# Master Thesis

Koen van der Sanden

March 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.

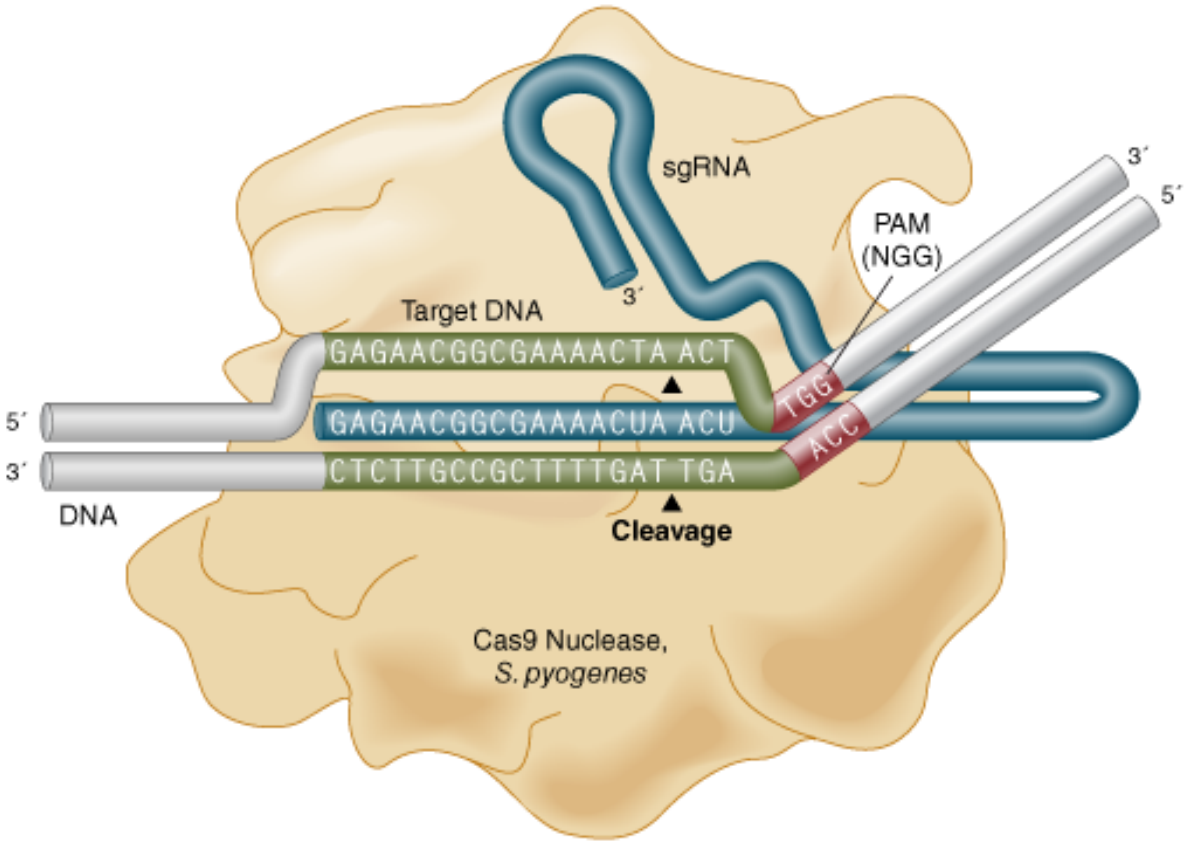


Figure 2.1: Schematic of (d)Cas9

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.

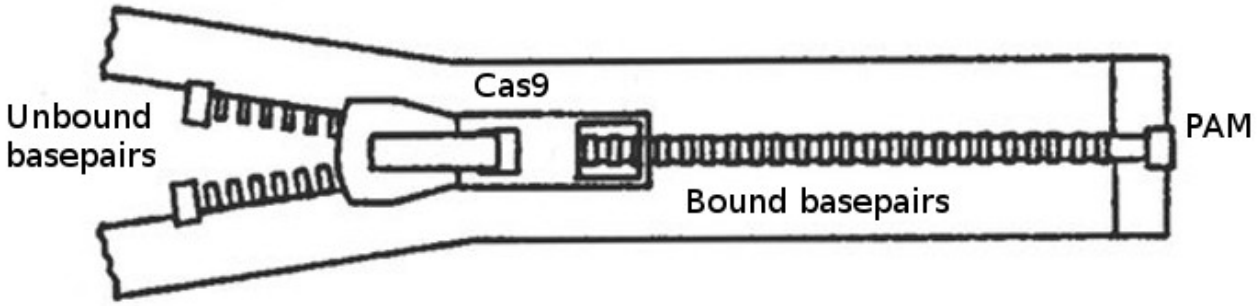


Figure 2.2: The model behaves just like a zipper.

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.

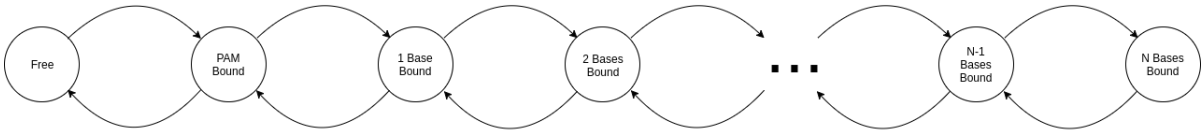


Figure 2.3: A schematic representation of the system.

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.

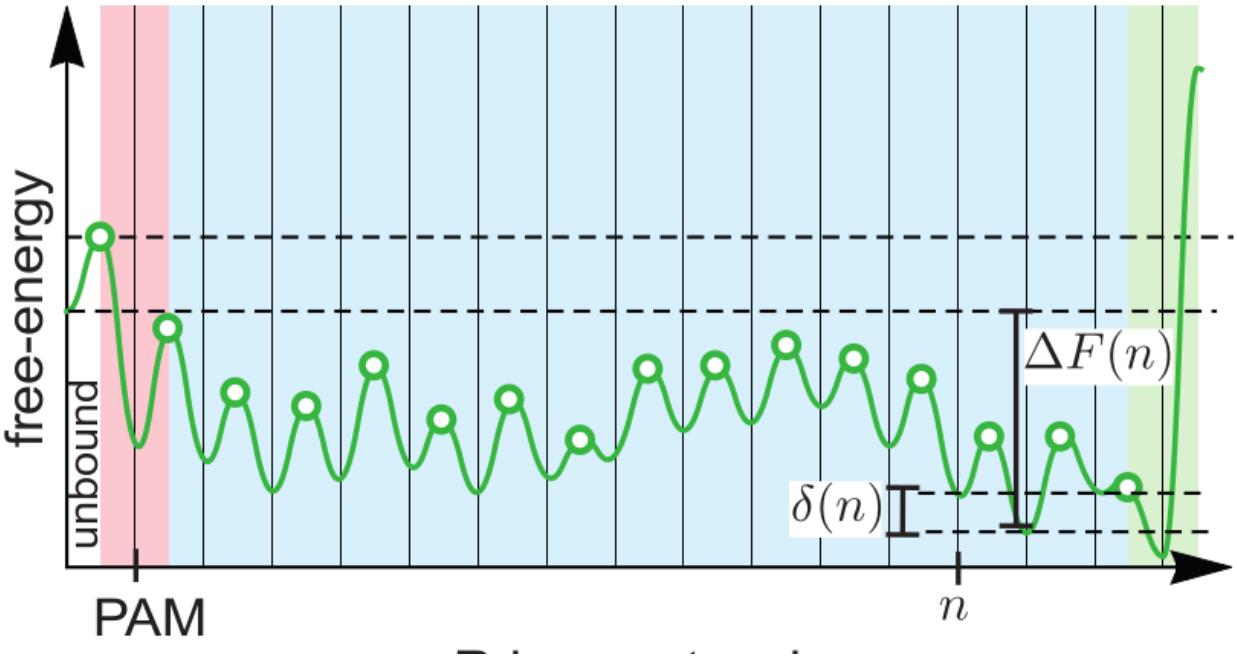


Figure 2.4: A visualization of how the energy landscape could look like.

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 only thing we need to know are the energy values of each state and the rate at which dCas9 transitions between states. These energies and rates together fully describe the system, both in equilibrium and out of equilibrium.

### 2.2.1 Equilibrium

The simplest way to calculate the fraction of bound Cas9 is to assume that the entire system is in equilibrium. In equilibrium there is no evolution over time and therefore the transition rates which determine how the system evolves over time drop out of the model. 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. This calculation is simple and can easily be done by hand, but that is not the case if the system is out of equilibrium.

### 2.2.2 Time dependent

When the system is not in equilibrium a time dependency is introduced which makes the problem significantly more difficult. Not only the calculations will be harder to do but since the system evolves over time the transition rates do not drop out. 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. We assume that the entire system has $N$ different states: the unbound state, the PAM bound state and one state for each subsequent base pair. Each state is able transition to the next or previous state only with a certain rate for each transition. Schematically this can be represented as in figure 2.2.2. 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.

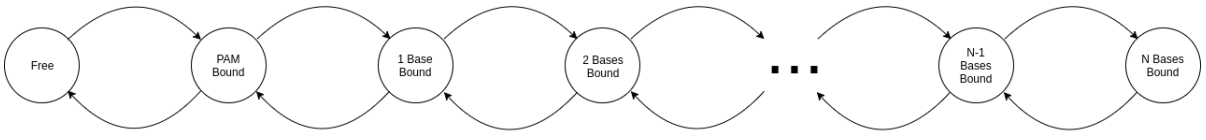


Figure 2.5: A schematic representation of the system.

Any particular dCas9 enzyme has a probability $P_i(t)$ to be in a specific state $i$ at a specific time $t$. If we look at the probability $P_i(t+dt)$ 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 $\lambda_i$ is the forward rate from state $i$ to state $i+1$ and $\mu_i$ 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 $1..N$ the states with bases $1..N$ 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 $t$. Following Kramers rate theory [SOURCE] the forward and backward rates contained within the matrix $M$ are given by:

$$k_f(i) = k_0 \cdot exp(F_i - T_i) \tag{2.7}$$
$$k_b(i) = k_0 \cdot exp(F_i - T_{i-1}), \tag{2.8}$$

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

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

The energy landscape therefore also contains information about the rates. In fact, if the energy landscape is known it is possible to calculate all forward or backward rates if all backward or forward rates are known respectively. To eliminate a lot of potential parameters we will assume for most models that the forward rates are all equal. 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 and the interaction between the PAM sequence and the protein is fundamentally different from all other bonds that are formed. Thus, if all forward rates are assumed to be constant and the same: $k_f(i) = k_f(i+1)$ for all $i \neq -1$, we can completely determine the transition matrix in equation 2.6 given two rates: the rate from solution to PAM (the on-rate) and a rate from any bound state to any other state (the attempt rate). The attempt rate is the $k_0$ factor that is shown in equations 2.7.

## 2.3 Minimal model

The simpelest model 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 ($\epsilon C$) the energy penalty of a mismatch ($\epsilon I$) and the energy gain or penalty of the PAM ($\epsilon PAM$) to fix the energy landscape. It also contains two parameters to describe the time evolution of the system: the on-rate ($k_{on}$) and the attempt rate ($k_0$). The minimal model will build the foundation for this entire report.

The minimal model encompasses everything described in section 2.2. 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 intuitive sense, it is also the same quantity that is measured in several experiments like (Boyle et al., 2017) and [SOURCE]. By defining binding like this the minimal model has several features that depend mostly on the energy landscape, which is determined by the free energy parameters.

### 2.3.1 Energy landscape

The energy landscape of the R-loop determines the fraction of the dCas9 that will be bound in equilibrium and also determines, to an extent, how fast binding happens. As discussed before, the minimal model has three free energy parameters: one for a matching PAM, one for a matching base pair and one for a mismatching base pair. When we refer to an energy landscape we refer to the entire landscape, so for all possible states, that is predicted by these three parameters. An example landscape is shown in figure **??**.

By looking at the energy landscape it is possible to predict quite a lot about the behavior of the system. We will note several of these features and describe them by referring to several energy landscapes. These explanations will focus on building an understanding about the system as described by the minimal model, therefore they will not contain much mathematics but refer to several drawn energy landscapes instead.

### 2.3.2 Seed region

The first feature that is predicted by the minimal model, which is also a known feature of Cas9 is that the closer a mismatch is to the PAM, the less likely the sequence is to cleave [SOURCE]. One could assume then that the sequence is also less likely to bind than another sequence without mismatches or with mismatches far removed from the PAM. The region 'close to the PAM' is often referred to as the seed region. If there is a mismatch in this seed region binding is hard and most dCas9 will stay unbound. If instead a mismatch occurs at the end of a sequence dCas9 will readily bind to that sequence. The switch from low binding probability to high binding probability happens at the end of the seed region. This seed region can be easily identified in any energy landscape. In figure **??** an energy landscape is drawn with a mismatch at position seven, while the used parameters would predict a seed region of ten base pairs.

In figure **??** the mismatch occurs within the seed region and this can easily be spotted by the fact that the mismatch occurs before any bound state becomes more favorable than the solution state. Since the solution state is still the most favorable state most of the dCas9 will occupy the solution state in

equilibrium. If instead the mismatch would be placed outside the seed, as in figure **??** then several bound states will have a free energy lower than the energy of the solution state. These states will be more favorable in equilibrium and therefore dCas9 will mostly occupy these states in equilibrium.

The length of the seed region is then determined by the values of the energy parameters $\epsilon PAM$ and $\epsilon C$. An increase in $\epsilon PAM$ lifts the energy of all states, thus making the seed region larger as it takes more matches to drop the energy below the solution energy. On the contrary a decrease in $\epsilon C$ (which is an increase in absolute value since $\epsilon C$ is negative) will make the seed region shorter, as it will take less matches to revert the energy gain from the PAM. A decrease in $\epsilon C$ has another effect on the seed region: it will make the transition from low to high binding sharper. This is because with each matching base pair the energy drops an amount $|\epsilon C$. If this is large amount then crossing the solution energy threshold can happen within a single base pair and if it is even larger then the energy of the most favorable bound state can, within one match, go from a large positive (unfavorable) energy to a large negative (favorable) energy. If there is such a sharp transition the dCas9 will go from mostly occupying the unbound state to mostly occupying the most favorable bound state if the mismatch is moved one position to the back.

Note that this definition of the seed region is fundamentally different than the definition given in (Klein et al., 2017), even though the same model is used. The reason for this discrepancy is that in this thesis binding instead of cleavage is considered. For cleavage the transition energies (the peaks in the energy landscape) are important while for binding the free energies (the valleys in the energy landscape) are important. For binding these free energies are important because those are the energies that the R-loops will have when they occupy those states in equilibrium. Why exactly the transition energies are important for cleavage we will leave to (Klein et al., 2017), for now it suffices to say that it has to do with the fact that Cas9 has to be able to make it over these energy barriers to the end of the sequence to cleave the sequence but this is not a necessity to bind the sequence.

### 2.3.3 Minimum binding

A second feature of the minimal model is that there will always be a minimum amount of binding for any sequence with any amount of mismatches. If we take a look at an energy landscape of a sequence containing two mismatches (shown in figure **??**) it becomes clear why this is the case.

As can be seen in figure **??** even though there are two mismatches some of the bound states are quite close or even below the solution energy. In equilibrium these states will have a significant amount of binding, even though the sequence contained two mismatches. There are two ways to limit this binding despite mismatched bases. The first one is quite obvious: simply add more mismatches. This will increase the energy of all subsequent states even more, therefore decreasing the amount of binding to the point where it becomes negligible. A second way is to move the mismatches closer to the PAM. Even though the total energy gained and dropped is still the same, since the parameters did not change and the amount of mismatches is constant, the fraction of bound dCas9 can still change. That the total energy gained and dropped is the same is easily proven by the fact that the final state is at the same energy, no matter where the mismatches occur. However, the states before the mismatches are not all at the same energy. Sequences that have mismatches occur nearer to the PAM contain more states where the energy increase from the mismatch has an influence. This is due to the nature of the model; where the energy of each state is the sum over the energies of all previous states plus the energy gain or penalty of the state itself. This is very easy to see if two energy landscapes with the same amount of mismatches are overlapped as in figure **??**. The sequence with mismatches more to the end clearly has more states with a lower energy, even though the final energy is the same.

The minimal amount of binding in equilibrium, according to the minimal model, is therefore the amount of bound dCas9 on the sequence where all mismatches occur at the very start of the R-loop. In the case of two mismatches the sequence where mismatches occur at position one and two would bind the minimum amount of dCas9 that will be bound by any sequence where in total two mismatches

occur. Naturally then if the amount of mismatches increases, the minimum amount of bound dCas9 decreases.

### 2.3.4 Maximum binding

The minimal model not only predicts a minimum amount of bound dCas9 given a number of mismatches, it also predicts a maximum amount of bound dCas9. The reasoning is much the same as the reasoning for the minimum binding but reversed. The maximum amount of binding for a sequence with $n$ mismatches is reached when those mismatches occur in the last $n$ states. This is also very clear if we overlap two energy landscapes of two sequences with the same number of total mismatches, shown in figure **??**. The sequence which contains the later mismatches has more states with a lower energy, even though the final state again has the same energy in both cases. Since there are more states with a lower free energy, a higher fraction of dCas9 will bind to that sequence.

In the same way as with the minimum binding the maximum amount of binding can be decreased in two ways. Firstly it is possible to increase the number of mismatches. Secondly it is possible to move the mismatches to the end of the sequence. No sequence with $n$ mismatches can therefore bind more dCas9 than a sequence with those $n$ mismatches at the final $n$ positions.

### 2.3.5 Pair region

The previous features were all easily demonstrated by looking at the energy landscape since they all depended on single things. The last feature of the minimal model that will be discussed here is different because it depends on an interaction within the energy landscape. The pair region refers to a sort of second-seed region. The seed region discussed before is the region where a mismatch prohibits much binding of any sort and can be identified in the energy landscape as the point where the energy of the matching bases dips below the energy of the solution. Consider now that the effect of a mismatch is not as large as in figure **??** but smaller as in figure **??**. Then the final states are still below the energy of the solution state and therefore there will still be significant binding in equilibrium, even though there is a mismatch in the seed.

This changes if a second mismatch is added. If the second mismatch is placed directly behind the first one then there may be no state below the solution state, as shown in figure **??**, however if the mismatches are spaced apart a similar situation as with the PAM occurs. The energy of the R-loop starts near the energy of the solution state and is increased. This increase is now due to a mismatch instead of the PAM but there is an increase nonetheless. Then there may follow a number of matching base pairs which create a second opportunity to decrease the energy below the energy of the solution. If the second mismatch is placed after this second-seed region then there will be a state or several states which are again more favorable than the solution state, therefore increasing the amount of binding significantly. This second seed region is named the pair region. An example of a sequence where the second mismatch is placed after the pair region is shown in figure **??**.

The pair region is very similar to the seed region. It gets larger as the energy increase which causes it gets larger; the mismatch energy penalty in the case of the pair region. It gets smaller and sharper as the energy gain from a match becomes larger, all for the same reasons as for the seed region. In theory there can be multiple pair regions if there are more than two mismatches present, however the length of the gRNA is limited so there can not be arbitrarily many pair regions without extending the gRNA.

# 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. Therefore it will function as a starting point for all our research.

## 3.1 The first steps of the minimal model

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 $\epsilon PAM$ 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 data set was chosen because it has several advantages over other data sets. 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 data sets and test our fit for a different experiment.

### 3.1.1 Methods of (Boyle et al., 2017)

To better understand why we did certain things in our fitting procedure it is useful to understand how (Boyle et al., 2017) performed their measurements. Firstly, Boyle et al. (2017) used a DNA sequencing chip over which they let flow a solution of dCas9. They knew exactly where each sequence was located on their chip. This allowed them to do fluorescence experiments and link the fluorescence signal to a specific sequence, by checking the location of the signal.

Secondly, it is good to note that all the experiments they performed were fluorescence experiments. This means all their experiments had all the issues that every fluorescence experiment has (calibration inaccuracies, diminishing signal, photobleaching, etc. [SOURCE]). For the occupation experiments the fluorescence is easily linked to the occupation of a certain sequence. If the signal is more intense then more of the enzymes are bound. If the light sensor is perfectly calibrated this allows you to directly link the fluorescence to the occupation. This is not as trivial for the effective on- and off-rates. These rates were also measured by linking the fluoresence to the occupation, but after this some fitting is necessary to extract the rate.

The effective on- and off-rates are measured by measuring the fluorescence signal at a 500 second interval and fitting a straight line through these points. The line is also forced to go through zero as, logically, at the start of the experiment none of the Cas9 is bound. To fit such a line using least-squares the following formula can be used: [SOURCE: $https://online.stat.psu.edu/\ ajw13/stat501/SpecialTopics/Reg_thru_o$
]

$$\beta = \frac{\sum_i X_i Y_i}{X_i^2}, \tag{3.1}$$

where $X_i$ is the x-value of the i-th data point and $Y_i$ is the y-value of the i-th data point. Solving this equation will give you $\beta$: the slope of the straight line which is fitted to the data. This line is forced to go through zero, so the offset $\alpha = 0$.

It is well-known and also logical that the this linear increase in occupation cannot hold for all times; a natural binding and unbinding process will follow an exponential which starts out as a linear process but eventually flattens gradually into an equilibrium, see figure **??**. The mathematical justification for the linear approximation is provided in section 3.1.1. It will turn out this approximation is only valid for very fast times compared to the rates of the system.

### The linear approximation

In this section a brief justification of the linear approximation to the exponential binding curve of dCas9 will be given. Consider the system as described in the time-dependent section of the theory chapter (section 2.2.2). This describes the entire system for all times, but if we are only looking at the system for the very early times then no dCas9 will have made it to the end of this system. We can conclude that some of the states in our system can be ignored as they will have no effect on these short timescales. We can say that **if** we can approximate the unbinding of dCas9 as a single exponential, with a single rate, **then** we can model the system as a two-state system: one unbound and one bound state.

For the very short timescales it is easily justifiable that we can model the system with only one unbound and one bound state. All dCas9 is unbound at the start and after some average time the PAM is bound to the DNA. If the measurement stops before any dCas9 has the oppertunity to move on to the first bound state then the solution and the PAM state are naturally all the occupied states in our system. For slightly longer times where dCas9 has the time to form an R-loop it is still possible that the two-state approximation holds, as long as the bound states together behave as a single buond state, meaning they can be modelled as one state with a single dissociation rate, which must necessarily be different for every sequence.

The model has now simplified to the one in figure **??**. We can solve this simplified model with the master equation in the same way as in section 2.2.2. However, since this system only has two states instead of numerically it is possible to solve it analytically too.

First of all the total probablity to be either unbound or bound is one:

$$P_{bound}(t) + P_{unbound}(t) = 1. \tag{3.2}$$

Furthermore we know the time differential equations, since they are dictated by the rates:

$$\frac{dP_{bound}(t)}{dt} = k_f \cdot P_{unbound}(t) - k_b \cdot P_{bound}(t), \tag{3.3}$$

$$\frac{dP_{unbound}(t)}{dt} = -k_f \cdot P_{unbound}(t) + k_b \cdot P_{bound}(t), \tag{3.4}$$

rewriting the top one using equation 3.2 gives:

$$\frac{dP_{bound}(t)}{dt} = k_f \cdot (1 - P_{bound}(t)) - k_b \cdot P_{bound}(t), \tag{3.5}$$

$$\frac{dP_{bound}(t)}{dt} = k_f - (k_f + k_b) \cdot P_{bound}(t). \tag{3.6}$$

Solving this gives the solution for both $P_{bound}(t)$ and $P_{unbound}(t)$:

$$P_{bound}(t) = c_1 \cdot \exp(-(k_f + k_b)t) + \frac{k_f}{k_f + k_b}, \tag{3.7}$$

which is easily solved by substituting the initial condition for $t = 0$ to obtain the constant $c_1$. If we choose to start with all dCas9 in the unbound state we retrieve the exponential in figure **??** for the occupation of the bound state.

So this two-state approximation lets us analytically solve the master equation and model the total exponential. To get the linear approximation we need to go one step further. We can do a taylor expansion around $t = 0$ for the occupation of the bound state to get:

$$P_{bound}(t) = (c_1 + \frac{k_f}{k_f + k_b}) + (c_1 \cdot -(k_f + k_b))t + \frac{1}{2}(c_1 \cdot (k_f + k_b)^2)t^2 + ..., \tag{3.8}$$

if we then ignore terms of second order or higher and plug in the initial condition that $P_{bound}(t = 0) = 0$, we obtain the linear equation:

$$P_{bound}(t) \approx (c_1 \cdot -(k_f + k_b))t = k_f \cdot t. \tag{3.9}$$

Here $c_1 = \frac{-k_f}{k_f + k_b}$ because $P_{bound}(t = 0) = 0$. Equation 3.9 is the approximation used to calculate the effective forward rate in the experiments of Boyle et al. (2017). Note that we have used two approximations to obtain this result:

- The system can be approximated by only two states: one bound state and one unbound state. This approximation is valid only if the unbinding can be approximated by a single exponential unbinding curve with a single unbinding rate.
- We can neglect all second order and higher terms in the Taylor expansion of equation 3.8. This is valid only if $\frac{(k_f + k_b)t}{2} \ll 1$.

### 3.1.2 Observations from Boyle et al. (2017)

In the previous section we have reiterated how Boyle et al. (2017) performed their measurements, In this section we will take a look at their observations by primarily examining the three heatmaps they have provided: figures a, b and c in Boyle et al. (2017). These heatmaps show the effective on-rate, the effective off-rate and the occupation in equilibrium respectively. If the minimal model is to hold then it has to be able to explain the observations that are made in these heatmaps. Since the minimal model is as simplistic as possible it is unlikely that it is able to explain every feature in the data, however assuming that the general underlying model is correct, even the minimal model should be able to qualitatively describe most of the features found in the data. Let us first take a look at the effective on-rate.

**Effective on-rate**

First of all note that there is a sequence dependent effective on-rate at all. This does not comply with the requirements for the linear approximation we have established at the end of section 3.1.1. Specifically the second requirement $\left(\frac{(k_f+k_b)t}{2} \ll 1\right)$ does not hold, since evidently equation 3.9 does not hold. Therefore we are not in the purely linear, and therefore sequence independent, regime. Since the sequence can only alter the backward rate from the bound state there must be at least some influence of higher order terms. The second order term is:

$$\frac{-k_f}{2} \cdot (k_f + k_b)t^2. \tag{3.10}$$

Here we see that the influence of the backward rate is to decrease the amount of bound dCas9, as expected. Sequences with closer mismatches have to cross the mismatch sooner than sequences with later mismatches. Therefore they run into a barrier sooner and have to backtrack only a few steps to unbind, in contrast with the sequences which have a terminal mismatch. These terminal mismatch sequences have to backtrack almost the entire R-loop to unbind and are therefore less likely to unbind before making it over the mismatch barrier. Consequently the earlier the mismatch the higher we can approximate the backward rate from this single bound state. From Klein et al. (2017) we know that the probability to make it past the mismatch before unbinding follows a sigmoidal curve. If we assume it is almost impossible to unbind once a dCas9 has made it past the mismatch it is logical that the backward rate $k_b$ also follows a sigoidal curve, but inverted (starting high and ending low). Therefore the minimal model would also, very generally, predict a sigmoidal curve (from low to high) for the effective on-rate.

**Effective off-rate**

The analysis for the effective off-rate is made more complicated since the initial condition is not as simple as the one for the effective on-rate. The initial condition for the dissociation experiment is the equilibrium of the system. This equilibrium can easily be calculated using boltzmann factors as explained in section 2.2.1. It is hard to justify taking all these different bound states together and collpasing them into one single bound state, since the argument that only very little time has passed and the last states do no contribute anything does not hold anymore. If we simply ignore these issues for now we can at least try to compare the data to the simplified two-state model.

Boyle et al. (2017) make a remarkable discovery in their dissociation experiment: the existence of a special, unstable region between the seed region and the terminal bases. They name this region the reversibility determining region (RDR). This special region is characterized by an increased dissociation rate compared to the seed region when a the first mismatch of the sequence is placed in this region. In other words: the most unstable R-loop is formed when the first mismatch of a sequence is placed in the RDR.

This observation is not in agreement with the minimal model. We have established that the dissociation rate from the bound state is higher is the mismatch is at the beginning of the DNA sequence and gets lower as the mismatch is moved to the end. Therefore one would expect that the effective off-rate that is measured also decreases as the mismatch is moved towards the end of the DNA sequence. Note that the simplified two-state system can never explain both the effective on-rate and effective off-rate ersults from Boyle et al. (2017). If the actual on-rate is the same for all sequences then the difference in effective on-rate between sequences must stem from a difference in dissociation rate. This assumption is easily justified in the zipper model as the unbound dCas9 enzyme has no knowledge of the sequence beyond its interaction with the PAM.

The RDR is even more problematic than that as, even if we include all possible bound states, the minimal model always predicts that, the closer the mismatch, the more unstable the entire R-loop is.

Therefore the dissociation results are harder to explain than the association results.

#### Occupation

The final heatmap that is published in Boyle et al. (2017) is supplementary figure . Here the authors show the intensity of the fluorescence signal after twelve hours. For the purposes of this report it is assumed that the mixture of DNA and dCas9 is in local equilibrium after twelve hours. It is hard to justify this assumption definitively, but we will try to make it at least believable.

First of all, in order to function as an effective immune system for the bacteria who use CRISPR as a means to fend off virusses it is necessary that Cas9 is able to cleave the correct sequences quite quickly. If Cas9 is able to cleave the on-target then it must have reached the fully bound R-loop state. It is then a reasonable assumption that the on-target is able to reach equilibrium after twelve hours. Of course, adding mismatches slows down the entire binding process. However it is also known that Cas9 is able to cleave invasive virus RNA containing a mismatch as a means of defending itself. Therefore the mismatch barrier cannot be insurmountable on these, relatively short, timescales. It is then not far-fetched to also assume that double mismatches could be overcome in twelve hours.

Secondly, we assume that the system is in a local equilibrium, instead of a global equilibrium. This means that the individual clusters on the DNA sequencing chip are in equilibirum with their immediate surroundings, but the entire chip is not in equilibrium. This is harder to justify than the claim that the chip is likely in equilibrium. One argument is that, if the entire chip was in equilibrium, all double mismatch sequences would be outcompeted almost entirely by the on-target sequence. This is not what happens in the experiment and therefore it is unlikely that the entire chip is in global equilibrium. This argument ignores the effect of competition between sequences and assumes there is an infinite amount of on-target sequences (and all other sequences) to bind to, which is obviously not the case on the actual chip. Nevertheless one might expect a more skewed image, with much more fluorescence concentrated at the on-target and single mismatch sequences, if the chip was in a global equilibrium. A second argument in favor of a local equilibrium instead of a global equilibrium is that to establish a global equilibrium every dCas9 enzyme must have had the oppertunity to explore the entire chip and all available states. The diffusion coefficient of a small protein in water is about $100\mu m^2/s$. This means in twelve hours every protein has had the ability to explore $4.32mm^2$ of the chip. However while diffusing dCas9 also binds and unbinds to the DNA sequences it encounters, so this estimation is necessarily an overestimation. Since this explored surface area is on the same order of magnitude as that of the chip, combined with the fact our calculation is an upper limit it is not unreasonable to say that not every dCas9 has the oppertunity to explore all the available states.

The general zipper model is able to predict the fraction of bound dCas9 by calculating the boltzmann weight of every available state. Using these weights it is possible to calculate the probability to be in any state, given that the system is in equilibrium, as described in section **??**. The minimal model merely provides a way to prescribe the energy of each state so the boltzmann weights can be calculated. The minimal model predicts that the further down the mismatch, the more stable the R-loop and therefore a larger fraction of bound dCas9. It also predicts an interaction between mismatches which form a region analogous to the $n_{pair}$ region from Klein et al. (2017). Both these observations can be seen in the occupation heatmap from Boyle et al. (2017), albeit that the seed region is much more pronounced than the $n_{pair}$ region. This can however simply be a matter of parameter values.

### 3.1.3 Initial expectation of the minimal model

So far we have taken a closer look at how Boyle et al. (2017) measured their data and some of the conclusion they have drawn from that. We established that the minimal model is likely able to predict similar features as the ones shown in the association data and the occupation data. However the RDR in the dissociation data is not predicted at all by the minimal model. Furthermore we have established

that the association and dissociation data cannot simultaneously be explained by a simply two-state model. In the next section the minimal model will be put to the test and we will attempt to reproduce the results from Boyle et al. (2017) using the minimal model, starting with the occupation data.

## 3.2 Occupation fit

Since we expect the effective on-rate data and the occupation data to be more easily explained by the minimal model than the dissociation data and the occupation data contains two parameters less, the first fits will be on the occupation data from the supplementary information in Boyle et al. (2017). For the fitting of the minimal model to the occupation data the simulated annealing algorithm, described in chapter **??** is used.

### 3.2.1 Model for the fit

To fit the minimal model to the occupation data the model explained in section 2.2.1 is used. The minimal model assigns a specific energy to each state in the model based on whether the base pairs match and every match or mismatch before. With these energies it is possible to calculate the equilibrium distribution of dCas9 over all available states with the Boltzmann weights. It is therefore possible to calculate the probability to be bound given any sequence.

### 3.2.2 Data for the fit

The provided occupation data from Boyle et al. (2017) contained the fluorescence intensity of single sequences before association and before dissociation. Of course the fluorescence intensity is linked to the amount of bound dCas9, however the nature of fluorescence necessitates some form of normalization before the measured intensity reflects the actual fraction of bound dCas9. Furthermore it is helpful to be aware of the fact that there is a non-negligible amount of noise in the measurements. To circumvent these issues with the data, the data was transformed before the fit. First the pre-association intensity was subtracted from the pre-dissociation value for each data point separately. Since before any association no dCas9 should be bound the pre-association intensity value is attributed to noise. After this the median intensity was calculated for every sequence. Since the variance in each sequence is quite large, taking the median and fitting to that will most likely produce better results as the noise is reduced, assuming the noise is symmetric. Finally the median intensity value for each sequence is divided by the median intensity value for the corresponding on-target sequence. This allows us to see how much each sequence binds compared to the on-target sequence. This ratio is no longer in arbitrary units; no matter the units the ratio is the same. Another advantage of this ratio is that it makes the results easier to interpret. The on-target should, logically, bind the most out of any sequence so every mismatched sequence should result in a ratio of less than 1 and the closer to 1, the less impact the mismatch has had on the binding.

### 3.2.3 Fit results

The fit was performed nultiple times on replica 1 and replica 2 from Boyle et al. (2017) separately. Furthermore the fit was also repeated several times on different subsets of data. This allows for testing of the model within the total data set as well as testing on other data sets. We will start by showing the fit on the total data for replicas one and two separately; these results can be found in figures ... and table ...

Figures ... may be even more useful. They show the relative and absolute error in the prediction. When examining these errors it is important to realize that the fitting function that was used to obtain the model parameters minimizes the *absolute* error, not the relative error.

Several specific features of these results will be looked at in greater detail in following sections. One surprising result we would like to bring to your attention now however is the positive value for the PAM energy. It turns out that the best fit to the data is one where binding to the PAM is unfavorable compared to the solution state. This is allowed within the minimal (and general) model as it violates no assumption that was made in section 2.2. It was mentioned that a matching PAM sequence should be favorable compared to a non-matching PAM sequence. Since this can still be the case a positive $\epsilon PAM$ is allowed within the model. In section ... it is also explained that the PAM, and therefore all energies, are concentration dependent. Thus the actual PAM energy in-vivo could possibly be much lower.

Note that we have not allowed for a positive $\epsilon C$. While one could argue that a match should merely be favorable compared to a mismatch and is not necessarily an absolute energy decrease, we would like to justify our assumption by reminding people of the origins of Cas9. Cas9 is naturally used to protect bacteria against viruses. If a match would be an absolute energy increase, it would be unlikely or take long to cleave any target, including the on-target. Therefore Cas9 would not be very effective as an immune system if a match would be an absolute energy increase. Finally one could also argue that there is an energy increase but it is tiny. If that was the case however the fitting algorithm would most likely have found a very low decrease in energy to match the reality as well as possible. Since the optimal parameters are not near zero it is probable that a match is an energy decrease.

**Seed region**

In the fit the seed region corresponds to the seed region in the data. In both experiment and theory the shift from a low to a high occupation occurs around base pair ten. In the theory this is very naturally explained by the fact that around base pair ten the energy of the formed R-loop dips below the energy of the solution. This is illustrated in figure **??**...

Figure **??**.. provides a nice geometric and intuitive way to see the transition from the low occupation to the high occupation. It is also possible to make some simplifications and calculate the approximate gain in occupation for each subsequent matching base. It is known that in equilibrium the occupation of each state is given by

$$P(s) = \frac{\exp(-E_s)}{\sum_{s=-1}^{s=20} \exp(-E_s)}, \tag{3.11}$$

where $s$ is the state, numbered $-1$ to 20 in this case. The ratio between the occupation of two states is

$$\frac{P(a)}{P(b)} = \frac{\exp(-E_a)}{\exp(-E_b)}. \tag{3.12}$$

If we assume that the energy gain for a match is large and that $a$ and $b$ are subsequent states ($b = a+1$), which are both matches, then the ratio of the occupancy becomes:

$$\frac{P(a)}{P(b)} = \frac{\exp(-E_a)}{\exp(-(E_a + \epsilon C)} = \frac{1}{\exp(-\epsilon C)} = \exp(\epsilon C) \approx 0, \tag{3.13}$$

where the last approximation is made since $\epsilon C$ is large and negative. In this case all dCas9 would occupy state b instead of state a, we will call state b the dominant state.

If we assert, on top of the assumption that $\epsilon C$ is large, that the first mismatch in a sequence occurs at position $n + 1$ then there will be two dominant states competing for the dCas9: the solution state and state $n$. This is because $\epsilon PAM$ is positive, allowing the solution state to be the dominant state

even though the PAM and the first $n$ base pairs matched. The ratio between the two dominant states is then:

$$\frac{P(-1)}{P(n)} = \frac{1}{\exp(-E_n)} = \exp(E_n), \tag{3.14}$$

so with these assumptions the ratio of unbound to bound decreases exponentially as the first mismatch gets pushed back further, since $E_n$ becomes lower and lower. More exactly: if the mismatch is pushed back one base pair then the ratio of the unbound to the bound fraction decreases with a factor $\exp(\epsilon C)$. This also proves that, as the geometric insight already implied, once the energy dips below the energy of the solution state the occupation rapidly increases, which marks the end of the seed region.

### Stripes

A significant difference between the prediction of the minimal model and the experimental results are several 'stripes'. These features are rows and columns in the occupation heatmap for which the minimal model predicts a low occupation, while a high occupation is measured by Boyle et al. (2017). Less pronounced are also the opposite stripes, where the model predicts a high occupation while the experiment measures a low occupation. These features undoubtedly exist within the data but can never be predicted by the minimal model. The minimal model does not allow for an increase in occupation followed by a decrease in occupation or vice versa; the minimal model is monotonic. The reason for this is two-fold. Firstly, every match or mismatch affects the energy of all subsequent states. This is however true for all models that fall within the general zipper model described in chapter **??**. The reason the minimal model is monotonic is because the minimal model only has one single parameter for a match and another single parameter for a mismatch. Therefore the minimal model can not contain any local effects. Coupled with the first point the minimal model is necessarily monotonically increasing in occupation versus the first mismatch position.

The stripes visible in the data thus point towards some very fundamental properties of the energy landscape which determines the binding of dCas9. There need to be some local and/or neighbor effects incorporated into the model to better predict the results from Boyle et al. (2017). These stripes, especially the one where a mismatch occurs at position two, also show up in different data sets.

A simple addition to the model could be a position dependent $\epsilon I$. For example, if the $\epsilon I$ at position two is much lower than other $\epsilon I$, a mismatch at position two would prohibit binding much less than mismatches at other positions, resulting in a higher occupancy. This is a simple addition to explain a single striped pattern. There are other models that would be able to explain this striped pattern, we merely bring up the position dependent model to show the stripes could be explained by a more complicated energy landscape.

### Upper plateau

A third feature that is apparent in the data but not in the model is hidden in the region where the occupancy is highest. In this region the model predicts an occupancy of almost 1, meaning the same occupancy as the on-target. This is a logical prediction within the minimal model as a double mismatch at positions nineteen and twenty would still have 18 matching base pairs before any mismatches occur. If the seed region extends to base pair ten then ten matching base pairs approximately bring the energy of the R-loop down to the energy of the solution, resulting in about an even split between bound and unbound dCas9, as we have seen in section 3.2.3. If another eight matching bases are then added beyond the seed region the bound fraction is approximately $0.5 \cdot \exp(-8\epsilon C)$, which approximately equals 1 if $|\epsilon C|$ is not absurdly low.

The model is quite clear with its results in this high occupancy region, but the data is not so clear. According to the data these sequences with mismatches at the final few bases are actually easier to bind to than the on-target; resulting in measured occupancies higher than 1. This is strange behavior when you consider the original role of Cas9 in the immune system of a bacterium. Logically one would say that Cas9 would be best at cleaving the on-target sequence, as that is the sequence which corresponds exactly to the gRNA. This property would also make it perform best at defending against invading viruses which match the gRNA sequence. Of course it could be that these sequences are more favorable to bind than the on-target sequence but it would contradict some of the assumptions we have made at the very start for the general zipper model. These abnormally high occupancies could be explained be an (position dependent) energy penalty for a match or a (position dependent) energy gain for a mismatch. However other models which would not contradict the general assumptions so blatantly could perhaps also explain this behavior. Effects from neighboring bases, sequence related effects or protein related effects can all also be causes for the increased occupancy of these off-targets. Finally, although the data quality is good the increased occupancy measurements could also simple be the result of noise. If the noise is relative to the signal strength then this would explain why this effect is less apparent in the low occupancy region. However the increased occupancy seems to be consistent throughout the entire high occupancy region, which is not what you would expect if the effect was purely caused by symmetric noise.

#### Pair region

At the very edges of the heatmaps there are again regions with a higher occupation even though they contain mismatches within the seed. These regions are the pair regions where the mismatches interact in such a way that there can still be significant binding even though there is a mismatch early in the sequence. This 'second seed region' effect is described in 2.3.5. For some sequences it can easily be understood; for example a sequence with mismatches at positions one and twenty. This sequence essentially behaves as a sequence with only a single mismatch at position one, since the mismatch at position twenty has a negligible effect on the binding. For other sequences it is less clear and size of the pair region is actually determined by the ratio of mismatch energy penalty to match energy gain. In the case of the minimal model for replica one this ratio is .. Therefore the pair region should be another .. bases after the seed region, which is indeed the case.

## 3.3 Effective on-rate fit

Fitting to the occupation data fixed all energies in the minimal model. Knowing these three parameters allows you to predict the equilibrium state of the system, however to predict the time evolution of the system more parameters need to be known, as discussed in section **??**. These parameters are the 'on-rate' and the 'attempt rate' of the system: the rate of binding from solution to PAM and the rate of moving from any bound state to any other state respectively. In this section these two parameters will be fit to the effective on-rate data gathered by Boyle et al. (2017).

### 3.3.1 Model for the fit

The model that is used will be based on the master equation which is described in section **??**. To be able to fit the model to the data from Boyle et al. (2017) it is necessary to model the experiment as accurately as possible. The details of the experiment are repeated in section 3.3.2. To model the experiment as accurately as possible we followed the procedure:

1. Nothing is bound at the start of the experiment.
2. The occupancy is calculated at three evenly spaced timepoints.

3. The bound fraction at each timepoint is calculated by summing over the occupancy of each state except the solution state.
4. A straight line is fitted through the bound fraction values. This line is forced to go through $(0, 0)$.
5. The slope of the fitted line is the effective on-rate.

This is the procedure followed by the experiment translated to the minimal, time-dependent, model. Of course in the experiment step two and three are simply replaced by measuring the intensity, which should be equal to the occupancy of the bound states up to some constant factor and noise.

The general procedure is clear, however some details of the model still need to be determined. We will go through the general procedure step by step and see what each statement means for the calculation. The first step is that nothing is bound at the start of the experiment, you start off clean. This is easily translated to the model, where it is possible to simply intialize the occupation vector with all of dCas9 in the unbound state. The occupation vector in this context is the vector which contains the occupancy of each state. Assuming the first state of the occupation vector is the solution state we initialize the vector as:

$$\vec{P}(t = 0) = [1, 0, 0, 0, ..., 0, 0]\,. \tag{3.15}$$

The second step of the general procedure is to calculated the occupancy of each state at three evenly spaced timepoints. Within this step essentially the entire minimal model is hidden. To calculate the occupancy at different timepoints two things are needed: the occupancy of each state at a single timepoint, the initial condition, and the rate matrix, which describes the evolution of the occupancy vector through time. If these two things are known it is possible to evaluate equation **??** at any timepoint. Luckily the initial condition is known from the previous step. The rate matrix is not entirely known, it depends on both the energy landscape and the rates from the minimal model. In this fit the energy landscape is fixed from the occupancy fit, however the rates are still free. The rates are therefore the parameters that will be fitted. In the minimal model we will make the assumption that all attempt rates are equal as described and justified in section **??**. This reduces the number of rates that need to be fitted to only two. Knowing the energy landscape and the fitted rates allows us to calculate the occupancy at the three different timepoints and move on to step three. Step three, four and five are all relatively straightforward. The occupancies that are calculated in step two will be used to calculate the bound fraction of dCas9. In this case again it is assumed that dCas9 is bound when it is not in the solution state. In step four a line is fitted to the occupancies, this fitting procedure is described in section **??** and is the same fitting procedure as is carried out by Boyle et al. (2017). Finally in step five the effective on-rate is obtained from the model which can be compared to the measured effective on-rate.

### 3.3.2 Data for the fit

The provided effective on-rate data from Boyle et al. (2017) contains the slopes of the intensity curve for each sequence. To extract a slope of the fluorescence curve per sequence out of the fluorescence intensity for each single DNA strand the median fluorescence value per tile was taken. This was then normalized by the median fluorescence of the on-target after 12 hours of binding. After 12 hours of binding the on-target is expected to be fully bound and this is confirmed by the experiment. The median intensity value per tile is used to *'account for systematic differences in focus, cluster formation efficiency and illumination across tiles'* (Boyle et al., 2017). The intensity values are then normalized by the on-target fluorescence in equilibrium to convert the intensity to an occupation. This normalization by the on-target intensity in equilibrium is absent in the minimal model that was used as the model is able to directly work with occupancies and has no need for any conversions. As long as the bound fraction of dCas9 for the on-target in equilibrium is approximately equal to one this is a justifiable omission and as confirmed in section 3.2.3 this is indeed the case.

### 3.3.3 Fit results

The fit was performed on replica 1 and 2 from (Boyle et al., 2017) separately and repeated several times. The results of the fit are more conclusive than the the results of the occupation fit, as can be seen in table **??**. The fit results are shown as a heatmap in figures **??** and **??**. As with the occupancy fit the absolute and relative errors are also plotted in heatmap format in figures **??** and **??**. Again, these fits were performed with the simulated annealing algorithm described in chapter **??** and minimize the absolute error.

The effective on-rate results contain less features in their heatmap, for the data as well as for the fit. Some form of seed region can indeed be seen in the data and there is definitely some position dependence, on account of the striped pattern also present in the occupation fit. This position dependence is not contained within the minimal model. However it may be more interesting to discuss the features that are absent in this effective on-rate heatmap: the pair region where the two mismatches interact.

Note that the attempt rate between fits is really consistent but the on-rate is not. The reason for this becomes clear if we draw the $\chi^2$ landscapes for both parameters leaving one fixed (figures **??** and **??**). Both curves have the shape of a Lennart-Jones potential, however the flat part for the attempt rate is higher than its lowest point at the predicted attempt rate. For the on-rate the valley seems to be missing entirely and the sharp decrease at the start is simply followed by a flat line which extends far. This suggests that the on-rate simply needs to be large compared to the attempt rate in order to have a good fit; the actual value does not matter as much. This is also quite logical since the on-rate determines only two things: the rate at which dCas9 moves from solution to the PAM state and the probability to go from the PAM state to the first base pair state. Once the on-rate is large compared to the attempt rate these two effects cancel out. If the on-rate doubles the probability to go from PAM to first base pair is halved, but at the same time the time it takes to go from solution to PAM is also halved, meaning there are twice as many opportunities to move past the PAM state. This effect ensures that as long as the on-rate is large compared to the attempt rate its exact value does not matter as much.

#### Seed region

It is clear from section 3.2.3 why the occupation heatmap should show a seed region, however this does not explain why the effective on-rate map should also contain such a seed region. When out of equilibrium the binding to each sequence does not know how many states there are below the energy of the solution state down the line. The most important interaction of the system is the one of the binding base pair. The reason a seed region is still visible is therefore unlikely to be due to some global effects of the system. In other words the actual on-rate from solution to bound PAM is the same for every sequence with the same, matching, PAM. The reason a seed region is visible is because we are measuring the effective on-rate. The dCas9 has the time between measurements to move between several states. This allows dCas9 to bind to the sequence, go several base pairs in and make it across a mismatch barrier or be bounced back from the mismatch barrier. Essentially dCas9 has the time to, in between measurements, explore the energy landscape. This allows the enzyme to feel the interaction of more than the first few base pairs and can result in some form of seed region. To make it intuitive why this exploration of the energy landscape can lead to some form of seed region it might be informative to push some of the energy landscape parameters to the extreme. Consider the case where $\epsilon I = \inf$ and $\epsilon C \ll -1$. Here the PAM pushes the energy of several states up and the energy gets a lot lower whenever another match is added to the sequence. However when dCas9 encounters a mismatch it is impossible to make it over the mismatch energy barrier. The enzyme has only one choice: to unbind again. Assume now that the attempt rate is fast $k_0 \gg 1$, then once an enzyme is bound it quickly moves down the sequence until it encounters a mismatch. When it does it will move between the bound states and unbind eventually. If this insurmountable mismatch occurs in the seed region, on average almost no dCas9 will be bound since the energy of the most stable state is above

the solution energy. However if the mismatch is moved down the sequence on average more and more dCas9 will be bound since the most stable state becomes more and more stable. Eventually the most stable bound state is more stable than the solution state and there is a transition from a high solution occupancy to a high bound occupancy. In the measurements of the effective on-rate you will measure a very slowly increasing occupancy for sequences with seed mismatches and then at some point a much quicker increasing occupancy for non-seed mismatches. In this example every sequence binds equally fast but the maximum binding of the non-seed sequences is simply much higher, resulting in a higher effective on-rate. This effect is also seen in a slightly more complicated manner in the minimal model. Here the mismatch barrier is not infinite so there is an interplay between time to bind and maximum binding of a sequence but the core ideas remain the same.

#### Absence of pair region

Since there is some sort of seed region present on the effective on-rate data one might suspect that there is also some sort of pair region. In the data however no such pair region is visible and in the same way the model does not predict such a pair region. Why does the pair region not appear in the heatmap while the seed region does? That has everything to do with the time dependent nature of the problem. The experiment and the simulation run for some time and measure the bound fraction of dCas9 at three timepoints. After the third timepoint the measurement is stopped. This inherently limits how much of the energy landscape the enzyme is able to explore since essentially it only has a limited amount of transitions it can make. For the pair region to appear however there is one more requirement. One of the transitions that has been made necessarily is a transition past a mismatch, for the pair region is the region where the two mismatches interact. Since the mismatch barrier is much larger than the match barrier it takes much longer to pass a mismatch than a match. This means that when the total time of the experiment is shortened the pair region starts to shrink, as the dCas9 no longer has the option to explore the energy landscape past a mismatch. In the extreme cases this can even lead to shrinking of the seed region since dCas9 also needs time to explore the matching energy landscape. Severely limiting the time of dCas9 to do so will lead to everything past a certain number of matches to be labeled as having the same effective on-rate. This can be confirmed by the obtained minimal model, as shown in figure **??**.

## 3.4 Effective off-rate results

At this point two separate fits have been done: the first on the occupation data, which fixed the entire energy landscape, and the second on the effective on-rate data, which fixed the rates. With these fits completed the entire minimal model is determined. The effective off-rate data thus only allows us to test the obtained model. Since the dissociation data has not been used in any fit it can provide powerful clues as to where the minimal model fails.

### 3.4.1 Model

To model the dissociation the same general principle as for the association model is used; the master equation is solved numerically for three evenly spaced timepoints. There are however two important differences:

1. The initial condition is the equilibrium distribution as opposed to everything unbound.
2. The fitted line is not forced through $(0, 0)$.
3. Before starting the simulation the occupancy of the solution state is set to zero and the occupancy vector is re-normalized.
4. The on-rate is set to zero.

The first point is simply because the dissociation experiment is started after twelve hours of binding so the system is in equilibrium. The second point is because Boyle et al. (2017) also do not force their fit through $(0, 0)$ for the effective off-rate fit. The third and fourth point are more fundamental. The occupancy of the solution state is set to zero since the experimenters flush out all dCas9 in their experiment before starting the dissociation experiments. It is assumed that this flushing out happens perfectly and no dCas9 remains in the solution state. After setting the solution state to zero in the model the occupancies are re-normalized so they sum to one again, as probabilities should. The fourth point is the most disputable: the on-rate is set to zero. If the occupation of the solution state is negligible almost no dCas9 should rebind to the DNA in theory, however in practice this is hard to guarantee. More compelling is the argument that Boyle et al. (2017) also performed an experiment where they introduced a surplus of unlabeled dCas9. This meant that the labeled dCas9 had to compete with unlabeled dCas9 to rebind to the DNA. Since there was a significant surplus of unlabeled dCas9 it is unlikely that much of the unbound, labeled, dCas9 was able to rebind before any unlabeled dCas9 could claim the free spot on the DNA. This effectively sets the rebinding rate (or on-rate) of the labeled dCas9 to zero. Since only the labeled dCas9 is measured this is equivalent to setting the on-rate to zero for the simulation.

### 3.4.2 Data

The effective off-rate data that was provided by Boyle et al. (2017) was also obtained by fitting a line through the fluorescence signal, similar to the effective on-rate data. The main differences are in the normalization of the data. For the effective off-rate data the fluorescence at each timepoint was divided by the median on-target fluorescence at that timepoint. This is to *'account for additional variation in signal assuming that the dissociation of dCas9 from its canonical target is negligible'* (Boyle et al., 2017). For the simulation this adjustment is not necessary since for the calculations there is no variation in signal as the formalism is exact and we are not dealing with intensities but with actual probabilities. Secondly the data from the experiment was normalized by the first datapoint in the series *'such that the corrected values represented the proportional decrease in fluorescence signal.'* (Boyle et al., 2017). This normalization is done in the model, albeit slightly hidden. It happens when the occupancy of the solution state is set to zero and the remaining occupancies are re-normalized. This re-normalization is the same normalization as normalizing the fluorescence signal by its first datapoint.

### 3.4.3 Results

The heatmaps of both the data and the model results can be seen in figures **??** and **??**. The data and the model both contain some form of so-called Reversibility Determining Region (RDR). The fact that the model produces any form of RDR is surprising at first since in the minimal model the first bound states are quite explicitly less stably bound than the later states. If you would place a mismatch later in the sequence you would therefore expect the resulting R-loop to be more stable, why then does even the minimal model produce such a RDR? The answer is two-fold but it again has everything to do with the fact that the experiment is inherently time-dependent.

It is important to realize two facts about the formed R-loops in the minimal model and they both have a connection to meta-stable states. Meta-stable states are states just before a mismatch. These states have a lower energy than the states before them and also a lower energy than at least some of the states after them due to the mismatch. They are however not necessarily the most stable states in the system. Consider the energy landscape of a sequence with a single mismatch at position five, so in the seed region and outside the RDR (figure **??**). State number four has a lower energy than all states before it and most states after it, however it is neither below the solution state nor is it the most stable bound state in the system, since state twenty has a much lower energy still. In equilibrium there are therefore three competing dominant states: the solution state, state four and state twenty. Obviously in equilibrium most dCas9 will reside in state twenty, a small fraction in state

four and a slightly larger fraction in the solution state. However now consider what happens when the dissociation experiment is started. The small fraction of dCas9 in state four is able to quickly unbind as it only has to overcome a small energy barrier to reach the PAM state, from which it is easy to unbind. Therefore a small fraction of bound dCas9 quickly unbinds. However the majority of dCas9 is still bound at state twenty. These bound enzymes unbind much slower since they have to overcome a much larger energy barrier to reach the PAM state. Since the fluorescence measurements of Boyle et al. (2017) are only done once every 500 seconds the quick unbinding from the very unstable state four is missed since it happens so quickly **and** is a small drop. This effect can clearly be seen in figure **??**, where the under-sampled dissociation curve with only three points and its fit is plotted together with the actual dissociation curve according to the minimal model.

If we compare the scenario for a sequence with a mismatch at position four to the scenario for a sequence with a mismatch at position ten, so in the RDR, we clearly see the difference. First let us take a look at the energy landscape of a sequence with a mismatch at position ten (figure **??**). Obviously the meta-stable state is now at position nine, but that is not the only difference. Since the mismatch is placed at position ten the meta-stable state now has a much lower energy than the one in the previous example. This means it is much more stable but also contains a much larger fraction of all bound dCas9. When the dissociation experiment is started this has two effects. The dCas9 unbinds slower from the meta-stable state, but still faster than unbinding from the most stable state (state twenty). There is also a much larger drop in the bound fraction of dCas9 due to unbinding from the meta-stable state since it contained so much more of the bound dCas9 in the first place. These two effects together mean that the effective dissociation rate is no longer under-sampled, as can be seen in figure **??**. Essentially the difference between the two curves that are shown here is that one is under-sampled and one is not. Therefore the dissociation from the most stable state is measured in the case of a mismatch very early in the seed region, while the dissociation from the meta-stable state is measured if the mismatch is further down the sequence in the RDR. The RDR is therefore, according to the minimal model, not a physical region for dCas9, it is merely a result of this experiment with these exact parameters, especially the time interval between measurements.

That being said, it is clear that while the minimal model does predict some sort of RDR phenomenon it does not predict the region at the same location in the heatmap and it also does not have the same shape as the RDR in the data. This shows that the minimal model on its own is not capable of explaining all features of the data that is provided in Boyle et al. (2017). This dissociation data and the discrepancies of the model can point towards some underlying fundamental features of the energy landscape which we will discuss in a later chapter.

## 3.5 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:

$$K_D = \frac{1}{\sum_{n=0}^{N} \exp\left(-\Delta F(n)\right)}, \tag{3.16}$$

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:

$$\begin{aligned} \delta C &= -1.41, \\ \delta PAM &= -2.02, \\ \delta I &= 8.87, \\ k_0 &= 1000, \\ k_{on} &= 0.425 \text{ nm/s}, \\ [Cas9] &= 1 \text{ nm}, \end{aligned} \tag{3.17}$$

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

# Simulated Annealing Algorithm

To fit parameters to the data a fitting algorithm had to be chosen. For almost all of the fitting we used a simulated annealing algorithm. In this chapter we will explain the inner workings of this algorithm and we will explain why we specifically used this algorithm.

## 4.1 Basic principles of simulated annealing

Kirkpatrick et al. (1983)

Simulated annealing is a form of an optimization algorithm. It draws inspiration from statistical mechanics. In statistical mechanics systems with many particles and therefore many degrees of freedom reach a thermal equilibrium. This behaviour is similar to a function with many degrees of freedom for which one tries to find the global minimum.

In statistical mechanics a system can be in a finite number of states. Each of these states has a specific energy associated with it. In thermal equilibrium the probability to be in any of these states follows the boltzmann distribution, as we have discussed in section 2.2.1. Naturally then, if a system is cooled down to a low enough temperature the only occupied state will be the state with the lowest energy. In the optimization analogy this lowest energy state would be the minimum of the function we are trying to minimize. There is one catch to this idea; if a system is cooled down fast enough it will reach a local equilibrium but not a global equilibrium. As an example take a liquid which is cooled down so it forms a solid. If the cooling is quick, the solid will form but it will have many defects. Essentially the molecules are simply locked into place. This configuration does have less energy than the liquid state, however it is not the ground state of the system. If instead the cooling process is very gradual the liquid has the time and oppertunity to form a perfect crystal. This crystal is the actual ground state of the system. This gradual cooling is the 'annealing' part of the simulated annealing algorithm. It refers to the very gradual annealing of a crystal from a melt. By 'cooling down' the cost function gradually we are able to obtain better solutions than just the local minima.

That there is an analogy between minimizing the cost function and cooling down an ensemble of particles to its ground state is clear. However, we conveniently left out the temperature analogy up to this point. While clearly there is a temperature when cooling down a physical system, the equivalent of the temperature or the energy of the system is not as clear in the case of the cost function. In fact, simulated annealing algorithms explicitly introduce this temperature. When simulated annealing finds a better solution to the cost function, it accepts that solution until a better one is found. If only these better solutions were accepted that is akin to flash-freezing our system, it will simply move towards the nearest local minimum. Instead, if the algorithm discovers a worse solution it still has some chance to accept this solution. The probability to accept a solution is given by the following formula:

$$P_{accept} = \exp\left(-\frac{C_1 - C_0}{k_B T}\right), \tag{4.1}$$

where $T$ is the temperature, $k_B$ is the boltzmann constant, $C_0$ is the cost of the old solution and $C_1$ the cost of the new solution. The consequence of choosing this particular probability function is that the system will evolve into a boltzmann distribution over time. It is clear that if we lower the temperature in this case, it becomes harder for worse solutions than the current one to be accepted.

## 4.2 Simulated annealing protocol

The simulated annealing protocol is simple:

1. Define some cost function.
2. Initialize the algorithm by giving it a starting guess for the parameters and an initial temperature.
3. Vary the parameters slightly and calculate the cost.
4. Accept or reject the new parameter set based on equation 4.1.
5. Repeat step 3 and step 4 a number of times.
6. If the stop condition is not reached; lower the temperature and return to step 3. If the stop condition is reached continue to step 7.
7. Return the final parameter set; this should be the minimum of the cost function.

Although this protocol is simple, there are several points which the designer of the algorithm has to take into consideration. First of all the cost function should be defined. This function is the function that is ultimately minimized (or maximized) and should be representative of the problem which is solved. Secondly the user has to initialize the algorithm with an initial guess and an initial temperature. In theory this initial guess is not important, but the initial temperature is. If the initial temperature is too low, the cost function could already be in a local minimum from the beginning and therefore never reach the global minimum of the cost function. The intial temperature should therefore be high enough that the algorithm has the oppertunity to explore all possible parameters. Thirdly the parameters should be varied slightly when choosing a new parameter set. The issue is that 'slightly' varying the parameters is relative and could also change during the running of the simulated annealing algorithm. Fourthly, the parameters should be changed a a certain number of times before lowering the temperature to simulate the waiting and settling down of the system before continuing to cool it. This effectively makes the algorithm behave quasi-equilibriated. The user should determine how long the algorithm should wait before lowering the temperature again. Finally the user should also determine the stop condition. A simple stop condition could be to stop the algorithm when it reaches a low temperature, however what constitutes a low temperature can change depending on the cost function, the data, etc.

In summary, even though the protocol is simple in essence, there are several things the user needs to consider before using a simulated annealing algorithm.

## 4.3 Version used in this thesis

To overcome the issues laid out in section 4.2 the version of simulated annealing that is used in this thesis uses additional metrics and safeguards to ensure the solution found by simulated annealing is the optimal solution.

– Sidenote: the algorithm is not perfect, fitting is hard. –

### 4.3.1 Cost function

The cost function is at the heart of the simulated annealing algorithm. Since we are fitting to experimental data the cost function has to be representative of how well the model matches the input data, given a certain parameter set for the model. The metric we use for this is $\chi^2$:

$$\chi(x)^2 = \left( \frac{model(x) - data(x)}{error(x)} \right)^2 . \tag{4.2}$$

Here *model* refers to the prediction of the model at point $x$, *data* refers to the value of the data at $x$ and *error* to the error of that specific datapoint.

# 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.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, 220(4598):671–680.

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