CRISPR/Cas9 LockOn

A project by Meryl Stav and Mike Tung

Northeastern University

BIOL6309 Bioinformatics Computational Methods 2

Table of Contents

[Specific Aims 3](#_Toc448013765)

[Background and Significance 3](#_Toc448013766)

[Preliminary Studies 4](#_Toc448013767)

[Research Plan 5](#_Toc448013768)

[Data Fetching: 5](#_Toc448013769)

[Pam Finder: 5](#_Toc448013770)

[Gene Selection: 5](#_Toc448013771)

[Proposal 5](#_Toc448013772)

[Works Cited 7](#_Toc448013773)

Table of Figures

Figure 1: CRISPR/Cas9 Endonuclease System. Cas9 and a synthetic gRNA with the spacer sequence is coupled to form a Cas9-gRNA complex. The Cas9-gRNA complex binds to the gene of interest (blue) that is immediately followed by a PAM (orange). If there is sufficient homology between the target site and the gRNA and there exists a PAM, Cas9 cuts the DNA and exits, leaving the damaged DNA to repair itself either by Non-Homologous-End-Joining or Homology-Directed-Repair (Not Shown) to introduce mutations that will disrupt and silence the targeted gene. 4

Table of Tables

Table 1: Division of Labor between Mike Tung and Meryl Stav 6

# Specific Aims

The goal of the proposed project is to provide a command-line tool, which will help facilitate the experimental designs of biologists in the lab by providing the locations of Protospacer-Adjacent-Motifs(PAMs) in a given genome for CRISPR/Cas9 targeting.

1. Data Fetching: Provide a script to download and fetch DNA data from NCBI.
2. Pam Finder: Build a command-line tool to search PAMs of a given variant of Cas9.
   1. Start out with SP variants (grouped) but later move on to other strains.
3. Gene Selection: Find a set of genes for targeting and show that the tool can be used to determine if it is possible to target the gene using CRISPR/Cas9.
   1. Start out in Saccharomyces *cerevisiae* (SC282) for pilot.
   2. Scale up to Escherichia *coli*.

# Background and Significance

Clustered regularly interspaced short palindromic repeats or CRISPRs are a particular family of tandem repeats found in a wide range of prokaryotic genomes (Grissa, Vergnaud, & Pourcel, 2007). They consist of a succession of highly conserved regions varying in size from 23 to 47bp separated by similarly sized unique sequences termed spacers. Associated with CRISPRs are endonucleases called Cas9 (CRISPR Associated Proteins) and short “guide” RNA (gRNA) which when coupled together in the CRISPR system can become an essential tool for molecular scientists interested in targeting genes.

Starting with the gRNA, scientists can design a synthetic gRNA of about length 20bp that is complementary to a gene of interest. This targeting sequence or “spacer” sequence can be changed to be complementary to any part of the genome, i.e. target any gene of choice. However, two conditions must be met in order for the CRISPR/CAS9 system to knock a gene out. The first condition is that the spacer must be unique when compared to the rest of the genome, and the second condition is that the target is immediately upstream of a Protospacer Adjacent Motif (PAM) site. The PAM site is mandatory for target binding and the exact sequence of the PAM depends on the species of Cas9, i.e. the source bacteria from which Cas9 was derived (**Figure 1**).

Once the two conditions have been met, CRISPR/Cas9 binds to the DNA at the target site and performs a double stranded break and the resulting damaged DNA is then subject to either Non Homologous End Joining or Homologous Directed Repair depending on the experimental design.

The ability to effectively target a gene in the genome proves to be an invaluable tool for molecular biologist. One real world application would be in personalized medicine, if there is sufficient knowledge of genes associated with a disease such as Alzheimer's then scientists could cure Alzheimer's by designing a CRISPR/Cas9 system that can target and silence all genes associated with Alzheimer’s disease. Another real world application would be in population genetics, where CRISPR/Cas9 could be used in “driving” a gene in a given population into extinction or resurrection. Of course these real world applications pose ethical issues, but these applications also demonstrate the powerful utility of these “molecular scissors”.

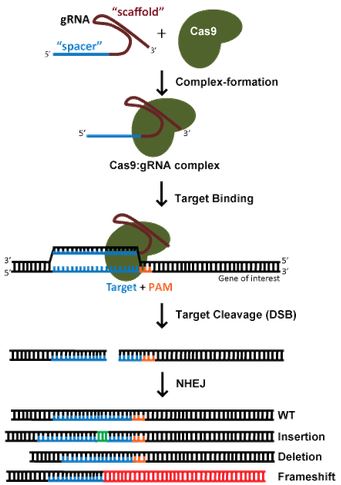


Figure 1: CRISPR/Cas9 Endonuclease System. Cas9 and a synthetic gRNA with the spacer sequence is coupled to form a Cas9-gRNA complex. The Cas9-gRNA complex binds to the gene of interest (blue) that is immediately followed by a PAM (orange). If there is sufficient homology between the target site and the gRNA and there exists a PAM, Cas9 cuts the DNA and exits, leaving the damaged DNA to repair itself either by Non-Homologous-End-Joining or Homology-Directed-Repair (Not Shown) to introduce mutations that will disrupt and silence the targeted gene. (Addgene, 2015)

# Preliminary Studies

We programmatically downloaded each of the yeast chromosomes from NCBI in FASTA format for our preliminary study using a Perl script.

In addition, we have written a Perl script, “pamFinder.pl”, which accepts FASTA files, a given strain of Cas9 and a file of guide RNAs. The script searches the sequence located in the FASTA file for a reverse complementary match for each of the guides and checks to see if a PAM site is immediately downstream of the matched region. The output is a tab separated values file or TSV file which contains the location of the match in the FASTA, the guide matched, and the DNA strand (sense or anti-sense).

We then went to Saccharomyces Genome Database (SGD) (Standford University, 2011) and looked for a gene to target with our tool. The gene selected was ARS103, which is an Autonomously Replicating Sequence responsible for yeast growth. The sequence spanned 560bp and we developed a short utility tool to subset the sequence into subsequences of length 20bp with sliding window.

# Research Plan

To create a command-line based tool to help design experiments regarding gene studies, we propose the following plan:

## Data Fetching:

NCBI provides a wealth of data regarding various organisms and provides an API to programmatically fetch data in BioPerl. Cross referencing NCBI with SGD, Saccharomyces Genome Database, helped find all the corresponding RefSeq IDs needed to fetch the corresponding Yeast chromosomes which make up the Yeast Genome. In addition, we used SGD to find candidate genes to test our tool in Yeast Strain S288C.

## Pam Finder:

Pam Finder is an essential component of our study, and all downstream actions are reliant on the tool. We built the tool with the mindset of maintaining scalability to bigger and much more complex genomes. In addition, the tool is very modular in the sense of workflows involving analyses of multiple genomes or multiple regions of a given genome. In our preliminary study we wrote wrappers to demonstrate the automation of workflow via a set of Perl scripts to automate tasks such as searching multiple genomes for targetable gene regions or filtering the output file for specific genomic regions.

## Gene Selection:

Preliminary research confirmed the feasibility of our Pam Finder tool as we had chosen to see if ARSE103 a gene involved in yeast growth was targetable by CRISPR/Cas9. The result was that CRISPR/Cas9 could target the gene because there were PAM sites present immediately downstream of some of our guide RNAs. The resulting output file could then be translated into experiments done in the wet lab setting to generate data for downstream analyses.

We hope in future work to test and see how close our tool comes in prediction of CRISPR/Cas9 targeting. The results of the future work would help improve our tool and streamline laboratory workflows involving CRISPR/Cas9 for gene studies.

# Proposal

We propose a tool for designing a CRISPR/Cas9 system that follows the rules for gene targeting and silencing. The programming project, named CRISPR/Cas9 LockOn, will consist of three pieces: a database, a user friendly script, and tests. The database portion of the project will consist of curating information about various Cas9 PAMs available for endonucleases and storing them in some database system along with the outputted data from our program. The second component of the project will be our program CRISPR/Cas9 Lock-On! that will be implemented in Perl. The preliminary aim of the program will be to effectively find all valid CRISPR/Cas9 targetable sites in Saccharomyces *cerevisiae* (baker’s yeast) strain S288c, but we do plan to scale up to more genomes in the future such as humans. Finally, the last part of our project will be testing and automating the program to ensure high quality user-interface (UI) experience.

The workload for each of the pieces will be divided evenly between Meryl Stav and Michael Tung. (Table 1) All documentation of the project will be stored on github and will remain Open Source.

Table 1: Division of Labor between Mike Tung and Meryl Stav

|  |  |  |
| --- | --- | --- |
| Task | Mike Tung | Meryl Stav |
| Data Fetching | Setup framework and rough prototype | Debugged and updated code with finished design |
| Research Biological Background | Searched literature and web for CRISPR/Cas9 Strains and Variants |  |
| PAM Finder | Setup framework and wrote function to output file, search FASTA for matches | Setup framework and wrote functions to check guide RNA sequences, updated code with more CRISPR/Cas9 Strain options, wrote wrappers and tools to assist in testing tool |
| Unit Testing |  | Unit Tested each component and made necessary updates |
| Future Work | Perform the wetlab experiments | Test to see how accurate tool was in relation to real results, perform benchmark tests, apply parallel processing techniques for optimizations |

# Works Cited

Addgene. (2015, 5 13). *CRISPR/Cas9 Guide.* Retrieved 4 7, 2016, from Addgene: http://www.addgene.org/CRISPR/guide/?gclid=Cj0KEQjwipi4BRD7t6zGl6m75IgBEiQAn7CfF-oizWaEgV-jqzW0L5vbmTRv5Nv1BJoUktHPny1whUcaAop28P8HAQ

Grissa, I., Vergnaud, G., & Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research, 35*(Web Server), 6.

National Center for Biotechnology Information. (n.d.). *National Center for Biotechnology Information*. Retrieved 04 07, 2016, from U.S. National Library of Medicine: http://www.ncbi.nlm.nih.gov/

Standford University. (2011, 12 16). *Sacharromyces Genome Database S288c*. Retrieved 2016, from Sacharromyces Genome Database: http://www.yeastgenome.org/