Construction of Free Energy Model

Zheng Hu, Sherry Dongqi Bao
Tianjin University

September 24, 2018

Abstract

Our model aims to find the intricate reason in gene cleaving behavior and find efficient method to enhance the genome-editing tools reliability, i.e. reducing the off-targeting possibility. The basic knowledges of this model are mostly simple probability and dynamic deduction, which make our model both convincing and pellucid.

1 Introduction

In the appearing researching, people have construction similar model as illustrating in the following figure (Figure.1). There are four common rules when cas nuclease cleave the DNA[].

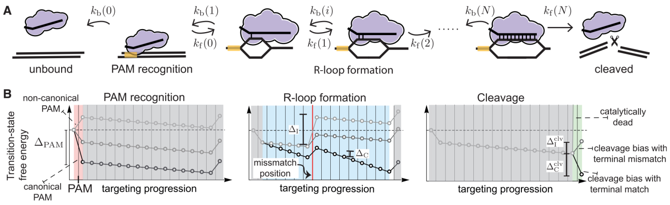

Figure 1: schematic diagram

(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 to mismatches than cleavage

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

Based on these four rules, possibility theory was applied in to explain it. As we know, there are always only two results in experiment, which are successful cleavage and unsuccessful cleavage. In math view, it can be on-hot encoded and they are corresponding to 1 and 0. To solve it, we may consider it as a cluster problem, however, it is easier to find a quantitative continuous function rather than

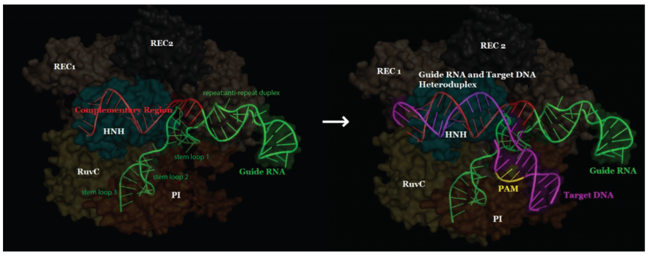

Figure 2

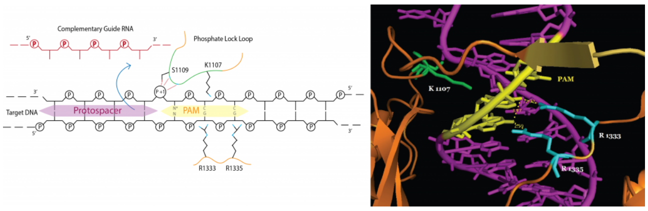

Figure 3

try hard to find a suitable cluster distance function. Naturally, finding an approximate possibility distribution is a good choice. In many target design toolkits, they have a set of parameters to get a score to evaluate whether it's good or bad. Here we think the score function has similar ability to possibility which is a description of better or worse and can't affirm whether successful cleavage appear. Considering this problem, it roots in the question itself because future can't be predicted. But possibility can. Hence, there we tend to find a function so that we can know what target is better.

Considering difference between model predicting data and experimental data, our model consists of two aspects, which are kinetic inference and updating module.

2 Model

2.1 kinetic model

From the figure.2, this process is begun with binding between PAM and protein. Therefore, it corresponds to rule 1 mentioned before. And with the reaction proceeding, every steps in it are reversible and it's irreversibility mainly depended on the banding energy of two DNA bases.

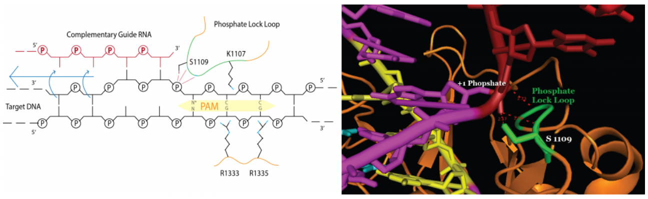

Figure 4

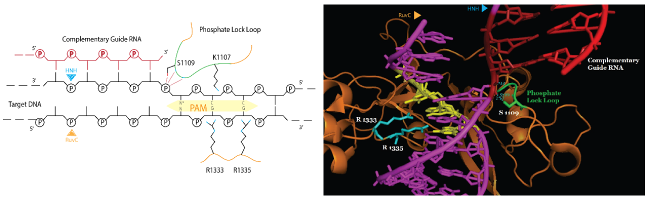

Figure 5

The boundary probability Pclv;N , representing the probability to cleave s-taring with a full R-loop and without reducing the R-loop's length, is given by a simple splitting probability:

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

$$P_{clv} \equiv P_{clv,0} = \frac{1}{1+\sum_{k=1}^{N}\prod_{i=1}^{n}\gamma_i}$$

free-energy:F_i ; the transition state energy:T_i, from i to $i+1$, we can get the relationship between reaction constant and free energy:

$$k_f(i) = k_0 exp(-(T_i - F_i)), k_b(i) = k_0 exp(-(-T_{i-1} - F_i))$$

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

$$\Rightarrow P_{clv} = \frac{1}{1+\sum_{k=1}^{N}\prod_{i=1}^{n}\gamma_i} = \frac{1}{1+\sum_{k=1}^{N}\prod_{i=1}^{n} exp(-\Delta_i)} = \frac{1}{1+\sum_{k=1}^{N} exp(-\sum_{k=1}^{n}\Delta_i)}$$

define:

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

$$P_{clv} = \frac{1}{1+\sum_{k=1}^{N} exp(-\Delta T_n)}$$

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

$$P_{clv} \approx \frac{1}{1+exp(-\Delta T_{n)}}$$

As figure shows, we analyst each process energy change:

$$\left\{\begin{array}{l}\text{for the PAM state (i = 0) we have } \Delta_0 = PAM\\ \text{for a partial R-loop we have } \Delta_i = \Delta_C \text{ and } \Delta_i = \Delta_I \text{ if mismatched}\end{array}\right.$$

$$\Delta T_n = \Delta_{PAM} + n_C(n)\Delta_C(nn_C(n))\Delta_I\delta_{n,N}\Delta_{clv}; n = 0\ldots N$$

where $\delta_{n,N}$ represents the $Kronecker$ delta: $\delta_{n,N}=\left\{\begin{array}{l}1, n = N;\\ 0, n \neq N.\end{array}\right.$

For PAM independent systems (such as cas13), we instead use:

$$\Delta T_n = n_C(n)\Delta_C(nn_C(n))\Delta_I\delta_{n,N}\Delta_{clv}; n = 0\ldots N$$

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

So the kenetic module set up a form to regression the relationship between cleavage and the numbers of nucleotide matches and mismatches. Considering this problem more carefully, the cleavage possibility become equal to analysis energy change and simply 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 mismatched base pair was classified by group volume, i.e. two pyrimidine (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 following formation.

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

2.2 Optimization model

It is common sense that experiment result are facts but theorical results are only conjectures. In the model 1, we can get an output, which means the cleavage possibility. The parameter we chooce only aim to make results has discrimination, while it's not quantitative. And in a cleavage experiment, we only have two results, 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 choose cross entropy. And their principle can be concluded as following.

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

$$loss = \sum_i y_i \ln{}_i$$

Considering the difference in gradient calculation, we use difference to substitute differential aim to accelerate operating speed.

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

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.

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

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

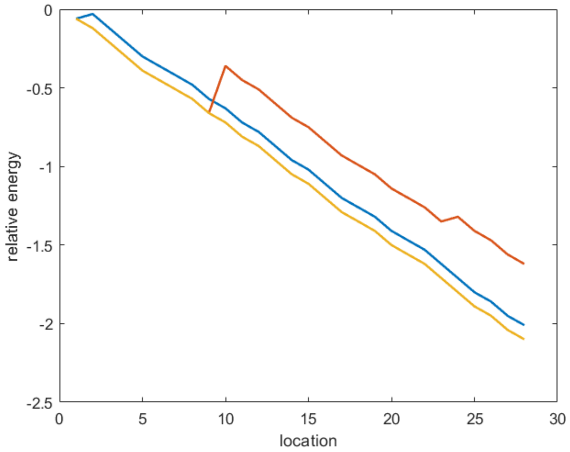

Figure 6

Figure 7

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

3 Result

Here, we set this parameters as default values and observe its performance. As the following figure shows, the energy is always declined or has some turning point and is alway negative. Corresponing to the other figure's two special location(a ad b), only in these point their energy are all negative (because we want to see the idea target series, only the locaitons which corresponds to negative energy are collected).

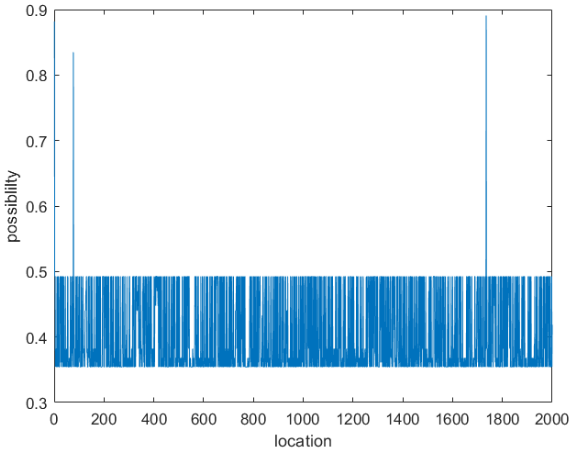

Figure 8

Figure 9