Bovine leukemia virus sequences and rate of nucleotide exchange

Sotnikova E.A., Kosovsky G.Yu., Levitskaya O., Glazko V.I.

Centre of experimental embryology and reproductive biotechnology Russian Academy of Agricultural Sciences

e-mail: sotnikova.evgeniya@gmail.com

For a comparative analysis of the prevalence of nucleotide substitutions in different genes of the proviral DNA of bovine leukemia virus in our studies performed comparing the frequency nucleotide substitutions and nonsynonymous to synonymous substitutions ratio (Kn/s) in the coding sequences of genes *env* and *pol*. In the *env* gene frequency of nucleotide substitutions calculated separately in encoding *env* sequences, as encoding glycoprotein gp51, involved in the reception of viral particles to target cells, and the glycoprotein gp30, witch have a transmembrane domain. The obtained data is following. The frequency of nucleotide substitutions per nucleotide in sequence encoding gp51, was 0.1766 for Kn/s 0.5955, encoding gp30 - 0,1636 in Kn/s - 0.3636 (based on 52 sequences). The data indicate that the frequency of nonsynonymous substitutions in gp51 sequence slightly higher compared to gp30, despite the general similarity of average frequencies of nucleotide substitutions per nucleotide. This suggests the existence of a more intense effects of “pure” selection on the sequence of gp30 compared with gp51, which appears to be due of gp30 belonging to transmembrane proteins.

In areas of the *pol* gene with coordinates 2325-2818 (29 sequences) 3986-4218 (35 sequences) 4499-4631 (33 sequences) the frequency of nucleotide substitutions on the one nucleotide were similar (0.1417; 0.1552; 0.1654), with the exception of a fragment with coordinates 4257-4393 (38 sequences), where the average rate of nucleotide substitutions reached high values - 0.2774. The smallest value of Kn/s found in the area 3986-4218 (0.2414). Similar to those observed in the gene gp30 - the areas of the *pol* gene with coordinates 2325-2818 and 4499-4631 (Kn/s 0.4000 and 0.4667, respectively). The highest values of Kn/s (0.9000) identified in the *pol* region with coordinates 4257-4393, also there was the highest average frequency of nucleotide substitutions per nucleotide. Similar values of the frequency nonsynonymous and synonymous substitutions in this site may indicate the absence of intense effect of pure selection, and, therefore, increased the likelihood of erroneous results when it used to diagnose the integration of proviral DNA into the animals genome.

In order to evaluate the possible contextual features of the distribution of nucleotide substitutions in the genes examined, analyzed the incidence of non-canonical DNA structures such as quadruplex G4 was made. Localization of G4 was assessed with the using the program QGRS (Quadruplex forming G-Rich Sequences) Mapper. 4 non-overlapping potential G4 sequences were revealed in *env*, most of which are localized in three segments of the gene. The highest density of G4 found in the gp51 coding sequence and terminal region of the gp30 coding sequence. In gp51 coding sequence gene observed prevalence of nonsynonymous substitutions relative to synonymous (Kn/s = 1.3333). 10 non-overlapping sequences with potentially susceptible to formation of G4 quadruplex found in *pol* gene and 49 overlapping, that is substantially less than the density of their distribution in comparison with the *env* gene. There is detected higher frequency of nucleotide substitutions per nucleotide (0.6667) with relatively high values of Kn/s (0.7143) In the *pol* gene region coordinates 4256-42780, which is localized in two sequences potentially susceptible to the formation of G4 quadruplex. The data indicate significant differences on relations nonsynonymous and synonymous nucleotide substitutions in the areas of gene *env*, *pol*, presumably reflecting the different impact of "pure" selection. There were no definitive links between the higher density of nucleotide sequences prone to forming G4 quadruplex, and high frequency of nucleotide substitutions. Using the program «Giri» search for areas of homology in dispersed repeats was made. In the *pol* gene (unlike *env*) revealed regions of homology with the *pol* segment coordinates 2340 - 2617 (278 bp) coding for and a transmitting retrotransposon sequence inner portion primates (ERV-2Cla-1) in a segment of coordinates 3824 - 3888 (63 bp), and transmits to the coding sequence LTR retrotransposon Danio rerio (DIRS-14 DR); plot coordinates 3964 - 4287 (324 bp) and translates to a coding sequence of the guinea pig endogenous retrovirus (ERV2-4CPo - 1). (63-75%).

The received data indicate marked differences in the frequencies of nucleotide substitutions, the ratio Kn/s, which must be considered in the search for sites BLV proviral genome in order to develop the most versatile test system for provirus genome identification in cattle genome.