Island Periodicity in Genomes

Michael G. Sadovsky[1] and Eugene M. Mirkes[2]

[1] Institute of computational modelling SB RAS msad@icm.krasn.ru
[2] Krasnoyarsk institute of railway engineers mirkes@bk.ru

A retrieval of various patterns and, in general, an ordering search within completely sequenced genomes is of great interest. In particular, the sequence inhomogeneity manifesting in the difference of frequency dictionaries of non-overlapped coherent triplets counted for three different starting positions indicates the presence of protein coding regions in a genome; more exactly, non-coding regions are invariant against the frame shift of the triplet pattern, while the coding ones lack these invariance [1–3]. Some recent results on the complexity and patterns observation in human genome see in [4].

A complexity of patterns observed in a genetic sequence may vary significantly. Screening a genome with respect to a complexity of different fragments, a researcher may find various biologically important peculiarities in nucleotide sequences. Here a new approach to figure out some patterns in the mutual distribution of triplets observed alongside a sequence is present.

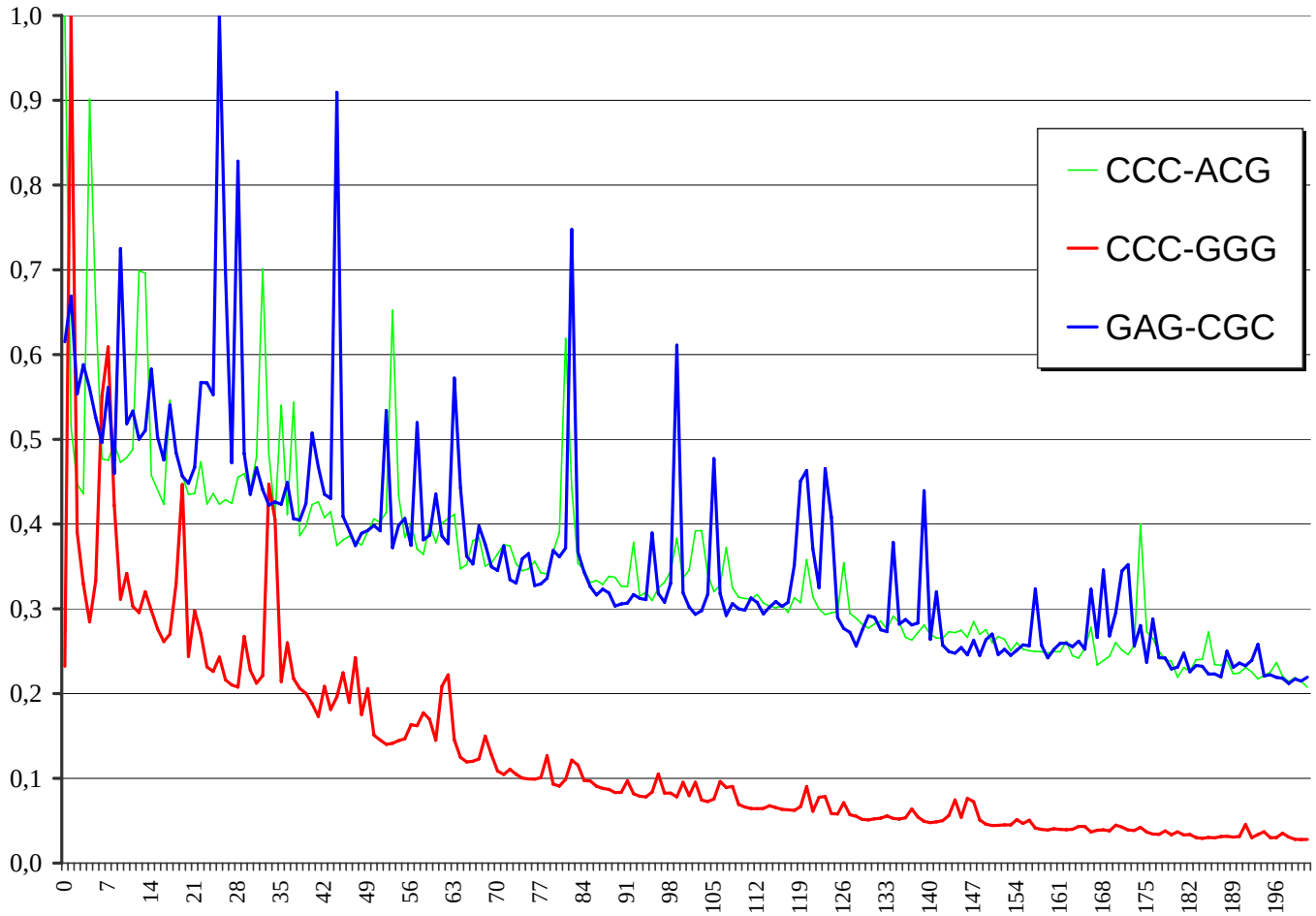

Fig. 1. Distribution function $f_{\langle\omega_1,\omega_2\rangle}(r)$ for three couples of triplets, of 11^{th} chromosome of *B. taurus*.

The approach is based on the consistent and comprehensive study of the distribution of the triplets alongside a nucleotide sequence, where the distribution is developed for the nearest neighbours. More exactly, consider a nucleotide sequence of the length N (we studied more than 300 genomes of various organisms, with total abundance of slightly less than 4000 nucleotide sequences; the results are illustrated on the sequence of *Bos taurus* chromosome 11, accession # CM000187 in EMBL–bank; $N = 110171769$ b.p.); further we shall suppose that there are no other symbols in a sequence, but $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$.

Fix, then, two triplets $\omega_1 = \nu_1\nu_2\nu_3$ and $\omega_2 = \overline{\nu_1\nu_2\nu_3}$. Find the first embedment of $\nu_1\nu_2\nu_3$; then a distance (in base number) from that former triplet to the nearest embedment of $\overline{\nu_1\nu_2\nu_3}$ is determined. Next, change for the second embedment of $\nu_1\nu_2\nu_3$ and find the other nearest occurrence of the second triplet, and so no. Finally, the distribution function

$$F_{\langle\omega_1,\omega_2\rangle}(r) = n_{\langle\omega_1,\omega_2\rangle}(r)\ , \qquad r = -2, -1, 0, 1, 2,\ \ldots \tag{1}$$

would be developed. Here n is the number of couples $\omega_1 \div \omega_2$ occurred at the distance r, alongside a sequence, so that the triplets make the closest neighbours. That latter means that there is no other triplet somewhere in between a current couple $\omega_1 \div \omega_2$. Obviously, the function (1) figures depend on the length of a sequence; so, a normalization must be implemented. It is very natural to change the function (1) for frequency function $f_{\langle\omega_1,\omega_2\rangle}(r)$, so that

$$\sum_{r\in\mathsf{R}} f_{\langle\omega_1,\omega_2\rangle}(r) = 1\,.$$

Here R is the definition area of (1). The case of $r = -2$ corresponds to the overlapping of the triplets over a dinucleotide; the case of $r = -1$ corresponds to the overlapping of the triplets over a nucleotide, $r = 0$ means that the triplets ω_1 and ω_2 neighbours each other immediately, and so on.

In spite of a simplicity of the problem, yet there are no one study in this area. Here we present the results towards the behaviour of the function $f_{\langle\omega_1,\omega_2\rangle}(r)$ illustrated with the observations on *B. taurus* eleventh chromosome. A general pattern of $f_{\langle\omega_1,\omega_2\rangle}(r)$ variation looks like an exponentially decaying curve; that is quite natural, since the strong combinatorial constraints are standing behind.

More intriguing are the deviations from the exponential decay; Fig. 1 shows the patterns of the function $f_{\langle\omega_1,\omega_2\rangle}(r)$. To improve the vision, we have renormalized the function to the maximal. Probably, the most essential is the question of the origin of such long ranged correlations in triplets distribution. The correlations themselves make the island periodicity that is an increased probability to meet the given couple (as the nearest neighbours), at the given distance, and, probably, in some peculiar regions of a genome.

To make sure whether the pattern results from the long repeats, we have checked the structure of these latter; it makes no result. Neither the Markov models of the original nucleotide sequences of the order two to eight yielded a reasonable coincidence to the observed patterns.

In brief, below are enlisted the most general properties of the function $f_{\langle\omega_1,\omega_2\rangle}(r)$.

1. Any genome exhibits the middle and long-ranged correlation in the nearest neighbouring triplets distribution;
2. Bacterial genomes, and yeast genomes show significantly lower level of the long-ranged correlations, and the pattern of the function $f_{\langle\omega_1,\omega_2\rangle}(r)$ is smoother and more regular, in comparison to those observed for higher eukaryotes.
3. The function $f_{\langle\omega_1,\omega_2\rangle}(r)$ is asymmetric: $f_{\langle\omega_1,\omega_2\rangle}(r) \neq f_{\langle\omega_2,\omega_3\rangle}(r)$. Moreover, this asymmetry is observed for various types of the correspondence of the triplets (i. e., for complementary palindromes, as well).
4. The structure revealed through the observations towards the behaviour of the function $f_{\langle\omega_1,\omega_2\rangle}(r)$ has the origin other than long (or super long) repeats; neither it might be explained due to Markovian model of a sequence of the order 2 to 8.
5. The set of 4096 couples of triplets is separated into the subsets according to the exponential decay factor figure observed for them. The pattern of this separation is specific for each species.

References

1. A.Yu.Zinovjev, A.N.Gorban, T.G.Popova: In Siliko Biology **3**, 471 (2003).
2. A.N.Gorban, A.Yu.Zinovjev, T.G.Popova: Open Systems & Information Dyn. **10**, 321 (2003).
3. A.Carbone, A.Zinovyev, F.Kepes: Bioinformatics **19**, 2005 (2003).
4. W.Li: J.Theor.Biol. **288**, 92 (2011).