

Potential bacteriophage recombination sites inside genes containing cysteine repeats

Sorokina N.¹, Daugavet M.²

¹ Applied Genomics Laboratory, SCAMT Institute, ITMO University, St. Petersburg, Russia

² Laboratory of Noncoding DNA, Institute of Cytology, RAS, St. Petersburg, Russia

* sorokina@scamt-itmo.ru

Key words: cysteine repeats, genes, horizontal transfer

Motivation and Aim: Prokaryotic-specific domains have been found in eukaryotic organisms frequently. This event is often associated with horizontal gene transfer and some studies have considered viruses as putative vectors for gene transfer. A previous study identified specific eukaryotic cysteine repeats adjacent to bacteriophage recombination sites (Atp) in the ascidian genome [1, 2]. A pattern of cysteine repeats was retrieved and used to create an experimental set of proteins to search for bacteriophage recombination sites in various eukaryotic genomes.

Methods and Algorithms: The work was done based on Python algorithms developed. An experimental set of proteins was created using the cysteine pattern $C*\{4,9\}C*\{4,9\}C*\{4,9\}C*\{4,9\}CC$ for the study. All proteins matching the pattern in the Uniprot Trembl database were selected. Also, control dataset was generated: for each species from experimental dataset proteins without specific cysteine repeats were extracted. The second experimental set - sequences of bacteriophage recombination sites were collected earlier from the NCBI database. The resulting proteins were back-translated into nucleotide sequences using the Entrez database. Recombination sites in the nucleotide sequences of proteins with cysteine repeats without mismatches and indels were searched. Atp was also searched in the control dataset.

Results: Following a cysteine pattern search, 47,622 proteins were selected from 219,174,960 Uniprot Trembl proteins at the time the database was uploaded. Out of these, for 3,010 proteins nucleotide sequences of the genes were available. The experimental set of bacteriophage Atp recombination sites amounted to 186 sequences ranging from 10 to 491 nucleotides. By searching for Atp in the genes of proteins with cysteine repeats without mismatches and indels the following results were obtained. In 1,294 sequences of 3,010 (43 %) proteins 2,700 recombination sites were found, and in some protein genes several Atp were found. The length of the atp detected was between 10 and 16 nucleotides. Genes of proteins with cysteine repeats were obtained for 122 organisms. The control dataset consisted of one protein without a cysteine repeat for each species. Out of 122 genes of proteins, in 17 sequences (13,9 %) we found 28 short atp of length from 10 to 13 nucleotides.

Conclusion: Consequently, potential recombination sites of bacteriophages within genes containing cysteine repeats were found. Compared with the control group, it can be concluded that atp was more frequently found in the dataset containing cysteine repeats. Further, it is necessary to search for bacterial domains in eukaryotic genes containing recombination sites. It is reasonable to continue investigating the hypothesis of horizontal transfer associated with bacteriophages.

Acknowledgements: This work was supported by the Russian Science Foundation (grant No. 19–74-20102) and Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075-15-2021-1075, signed 28.09.2021).

References

1. Daugavet M.A., Shabelnikov S.V., Podgornaya O.I. Amino acid sequence associated with bacteriophage recombination site helps to reveal genes potentially acquired through horizontal gene transfer. *BMC Bioinformatics*. 2020;21(Suppl 12):305.
2. Daugavet M.A., Shabelnikov S., Shumeev A., Shaposhnikova T., Adonin L.S., Podgornaya O. Features of a novel protein, rusticalin, from the ascidian *Styela rustica* reveal ancestral horizontal gene transfer event. *Mobile DNA*. 2019;10:4.