Journal of Siberian Federal University. Mathematics & Physics 2014, 7(4), 530–532

УДК 57:015 + 573.2

Very Low Ergodicity of Real Genomes

Michael G. Sadovsky*
Institute of computational modelling SB RAS
Akademgorodok, Krasnoyarsk, 660036 Russia

Ksenia A. Nikitina†
Institute of Space and Information Technology
Siberian Federal University
Kirenskogo 26, Krasnoyarsk, 660074
Russia

Received 30.08.2014, received in revised form 30.09.2014, accepted 05.11.2014

A distribution of the longest distances to the nearest neighbour alongside a genetic sequence is investigated. It is shown the real DNA sequences differ basically from any model one, while the figures of the distribution prove almost lack of ergodicity in these latter.

Keywords: order, triplet, Markov process, distance, mutual distribution.

Introduction

Finite symbol sequence is a typical mathematical object; eventually, it is natural for any living being as DNA sequence. Further we consider finite symbol sequences of (various) length N, from four-letter alphabet $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. No other symbol takes place in a sequence; a sequence is also supposed to be coherent (i. e. consisting of a single piece).

We study the mutual distribution of short strings (the words of the length $q = 3$, to be exact) alongside a sequence. So called the distribution to the nearest neighbour was studied. That latter is a distribution of lengthes of the fragments to be found between two given triplets $\nu_1\nu_2\nu_3$ and $\nu'_1\nu'_2\nu'_3$ meeting the following constraint: there is no other triplet $\nu'_1\nu'_2\nu'_3$ somewhere in between the given couple, so this triplet is the closest (right) neighbour to the starting one $\nu_1\nu_2\nu_3$. Paper [1] reports on this kind of studies firstly.

The distributions as described above show "heavy tails": an expectedly high probability (estimated through the comparison to Markov chain models, of various orders) of long gaps between the triplets is tremendously less in comparison to the observed one. Skipping the details of the distribution (while interesting and important themselves), we concentrate here on the study of the longest possible gap to be observed between two given triplets.

1. "Long tails" as it is

We studied the genomes of various organisms with taxonomy ranging from bacteria to man. Very few results are present here, due to the paper size restriction. For each sequence, a matrix

*msad@icm.krasn.ru
†ksklokova@yandex.ru

of the longest distances to be observed between all 4096 couples of the triples was developed. Since genetic sequences differ drastically in length, for various organisms, we used a figure of the relative length of the longest gap. To take into account the variation in genetic entity length, we introduced the following measure to compare different entities:

$$r^*_{\langle\omega_1,\omega_2\rangle} = \frac{l_{\langle\omega_1,\omega_2\rangle}}{\ln N}\,, \tag{1}$$

where $l_{\langle\omega_1,\omega_2\rangle}$ is the longest gap observed for the given couple $\langle\omega_1,\omega_2\rangle$, and N is the length of a sequence. The normalization (1) is not unique, while we shall not here discuss the other ones. The motivation standing behind the normalization (1) comes, in some sense, from [2–4].

In general, pattern of the distribution of the longest gaps observed over the ensemble of the triplet couples resembles rather similar, for different organisms. Of course, different organisms exhibit different order of the couples, when ordered according to (1) figure. Fig. 1 illustrates this fact; surely, the order of couples is specific, for each genetic entity, so different couples occupy the same position at the horizontal axis in this figure.

2. Results and discussion

The paper presents an approach to figure out a new structure in genetic sequences manifested through the mutual distribution of triplets observed alongside a sequence. To be exact, we studied not the distribution of the nearest neighbours, but the longest tail of that latter. The figures of this ultimate distances are very big, so raising the frequencies of extra long tails to be found in real sequences.

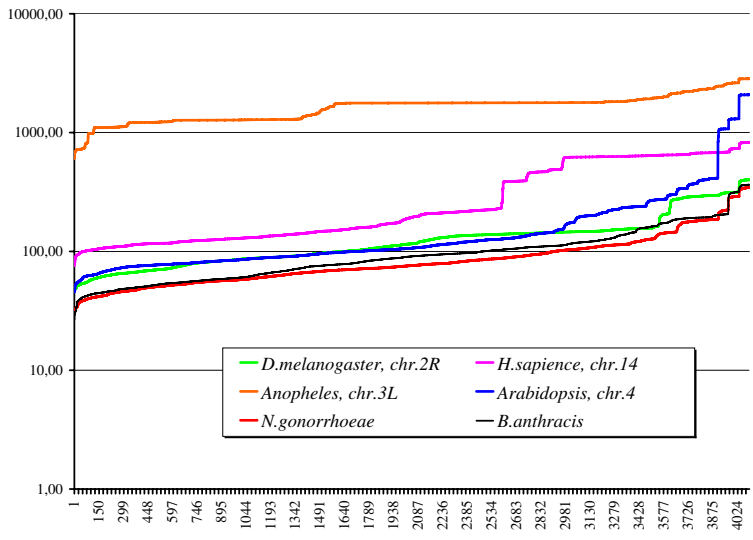

Fig. 1. Relative gap length (1) observed for five genetic sequences

Ergodicity is fundamental feature for a number of physical, chemical etc. processes. In simple words, it means that any state of a system must be reached. In such capacity, the genetic sequences meet this constraint: there are no genetic sequences[‡] lacking some triplet. Definitely, the longer string is taken into consideration, the greater is the probability not to find it in a sequence. This is not a proof of non-ergodic character of a genetic sequence. The controversy is easily overcome, as one changes a sequence for an ensemble of strings of the given length, to analyze. So, there exists an ensemble of fragments of a sequence where no ergodicity takes place.

Markov process is typical example of an ergodic entity. Thus, the question arises where a Markov process relevant to a genetic sequence yields a similar pattern of the occurrence of

[‡]Unless one considers rather short sequences.

triplets, or not. The main result of the paper is that is not true, for genetic sequences. Any Markov process modelling a genetic sequence yields a probability of the longest gap as long as 100 symbols equal to $p \sim 10^{-10}$ or less. The similar figure for a gap of the length 1000 symbols becomes less than 10^{-250}. On the contrary, real genetic sequences show the probability figures for the longest gap equal to $p = N^{-1}$, where N is the length of a sequence. For any real sequence, this real figure exceeds tremendously the evaluation provided by any Markov chain. This contradictory allows to call the real genetic sequences to be weakly ergodic.

Typically, the longest gap between two nearest neighbours exists in a single copy. Nonetheless, some of them exists in two copies; strictly speaking, nothing prohibits an occurrence of these gaps in three and more copies, while we have not yet found such pattern. Eventually, a multiple occurrence of the longest gap is extremely low-probable event, in a symbol sequence of any nature.

Since the longest gaps strongly violate typical ergodicity pattern, then their mutual interlocation alongside a genetic sequence becomes of a great interest. In articular, a (local) density of the starting point of such gaps, for various couples of triplets, is rather important question. Also, the relation between the patterns observed in different genomes is important: do two (or several) genomes of taxonomically close organisms exhibit proximal pattern of the longest gaps distribution? And, if yes, what is the measure of the proximity in this case? Moreover, an intragenomic variability of the patterns is also a key question here. All these issues require further deeper study, and fall beyond the scope of this paper.

References

[1] Eu.Yu.Bushmelev, Eu.M.Mirkes, M.G.Sadovsky, On the structures revealed from symbol sequences, *Journal of Siberian Federal University. Mathematics & Physics*, **5**(2012) no. 4, 507–514 (in Russian).

[2] P.Pollack, Long gaps between deficient numbers, *Acta Arith.*, **146**(2011), no. 1, 33–42.

[3] P.Allegrini, M.Barbi, P.Grigolini, B.J.West, Dynamical model for DNA sequences, *Phys. Rev. E*, **52**(1995), 5281–5296.

[4] Yaw-Hwang Chen, Su-Long Nyeo, Chiung-Yuh Yeh, Model for the distribution of k-mers in DNA sequence, *Phys. Rev. E*, **72**(2005), 011908.

Реальные геномы обладают очень низкой эргодичностью

Михаил Г. Садовский
Ксения А. Никитина

Исследовалось распределение наидлиннейщих расстояний до ближайшего соседа во взаимном распределении триплетов (подпоследовательностей длины 3). Показано, что реальные последовательности существенно отличаются от любых модельных, построенных на основе каких-либо случайных процессов.

Ключевые слова: порядок, триплет, Марковский процесс, расстояние, взаимное расеределение.