Master Thesis

Koen van der Sanden

February 26, 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 keep up with our genome editing needs. One of the newest tools in our collective toolbox is an enzyme named Cas9. It is a CRISPR enzyme that allows targeting of a specific spot on the DNA and cleave the DNA at that exact location. This precision instrument 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. This could lead to disastrous side-effects if we ever plan to use Cas9 to edit, for example, human embryos.

To solve this issue it is important to understand why some off-targets are cleaved while others are not. A lot of experiments have been done to better understand the off-targeting and people have draw empirical conclusions from the data. However, a physical model would providemore insight and therefore, hopefully, better predictions for the off-targets. The fundamentals for such a physical model have been laid by Klein et al. (2017). They are attempting to create a kinetic model which predicts the probability a specific DNA sequence is cleaved by Cas9. This work builds on the fundamentals that have been laid by Klein et al. (2017), but will try to answer a slightly different question:

*"Is it possible to build a more accurate algorithm for predicting the **binding** of dCas9, based on the physical model from Klein et al. (2017), than the existing models based on the empirical observations from experiments, without any underlying physical model?"*

There are several differences between this question and the work done by Klein et al. (2017). The main difference is that, while Klein et al. (2017) predicts the *cleaving* probability of Cas9, this work will attempt to predict the *binding* probability of dCas9. Note the cleavage is done by the enzyme Cas9, while binding is done by dCas9 (dead Cas9), which for all intents and purposes is assumed to be exactly the same enzyme but incapable of cleaving. Even though we are in essence discussing the same process with the same enzyme, Klein et al. (2017) already notes that 'only binding' and 'binding & cleaving' are expected to behave differently based on the exact circumstances.

One might ask that if our end goal is genome editing why should we be concerned with the binding probability of dCas9 at all? Would the cleavage probability not be the only parameter of interest? Since Cas9 targets specific DNA sequences in the same way as active Cas9 it has applications outside genome editing. dCas9 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 on a DNA strand can be identified. This is usefule for sequencing, blocking specific genes, testing DNA strands for specific genes, etc.

In short, the main goal of this work is to answer the question if we can build a binding prediction algorithm for dCas9 based on Klein et al. (2017). While this is the main goal, since it is based on a

physical model which requires an understanding of the binding process, implicitly we also hope to get a better understanding of the way (d)Cas9 binds to the DNA.

Chapter 2

Chapter 1 - Theory

In this chapter we will discuss the general model used to predict binding and cleavage. This entire work is theoretical so this chapter could be very long. Instead this chapter will contain the very fundamentals of the model, some important insights and some expressions and formulas that will be used throughout this thesis. This necessarily means more theory, insights and slight changes to the model will follow in sections outside of this one, but the fundamentals of the model will be laid out here.

2.1 Definitions

Since the model is build on the physics underlying the binding and cleaving by Cas9 it is useful to first get a simplistic understanding of what the enzyme exactly is and how it works. In figure 2.1 a schematic of (d)Cas9 is depicted. It shows the general shape of the enzyme and some important components.

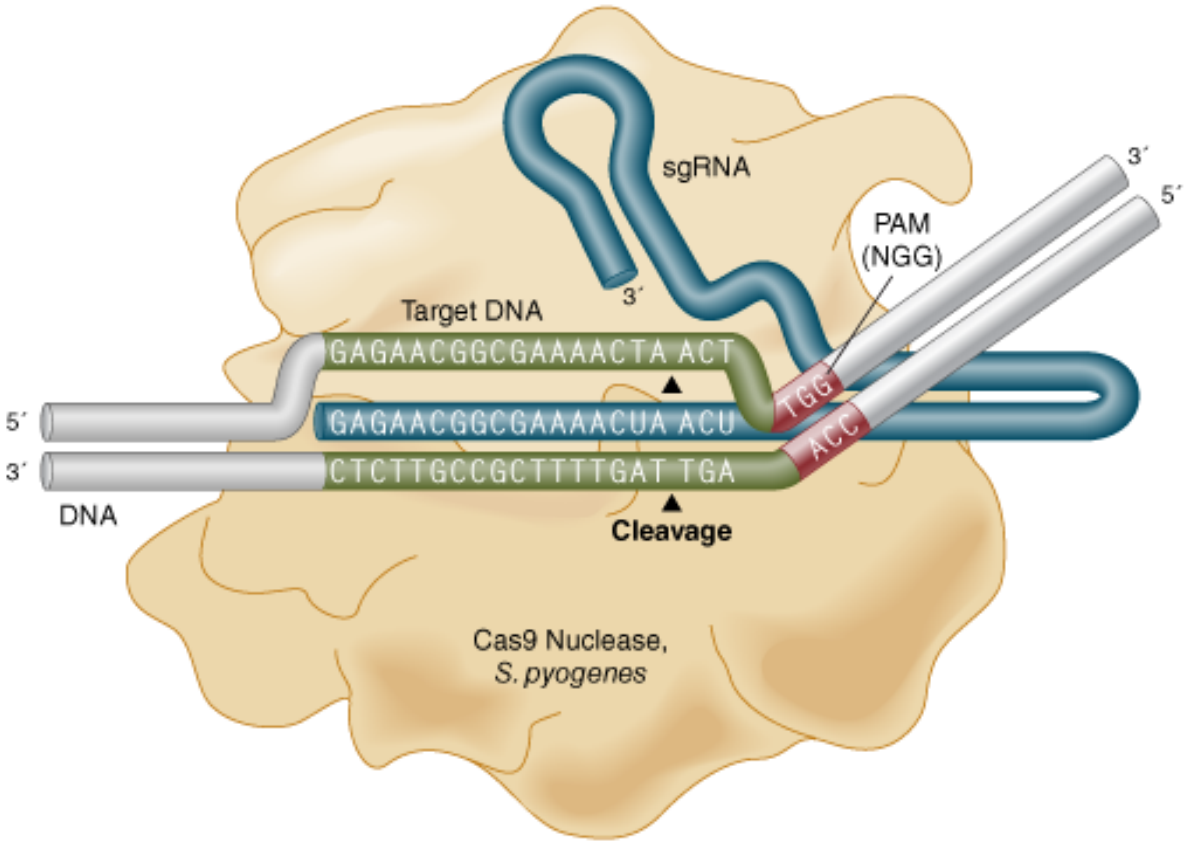

Figure 2.1: Schematic of (d)Cas9

In figure 2.1 several components of Cas9 are defined, we will take a closer look at what each of these is. In the schematic there are two strands of DNA (white/beige) and one strand of RNA (blue). The RNA is called sgRNA or gRNA, short for (single stranded) guide RNA. The gRNA is connected to the enzyme but it is fabricated separately and is attached later to the Cas9. Therefore the gRNA is also easily fabricated in a lab. One of the DNA strands contains the target sequence (and naturally the other contains the complement to the target sequence). Finally the DNA contains the PAM (protospacer adjacent motif) sequence. This is a sequence (NGG) just before the target sequence (sometimes called protospacer) that does not match to the gRNA but it matches to the exact shape of the enzyme. Because this sequence depends on enzymatic interactions, it is practically unchangeable.

2.2 General Model

The model from Klein et al. (2017) is quite simple at its core. Cas9 binds to the DNA in essentially two ways. One section forms bonds between the target DNA and gRNA. The gRNA has a length of twenty bases. The second section forms 'bonds' between the target DNA and the protein itself, the PAM section. 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.

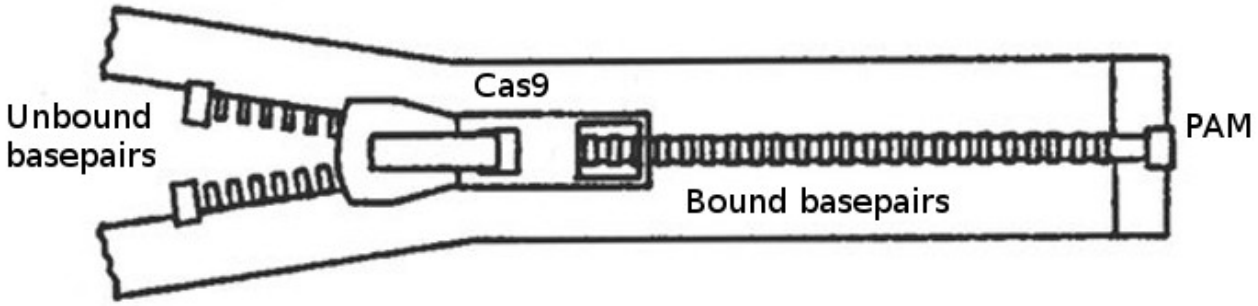

Figure 2.2: The model behaves just like a zipper.

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 only 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 twenty pairs of teeth. Finally we have the very last state where the Cas9 cleaves the target DNA. This last cleaved state is special because it is irreversible. Therefore it does not have an analogous state in the zipper picture. A schematic drawing of the separate states is shown in 2.2.

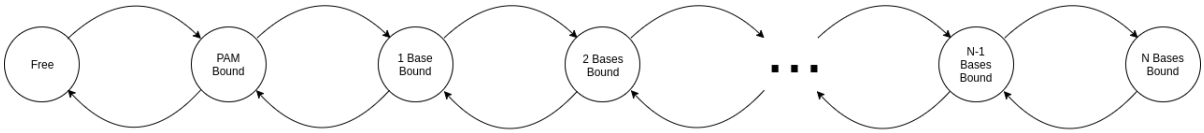

Figure 2.3: A schematic representation of the system.

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 the teeth will not interlock correctly. However once the zipper is pulled over the broken tooth it is again easy to tighten the rest of the zipper. This broken tooth is a mismatch on the target DNA. At first Cas9 has no way of knowing which sequences are off-targets, so it will bind to the off-target. However, 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 since the mismatch is difficult to pass, the Cas9 can also simply unbind from that particular DNA sequence.

This zipper picture 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 state with the lowest possible energy.

For now we will not worry about the precise values of the energies 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. This is justified by the fact that we mostly measure canonical PAM sequences cleaved [SOURCE] and by the fact that Cas9 has a specific shape to recognize the canonical PAM sequence. Naturally non-canonical PAM sequences therefore have less affinity to be bound.
3. The PAM energy and therefore all subsequent energies, depend on the concentration of (d)Cas9. This is the entropic effect. Logically, if there is a surplus of Cas9 in solution, more of it will be bind to less favourable sequences.
4. Matching bases increase the likelihood of cleaving the DNA, therefore a matching base must be an energy decrease. If matching bases did not increase the chance of cleavage then almost no sequences would be cleaved in a short timeframe.
5. Non-matching bases decrease the likelihood of cleaving the DNA, therefore a mismatch must be an energy increase. This is justified because, if mismatched bases also decreased the energy then it would be favourable to cleave all sequences.
6. 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 in the system. This is the activation energy.

From these assumptions we can draw some general energy landscapes, as seen in figure 2.2.

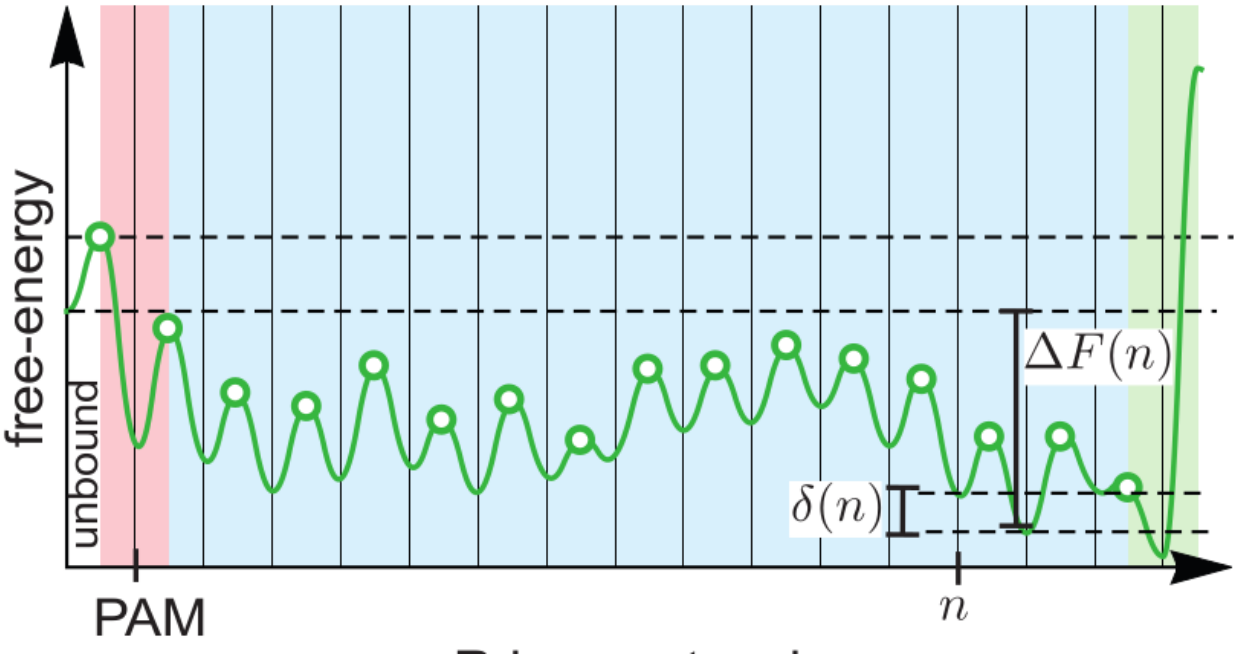

Figure 2.4: A visualization of how the energy landscape could look like.

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. Similar conclusions can be drawn for binding instead of cleaving, but they are not as straightforward. For binding it is possible to have metastable states that allow stable binding, even on sequences that will never be cleaved. This is because, for cleavage the only important question is if Cas9 is able to reach the final state. For binding the question is if there are any stable states (lower energy than solution).

With this rudimentary pictuer of the zipper in mind we can create countless different models. The simpelest of all is what will be called 'the minimal model'. This model is the one which captures all our assumptions in the least amount of parameters possible. In fact it only requires three parameters: the energy gain from a match (ϵC) the energy penalty of a mismatch (ϵI) and the energy gain or penalty of the PAM (ϵPAM). The minimal model is already described for cleavage in Klein et al. (2017) and like there it will set the foundation for this entire report.

2.3 Minimal model

The minimal model is the simpelest model that encompasses everything described in section 2.2. It only has three different parameters: ϵC, ϵI and ϵPAM, the energy gain from a match, energy penalty from a mismatch and energy gain or penalty from the PAM respectively. This gives complete information about the energy of each state in the minimal model. To predict the binding probability of any sequence we only have one question left to answer: 'When is a Cas9 enzyme bound?' This may seem like a silly question but it is important to agree on what constitutes binding. One option is to only mark Cas9 as bound when it reached the final state in the system, meaning a fully formed R-loop. Another option is to mark Cas9 bound when the entire seed-region is bound. A third proposition could be to mark Cas9 as bound when it is simply not unbound, so every state except solution would be a 'bound state'. None of these definitions is necessarily better than any other but it is important to be consistent, even more so between theory and experiment. Therefore we will consider all states bound except the solution state. Not only does this make ituitive sense, it is also the same quantity that is measured in several experiments like (Boyle et al., 2017) and [SOURCE].

2.3.1 Equilibrium

The 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:

$$P_i = \frac{\exp\left(-E_i\right)}{\sum_{i=-1}^{N} \exp\left(-E_i\right)}, \tag{2.1}$$

where N is the total number of states available and E_i is the energy of state i. If we consider only the solution state as unbound, then the probability of being bound (P_b) in equilibrium is:

$$P_b = \frac{\sum_{i=0}^{N} \exp\left(-E_i\right)}{\sum_{i=-1}^{N} \exp\left(-E_i\right)}, \tag{2.2}$$

where we have set 'state -1' to be the solution state.

2.3.2 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: 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 2.3.2. 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.

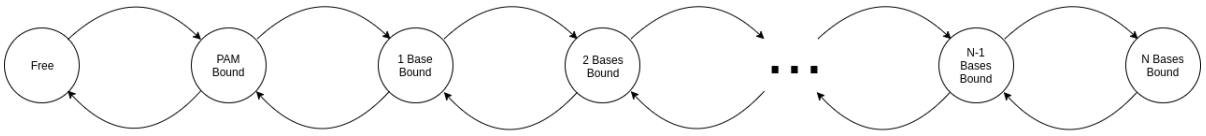

Figure 2.5: A schematic representation of the system.

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

$$P_i(t + dt) = P_i(t) - P_i(t) \cdot (\lambda_i + \mu_i) \cdot dt + \lambda_{i-1} P_{i-1}(t) \cdot dt + \mu_{i+1} P_{i+1}(t) \cdot dt, \tag{2.3}$$

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

$$\frac{\partial P_i(t)}{\partial t} = (-\lambda_i - \mu_i) P_i(t) + \lambda_{i-1} P_{i-1}(t) + \mu_{i+1} P_{i+1}(t), \tag{2.4}$$

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 2.4 in matrix form

$$\frac{\partial \vec{P}(t)}{\partial t} = M \cdot \vec{P}(t), \tag{2.5}$$

where M is the transition matrix containing the forward and backward rates. Note that we have basically rewritten the master equation in matrix form. One can easily solve this equation:

$$\vec{P(t)} = exp(M \cdot t) \cdot \vec{P}(0). \tag{2.6}$$

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 [SOURCE] the forward and backward rates are given by:

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

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

$$k_b(i) = k_f(i-1) \cdot exp(F_i - F_{i-1}). \tag{2.7}$$

So by assuming that all forward rates are equal, we can completely determine the transition matrix in equation 2.6 and therefore follow the occupancy of each state through time, giving us the binding probability.

Chapter 3

Minimal Model

In this chapter we will discuss the minimal model in depth. The minimal model, although it has only a few parameters and is quite simple in concept is able to explain a lot of the behaviour of Cas9. Furthermore, since it is the simplest model it is more intuitive and easier to pinpoint what exactly it does and does not do and why. Therefore it will function as a starting point for all our research.

3.1 Fitting the parameters

The first step in analyzing the minimal model is getting the correct values for the parameters in our model. The exact values of $\epsilon C, \epsilon I$ and ϵPAM 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 it has 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 and test our fit for a different experiment.

3.1.1 Methods of (Boyle et al., 2017)

3.2 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 Klein et al. (2017). However it turned out this is not the case. The expression for the dissociation constant as reported in Klein et al. (2017) 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.

To support this claim we use the minimal model to simulate the system. To do this we numerically solve 2.6, which gives us the probability to be in any specific state at any specific time, given an initial condition. For the dissociation experiment as performed by Boyle et al. (2017) this initial condition is simply the equilibrium as given by the boltzmann weights. We will show with this that for certain parameter sets we can reproduce the characteristics of the dissociation heatmap as shown in figure 3 of Boyle et al. (2017). The parameter set we will use is:

$$\begin{aligned} \delta C &= -1.41, \\ \delta PAM &= -2.02, \\ \delta I &= 8.87, \\ k_0 &= 1000, \\ k_{on} &= 0.425 \text{ nm/s}, \\ [Cas9] &= 1 \text{ nm}, \end{aligned} \tag{3.2}$$

where the delta's are the energy differences for a match, mismatch and a PAM match, while k_0 is the attempt rate, k_{on} is the on-rate and [Cas9] is the concentration of Cas9 enzymes.

Before we check the entire heatmap, we first take a look at the dissociation figures as represented in figure 1 in Boyle et al. (2017). Here we see the on-target disscociation compared to the dissociation of a sequence with a mismatch at position 16. We can reproduce these curves with our simuated minimal model, resulting in figure **??**.

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.

Klein, M., Eslami-Mossallam, B., Arroyo, D. G., and Depken, M. (2017). The kinetic basis of crispr-cas off-targeting rules. *bioRxiv*, page 143602.