Research Project EPSD 4308

Non-equilibrium Thermodynamics Linking the Structure and Functionality in DNA

by

Astero Provata

August 22, 2013

Scientific Report

Purpose of the visit

The aim of this project was to build on ideas initiated on September 2012, during a short sabbatical stay of the grantee at the Interdisciplinary Center for Nonlinear Phenomena and Complex Systems, Univerite Libre de Bruxelles (ULB). These ideas concern a novel characterisation of the primary structure of DNA sequences using the theory of Non-equilibrium Thermodynamics and of Complex Systems in order to connect the local basepair structure and statistics with the DNA functionality, building on the recent advances in the field of Symbolic Dynamics. These ideas are also likely to be applicable to any symbol sequences with finite alphabets.

Description of the work carried out during the visit

A multitude of methods borrowed from the field of Non-equilibrium Thermodynamics were used to test stationarity, irreversibility and correlations in human DNA. As a working data a long contig of Homo Sapiens, Chromosome 20 was used. This genomic contig is the locus N1_ 011387 (primary assembly) and contains 26259569 base pairs (bps), while the entire Chromosome 20 contains ~ $63\cdot 10^6$ bps. N1_ 011387 contig is a DNA entity long enough to ensure good statistics, when addressing both the short and long range spatial properties. Moreover it is representative of the entire DNA molecule, since it contains both coding and non-coding sequences and other functional elements in similar densities as for all other human chromosomes. The contig was represented either as a two-letter sequence based on purines-pyrimidines or on a four-letter representation using the {A,C,G,T} nucleotides.

A preliminary statistical analysis was first undertaken to find symbol frequencies and transition matrices between the different symbols . To test the stationarity properties of the sequence a Markov chain analysis was used of order up to 6. To test spatial asymmetry and detailed balance the probability fluxes were calculated and compared both for the two- and the four-letter alphabet. A series of entropy related quantities were evaluated the dynamical complexity and information transfer along different parts of the chain. Exit length and recurrence length distributions were computed to test for long or/and short range correlations along the chain. To further verify quantitatively the existence of long-range correlations the Hurst exponent was computed. Finally, a construction algorithm of a “model DNA” was discussed. This algorithm is based on the Monte Carlo rejection sampling method and is appropriate for transferring all the statistical features of the natural sequence on the artificial one. The structure of the natural and artificial sequence were compared using different criteria.

Furthermore I had interesting discussions with several members of the Interdisciplinary Center for Nonlinear Phenomena and Complex Systems. With Prof. Anne de Wit we discussed on fractal pattern formations in 2D reaction-diffusion systems and on viscous fingering with applications to environmental sciences. With Prof. Pierre Gaspard we discussed on irreversibility issues related to the human genome. With Prof. Yannick De Decker and his PhD student Domenico Bullara we discussed on reactive dynamics on low dimensional lattices with intention to open a collaboration in this direction.

Description of the main results obtained

Our investigations have shown that human DNA represented as a symbolic sequence (either in the two- or the four-letter representation) presents intricate correlations influencing many spatial scales. In particular Markov chain analysis has shown that human DNA can not be describe as a low order Markov chain, as orders below r=6 are not enough for its faithful description. The test for detailed balance has given very interesting results. Although detailed balance seems to hold in the case of the two-letter representation it does not hold in the case of the four-symbols. Likewise, block-entropy increases nonlinearly with the block size but presents a power law increase with exponent of the order 1.339 for block size n=1, to 1.273 for block size n=8. Information-like quantities demonstrate information exchange down the chain for at least 100 bps.

The distribution of exit-distances of a certain symbol along the sequence were computed for two purposes: a) for the search of long range correlations and b) as a basic ingredient for the modelling of the DNA. The exit-distance distributions were shown to include power-law tails and this was further corroborated by similar power-tails in the recurrence-length distributions. The study of the Hurst exponent verified further the existence of long-range correlations giving a Hurst exponent H=0.6145, different from the random and uncorrelated sequence where H=0.5.

Based solely on the results on the exit-length distributions for the case of the four-letter alphabet we were able to reconstruct an artificial DNA sequence using the Monte Carlo rejection sampling method. The artificial sequence obtained with this method contains all the statistical correlations as the original natural sequence (i.e. same Hurst exponent, same entropic, recurrence and information properties, some correlations) but fails in the local details of the structure as these are depicted by the point-to-point Hamming distances.

All above results are still under discussion principally for their interpretation with respect to the spatial characteristics of the primary human DNA structure and for understanding the possible connection between structure and functionality.

Future collaboration with host institution

All results obtained during this collaboration period concern solely the human genome. I intent to continue my collaboration with Profs. G. and C. Nicolis and explore the ideas of irreversibility and information content on particular DNA elements, such as coding DNA, noncoding DNA, repeats etc. Similarly, we would like to extend our research in the field of comparative genomics applying these ideas in organisms with intrinsically different genomic structure (eg. comparison between procaryotic and eucaryotic genomes). For this purpose, Prof. G. Nicolis is planning to visit my group in Athens in October 2013.

I addition, I plan to pursue further possibilities for collaboration with other members of the Interdisciplinary Center for Nonlinear Phenomena and Complex Systems, Univerite Libre de Bruxelles. Related to these collaborations Dr. Florence Haudin from the group of Prof. Anne de Wit is planning to visit my group in November 2013 and my PhD student Evelyn Panagakou is planning to visit Profs. Yannick De Decker and Anne de Wit in January 2014.

Projected publications

My collaboration with Prof. G. Nicolis and Prof. C. Nicolis is a "work in progress" at the moment, as we are in the process of composing a manuscript entitled *"DNA viewed as an out-of-equilibrium structure"* to be submitted for publication. An acknowledgement to the support received by the European Science Foundation will appear in this publication.

Related Literature

[EN1991] W. Ebeling and G. Nicolis G, ``Entropy of symbolic sequences: the role of correlations", Europhys. Lett., **14**, 191-196 (1991).

[AV2011] A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. d'Aubenton-Carafa and C. Thermes, Physics Reports, **498** 45–188 (2011).

[NN2012] G Nicolis and C. Nicolis, "Foundations of Complex Systems", 2nd edition, World Scientific, Singapore (2012).

[PB2011] Provata A., Beck C., "Multifractal analysis of nonhyperbolic coupled map lattices: Application to genomic sequences", *Phys. Rev. E,* **83**, 066210 (2011).

[BP2011] Beck C., Provata A., "Multifractal information production of the human genome", *Europhys. Letts.,* **95**, 58002 (2011).

[PB2012]Provata A., Beck C., "Coupled intermittent maps modeling the statistics of genomic sequences: A network approach", *Phys. Rev. E,* **86**, 046101 (2012).