Computational Biology and Chemistry 53 (2014) 5–14

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Complexity measures for the evolutionary categorization of organisms

A. Provata[a,*], C. Nicolis[b], G. Nicolis[c]

[a] Institute of Nanoscience and Nanotechnology, National Center for Scientific Research "Demokritos", 15310 Athens, Greece
[b] Institut Royal Météorogique de Belgique, 3 Avenue Circulaire, 1180 Bruxelles, Belgium
[c] Interdisciplinary Center for Nonlinear Phenomena and Complex Systems, Université Libre de Bruxelles, Campus Plaine, C.P. 231, 1050 Bruxelles, Belgium

ARTICLE INFO

Article history:
Accepted 11 July 2014
Available online 28 August 2014

MSC:
00-01
99-00

Keywords:
Genomic sequences
Irreversibly
Probability fluxes
Block entropy

ABSTRACT

Complexity measures are used to compare the genomic characteristics of five organisms belonging to distinct classes spanning the evolutionary tree: higher eukaryotes, amoebae, unicellular eukaryotes and bacteria. The comparisons are undertaken using the full four-letter alphabet and the coarse grained two-letter alphabets AG-CT and AT-CG. We show that the conditional probability matrix for the four-letter and AT-CG alphabet is markedly asymmetric in eukaryotes while it is nearly symmetric in bacterial genomes. Spatial asymmetry is revealed in the four-letter alphabet, signifying that the probability fluxes are nonvanishing and thus the reading sense of a sequence is irreversible for all organisms. Calculations of the block entropy and excess entropy demonstrate that the human genome accommodates better all possible block configurations, especially for long blocks. With respect to point-to-point details and to spatial arrangement of blocks the exit distance distributions from a particular letter demonstrate long distance characteristics in the eukaryotic sequences for all three alphabets, while the bacterial (prokaryotic) genomes deviate indicating short range characteristics. Overall, the conditional probability, the fluxes, the block entropy content and the exit distance distributions can be used as markers, discriminating between eukaryotic and prokaryotic DNA, allowing in many cases to discern details related to finer classes. In all cases the reduction from four letters to two masks some important statistical and spatial properties, with the AT-CG alphabet having higher ability of discrimination than the AG-CT one. In particular, the AT-CG alphabet reduction accentuates the CpG related properties (conditional probabilities w_{32}, long ranged exit distance distribution for A and T nucleotides), but masks sequence asymmetry and irreversibility in all examined organisms.

1. Introduction

Traditionally the evolutionary tree construction is based on the phylogenetic characteristics determining the lineage between species and on affinity in their morphology. Recently, with the revolutionary developments in the field of Molecular Biology and in particular with the continuous recoding of the genome of many species, details on the structure of the evolutionary tree have been adjusted or modified as a result of genomic comparison between species. The continuous increase of computational speed in data processing and data storage capacity has greatly contributed to this purpose. Although point-to-point genomic comparison of organisms, an easy task for today's computers, can reveal common properties and differences between organisms, this can be used only locally and is tedious for large data sets, is time consuming and in most cases produces enormous amounts of information not particularly useful and in many cases erroneous. For this reason it is convenient to use quantitative indices pertaining to global properties, which together with local criteria can be used to understand the species evolution, to reveal relations between organisms and to help in more accurate classification of organisms.

Attempts to use collective quantitative indices to characterize globally DNA structures date from early 1990s. In their pioneering works, Peng et al. (1992), Li and Kaneko (1992) and Voss (1992) have demonstrated the presence of long range correlations in noncoding DNA sequences, differentiating them from structures related to Markov-type stochastic processes. These findings were further enriched during recent years. In particular, using the chaos game representation (Hao, 2000; Deng and Luan, 2013) Hao demonstrated the existence of fractal patterns in DNA sequences covering many hierarchical orders (Wu, 2003). Fractality was demonstrated in primary structure of DNA in the alternation of coding and noncoding segments seeing as a finite Cantor-like

* Corresponding author.
E-mail addresses: aprovata@chem.demokritos.gr (A. Provata), cnicolis@oma.be (C. Nicolis), gnicolis@ulb.ac.be (G. Nicolis).

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.004
1476-9271/

construction (Provata and Almirantis, 2000), while wavelet fractal analysis has demonstrated the scaling properties inherent in the noncoding parts of the genome (Arneodo et al., 1995, 1996, 2011). In parallel, interesting correlations related to the presence of the CpG islands were discovered (Arneodo et al., 2011; Polak and Arndt, 2009; Li and Miramontes, 2006; Zhang and Chen, 2005; Freudenberg et al., 2009; Li and Holste, 2005; Clement and Arndt, 2011; Carpena et al., 2011) and their occurrence was connected with the central role of the CG complex as a structural element of the promoter sequences (Katsaloulis et al., 2006, 2009). Long range correlations were further demonstrated in the word frequency (Ebeling and Nicolis, 1991), in the information content (Roman-Roldann et al., 1996; Provata et al., 2014), in the size distribution of noncoding sequences (Almirantis and Provata, 1999), in the distribution of repeats (Massip and Arndt, 2013) and in the persistence and antipersistence of symbols (Almirantis and Provata, 1999; Melnyk et al., 2005; Zhou et al., 2004; Marx et al., 2006). Strand asymmetry has also been studied extensively in DNA literature, in the sense of a departure from intrastrand equifrequency between the A and T nucleotides and the C and G ones (Elson and Chargaff, 1952; Rudner et al., 1968; Lobry, 1996; Francino and Ochman, 1997). Equifrequency between A and T nucleotides (and C and G ones), in the double stranded DNA is now known to hold exactly and is termed as the first Chargaff rule. Instead, the second Chargaff rules which claims equifrequency between (A,T) and (C,G) within one strand is only approximate (Lobry, 1996; Francino and Ochman, 1997).

In this work we further demonstrate the universality of the above results across organisms. We examine long DNA sequences and whole chromosomes from five organisms belonging to distinctly different evolutionary classes, the primates (one organism, *Homo sapiens*), the transition single- to multi-cell eukaryotes (one organism, *Polysphondylium pallidum*), the single cell eukaryotes (one organism, *Saccharomyces cerevisiae*) and bacteria (two organisms, *Escherichia coli* and *Bacillus subtilis*). For their comparison we use the evolutionary measures previously applied to the human chromosome 10 (Provata et al., 2014). These measures are borrowed from the theory of nonlinear dynamics, statistical mechanics and information theory and they are investigated here as potential indicators of evolution, permitting to discriminate between organisms belonging to different evolutionary classes. In particular, regarding strand asymmetry we develop a different view than the one usually adopted in the literature. Specifically, as argued in a recent work by the present authors (Provata et al., 2014), spatial asymmetry is probed in single DNA strands and is found to give different information content when read from left-to-right than from right-to-left, indicating that DNA chromosomes can be regarded as out-of-equilibrium structures. It must be noted that the difference in the A and T content (or the C and G one) is not identical with the asymmetry produced in the reading sense. One particular factor which accentuates asymmetry is clustering of homologous nucleotides. Indeed, in Provata et al. (2014) long range correlations were recorded in the clustering properties (exit distance and recurrence distributions), a characteristic frequently observed in out-of-equilibrium structures. In the same work the natural human DNA strands were compared against different stochastic processes using entropy and information measures, demonstrating marked deviation from stochasticity.

We use for the comparison three types of alphabets, the full four-letter alphabet and two coarse grained ones. In the four-letter alphabet the four nucleotides (Adenine = A, Cytosine = C, Guanine = G, Thymine = T) are the elementary symbols. In the two coarse grained alphabets the nucleotides are classified in two groups. In the classic two-letter alphabet of purines (A,G) and pyrimidines (C,T), called for short AG-CT alphabet, the classification is based on chemical affinity of the nucleotides. Indeed, the chemical structure of the two purines is very similar and the same holds for the pyrimidines. This similarity has led to the general belief that the two purines the two pyrimidines originate from the same primitive units. This AG-CT classification retains the most primitive evolutionary characteristics in the sequence. The second two-letter alphabet analyzed is the AT-CG alphabet where the classification is based on the observation that (A,T) are grouped together to form a weak H-bond group, whereas (C,G) form a strong H-bond group. Because of this particular grouping the AT-CG alphabet accentuates the presence of the CpG islands which is frequent in the human genome (Deng and Luan, 2013).

In the next section we summarize the principal genomic characteristics of the five organisms, for the purposes of the interpretation of the quantitative indices and the comparison between the organisms carried out in the subsequent sections. In Section 3 we compute the conditional probabilities for the presence of doublets and compare them between the different organisms. The conditional probabilities are used as input in Section 4 to address questions of asymmetry and sequence irreversibility. In Section 5 entropic quantities are used to assess DNA as an information repository and to compare the DNA information content of the five organisms. In Section 6 the long range properties of the organisms are compared and they are shown to follow some evolutionary path: shorter ranges for bacteria, intermediate ranges for the lower eukaryotes and longer ranges for *H. sapiens*. In the concluding section we recapitulate our main findings and discuss results and open problems.

2. Description of representative genomes and organisms

For the genomic analysis we have chosen to compare test organisms which are distant in the phylogenetic tree and in their phenotypes. The question is whether the complexity measures of their genome can reflect the evolutionary characteristics of the organisms. The choice of the five organisms was based on:

- The availability of the DNA sequences in the international databases.
- The size of the genome. Large genomes are preferable for reliable statistical analyses. In this analysis we do not use merging of data or averaging over many sequences/organisms since we cannot assume that all organisms follow the same statistics.
- Quality of the data. Genomes with the lowest possible percentage of unknown base-pairs (bps) are preferable.
- Whole chromosomes are preferable when available.

Although many organisms are found listed as sequenced in the databases, many of them are partially sequenced and their records contain a large percentage of unknown bps, generally denoted by the letter *N*. In many cases *N*s are scattered in the genome while in other cases they form large gaps between successive contigs.

Taking all these into consideration we selected the largest available sequences containing the lowest possible percentage of unknown bps of the following five test organisms: the higher eukaryote *H. sapiens*, the amoeba *P. pallidum* which is an intermediate eukaryote between unicellular and multicellular organisms, the unicellular eukaryote yeast *S. cerevisiae* and the two model bacteria *E. coli* and *B. subtilis*. Here we describe the genome characteristics of the representative organisms and more specifically we give biological details on the data analyzed below.

2.1. H. sapiens

From the higher eukaryotes we have chosen to work with two large contigs from chromosomes 10 and 14 of *H. sapiens*, denoted