Master Thesis

Koen van der Sanden

January 9, 2018

Chapter 1

Introduction

Genome editing is one of the major technological and biological developments of the past few years and will certainly play a major role in the society of the future. In the past few years the advancements in genome editing have sped up enormously and we are looking for ever more precise tools to perform this task. One of the newest tools in our collective toolbox is an enzyme named Cas9. It is a CRISPR enzyme that allows targeting a specific spot on the DNA and cut the DNA at that exact location. This could open the way for a lot of practical uses, for example targeted editing of plant genomes to generate more crops, editing plant genomes to be resistent against pesticides and removing heriditary diseases from embryos. The current issue with Cas9 is that the enzyme is not perfect in its targeting of the DNA, resulting in off-targets which are also cleaved.

To solve this issue it is important to understand why some off-targets are cleaved while others are not. From experiments we have a lot of data from which we can draw empirical conclusions, however a physical model would tell you much more than just a rule-of-thumb. Such a model is being thought out by [SOURCE] . They are attempting to create a kinetic model which predicts the cleavage of Cas9. However since Cas9 targets specific DNA sequences it can be used for more than just cleaving. dCas9 is the dead version of Cas9; it also targets sequences on the DNA but it is incapable of cleavage. This dead enzyme can be used for all kinds of research. Usually a fluorescent molecule is attached to the enzyme so it can be tracked and bound (off-)targets can be identified. In this thesis I will attempt to adapt the cleavage model from [SOURCE] to predict binding and bound off-targets.

Chapter 2

Chapter 1 - Theory

In this chapter we will discuss the general model used to predict binding and cleavage. We will start by describing the cleavage model as described in [SOURCE]. However instead of calculating the probability to cleave a sequence we will calculate the probability to bind to a certain sequence.

2.1 General Model

The model from [SOURCE] is quite simple at its core. Cas9 binds to the DNA in essentially two ways. One section forms bonds between the target DNA and RNA attached to the protein called 'guide RNA' or gRNA for short. The gRNA has a length of twenty bases. The second section forms 'bonds' between the target DNA and the protein itself, this section is called the 'protospacer adjacent motif' or PAM for short. Once everything is bound, the PAM and all twenty bases in the gRNA, active Cas9 is able to cleave the DNA.

The model used to predict binding and cleavage is analogous to a zipper. In a zipper there are separate teeth that get locked together one after another when the zipper is tightened. In a similar way the teeth let loose one set after another when the zipper is loosened. Analogous when the DNA and Cas9 bind to each other first the PAM binds, then the first base pair of the gRNA, then the second base pair, then the third, etcetera. This continues until all twenty bases are bound to each other and only then Cas9 can cleave. Not only is the one-by-one binding of the bases similar to the one-by-one interlocking teeth of a zipper, the entire process is also reversible like a zipper. The bound bases can also one-by-one unbind from each other just like the teeth of a zipper can be separated again. The only difference here is that once the DNA is cleaved, the process can not be reversed.

With this zipper model in mind we can describe the binding of Cas9 to the DNA as a number of distinct states. The first state in our model is simply unbound Cas9; the DNA and the enzyme are separate. This is similar to a zipper which is entirely separated. The second state in our model is the one where the PAM is bound, similar to attaching the very first part of the zipper; no teeth are interlocked yet but the first connection is made and the two strands are attached to each other. After that we have twenty states corresponding to the binding of each of the base pairs, similar to a zipper which has tweny pairs of teeth. Finally we have the very last state where the Cas9 cleaves the target DNA, which is a special state since it is irreversible.

A well-functioning zipper is similar to the actual target of a Cas9 enzyme, called the on-target. The start of the DNA sequence of the on-target fits the PAM of Cas9 and every base on the DNA perfectly corresponds to its complement on the gRNA. An off-target however is similar to a zipper with a broken tooth somewhere along the way. Everything up to the broken tooth is similar to a well-functioning zipper but it is hard to pull the zipper over the broken tooth since it will not interlock correctly with the tooth on the other strand. However once the zipper is pulled over the broken tooth it is again very easy to tighten the rest of the zipper. This broken tooth is a mismatch on the target DNA. At first

Cas9 does not know about the mismatch and will bind to the off-target but then somewhere along the way it will hit the mismatch. If it makes it over the mismatch it is then easy to continue further and eventually cleave the DNA, but if the mismatch is hard to get over the Cas9 can also simply unbind from that particular DNA sequence, since the binding process is reversible.

This zipper image tells us how to think of Cas9 binding and cleavage but it does not yet allow us to predict which DNA sequences will be cleaved. To make that prediction we prescribe every state with a certain energy. We know that processes always tend to the configuration or state with the lowest possible energy. For now we will not worry about what the precise value of this energy is but we can assume certain things from what we know about Cas9:

1. The solution state has a certain energy associated with it, but since all energy changes are relative we can set the solution energy to any value we want. Therefore only increases and decreases in energy matter.
2. A sequence with a matching PAM is more likely to cleave than a non-matching PAM. Therefore a matching PAM has a lower energy than a non-matching PAM.
3. Matching bases increase the likelihood of cleaving the DNA, therefore a matching base must be an energy decrease.
4. Non-matching bases decrease the likelihood of cleaving the DNA, therefore a mismatch must be an energy increase.
5. The process of binding one base pair involves several things. First the DNA pair must be separated, then the DNA base must turn to the RNA base and then the RNA and DNA base must bind to each other. The unbinding of the DNA base pair and the turning of the DNA base will at first increase the energy of the state. This is the activation energy.

These things tell us that there are three parameters in the minimal model that we have tot take into account: $\Delta C, \Delta I$ and ΔPAM, the energy gain from a match, the energy penalty from a mismatch and the energy gain from the PAM respectively. From these assumptions we can draw some general energy landscapes.

At this point we know that Cas9 behaves as a sort of zipper and each state has an associated energy. It is also known that nature tends towards the lowest energy state. From this it is clear that some sequences will be cleaved; the ones with a low final energy, and some will not be cleaved; the ones with a high final energy. We now want to quantify these initial conclusions.

From Kramers rate theory [SOURCE] we know:

$$k_f(i) = k_0 \exp\left(F_i - T_i\right), \tag{2.1}$$

$$k_b(i) = k_0 \exp\left(F_i - T_{i-1}\right), \tag{2.2}$$

where k_f and k_b are the forward and backward rate respectively. They depend linearly on a certain attempt rate k_0 and exponentially on the free energy of state i and the transition energy away from state i. The free energy of state i is F_i, the transition energy from state i to state $i+1$ is T_i and the transition energy from state i to state $i-1$ is T_{i-1}. Our system can then be pictured schematically as in figure

This couples the rates to move between states to the energies associated with those states. Using this we can say something about the probability to cleave and/or bind certain sequences. In this thesis we will focus on the binding instead of the cleaving, see [SOURCE] for a discussion on cleavage.

The first step in using this model to determine whether a specific sequence will be bound by Cas9 is to establish what we will classify as 'bound'. One option is to classify every not-unbound state as bound, in other words: as long as Cas9 is connected to the DNA, albeit only the PAM, we will classify it as bound. Another option is to only classify a Cas9 as bound when it has a certain amount

of bases of the gRNA bound. A logical choice would be to classify only Cas9 as bound if the entire gRNA is bound. A third option would be to classify Cas9 as bound when it is likelier to fully bind the gRNA than to unbind. All choices could be justified, however the choice has to correspond to the measurement that is done in the data which will eventually be used to fit the parameters in the model.

2.1.1 Equilibrium

No matter which definition of 'bound Cas9' we pick we have several options to calculate the fraction of bound Cas9. One simple way to calculate the fraction of bound Cas9 is to assume that the entire system is in equilibrium. If the system is in equilibrium and there are no exclusion effects then we can use boltzmann statistics to determine the fraction of Cas9 in each available state by only knowing the energy of each state. The probability to be in a single state is as follows:

$$P_i = \frac{\exp\left(-E_i\right)}{\sum_{i=0}^{N} \exp\left(-E_i\right)}, \tag{2.3}$$

where N is the total number of states available and E_i is the energy of state i. Let us now take the example where every state is considered bound except the solution, or unbound state. Then the probability of being bound (P_b) in equilibrium is:

$$P_b = \frac{\sum_{i=1}^{N} \exp\left(-E_i\right)}{\sum_{i=0}^{N} \exp\left(-E_i\right)}, \tag{2.4}$$

where we have set state 0 to be the solution state.

2.1.2 Time dependent

When the system is not in equilibrium a time dependency is introduced which makes the problem significantly more difficult. As far as we know there is no exact solution to the problem when it is not in equilibrium, however we can calculate the solution numerically. To do this we solve the master equation numerically. As with the minimal model we assume that the entire system has 22 different states (for a gRNA with 20 bases): the unbound state, the PAM bound state and one state for each subsequent base pair. Each state can transition to the next or previous state only with a certain rate for each transition. Schematically this can be represented as in figure 2.1.2. Each circle in this diagram represents a state of the system and each arrow represents a forward or a backward rate to go from a state to another.

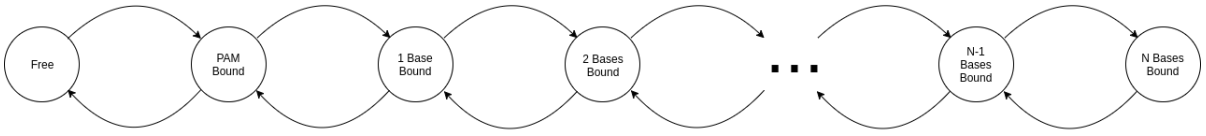

Figure 2.1: A schematic representation of the system.

We can say that a particular dCas9 enzyme has a probability to be in a specific state P_i at a specific time t. If we look at the probability $P_i(t)$ to be in the state i at a slightly later time $t + dt$ then the probability has changed to

$$P_i(t + dt) = P_i(t) - P_i(t) \cdot (\lambda_i + \mu_i) \cdot dt + \lambda_{i-1} P_{i-1}(t) \cdot dt + \mu_{i+1} P_{i+1}(t) \cdot dt, \tag{2.5}$$

which holds as long as dt is small enough to only allow a single transition. Here λ_i is the forward rate from state i to state $i+1$ and μ_i is the backward rate from state i to state $i-1$. Rewriting this equation and letting $dt \to 0$ we get

$$\frac{\partial P_i(t)}{\partial t} = (-\lambda_i - \mu_i)P_i(t) + \lambda_{i-1}P_{i-1}(t) + \mu_{i+1}P_{i+1}(t), \tag{2.6}$$

for $i \in [-1, N]$. Where we call state -1 the free state, state 0 the state with only the PAM bound and states 1..20 the states with bases 1..20 bound. We can write equation 2.6 in matrix form

$$\frac{\partial \vec{P}(t)}{\partial t} = M \cdot \vec{P}(t), \tag{2.7}$$

where M is the transition matrix containing the forward and backward rates. One can easily solve this equation

$$\vec{P(t)} = exp(M \cdot t) \cdot \vec{P}(0). \tag{2.8}$$

This gives the probability of a specific molecule to be in each state at a time t. In other words, this is the fraction of molecules in each state at a time t.

For the specific rates contained in the matrix we have assumed that each forward rate is constant and the same: $k_f(i) = k_f(i+1)$ for all $i \neq -1$. This is a reasonable assumption since physically the molecules only feels the interaction of the forward barrier for its forward rate. This energy barrier is determined by the energy it takes to break up the DNA-DNA bond and, approximating all DNA bonds as equal, this is always the same. The only forward rate we have not assumed the same is that from solution to the first (PAM) bound state, since this can vary with concentration of dCas9. Following Kramers rate theory (equations 2.1 and 2.2) the forward and backward rates are given by

$$\begin{aligned} k_f(i) &= k_0 \cdot exp(F_i - T_i) \\ k_b(i) &= k_0 \cdot exp(F_i - T_{i-1}), \end{aligned}$$

With F_i the free energy of state i and T_i the transition energy from state i to state $i+1$. Therefore

$$k_b(i) = k_f(i-1) \cdot exp(F_i - F_{i-1}). \tag{2.9}$$

With this it is possible to calculate the occupancy of each state over time and therefore the predicted on- and off-rates. As Boyle et al. (2017) we will call every state other than the unbound state a bound state.

Chapter 3

Findings

Now that we are able to model the system the next step is to get the correct values for the parameters in our model. The exact values of $\delta C, \delta I$ and δPAM determine the behaviour of the model, therefore it is essential to get these values as accurate as possible. To obtain these values we will fit the model to data from Boyle et al. (2017). This dataset was chosen because of several advantages over other datasets. For one it is very large, it contains all single and double mismatch and on top of that also sequences with more than two mismatches. Furthermore the authors have measured several properties of the enzymes: the on-rate, the off-rate and the occupancy. This allows us to fit on multiple datasets or test our fit for a different experiment. In the end we have tried to fit the minimal model to all three types of data, because for the first two datasets we tried we ran into problems. Still we will report on those problems here since they provided us with insight into the workings of the enzyme and the model.

3.1 Dissociation Rate

The first dataset we tried to fit was the dissociation rate (figure [NUMBER] in Boyle et al. (2017)). The reason for this is that we figured it would be similar to the dissociation constant, for which an expression is already reported in [SOURCE]. However it turned out this is not the case. The expression for the dissociation constant as reported in [SOURCE] is:

$$K_D = \frac{1}{\sum_{n=0}^{N} \exp\left(-\Delta F(n)\right)}, \tag{3.1}$$

where $F(n)$ is the energy difference between the solution ($n = -1$) and the nth state. This dissociation constant is udeful when considering a system that is in equilibrium. This is where the first issue with the dissociation dataset comes in; the reported dissociation rates are definitely not in equilibrium due to the way they are measured. If everything in the experiment were performed perfectly the reported dissociation rates would be the actual, concentration-independent dissociation rates. These are however closely linked to the dissociation constant so one might think that we could still use the dissociation constant for the fitting procedure. However in the data is also something peculiar: the sequences with mismatches in the seed are more stable than sequences with a mismatch at the end. This contradicts what we already know of dCas9: that mismatches in the seed make the complex less stable [SOURCES]. It also contradicts the association dataset of Boyle et al. (2017), which tells us that these sequences with a mismatch in the seed have a much lower effective on-rate. Assuming there are no interactions at a distance, the on-rate is only determined by the first bases: the PAM. If this is all the same for these sequences then a lower effective on-rate can only be due to a higher effective off-rate.

The reason for this inverted picture is, we argue, not the result of some special region the authors name the Reversibility Determining Region (RDR for short), but the inverted picture is entirely due to the inherit time-dependence of the problem and experiment. Boyle et al. (2017) measure the effective off-rate by measuring the intensity of a fluorescent signal every 500 seconds. This fluorescent signal is emitted by bound dCas9 enzymes, so a decreasing signal points to unbinding of dCas9. However it is quite hard to extract a dissociation rate from only the intensity of a fluorescent signal. When Boyle et al. (2017) measure the intensity every 500 seconds they naturally lose any direct signs of events that happen within 500 seconds. Now the unbinding of dCas9 is quite slow, as proven by the dissociation rate from the on-target. However when we introduce mismatches meta-stable states are created which can contain a significant portion of the total dCas9 population in equilibrium (and even more before equilibrium). These meta-stable states can have much faster dissociation rates, since they are inherently less stable than the full R-loop.

We will start by fitting the minimal model to the occupation data from Boyle et al. (2017) (figure S2).

For this initial fit we want to keep our options very broad so we do not limit any of the parameters.

Chapter 4

Chapter 3

Chapter 5

Chapter 4

Boyle et al. (2017)

Bibliography

Boyle, E. A., Andreasson, J. O., Chircus, L. M., Sternberg, S. H., Wu, M. J., Guegler, C. K., Doudna, J. A., and Greenleaf, W. J. (2017). High-throughput biochemical profiling reveals sequence determinants of dcas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences*, 114(21):5461–5466.