Measurement of genomic complexity in Eukarya using Biobit

Nuria De Frutos-Andicoechea [1][2], Pablo Román-Escrivá [1],[2], Vicente Arnau [1],[2],[3], Wladimiro Díaz-Villanueva [1],[2],[3]), Andrés Moya [1],[2],[3].

(1) *Institute for Integrative Systems Biology (I2SysBio), University of Valencia and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain.*

(2) *Foundation for the Promotion of Sanitary and Biomedical Research of Valencian Community (FISABIO), Valencia, Spain.*

(3) *Centro de Investigación Biomédica en Red de Epidemiología y Salud Púbilca (CIBEResp), Madrid, Spain.*

Contact: nudefru@alumni.uv.es

INTRODUCTION

The tendency toward increasing complexity in biological evolution is controversial in biology. Having a complexity measure can help with its resolution. We suggest appealing to genomes to measure complexity because they store information about the biotic and environmental interactions of species in their evolutionary history.

The domain Eukarya contains a broad number of complete sequenced genomes and an extensive evolutionary history that has allowed it to diverge to different levels of complexity. The different groups have adopted various evolutionary strategies reflected in their genomes. An example is the inclination towards polyploidy in plants, tandem and proximal duplications characteristic of amphibians, and genome reduction in birds to decrease energy expenditure. All of them are reflected in the percentage of hapaxes and the value of the anti-entropic component.

MATERIALS & METHODS

Biobit is a k-mers-based metric that measures genomic complexity, establishing as a basis the balance between the anti-entropic and entropic components of the sequence. To do this, the metric establishes two terms based on the anti-entropic (AC) component.

$$BB = \sqrt{\mathrm{AC}}\,(1 - 2AC/LG)^3$$

The **anti-entropy** of the genome sequence is high when there is a high percentage of repeated k-mers. Therefore, to calculate it, we will use the following formula.

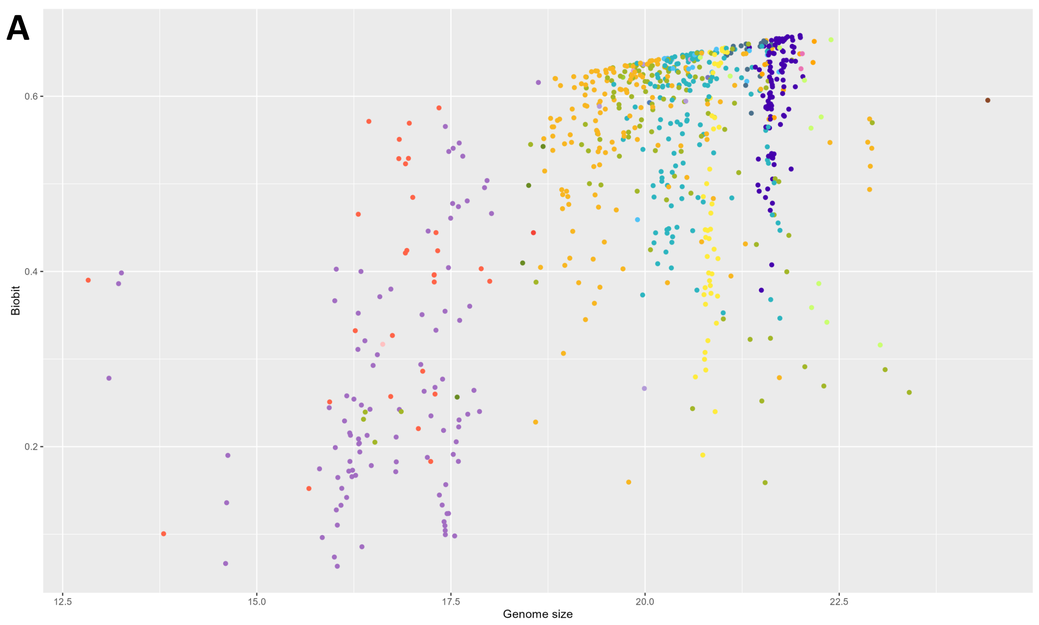
$$\mathrm{AC} = 2\mathrm{LG} - \mathrm{E2LG(G)}$$

Where 2LG is log4(G). It is the upper limit of entropy that can be reached in genomes with the same length, and E2LG(G) is the entropy found in the sequence.

The **hapaxes** are unique k-mers in the sequence. 2LG is used as the k value. When 2LG is not an integer, it is interpolated between k1 and k2, so that k1 < 2LG < k2. Therefore, the percentage of hapaxes will be the number of unique words found relative to the total number of k-mers.

742 eukaryotic genomes were measured at the whole genome and chromosome sequencing level.

RESULTS



The graphs A, B, and C use the legend located on the right side of graph B to characterize the represented taxonomic groups.

- The anti-entropic component in the formula of the Biobit has more effect on the Biobit equation until the value of AC reaches 1. At that point, AC is in a root, and there is almost no increase in this term in the equation, so the term $(1 - 2AC/LG)^3$ has more impact, decreasing at high AC values.

- The anti-entropic component provides a measure inversely proportional to the percentage of hapaxes.

- The highest values of the anti-entropic component correspond to Viridiplantae, specifically to streptophytes plants, just as we expected due to polyploidy. In BB vascular plants also show high values, but lower than Mammalia.

- Another noteworthy case is birds, which have a low variation in the anti-entropic component, therefore on the percentage of hapaxes. The average value is lower than most metazoans. Birds have reduced genomes and smaller cells to increase energy efficiency, so one would expect a reduction in repeated k-mers and a high AC, as observed.

CONCLUSIONS & NEAR FUTURE

- The measurement of hapaxes in the sequence seems to show a close relationship with complexity. It could be a factor to use in evolutionary studies in future investigations.

- This measure allows us to capture various evolutionary strategies present in eukaryotes, so studying its relationship in other groups can help us understand how their sequences have varied and evolved over time.

- The compensation of the anti-entropic component of the Biobit with the second factor of the formula is still an imprecise factor, however the Biobit metric appears to be a good candidate for measuring complexity.

REFERENCES

[1] Moya, A., Oliver, J.L., Verdú, M. et al. 2020. Driven progressive evolution of genome sequence complexity in Cyanobacteria. Sci Rep, 10, 19073, https://doi.org/10.1038/s41598-020-76014-4

[2] Bonnici, V., Manca, V. 2016. Informational laws of genome structures. Scientific Reports, 6(1). https://doi.org/10.1038/srep28840

[3] Tate, J. A., Soltis, D. E., & Soltis, P. S. (2005). Polyploidy in Plants. En T. R. Gregory (Ed.), The Evolution of the Genome (pp. 371-426). Academic Press. https://doi.org/10.1016/B978-012301463-4/50009-7

[4] Chen, W., Chen, H., Liao, J., Tang, M., Qin, H., Zhao, Z., Liu, X., Wu, Y., Jiang, L., Zhang, L., Fang, B., Feng, X., Zhang, B., Reid, K., & Merilä, J. (2023). Chromosome-level genome assembly of a high-altitude-adapted frog (Rana kukunoris) from the Tibetan plateau provides insight into amphibian genome evolution and adaptation. Frontiers in Zoology, 20(1). https://doi.org/10.1186/s12983-022-00482-9

[5]Kretschmer, R., De Oliveira, T. D., De Oliveira Furo, I., Silva, F. A. O., Gunski, R. J., Del Valle Garnero, A., De Bello Cioffi, M., De Oliveira, E. H. C., & De Freitas, T. R. O. (2018). Repetitive DNAs and shrink genomes: A chromosomal analysis in nine columbidae species (Aves, columbiformes). *Genetics and Molecular Biology*, *41*(1), 98-106. https://doi.org/10.1590/1678-4685-gmb-2017-0048

ACKNOWLEDGEMENT

This work was supportedby PMPTA22/00037, CIPROM/2021/042 and CIBEResp. The computations were performed on the HPC cluster Garnatxa at Institute for Integrative Systems Biology (I2SysBio).