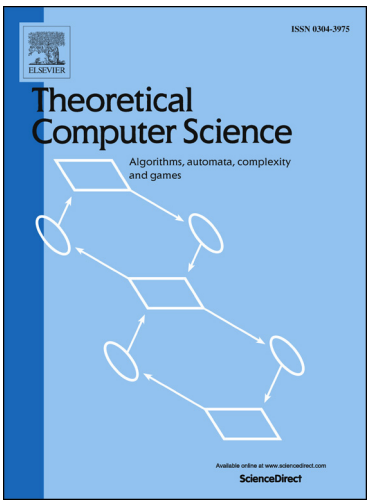
Accepted Manuscript



The principles of informational genomics

Vincenzo Manca

PII: S0304-3975(17)30276-1
DOI: http://dx.doi.org/10.1016/j.tcs.2017.02.035
Reference: TCS 11144

To appear in: *Theoretical Computer Science*

Received date: 29 October 2016
Revised date: 11 February 2017
Accepted date: 19 February 2017

Please cite this article in press as: V. Manca, The principles of informational genomics, *Theoret. Comput. Sci.* (2017), http://dx.doi.org/10.1016/j.tcs.2017.02.035

The Principles of Informational Genomics

Vincenzo Manca

University of Verona

Abstract

The present paper investigates the properties of genomes directly related with their long linear structure. A systematic approach is introduced that is based on an integration of string analysis and information theory, applied and verified on real genomes. New concepts and results are given in connection with genome empirical entropies (and related indexes) [5], genome dictionaries and distributions, word elongations, informational divergences, genome assemblies, and genome segmentations.

Keywords: Genomes, Information Theory, Genomic Distributions, Genomic Indexes, Genomic Dictionaries, Genomic Entropies

1. Introduction

Genomes are containers of biological information that direct the functions of the organisms and transmit biological information along generations, but at same time they evolve by producing new species from previous ones in the tree of the life. For this reason, in a sense, they are both instruments of biological tradition and of biological innovation. This can be realized without contradiction because the two phenomena follow very different time scales. The transmission of individual genomes from parents to children follow the scale of individual generation, whereas the generation of derived species from primitive species follow scales of time of many orders larger. Individuals are instances of species, but species exist only by means of individuals and, in a sense, they are only an abstract concept.

From a mathematical point of view (individual) genomes are strings over an alphabet of four symbols, and more precisely, they are "long" strings (typically of lengths in the range of $10^5 - 10^{10}$). Informational genomics (shortly *infogenomics*) is the study of such a kind of strings, in order to discover which