PAIR CHANGE POINTS OF THE TRIPLET PERIODICITY OF GENES

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

Abstract

The triplet periodicity (TP) is a distinguished property of protein coding sequences. There are complex genes with insertions of another TP type. We say that these genes contain a pair change points of the triplet periodicity (PCP). The aim of the work is to study such genes and try to understand the nature of the PCP phenomenon. We introduced a new mathematical measure of similarity between triplet matrixes and developed a mathematical method to identify PCP in a sequence. We identified 2700 genes with PCP among 66936 genes from 17 bacterial genomes. We developed a mathematical approach to visualize the presence of PCP in the genes and PCPs are easily distinguishable. At the same time, we found 6459 genes with a single change point, which is consistent with our previously obtained results.

1 Motivation

The process of gene evolution and formation of new genes has been discussing for many years. Currently there exists a theory claiming that in some phase of evolution nature began to rearrange the existed elements to increase the complexity instead of creating genes de novo. Shuffling of existing DNA coding blocks, so-called domains, provided proteins with new architecture, activity and features. This diversity of proteins could arise as a result of domain recombination. The process of forming new proteins by means of elements shuffling may take place both at mRNA and DNA level. At the level of DNA strand there exist such mechanisms as unequal crossing over, DNA-strand breakage and repair, transposion, which potentially can lead to different rearrangements of DNA subunits. On the other hand, triplet periodicity (TP) of DNA coding sequences is a common property of all known living organisms and also it is associated with a gene reading frame (RF). Classification analysis of TP of the genes from the KEGG database previously showed that most of them belonged to relatively small set of TP classes (about 2500 classes) [1]. These classes may differ greatly. This means that if gene has insertion of DNA fragment with different types of TP, then this event can be relatively easy to detect. The DNA insertion in the genes can be revealed as pair change point (PCP) of triplet periodicity of the gene [2]. Change points could be very important for creation of artificial proteins by the way of gluing of the protein parts.

2 Methods

The main subject of the present study is gene sequences that contain PCP of the triplet periodicity which could be the insertions DNA sequences with one type of triplet periodicity to the gene sequence with different TP type. PCP could show us the events where part of gene with other biological function is inserted in gene. To find PCP we introduce new mathematical measure of the similarity between TP matrixes in adjacent regions of the analyzed gene [1]. We have developed new mathematical method to identify the single change points and the PCP along a gene sequence. The Monte-Carlo calculations were used for finding of the statistical importance of the PCPs. We tested developed method with help of the sequences with the artificial PCPs. This test shows that the developed mathematical approach allows to detect of the pair change points of the gene triplet periodicity. The efficiently of the PCP registration was more that 40%. For illustration of PCP we created the contour plots. Contour plot shows the presence of PCP in a gene sequence.

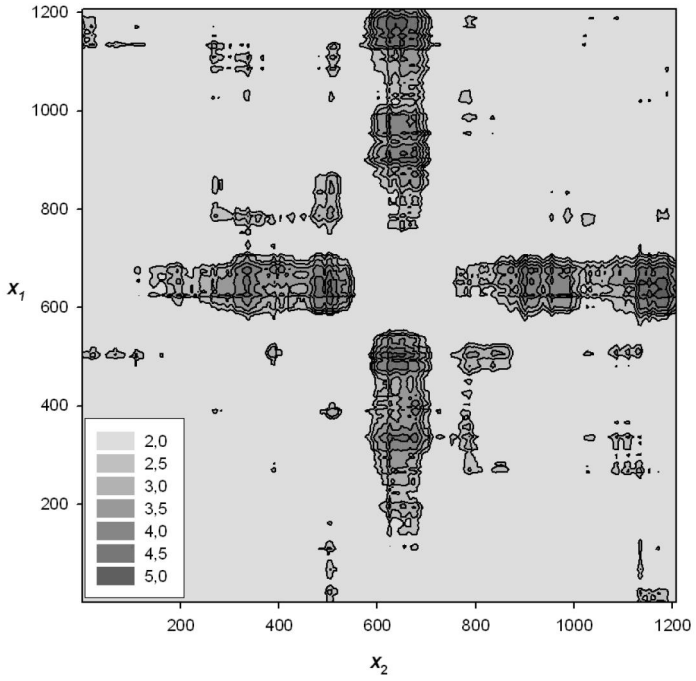

Figure 1: The contour plot for gene coding the glycerol-3-phosphate permease from *B.subtilis* genome. Gene ID of the KEGG data base is BSU02140. It is possible to see the pair change points in the position 600 and 700 nt

3 Results

Firstly we applied developed method for analyze of the genes from 17 bacterial genomes, total number of genes is 66936. It was found 6459 genes that contain the single change points which is about 10% from total number of the studied genes. These data are consistent with our earlier results [3]. The overlap in these data is approximately 50%. In this study we used a measure of the similarity of the triplet matrixes and more adequate results for the search of the single change points were received [4]. The number of paired change points is 2700 that is 4% form analyzed genes. Examples of paired change points are shown in Fig.1. This figure shows the counter plot for gene coding the glycerol-3-phosphate permease (BSU02140 in Kegg) from *B.subtilis* genome is shown. It is possible to see the pair change points in the position 600 and 700 b.p. It is possible to assume that genes with PCP actually were formed by insertion of DNA fragment from some gene in sequences of other gene. The triplet periodicity of inserted DNA fragment is different from triplet periodicity of the gene sequences and it permits to us to detect these PCP.

References

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

[2] Korotkova M.A., Kudryashov N.A., Korotkov E.V. (2011). An approach for searching insertions in bacterial genes leading to the phase shift of triplet periodicity. *Genomics, Proteomics &Bioinformatics*. Vol. **9**, pp. 158-170.

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

[4] Pugacheva V.M., Korotkov A.E., Korotkov E.V. (2012). Searching for pair points of triplet periodicity phase shifts in the genes of 17 bacterial genomes. *Mathematical Biology and Bioinformatics*. Vol. **7**, pp. 461-475