2nd Bioinformatics Meeting
October 13 and 14, 2016, Universidad de Talca

SCIFORUM
BMEICB-02

Efficient computer implementation to test the validity of the generalized version of Chargaff's second parity rule.

Camilo Fuentes [1], Karen Orostica [2] , Eduardo Alarcón [3] , Ignacio Vidal [3], and Gonzalo Riadi [1,*]

[1] Centro de Bioinformática y Simulación Molecular, Facultad de Ingeniería, Universidad de Talca, 2 Norte 685, Casilla 721, Talca, Chile.
[2] Facultad de de Ciencias Físicas y Matemáticas de la Universidad de Chile, Beaucheff 850, Santiago, Chile.
[3] Instituto de Matemáticas y Física, Universidad de Talca, 2 Norte 685, Casilla 721, Talca, Chile.
* Author to whom correspondence should be addressed; E-Mail: griadi@utalca.cl; Tel.: +5671 2201685; Fax: +5671 2201685.

Abstract: Chargaff's second parity rule holds that for each of the two DNA strands in a genome, the %A is similar to %T and %G is similar to %C. Although the validity of the second rule is still in debate and the biological cause is unknown, a generalized form of the second parity rule has already been proposed. The generalization states that the frequency of a string of a particular length is similar to the frequency of its reverse complement in the same strand. In a previous work, we have developed a statistical hypothesis test for the generalized second Chargaff parity rule for any particular string in a genome. One obstacle to test all available genomes with this statistical test was the efficiency of the computational implementation. In this work, we circumvent this issue, implementing our statistical test in an efficient and multi-processing computer program. The development was carried out in Python, using packets for handling sequences in FASTA format and the required calculations. For each input genome, a database SQLite is generated holding the absolute frequencies of all existing strings in the genome up to a user defined length, and their reverse complements. Multiple sequences are accepted as input, each sequence being analyzed in one CPU. Thus, a bacterial genome, ~4M characters, with 4 sequences takes about 14 seconds to process entirely in 4 CPUs. This computer program will allow our test to be carried out in all

available completely sequenced genomes and assess the validity of Chargaff's second parity rule and its generalized version.

Conflicts of Interest

The authors declare no conflict of interest.