An algorithm to find clusters of information in genome sequences using Shannon Entropy

Rafael P. Simões[a 1], Lauana Fogaça[b], Bruno A. Corrêa[a] and Guilherme T. Valente[a]

[a]UNESP - Univ Estadual Paulista, Campus Botucatu, Departamento de Bioprocessos e Biotecnologia, Botucatu, SP, Brazil

[b]UNESP - Univ Estadual Paulista, Campus Botucatu, Instituto de Biociências, Botucatu, SP, Brazil

Abstract

The genomic is currently subject to studies not only of biologists. Physicist, chemist, mathematician and statistician researchers are also focusing on this topic. This occurs because some genomic sequence carries information that can be analysed under perspectives of different areas, which together can corroborate to understanding the biological functions of these sequences. One of goals of this research area is the correct identification of the organized genomic segments, which must be the key to find the clusters that could be related to some biological activities. The Shannon entropy was the first concept concerning entropy using information theory. Usually, the Shannon Entropy model has been performed by similarity or yet by the association with tools derived from the chaos dynamics and repeated random walks. This work presents a computational algorithm applying the physical concept of Shannon entropy to identify the organized nucleotides portions and to better choose informative regions in a genome for single nucleotide sequences, what the traditional treatments for a DNA sequence are unable to produce. This algorithm suggests the overlapping of nucleotides segments of variable length from the same genome sequence. The algorithm was implemented as a Fortran script which was used to perform a thorough analysis of four classical and well-known genome sequences, the: *Arabidopsis thaliana*, *Drosophila melanogaster*, *Escherichia coli* and *Saccharomyces cerevisiae*. The results reveal that there are high and low entropy regions in all these genome sequences, and the low entropy genome segments can contain relevant biological information. This denotes the algorithm as a potential bioinformatics tool to increase the genome annotation qualities and so on.

Keywords: Shannon entropy, information, genomic, computational physics, bioinformatics.

[1]E-mail Corresponding Author: rafael@fca.unesp.br