An efficient command-line tool for CRISPR-CAS9 experiment design

Srinivas Suresh, Srivathsa Pasumarthi, Juhi Malani, Shubha Tirumale

October 26, 2017

Introduction

CRISPR-CAS9 system is a heritable natural immune system in prokaryotes which has kindled interest in the last few years for its unique applications across different life systems. The range of applications include knock-in/knock-out of functional and non-functional genes, activation of silenced genes, introduction of pre-designed mutations and promoter studies or screening in genomes. Owing to the great potential of CRISPR system, there is a growing need for tools that help the researchers to make informed decisions when designing the CRISPR-CAS9 experiment.

Methods

Many tools like CHOPHOP[1], CRISPOR[2], FlashFry[3], CasFinder/ CasValue [4] and Cas-OFFinder[5] are available for facilitating the $sgRNA$ design as well as for finding putative genome sites for a given $sgRNA$. These tools employ various computational techniques, specialized algorithms and indexing data structures to improve on the computational efficiency while producing accurate results. During the course of this project, we would like to explore various algorithms, indexing data structures and querying mechanisms which will attempt to improve the memory footprint and/or the computational efficiency of the aforementioned tools. In particular, we would like to investigate the applicability of the probabilistic data structure - $BloomFilter$[6] as a means of improving the performance of the system.

The efficiency with which an $sgRNA$ binds to a target sequence determines it's *on-target* score. Though there are a few basic characteristics of a target sequence (like the NGG PAM requirement for the SP - *Streptococcus pyogenes* - CAS9, the GC-content etc) that makes an $sgRNA$ to bind efficiently to it, a lot of emperical evidence is being collected recently[7] to understand the more subtle homological relation between the $sgRNA$ and the target sequence to create the desired cleavage. Based on the emperical evidence, *Doench et al*[7] has built a predictive logistic regression model using SVM to build a more reliable $sgRNA$ scoring function. As the second goal for this project, we would like to avail the existing experimental data and build a more robust predictive model using at least one of - Decision Trees, Random Forests and Neural Networks.

Assumptions

The exact conditions under which a target sequence will efficiently bind with an $sgRNA$ is not clearly and completely defined and is still being figured out by researchers by performing experiments with various $sgRNA$s and genes of interest. Having said that, there are a few well-defined rules under which there is a high probability that a cleavage will happen in a desired manner and for the purpose of simplicity, we would like to consider only these rules in the design of the proposed tools. We would be using these rules to determine both the *on-target* and the *off-target* scores. It is worth mentioning that it is safe to make such an assumption given the fact that most of the existing computational tools for CRISPR experiment design use only a subset of these rules. The rules are as follows:

1. **Necessary Condition**: We would like to consider the case where only the *SP-CAS9* endonuclease enzyme is used in the experiment. This means that we will look only for the NGG PAM (Protospace Adjacent Motif) to be present downstream (towards the 3' end).

2. Prefer target sequences having a Guanine as the first (in the 5' end) nucleotide.

3. Prefer target sequences having a Guanine-Cytosine (GC) content between 40-80%.

4. Prefer target sequences having minimal or no base-pair mismatches in the *seed-region* (12-13 nt upstream of PAM) when compared to the *non-seed region*.

Milestones

1. To develop a CLI (command line interface) to find putative genome sites for a given $sgRNA$. The tool will use $BloomFilter$ in order to attempt to improve on the computational efficiency and/or the memory bandwidth of at least one tool among [1], [2], [3], [4], [5].

2. To develop a CLI to find the most effective $sgRNA$(s) given a region of interest in the reference genome. The tool will use $BloomFilter$ in order to attempt to improve on the computational efficiency and/or the memory bandwidth of at least one tool among [1], [2], [3], [4], [5].

3. To develop a robust scoring algorithm by building a predictive model using at least one of Decision Trees, Random Forests and Neural Networks. The aim is to improve on the predictive capability of the existing SVM logistic regressor [7].

Stretch Goals

1. To develop an intuitive web-based visualization/simulation tool for CRISPR experiment design, improving on the various aspects such as readability, ease-of-use etc., of existing tools [1], [8].

2. To develop a parallelized GPU accelerated version of the aforementioned CLIs.

Evaluation

Accuracy

We would like to assess the *on-target* and *off-target* scores of the results (computed from CasValue [4]) produced by our tools and compare the same

with that of [1] and [4] to show that the accuracy of our tool is on par with the existing ones.

We would like to assess the accuracy of our score prediction model by comparing its *F1 Score* with that of the existing SVM Logistic Regressor[7].

Computational Efficiency

We would like to assess the CPI (Cycles per instruction) or IPC (Instructions per cycle) metrics of our tools and compare the same with that of [1] and [4] to find if our tools computationally perform better than the existing ones.

Memory Performance

We would like to assess the memory footprint of our tools by monitoring the RAM usage and compare the same with that of [1] and [4] to find if our tools' memory performance is better than the existing ones.

Datasets

Reference Genome

For both our CLIs, we would like to start with a smaller *prokaryotic* genome and then extend to the larger *human reference genome* (H. Sapien hg38/GRCh38).

Guide RNA

We would like to use a subset of the $1841 - sgRNAs$ from [7] for which the *on-target* and *off-target* scores are already computed.

Data for predictive model

In order to build the predictive model we would use the featured list of *sgRNA*s along with the most compatible target sequences from [7].

References

[1] Kornel Labun and Tessa G. Montague. *CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering.* Nucleic Acids Research, 2016, Vol. 44, Web Server issue, Bergen, Norway, 2016.

[2] Maximilian Haeussler, Kai Schonig, Helene Eckert, Alexis Eschstruth, Joffrey Mianne, Jean- Baptiste Renaud, Sylvie Schneider-Maunoury, Alena Shkumatava, Lydia Teboul, Jim Kent, et al. *Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool crispor.* Genome biology, 17(1):148, 2016.

[3] Aaron McKenna and Jay Shendure. *FlashFry: a fast and flexible tool for large-scale CRISPR target design.* bioRxiv, Seattle, WA, USA 2017.

[4] John Aach. *CasFinder: Flexible algorithm for identifying specific Cas9 targets in genomes.* bioRxiv, Boston, MA, USA, 2014.

[5] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion.* Addison-Wesley, Reading, Massachusetts, 1993.

[6] Bloom, Burton H. *Space/Time Trade-offs in Hash Coding with Allowable Errors.* Commun. ACM. New York, NY, USA, 1970.

[7] John G. Doench, Ella Hartenian, Daniel B. Graham, Zuzana Tothova, Mudra Hegde, Ian Smith, Meagan Sullender, Benjamin L. Ebert, Ramnik J. Xavier, and David E. Root. *Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation.* NIH-PA, Boston, MA, USA 2014.

[8] UCSC Genome Browser - https://genome.ucsc.edu/cgi-bin/hgGateway