[entry]none/global/

Construction of Free Energy Model

Zheng Hu, Sherry Dongqi Bao
Tianjin University

October 10, 2018

1 Introduction

Nowadays, the analysis of cleavage possibility can be devided into to type, i.e. meta-empirical and empirical. For the first one, many people develop the various score function based on experiment data to evaluate if a sgDNA is good or bad. Correspondingly, the other group chooce set up a theoretical model based on kinetic theory. But because using many approximations, it has drawbacks inevitably.

Our model aims to investigate the off-target problem in gene editing by the CRISPR-Cas system, therefore finding efficient ways to enhance the reliability of gene editing. The foundations of this model are mostly simple probability theory and dynamic deduction, which make our model both convincing and pellucid.

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].
.

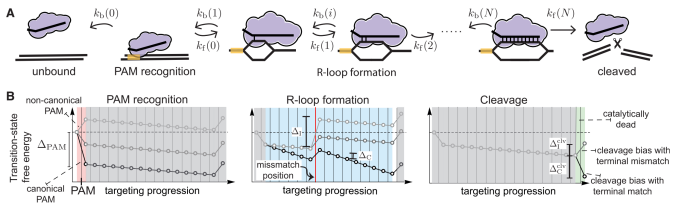

Figure 1: schematic diagram

(1) Seed region: single mismatch(es) 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 only 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.

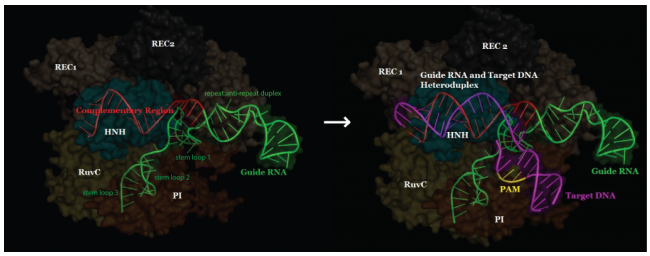

Figure 2

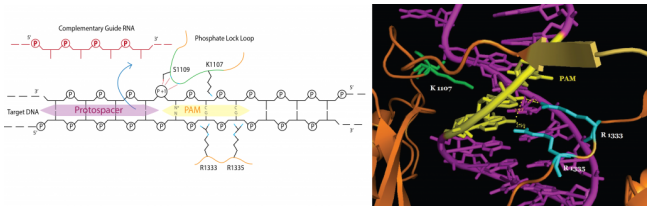

Figure 3

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.

In many target design toolkits, they use a score function with several parameters which can generate a score 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 can't affirm 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 inference and an updating module.

2 Methods

2.1 Kinetic module

Figure 2 shows that the whole binding-cleavage process begins with the binding between PAM and protein. Therefore, 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.

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

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

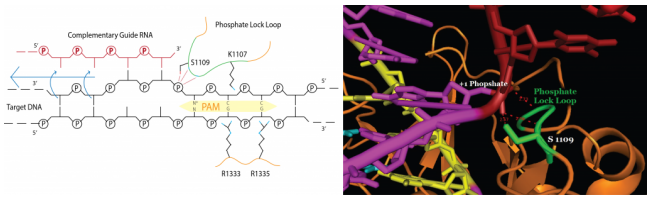

Figure 4

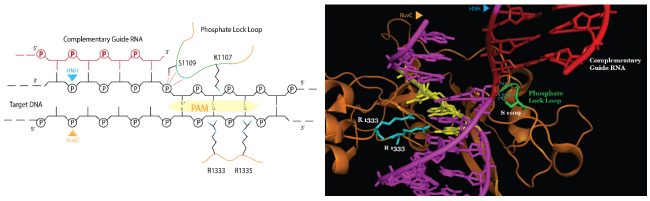

Figure 5

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 complete match:

$$P_m \equiv P_{m,0} = \frac{1}{1 + \sum_{n=1}^{N} \prod_{i=1}^{n} \gamma_i} \tag{3}$$

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.

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

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

We define

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

so

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

From the above, it is clear that the matching probability depends only on the state transition energy, not on the free energy of the metastable states. If we assume there is one dominant minimal bias, say for $n = n^*$, then this equation can be approximated as:

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

As figure shows, we analyze each process energy change:
$\begin{cases}$ for the PAM state (i = 0) we have $\Delta_0 = \Delta PAM$
for a partial R-loop we have $\Delta_i = \Delta_C$ and $\Delta_i = -\Delta_I$ if mismatched

$$\Delta T_n = \Delta_{PAM} + n_C(n)\Delta_C - (n - n_C)\Delta_I - \delta_{n,N}\Delta_{clv};\ n = 0...N \tag{9}$$

where $\delta_{n,N}$ represents the Kronecker delta: $\delta_{n,N} = \begin{cases} 1, n = N; \\ 0, n \neq N. \end{cases}$
For PAM independent systems (such as Cas13), we instead use:

$$\Delta T_n = n_C(n)\Delta_C - (n - n_C)\Delta_I - \delta_{n,N}\Delta_{clv};\ n = 0...N \tag{10}$$

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

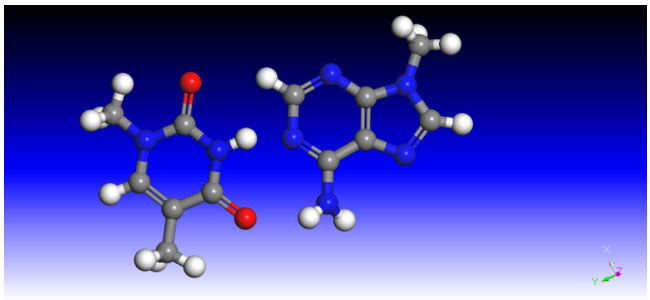

Figure 6: AT

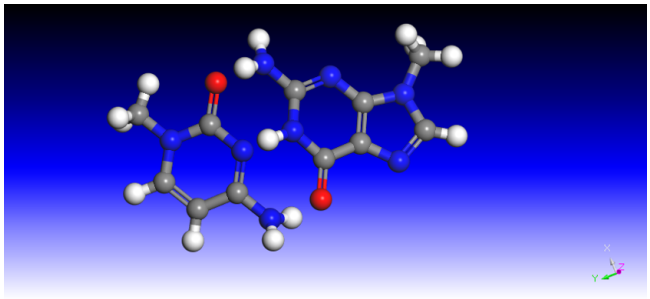

Figure 7: CG

So the kinetic module set up a form to regress the relationship between cleavage and the numbers of nucleotide matches and mismatches. In consideration of this problem more carefully, the cleavage possibility becomes equal to analysis

energy change, and we know the binding energy of A/T and C/G is different due to the different hydrogen bond between them. However, in appearing kinetic model, research tend to describe them in a rough definition as "matched base pairs", and the energy incline in C/G is approximately 1.5 folds as A/T. Similarly, the mismatch has more difference because the size of nucleotides is various. Hence, the combination of the mismatched base pair was classified by group volume, i.e. two pyrimidines (such as C/T, "L"), pyrimidine and purine (such as C/A, "M"), two purine (such as G/T, "S"). Hence, our possibility can be calculated using the following formation.

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

2.2 Optimization module

It is a common sense that experimental results are facts, but theoretical results are only conjectures. From kinetic module, we can get an output, which is the cleavage possibility. The parameter we choose only aims to make results have discrimination, while it's not quantitative. And in a cleavage experiment, we only have two outcomes, successful and unsuccessful. To make our prediction possibility more approximate to experiment, we regard this as a regression problem.

Here, the method we choose is stochastic gradient descent (SGD) and cross entropy. And their principle 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 parameters array and J means the loss function.

Considering the difference in gradient calculation, we use difference to substitute differential aim 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, updating using newest data and becoming more reliable.

2.3 Pre-selector

It's obviously that the algorithm is too complex to applying in slide in a huge DNA array. To solve this problem, we use a pre-selector to get some candidates and use previous model to compare them so that we could get a greatest target.

And here this pre-selector structure is very simple. Considering use this map to reflect the similarity between target and full DNA.

$$f(a_{\text{target}} - a_{\text{full}}) = \mathbf{x} = (x_i) \tag{15}$$

$$x_i = \begin{cases} 1, & matched \\ 0, & mismatched \end{cases} \tag{16}$$

Here, we use PAM as an input and collect the array which contain the same beginning code as PAM.

3 Result

Here, we set the parameters as default values and observe its performance. As the following figure shows, the energy always decreases or has some turning point and is always negative. Such as the red line, it has a peak due to a mismatch here, and in our model, we find that it doesn't make the energy positive. That means that in this reaction process there is some force like "momentum" pushing it to proceed and cross the energy peak. Corresponding to the other figure's two particular locations (a and b), only in these points their energy are all negative (because we want to see the idea target series, only the locations which correspond to negative energy are collected). After testing our code run time, its manage rate

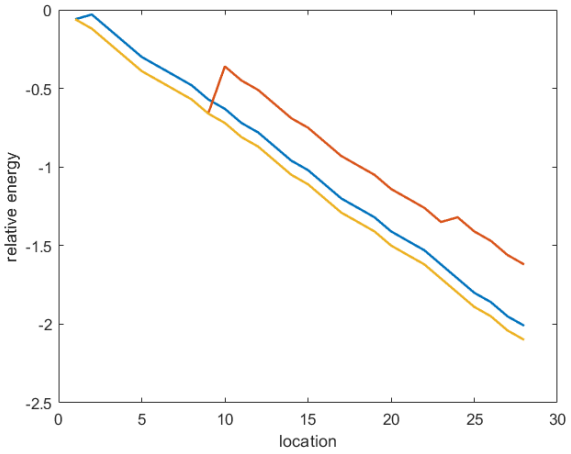

Figure 8: energy change

can reach approximate $2 \times 10^8\ base/h$ (under parallel calculation in 4 cores) and have somewhat applicaiton value. Besides the default parameters, we hope our model can hit more true data. So if we get the experiment data, we can use model 2.2 to get greater parameters. (@@no experiment data)

References

[1] family=Klein, familyi=K., given=Misha, giveni=M., "Hybridization kinetics explains CRISPR-Cas off-targeting rules". In: *Cell reports* 22.6 (2018), pp. 1413–1423.

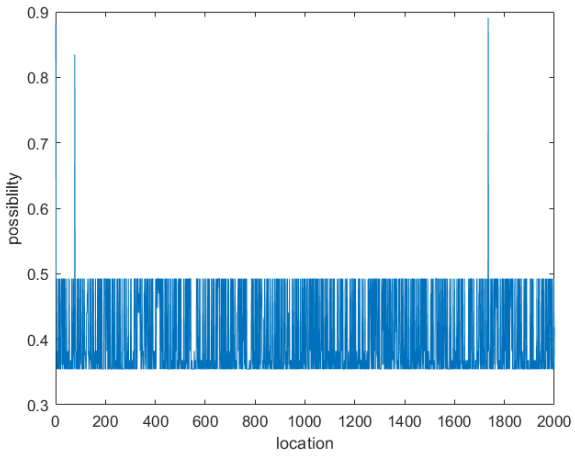

Figure 9: the possibility of target binding to nucleotide array in different location