Free Energy Model of Off-target Problem

Zheng Hu, Sherry Dongqi Bao
Tianjin University

October 15, 2018

1 Background

The current models of the off-target problem in the CRISPR-Cas system can be divided into three types. Firstly, it can be modeled as purely ”sequence alignment problem” similar to BLAST, which doesn’t consider either the energy change or the experiment data; Secondly, it can be modeled with the help of experiment data; Thirdly, in a kinetic model, the energy change will be considered.

Our model aims to investigate the off-target problem in the CRISPR-Cas system in an innovative kinetic model, therefore finding efficient ways to enhance the reliability of gene editing. Our motivation is based on our project Cell-free cancer detection. Firstly, there are no existing tools for evaluating sgRNA candidates’ off-target probability with Cas13a, and the current tool for Cas12a is BLAST-based, which means that it is a rough model and can hardly take the case of single nucleotide mutation of cancer into account. The foundations of our model are mostly simple probability theory and energy theory, which make our model both convincing and pellucid.

2 Introduction

Currently, people have constructed a similar model as illustrated in the following figure 1. There are four common rules when Cas nuclease cleaves the DNA[1].

.

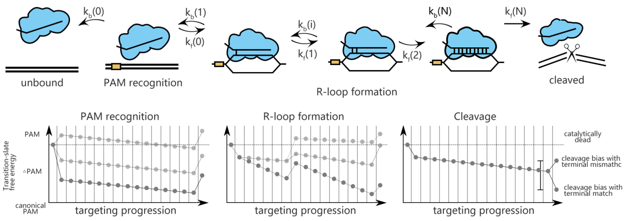

Figure 1: schematic diagram

(1) Seed region: single mismatches within a PAM proximal seed region can completely disrupt interference;

(2) Mismatch spread: when mismatches are outside the seed region, off-targets with spread-out mismatches are targeted most strongly;

(3) Differential binding versus differential cleavage: binding is more tolerant of mismatches than cleavage;

(4) Specificity-efficiency decoupling: weakened protein-DNA interactions can improve target selectivity while still maintaining efficiency;

Based on these four rules, probability theory is applied in to explain it. As we know, there are always two results in an experiment, which are successful cleavage and unsuccessful cleavage. In math view, it can be one-hot encoded, and they are corresponding to 1 and 0.

However, giving a 0/1 prediction is hard and unreliable. To solve this problem, one choice is to consider it as a cluster problem; however, it is easier to find a continuous quantitative function rather than to find a suitable cluster distance function. So naturally, finding an approximate probability distribution is a good choice.

Many current off-target evaluating websites use a score function to evaluate whether the target is good or bad. Here we consider the score function has the similar ability to probability, which is a description of "better" or "worse" while it can't guarantee whether successful cleavage will appear. For our case, our goal is to find a function indicating which target is BETTER.

Considering the difference between model prediction and experimental data, our model consists of two aspects, which are kinetic model and parameter updating module.

3 Methods

3.1 Kinetic model

The whole binding-cleavage process begins with the binding between PAM and protein. It corresponds to rule#1 mentioned before. And as the reaction proceeds, every step of it is reversible, and its irreversibility mainly depends on the binding energy of two DNA bases.

Here we use the term binding probability P_{bind} represent the probability of sgRNA binding with DNA, with a rough assumption that the probability of cleavage is the same for every kind of binding. In total, the probability of off-target cleavage is positively correlated with the binding probability.

The probability $P_{bind,N}$, representing the probability of binding at the Nth position (the last position of sgRNA) of nucleotide base, is given by:

$$P_{bind,N} = \frac{k_f(N)}{k_f(N) + k_b(N)} = \frac{1}{1+\gamma_N} \tag{1}$$

where k is the reaction rate constant; f represents the forward reaction; b represents the backward reaction. And

$$\gamma_N = \frac{k_b(N)}{k_f(N)} \tag{2}$$

So for a the whole sgRNA sequence:

$$P_{bind} = \frac{1}{1 + \sum_{n=1}^{N} \prod_{i=1}^{n} \gamma_i} \tag{3}$$

Consider the rate constant $k_f(i)$ and $k_b(i)$:

$$k_f(i) = k_0 exp(-(T_{i,i+1} - F_i)), k_b(i) = k_0 exp(-(T_{i,i-1} - F_i)) \tag{4}$$

where F_i means free energy of each metastable state, $T_{i,i+1}$ means the highest free energy point on the reaction path from position i to position i + 1. Therefore, $T_{i,i+1} - F_i$ is the activation energy of forward reaction and $T_{i,i-1} - F_i$ is activation energy of the backward reaction.

$$\Rightarrow \gamma_i = exp(-\Delta_i), \Delta_i = T_{i,i+1} - T_{i,i-1} \tag{5}$$

$$\begin{aligned}\Rightarrow P_{bind} = \frac{1}{1 + \sum_{n=1}^{N} \coprod_{i=1}^{n} \gamma_i} &= \frac{1}{1 + \sum_{n=1}^{N} \coprod_{i=1}^{n} exp(-\Delta_i)} \\ &= \frac{1}{1 + \sum_{n=1}^{N} exp(-\sum_{i=1}^{n} \Delta_i)}\end{aligned} \tag{6}$$

We define

$$\Delta T_n = \sum_{i=1}^{n} \Delta_i$$

so

$$P_{bind} = \frac{1}{1 + \sum_{n=1}^{N} exp(-\Delta T_n)} \tag{7}$$

From the above, it is clear that the binding probability depends only on the state transition energy, not on the free energy of metastable states. If we assume there is one dominant bias ΔT_{n^*}, then this equation can be approximated as:

$$P_{bind} \approx \frac{1}{1 + exp(-\Delta T_{n^*})} \tag{8}$$

As figure shows, we analyze the energy change of every process:

$$\begin{cases} \text{for the PAM position (i = 0) we have } \Delta_0 = \Delta PAM \\ \text{for a partial R-loop we have } \Delta_i = \Delta_M \text{ if matched and } \Delta_i = -\Delta_U \text{ mismatched} \end{cases}$$

$$\Delta T_n = \Delta_{PAM} + n_M \Delta_M - (n - n_M)\Delta_U - \delta_{n,N}\Delta_{clv};\ n = 0...N \tag{9}$$

where n_M is the number of matches, Δ_{clv} is the energy change of cleavage after the process of binding is finished.

$\delta_{n,N}$ represents the Kronecker delta: $\delta_{n,N} = \begin{cases} 1, n = N; \\ 0, n \neq N. \end{cases}$

For PAM independent systems (such as Cas13a), we instead use:

$$\Delta T_n = n_M \Delta_M - (n - n_M)\Delta_U - \delta_{n,N}\Delta_{clv};\ n = 0...N \tag{10}$$

To sum up, the binding probability mainly relies on the free energy change, and PAM appears as a significant energy decline.

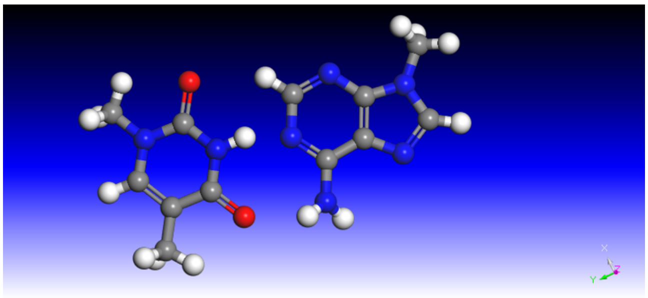

Figure 2: Hydrogen bonds between AT

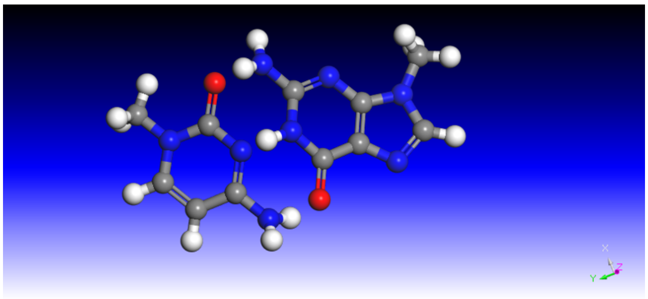

Figure 3: Hydrogen bonds between CG

The kinetic model sets up a framework to build the relationship between binding probability and the numbers of nucleotide matches and mismatches. In consideration of this problem more carefully, the binding probability becomes equal to the analysis of energy change, because we know the binding energy of A/T and C/G is different due to the different hydrogen bonds between them (figure 2 and 3), and the energy decrease in C/G is approximately 1.5 folds as A/T. Similarly, the mismatch has more types of variance because the sizes of nucleotides are various. Hence, the types of the mismatched base pair are classified by group volume, i.e., two pyrimidines (such as C/T, "Large"), pyrimidine and purine (such as C/A, "Medium"), two purine (such as G/T, "Small"). Hence, the probability can be calculated using the following equation:

$$P_{bind} = \frac{1}{1 + e^{-n_1\Delta_{A/T} - n_2\Delta_{C/G} + n_3\Delta_L + n_4\Delta_M + n_5\Delta_S}} \tag{11}$$

3.2 Parameter Optimization

From the kinetic model, we can get an output, which is the binding probability. It needs to be noticed that the parameters we choose should make results well discriminated, because in a cleavage experiment, we only have two outcomes, successful(1) and unsuccessful(0).

To make our predictions from the model more approximate to experiment(facts), we set a regression module and implement parameter optimization. Here, the method we choose is stochastic gradient descent (SGD) and cross entropy. Their principles can be concluded as follows:

$$\theta = \theta - \eta \nabla_\theta J(x^{(i)}, y^{(i)}, \theta) \tag{12}$$

$$loss = \sum_i y_i \ln y_i \tag{13}$$

where θ means the parameter array and J means the loss function.

Considering the difficulty in gradient calculation, we substitute the differential term with a difference term to accelerate operating speed.

$$\frac{dy}{dx} \approx \frac{\Delta y}{\Delta x} = \frac{y(x+\delta x) - y(x)}{\delta x} \tag{14}$$

By using this simple method, our model can be more vibrant and more reliable with the ability of updating using newest data.

3.3 Generating sgRNA Candidates

Meanwhile, we designed a program to generate all the sgRNA candidates for a target gene, and combined with the previous model, we can compare and rank all the candidates.

The principle of the program is very simple. We use PAM as the input and collect the arrays with a certain length which contain the same beginning code as PAM.

4 Results

Here, we set the parameters as default values and observe its performance. Our default parameters are set based on three simple rule. Firstly, because of the different numbers of hydrogen bonds between A/T and C/G, the energy decrease due to A/T binding are 1.5 times to C/G binding. So the parameter#1 is 1.5 times parameter#2. Secondly, it's universally known that big compounds with small distance will have a high energy because of repulsion. The mismatching combination between two nucleotides can be classified into three parts–large, medium and small, and we should set other parameters in the same order of values. Thirdly, we use the parameters from [1] as reference. In total, we set the parameters default values as [0.06,0.09,0.3,0.6,0.03].

As the following figure shows, the energy always decreases or has some turning point because of mismatch and is always negative. The yellow line represents complete match. The blue line and the red line are two examples of mismatches at different positions. For example, the red line has a peak due to a mismatch, and in our model, we find that it doesn't make the energy positive. It means that in this reaction process there is some force like "momentum" pushing it to proceed and cross the energy peak. For a off-target examination to every location of all the DNA in a system (a genome), there will also be positive energy result, which is obviously not a feasible solution (will not lead to off-target). This kind of results is omitted in the plot.

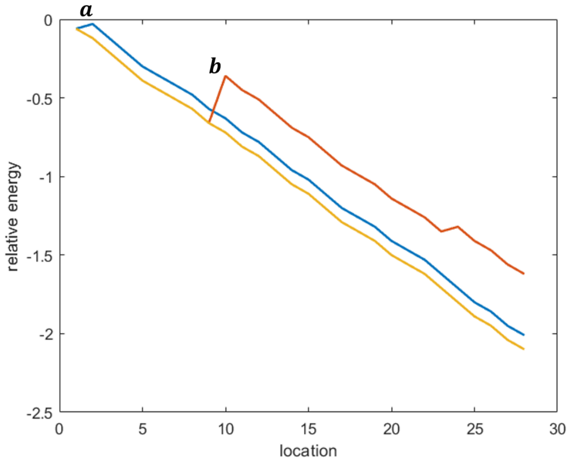

Figure 4: Energy change of sgRNA candidates binding to DNA

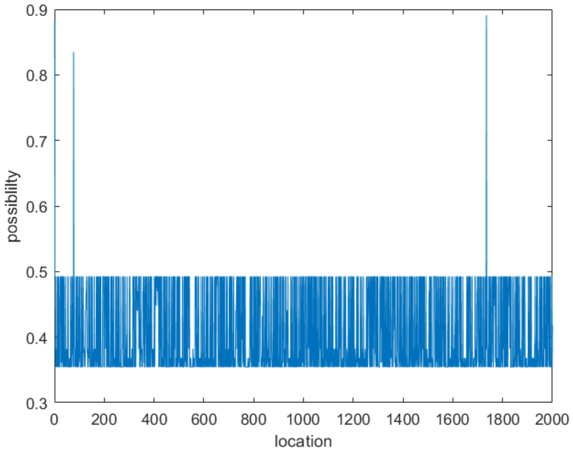

Figure 5: The possibility of sgRNA candidates binding to DNA's different locations

After testing our code's running time, we find its rate can reach approximately 2×10^8 $base/h$ (under parallel computing in 4 cores).

Besides the default parameters, we hope our model can hit more true data. So after we get the experiment data, we can use the parameter optimization method mentioned before to get more precise parameters.

References

[1] Misha Klein et al. "Hybridization kinetics explains CRISPR-Cas off-targeting rules". In: *Cell reports* 22.6 (2018), pp. 1413–1423.