STUDYING OF GENOME–SPECIFICITY OF TRIPLET PERIODICITY

Y.M. Suvorova[1], E.V. Korotkov[1,2]
[1]*Center of Bioengineering Russian Academy of Sciences*
Moscow, RUSSIAN FEDERATION
[2]*National Nuclear Investigational University (MIFI)*
Moscow, RUSSIAN FEDERATION
e-mail: suvorovay@gmail.com

Abstract

Triplet periodicity (TP) is distinguishing property of protein coding sequences of both prokaryote and eukaryote genomes which studied for decades. In the work we explored distributions of triplet periodicity difference inside and between bacterial genomes. We constructed two hypothesis of TP distribution on set of coding sequences and simulated corresponding artificial datasets. We found that triplet periodicity is more similar inside genome than between genomes and that TP distribution inside genome corresponds to hypothesis which imply common TP pattern for majority of sequences.

1 Motivation

The most-known periodicity type presented in all living species is triplet periodicity (TP) [1] of genes encoding proteins and periodicity of larger periods divided by three [2]. This feature is widely used in practice for revealing coding regions [3], for detecting mutation in the sequence like frame-shifts [4][5] and fusions [6]. TP is characterized by non-equal nucleotide distribution of different codon positions. In the work we aimed to study difference between TP of coding sequences inside and between prokaryotic genomes. Also we wanted to find out is there some genome-specificity of TP - if so one can use it to determine the genome to which a considered gene or part of the gene most likely belongs. Although the TP was classified previously [7] but in that study classes had minor difference (about 5-10%) so this classification could not detect any genome-specificity and the question was not answered.

2 Methods

Consider protein coding sequence S of lenght L from the alphabet $A = \{'A', 'C', 'T', 'G'\}$. One can presents it as as triplet periodicity matrix M (4×3 matrix with rows correspond to four symbols of DNA and columns to period 1, 2 or 3). An element of M $m(i,j)$ is equal to the number of occurence of nucleotide i ($i = 1$ for $'A'$, $i = 2$ for $'C'$, $i = 3$ for $'T'$ and $i = 4$ for $'G'$) in the position j of codon.Then each element was normalized.

$$n(i,j) = \frac{m(i,j) - Lp(i,j)}{\sqrt{Lp(i,j)(1 - p(i,j))}}$$

where

$$p(i,j) = \frac{(\sum_{i=1}^{4} m(i,j)) \times (\sum_{j=1}^{3} m(i,j))}{L^2}$$

To compared two normalized matrixes N_1 and N_2 we used a measure

$$D = \sum_{i=1}^{4}\sum_{j=1}^{3}\left(\frac{n_1(i,j) - n_2(i,j)}{\sqrt{2}}\right)^2$$

$$X = \sqrt{2D} - \sqrt{2df - 1}$$

To study the distribution of TP difference inside genomes we performed paiwise comparison of TP matrixes of the entire set of genes from one genome and constructed the distribution. As a control the same distributions were constructed for two modeling sets. We explored modeling sets of two types: the first one where all sequences TP were obtained from one TP pattern (*Perf*) and the second one where conversely TP of all sequences were independent and random (*Rand*). An illustration of such distribution for *E.coli* genome is shown in Figure 1. To investigate the distribution of TP differ-

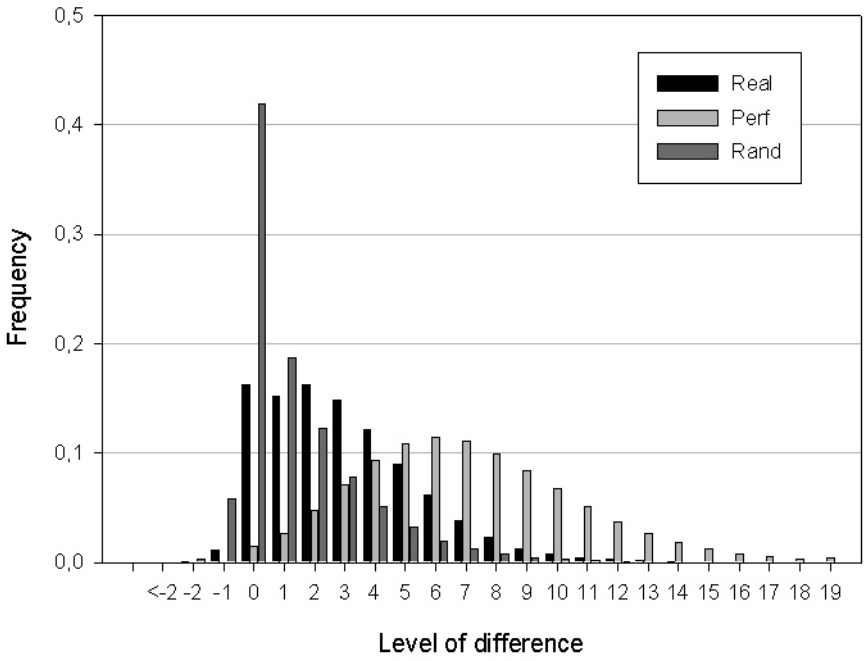

Figure 1: The distribution of TP difference (X) among total set of genes from *E.coli* genome (*Real* and *Rand*), and two artificial datasets simulated for this genome.

ence between genomes we calculated difference between TP matrixes of genes from two different genomes using the same measure of difference and constructed corresponding

distribution. As a control the same distributions were constructed for the corresponding modeling sets. As it shown in Figure 2 the difference TP of genes between *E.coli* and *B.avium* genomes is greater than between genomes. Comparing TP matrixes from two genomes one can see that *Real* distribution shifts left (towards *Rand* distribution), while the positions of distributions on the simulated sets do not changed compared with those obtained inside a genome. This implies that TP between genomes differs more than inside. Further pairwise comparisons of different genomes confirm the result.

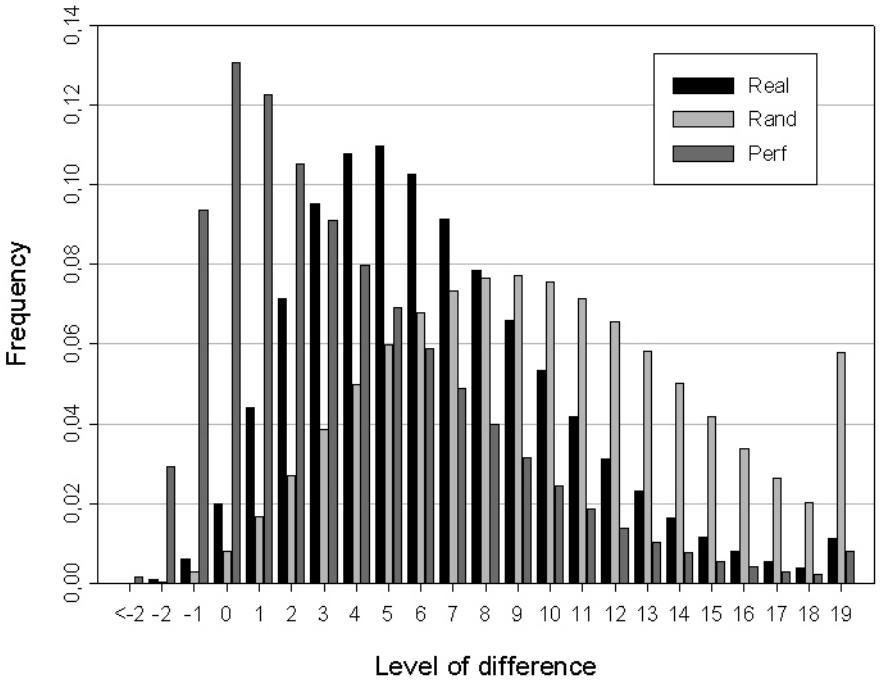

Figure 2: The distributions of TP difference (X) between *E.coli* and *B.avium* for real dataset and two simulated sets.

3 Results

We constructed distribution histograms of TP difference on the real and simulated datasets. The histograms allow to conclude that TP difference in real dataset more similar to the *Perf* dataset, which assumes that TP most of genes inside genome set is similar. We showed that (1) the distribution of TP difference inside a genome is more similar to the corresponding distribution inside *Perf* dataset (except about 10% of genes); (2) TP matrixes inside genome are closer than between genomes. Our results suggest that there is some process inside bacterial genomes that formed and maintained special TP type of genes inside one genome. Without such process it is hard to explain how TP could persist in the context of mutation process even if all genes inside genome

initially had the same TP type. In practice genome-specificity of TP could be useful for pathological genome identification in medicine and homogeneity of TP inside genome - for prediction of horizontally transferred genes.

References

[1] Trifonov E.N. (1998). 3-, 10.5-, 200- and 400-base periodicities in genome sequences.*Physica A*. Vol. **249**, pp. 511-516.

[2] Korotkov E.V, Korotkova M.A., Kudryashov N.A. (2003). Finding borders between coding and noncoding DNA regions by an entropic segmentation method. *Physics Letters A*. Vol. **312**, pp. 198-210.

[3] Bernaola-Galvan P., Grosse I., Carpena P., Oliver J.L., Roman-Roldan R., Stanley H.E. (2000). Information decomposition of symbolic sequences. *Physical Review Letters*. Vol. **85**, pp. 1342-1345.

[4] Frenkel F.E., Korotkov E.V. (2009). Using Triplet Periodicity of Nucleotide Sequences for Finding Potential Reading Frame Shifts in Genes.*DNA Research*. Vol. **16**, pp. 105-114.

[5] Rudenko V.M., Suvorova Y.M., Korotkov E.V. (2011). Detection of Possible Reading Frame Shifts in Genes Using Triplet Frequencies Homogeneity. *Austrian journal of statistics*. Vol. **40**, pp. 137-146.

[6] Suvorova Y.M., Rudenko V.M., Korotkov E.V. (2012). Detection change points of triplet periodicity of gene. *Gene*. Vol. **491**, pp. 58-64.

[7] Frenkel F.E., Korotkov E.V. (2008). Classification analysis of triplet periodicity in protein-coding regions of genes.*Gene*. Vol. **421**, pp. 52-60.