

**COMPUTATIONAL INSIGHTS INTO CRISPR/CAS9 SYSTEM
FOR IMPROVED GENOME EDITING**

JASPREET KAUR DHANJAL



**DEPARTMENT OF BIOCHEMICAL ENGINEERING AND BIOTECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI
APRIL 2019**

©INDIAN INSTITUTE OF TECHNOLOGY DELHI (IITD), NEW DELHI, 2019

**COMPUTATIONAL INSIGHTS INTO CRISPR/CAS9 SYSTEM
FOR IMPROVED GENOME EDITING**

by

Jaspreet Kaur Dhanjal

Department of Biochemical Engineering and Biotechnology

Submitted

**in fulfilment of the requirements for the degree of Doctor of Philosophy
to the**



Indian Institute of Technology Delhi

April 2019

Certificate

This is to certify that the thesis entitled '**Computational insights into CRISPR/Cas9 system for improved genome editing**' being submitted by **Ms. Jaspreet Kaur Dhanjal** to the Indian Institute of Technology Delhi for the award of the degree of '**Doctor of Philosophy**', is a record of the bonafide research work carried out by her, which has been prepared under my supervision in conformity with the rules and regulations of the Indian Institute of Technology Delhi. The research reports and the results presented in this thesis have not been submitted for any degree or diploma in any other University or Institute.

Dr. D. Sundar

Professor

Department of Biochemical Engineering and Biotechnology

Indian Institute of Technology Delhi

Acknowledgements

First and foremost I want to thank my supervisor, **Prof. D. Sundar**. I appreciate all his contributions of time and ideas to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I am also thankful for the excellent example he has provided as a successful professor and a great human.

Besides my supervisor, I would like to thank the rest of my research committee- Prof. G.P. Agarwal, Prof. Ritu Kulshreshtha and Prof. Samudrala Gourinath (JNU, New Delhi), for their insightful comments and encouragement, and also for the hard questions which incited me to widen my research from various perspectives. I am also grateful to Dr. Manish Agarwal, Administrator for The Central Hybrid Supercomputing Cluster at IIT Delhi, for being very kind, patient, and always willing to support whenever I approached him.

The members of my research group (DAILAB @ IIT Delhi) have contributed immensely to my personal and professional time at the institute. The group has been a source of friendship as well as good advice and collaboration. It was a pleasure to work with all the scholars and the numerous other grad students and interns who have come through the lab. I would like to specially mention Mr. Shashank, Navaneethan, Shreya, Samvit, Yugesh and Vivek for helping me in carrying out my work. Learning from them was a great experience.

Lastly, I would like to thank my family for all their love and encouragement, for my parents and brothers who supported me in all my pursuits.

Jaspreet Kaur Dhanjal

Abstract

In spite of a decade of research in the area of targeted genome editing, which has revolutionized the field of science, technology and medicine, there are still many aspects that limit the widespread use of genome editing tools for targeted cleavage of the region of interest in genomic DNA. Various issues associated with these tools include the lack of specificity and precision, low efficiency in mammalian cells, delivery into the cells and immunogenicity. In this thesis, I have studied the problem of specificity and precision of one of the most recent genome editing tool-CRISPR/Cas9 system.

The ability to direct the CRISPR/Cas9 nuclease to a unique target site within a genome holds great promise in the field of targeted genome engineering. However, CRISPR RNA is reported to bind to other genomic locations that differ from the intended target site by a few nucleotides, demonstrating significant off-target activity. Previous studies have sought to predict CRISPR target sites and their corresponding possible off-targets within a genome that are mainly based on sequence similarity between CRISPR RNA and the genomic DNA. Moreover, the intricacies of RNA-DNA interaction recently reported in literature, namely the mismatch, single guide RNA and DNA bulge tolerance pattern are not yet explored for screening the off-targets. This thesis is an attempt to understand the working of CRISPR/Cas9 system, mainly focusing on the factors governing its on-target and off-target activity. A webserver called *CRISPCut* has been developed for prediction of optimal sgRNAs for targeting the human genome in different cell types. The predictions made by this webserver were further improved using a machine learning-based model that gives the probability of a human genomic region (similar to the target site) to be an off-target location for cleavage. Finally, the applicability of CRISPR/Cas9 system to target cancer-specific cell vulnerabilities for designing new therapeutic interventions was investigated.

सारांश

लक्षित जीनोम संपादन के क्षेत्र में एक दशक के अनुसंधान ने विज्ञान, प्रौद्योगिकी और चिकित्सा के क्षेत्र में क्रांति लाई। इसके बावजूद अभी भी कई पहलू हैं जो जीनोम संपादन उपकरण के व्यापक उपयोग को उद्देशित जीनोमिक डीएनए के लक्षित अनुभेदन के लिए सीमित करते हैं। इन उपकरणों से जुड़े विभिन्न मुद्दों में विशिष्टता और परिशुद्धता की कमी, स्तनधारी कोशिकाओं में कम दक्षता, कोशिकाओं में वितरण और अनियंत्रितता शामिल हैं। इस थीसिस में, मैंने सबसे आधुनिक जीनोम संपादन साधन- क्रिस्पर/कैस9 प्रणाली की विशिष्टता और सटीकता की समस्या का अध्ययन किया है।

जीनोम में एक अद्वितीय लक्ष्य स्थल के लिए क्रिस्पर/कैस9 न्यूक्लियस को निर्देशित करने की क्षमता लक्षित जीनोम अभियांत्रिकी के क्षेत्र में अद्भुत क्षमता दर्शाती है। यद्यपि क्रिस्पर आरएनए को कुछ न्यूक्लियोटाइड्स द्वारा लक्षित स्थल से भिन्न अन्य जीनोमिक स्थानों से बाँधने की सूचना दी गई है, जो महत्वपूर्ण ऑफ-टारगेट गतिविधि का प्रदर्शन करते हैं। पिछले अध्ययनों ने एक जीनोम में क्रिस्पर लक्ष्य साइटों और उनके संबंधित संभावित लक्ष्य का अनुमान लगाने का प्रयास किया है, जो मुख्य रूप से क्रिस्पर आरएनए और जीनोमिक डीएनए के बीच अनुक्रम समानता से संबंधित हैं। किंतु हाल ही में साहित्य में सूचना दी गई है कि आरएनए-डीएनए परस्पर क्रिया की पेचीदगियों, अर्थात् बेमेल, एसजीआरएनए और डीएनए उभार सहिष्णुता पैटर्न को ऑफ-टारगेट की जांच के लिए नहीं खोजा गया है। यह थीसिस क्रिस्पर/कैस9 प्रणाली के कार्य को समझने का एक प्रयास है, मुख्य रूप से इसके ऑन-टारगेट और ऑफ-टारगेट गतिविधि को नियंत्रित करने वाले कारकों पर ध्यान केंद्रित कर रहा है। विभिन्न प्रकारों में मानव जीनोम को लक्षित करने के लिए इष्टतम एसजीआरएनए की पूर्वानुमान के लिए क्रिस्पकट नामक एक वेबसर्वर विकसित किया गया है। इस वेबसर्वर द्वारा किये गए अनुमानित परिणाम को एक मशीन लर्निंग-आधारित मॉडल का उपयोग करके और भी बेहतर बनाया गया, जिसने मानव जीनोमिक क्षेत्र (लक्ष्य साइट के समान) को अनुभेदन के लिए एक ऑफ-टारगेट स्थान होने की संभावना बताता है। अंततः नए चिकित्सीय हस्तक्षेपों की रचना करने के लिए कैंसर-विशिष्ट कोशिकाओं की कमियों को लक्षित करने के लिए क्रिस्पर/कैस9 प्रणाली की प्रयोज्यता की जांच की गई है।

Table of Contents

List of figures	i
List of tables	iii
List of abbreviations	v

CHAPTER 1. Introduction to CRISPR/Cas9 system

1.1. Genome engineering	1
1.2. CRISPR/Cas9 system, molecular scissors for targeted genome engineering	4
1.2.1. Evolution of CRISPR/Cas9 system from a bacterial defence mechanism to a promising genome editing tool	4
1.3. Genetic perturbations using CRISPR/Cas9 system	5
1.3.1. Gene knock-out.....	6
1.3.2. Gene knock-in.....	6
1.3.3. Gene expression control.....	7
1.4. Major applications of CRISPR/Cas9 system in diagnostics and therapeutics....	7
1.5. Off-target activity of CRISPR/Cas9 system - Definition of problem.....	8
1.6. Role of computational approaches in finding optimal solutions	9
1.7. Thesis organization	9

CHAPTER 2. Governing rules for design of sgRNA

2.1. Design rules for highly efficient on-target sgRNAs	13
2.1.1. Sequence profile of the target DNA and sgRNA.....	13
2.1.2. PAM and sequence flanking the 23 nucleotide long target DNA.....	14
2.1.3. Location targeted within the gene.....	14
2.1.4. Accessibility of the target DNA.....	14
2.1.5. Microhomology profile of the target site	16
2.1.6. Epigenetic alterations in the chromatin.....	16
2.1.7. Experimental conditions	17
2.2. Computational approaches for designing sgRNA with high on-target efficiency.....	18
2.2.1. sgRNA design tools for knock-out and knock-in experiments.....	18
2.2.2. Species-specific tools for design of sgRNA	21

2.2.3. Tools for analysis of data obtained from CRISPR/Cas9-based experiments	22
2.2.4. CRISPR/Cas9 genome editing databases.....	22
2.3. Current challenges associated with sgRNA design tools.....	22

CHAPTER 3. *CRISPCut* : A novel tool for designing optimal sgRNAs for CRISPR/Cas9-based experiments in human cells

3.1. Background.....	25
3.2. Implementation of the prediction algorithm	26
3.2.1. Input options	26
3.2.2. Workflow	26
3.2.3. <i>CRISPCut</i> server	28
3.3. Visualization	28
3.4. Results and Discussion	30
3.5. Conclusion	35

CHAPTER 4. Machine learning-based model for the evaluation of off-targets predicted by *CRISPCut*

4.1. Background.....	37
4.2. Methodology.....	38
4.2.1. Preparation of target and off-target dataset.....	38
4.2.2. Calculation of sequence descriptors.....	39
4.2.3. One Hot Encoding of categorical variables	39
4.2.4. Testing different classifiers for model building.....	40
4.2.5. Gradient boosted regression algorithm	41
4.2.6. Feature importance.....	41
4.3. Results.....	41
4.3.1. Data summary	41
4.3.2. Model building using different algorithms	41
4.3.3. Use of gradient boosted regression trees	42
4.3.4. Sequence descriptors or features contributing in accuracy of prediction ..	42
4.4. Discussion.....	44
4.5. Conclusion	47

CHAPTER 5. Designing targeted therapies using cancer-specific synthetic lethal genetic interactions

5.1. Background	49
5.1.1. Cancer biology and drug design	49
5.1.2. The concept of synthetic lethal interactions.....	50
5.2. Exploring synthetic lethal genetic interactions	51
5.2.1. Data for genomic mutations in cancer patients.....	51
5.2.2. Investigating mutually exclusive genetic alterations and inferring synthetic lethal combinations.....	51
5.2.3. Comparison of predicted results with genome-wide essentiality screens across different cancer cell lines	52
5.2.4. sgRNA designs for targeting genes comprising the synthetic lethal pairs.....	52
5.3. Results.....	52
5.3.1. Weighted exact test for fishing synthetic lethal gene pairs.....	52
5.3.2. Genes in synthetic lethal combination with the most frequently mutated gene, PIK3CA	53
5.3.3. Synthetic lethal combination of different genes with TP53 as partner	54
5.3.4. sgRNA designs for the knock-out of predicted genes	56
5.4. Discussion	56
5.4.1. Induction of lethality in breast cancer cells harbouring gain-of-function mutation in PIK3CA	56
5.4.2. Gene candidates for synthetic lethality in breast cancer cells with TP53 mutations.....	58
5.5. Conclusion	59
CHAPTER 6. Conclusions	60
References.....	65
Appendices.....	81
Publications from the thesis	87
Resume of the PhD candidate	89

List of figures

Figure 1.1. Structure of CRISPR/Cas9 system.	5
Figure 1.2. Off-target activity of CRISPR/Cas9 system.....	8
Figure 3.1. The basic work flow of <i>CRISPCut</i>	29
Figure 3.2. Snapshot of <i>CRISPCut</i> web interface.....	30
Figure 3.3. Off-targets reported from GUIDE-seq experiments that have been predicted by <i>CRISPCut</i> in comparison with other prediction tools, namely CcTop, Optimized CRISPR Design-MIT, CHOPCHOP, and ZiFit.....	34
Figure 4.1. CRISPR/Cas9 system bound to target DNA.	39
Figure 4.2. Example showing One Hot Encoding of categorical sequence features for data preparation.....	40
Figure 4.3. <i>SHAP</i> values depicting the importance of sequence features contributing to the efficacy of prediction model with respect to each data point of the training set.	45
Figure 4.4. Sequence features with significant contribution in the overall performance of the prediction model.	45
Figure 5.1. Graphical representation of genetic and chemical means of synthetic lethal interactions.	50
Figure 5.2. Frequently mutated genes across a cohort of patients suffering from breast invasive carcinoma.....	53

List of tables

Table 2.1. List of computational resources available for designing sgRNAs.....	19
Table 3.1. Genomic sequences targeted by 10 sgRNAs used to evaluate the performance of <i>CRISPCut</i>	32
Table 3.2. Number of off-targets predicted to be accessible to CRISPR under <i>in vivo</i> conditions.....	35
Table 4.1. Accuracy of prediction of <i>positive</i> off-targets in training/testing split data and validation data for supervised learning algorithms.	42
Table 4.2. List of sequence features contributing in accurate prediction of <i>positive</i> off-targets.....	43
Table 4.3. Confusion matrix showing the false-positive and false-negative predictions from the developed model after running training split data and validation data.	44
Table 5.1. Cell lines harbouring mutations in PIK3CA and TP53 for analysis of GARP score.	52
Table 5.2. Genes predicted to be in synthetic lethal combination with PIK3CA.	54
Table 5.3. Synthetic lethal gene pairs involving TP53.	55
Table 5.4. sgRNA designs for targeting genes in synthetic lethal combination with PIK3CA and TP53.	56

List of abbreviations

Cas	CRISPR associated protein
CDS	Protein coding sequence
<i>CRISPCut</i>	sgRNA design tool developed as a part of this thesis
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPRa	CRISPR activation
CRISPRi	CRISPR interference
crRNA	CRISPR RNA
dCas9	Catalytically dead Cas9
DSB	Double strand break
GARP	Gene Activity Rank Profile
HDR	Homology directed repair
iPSCs	Induced pluripotent stem cells
LINE	Long interspersed nuclear elements
LTR	Long terminal repeats
MMEJ	Microhomology-mediated end joining
NCBI	National Center for Biotechnology Information
NHEJ	Non-homologous end joining
PAM	Protospacer adjacent motif
RNP	Ribonucleoproteins
sgRNA	Single guide RNA
SINE	Short interspersed nuclear elements
SpCas9	<i>Streptococcus pyogenes</i> Cas9
SQL	Structured Query Language
TALE	Transcription activator-like effectors
tracrRNA	Transactivating CRISPR RNA
UTR	Untranslated region
ZFN	Zinc finger nuclease