Bioinformatics, 35(13), 2019, 2309–2310
doi: 10.1093/bioinformatics/bty935
Advance Access Publication Date: 16 November 2018
Applications Note

Sequence analysis

PAVOOC: designing CRISPR sgRNAs using 3D protein structures and functional domain annotations

Moritz Schaefer[1,*], Djork-Arné Clevert[2], Bertram Weiss[2] and Andreas Steffen[2]

[1]Computer Science, TU Berlin, D-10623 Berlin, Germany and [2]Bioinformatics, Bayer AG, D-13342 Berlin, Germany

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on August 22, 2018; revised on October 1, 2018; editorial decision on October 12, 2018; accepted on November 9, 2018

Abstract

Summary: Single-guide RNAs (sgRNAs) targeting the same gene can significantly vary in terms of efficacy and specificity. PAVOOC (Prediction And Visualization of On- and Off-targets for CRISPR) is a web-based CRISPR sgRNA design tool that employs state of the art machine learning models to prioritize most effective candidate sgRNAs. In contrast to other tools, it maps sgRNAs to functional domains and protein structures and visualizes cut sites on corresponding protein crystal structures. Furthermore, PAVOOC supports homology-directed repair template generation for genome editing experiments and the visualization of the mutated amino acids in 3D.

Availability and implementation: PAVOOC is available under https://pavooc.me and accessible using modern browsers (Chrome/Chromium recommended). The source code is hosted at github.com/moritzschaefer/pavooc under the *MIT License*. The backend, including data processing steps, and the frontend are implemented in Python 3 and ReactJS, respectively. All components run in a simple Docker environment.

Contact: mail@moritzs.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The discovery of the CRISPR/Cas system (Cong *et al.*, 2013; Jinek *et al.*, 2012) was a breakthrough in the area of genome editing. An important application of CRISPR/Cas is to induce a targeted knockout (KO) of a gene of interest. Such KO experiments can help to study the essentiality of the targeted genes in given cellular contexts (e.g. a cancer cell line bearing certain genomic alterations) and ultimately support the validation of a new drug target (Moore, 2015). Shi *et al.* (2015) showed, that the effect of a CRISPR based KO can be boosted by targeting functionally relevant regions of a protein, as in these regions in-frame mutations (indels) are more likely to induce a significant effect than in non-functional regions. Another application of CRISPR/Cas is to precisely introduce missense mutations into a genome and study the resulting effects of the perturbations. In both applications single-guide RNAs (sgRNAs) are used to direct the Cas9 enzyme towards the genomic region of interest, such that the Cas9 can cut the DNA at the targeted position. For the genome editing experiments in addition a template sequence needs to be provided that contains the desired nucleotide sequence.

A number of tools have been published that facilitate and automate the design of sgRNAs for CRISPR KO experiments (Hough *et al.*, 2016; Listgarten *et al.*, 2018; Meier *et al.*, 2017; Stemmer *et al.*, 2015). In this application note, we present PAVOOC (Prediction And Visualization of On- and Off-targets for CRISPR)—a modern web application to support wet lab biologists in designing and selecting optimal sgRNAs and template sequences for KO and genome editing experiments using machine learning-based on- and off-target scoring, multi-attribute ranking, protein structure mapping of the cut sites and integration of cancer cell line data.

2 Materials and methods

PAVOOC is a web application that allows to design and visualize sgRNAs for gene KO and genome editing experiments. For KO experiments, a set of genes has to be provided (in form of symbols or Ensembl identifiers). PAVOOC then generates a table that contains a user-defined number of sgRNAs for each of these genes. These sgRNAs are prioritized based on the scoring function in Equation (1), which combines weighted on- and an off-target scores as well as whether the targeted regions lies within a protein domain. The on-target score is calculated using the Azimuth model, whereas the cutting frequency determination score is used to assess off-target effects (Doench *et al.*, 2016).

$$\begin{aligned} ranking\_score &= 0.6 \cdot on\_target\_score \\ &\quad + 0.3 \cdot (1 - off\_target\_score) \\ &\quad + 0.1 \cdot is\_in\_domain \end{aligned} \tag{1}$$

It is possible to further analyze and modify the sgRNA selection for a gene in a detail view (see Supplementary Fig. S1). The detail view consists of three synchronized panels: The 'LineUp' ranking table on the upper right, the protein structure view on the upper left and the sequence view on the bottom panel of the page. The LineUp (Gratzl *et al.*, 2013)-based sgRNA ranking table allows an individual adjustment of the weights for the on- and off-target scores in order to prioritize the sgRNAs accordingly. For each sgRNA, the LineUp table displays whether the targeted genomic region lies within a protein domain and whether the optionally selected cancer cell line contains a single nucleotide variation at that position. The sequence view on the bottom is based on Biodalliance (Down *et al.*, 2011) and shows the gene annotation, all targeted regions of the sgRNAs, protein domains and cancer cell line alteration data in order to support the tailored sgRNA design for a cell line under study. On the left side, available 3D protein structures from RCSB (Berman *et al.*, 2000) are shown and sgRNA-related cut sites are mapped and highlighted on the structure using the NGL viewer (Rose and Hildebrand, 2015). In this way, the user can assess the position of the Cas9 cut position on the protein structure and thus prioritize sgRNAs that are more likely to affect functionally relevant regions of a gene. Furthermore, when designing genome editing experiments, the structure view enables amino acid editing and displaying the designed alterations directly on the protein structure.

We integrated genomic sequence data from UCSC in version hg19 (Consortium *et al.*, 2001). The genomic annotations, including genes, transcripts and exons were taken from the GENCODE project (Harrow *et al.*, 2012). Cancer cell line alteration data was taken from the Cancer Cell Line Encyclopedia (Barretina *et al.*, 2012) (based on hg19). In order to facilitate the mapping between genomic and protein coordinates we used the canonical transcript from APPRIS (Rodriguez *et al.*, 2013) only. Exons which are not contained in that transcript are not considered in our application. SIFTS (Velankar *et al.*, 2012) mappings are used to derive genomic coordinates of PDB structures. A structured overview of our pipeline is shown in Supplementary Figure S2.

The data shown in the application is all pre-processed offline and stored in a non-relational database. Guide search and off-target scoring is performed using FlashFry (McKenna and Shendure, (2017).

3 Discussion

Our new tool PAVOOC provides a convenient means to design optimal sgRNAs for KO and genome editing experiments. A machine learning-based scoring system guides the user to select sgRNAs with possibly strong on- and low off-target effects. Through the integration of structural data, PAVOOC is able to display cut sites on corresponding protein crystal structures such that sgRNAs can be selected which cut in functionally relevant regions. Integration of cancer cell line data ensures that existing genomic alterations are considered during sgRNA selection. The tool was used internally to design a domain-targeting genome-wide sgRNA library.

PAVOOC is hosted on GitHub and is an actively maintained project. As such, it provides an open platform to build and integrate use cases of CRISPR that are not part of the current state. The PEP8 compliant Python code and the react.js-based frontend simplify the entry for developers. The application runs in a Docker environment which makes it easy to host the application on premise.

Acknowledgements

We thank Robin Winter, Claudia Noack and Barbara Nicke for useful suggestions and discussions throughout the project.

Funding

This work was supported by the Bayer AG.

Conflict of Interest: none declared.

References

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Cong,L. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

Consortium,I.H.G.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860.

Doench,J.G. *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.

Down,T.A. *et al.* (2011) Dalliance: interactive genome viewing on the web. *Bioinformatics*, **27**, 889–890.

Gratzl,S. *et al.* (2013) LineUp: visual analysis of multi-attribute rankings. *IEEE Trans. Vis. Comput. Graph.*, **19**, 2277–2286.

Harrow,J. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760.

Hough,S.H. *et al.* (2016) Desktop genetics. *Personalized Medicine*, **13**, 517–521.

Jinek,M. *et al.* (2012) A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.

Listgarten,J. *et al.* (2018) Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat. Biomed. Eng.*, **2**, 38.

McKenna,A. and Shendure,J. (2018) FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC biology*, **16**, 74.

Meier,J.A. *et al.* (2017) GUIDES: sgRNA design for loss-of-function screens. *Nat. Methods*, **14**, 831.

Moore,J.D. (2015) The impact of CRISPR–Cas9 on target identification and validation. *Drug Discov. Today*, **20**, 450–457.

Rodriguez,J.M. *et al.* (2013) APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.*, **41**, D110–D117.

Rose,A.S. and Hildebrand,P.W. (2015) Ngl viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.

Shi,J. *et al.* (2015) Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol.*, **33**, 661.

Stemmer,M. *et al.* (2015) CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS One*, **10**, e0124633.

Velankar,S. *et al.* (2012) SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.