if genoma sequence complexity in Cyanobacteria. *Nature: Scientific reports.*
3.Roman-Roldan, R. Bernaola-Galvan, P. Oliver, J. 1998 Sequence compositional complexity of DNA through an entropic segmentation method. *Physical review letters*. **80**(6).
4.Almeida, J. Carriço, J. Maretzek, A. Noble, P. Fletcher, M. 2001 Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics.* **17**(5), pp. 429-437.
5.Bonicci, V. Manca, V. 2016 Informational laws of genoma structures. *Scientific Reports.* **6**.
6.Vargas, P. Zardoya, R. (Ed) 2012 *El Árbol de la Vida: Sistémica y evolución de los seres vivos.* Madrid.