Cas9 Off-target Analysis Model.

Content:

1. Overview

This model aims at predicting the off-target rate of any Cas9-based system in vivo. It has the following key ideas.

$$\text{Cas9+DNA} \xleftrightarrow{\text{Reversible binding}} \text{Cas9-DNA} \xrightarrow{\text{Irreversible enzymatic reaction}} \text{Double strand break DNA + Cas9}$$

The Cas9 cleaving process is divided into two separated reactions - the reversible binding reaction and the irreversible cleaving reaction.

First, The probability of Cas9-DNA binding is majorly determined by the affinity of the gRNA and DNA. A Δ G' is assumed to indicate this affinity. The Δ G' is determined by Δ G(i), which is calculated by NN nearest neighbor model of nucleic acid thermodynamics.

Second, by analyzing binding equilibrium, dCas9 inhibition data and aCas9 activation data, the model to predict the possibility of gRNA-d/aCas9 binding to certain target in vivo can be constructed. The fitting result of this model reveals the equation to calculate Δ G' from Δ G(i).

Finally, By analyzing the Cas9 cleaving process, the link between Cas9-DNA binding probability and editing efficiency can be established.

Data from six papers were analysis and/or used for model fitting. The data for Cas9 editing model fitting is generously provided by Vikram Pattanayak and Prof. David Liu, who has published the paper - High-throughput profiling of off-target DNA cleavage reveals RNA- programmed Cas9 nuclease specificity - on Nature Biotechnology, 11 Aug 2013.[1] The data for Cas9 binding model derivation and fitting is extracted from the following figures, Fig 2C, S7B, S7E of [2], Fig 5C of [3], Fig 2AB of [4]. The software used to extract high fidelity data is GetData Graph Digitizer V2.22.

2. Symbol table, Assumption and reasons.

Symbol	
[ ]	The symbol of concentration, i.e. [A] means the concentration of A
Δ G'	Difference in Modified Gibbs Free Energy. It's assumed to determine the binding constant between gRNA-Cas9 and DNA
Δ G(i)	The calculated Δ G for the i^{th} position of gRNA-DNA interaction
a	The input of Δ G', a vector consist of Δ G'(1)-(21)
b	The constant representing all interaction in the binding process other than the gRNA-DNA interaction.

ω	The weight vector
F()	Relation function
Ka	Association constant of gRNA-Cas9 and DNA
Kd	Dissociation constant of gRNA-Cas9 and DNA
$[A]_0$	the concentration of certain sequence in the pre-selection library,
$[A]_{tot}$	the concentration of all DNA sequence in the pre-selection library
[C]	the concentration of certain sequence in the post-selection library
$[C]_{tot}$	the concentration of all DNA sequence in the post-selection library
A'	the number of certain sequence we sampled from the pre-selection library
A_{tot}'	the number of all sequence we sampled from the pre-selection library
P'	the number of certain sequence we sampled from the post-selection library
P_{tot}'	the number of all sequence we sampled from the post-selection library
θ	Cas9 targeting efficiency
S	Substrate, DNA
E	Enzyme, gRNA-Cas9
P	Product, double strands broken DNA
A	The intact DNA duplex
B	the DNA molecule in which one of the two strands has been cleaved at the recognition site for the restriction enzyme
C	the DNA molecules in which both strands have been cleaved at the recognition site
k_a,k_b	The two apparent first-order reaction constant of the two steps of cleaving of Cas9
k_1,k_{-1}, k_{cat}	Reaction constants
K_M	MM constant
R	Gas constant
T	Absolute temperature
p_b	Binding probability
p_c	Cutting probablity
Abbreviation	
dCas9	Deactivated Cas9, Cas9 inhibitor, a Cas9 with two mutations D10A and H841A
aCas9	Cas9 activator, a dCas9 that fused with a activator domain like VP64, TAL and omega subunit of RNAP.
d/aCas9	Deactivated Cas9, no matter whether it's an activator or inhibitor

Table 1. Symbol table of Cas9Off Model

1. As Cas9 need the guiding of gRNA to cut DNA, the unbounded gRNA and Cas9 are ignored in the analysis, and other gRNA and Cas9 are considered to constantly bind to each other.
2. The model does not take the 3D structure of DNA, gRNA and DNA-gRNA complex into consideration. As the data is not sufficient to take these factor into consideration.
3. The model is based on NN nearest neighbor model of base pairing energy[5]. This model was built for thermodynamic energy calculation of DNA strand interaction. But we employ it to model the gRNA-DNA interaction. This will bring in some inherent flaws. The most prominent one will be when the RNA side is a U and the DNA side is a G. In the NN model it's considered as a T-G pair, which is not as energetically

favorable as U-G [6]. However, there is no model available for DNA-RNA interaction energy calculation yet. So it's assumed that the energy (Δ G(i)) calculated from the NN model is to some degree consistent with reality. In fact, [7] suggested a rough sort of the tolerance of base mismatch: CC<UC<AG<AA<GA<CA<UG<CT<GG<UT <AC<GT, while the model suggested that CC<AC<TC<AA<TT<GA<GT<GG.

4. We believe by employing a better model of energy prediction, the whole Cas9 off-target model will be improved.

5. We assume Δ G' takes up a form of $\Delta G' = F(\vec{\omega} \bullet \vec{a} + b)$. Where "a" is an 1×21 vector that contain Δ G(1) to Δ G(21) as its value, ω is the weight vector. Only the impact of DNA-gRNA interaction ("a") is counting as a variable, and the Δ G contributed by other interaction(eg. protein-DNA interaction) are considered as a constant b. This is also why this model cannot predict Cas9 off-target rate of a target without PAM(NGG), which interact with Cas9 rather than gRNA. F() is the function that relate $\vec{\omega} \bullet \vec{a} + b$ with Δ G'.

6. Both cleaving steps of Cas9 are assumed as classic Michaelis-Menten enzyme reaction.

7. The dCas9, aCas9 and normal Cas9 are assumed to share a same K_a with DNA, given they are guided by the same gRNA. This is reasonable as the only changes in Cas9 are D10A and H841A.

3. Modeling result

We employ a NN nearest neighbor model to calculate the Δ G(i) between gRNA and DNA on each NN position. From the 1st nucleotide of the target area of gRNA to the 20th, Δ G(i) of totally 19 position are calculated. The feasibility of our idea is proved by calculating the correlation between Δ G(i) and cutting efficiency (employing data from [1] CLTA1,2,3).

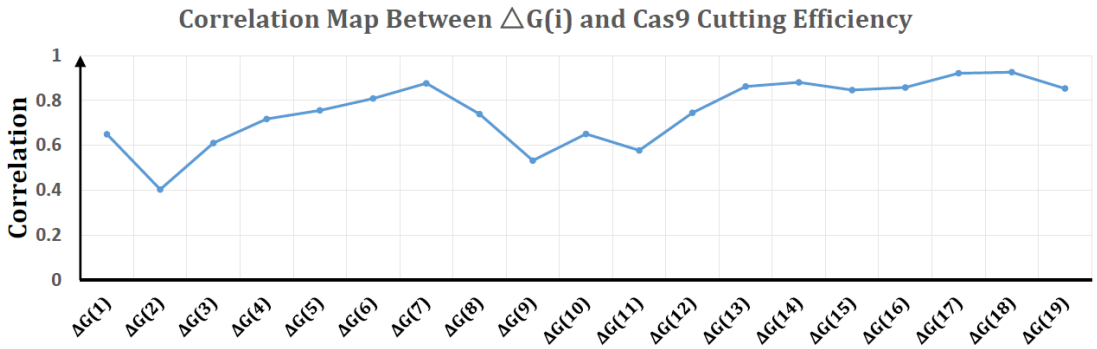

Figure 1. Correlation map between Δ G(i) and Cas9 cutting efficiency

The result shows that roughly the closer the position to PAM the larger the correlation. This discovery is consistent with previous studies [1,2,3,4,7,8]. Therefore we confirm that Δ G(i) do influence the targeting efficiency of Cas9.

But the data from [1,2,3,4,8] also revealed that Δ G(i) is not proportional to targeting efficiency. In most high single mismatch tolerance cases, the correlation between Δ G(i) and targeting efficiency is not significant. The following table can be concluded.

Sequence	Single Mismatch tolerance	G/C	Ref.

TCATGCTGTTTCATATGATC	low	7	[4]
AACTTTCAGTTTAGCGGUCU	low	8	[3]
TGTGAAGAGCTTCACTGAGT	low	9	[1]
GATGCCGTTCTTCTGCTTGT	low	10	[8]
AGTCCTCATCTCCCTCAAGC	low	10	[1]
GAGATGATCGCCCCTTCTTC	low	11	[2]
CTCCCTCAAGCAGGCCCCGC	low	15	[1]
		Ave. G/C 10	
GCAGATGTAGTGTTTCCACA	high	9	[1]
GGTGGTGCAGATGAACTTCA	high	10	[8]
GGGGCCACTAGGGACAGGAT	high	13	[2]
GTCCCCTCCACCCCACAGTG	high	14	[2]
GGGCACGGGCAGCTTGCCGG	high	16	[8]
		Ave. G/C 12.4	

Table 2. The relation of G/C frequence and single mismatch tolerance

The mismatch tolerance is roughly determined from the data of the references, for details please click here. A low tolerance sequence with single mismatch on at least 7 positions has significant performance drop. A high tolerance sequence with single mismatch at more than 16 position can perform as well as the original sequence in guiding Cas9.

The relationship of G/C frequence with single mismatch tolerance can be explained by the fact that the abundance in G/C make the gRNA binds to DNA more stable, and single mismatch is not strong enough to disturb the binding. This suggests that the F() may be a reversed sigmoid function. But to determine this sigmoid function (and to determine b) requires more specific experiment data of d/aCas9 binding kinetics, which is not available.

This guess is also supported by later analysis in the comparison of the Cas9 binding model, which assumed a normal proportional relationship between Δ G(i) and Δ G'. The results shows that the Cas9-gRNA not only is not sensitive to energy change in DNA-gRNA binding when Δ G' surpass some threshold, but also not sensitive to such energy flux when Δ G' is lower than some threshold.

A rough ω is calculated from Fig S7C and Fig.5D of [3] (both cases are intolerant to single mismatch).

$$\omega = [0.21,\ 0.25,\ 0.30,\ 0.39,\ 0.36,\ 0.32,\ 0.35,\ 0.39,\ 1.04,\ 1.19,$$
$$1.20,\ 1.05,\ 1.22,\ 2.80,\ 1.83,\ 1.92,\ 2.30,\ 2.36,\ 2.09\ ]$$

The performance of these parameters and the model are checked by compared the predicted value with the data from Fig.2B of [4] and Fig.5C of [3]. Noticed that all the following figure using data normalize by the activity of "wildtype" gRNA, thus the b is not required for the prediction.

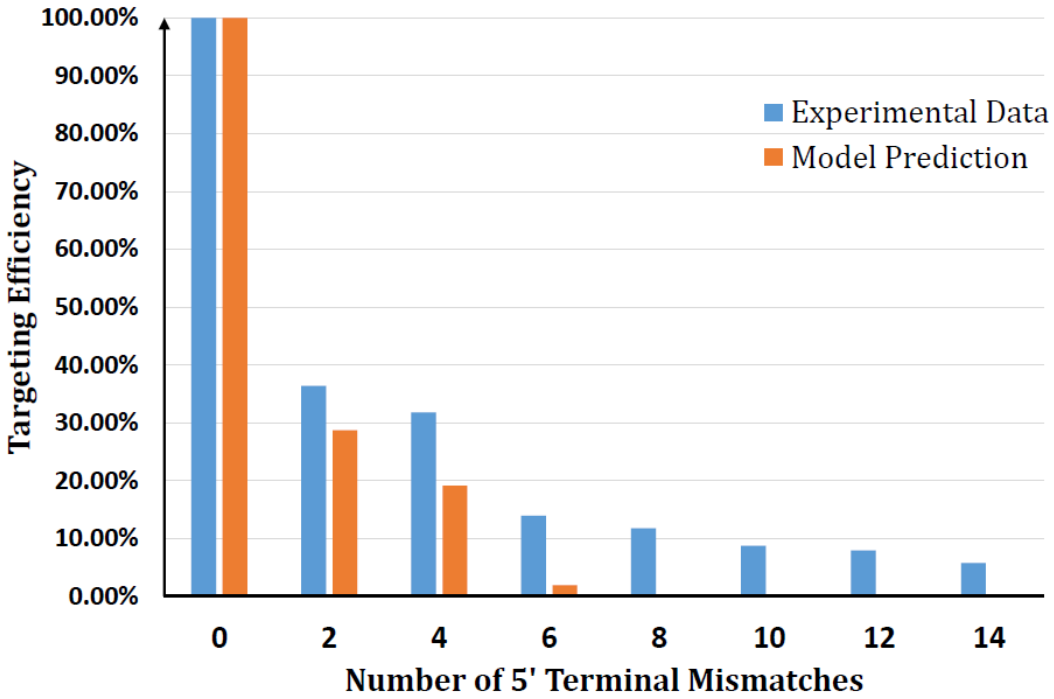

Figure 2. Model prediction compared with data from Fig.2B of [4]

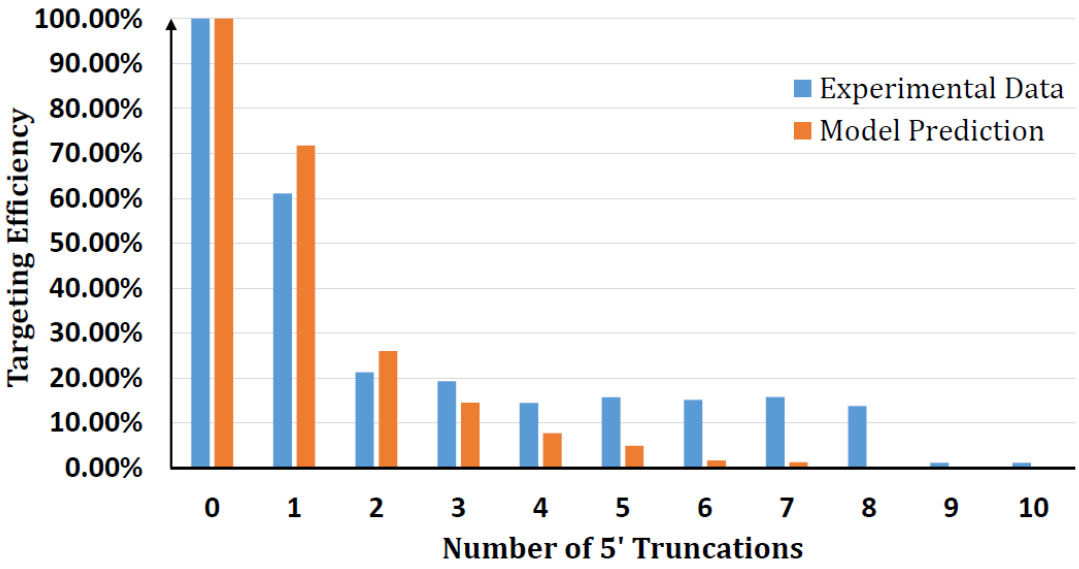

Figure 3. Model prediction compared with data from Fig.5CB of [3]

These data are collected from 1’ end truncation or consecutive mutation experiment of gRNA. In Both figure, as the column number grows, the end truncation/end mutations become more serious, and the total energy of DNA-gRNA binding drops. The data show great non-lineality. Obvious platforms formed in the 4-8 column of Fig.2 and column 3-9 of Fig.3, which suggest the gRNA-Cas9 complex is not sensitive for the energy loss cause by the continuous mismatch / truncation at these stage.

On the las part of the model. Kinetic analysis reveals that both concentration and reaction time are important for off-target control.

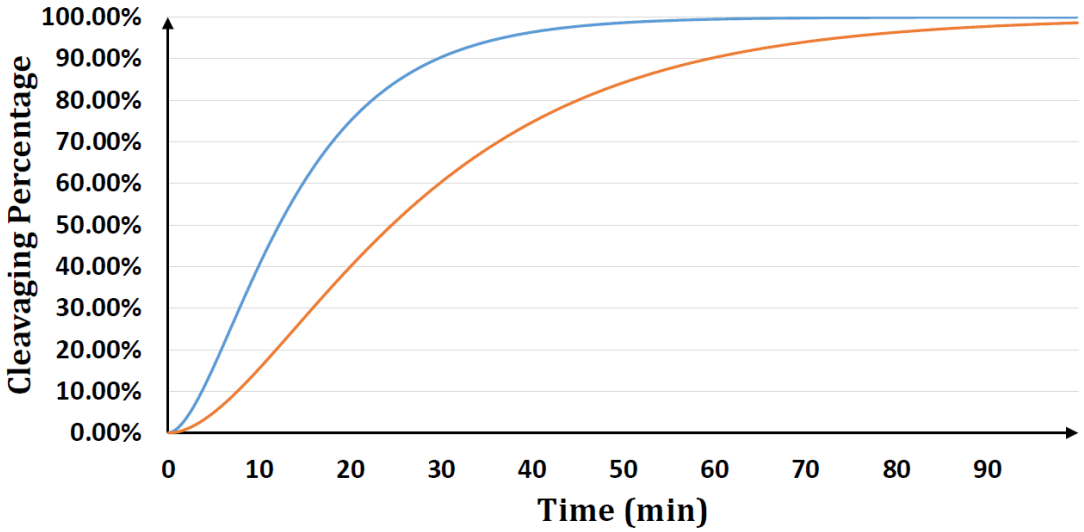

Figure 4. Theoretical curves from the Cas9 cleaving reaction
The curves displaying changes of two different cleaved products. Boundary conditions were set as $[A_0]=1.0$, $[B_0]=[C_0]=0$, $k_a=0.2\ min^{-1}$,$k_b=0.1\ min^{-1}$ for red line; And $[A_0]=1.0$, $[B_0]=[C_0]=0$, $k_a=0.1\ min^{-1}$,$k_b=0.05\ min^{-1}$ for blue line.

4. Model derivation

4.1. Calculation of Δ G' of DNA-gRNA binding

The calculation method of Δ G(i) and Δ G' is modified from the NN nearest neighbor model introduced in [2].

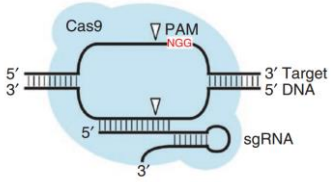

Figure 5. schematic picture of Cas9 digestion, modified from [1]

Step1. Set up the binding sequence
The input will be the 21nt of the target prior to the GG of the PAM, and the corresponding 21nt of the potential off-target sequence. The reason for why we need a 21nt sequence rather than 20nt is that the NN model using the adjacent 2nt as inputs. In order to completely consider the impact of the 20nt targeting sequence of gRNA, we need to consider the 21st base to make the calculation comprehensive. Hereby we explain our way to process inputs using an example. Mismatch base pairs are highlighted in red.

Example:
Target sequence and gRNA sequence: ATCG.....CCGG (20nt)
Possible off-target sequence: ACCG.....CGGG (20nt)
Change the off-target sequence to its complementary sequence:
TGGC.....GCCC (20nt)
The binding double strand will be:
ATCG.....CCGG (G) gRNA
TGGC.....GCCC (A) potential off-target DNA
The base in brackets is the 21st nucleotide on each chain. Notice that the 21st nucleotide of gRNA is always G.

We then divide the chains into the following form

(AT)terminal+ (GG) + CG +..... CC + CC + GG terminal
(TG)terminal+ (CT) + GC +..... GC + GC + CC terminal
: : : : : :
: : : : : :
Δ G(1), Δ G(2), Δ G(3) Δ G(17),Δ G(18),Δ G(19)

Step2. Terminal energy calculation

Determine the first mismatch from both direction of the chain. If the mismatch happen within 2nt from end (i.e. at position 1,2), consider the corresponding end as an dangling end. There is a reason for only consider terminal mismatch and dangling end effect on the "1' end". These terminal stabilizing effect originate from the fact that if the two chains are not suitable to bind at the terminal, they can simply not bind in the classic way, which is energetically unfavorable, but just floating around. But the Cas9 is "grasping" at the "20' end" of gRNA and DNA binding, as the protein needs to anchor on the PAM immediately following the "20' end". This spatial constraint make the gRNA and DNA has no other way but the "normal" way of binding. Therefore, we use the energetically unfavorable single mismatch table(in later steps) to calculate the energy here, rather than the relatively more stabilizing dangling end table.

Therefore, in the example, we consider the left end as a dangling end, the right end as a normal end.

If a dangling end is determined, determine the first match position following the mismatch position. In the example, this will be position 2 (△G(2)). Set all dangling end position energy as 0, i.e. △G(1)=0, and calculate the first match according to Table 3, i.e. △G(2)=5'TC/G+3'GG/C=-0.58-0.44=-1.02 kcal/mol,

Dangling end sequence	ΔG_{37}^{0}			
	X=A	X=C	X=G	X=T
5'-dangling ends				
X~A ‖ T	-0.51	-0.42	-0.62	-0.71
X~C ‖ G	-0.96	-0.52	-0.72	-0.58
X~G ‖ C	-0.58	-0.34	-0.56	-0.61
X~T ‖ A	-0.50	-0.02	0.48	-0.10
3'-dangling ends				
A~X ‖ T	-0.12	0.28	-0.01	0.13
C~X ‖ G	-0.82	-0.31	-0.01	-0.52
G~X ‖ C	-0.92	-0.23	-0.44	-0.35
T~X ‖ A	-0.48	-0.19	-0.50	-0.29

Table 3. Nearest-neighbor model for terminal dangling ends next to Watson-Crick pairs in 1 M NaCl, modified from Table 3 of [5]

If no dangling end appears. Determine whether the terminal pair is A-T. If yes, add a terminal AT penalty(+0.05) to the △G(i), and calculate all △G(i) according to Table 3.

Step3. Internal energy calculation

Calculate all position except for first match and dangling end position according to Table 4, in our example, this set contains ∆ G(3) to ∆ G(19).

The result will be

∆ G(3)=-2.17 kcal/mol

∆ G(17)=0.70 kcal/mol

∆ G(18)=0.70 kcal/mol

∆ G(19)=-1.84 kcal/mol

Propagation sequence	X	Y			
		A	C	G	T
G ~ X ‖ ‖ C ~ Y	A	0.17	0.81	-0.25	-1.30
	C	0.47	0.79	-2.24	0.62
	G	-0.52	-1.84	-1.11	0.08
	T	-1.44	0.98	-0.59	0.45
C ~ X ‖ ‖ G ~ Y	A	0.43	0.75	0.03	-1.45
	C	0.79	0.70	-1.84	0.62
	G	0.11	-2.17	-0.11	-0.47
	T	-1.28	0.40	-0.32	-0.12
A ~ X ‖ ‖ T ~ Y	A	0.61	0.88	0.14	-1.00
	C	0.77	1.33	-1.44	0.64
	G	0.02	-1.28	-0.13	0.71
	T	-0.88	0.73	0.07	0.69
T ~ X ‖ ‖ A ~ Y	A	0.69	0.92	0.42	-0.58
	C	1.33	1.05	-1.30	0.97
	G	0.74	-1.45	0.44	0.43
	T	-1.00	0.75	0.34	0.68

Table 4.Nearest-neighbor model, modified from Table 2 of [5]

Step4. Further analysis of internal loops and bulges.
We will complete this step in the future. For the model V1.0, the algorithm will skip this step.

Step5. Adjust △G(i) according to ion concentration
Empirical salt correction equations have been derived,

$$\Delta G(i)[Na^+] = \Delta G(i)[1\mathrm{M}\,NaCl] - 0.114 \times N / 2 \times \ln[Na^+]$$

where N is the total number of phosphates in the duplex, and [Na^+] is the total concentration of monovalent cations from all sources (the same equation works for sodium, potassium, and ammonium)over a range of monovalent concentration of 0.05 to1M.

Step6. Calculate △G' We assume Δ G' takes up a form of $\Delta G' = \vec{\omega} \bullet \vec{a} + b$. Where "a" is an 1×19 vector that contain Δ G(1) to Δ G(19) as its value, ω is the weight vector. Only the impact of DNA-gRNA interaction ("a") is counting as a variable, and the Δ G contributed by other interaction(eg. protein-DNA interaction) are considered as a constant b. This is also why this model cannot predict Cas9 off-target rate of a target without PAM(NGG), which interact with Cas9 rather than gRNA. (Assumption 4)

According to the previous steps, the △G' of our example should be

$$\Delta G' = F(\vec{\omega} \cdot \vec{a} + b) = F(\vec{\omega} \cdot [0, -1.02, -2.17 \ldots\ldots 0.70, 0.70, -1.84]^T + b)$$

4.2. Correlations between Δ G' and Cas9 targeting efficiency

Vikram Pattanayak et al. used *in vitro* selection and high-throughput sequencing to determine the propensity of eight guide-RNA:Cas9 complexes to cleave each of 10^{12} potential off-target DNA sequences. This size is sufficiently large to include tenfold coverage of all sequences with eight or fewer mutations relative to each 22-base-pair target sequence.

$$\text{Pre-selection library} \xrightarrow{\text{Cas9 digestion}} \text{Post-selection library}$$

The DNA of target sequences and their corresponding potential off-target sites were produced as substrates by PCR and rolling circle amplification. The abundance of each kind of sequence in the pre-selection library will differ from their abundance in post-selection library. This abundance changes reveals the relative targeting efficiency of the Cas9 on certain target.

Let us define the following variant.
$[A]_0$ is the concentration of certain sequence in the pre-selection library,
$[A]_{tot}$ is the concentration of all DNA sequence in the post-selection library
[C] is the concentration of certain sequence in the post-selection library,
$[C]_{tot}$ is the concentration of all DNA sequence in the post-selection library
A' is the number of certain sequence we sampled from the pre-selection library
A_{tot}' is the number of all sequence we sampled from the pre-selection library
P' is the number of certain sequence we sampled from the post-selection library
P_{tot}' is the number of all sequence we sampled from the post-selection library

So we have

$$\frac{A'}{A_{tot}'} \approx \frac{[A_0]}{[A]_{tot}} \quad \frac{P'}{P_{tot}'} \approx \frac{[C]}{[C]_{tot}}$$

And the Cas9 targeting efficiency

$$\theta = \frac{P'}{P_{tot}'} / \frac{A'}{A_{tot}'} = \frac{P' A_{tot}'}{P_{tot}' A'}$$

If Δ G' really determine the probability of Cas9 digest certain DNA. There must be some kind of correlation between each Δ G(i) and the Cas9 targeting efficiency θ. We can calculate the Δ G' and θ of all sequence contained in the library, and calculate the Pearson's product-moment coefficient of Δ G' and θ. So we analyzed CLTA1,2,3 one-mutation pre-selection library and "v2.1 gRNA 100nM Cas9" post-selection library, and get Figure 1.

4.3. Derivation of Cas9 binding model, for off targe prediction of d/aCas9

Cas9 must first binds to DNA to cut them. For d/aCas9,

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES$$

Where E stands for the enzyme - gRNA-Cas9, S the substrate - certain DNA of specific sequence, ES the gRNA-Cas9-DNA complex. We keep calling Cas9 a enzyme for uniformity in this article, though all Cas9 considered in this part(4.3) is deactivated and is actually not an enzyme.

First, we link Δ G' with [S], [E] and [ES] through

$$K_d = \frac{k_{-1}}{k_1} = \frac{[S][E]}{[ES]} \quad \text{and} \quad K_d = e^{\frac{\Delta G'}{RT}}$$

In a living cell at steady state, the protein concentration is usually kept in a constant. In *E*.coli this constant is approximately 1nM[9]. The substrate concentration is also fixed, as certain sequence usually has relative fixed copy number in a cell, especially in prokaryote like *E*.coli. The concentration of certain DNA sequence in a cell is typically

$$\mathrm{DNA}_{vivo} = 1\sim 100/E.\mathrm{coli} \le \frac{100/(6\times 10^{23})\mathrm{Mole}}{1\mu\mathrm{m}^3} \approx 10^{-7}\mathrm{M}$$

$$\therefore [S] \ll K_d$$

$$\therefore [ES] = \frac{[E_0][S]}{K_d + [S]} = \frac{[E_0][S]}{K_d}$$

For Cas9 guided by two different gRNA targeting at the same sequence,

$$\frac{[E_1S]}{[E_2S]} = \frac{K_{d2}}{K_{d1}} = \frac{e^{\frac{\Delta G_2'}{RT}}}{e^{\frac{\Delta G_1'}{RT}}} = e^{\frac{\Delta G_2' - \Delta G_1'}{RT}}$$

Lei et.al. and Bikard et.al. use the inhibitory effect of dCas9 to measure the targeting efficiency of different gRNA [3,4]. The regulated florescence represents the inhibitory effect of dCas9. Prashant et.al. employ aCas9 for the same purpose[2]. But they sequence mRNA to measure the activation of the aCas9 guided by various gRNA. Anyway, the concentration of fluorescence protein and mRNA both obey following ODEs (detailed explanation in our TP model, the equation is the same as the equations in [10]).

$$\mathrm{DNA} \rightarrow \mathrm{RNA} \rightarrow \mathrm{Protein}$$

$$\frac{d[mRNA]}{dt} = \alpha[TF] - \lambda[mRNA]$$

$$\frac{d[protein]}{dt} = \nu[mRNA] - k[protein]$$

These equation can reach steady state quickly when compared with the time scale of any in vivo or in vitro experiment. Because according to the data from [3], Cas9-DNA binding can achieve equilibrium within 0~3 min. The reasoning is as follow.

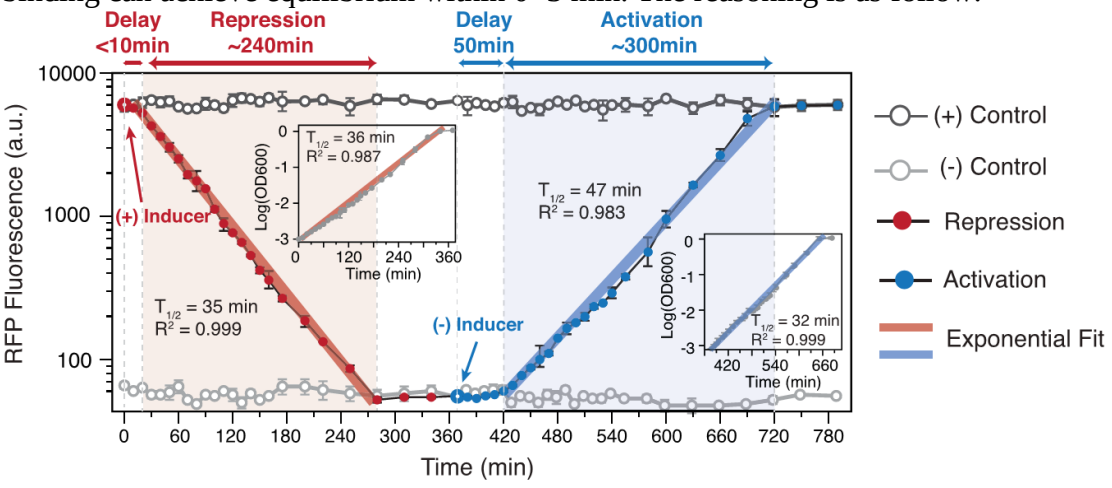

Figure 6. dCas9 regulation on promoter J23119 (extracted from [3])

Notice that, on Figure 5, the RFP started to decrease exponentially 10min after the adding of inducer. This is only possible, when v[mRNA] is hold as a constant. So d[mRNA]/dt=0, which means [TF] is a constant. In this equation, [TF] means the concentration of transcription factor that binding to the promoter, while dCas9 is the only transcription factor in this experiment. According to table 2.1 and 2.2 in [9], the typical mRNA lifetime in *E*.coli is 2-5 min, the time for protein (Cas9) transcription and translation is 5 min. So the Cas9-DNA binding can achieve equilibrium within (10-5-5~10-5-2) 0~3 min in vivo. So the time needed to achieve equilibrium is much shorter than the experiment time-scale both in vivo and in vitro.

So we can consider the equations are in steady state.

$$\frac{d[mRNA]}{dt} = \frac{d[protein]}{dt} = 0$$

$$\therefore [mRNA]_{eq} = \frac{\alpha}{\lambda}[TF]$$

$$\therefore [protein]_{eq} = \frac{\nu\alpha}{\lambda k}[TF]$$

Because dCas9 or aCas9 is the only transcriptional factor for the promoter of measurement, the concentrations of mRNA and fluorescence protein are proportional to the concentration of d/aCas9 binding to the target promoter. So the relative repression or activation activity of d/aCas9 guided by two different gRNA is,

$$\frac{[\text{measurement}_1]}{[\text{measurement}_2]} = \frac{[TF_1]}{[TF_2]} = \frac{[E_1S]}{[E_2S]} = \frac{K_{d2}}{K_{d1}} = \frac{e^{\frac{\Delta G_2'}{RT}}}{e^{\frac{\Delta G_1'}{RT}}} = e^{\frac{\Delta G_2' - \Delta G_1'}{RT}}$$

These data can never give us the exact value of Δ G' as they only indicate the difference between Δ G'. So we can assume that the exact match gRNA result in a Δ G'_{norm}=0 to calculate ω. (The norm shall be reset for every different set of data)

$$\therefore \Delta G_1' = -RT \ln \frac{[\text{measurement}_1]}{[\text{norm}]} = F(\vec{\omega} \cdot \vec{a} + b)$$

Therefore ω can be calculated (See result).

In order to predict the off-target rate of d/aCas9. Following equation can be derived. At equilibrium,

$$K_d = \frac{[S][E]}{[ES]} \quad and \quad [S] \ll K_d$$

$$\therefore [ES] = \frac{[E_0][S]}{K_d + [S]} = \frac{[E_0][S]}{K_d}$$

$$\therefore p_b = \frac{[ES]}{[ES]+[S]} = \frac{[E_0]}{[E_0]+K_d}$$

So at equilibrium, the probability of a substrate binding with a Cas9 is $[E_0]/([E_0]+K_d)$. If we set p_{bw} as the probability of d/aCas9 binding to the wrong target, pbr as the probability of d/aCas9 binding to the right target. The off-target rate will be,

$$\therefore \frac{p_{bw}}{p_{br}} = \frac{[E_0]}{[E_0]+K_{dw}} / \frac{[E_0]}{[E_0]+K_{dr}} = \frac{[E_0]+K_{dw}}{[E_0]+K_{dr}} = \frac{[E_0]+e^{\frac{\Delta G'_w}{RT}}}{[E_0]+e^{\frac{\Delta G'_r}{RT}}}$$

This equation can also be employed to calculate the best enzyme concentration of gRNA-Cas9 for an ideal balance between regulation and off-target.

4.4. Derivation of Cas9 cutting model, for off targe prediction of Cas9

Cas9 contains two nuclease domain - a RuvC-like domain and a HNH motif flanked by two RuvC-like domains. Each of them responsible for cutting one of the two nucleotide chains[11]. The kinetic of endonuclease catalyzed DNA double strand break is very complex. But fortunately, experiments have showed that most double strand break process can be approximated by a consecutive first-order reaction as below[12,13,14]. RuvC itself also show enzymatic activity consistent with first-order reactions based prediction[15].

$$A \xrightarrow{k_a} B \xrightarrow{k_b} C$$

in the equation A represents the intact DNA duplex, B the DNA molecule in which one of the two strands has been cleaved at the recognition site for the restriction enzyme and C the DNA molecule (or molecules) in which both strands have been cleaved at this site.

In order to link the apparent first-order rate constant to ∆ G'. We assume both steps of cleaving is classic enzymatic reaction as follow.

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_{cat}} P + E$$

With S as the substrate, E the enzyme and P the product.

One can derive the concentration-time function of C following enzyme kinetic equations. The equation will be like following (derivation details in Addendum):

$$\therefore [C] = [A_0][1 + (\frac{1}{k_a - k_b})(k_b e^{-k_a t} - k_a e^{-k_b t})]$$

$$k_a = \frac{k_{cat1}[E]^*}{K_M};\ k_b = \frac{k_{cat2}[E]^*}{K_M}$$

This equation is hard to link with ∆ G', as

$$K_M = \frac{k_{-1} + k_{cat}}{k_1}, \quad K_a = \frac{k_1}{k_{-1}} = e^{-\frac{\Delta G'}{RT}}$$

It's also hard to fit into present data, as there is no kinetic data for Cas9 available now. But we can employ this function to draw figure 4.

The Figure shows that even the k_a and k_b of the on-target binding is twice as large as ka and kab of the off-target binding, the off target rate will still grows drastically as the time goes on.

So in addition to control the concentration of Cas9, control the expressing time of Cas9 is also important for off-target rate control. Cas9's expression can be stoped as soon as possible when acceptable theoretical editing rate is reached, in order to reduce off-target rate.

5.Discussion

There may be three reason for the correlation variation throughout output 1-19.

First, the Cas9 has a mismatch tolerance for the 5' end of gRNA. This is backed by all studies [1,2,3,4,7,8].

Second, there are flaws in the calculation of terminal energy. As all terminal mismatch of RNA and most for DNA are stabilizing [5,6]. The NN model may fail to catch all these stabilizing effect. So improvement of the energy calculation rules may help to fix the negative correlation.

Three, the NN model is derived from the binding energy database of free binding DNA double strands, while we employed it to calculate Cas9 influenced RNA-DNA binding. We considered this as the prime source of error in our model. And it may contribute to the funny correlation valley in Δ G(8)~Δ G(12). Or, maybe the valley means that Cas9 is indeed relatively insensitive to energy changes in these position.

6.Addendum

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_{cat}} P + E$$

$$\frac{d[\mathrm{S}]}{dt} = -k_1[\mathrm{S}][E] + k_{-1}[ES]$$

$$\frac{d[E]}{dt} = -k_1[\mathrm{S}][E] + (k_{-1} + k_{cat})[ES]$$

$$\frac{d[ES]}{dt} = k_1[\mathrm{S}][E] - (k_{-1} + k_{cat})[ES]$$

$$\frac{d[P]}{dt} = k_{cat}[ES] = v$$

In a typical endonuclease environment, $[A] \ll K_M$ and $[A] \leq 10[E]$ are always hold. Even in Pattanayak's paper[1], though the total DNA concentration is 200nM, the concentration every single kind of DNA(with certain sequence) is lower than 0.1nM, which is much lower than K_M of any typical restriction enzyme,

But still, the MM equation remains valid. Because, first, under these conditions, [E] (free E concentration) doesn't change much, because most "enzymes" are in free form and they don't do anything; second, some time after enzyme and substrate are mixed the concentrations of free enzyme sites and of substrate complexed will reach a steady state.[17]

$$\therefore \frac{d[E]}{dt} = -k_1[A][E] + (k_{-1} + k_{cat})[EA] = 0$$

$$\therefore v = \frac{k_{cat}[E]^*[A]}{K_M + [A]} = \frac{k_{cat}[E]^*[A]}{K_M} \text{ (Michaelis-Menten equation)}$$

$$\therefore k_a = \frac{k_{cat1}[E]^*}{K_M};\ k_b = \frac{k_{cat}[E]^*}{K_M}$$

$$A \xrightarrow{k_a} B \xrightarrow{k_b} C$$

$$\therefore [A] = [A_0] e^{-k_a t}$$

$$\therefore [B] = [A_0] (\frac{k_a}{k_b - k_a})(e^{-k_a t} - e^{-k_b t})$$

$$\therefore [C] = [A_0][1 + (\frac{1}{k_a - k_b})(k_b e^{-k_a t} - k_a e^{-k_b t})]$$

Pattanayak's in vitro experiment can reveal the off-target rate in vivo. Because in the experiment the DNA and gRNA-Cas9 concentration is 200nM and 100nM respectively. Every single kind of DNA has a abundance equals to or less than 0.1% (which is approximately the abundance of wild type sequence, the most abundant one), so the concentration of a specific DNA is on the same power(or less than) 0.1nM. Therefore,

$$\frac{\mathrm{DNA}_{\mathrm{exp}}}{\mathrm{Cas9}_{\mathrm{exp}}} = \frac{0.1\mathrm{nM}}{100\mathrm{nM}} = \frac{10^{19}/\mathrm{L}}{100\mathrm{nM}} = \frac{10^{13}/\mathrm{L}}{0.1\mathrm{pM}} = \frac{10^{-2}/\mu\mathrm{m}^3}{0.1\mathrm{pM}}$$

$$\frac{\mathrm{DNA}_{vivo}}{\mathrm{Cas9}_{vivo}} = \frac{1/\mathrm{Nucleolus}}{0.1\mathrm{pM}} = \frac{1/100\mu\mathrm{m}^3}{0.1\mathrm{pM}} = \frac{10^{-2}/\mu\mathrm{m}^3}{0.1\mathrm{pM}}$$

Nucleolus size according to [16], in vivo protein concentration of mammalian cell from [9]

The DNA-Cas9 ratio is of the same order, so it's reasonable to use the experimental data to predict the Cas9 behavior in vivo.

Reference

1. Pattanayak, Vikram, et al. "High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity." *Nature biotechnology* (2013).
2. Mali, Prashant, et al. "CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering." *Nature biotechnology* 31.9 (2013): 833-838.
3. Qi, Lei S., et al. "Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression." *Cell* 152.5 (2013): 1173-1183.
4. Bikard, David, et al. "Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system."*Nucleic Acids Research* (2013).
5. SantaLucia Jr, John, and Donald Hicks. "The thermodynamics of DNA structural motifs." *Annu. Rev. Biophys. Biomol. Struct.* 33 (2004): 415-440.
6. Mathews, David H., et al. "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure." *Journal of molecular biology* 288.5 (1999): 911-940.
7. Hsu, Patrick D., et al. "DNA targeting specificity of RNA-guided Cas9 nucleases." *Nature biotechnology* 31.9 (2013): 827-832.
8. Fu, Yanfang, et al. "High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells." *Nature biotechnology* 31.9 (2013): 822-826.
9. Alon, Uri. *Introduction to Systems Biology: And the Design Principles of Biological Networks*. Vol. 10. CRC press, 2007. Page 6.

10. Buchler, Nicolas E., Ulrich Gerland, and Terence Hwa. "Nonlinear protein degradation and the function of genetic circuits." *Proceedings of the National Academy of Sciences of the United States of America* 102.27 (2005): 9559-9564.
11. Jinek, Martin, et al. "A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity." *Science* 337.6096 (2012): 816-821.
12. Halford, Stephen E., Nicola P. Johnson, and John Grinsted. "The reactions of the EcoRi and other restriction endonucleases." *Biochemistry*. J 179 (1979): 353-365.
13. Halford, Stephen E., Nicola P. Johnson, and John Grinsted. "The EcoRI restriction endonuclease with bacteriophage lambda DNA. Kinetic studies."*Biochemistry*. J 191 (1980): 581-592.
14. Fogg, Jonathan M., et al. "Yeast resolving enzyme CCE1 makes sequential cleavages in DNA junctions within the lifetime of the complex." *Biochemistry* 39.14 (2000): 4082-4089.
15. Fogg, Jonathan M., and David MJ Lilley. "Ensuring productive resolution by the junction-resolving enzyme RuvC: large enhancement of the second-strand cleavage rate." *Biochemistry* 39.51 (2000): 16125-16134.
16. Alberts, Bruce. *Molecular biology of the cell* (4th edition). Garland Science, (2000): 191-234.
17. Gutfreund, Herbert. *Enzymes: physical principles*. London: Wiley-interscience, 1972.

Thanks for
the consulting by Kenji Adzuma
the data processing by Yancheng Liu
the data collection by Lei Yang
the debugging by Yao Yang
Otherwise the model can never be done in time.