Bioinformatics, Proteomics and Immaging Analysis

OMMEGA Publishers
Answering Intelligence
www.ommegaonline.com

Editorial

OPEN ACCESS

Research Lines in Infogenomics

Vincenzo Manca*

Department of Computer Science, University of Verona, Strada le Grazie, Verona, Italy

Corresponding Author: Manca, V. Department of Computer Science, University of Verona, Italy.
E-mail: vincenzo.manca@univr.it

Received Date: August 25, 2015
Accepted Date: August 26, 2015
Published Date: August 27, 2015

Citation: Manca, V. Research lines in Infogenomics (2015) Bioinfo Proteom Img Anal 1(1):25- 26.

Infogenomics intends to develop genome analyses based on a pure informational perspective. The main objectives of Infogenomics[1,2] are the following: i) the definition of parameters, with their computation in real genomes, that can discriminates interesting genomic features and properties, ii) the definition, with their computation in real genomes, of statistical distributions providing interesting profiles associated to genomic elements, and finally iii) the extraction of informationally relevant genomic words, in genomes or classes of genomes. Concerning the last point, many possible methods can be developed, and an adequate notion of informational relevance needs to be elaborated. Some ingredients are important to this end. Firstly, the basic notions of Shannon information theory (and their following elaborations) as information source, entropy, entropic divergences, mutual information, randomness and typical distribution of random processes (Gaussiam, Poisson, and exponential distributions). Moreover, basic concepts from formal language theory are useful, in order to apply probabilistic concepts to strings and sets of strings. “Short” strings are the main constituents of the “long” strings constituting genomes, and surely some of them are the components of genome constructions and developments, during the evolutive process that produced them (assembling longer sequences from some initial minimal genomes, or RNA genomes of proto-organisms). In fact, if genomes convey information, then minimal units (words) have to exist that increased by making evolution possible by realizing more complex biological functions. The notion of “word”, defined in terms of pure distributional and syntactical analysis, was at the origin of many linguistic debates. For genomes the challenge is even more difficult, because many clues typical of human communication are missing. However, the power of computational analysis and the availability of genomic data suggests to claim that now it could be time to reconsider the problem of discovering genomic codes (probably, for specific genomes or classes of them), in order to put forward what genetic code started, but at a deeper level of comprehension. The starting point of such an approach is the identification of mathematically grounded computational tools for constructing suitable genomic dictionaries[3] (sets of strings occurring in given genomes). In this way, starting from simple dictionaries consisting of words with fixed length, we will need to generate, by means of suitable operations (selection, combination, integration), dictionaries of variable length that have to be evaluated according to specific properties of adequacy (coverage, multiplicity, length, repeatability, localization, context dependence or independence, recurrence). In this process, some statistical distributions will guide the choice of parameters and the ranges of their values. An essential perspective of this approach is the interplay between theoretic informational analysis, inspired by general and abstract notions, definable on long strings, and by their experimental evidence in real genomes, obtained by means of suitable algorithms, data representations, and software. We want to remark that the notion of dictionary is implicitly present in the crucial problem of genome sequencing. Here the terminology is quite different, but essentially the main problem solved by the sequencing methods is always the reconstruction of a sequence hosting a given set of “reads”, that is a dictionary of words (in many copies) coming from random fragmentation of the original genome. This is an indirect proof that all the information contained in a genome can be entirely recovered (within an approximation threshold) from a suitable dictionary of its words. Another simple argument can show the crucial aspect that dictionaries play in the analyses of genomes. Let us consider a genome G with a length of 10^6 bases. All the sequences of length 30 which we encounter by scanning all the genomes are (10^6 - 29), moreover possibly many of them occur many times. However, the number of possible different words of four letters having length 30 is 4^{30} which is a value greater than 10^{18}. This means that the part of sequences of length 30 which occur in that genome is a portion 10^{-12} smaller than the whole set of possible words. This simple numerical evaluation tells us that surely sequences of length 30 were selected among a huge number of other possible sequences. On the side of molecular biology and biochemistry, many international projects are active for deciphering genomes, in order to pass from the knowledge of genome sequences to their biological functions. In particular, the project ENCODE (ENCyclopedia of DNA Elements) is mainly aimed to extract lexicons, and catalogs of biochemically annotated fragments in human genome. A very complex dynamics of interactions results among DNA regions, proteins, and RNAs, with a lot of newly identified elements, and with a huge number of data (see websites: http://nature.com/encode, http://epd.vital-it.ch). Infogenomics, which relies on a completely different approach, intends to realize a similar task starting from Information theory. In fact, if genomes are containers (and generators) of biological information, an informational annotation of genomes could be a complementary perspective with respect to the "biochemical" identification of relevant genome fragments.

References

1. Castellini, A., Franco, G., Manca, V. A dictionary based informational genome analysis. (2012) BMC Genomics 13:485.
2. Manca. V. Infobiotics: information in biotic systems (2013) Springer.
3. Bonnici, V., Manca, V. Infogenomics Tools: A Computational Suite for Informational Analyses of Genomes. (2015) J of Bioinfo Proteomics Rev 1(1): 7-14.