Journal of Bioinformatics and Computational Biology

PROKARYOTE AUTOIMMUNITY IN THE CONTEXT OF SELF-TARGETING BY CRISPR-CAS SYSTEMS

TATIANA LENSKAIA

Department of Computer Science and Engineering, University of Minnesota, 4-192 Keller Hall, 200 Union Street SE, Minneapolis, Minnesota 55455, USA
lensk010@umn.edu

DANIEL BOLEY

Department of Computer Science and Engineering, University of Minnesota, 4-192 Keller Hall, 200 Union Street SE, Minneapolis, Minnesota 55455, USA
boley@umn.edu

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Prokaryote adaptive immunity (CRISPR-Cas systems) can be first a threat to its carriers. We analyze risks of autoimmune reactions in prokaryotes by computational methods. We found important differences between Bacteria and Archaea with respect to manifestations of autoimmunity. According to the results of our analysis, CRISPR-Cas systems in Bacteria are more prone to self-targeting even though they possess several times less spaces per organism on average than Archaea. The results of our study provide opportunities to use self-targeting in prokaryote for biological and medical applications, e.g., for treatment of bacterial infections.

Keywords: genome dictionary; computational methods; spacer memory.

1. Introduction

Adaptive immunity was first demonstrated in prokaryotes in 2007. Many important findings led to this discovery and helped put the puzzle of this enigmatic mechanism together[1]. In 2007, Barrangou et al. found experimental evidence[2] of the hypothesized function of the segments consisting of repetitive structures (spacers-repeats) and associated genes previously found in prokaryote genomes. Later practical protocols using these systems for precise genome editing attracted great attention[3]. However, many questions regarding the fundamental mechanism of adaptive immunity still remain open.

The most poorly understood part of this immunity mechanism is spacer acquisition[4]. Criteria that bacteria use for spacer selection are under investigation[5]. Researchers suggest that molecular mechanisms can play a role in determining the size of spacers at least for some bacterial species[6]. However, the question of "wise" spacer selection

remains open given that bacteria utilize very rapid and extensive exchange of genetic materials[7]. All these findings raise a question of how self-targeting occurs in prokaryotes.

Stern et al. carried out the first systematic search[8] of self-targeting spacers in 2010 using the information from CRISPRdb[9] about CRISPR structures found in the sequenced genomes available at that time. The authors found over a hundred of self-targeting spacers in 330 organisms (0.4% of 23550 spacers in total). After previous sketchy reports about the observed self-targeting events, that study provided the first systematic estimate of self-targeting rate in prokaryotes: "59 of 330 (18%) CRISPR-encoding organisms possess at least one array with at least one self-targeting spacer". Stern et al. also explored the hypothesis about a suggested role of self-targeting spacers in gene regulation and rejected it. Their conclusion was that self-targeting is a form of autoimmunity with a negative fitness cost. They also outlined possible ways to escape autoimmunity for prokaryotes including inactivation of self-targeting spacer, inactivation of CRISPR-Cas system, mutation of self-protospacer.

Subsequent researchers demonstrated that self-targeting spacers can be a marker of the presence of CRISPR-Cas inhibitors[10,11]. These inhibitors mostly encoded by phages represent anti-CRISPR mechanisms that can help phages overcome CRISPR systems. These inhibitors may be used to control artificial CRISPR-Cas systems in the process of genome editing in eukaryotic cells.

These observations indicate that self-targeting events deserve further exploration. We use newly developed dictionary-based methods to facilitate this analysis. First, we repeat the initial analysis made by Stern et al. in 2010 to benchmark our methods. Second, we apply the same analysis to the current data available in CRISPRdb (3261 prokaryotes, 167,583 spacers). Our analysis aims to answer the following questions: (1) Are Archaea more prone to self-targeting compared to Bacteria? (2) Is there a difference in spacer length with respect to self-targeting between Bacteria and Archaea; (3) Are self-targeting spacers more often located on plasmids than on chromosomes? The answers to these questions help us to better understand self-targeting mechanism in the context of our current knowledge about CRISPR-Cas systems. In turn, it will provide opportunities to utilize self-targeting for biological and medical applications.

2. Results

We repeated the analysis of self-targeting events carried out by Stern et al. (2010) using our newly developed dictionary methods. Comparison of the results of our analysis and the results of Stern et al. is shown in Table 1. In 87.07% of cases (101 of 116 spacers), the number and localization of additional copies of spacers coincided. However, only in 42% (49 cases) the polarity (sense/antisense) of the protospacer found by us and reported by Stern et al. agreed. In 24% (28 spacers) the position we found matched the position(s) reported by Stern, but with the opposite polarity. In 20.69% (24 spacers) the information about the polarity was missing in the report of Stern et al., but our methods were able to identify the polarity for these self-targeting events. In the remaining 12.93% of cases (15 spacers) we found mismatches: 5 spacers had the mismatched position(s) due to the

sequence updates, 8 spacers had copies only within the identified CRISPR arrays, and the information about the rest 2 spacers was missing in the current version of CRISPRdb. The results of this comparative analysis are to validate the dictionary method as a viable approach to identify self-targeting spacers: both location and polarity.

Table 1. The correspondence between Stern et al. results and the results of our analysis.

Category	Number of spacers	%
1. Exactly matched (position(s) and directionality)	49	42.24%
2. Matched position(s) except for directionality	28	24.14%
3. Matched position(s), directionality was undetermined by Stern et al.	24	20.69%
Total agreement on position(s) of self-targeting events	***101***	***87.07%***
4. Mismatched position(s) of found self-targeting events	5	4.31%
5. No self-targeting events within the analyzed sequence	8	6.90%
6. No information about CRISPR structure in CRISPRdb	2	1.72%
Total disagreements	***15***	***12.93%***

Having validated our dictionary approach, we applied it to all the spacers currently stored in CRISPRdb. We found 2488 self-targeting spacers, approximately 1.5% of all 167,581 spacers. We analyzed 3261 CRISPR-encoding prokaryotes of which 957 (29.35%) have self-targeting spacers (Table 2).

Table 2. The number of organisms with self-targeting spacers in Bacteria and Archaea.

Organisms	Bacteria	Archaea
Self-targeting	892	65
No self-targeting	2166	138

2.1. *The comparison of self-targeting spacer rates in Bacteria and Archaea*

There is a significant difference between Bacteria and Archaea with respect to the rate of self-targeting spacers (Table 3, Chi squared test, $p < 2.2e\text{-}16$). The rate of self-targeting spacers in Archaea is (0.59%) is lower than the rate of self-targeting spacers in Bacteria (1.66%).

Table 3. The number of self-targeting spacers in Bacteria and Archaea.

Spacers	Bacteria	Archaea
Self-targeting	2325	163
No self-targeting	137514	27581

The comparison of the distributions of the number of spacers per organism in Archaea and Bacteria demonstrates that these distributions are quite different (Fig.1A). For Bacteria, most organisms have less than 50 spacers with the median of 28 spacers; for Archaea, most organisms have 100-150 spacers with the median of 116 spacers (Fig.1B).

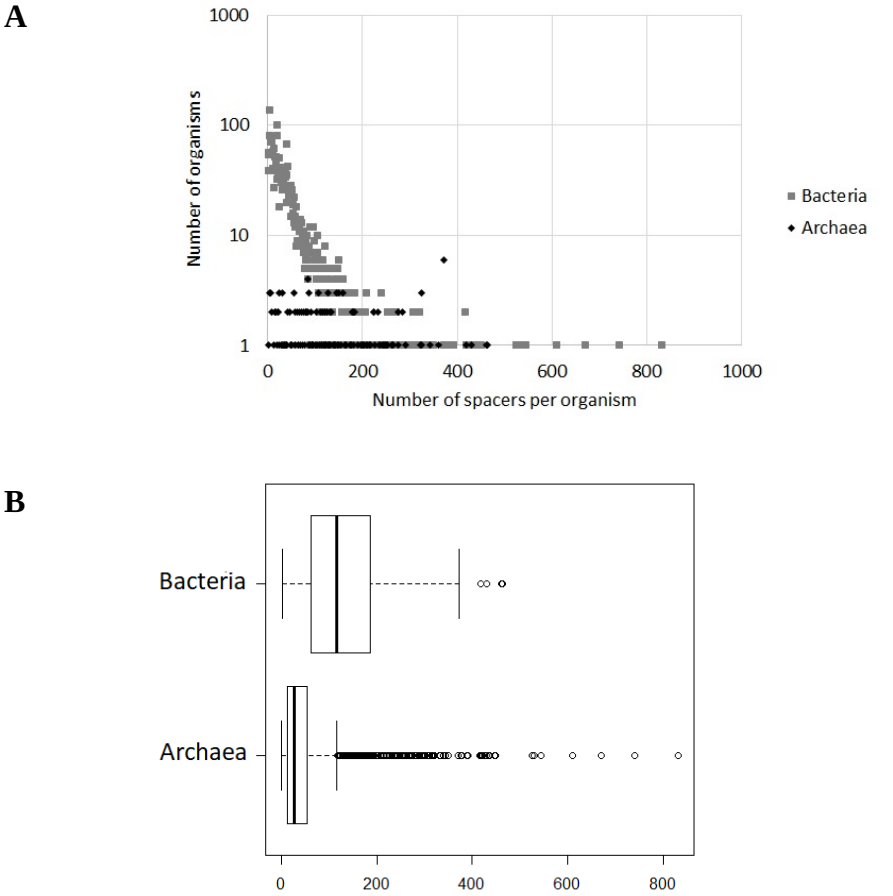

Fig. 1. The number of spacers per organism for Archaea and Bacteria: (A) the distributions are plotted using the semi-log scale, (B) the boxplots for both distributions.

2.2. *The spread of self-targeting spacers in Archaea*

We found that 163 self-targeting spacers were spread across 65 of 203 (32.02%) archaeal organisms (Figure 2). The organisms contained from 1 up to 20 self-targeting spacers. More than half of the organisms with self-targeting spacers, 34 of 65 organisms (52.3%) had only one self-targeting spacer. Another 16 organisms (24.62%) had exactly 2 self-targeting spacers. The remaining 15 organisms (23.08%) had from 3 to 20 self-targeting

spacers. The number organisms with exactly 2 self-targeting spacers was almost the same as those with 3 or more such spacers. Separately, only 5 of 163 (3%) self-targeting spacers were found on plasmids, and the remaining spacers being located on chromosomes.

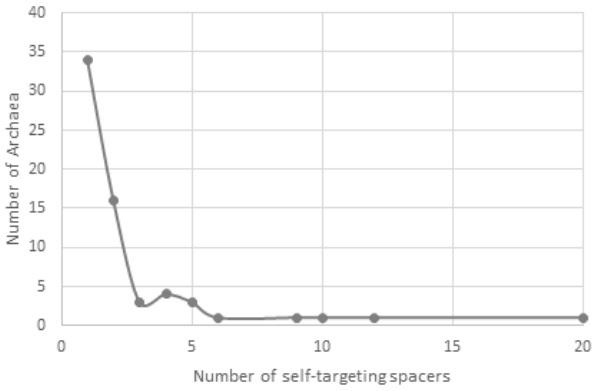

Fig.2. The distribution of self-targeting spacers in Archaea

2.3. *The spread of self-targeting spacers in Bacteria*

We found that 2325 self-targeting spacers were spread across 892 of 3058 (29.17%) bacterial organisms (Figure 3). More than a half of the organisms with self-targeting spacers, 473 of 892 organisms (53.03%) had only one self-targeting spacer, and 167 organisms (18.72%) had exactly 2 self-targeting spacers. The remaining 252 organisms (28.25%) had from 3 to 47 spacers. Separately, we noted that only 32 of 2325 (1%) self-targeting spacers were located on plasmids, all the remaining spacers were found on chromosomes.

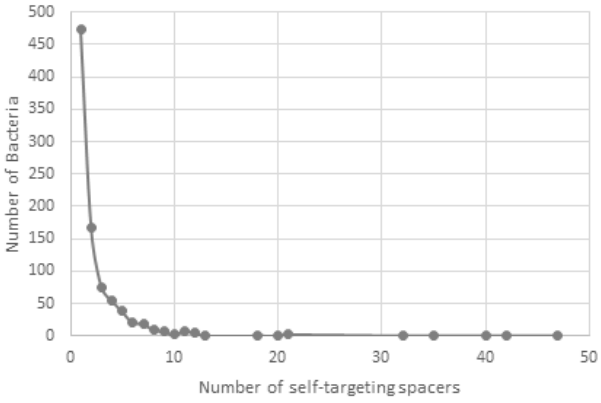

Fig.3. The distribution of self-targeting spacers in Bacteria

2.4. *The average length of spacers in CRISPR arrays of Bacteria and Archaea*

We found that Archaeal spacers are longer on average than Bacterial spacers (Fig.4 A and B). The difference between the means evaluated using a two-sample two-sided t-test with unequal variance is statistically significant ($p < 2.2e-16$); the 95% confidence interval for the difference between the means is (3.54,3.64). Interestingly, self-targeting spacers in Archaea tend to be shorter than archaeal spacers overall and spacers without self-targeting (Table 4.). Self-targeting spacers in Bacteria is about the same length as bacterial spacers overall and spacers without self-targeting. However, the standard deviation is high compared to the difference in the means, so the difference is likely not statistically significant.

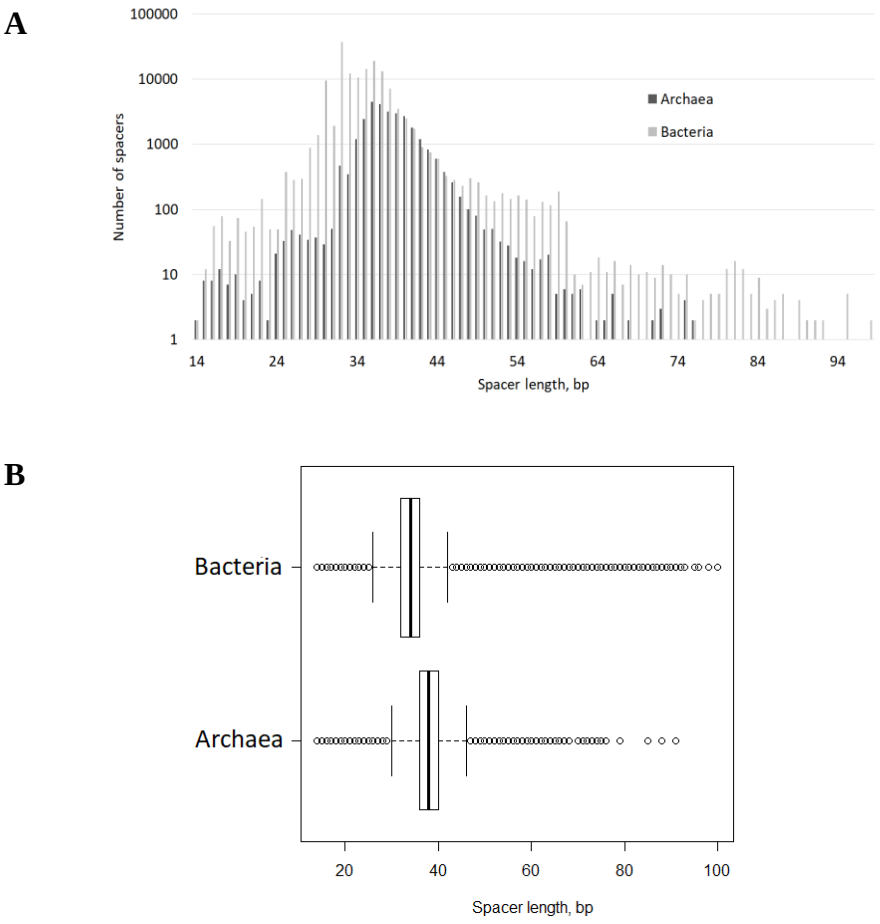

Fig.4. The distribution of spacer lengths in Bacteria and Archaea (spacers < 100 bp are shown): (A) The number of spacers for each spacer length plotted on the semi-log scale; (B) boxplots for Bacteria and Archaea.

Spacer length (mean ± sd)	Archaea	Bacteria
All spacers	38.22 ± 3.96	34.66 ± 5.22
Self-targeting spacers	33.83 ± 7.28	35.33 ± 8.57
Spacers without self-targeting	38.24 ± 3.92	34.64 ± 5.15

Table 4. The average length of spacers in CRISPR arrays of Bacteria and Archaea.

3. Methods

Information on the found CRISPR structures was obtained from CRISPRdb (the latest update, May 9, 2017). We downloaded the xml file for all analyzed prokaryotes that have at least one confirmed CRISPR array. Based on these data, we have compiled a list of organisms in the genomes of which CRISPR structures were detected. We downloaded the genomes of these organisms from the NCBI Nucleotide database. If genomes of organisms contained several replicons (i.e., chromosomes and plasmids), then each replicon was analyzed separately, and then the results were summarized at the organism level.

We extracted information about 330 organisms analyzed by Stern et al.[8]. As a reference, we used the list of the analyzed CRISPR arrays and the list of found self-targeting spacers provided by Stern et al. in the supplementary materials. We re-analyzed these data, using the most current information available at CRISPRdb. Stern et al. searched for self-targeting spacers using BLAST alignment[12] with a high similarity threshold to find 100% identity matches. However, many self-targeting events found by Stern et al. and included in the list of self-targeting spacers did not contain information about polarity. Moreover, BLAST utilizes a heuristic approach; it does not guarantee the search for all possible solutions. Instead of using an alignment-based approach, we use an "exact matching" approach inspired by the CRISPR mechanism itself. To search for exact matches and to accurately determine the polarity, we utilized our dictionary methods.

This approach is made efficient by using a dictionary (hash table) data structure. To search for self-targeting spacers in the genome of prokaryotes, we took information about all found CRISPR structures. We grouped all the found spacers by length. For each of their possible lengths, we compiled a dictionary with the unique strings of a given length as the keys and the lists of positions of these strings in the genome as the values. Then, we searched the dictionary for all spacers of that given length. To find copies on the forward and reverse strands, we searched the dictionary for the spacer (copies of the spacer on the direct strand) and its reverse complement (copies of the spacer on the reverse strand). As a result, for each spacer, we recorded into the output file its content, length, its position in the sequence and position(s) of the found copies of this spacer on the forward and reverse stands in the sequence. This helped us accurately identify the localization and polarity for self-targeting spacers. Then we compared all the found self-targeting spacers to those reported by Stern et al. Next, we conducted a similar analysis on all the data currently available at CRISPRdb.

4. Discussion and conclusion

The autoimmunity problems are a factor of evolutionary pressure on prokaryotes that possess CRISPR systems. Our findings demonstrate that about a third of prokaryotes carry self-targeting spacers even though the fraction of self-targeting spacers in a pool of all spacers is rather small (~ 1.5%).

We found a significant difference in self-targeting rates between Bacteria and Archaea ($p < 2.2e\text{-}16$). Although Archaea on average possesses several times more spacers in their genome on average than Bacteria, the rate of self-targeting spacers in Archaea is almost three times lower than in Bacteria. This suggests that Archaea have developed more robust mechanisms of CRISPR systems and can manage larger spacer memory. Consequently, Archaea may accumulate more spacers and have a lower turn-over of spacers than Bacteria.

We also found that Archaea tend to have slightly longer spacer on average than Bacteria. It means archaeal spacers are more specific in capturing potential invaders. The longer spacer can also explain the decrease in the number of self-targeting events since higher spacer specificity protects better from spurious matches.

In addition, we found that self-targeting spacers in Archaea have shorter length in comparison to the average length of spacers overall. Thus, self-targeting events might be driven by taking spacer with not enough specificity. However, for Bacteria, the problem of self-targeting might have a different origin since they have about the same average length for self-targeting and other spacers. Considering very intensive genomic exchange in Bacteria, the increased specificity might not be helpful because of extensive fragments shared between phages and Bacteria. In this case, self-targeting is an embedded cost of genome flexibility. Also, we found that only 1-3% of self-targeting spacers in prokaryotes are present on plasmids. The fitness cost of plasmids that bear self-targeting spacers is usually less that the cost of self-targeting spacers on chromosomes, and such plasmids are often eliminated from genomes.

Future studies may explore two possible directions: (a) how the pressure of autoimmunity shapes the evolution of bacteria and (b) how we can use autoimmunity manifestations for treatment of bacterial infections. The induction of autoimmunity during the operation of CRISPR-Cas systems represents a potential opportunity for the selective destruction of pathogenic microorganisms. Our finding that CRISPR systems in Bacteria are more prone to autoimmunity may provide important opportunities to develop new treatment methods that are alternative to antibiotics.

Acknowledgments

The work of the first author was supported by the University of Minnesota Interdisciplinary Doctoral Fellowship. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper. URL: http://www.msi.umn.edu

References

1. Ishino, Y., Krupovic, M., & Forterre, P. (2018). History of CRISPR-Cas from encounter with a mysterious repeated sequence to genome editing technology. *Journal of bacteriology*, 200(7), e00580-17, 2018.
2. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315:1709–1712, 2007. https://doi.org/10.1126/science.1138140.
3. Pickar-Oliver, A., & Gersbach, C. A. The next generation of CRISPR–Cas technologies and applications. *Nature Reviews Molecular Cell Biology*, 20(8), 490-507, 2019.
4. McGinn, J., & Marraffini, L. A. Molecular mechanisms of CRISPR–Cas spacer acquisition. *Nature Reviews Microbiology*, 17(1), 7-12, 2019.
5. Nasko, D. J., Ferrell, B. D., Moore, R. M., Bhavsar, J. D., Polson, S. W., & Wommack, K. E. CRISPR spacers indicate preferential matching of specific virioplankton genes. *mBio*, 10(2), e02651-18, 2019.
6. Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N., & Doudna, J. A. Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature*, 527(7579), 535, 2015.
7. Dutta, C., & Pan, A. Horizontal gene transfer and bacterial diversity. *Journal of biosciences*, 27(1), 27-33, 2002.
8. Stern, A., Keren, L., Wurtzel, O., Amitai, G., & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends in genetics*, 26(8), 335-340, 2010.
9. Grissa, I., Vergnaud, G., & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC bioinformatics*, 8(1), 172, 2007.
10. Rauch, B. J., Silvis, M. R., Hultquist, J. F., Waters, C. S., McGregor, M. J., Krogan, N. J., & Bondy-Denomy, J. Inhibition of CRISPR-Cas9 with bacteriophage proteins. *Cell*, 168(1-2), 150-158, 2017.
11. Watters, K. E., Fellmann, C., Bai, H. B., Ren, S. M., & Doudna, J. A. Systematic discovery of natural CRISPR-Cas12a inhibitors. *Science*, 362(6411), 236-239, 2018.
12. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17), 3389-3402, 1997.

Tatiana Lenskaia is a PhD candidate in the Department of Computer Science and Engineering with a specialization in Bioinformatics and Computational Biology at the University of Minnesota. Her primary research interests focus on modeling recognition systems in biology with an emphasis on genome organization and transitions between functional states in genome-genome interactions. Specifically, her work focuses on developing computational methods to explore the underlying mechanism of CRISPR-Cas systems.