doi 10.18699/bgrs2024-8.4-20

Bioinformatics analysis of the biophysical determinants of CRISPR array adaptation mechanism

Tumbas M.*, Đorđević M.
Quantitative Biology Group, Faculty of Biology, University of Belgrade, Belgrade, Serbia
** marko.tumbas@bio.bg.ac.rs*

Key words: CRISPR-cas; melting energy; extreme value distribution

Motivation and Aim: CRISPR-cas systems are incredibly diverse and currently are classified into six major types and over 30 subtypes [1]. Apart from their role in adaptive immunity, some of the CRISPR-cas subtypes are also involved in host gene regulation and collateral damage, leading to bacteriostatic or lethal outcomes for the host. CRISPR array spacers direct and influence canonical and non-canonical functions of the CRISPR-Cas system together with subtype Cas proteins. A better understanding of spacer adaptation mechanisms is crucial for uncovering the intricacies of an evolutionary arms race between prokaryotes and phages.
Methods and Algorithms: Bacterial and viral genomes were retrieved using NCBI Datasets API [2]. CRISPRidentify and CRISPRcasIdentifier tools were used for CRISPR array, Cas genes detection, and subtyping [3–4]. Viral genomes were mapped to their hosts using the latest version of the Virus-Host DB [5]. Mapping was performed on the genus level of the hosts' phylogenetic trees. Gumbel extreme value distribution was used to determine the statistical significance of each spacer Smith-Waterman alignment score. Mobile genetic elements (MGE) and prophage predictions were obtained for all complete bacterial genomes using the VRprofile2 tool [6].
Results: We present a large-scale analysis of CRISPR array spacers from 31,845 complete bacterial genomes. Differences in melting energy and GC content between identified spacers, origin bacterial genomes, and infecting bacteriophages were explored for different CRISPR-cas subtypes and bacterial genera. Significant GC content differences can be observed between bacterial genomes and their spacers and between spacers and infecting phage genomes. Spacers from the extremes of the GC content distribution were aligned to bacterial and infecting phage genomes to determine their origin. We obtain that high GC content spacers have better alignments to phage genomes than low GC spacers and that high and low GC spacers show better alignments to origin bacterial genomes than infecting phage genomes. We further tested if spacer alignment locations to origin bacterial genomes correspond to predicted MGE and Prophage locations and if alignment scores between those spacer groups differ. We found that low GC spacers do not preferentially target integrated prophage or MGE locations over the rest of the bacterial genomes, while high GC spacers tend to target integrated prophage sequences over MGE and remaining bacterial genome sequences.
Conclusion: The GC content of the spacers was smaller than the GC content of the source bacterial genome but larger than the infecting viral genome. This observation aligns with the hypothesis that the majority of CRISPR spacers were adapted from the bacteriophage genomes and serve a canonical function. Alignments of the spacers from GC-rich distribution tails have shown their preferential targeting of host genomes. This further supports the hypothesis that GC-rich spacers originated from the bacterial genome and

have a non-canonical function. The alignment of GC-rich spacers to predicted prophage genomes confirms their viral origin and further supports the hypothesis of GC-rich spacers' preferential adaptation to CRISPR arrays.

Funding: The study is supported by the q-bioBDS Science Fund of the Republic of Serbia (No. 7750294).

References

1. Makarova K.S. et al. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol*. 2020;18:67-83
2. Sayers E.W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2022;50:D20-D26
3. Mitrofanov A. et al. CRISPRidentify: identification of CRISPR arrays using machine learning approach. *Nucleic Acids Res*. 2021;49:e20
4. Padilha V.A. et al. CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems. *GigaScience*. 2020;9:giaa062
5. Virus-Host D.B., Virus-Host D.B. Website, 2020. https://www.genome.jp/virushostdb
6. Wang M. et al. VRprofile2: detection of antibiotic resistance-associated mobilome in bacterial pathogens. *Nucleic Acids Res*. 2022;50:W768-W73