Eugene Bushmelev
Siberian Federal university
e-mail: hoochie_cool@list.ru
Michael Sadovsky
Institute of computational modelling of SB RAS
e-mail: msad@icm.krasn.ru

Close order in triplet composition in genomes

We studied a two-particle distribution function $l(\omega_1, \omega_2)$ of a distance defined in the number of nucleotides between two given triplets $\omega_1 = \nu_1\nu_2\nu_3$ and $\omega_2 = \mu_1\mu_2\mu_3$. For each entry of a given triplet ω_1 the distance to the nearest given triplet ω_2 has been determined, thus revealing the distribution function $l(\omega_1, \omega_2)$ of the couples of triplets in a genetic entity. The function is defined in rather multi-dimensional space ($64^2 = 4096$) that makes the problems of its analysis and visualization rather acute.

The distribution function $l(\omega_1, \omega_2)$ was found to be rather complex; it has several maxima, and the number and location (relative distance) of those maxima are specific, for various couples of triplets. For yeast genome of *Pichia stipitis* CBS 6054, typical number of maxima was equal to three, for any chromosome. Intra-genomic variation of the shape of $l(\omega_1, \omega_2)$ is rather significant; at least, different chromosomes have indistinctively discrete types of the function.

Special attention has been paid to the couples of triplets that make so called complementary palindrome. That latter is a couple of triplets read equally in opposite directions with respect to the complimentary rule substitution, say, ATG $\leftrightarrow$ CAT of GCA $\leftrightarrow$ TGC. Such triplets (and longer strings) are well known for a kind of symmetry in genomes: the frequency of each string in a complementary palindrome is pretty close each other. Information charge of the triplets composing a complimentary palindrome is another important issue, for the analysis of the close order in genomes. This former is a ratio of real frequency $f_{\nu_1\nu_2\nu_3}$ to the mostly expected one $\widetilde{f}_{\nu_1\nu_2\nu_3}$, which is defined as

$$\widetilde{f}_{\nu_1\nu_2\nu_3} = \frac{f_{\nu_1\nu_2} \times f_{\nu_2\nu_3}}{f_{\nu_2}} .$$

Information charge $p_{\nu_1\nu_2\nu_3}$ is more sensitive to the biological peculiarities of the genetic entity under consideration.

We have examined more than 20 genomes with as many sequences, as one hundred. All the investigated genetic entities exhibit the close order of triplet composition. The pattern of the order was different for the different species (and higher taxa). Moreover, even an intra-genetic variability of the patterns was high enough to put on the problem of the comprehensive analysis of the pattern itself.

To verify the patterns observed at the real genetic entities, we have carried out several computational experiments. We have generated a surrogate random non-correlated sequence with the same frequencies of nucleotides and the same length, and developed similar patterns to figure out the deviation in the patterns observed over a real sequence from similar observed over a surrogate. Significant difference has been detected.

Some biological issues of the observed order are discussed. The work is a part of a greater project of a study of the distribution of longer strings with increased information charge alongside a genome.