Predicting Off-Target Effects of CRISPR-Cas9

Vicky Li[1,2], Advisor: Manuel Lladser[1,2], Co-advisor: Hubert Yin[1,3]
[1]BioFrontiers Institute/IQ Biology, [2]CU-Boulder Applied Math, [3]CU-Boulder Chemistry and Biochemistry

This work was supported in part by the Interdisciplinary Quantitative Biology (IQ Biology) program at the BioFrontiers Institute, University of Colorado, Boulder. IQ Biology is generously supported by NSF IGERT grant number 1144807.

What is CRISPR-Cas9?

CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR associated nuclease) is a bacterial adaptive immune system that is being exploited for genome modification since 2012. It is highly customizable, specific, irreversible and applicable to the three domains of life plus viruses. Proposed applications include gene knockout, genetic engineering, and gene therapy. However, off-target mutations (errors) are known to occur. We wish to predict where and how often off-target mutations occur.

In bacterial genomes, CRISPR loci integrate foreign DNA (protospacers) from viruses and bacteriophages, interspaced by short palindromic repeats. These are transcribed and processed into CRISPR RNA (crRNA), consisting of a single protospacer and enclosing repeats. The protospacer is a guide RNA (gRNA) that associates with Cas, binds, and cleaves foreign DNA complementary to the gRNA and adjacent to a protospacer adjacent motif (PAM). There are three main types of CRISPR-Cas, and many subtypes that vary depending on the species they originate from. We model the nuclease activity of Type II CRISPR-Cas derived from *Streptococcus pyogenes* (CRISPR-Cas9), the most well studied CRISPR-Cas system.

In genome modification, CRISPR-Cas9 consists of Cas9 nuclease, single-guided RNA (sgRNA) composed of gRNA and scaffold, and template DNA. The gRNA sequences determines the target DNA. The PAM is determined by the motif 5'-NRG-3' on the target strand, and the gRNA binds the target complement strand. Since NRG sequences occur on average every 4 base pairs (bp) in humans, this poses nearly no restriction on targetable DNA. Following cleavage by Cas, DNA repair mechanisms either substitute the target with the template via Homology Directed Repair (HDR) or—in absence of template—mediate an indel mutation via Non-Homologous End Joining (NHEJ). The result is genomewide modification of the target sequence.

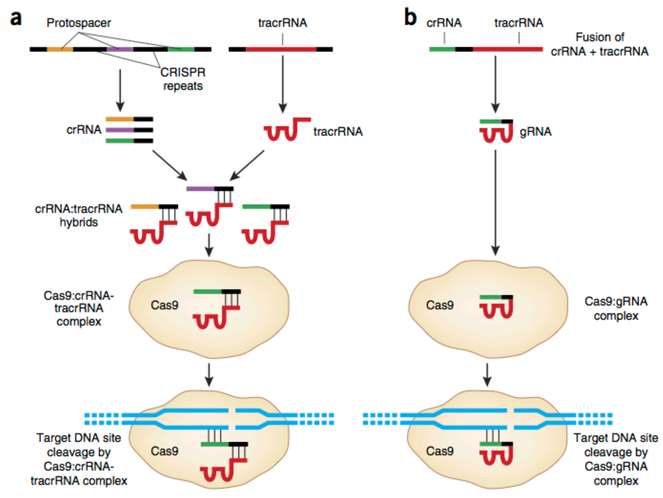


Figure 1: (a) CRISPR-Cas9 in bacterial adaptive immune system and **(b)** genome modification. [Sander, J.D. & Joung, J.K. (2014) CRISPR-Cas systems for genome editing, regulation and targeting. *Nat. Biotech.* **32**(4), 347-355]

Simple Model

1. gRNA recognizes valid PAM
2. binds (off)-target 1 nt at a time
3. successful binding = mutation

The log probability of successful binding (mutation) is:

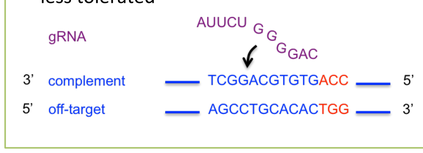
$$\ln P(\text{mutation} \mid \text{PAM}) = -s \cdot \sum_{r=\text{run of MM's}} \text{length}\,(r) - w \cdot \sum_{m=\text{MM}} \delta^{\text{distance}\,(m,\text{PAM})}$$

These factors determine successful binding:

1. Position - PAM-distal mismatches (MMs) are more tolerated i.e. lead more often to successful binding
2. Number – More MMs are less tolerated
3. Spacing – More consecutive MMs are less tolerated

Where:

$s \in (0,\infty)$ punishes consecutive MMs
$w \in (0,\infty)$ determines binding probability at $n = 0$
$\delta \in (0,1)$ punishes PAM-proximity



Preliminary Results

Fit model to synthetic data

Least squares regression determines parameters θ = (s, w, δ) from data (y, x):

$$y = \begin{pmatrix} l_1 \\ \vdots \\ l_N \end{pmatrix} \qquad x = \begin{pmatrix} u_t & m_{t,0} & m_{t,1} & \dots & m_{t,len-1} & 0 & \dots \\ \vdots & & & & & & \end{pmatrix}$$

$$l(x) = \begin{pmatrix} -u_1 s - w \sum m_{1,n} \delta^n \\ \vdots \\ -u_N s - w \sum m_{N,n} \delta^n \end{pmatrix}$$

Minimize SSE: $||y - l(x)||^2$

Example: $\theta = (s = 2, w = 0.333, \delta = 0.5)$

← $\text{PAM}_1 = 00010AGG$
← $\text{PAM}_2 = 00011AGG$
← $\text{PAM}_3 = 00110AGG$
← $\text{PAM}_4 = 00111AGG$
← $\text{PAM}_5 = 01110AGG$
← $\text{PAM}_6 = 01111AGG$
← $\text{PAM}_7 = 10010AGG$
← $\text{PAM}_8 = 10011AGG$
← $\text{PAM}_9 = 10110AGG$
← $\text{PAM}_{10} = 10111AGG$
← $\text{PAM}_{11} = 11010AGG$
← $\text{PAM}_{12} = 11011AGG$
← $\text{PAM}_{13} = 11110AGG$
← $\text{PAM}_{14} = 11111AGG$

Pseudo-data, error~N(0,0.1)

$$y = \begin{pmatrix} -6.59060083 \\ -6.23249472 \\ -2.32210308 \\ -1.84208075 \\ -0.39347391 \\ 0.10926778 \\ -2.38921238 \\ -2.01488366 \\ -0.31789753 \\ -0.10061098 \\ -0.200908 \\ -0.12668679 \\ -0.44859028 \\ 0.09763415 \end{pmatrix}$$

starting guess $\theta_0 = (-10, -10, -10)$
regression estimate
$\hat{\theta} = (2.01177801, 0.30898023, 0.42639936)$

Fit model to experimental data

First systematic mammalian study: Hsu, P.D., Zhang, F. et al. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotech.* **31**(9), 827-832. gRNA (hence targeted DNA) was varied and occurrence of indels signified (off)-target mutations. They measured R = total reads, n = reads w/ observed indels, q = fraction of negative control reads w/ indels, and calculated p = mutation frequency = MLE of a binomial error model.

System – *S. pyogenes* CRISPR-Cas9 (wt)
Cell – HEK293FT
Gene – human EMX1, 15 targets
sgRNA – sgRNA(+85)
Dosage – equimolar sgRNA cassette, Cas9 plasmid
Replicates – 2 biological replicates
Sequencing – Illumina MiSeq, target regions
Filtering – average Q scores, barcode /forward primer matches, Smith-Waterman alignment, indel Q scores

Since p, q, and R were reported but not n, we attempted to reproduce their p estimates by solving the inverse MLE problem. We successfully reproduced most p estimates with error ~10^{-5}, but several had minimal error 0.5 or more, up to 0.83.

p	n1	error1-1	error1-2	n2	error2-1	error2-2
0.13190476	[25595.]	[0.81541]	[0.81541]	[6373. 2]	[1.56889091e-05	[1.95944160e-05
0.13395092	[32369.]	[0.82398]	[0.82398]	[6714. 2]	[1.66231372e-05	[1.71389750e-05
0.11247965	[13963. 7]	[1.4846]	[1.42487]	[6926. 3]	[1.03634088e-05	[1.95964450e-05
0.11570527	[6191. 32]	[1.3784]	[2.21554]	[6027. 2]	[2.93342392e-05	[5.57099309e-06
0.11806734	[4654. 24]	[3.4233]	[1.40221]	[29719.]	[0.83068848]	[0.83068848]
0.11957842	[5756. 30]	[3.8367]	[8.77115]	[7308. 3]	[1.03371972e-05	[1.89256712e-05
0.1212302	[5858. 24]	[2.8662]	[1.0147]	[26164.]	[0.82092275]	[0.82092275]
0.10423285	[5365. 26]	[2.8239]	[1.13516]	[3744. 19]	[4.16796053e-05	[1.17791779e-05
0.12283053	[262.]	[0.00031]	[0.00031]	[7418. 3]	[1.83895519e-05	[1.08305632e-05
0.11148345	[3916. 39]	[0.11148]	[0.11148]	[7116. 3]	[2.06968271e-05	[8.33749151e-06

Mechanism and Structure

Biophysical studies have determined the structure and mechanism by which gRNA-Cas9 binds target DNA. However, as of 2015, the mechanism of Cas9 nuclease activation remains unknown.

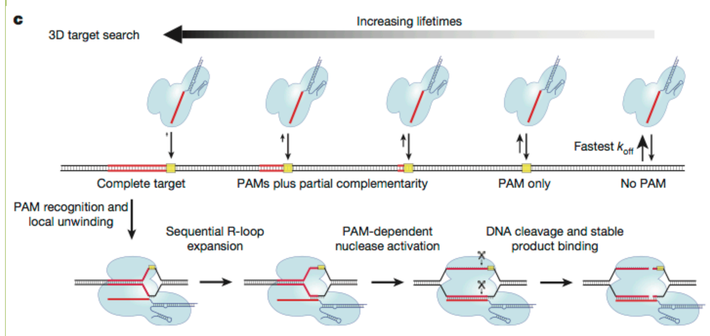


Figure 2: Experimentally verified model for CRISPR-Cas9 mechanism. [Sternberg, S.H., Doudna, J.A. et al. (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62-67]

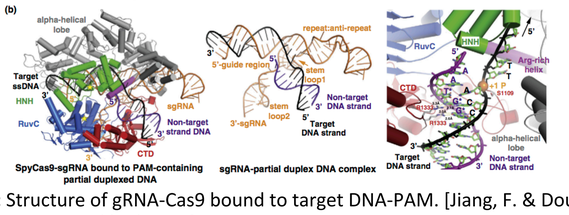


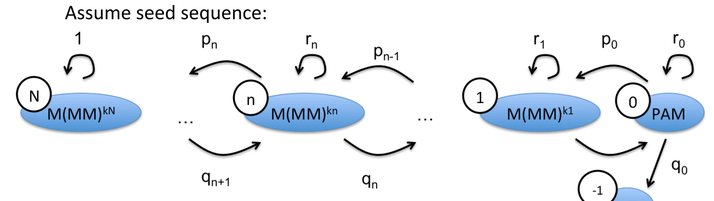
Figure 3: Structure of gRNA-Cas9 bound to target DNA-PAM. [Jiang, F. & Doudna, J.A. (2015) The structural biology of CRISPR-Cas systems. *Curr Opinion Struc Bio* **30**: 100-111]

Question: What is the mechanism of Cas9 nuclease activation?

Hypothesis: Cas9 cleaves (off)-target DNA upon recognition of the T-shape of a seed sequence of some length bound to sgRNA-Cas9. A mathematical model of this should successfully reproduce prokaryotic (off)-target data. Differences in eukaryotic and prokaryotic data may be due to less accessible DNA.

Markov Model

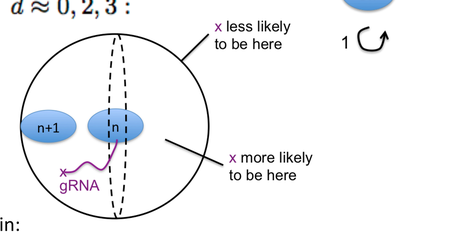
Assume seed sequence:



$$p_n = \frac{\alpha\beta}{(k_{n+1}+1)^d} \qquad d \approx 0, 2, 3:$$

$$q_{n\neq 0} = 1-\beta,\ q_0 = \gamma$$

$$r_n = 1 - p_n - q_n$$



Solve birth-death Markov chain:

$$P(\text{mutation} \mid \text{PAM}) = P(\text{M(MM)}^{k_N}) = h_0$$

$$h_n = P(n \to N) \qquad u_n = h_{n+1} - h_n \quad (1) \qquad \begin{cases} h_N = 1 \\ h_n = p_n h_{n+1} + q_n h_{n-1} + r_n h_n = \frac{p_n}{1-r_n} h_{n+1} + \frac{q_n}{1-r_n} h_{n-1} \\ h_{-1} = 0 \end{cases}$$

$$h_n = \left(\frac{p_n}{1-r_n} + \frac{q_n}{1-r_n}\right) h_n = \frac{p_n}{1-r_n} h_{n+1} + \frac{q_n}{1-r_n} h_{n-1} \to \frac{p_n}{1-r_n} u_n = \frac{q_n}{1-r_n} u_{n-1}$$

$$u_{n-1} = \frac{p_n}{q_n} u_n \to u_{n-1} = \gamma_n u_{N-1} \qquad \gamma_n = \Pi_n^{N-1} \frac{p_i}{q_i},\ \gamma_N = 1 \quad (2)$$

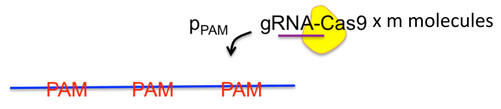
$$\sum_n^{N-1} u_i = h_N - h_n \to h_n = 1 - u_{N-1} \sum_{n+1}^{N} \gamma_i \quad (3)$$

Eqns 1,2 $\to h_0 = u_{-1} = \gamma_0 u_{N-1}$ Eqn 3 $\to h_0 = 1 - u_{N-1} \sum_1^N \gamma_i$

$$u_{N-1} = \frac{1}{\sum_0^N \gamma_j} \qquad \therefore h_n = \begin{cases} 1 - \sum_{n+1}^N \gamma_j / \sum_0^N \gamma_j = \sum_0^n \gamma_j / \sum_0^N \gamma_j & n = 0, 1, \dots, N-1 \\ h_N = 1 \\ h_{-1} = 0 \end{cases}$$

Model Prediction

Given:



What is the probability of mutating any one (off)-target?

$$p = P(\text{mutate (off)-target}) = mCp_{\text{PAM}} \cdot P(\text{mutation} \mid \text{PAM})$$

Where:

$$C = \sum_{i=0}^{N-1} b^i p_{\text{PAM}}$$

$N =$ max interrogations by CRISPR-Cas in time studied

$b = P(1 \text{ interrogation results in no mutation})$

$$p_{PAM} = \begin{cases} p_{NGG} & \text{PAM} = \text{NGG} \\ p_{NAG} & \text{PAM} = \text{NAG} \end{cases} \qquad (\#\ \text{NGG}) p_{NGG} + (\#\ \text{NAG}) p_{NAG} = 1 \qquad p_{NAG} = \frac{1}{5} p_{NGG}$$

A Method to Analyze Off-Target Data

Example

Simple Model
Hsu et al. Fig. 2 Data

Human genome GRCh38.p4
(1 copy of each chromosome):

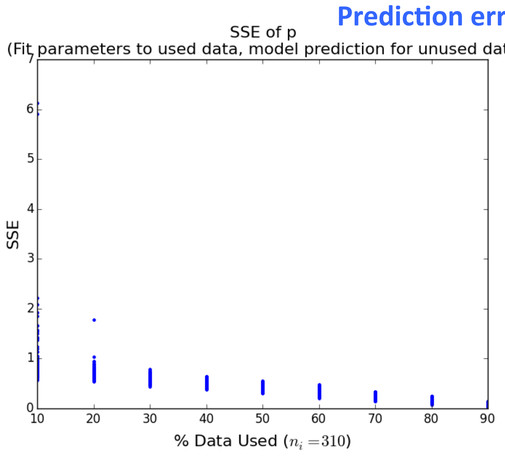
	# NAGs	# NGGs
autosomes	380,034,908	269,822,755
X	20,843,230	14,166,478
Y	3,535,774	2,474,436

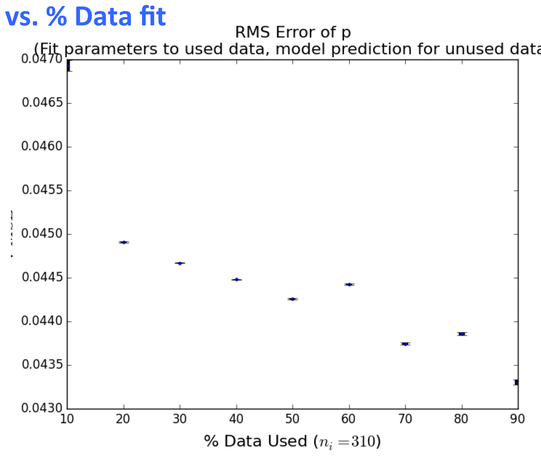
(2 copies, average XX and XY):

# NAGs	# NGGs
739,102,548	562,132,445

$$p_{PAM} = \begin{cases} p_{NGG} = 1.39 \times 10^{-9} \\ p_{NAG} = 2.77 \times 10^{-10} \end{cases}$$

Prediction error vs. % Data fit





Parameter estimate vs. Target

	EMX1.1	EMX1.2	EMX1.3	EMX1.6
C	1.26e+08	1.19e+08	7.50e+07	5.31e+07
s	6.00	6.00	6.00	6.00
w	7.28e-08	1.21e-08	1.61e-08	7.37e-07
delta	5.80	5.78	5.78	5.78
n	109	111	104	105
power	0.54	0.55	0.52	0.53

Power Analysis

Independent 2-sample 1-tailed t test
d = 0.2, α = 0.05, power = 0.8
Requires sample sizes n = 310

Power Analysis

1-sample 2-tailed t test
d = 0.2, α = 0.05, power = 0.8
Requires sample size n = 198

Future Plans

- Apply to prokaryotic data
- Eukaryotic data is confounded by DNA inaccessibility! (methylation, nucleosomes)
- Apply to Markov Model