Analysis of proposed complexity metrics in the context of the endosymbiosis process

Pablo Román-Escrivá[a,b], Eleonora Paganin[c], Moisès Bernabéu[a], Wladimiro Diaz-Villanueva[a,b,d], Vicente Arnau[a,b,d], Andrés Moya[a,b,d]

a. Institute for Integrative Systems Biology(I2SysBio), Universitat de València(UV) and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain
b. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region(FISABIO), Valencia, Spain
c. Alma Mater Studiorum of Bologna, Italy.
d. Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública(CIBEResp), Madrid, Spain

INTRODUCTION

Establishing the genome complexity of an organism is still an unresolved problem. Different metrics have been proposed in order to quantify the amount of order in a genome and its complexity. Applying those metrics to one of the major evolutionary transition event we can observe (endosymbiosis) has been the approach to test whether or not they are actually able to catch complexity and to try to understand more in deep what features of the genomes those metrics are actually capturing, knowing that free-living bacteria are by definition more complex than their endosymbiont counterparts

MATERIALS AND METHODS

SAMPLE: 159 genomes were used for the analysis, 80 were retrieved from endosymbiont organisms and 79 from their free-living counterparts. To this set two metrics were applied:

BIOBIT (BB)

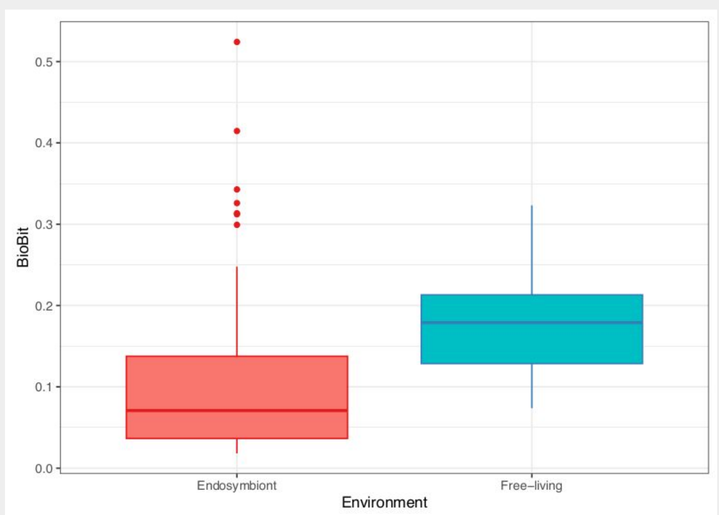
A metric based on the difference between the maximum entropy for a k-mer of a random genome of the same length as the genome under consideration and the entropy of that genome for such a k-mer[2,3]. Higher values of BB should be associated to higher levels of complexity.

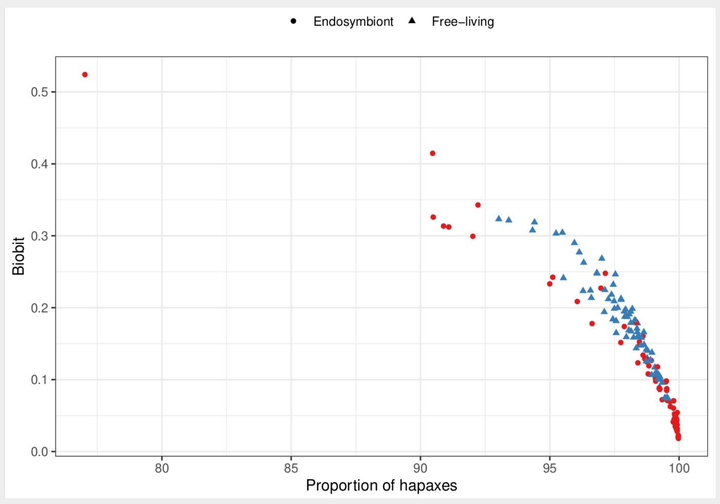
GENOMIC SIGNATURE

A metric that represents the value corresponding to the k-mer that maximizes the difference between observed and expected equi-frequent classes of mers. GS is based on the relative abundances of short oligonucleotides and chaos game representation applied to genomes[1,3]. Higher values of GS should be associated to higher levels of complexity

RESULTS

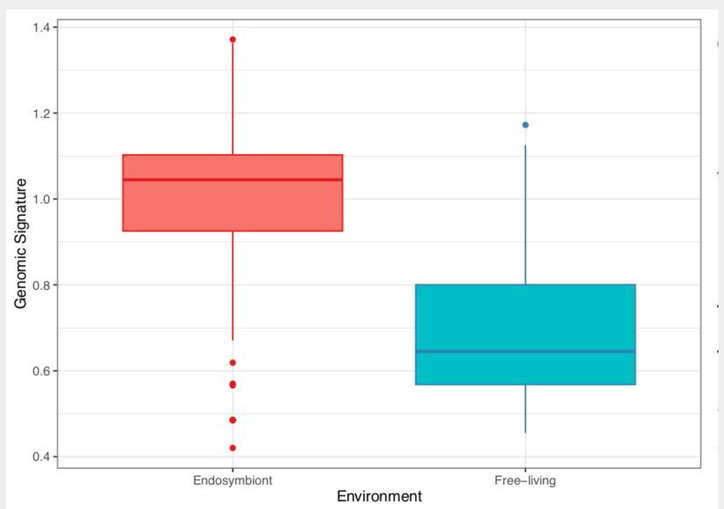
BIOBIT associates higher values to more complex organisms. As we can observe in the boxplots graph the metric is behaving as expected: the free-living bacteria set has BB values that are higher and statistically different than the ones of endosymbionts.





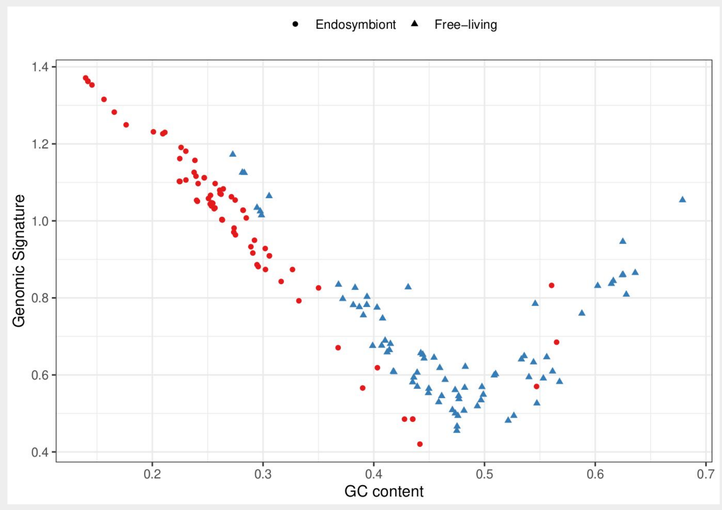
BIOBIT classifies free-living bacteria as more complex organisms since the high percentage of AT content in endosymbionts is recognised to be the result of a random process and not a process whose aim is to introduce order in the genome and increase its complexity.

Actually, the AT content increases the number of hapaxes in the genome, resulting in an increase in randomness instead of order. What's more, the metric allows higher values of entropy in the case of longer genomes, and by doing so it recognises that in short genomes the presence of repeated k-mers may results from randomness while in larger genomes higher frequencies of some k-mers are likely associated to order, higher values of complexity and maybe acquired functionalities.



GENOMIC SIGNATURE does not associate higher values to more complex organisms. As we can observe in the boxplots graph the metric is not behaving as expected: the free-living bacteria set has GS values that are lower and statistically different than the ones of endosymbionts.

GENOMIC SIGNATURE seems to not be effective in measuring the complexity of organisms. Instead, we can see a strong correlation between the GS value and the percentage of GC content we observe in the genome. Actually, the more the distribution of AT and GC content is far from a 50%-50% the more the GS value increases independently if the increase is in AT or GC content.



Likely, since the k value used by GS is lower, an increase of AT or GC content leads to the repetition of k-mers with such content resulting in a k-mers distribution that is far from a uniform distribution and by consequence in a higher GS value. The biological meaning of such correlation needs further studies.

CONCLUSIONS

BIOBIT seems effective in capturing the level of organization and complexity of the genomes thanks to the choice of the k value that is associated to the number of hapaxes in the genome and thanks to its capability of properly taking under consideration the genome length. Instead, the **GENOMIC SIGNATURE** lacks this capacity showing however a strong correlation with high levels of either GC or AT content; further studies need to be carried out in order to understand what characteristic of the genomes of organisms this metric is actually capturing.

CONTACT

pablo.roman@uv.es

ACKNOWLEDGMENTS

This work was supported by FPU21/03813, MICINN-PID2019-105969GB-I00, CIPROM/2021/042 and CIBEResp. The computations were performed on the HPC cluster **Garnatxa** at Institute for Integrative Systems Biology (I2SysBio).

REFERENCES

1. Almeida, J. S., Carrico, J. A., Maretzek, A., Noble, P. A., & Fletcher, M. (2001). Analysis of genomic sequences by chaos game representation. Bioinformatics (Oxford, England), 17(5), 429-437. doi:10.1093/bioinformatics/17.5.429
2. Bonnici, V., & Manca, V. (2016). Informational laws of genome structures. Scientific Reports, 6, 28840. doi:10.1038/srep28840
3. Moya, A., Oliver, J. L., Verdu, M., Delaye, L., Arnau, V., Bernaola-Galvan, P., et al. (2020). Driven progressive evolution of genome sequence complexity in cyanobacteria. Scientific Reports, 10(1), 19073-4. doi:10.1038/s41598-020-76014-4