From: AAAI Technical Report SS-95-03. Compilation copyright © 1995, AAAI (www.aaai.org). All rights reserved.

Genome Classification Systematics: Application to Human and Simian Immunodeficiency Viruses

Alexander A. Filyukov

Keldysh Institute of Applied Mathematics
Russian Academy of Sciences
Moscow 125047, Russia

Abstract

My contribution to the AAAI 1995 Spring Symposium on *Systematic Methods of Scientific Discovery* starts with the brief sketch of consequences which followed the discovery that genomic molecules are existing and operating in nonequilibrium excited steady state.

The proof of the latter statement was obtained through computer assisted investigations of genomic sequence of a particular virus. The strategy used for investigation was based on the general definition of any individual organism as a Gibbsian ensemble of identical personal DNA molecules. This approach provides application of methods of statistical thermodynamics of irreversible steady processes to genome informatics. The random processes theory and its Markov chains approximation lead in this approach directly to the definition of the generalized concept of evolution entropy and to the genuine measure of text information content in the sequences. Computational proofs of the existence of the nonequilibrium steady state conditions in particular viral genome molecule were obtained by investigation of special type triple cyclic balance relations which are specific to the steady-state systems. The main maxima of the text information content in the genomic sequence were established, decoded and denominated. It was found that coding principles are connected with deviations from equipartition of the nucleotides and with deviations from 'equilibrium conditions', expressed as violations of detailed balance for inverse dinucleotides in genomic sequences.

The independent confirmation of this discovery was provided by G.S. Mani from the Department of Theoretical Physics, Schuster Laboratory, Manchester University, UK. The verification of above mentioned results was obtained while my paper was in press. GS̈ Mani proposed a group of computational methods for identifying coding regions. They are related to codon usage and preference methods and, also, related to local non-random detecting methods. All the mentioned methods are based on the computational proof of the existence of the unique periodical cycle with a frequency of 1/3; that is a 3–point cycle that is displayed for all possible doublets. This result was provided by the application of a discrete Fourier transformations technique. It was found that only coding regions demonstrated these three-point cycles. Therefore, noncoding regions are similar to random sequences with a continuous spectra of frequences.