Review by Pierre Pontarotti , 06 May 2024 07:44

This is an updated version of the highly successful 'Defense Finder,' with more phage defense systems described and the availability of new web services. The new version of Defense Finder will be an important tool for scientists interested in anti-phage defense systems

We thank the reviewer for his encouraging comment.

Review by anonymous reviewer 1, 06 Jun 2024 09:32
REVIEW on
A Comprehensive Resource for Exploring Antiphage Defense: DefenseFinder Webservice, Wiki and Databases.
doi: https://doi.org/10.1101/2024.01.25.577194

Title and abstract
Does the title clearly reflect the content of the article?
Mostly, Yes. I suggest the following change since the systems do not only target phages, but also other mobile genetic elements.
“Comprehensive Resource for Exploring Prokaryotic Defense Systems: DefenseFinder Webservice, Wiki and Databases.”
Does the abstract present the main findings of the study?
Yes. Slight improvements are suggested in the text below.
Introduction
Are the research questions/hypotheses/predictions clearly presented?
Yes
Does the introduction build on relevant research in the field?
Yes
Materials and methods
Are the methods and analyses sufficiently detailed to allow replication by other researchers?
Yes.
Are the methods and statistical analyses appropriate and well described?
Yes (mostly). See comments below.
Results
In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)?
I don’t know
Are the results described and interpreted correctly?
Yes. But since this work is introducing a tool, databases and an encyclopedia, only some first (broad) insights are shown.
Discussion
Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument?
Yes.

Are the conclusions adequately supported by the results (without overstating the implications of the findings)?
Yes.

In the here presented work, done by Tesson et al., a website has been created to improve the bioinformatic detection of antiphage systems. This includes an updated version of DefenseFinder with a web service, plus three databases (a wiki on defense systems, a structure database with experimentally determined and AlphaFold2 predicted structures, and a precomputed DefenseFinder results database).
Overall, this is a great work and contribution for the community!
The arsenal and complexity of defense systems are huge and are expanding very fast, making them hard to track. This work facilitates transparency and a fast acquisition of knowledge on the systems by providing a clear overview. The creation of wiki pages was a lot of work and is highly appreciated. The experimental validation part is very nicely presented.

We appreciate that the reviewer enjoyed this work.

I have only a few comments and minor concerns:

· How are orphans (HMM-only hits) treated or should be treated. These orphans are genes involved in systems but do not make complete ones? This was not addressed in this or the previous article(s). Please add a paragraph that would be a recommended approach to deal with them.

We agreed with the reviewer that orphan HMM hits should be analyzed differently. We add a sentence line 137-139 “All results are displayed, including orphan HMM, which does not form a system. Those orphan HMMs should be used cautiously and analyzed using their score and profile coverage.”

· What about interactions or even synergy between defense systems? And, also, how is the concept of layers of defense systems considered and could be represented?

We agree that interaction between defense systems is one of the interesting directions of the field. However, right now there are examples of synergies between systems that can not be extrapolated to all detected defense systems. This is why we are not adding any information regarding this in DefenseFinder results.

These are just thoughts and suggestions that could be added in future versions. Please add some clarifications on these points since these topics are addressed in the community (as reflected in current literature).

Minor concerns
I tested the web service and it works well in my point of view. Is it possible to download figures of the genomic organizations (in good quality) without the need to make screenshots?

We agree that the export of the genomic locus is important. A new feature has been developed and the selected region is now downloadable both in png or in vector format (svg).

Moreover, hits on several contigs detected in a multi-FASTA file are all displayed on ‘one genome arrow’. Can this be displayed on separate contigs (which visualization should be limited to 10 or an adjustable number)?

We agree that this feature is important for the visualization of multi-contig fasta. We add the feature to view the different contig separately for multi-contig nucleic acid fasta and to visualize only one or multiple contig at the same time.

Abstract

- L.19-21: “all known antiphages defense mechanism” is in my point of view (slightly) overstated. There are also other tools that predict the presence of antiviral systems such as PADLOC (PMC9252829), which won’t fully overlap with DefenseFinder, indicating that one tool cannot find all systems. I suggest to remove ‘all known’.

We agree with the reviewer and change from “all known” to “known”. We are still saying known defense systems because we are only detecting families of defense systems that have been experimentally validated.

- L. 26-29 I suggest the following change for a better readability:

“To overcome these challenges, we present a hub of resources on defense systems, including: 1) an updated version of DefenseFinder with a web-service search function, 2) a community-curated repository of knowledge on the systems, and 3) precomputed databases, which include annotations done on RefSeq genomes and structure predictions generated by AlphaFold.”

We thank the reviewer for this rewriting suggestion and change the text accordingly.

Methods

- L39-42. Sentence is unclear. Verb is missing (starting from L41).

We agree that the sentence was too long and hard to understand. We splitted the sentence in two parts Line 40-43 “However, recent discoveries have revealed an important diversity of molecular modalities by which bacteria defend themselves against phages. This diversity of mechanisms englobe nucleotide depletion [4–10], membrane disruption[11–14], production of antiviral molecules [15].”

- Creation of a homepage is quite specific and I do not have the know-how to give any comments this part. The only question that I have is, if the pages will be maintained on the long-term? Is there a funding behind?

The defensefinder website has been created in collaboration with the Pasteur Bioinformatics and Biostatistics Hub, which is a platform dedicated to support this kind of project over the long term. The backend of the website is also managed with the IT department of Pasteur Institute. This allows a long-term maintenance of the website and tools, as it has been with other software such as MacsyFinder and IntegronFinder, which have been maintained by the Bioinformatics Hub for about 10 years now and going.
For the maintenance of the content of the website, the objective is to continue to update it frequently and also to make it more collaborative so it relies not on a single team.

o I suggest combining protein selection and structure prediction (L.304-319) together. What does "best hit" mean? Is the selection based on the 30% identity and 70% coverage? This threshold is not very high and may be biased by different structures. In other words, what are the maximum differences between sequences, and does it make sense to take just one sequence as representative?

We agree with the reviewer that the explanation is not explicit. We are only choosing to represent one system because of computational limitations. Indeed, there is a diversity of sequences that can result in different structure inside a single system or subsystem family.
We change the text to clarify the selection of the best hits Line 324-330: “For several subsystems, it was not possible to retrieve experimentally validated sequences for two reasons: no protein sequences or accessions on the discovery paper or subsystem with no experimental validation. For those system, one of the best system hits from DefenseFinder was randomly selected and used for the protein structure prediction. Best hits were selected based on their hit scores and profile coverage (fourth quantile of hit score for each gene of the system and more than 75% of profile coverage).”

o Pfam annotation (L.326-333): What is meaning of superimposed in this context? Per protein or domain/sequence position?

We agree that the wording was not good. We change the word to overlapping. If two PFAM hits are at the same place inside the protein we are only using the best one. We change to Line 353-354 “If two PFAM hits were overlapping in a single protein sequence, only the best hit (hit_score) was kept.”

Conclusion
o I recommend rephrasing the first section to avoid frequent repetition of 'antiphage.'

We removed some occurrences of “antiphage” by replacing it by either “antiviral”, “bacterial defense systems”.

o Additionally, I suggest to add a small section describing future plans for updating and improving this platform/hub. This could include plans for exploring interactions between defense systems, conducting docking simulations using the predicted structures, providing tutorials, and even organizing workshops or events to foster collaborations in computational work on defense mechanisms.

We add a paragraph in the discussion to explain the possible new update for the website. We added Line 284-290 “ We will continue to develop the community aspect of the knowledge base by providing tutorials and organizing workshops to encourage people to contribute to the project. New updates will be made to increase the information on the website (new predicted structure, alphafoldDB49, increase in the number of genomes, sequence availability). We plan to add in a future release, a new section where users can test whether a system is related to a known one or not. If the system is new, we will provide a form to add the new system both for DefenseFinder and the website.”

Review by Pedro Leão, 22 May 2024 16:23
Florian Tesson and colleagues make a significant contribution to the field of defense systems by developing and providing access to three comprehensive databases. This initiative can greatly impact researchers at all levels, offering experienced individuals the opportunity to contribute to the databases' future expansion. It also serves as an accessible platform for newcomers and enthusiasts in the rapidly growing field of defense system research. Below are my minor comments and suggestions:

We thank the reviewer for these comments.

Line 74-75: "The information of the web server is integrated into the rest of the website." It would be interesting to explain a bit more what this means for the general public, and potential users of the website.

We developed more links between all the components of the website. We change the text to : “All those website components are also integrated within the DefenseFinder webserver output to easily find information on a system found in a genome.”

Line 82-83: We suggest the authors add this sentence to the beginning of the next paragraph to improve readability.

We agreed with the reviewer and modified the text accordingly.

Line 117: "We use pyrodigal v3.0.131 to identify and annotate the coding regions..." . I believe this process identify the coding regions, and translate them, not annotate. Please double check.

We agree with the reviewer that pyrodigal does not compute functional annotations. Therefore, we change the text from “to identify and annotate the coding regions” to “to identify and translate the coding regions”.

Line 116-117: It's not clear how these tests are making "the development of future features more robust". Can you elaborate a bit more on this?

When adding new features or when optimizing the code to make it faster, it's easy to introduce new bugs. Having a non-regression test allows anyone to develop new features or make improvement to the code with greater confidence when they know that what they modify does not impact existing results. This is considered good practice in software development (it's like having a control in an experiment - you know how it's supposed to behave). We clarified the sentence.

Line 222-223: "For systems and subsystems where protein accessions were impossible to retrieve, we selected another representative." Could you please make this process more clear? What protein accessions were not possible to retrieve? The experimentally validade proteins structures? proteins with experimentally validade anti-phage function?

We agree that the text was not clear. There are two ways a protein sequence might be inaccessible : no sequence in the original paper or, untested subtype of defense systems. We change the text Line 230-234 to "For some systems, we could not find the original protein sequence or accession of the experimentally validated system. Some subsystems (CBASS43, Retron44,45, Lamassu29…) were not experimentally validated and are included in DefenseFinder. In those two cases, we selected a representative in DefenseFinder results (See Methods).."

The same on the methods session regarding this process (Line 306-307): "For systems with no accession available". We would recommend the authors to be more precise. For systems with no experimental validation?

We also changed the text regarding the method section to both develop which sequences were not retrieved and how the representative sequences were selected. We change the text Line 324-330 to "For several subsystems, it was not possible to retrieve experimentally validated sequences for two reasons: no protein sequences or accessions in the original paper or, it's a subsystem with no experimental validation. For those systems, one of the best system hits from DefenseFinder was randomly selected and used for the protein structure prediction. Best hits were selected based on their hit scores and profile coverage (fourth quantile of hit score for each gene of the system and more than 75% of profile coverage)."

Line 322-323: We believe the number (n) of archaeal and bacterial genomes are inverted. Please double check it.
We thank the reviewer for the correction and change the text accordingly Line 343-344 to "of both Bacteria (N = 22,422) and Archaea (N = 381) from July 2022"