Abstract for GR-TR Conference on Statistical Mechanics and Dynamical Systems

Topic: Complex Systems in Biology Preference: Oral

Non linear analysis of oligonucleotide distribution of evolutionary recent organisms

P. Katsaloulis[1*], A. Provata[1], T. Theoharis[2]

[1] Institute of Physical Chemistry, National Center for Scientific Research "Demokritos", Athens, Greece
[2] Department of Informatics and Communication, University of Athens, Greece
[*] Electronic Address: pkatsaloulis@chem.demokritos.gr

Computational DNA sequence analysis enhances our comprehension of complex biological systems and improves our knowledge on how biological information is stored and retrieved in the DNA [1, 2]. Following this approach we have focused on the statistical analysis of small DNA sequences (oligonucleotides) in evolutionary recent organisms. We have used chromosomes of most sequenced multi-celled organisms, and especially *Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana*.

Our approach is based on calculations of various scaling statistical parameters for all oligonucleotides of the same length [3]. Sequences of 5 and 6 nucleotides are used, depending on the size of the original chromosome. The distance distribution of consequent appearances of the same oligonucleotide is computed, for all possible nucleotides of a given length. Two parameters are used to model this size distribution, in order to find if long or short range tendency was present [4].

We have found that oligonucleotides bearing consensus promoter signatures follow power law distributions, while all others follow exponential distributions, or in general short range distributions. In a two dimensional plot of the two parameters, we have seen that oligonucleotides tend to cluster, depending on whether they contain the "CG" subsequence (a common consensus sequence of the promoter). In more resent organisms this phenomenon is more evident.

---

[1] W. T. Li and D. Holste, Universal 1/f noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome, *Phys. Rev. E*, **71**, 041910 (2005).

[2] P. Bernaola-Galvan, P. Carpena, R. Oman-Roldan, J.L. Oliver, Study of statistical correlations in DNA sequences, *Gene*, **200**, p.105 (2002).

[3] P. Katsaloulis, T. Theoharis, A. Provata, Statistical algorithms for long DNA sequences: oligonucleotide distributions and homogeneity maps, *Scientific Programming*, **13(3)**, p.177, (2005).

[4] P. Katsaloulis, T. Theoharis, W.M. Zheng, B.L. Hao, A. Bountis, Y. Almirantis, A. Provata, Long range correlations of RNA polymerase II promoter sequences across organisms, *Physica A*, **366**, p.308 (2006)