ChargaffCracker: A software for cracking the generalized version of Chargaff's 2nd rule.

Fuentes, C[1]., Orostica, K[3]., Vidal I[2]., Riadi G[1].
[1]Centro de Bioinformática y Simulación Molecular, Facultad de Ingeniería, Universidad de Talca, Talca - Chile.
[2]Instituto de Matemáticas y Física, Universidad de Talca, Talca - Chile.
[3]Doctorado de ciencias de la Ingeniería, Mención Ingeniería Química y Biotecnología,
Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago - Chile.

Chargaff's second parity rule

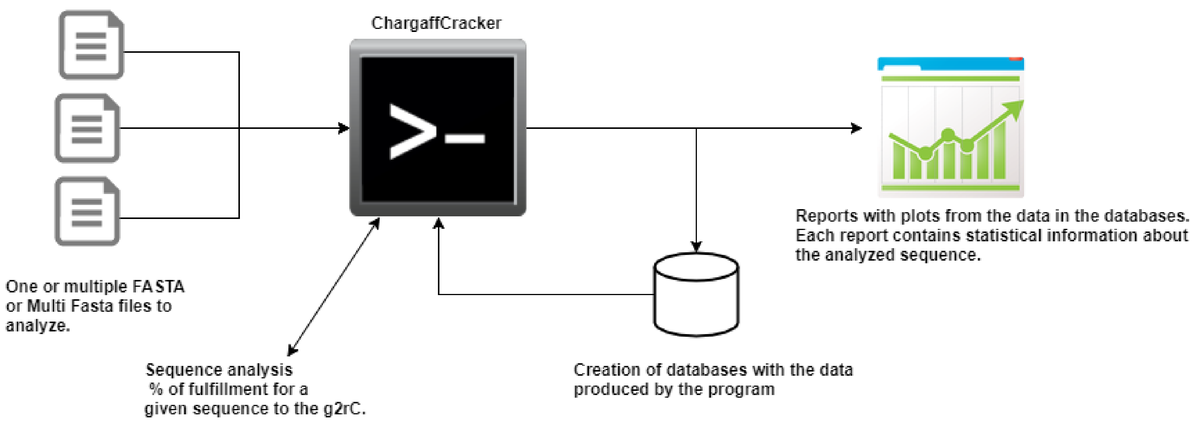
Chargaff's works derived in two laws that explain the proportions of nucleotides in the DNA, the first states that %A = %T and %C = %G, the second that %A ≈ %T and %C ≈ %G in one strand.
The generalization of Chargaff second law of (gC2r) states that for any frequency of an oligonucleotide of length k, in any strand, the frequency of his complementary reverse is similar.

Motivation

The studies around the gC2r have used and absolute measure to determine if certain genome complies with it. This leads to issues that in the end have made difficult to assert any proper explanation to the gC2r from a structural or biological point of view. We propose a new relative measure to assert the comply of a genome to the gC2r.

The software - ChargaffCracker

The only possible way to analyze multiple genomic sequences of all the kingdoms of life to test the compliance of the gC2r is with a software that uses state of the art programming techniques and all the technology available to efficiently compute the required data, this data will be later analyzed to unravel all the information contained within it.



The created test

The compliance is measured per pair of k-mer/k-merRC, using the natural logarithm of the number of times the k-mer is found, divided by the number of times its reverse complement is found in the genome or

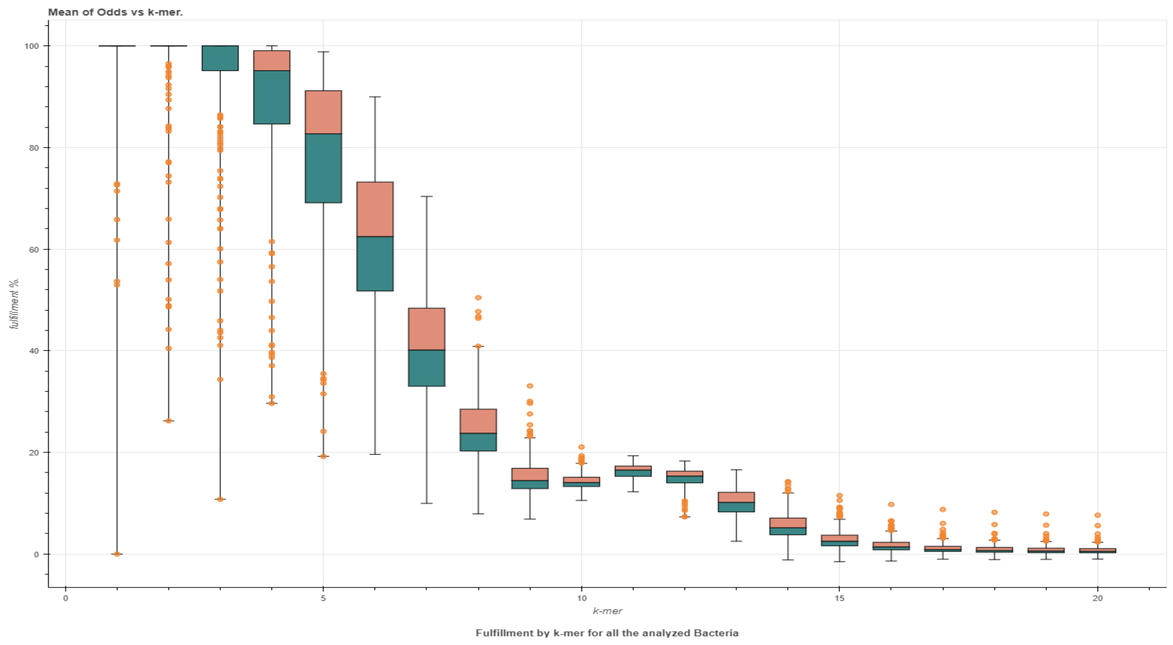
$$\ln\left(\frac{\#kmer}{\#kmer\_RC}\right)$$

This measure is independent of the size of the analyzed k-mer and the size of the genome.

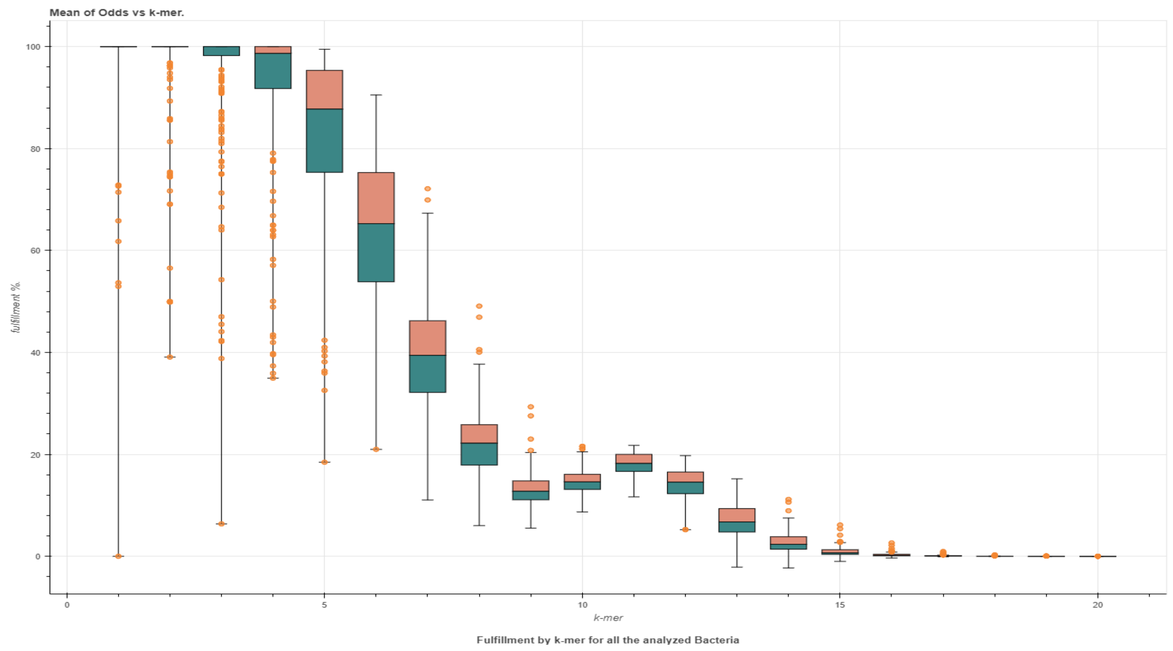
Results

We currently have made analysis to archaea, bacteria, virus and viroids genomes with their random data, and the results for the viruses are shown below.
The data contains:
- 19 Archaeas
- 231 Bacteria
- 36 Viroids
- 1951 Virus

The boxplot show the fulfillment percentage of all the data for each analyzed taxonomy.



Bacteria - Biological data



Bacteria - Random data

Questions

- Is there any difference between a biological sequence and a random one?
- There's a limit to the k-mer size in which all genomes fail to comply with the gC2r?
- If there's a limit, is because of a biological reason?
- It's gC2r a consequence or a cause of Chargaff's first proposed law?

Acknwoledgements

The authors would like to thank the Fondecyt grant #11140869