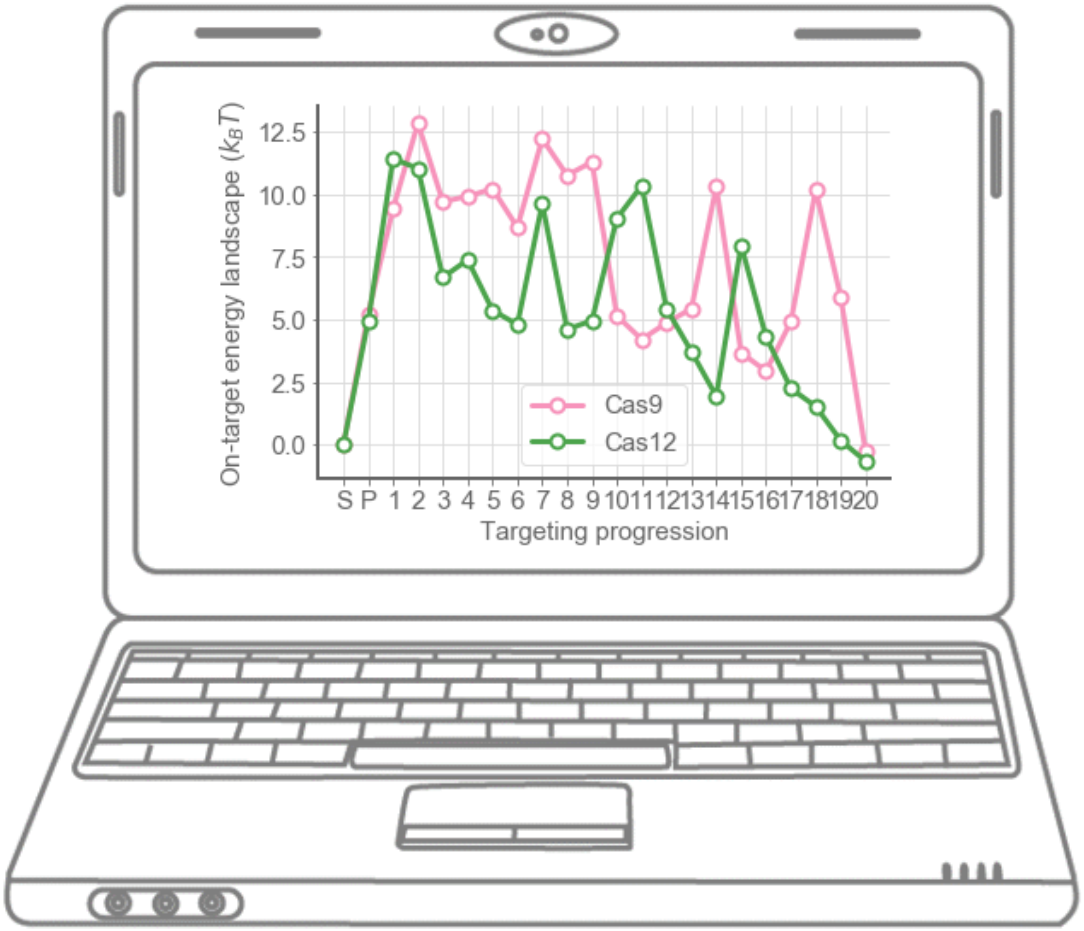
Predicting off-target binding of CRISPR-Cas systems

Diewertje Dekker

July 2019



TU Delft

A Model to predict off-target binding of CRISPR-Cas systems

by

Diewertje Dekker

In partial fulfilment of the requirements for the degree of

Bachelor in Science
in Nanobiology

At Delft University of Technology
and Erasmus University Rotterdam,
To be defended publicly on July 1st, 2019

Student number (TUD):	4592190	
Student number (EUR):	454854	
Project duration:	February 11th – July 1st, 2019	
Thesis supervisors:	MSc Misha Klein	
	Dr. Behrouz Eslami Mosallam	
Thesis committee:	Prof. Dr. M. Depken	TU Delft, supervisor
	Prof. Dr. H. Youk	TU Delft

Cover figure: Adjustment on a figure from [1]

TU Delft

ABSTRACT

The CRISPR-Cas system is used as a gene editing tool in many fields like biotechnology and clinical medicine, because it is easily programmed to target any sequence. However, one occurring problem is that CRISPR-Cas is not 100% precise. Next to the target sites, off-target sites get targeted, resulting in that it is unknown at which places the DNA will get cut. This hinders broad clinical application.

To understand what governs CRISPR-Cas fidelity, we need to understand the physics of CRISPR-Cas. Our group established a physical mechanism based on fitting to high-throughput data from Boyle et al. [2]. The fits are done by modelling the experiments, to obtain parameters that describe the free energy landscape of CRISPR-Cas upon binding to a DNA sequence.

For this thesis I trained the model on the dataset of Finkelstein et al. [3], which was obtained by a different experiment and performed for multiple different Cas proteins. The data of Finkelstein turns out to lack some information to judge if a Cas protein will be bound to a target site, after a certain amount of time, using a given concentration. Nonetheless, I confirmed that this data contains enough information to observe the same characteristics for Cas9 as we did with the dataset of Boyle. This allows us to apply our model on other Cas proteins as well, as will be showed for Cas12.

Although fitting our model to the data was successful, the fitting process is very demanding. During this thesis we will look at the validity of all the assumptions made, point out which information is lacking in the data and give some suggestions on how to obtain this information by combining data sets.

TABLE OF CONTENTS

Introduction:

Methods:

Results:

1 INTRODUCTION TO CRISPR:

Clustered Regularly Interspaced Palindromic Repeats combined with a CRISPR-associated operon, or better known as CRISPR-Cas, is a bacterial immune system that carries great promise for adoption as a genome editing tool. The latest example of CRISPR that reached the general public: The Chinese researcher He Jiankui of Southern University of Science and Technology in Shenzhen, claimed to have created the world's first genetically edited germline using CRISPR-Cas, that resulted in born babies [4]. Although most scientists share the opinion that CRISPR-Cas should not be applied to human germlines until it has been extensively tested [4], it is a useful toolkit for biotechnology, experimental setups and clinical medicine. It already has been proved to be useful in multiple biotechnological applications, including the generation of phage resistant dairy cultures [5] and phylogenetic classification of bacterial strains [6]. Furthermore, a modified version of the CRISPR-Cas9 system has been used to recruit heterologous domains that can regulate endogenous gene expression or label specific genomic loci in living cells [7]. In other experimental set-ups CRISPR-Cas9 is for example used to fluorescently tag specific DNA loci, which is a powerful live-cell-imaging alternative to DNA-FISH [8]. Moreover, Cas9 can by injected into fertilized zygotes to generate transgenic animal models [9,10], or it can be used for genomic screens by introducing loss-of-function mutations in human cells [11,12]. An example of a medical application is that it has been used to create antibiotics whose spectrum of activity is chosen by design [13].

1.1 THE BIOLOGY OF CRISPR-CAS

The first paragraph showed that CRISPR-Cas is a widely used gene-editing tool. But how does it exactly work? CRISPR-Cas is originally a defence mechanism in bacteria, where it is part of the immune system against viruses [14]. It recognises the viral DNA of the invader and degrades it.

CRISPR-Cas intervenes with the lytic cycle of the virus. In the lytic cycle the virus injects its DNA in the bacterium, in order to get its DNA replicated. A new virus will form around each replica, using components produced by the bacterium. This continues until the bacterium bursts open and releases the newly formed viruses into the surroundings (See Figure 1).

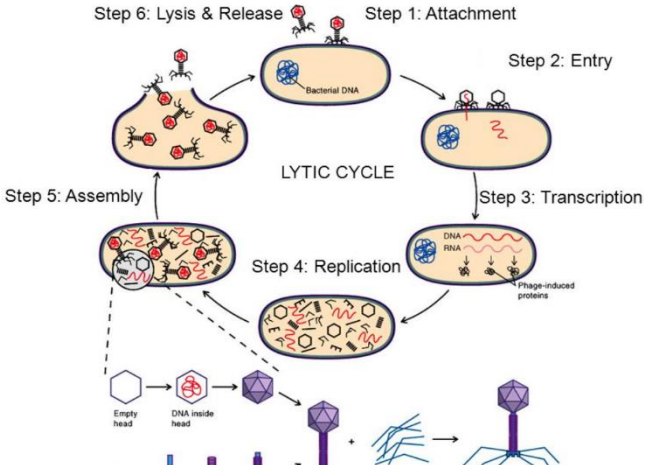


***Figure 1**: The Lytic cycle. CRISPR-Cas, a defence mechanism in bacteria, prevents this by degrading the viral DNA after it is injected.* [41]

CRISPR-Cas prevents above described process from happening by degrading the viral DNA. But how does CRISPR-Cas recognises this viral DNA? CRISPR-Cas is an adaptive immune system that stores memory of past infections [14]. This is done in the CRISPR array, a genomic locus which is composed of alternating identical repeats and unique spacers. These spacers are pieces of DNA that are identical to the viral DNA [15]. In front of this CRISPR array is a series of genes that drive the three phases of immunity [14]:
(See Figure 2 for a graphical representation.)

- Adaptation: Foreign nucleic acids are selected, processed and integrated into the CRISPR array to provide a memory of infection [16]. The viral DNA is recognised by Cas1/Cas2 when it enters the cell for the first time. How is not well understood. However, when it is recognised, a protospacer adjacent motif (PAM) will be identified. During integration of the DNA into the CRISPR array the PAM will not be included. Therefore, this PAM region is later used to distinguish between viral DNA and the CRISPR-array. The interference machinery (Cas protein) is only able to bind to the PAM and therefore, it is prevented from cutting the CRISPR array itself.
- CRISPR RNA (crRNA) biogenesis: The memory that is created during adaptation is retrieved when the CRISPR array is transcribed to produce a long crRNA. This crRNA gets loaded into the Cas protein and will guide the interference machinery (the Cas protein) to bind and cleave the viral DNA.

- Interference: The actual binding and cleaving of the viral DNA by the Cas protein, which is guided by the crRNA. The way the DNA is cut depends on the Cas protein.

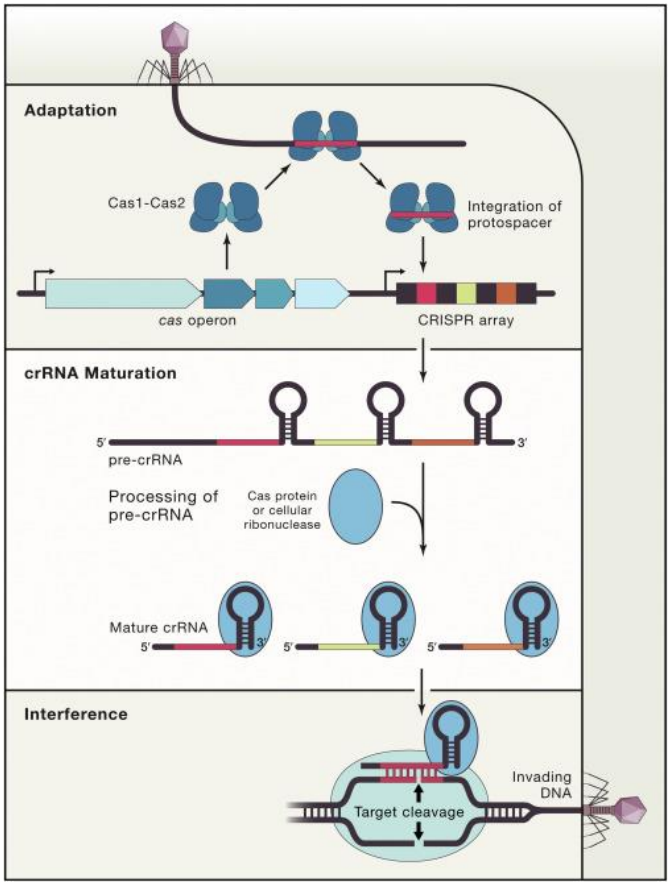


Figure 2: *The three stages of CRISPR immunity. During adaptation the DNA of a new virus gets selected, processed and integrated into the CRISPR array to provide a memory of infection. The memory that is created is retrieved by transcribing the CRISPR array into crRNA (crRNA Maturation), which will guide the Cas-protein to bind and cleave the viral DNA (interference).* [14]

1.2 CAS9 AND CAS12

In this thesis we will work with (d)Cas9 and (d)Cas12. A lot is already known about the crRNA biogenesis and interference performed by Cas9 [17–19]. Before Cas9 is bound to the guide RNA it maintains an autoinhibited conformation. Binding of the guide RNA to the protein induces a major conformational change, enabling recognition of the DNA. First the PAM region in the DNA will be recognised, after which binding between the target DNA and the guide RNA happens sequentially. For Cas9, target specificity is mostly determined by the seed region, lasting rom the first till approximately the 10th nucleotide. This region is very intolerant for mismatches. The PAM-distal regions are more tolerant of mismatches. After binding but before cutting the DNA, another conformational change occurs, enabling the protein to induce a double-stranded break at 3 nucleotides upstream of the PAM [20].

The process is similar for Cas12, also including 2 conformational changes at the same moments [21]. However, there is less known about the exact conformational changes that occur. The seed of Cas12 probably includes the first 5 nucleotides. Another difference is that upon cleaving, Cas 12 will generate a double stranded break with 5 nucleotide overhang at the PAM distant site (after the 18th nucleotide at the non-targeted strand and after the 23th nucleotide on the targeted strand).

1.3 CRISPR-CAS AS A GENE-EDITING TOOL

A series of studies [22–26] led to the realization that induced DNA double-strand breaks (DSBs) can stimulate genome editing through homologous recombination events [27]. This inducing of DSBs is exactly what we can achieve with CRISPR-Cas, as we described in section 1.1.

If we want to use CRISPR-Cas as a gene-editing tool, we can simply change the guide RNA, causing it to bind to a DNA sequence of choice and cut. In this way it is in theory possible to target every possible DNA sequence that is preceded by a PAM. But in reality this works slightly different due to the origin of CRISPR-Cas.

From an evolutionary point of view, it would be inconvenient to specifically target only one sequence. In that case the defence mechanism would become useless if the virus undergoes one slight mutation in its DNA. Therefore, CRISPR-Cas allows for multiple mismatches in the DNA compared to the guide RNA.

For the bacteria this was a convenient characteristic, but for us, when using it as a gene-editing tool, this causes a problem. We want to alter only one specific sequence and not cut the rest of our genome. This forms a big disadvantage of using CRISPR-Cas for gene-editing, since it can cause cancer or could be catastrophic when we use CRISPR-Cas for gene-editing of the human germline [4].

To help realise the therapeutic promise of CIRSPR-Cas, we want to understand this off-target binding better. Studies from Fu et al. [28], Hsu et al. [29], Mali et al. [30], and Pattanayak et al. [31], all show that Cas9 tolerates mismatches throughout the guide RNA sequence. They show that the amount of tolerated mismatches depend on their number, position and distribution. This gets supported by the data from Boyle et al. [2] and Finkelstein et al. [3] that we will use for this thesis. Hsu [29] and Pattanayak [31] also showed that mismatches appear to be better tolerated at high concentrations of Cas9. On top of that, Wu et al. [32] showed that Cas9 indeed has many off-target binding sites, but would only cleave a small fraction of them.

From these experiments people have tried to come up with empirical rules and models to predict off-targets, using different approaches. Several bioinformatics models have been produced to try and evaluate the potential off-target sites. For example, the model from Hsu et al. [29] and a model from Stemmer et al. [33], use empirically determined scoring algorithms to quantify off-target cleavage. Doench et al. [34] proposed the cutting frequency determination (CFD) score to calculate the off-target potential of sgRNA-DNA interactions. However, more physical approaches exist as well. In a recently published paper, Zhang et al. [35] show their model that is based on the free-energy scheme of Cas9. However, no generally accepted model exists so far.

1.4 THE RESEARCH OF THE MARTIN DEPKEN GROUP

The group of Martin Depken at TU Delft is trying to build a model for CRISPR-Cas as well. We are building a kinetic model, based on the amount of free energy, in the hope to be able to predict the probability of binding to an off-target sequence. This approach is in some ways similar to the approach of Zhang et al. [35]. The advantage of having a kinetic model instead of a bioinformatics/machine learning one, is the following:

- You start from a basic concept and try to add more and more parts until you are able to predict the experimental data. This provides you with an understanding of important features of the system, since you will discover which information will add predictive power to your model and what will not.
- A kinetic model is not molecule specific and therefore, also applicable to other molecules operating with the same type of sequential hybridization.

Our group already succeeded in building such a model that describes general features of CRISPR-Cas9 and it can be used to predict the off-target binding [36]. This model is trained on the experimental data of Boyle et al. [2] and so far only looks at the binding dynamics, under the assumption that the binding is sequence dependent and cleavage will only happen if CRISPR-Cas9 is fully bound.

1.5 THE FOCUS OF THIS THESIS

For this thesis we want to apply the same model to a different dataset from a different type of experiment. The experiment is from Finkelstein et al. [3] and this experiment is done for not only Cas9, but for a lot of other Cas proteins as well. Therefore, we can first check if we can apply the model on this type of data, by comparing the resulted energy landscape for Cas9, to the result from the model with the data of Boyle et al. [2]. If we indeed obtain the same energy landscape, we can use the same model for other Cas proteins and get a better understanding of their features. In this thesis we will take a look at Cas12.

The only problem that could appear is that the model is not applicable to this type of data, in other words: that there is important information missing in the experiments from Finkelstein et al. In that case we want to identify which information is missing.

The main questions for this thesis are:
"Do the experiments from Finkelstein et al. provide enough information to train our model? If not, what information is missing? If so, what can we tell about Cas12?"

We will first take a look at the general concept of our model in chapter 2. In chapter 3 we will take a look at the differences between the two experiments from Boyle et al. and Finkelstein et al. and how this influences our fit. Thereafter we will take a look at the results from the fit of Cas9 based on the data from Finkelstein et al. and see which conclusions we can draw about the data (Chapter 5). At last we will move on to Cas12 and see which properties we can determine about this protein (Chapter 6). We will conclude this thesis with a discussion and a recommendation for further developments for the model.

2 EXPLANATION OF THE MODEL

The off-target activity of CRISPR-Cas forms a problem for using it as a tool to modify DNA. To understand the targeting rules of CRISPR, the Martin Depken group at TU Delft kinetically models the guide-target hybrid formation [36]. Their main idea: View the hybrid formation between the guide RNA and target DNA as a random walk through a free-energy landscape. The process that occurs (cleavage or unbinding) gets determined by the surrounding energy barriers. The difference in barrier heights form the kinetic biases. We use the rule of thumb, that after binding to the PAM, unbinding before cleavage is likely if the highest barrier to cleavage is greater than highest barrier to unbinding, and visa versa.

In this section I will first explain the general concept of the model. After that we will take a deeper look into the kinetics, to understand how we modelled this using mathematics. In the end we will investigate the assumptions behind this model and try to justify them as far as possible. In the rest of this thesis, we will use this model to model experiments.

2.1 THE STATES OF THE SYSTEM

Since our main goal is to understand the off-targeting activity of CRISPR-Cas9, we try to build up the model using only already known facts and laws of statistical physics. From experiments [18] we know that Cas9 first needs to bind to the PAM region, then binds to all the other 20 nucleotides in the guide RNA, before it can finally cut. All those parts can be seen as a chain of 23 different states. The first state represents an unbound DNA molecule, the second represents the situation where the PAM of the DNA is bound to the CRISPR-Cas system, the 3th till the 22th as binding to every next base pair and the last state would be the state where CRISPR-Cas9 cuts the DNA (see Figure 3). In this way, a model with 23 states is obtained, containing the possibility to switch back and forward between the states at certain rates. In order to cut you must go through all the states, until the last one, from which you cannot come back anymore.

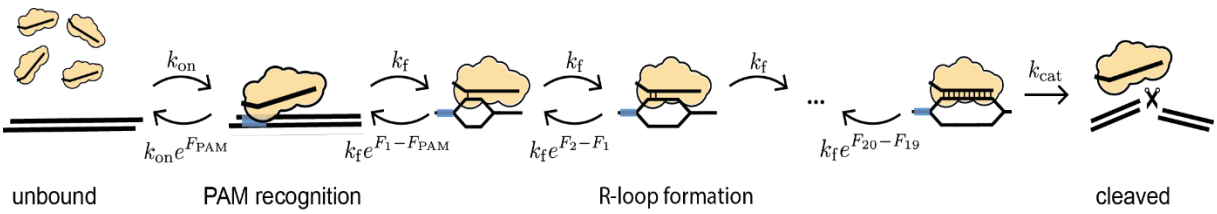


***Figure 3**: A graphical representation of the model, including 23 states with the probability to switch back and forward at certain rates. In this thesis we used dCas9, so k_{cat}=0.*

2.2 THE KINETICS

To understand the mathematics behind modelling these 23 states with their rates, let us look at a simpler, but comparable system. In this system we only distinguish 3 states: in solution, bound and cleaved. In Figure 4 this model is depicted, as well as its free energy landscape. The states are the minima in this landscape and the rates (k) represent the inverse average time to transition between adjacent states. In general terms, it is most likely to go to the adjacent state with the lowest free energy barrier.

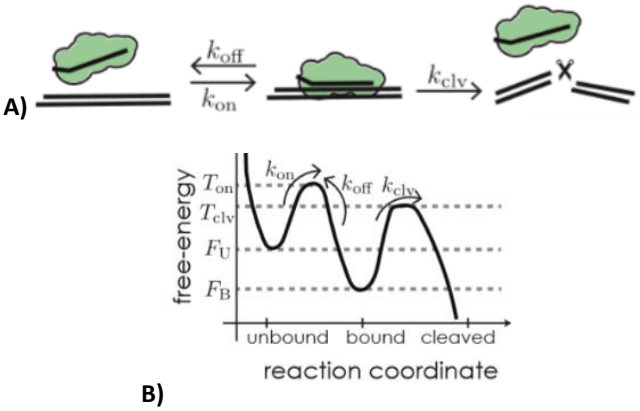


Figure 4: A) *A simple model of CRISPR-Cas with 3 states (in solution, bound and cleaved) and the switching rates in between.* ***B)*** *The free energy landscape for this model where the stable states from the model (as depicted in Figure 4A) are depicted as minima and denoted by F's, while the transition states between two configurations are indicated by T's.* [36]

We use kinetic modelling to describe the dynamics of the CRISPR-Cas system. The assumption is that the time (t) for any single reaction to get completed is exponentially distributed (equation (1)). This assumption is justified, since this is the distribution for a Poisson point process/stochastic process, describing random events. At the molecular scale, our rates originate from thermally activated transitions over the energy barriers and can therefore, be described as random.

$$\phi(t) = k\, e^{-kt} \tag{1}$$

In this equation, $\phi(t)$ is the probability density to complete the switching to a next state at time t and k is the rate to get into that specific state. If we now look at an even simpler system, with only 2 states and one switching rate k from state 1 to state 2. We can derive the probability to be in state 1 (p_1) from $\phi(t)$:

$$p_1(t) = 1 - \int_0^t \phi(t)dt = 1 - (1 - e^{-kt}) = e^{-kt} \quad \big(= 1 - p_2(t)\big) \tag{2}$$

$$\frac{dp_1}{dt} = -kp_1(t) \tag{3}$$

If we now move back to the system with 3 states as depicted in Figure 4, we can use this same principle. For example, the amount of unbound molecules decreases by the amount that will bind to the DNA and increases by the amount that unbinds from the DNA [37]. Doing this for all the states while using the rates indicated in Figure 4, gives the following set of differential equations:

$$\frac{dp_{ub}}{dt} = -\,k_{on}p_{ub}(t) + k_{off}p_{bnd}(t) \tag{4}$$

$$\frac{dp_{bnd}}{dt} = +\,k_{on}p_{ub}(t) - \big(k_{off} + k_{clv}\big)p_{bnd}(t) \tag{5}$$

$$\frac{dp_{clv}}{dt} = +\,k_{clv}p_{bnd}(t) \tag{6}$$

This can be rewritten in a matrix-vector form:

$$\frac{d\vec{P}}{dt} = M\vec{P}(t) \tag{7}$$

$$\text{Where: } M = \begin{pmatrix} -k_{on} & k_{off} & 0 \\ +k_{on} & -(k_{off} + k_{clv}) & 0 \\ 0 & k_{clv} & 0 \end{pmatrix},\ \vec{P}(t) = \begin{pmatrix} p_{ub}(t) \\ p_{bnd}(t) \\ p_{clv}(t) \end{pmatrix}$$

This equation we call the Master equation. Later in this thesis we will refer to matrix M as the rate matrix. Solving the Master equation gives for every state the probability of being in that state. In other words: the fraction of molecules that is in a certain state at timepoint t. This principle of looking at the states and the switching rates between them is exactly what we use in our model with 23 states as well.

2.3 ENFORCING A BINDING EQUILIBRIUM

For this thesis we work with dCas9 and dCas12, which have lost their ability to cut and therefore, we set the rate of going into the last state (k_{cat}) equal to zero. This results in a closed system in which a binding equilibrium will form between the unbound and bound states as $t \to \infty$. For any equilibrated system, the occupancy will start to follow the Boltzmann distribution (equation(8)) and the rates between state i and j must satisfy the relation in equation (9). We will use this 'detailed balance condition' to relate the rates to the free energy.

$$p_{ub,b}^{EQ} = \frac{e^{-F_{ub,b}}}{e^{-F_{ub}} + e^{-F_b}} = \frac{e^{-F_i}}{Z} \tag{8}$$

$$\frac{k_{i\to j}}{k_{j\to i}} = e^{-\Delta F_{ij}} \tag{9}$$

2.4 OTHER ASSUMPTIONS

So far we made the statements that our rates are originated by thermal fluctuations (section 2.2) and that a binding equilibrium will form as $t \to \infty$ (section 2.3). Next to these statements, there are 2 assumptions that we use for our model. Firstly, the rate from solution to PAM (k_{on}) is linearly dependent on the concentration of CRISPR-Cas9, which is a general rule of reaction rates in chemistry. Secondly, all the internal forward rates (k_f) are equal. The backward rate is determined by the relation as described in equation (9). This second assumption is made since we do not include any sequence dependency in our model. Therefore, opening any of the 20 dsDNA pairs should be more or less the same process and will occur by the same rate.

3 DESCRIPTION OF THE DIFFERENT EXPERIMENTS

The kinetic model of CRISPR-Cas from Klein et al. [36] is trained on the experimental data of Boyle et al. [2]. This data was used since the data itself already revealed certain patterns, indicating that features can be subtracted, and you are not only looking at noise. Using their data had 3 main advantages. At first, the measurements included the occupancy in binding equilibrium and the effective on- and off-rates, which can immediately be used to calculate the model parameters of Klein et al. Secondly, their data included all single and double mismatch sequences. Thirdly, not only the mismatch type, but also the exact mismatch base that occurred was recorded. Although this last point is not yet included in the model, it could be used in the future.

For this thesis we want to be able to train the model on the data of Finkelstein et al. [3] in order to be able to switch to other Cas proteins, as we have such data available for us. The fit on Boyle was performed on the data containing the effective on-rate at one concentration. The data from Finkelstein contains the occupancy after 10 minutes for different concentrations and therefore, does not include any information about time. In order to do train the model on this data, we need to understand the measured quantities and their data processing, since we want to exactly copy this during our fit.

3.1 FINKELSTEIN ET AL.:

Finkelstein et al. [3] developed a chip-hybridized association mapping platform (CHAMP) that combines sequencing with mapping the interaction between proteins and ~10^7 unique DNA clusters, equally spaced over the chip (See Figure 5). All the equally spaced DNA clusters will first be regenerated, in order to remove any fluorescent nucleotides that would otherwise confound imaging. After that, half of the clusters will be labelled with a fluorescent protein to be able to determine the position. Finkelstein et al. let the fluorescent labelled proteins flow through. Using high-throughput fluorescence imaging (done with Total Internal Reflection Fluorescence microscopy (TIRF)) they can measure the association between fluorescently labelled protein complexes and each DNA cluster. They analyse the images using the CHAMP software pipeline, which maps each fluorescent cluster to the underlying DNA sequence.

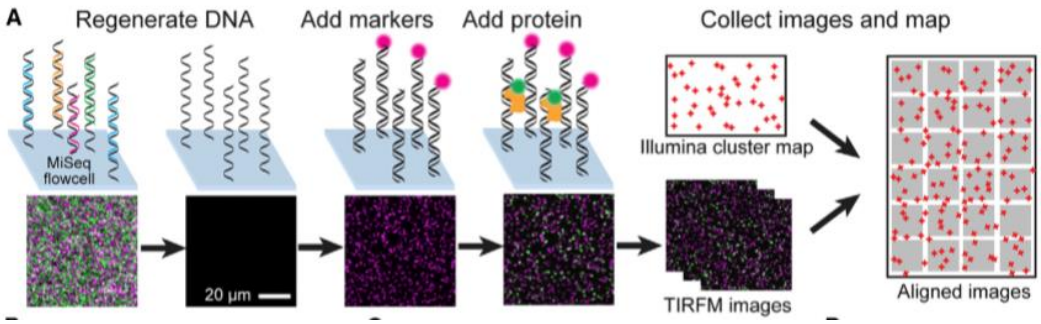


***Figure 5**: Overview of the CHAMP workflow, used to measure the association between fluorescently labelled protein complexes and each DNA cluster on the chip.* [3]

The experiment consists of measuring the fluorescence for a library of DNA strands/clusters (the same as Boyle et al.) with all possible single, double and block mismatch configurations. The measurements are taken at 10 minutes after introducing dCas-sgRNA for varying dCas9 concentrations (0 nM, 0.1 nM, 0.3 nM, 1 nM, 3 nM, 10 nM, 30 nM, 100 nM, 300 nM). The data they produced for dCas9 and dCas12 is not published yet.

3.1.1 PROCESSING OF THE DATA FROM FINKELSTEIN ET AL.

From the measured fluorescence intensity, one can calculate the occupancy.

$$P_{bound} = \frac{I - I_{min}}{I_{max} - I_{min}} \tag{10}$$

Here I_{min} is the background fluorescence, which is included to normalise P_{bound}. I_{max} is the intensity from the fully saturated on-target at 300 nM, thereby setting the on-target occupancy to one. From the P_{bound} one can calculate the k_d by fitting the resulting curve (of P_{bound} VS concentration) to the Hill-equation (Equation (11)).

$$P_{bnd} = \frac{1}{1 + \frac{k_d}{c}} \tag{11}$$

In general terms, the Hill equation describes the fraction of macromolecules saturated by ligand as a function of the ligand concentration during equilibrium. For the justification of the use of the Hill-equation in our multi-state model, see the derivation in section 3.1.2. The k_d reflects the concentration at which $P_{bnd} = 0.5$ (see Figure 6). From the k_d the Apparent Binding Affinity (ABA) can be calculated.

$$ABA = \log(k_d) \tag{12}$$

This is the difference in apparent ΔF (free energy) of a bound molecule to a molecule in solution (at 1 nM). Note the used convention: lower ABA values indicate stronger binding. However, for this definition of ABA to be true, equilibrium should be reached after 10 minutes, since the Hill-equation is strictly only valid after equilibrium has been reached (section 3.1.2). Finkelstein assumed that 10 minutes would be enough, but if it is not, we cannot find the right ABA value (see section 3.3).

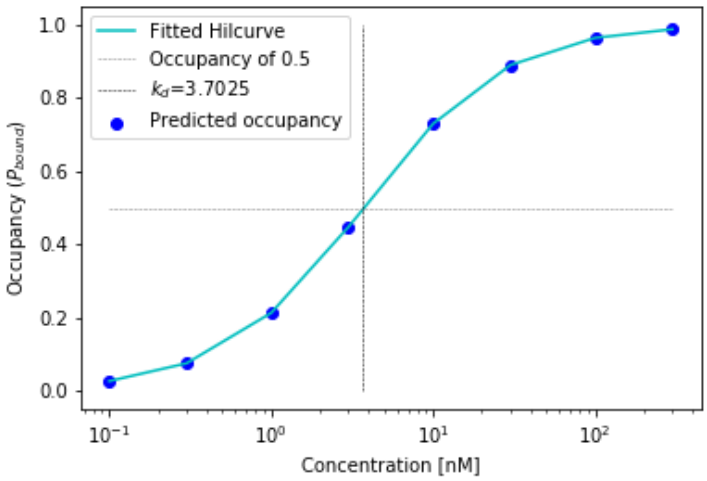


Figure 6: *To obtain the k_d value, first the occupancy is calculated based on the model parameters. The occupancy is plotted against the concentration, through which the Hill-curve can be fit. At the concentration where the occupancy is equal to 0.5, the k_d is found. This k_d can be used to calculate the ABA, the reported quantity by Finkelstein et al.*

3.1.2 DERIVATION OF THE HILL EQUATION

We start this derivation with the 22 different states as pictured in Figure 3 (without cleavage). These states are again: in solution, bound to PAM and the 20 nucleotides (N). The Hill-equation describes the occupancy in equilibrium as a function of enzyme concentration. Since we are in equilibrium, the occupancy can be described by the Boltzmann factors (equation (8)) and the following will hold:

$$\frac{p_n^{eq}}{p_{n-1}^{eq}} = \frac{k_{(n-1)\to n}}{k_{n\to(n-1)}} = e^{-(F_n - F_{n-1})} \tag{13}$$

$$p_{bnd} = \frac{bound\ states}{all\ states} = \frac{\sum_{n=PAM}^{N} e^{-F_n}}{e^{-F_{sol}} + \sum_{n=PAM}^{N} e^{-F_n}} \overset{F_{sol}=0}{=} \frac{\sum_{n=PAM}^{N} e^{-F_n}}{e^{-0} + \sum_{n=PAM}^{N} e^{-F_n}} = \frac{\sum_{n=PAM}^{N} e^{-F_n}}{1 + \sum_{n=PAM}^{N} e^{-F_n}} \quad (14)$$

As we take the binding rate from solution to PAM to be linearly dependent on concentration (section 2.4), such that if the concentration is raised, the rate increases, we can state that $k_{on} \propto \frac{c}{c_0}$. Using this and equation (13), we get the following expression for the energy at state PAM for every concentration compared to the reference concentration (c_0) :

$$F_{PAM} = F_{PAM}^0 - \log\left(\frac{c}{c_0}\right) \quad (15)$$

Since F_n is the cumulative sum of the amount of energy that every state up till state n adds, every consecutive state will get shifted by the same shift.

$$F_n = \sum_{i=0}^{n} (F_i - F_{i-1}) = F_n^0 - \log\left(\frac{c}{c_0}\right) \quad (16)$$

Plugging this back into our equation for p_{bnd} and rewriting gives the following expression:

$$p_{bnd}(c) = \frac{\sum_{n=PAM}^{N} e^{-F_n^0 + \log\left(\frac{c}{c_0}\right)}}{1 + \sum_{n=PAM}^{N} e^{-F_n^0 + \log\left(\frac{c}{c_0}\right)}} = \frac{1}{\frac{1}{c/c_0}\frac{1}{\sum_{n=PAM}^{N} e^{-F_n^0}} + 1} \quad (17)$$

Here we take:

- $\frac{c}{c_0} = c$ since our reference concentration $c_0 = 1\ nM$ in the experiments
- $$k_d = \frac{1}{\sum_{n=PAM}^{N} e^{-F_n^0}} \quad (18)$$

Which leaves us with the Hill-equation:

$$p_{bnd}(c) = \frac{1}{1 + \frac{k_d}{c}} \quad (19)$$

3.2 BOYLE ET AL.:

To be able to understand the differences between the datasets and the difference in the results of the fits that we will discuss in section 5.4 and 7, we will also take a short look at the experiment and data processing of Boyle et al. [2]. They did in principle the same, only using a slightly different chip and DNA library. The main difference is that they measured the fluorescence for every strand at 3 different timepoints (500 s,1000 s, 1500 s) after adding the RNP complexes to the flow cell, at a fixed concentration of 10 nM Cas9. Note that for the fit to our model we only used the measured on-rates, therefore, this is the only part of the experiment that we will discuss.

The florescence signals were converted into the occupancy in the same way as explained for Finkelstein et al. (equation (10)). To obtain the effective on-rates, a straight line was fitted through the occupancy over time and forced to go through origin. Note that their measured effective on-rate is not the same as our rate from solution to PAM. Their on-rate reflects the change in the occupancy. Therefore, they can simplify the binding of Cas9 to a two-state system as discussed in section 2.2. This makes that the start of the binding curve can be described as linear, since according to the Taylor series, $p_2(t) = 1 - e^{-kt} \approx kt$ for small t (equation (2)). They assume that the 3 measured timepoints are still in the linear domain and therefore, reflect the effective on-rate.

In short, the dataset from the experiments from Boyle et al. [2] contained the on-rate of dCas9. But we should think of the following notions about the assumptions they used:

- dCas9 strand invasion behaviour is not expected to obey a simple two-state binding dynamics (as discussed in section 2.1), which underlines the use of linearization to get the association rates. Boyle et al. did realise this and therefore, called the measured quantity an 'effective association rate'.
- The measured timepoints are assumed to be in the linear domain. However, the rates are dependent on the concentration and the ΔF for every target DNA sequence (see section 2.3). To stay in this linear domain, you would need a different concentration for every target DNA sequence. The experiment is performed at one fixed concentration of 10 nM, resulting in the probability to have passed the linear domain, so that you are actually fitting a line through 3 horizontal points in the plateau (Figure 7B).

3.3 Comparing the methods of Finkelstein and Boyle

In section 3.1 and 3.2 the experimental methods of Finkelstein and Boyle are discussed. A better understanding of the exact difference can be obtained by Figure 7A. This depicts the binding curves over time for different concentrations. It includes the measurements from Boyle trough which a straight line will be fitted. The measurements from Finkelstein will be fitted against the Hill-equation (see Figure 6).

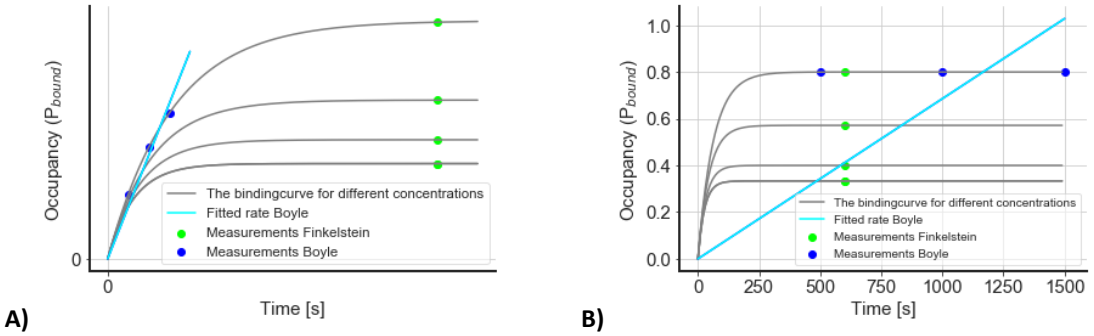


Figure 7: *A comparison of the experimental methods of Finkelstein and Boyle.* ***A) In theory****, Boyle fits a straight line through 3 measured points (at 1 fixed concentration of 10 nM), that still should be in the linear domain of the binding curve. Whereas Finkelstein measures 9 points (only 4 depicted in the figure) at different concentrations, that should be in the plateau in order to be fitted to the Hill-curve as in Figure 6.* ***B) In reality,*** *if 600 seconds is in a plateau, Boyle will fit a line through 3 horizontal points, if the assumption of Finkelstein is correct.*

This shows the importance of being in equilibrium for Finkelstein, since otherwise the measurements are not performed in the plateau, causing them to not lay on the Hill-curve. It also shows the importance of measuring at the right concentration for Boyle, as stated in section 3.2 (depicted in Figure 7B).

In theory both approaches could work. However, they contradict each other. Boyle assumes being in the linear domain until 1500 seconds after adding Cas9, while Finkelstein assumes already being in a plateau at 600 seconds. This can never both be true (Figure 7B).

4 WHAT WE ACTUALLY FIT

Now that we do have an idea what kinetics and assumptions we use for our model and know how the experimental data is obtained, we can bring it together into what we actually fit. We fit a parameter set of 43 parameters:

1st parameter:	ϵ_{PAM}	the free energy needed for binding of CRISPR to the PAM region
2-21:	ϵ_C^i	the free energy needed for the binding of every nucleotide position i (if matches)
22-41:	ϵ_I^i	the penalty in free energy for every mismatch at position i (these are added to ϵ_c)
42:	k_{on}	the rate for binding to the PAM region
43:	k_f	the internal forward rates

It is important to note that we only use 1 mismatch penalty for every position instead of 3. This means that we assume every mismatched nucleotide combination to cost the same amount of energy. The reason for this, is that this model still shows the patterns in the data, indicating that there is no need to include more parameters for the different mismatches. In this way we have a simpler, more understandable model.

The parameters we fit for the experiment of Boyle et al. and Finkelstein et al. are the same, but the method is different. In both cases we try to copy their experiment the best possible. Therefore,, we follow the exact same procedure as described in section 3.1.1, except for the beginning where we replace equation (10) with solving for the occupancy after 10 minutes using the master equation (equation (7)).

To solve the master equation, we need the rate matrix (M), which can be obtained from the parameters. For each configuration the free energy landscape is built by using the ϵ values and the information where the matches or mismatches are present. This free energy landscape describes the ΔF for every position at that target site, so for every state in our model. By using the relation in equation (9), we can calculate the backward rates from these ΔF and forward rate given in the parameter set. With both known we can build the rate matrix.

The obtained ABA value can then be compared with the experimental data by taking the χ^2 value (Equation (20)), in order to judge how well the fit worked and if the parameters describe the reality. This is done for every DNA strand separately. All these χ^2 values for every strand are added and this gives an overall measure for how well these parameters describe the data.

$$\chi^2 = \sum_{data} \left(\frac{ABA_{exp} - ABA_{mod}}{\sigma_{exp}} \right)^2 \tag{20}$$

4.1 SIMULATED ANNEALING

The process described in the section before, is done repeatedly in a loop, to find the parameters that describe the data the best. The specific optimization scheme we use is called Simulated Annealing and was described by Kirkpatrick et al. in 1983 [38].

This algorithm is inspired by principles of statistical mechanics. Statistical mechanics applies probability theory to a system, consisting of many particles, to study its thermodynamic behaviour. It uses the principle that a system can be in any number of states, which all have their own specific energy level. In thermal equilibrium, the probability to be in any of these states follows the Boltzmann distribution. By gradually lowering the temperature the system finds its minimum. Linking this back to our optimization problem, this would be the fit with the lowest error (in our case the χ^2 (equation (20))). One should keep in mind, that if you cool down the system too quickly, it can get stuck in a local minimum, instead of finding its global minimum (See Figure 8).

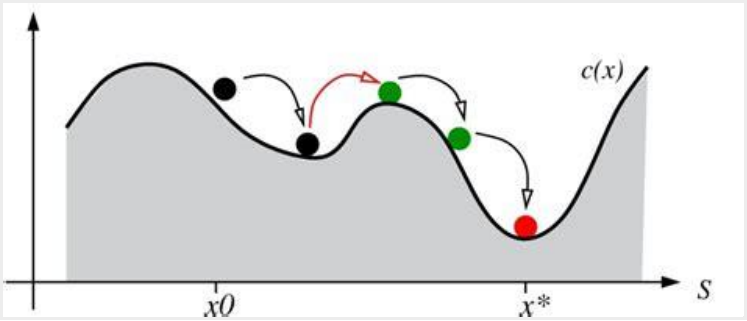


Figure 8: *from the local minima (black) you will not see the global minima (red), unless you manage to overcome the barrier (green). This is only possible if you do not cool down to quickly and apply equation (21).* [39]

The concept of temperature of a physical system has no obvious equivalent in the systems that are being optimised. In fact, this temperature term just gets introduced in the algorithm to mimic this effect of slowly cooling down the system. This is done by having a variable T that decreases every iteration, which is used in the decision function for accepting the new parameter set or not.

$$P_{accept} = \begin{cases} 1 & c_{new} \leq c_{old} \\ e^{-\frac{c_{new}-c_{old}}{T}} & c_{new} > c_{old} \end{cases} \quad (21)$$

C is the cost value of the new or old solution. In our case this is the χ^2 value of the new and old parameter set. P_{accept} is the chance for the solution to be accepted. This means that not only better solutions can get accepted, also worse solutions have a chance. The reason for this is that otherwise you would just directly walk towards your local minima and there is no chance to overcome a small barrier to a lower global minimum.

The simulated annealing protocol consists of the following steps:

1) Define some cost function to be minimised
   For us this is the χ^2 function
2) Initialize the algorithm by giving it a starting guess for the parameters and an initial temperature
   *Parameters: initial guess = [5.0] + [3.0]*40 + [1.5]*2*
   Temperature: determined by the code itself. It starts with an initial guess from the user. Then it computes 1000 iterations and calculates the acceptance ratio for the parameters. We want this to be around 50%, since then you are efficiently scanning the full landscape. If this is the case, this T is accepted as the initial temperature. If the acceptance ratio is lower, the energy barriers in the landscape are too high to overcome, so the initial temperature is raised. If the acceptance ratio is higher, the initial temperature is lowered.
3) Consider some neighbouring state by slightly altering the parameters
   The step size of the taken step is a random number between -1 and 1, multiplied by the maximal step size. This is done for every parameter individually. With these parameters the cost function is calculated for every DNA strand in parallel, using multiple processing. All these cost values are then added together.
4) Accept or reject the new parameters based on equation (21)
5) Repeat step 3 and 4 several times
   In our case 1000x
6) If the stop condition is not yet reached → lower the temperature and return to step 3
   Exponential cooling with cooling rate = 0.99
   If the stop condition is reached → return the final parameter set
   Temperature too low (0 degrees) OR the change in average χ^2 between temperature cycles is lower than tolerance (1e-5) + temperature low enough (1% of initial temperature)

Although these steps are not that complicated, it is still quite hard to make the algorithm work efficiently and prevent it from ending up at a local minimum. In order to do so, the right settings must be found for the initial temperature, the step size for changing the parameters, the cooling rate and stop condition.

4.2 SELECTION OF THE FITS

Given the complexity of the data (and model), we typically notice that not just a unique set of parameters results in a good fit to the data. In order to characterise what 'wiggle room' we have for each parameter we follow the following procedure. First, we repeat the simulated annealing algorithm several times, obtaining fits describing local minima and the global minimum. In order to select the fits that ended up in the global minimum, we want to select the ones with the smallest χ^2. However, instead of just taking the smallest χ^2, we realised that we can come up with a 'golden standard' for any sequence independent model, in saying that the derivative of the χ^2 function should equal zero. This results in a function which allows us to visualise the data, which makes it possible to visually compare the fit to the data.

$$\chi^2 = \sum_{data}\left(\frac{y_{exp} - y_{mod}}{\sigma_{exp}}\right)^2 = \sum_{i=\,mm\ configuration}\left(\sum_{j=DNA\ strand}\left(\frac{y_{exp}^{i,j} - y_{mod}^{i}}{\sigma_{exp}^{i,j}}\right)^2\right)$$

$$\frac{\partial \chi^2}{\partial y_{mod}} = \sum_{j=DNA\ strand} 2\left(\frac{y_{exp}^{j} - y_{mod}}{\sigma_{exp,j}^2}\right) * -1 = 0$$

$$\sum_{j=DNA\ strand}\frac{y_{exp}^{j}}{\sigma_j^2} = \sum_{n=DNA\ strand}\frac{y_{mod}}{\sigma_n^2} = y_{mod} * \sum_{n=DNA\ strand}\frac{1}{\sigma_n^2}$$

$$y_{mod} = \sum_{j=DNA\ strand} w_j y_{exp}^{j} \qquad with\ w_j = \frac{\frac{1}{\sigma_j^2}}{\sum_n \frac{1}{\sigma_n^2}} \tag{22}$$

In this equation y is the ABA value and σ is the measurement error from the experiment. y_{mod} would be the model that describes the dataset the best. This formula is also known as the weighted average, which now becomes the 'golden standard', so the ultimate model. We select our fits based on their similarity to this 'golden standard' by giving it a score using equation (23).

$$score = mean\left(\frac{abs(WA_{data} - WA_{model})}{WA_{data}}\right) \tag{23}$$

Here the WA is a list of the weighted average ABA for every target DNA sequence. Although there is no guarantee that our model can achieve a score of zero, selecting fits that result in the lowest x% of scores is a way for us to see what different set of microscopic parameters result in (more or less) equally good fits.

Although this seems a reasonable 'golden standard', the fit with the lowest score is not per definition the score with the lowest χ^2. This is something we must investigate further (outside the extend of this thesis). However, this forms a problem for the selection of fits for Cas12, hence we will select them based on the lowest χ^2.

5 DISCOVERED PROPERTIES OF THE DATA BY FITTING

In this thesis we try to train the model to the data of Finkelstein et al. In order to judge if the fit worked, we will compare the obtained on-target energy landscape (described by the parameters) to the landscape we obtained from the data of Boyle et al. The reason why we compare with this fit is that we have quite some confident in this fit. We can use it to predict the dataset from Finkelstein. Moreover, it also shows characteristics of Cas9 like the seed region and indications of conformational changes. If the model is able to fit on the data of Finkelstein, we should be able to observe the same characteristics of Cas9.

The on-target energy landscape obtained by training the model on the data of Boyle et al (Figure 9), contains 2 effective energy barriers, dividing the landscape in 2 parts. The first part contains the first big barrier in the beginning, reaching from nucleotide 1 till 11. This area therefore, immediately represents the seed region of Cas9, since obtaining a mismatch after the first barrier will have less effect than a mismatch on top of the first barrier (this is in agreement with section 1.2). This is followed by the second part, a smaller barrier before reaching the end of the target DNA sequence and providing the ability to cut (if it would be Cas9 instead of dCas9). This barrier could maybe relate to the conformational change upon cleavage. Moreover, the ϵ_{PAM} is set to 1.31 and therefore, the height of the energy landscape goes from approximately -10 till 10. This height is linked to the rates, higher rates mean higher barriers.

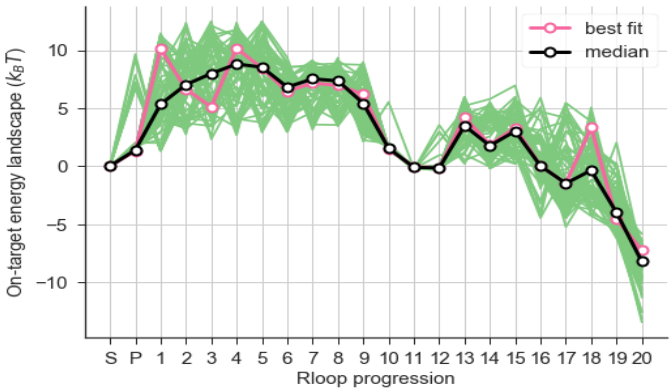


Figure 9: *On-target energy landscape obtained by fitting to the data of Boyle et al. It presents the energy level of the system for binding to every nucleotide in the target sequence. This landscape is used as reference for the fits to the data of Finkelstein et al.*

5.1 TIME EXPENSIVE FITTING TO THE DATA OF FINKELSTEIN

The data used for the first set of fits, includes the by Finkelstein measured ΔABA (= ABA − on-target ABA) for all double and single mismatches at (almost) every possible mismatch positions (we leave out the block mismatches). We included the data for 8 concentration points, with the uncertainty in the measurement. The 9th measured value (as mentioned in section 3.1) was at 0 nM. This is included by subtracting this measured intensity from all the other measurements, since it should only reflect noise.

The best fit seemed to get the first part of the overall energy landscape as we saw for Boyle (Figure 9), but the different fits did not have a lot of agreement. However, a more urgent problem was that the fit took between 40 and 60 hours to finish, while the fit on the data of Boyle et al. only needs 14 hours. This difference in time is caused by the amount of points we fit, for Boyle et al. [2] this were only 3 timepoints, while for Finkelstein we have 8 concentration points (see section 3.1 and 3.2). Those 60 hours is too long if you want to be able to adjust properties and test it again. Therefore, we decided to solve this problem first before we tried to improve the fit efficiency and quality. We came up with 3 ideas to make the code faster, which would give us more insight in the data at the same time:

- *Usage of multiple processing with mpi4py in python:* Both fits were performed using 20 cores on 1 node of the HPC05 cluster of TU Delft. To be able to run the code on more than 20 cores spread out over different nodes. Due to technical difficulties, probably with the settings of the cluster, this was impossible to achieve, despite extensive efforts.

- *Use less concentration points for the fit:* Since we fit the Hill-curve through these points, it maybe is possible to use less points and still be able to fit the correct k_d.
- *Assume equilibrium*: Right now, the most time expensive step for this fit is the calculation of the rate-matrix. If the system is equilibrated, the ABA values can be determined analytically (see section 5.3), allowing us to bypass the rate-matrix calculation.

5.2 USING LESS CONCENTRATION POINTS

Using less concentration points requires us to build the Master Equation (7) less frequently, making the evaluation of ABA, thereby the entire simulated annealing algorithm, faster. Now the question remains which and how many points you need to still make a good prediction for the ABA. In order to answer this question, we looked back at the unprocessed data from Finkelstein et al. from which we can calculate the occupancy (P_{bound}) for every concentration. These values for the occupancy were then used in the Hill-equation to find the K_d and calculate the ABA as was explained in section 3.1.1. For our previous fit this was done with all the 8 concentration points. Now we will test, if we get the same while using less concentration points, by evaluating the error and the correlation coefficient between the ABA value of all concentration points and the ABA with less points. Note that since we now got the unprocessed data from Finkelstein, we have the absolute ABA value, which contains more information, like the on-target ABA. Therefore, we will from now on train the model to predict this absolute value instead of the ΔABA.

The best combination of 2 concentration points [10 nM,100 nM], still gives an error in predicted ABA value of 10% for all the target DNA sequences that were tested (7395 in total). The best combination of 3 concentration points gives a lower error of only 7.7% for the concentrations [1 nM, 30 nM, 100 nM].

Trying a fit with only the 2 concentration points [10 nM,100 nM] results in an average fitting-time of 36.58 hours. Trying the one with 3 concentration points [1 nM,30 nM,100 nM] results in a n average fitting-time of 23.93 hours. We think that the reason for the longer time with only 2 concentration points is that the curve_fit() function in python has more difficulties with fitting the Hill-curve through these points, making this now the most time expensive step and slowing down the fit.

On top of that, we could also see that the fits with less concentration points do predict the ABA value less accurate, but do still get the overall pattern of the data (See Figure 10). Note that the values do differ, since the fit with 8 concentration points that we use as the reference (Figure 10A) was still fitted on the ΔABA, while we now fit on the absolute ABA. However, obtaining the same pattern indicates that we can indeed use 3 concentration points for our fit.

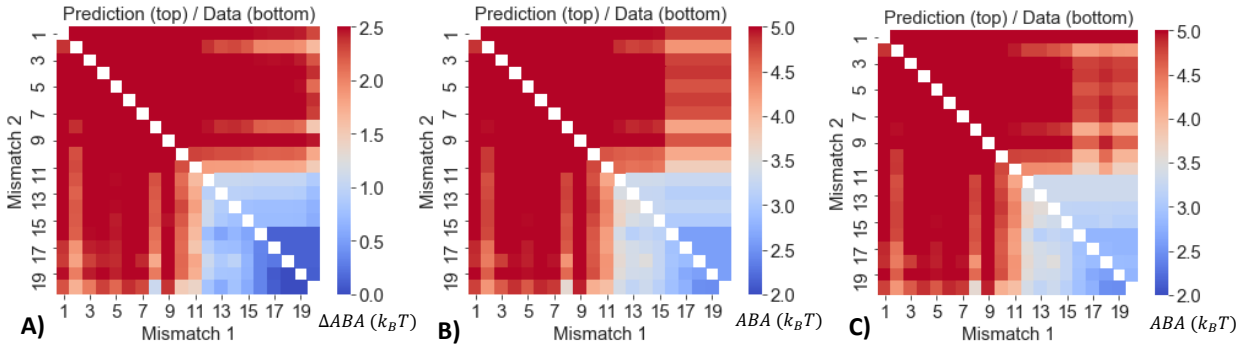


Figure 10: *The double mismatch heatmap showing the ΔABA or ABA for all possible combinations of double mismatches in the target DNA compared to the guide RNA. The predictions are from a fit based on:* ***A)*** *8 concentration points (ΔABA),* ***B)*** *2 concentration points (ABA),* ***C)*** *3 concentration points (ABA). It can be observed that using less concentration points results in less accurate predictions of the ABA, but does capture the overall pattern.*

5.3 ASSUMING EQUILIBRIUM

In the derivation of the Hill-equation in section 3.1.2 we saw that k_d is equal to the sum of the Boltzmann factors (see equation (18)). Therefore,, if equilibrium is actually reached, we can substitute this in the equation for ABA (12) to obtain equation (24).

$$ABA = \log(k_d) = \log\left(\frac{1}{\sum_{n=PAM}^{N} e^{-F_n^0}}\right) = -\log\left(\sum_{n=PAM}^{N} e^{-F_n^0}\right) = -\log\left(\sum_{i=bound\ state} e^{-F_i}\right) \tag{24}$$

With this expression we can immediately calculate the ABA from the energies that we can obtain directly from the parameters. Therefore, there is no need any more to do the calculation with the rate matrix (equation (7)) and the fitting to the Hill-curve (equation (11)), which will make the fit a lot faster. However, by excluding the rate matrix calculation, this method does not fit the rates anymore.

Running the fit did indeed result in an average fitting time of only 16.23 hours instead of 23.93 hours (we used 3 concentration points). Moreover, the fit results look quite reasonable, both the results for the single as well as double mismatches (See Figure 11). For both the seed region can clearly be distinguished in the double mismatch graphs (Figure 11 C&D), since the ABA drops significantly after nucleotide 10. Also, both fit the mismatch penalties ϵ_i approximately the same: values between 2 and 4 until position 10. However, there is much more difference in the fitted ϵ_c values. Due to these differences we get a different height of the on-target energy landscape as can be seen in Figure 12.

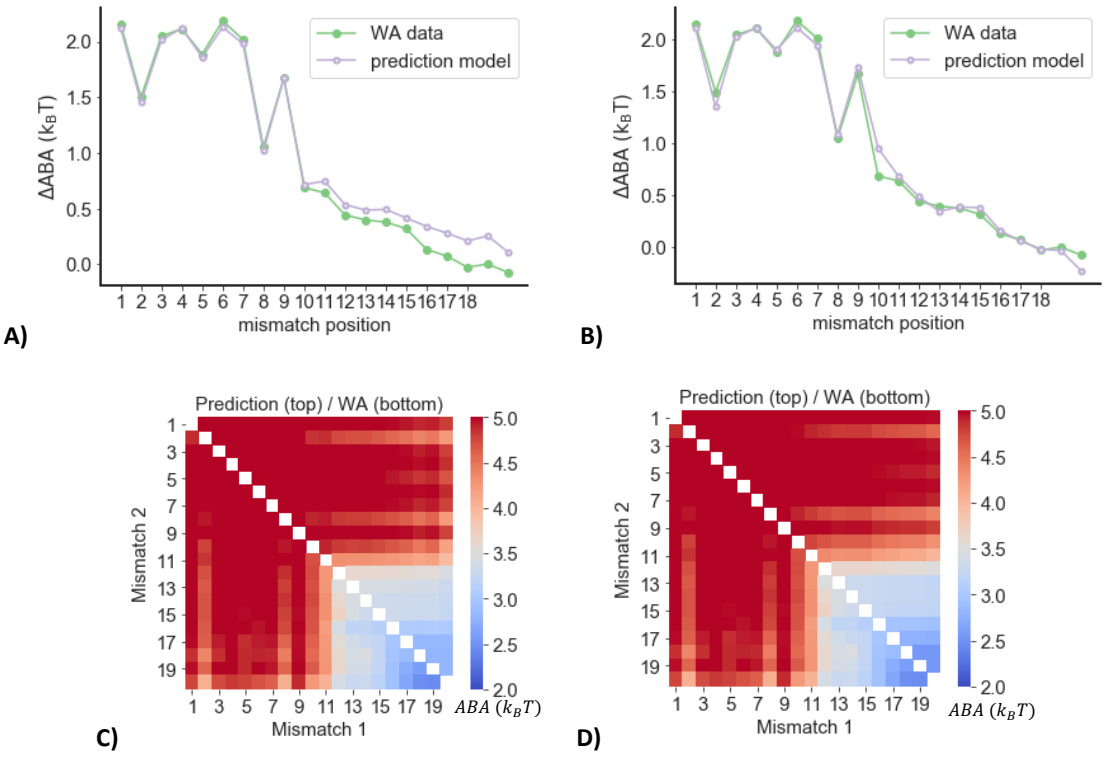


***Figure 11**: The Single mismatch graphs **(A&B)** and double mismatch heatmaps **(C&D)** for the fit using the master equation with 3 concentration points **(A&C)** and the fit where we assumed equilibrium with 3 concentration points **(B&D)**. The graphs display the ABA for all possible combinations of single or double mismatches in the target DNA compared to the guide RNA. It can be observed that both calculation methods result in accurate predictions of the data.*

The on-target energy landscape shows us that the fit assuming equilibrium keeps the overall general shape of the landscape: one big barrier, ending at nucleotide 10/11 where the energy level goes underneath the energy of the PAM, after which a second but smaller barrier follows, ending at the end of the target sequence. We also

see that the fits do still agree on base pair position 11, the end of the first barrier, indicating the end of the seed region at a similar position as the fit to Boyle (see beginning chapter 5).

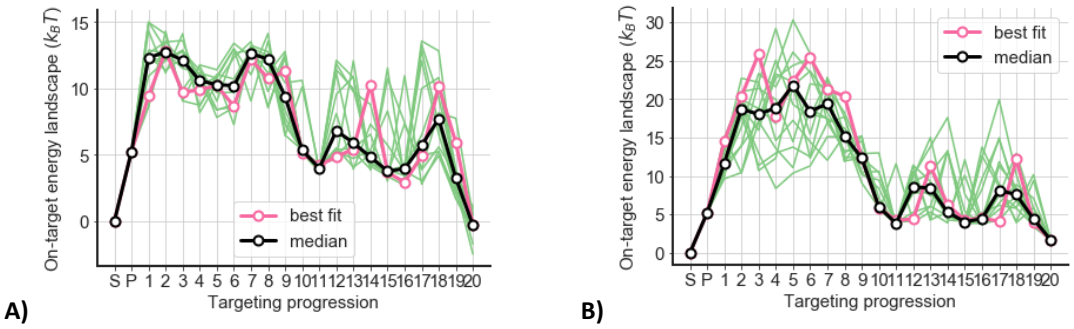


***Figure 12**: The on-target energy landscape for the fits selected on similarity with the weighted average of the data. A) The landscape for fits using the Master equation, selection threshold = 0.018. B) The landscape for fits using equation (24), selection threshold = 0.022.*

However, the fit assuming equilibrium seems to just raise the effective barrier compared to the solution state, resulting in a much bigger need of energy to overcome the first boundary and be able to bind. The existence of this possibility for the fit can be explained by the fact that the rates are not fitted this time, therefore, it can assume every possible rate. Since the rates are related to the barrier height, realising such a landscape (Figure 12B) by the kinetic model, would ask for much higher rates. Obtaining such a different height of the landscape hints that there is no equilibrium reached yet during the experiments.

To check if there is an equilibrium after 10 min, we calculated the ABA in two different ways, using the parameters that we got out of the best fit using the master equation:

1) Use the parameters to calculate the rate matrix from which we calculate the occupancy. The occupancy will be fitted to the Hill-curve to find the k_d value, from which we calculate the ABA.
2) Use the parameters to find the energy landscape, that we use to calculate ABA with equation (24)

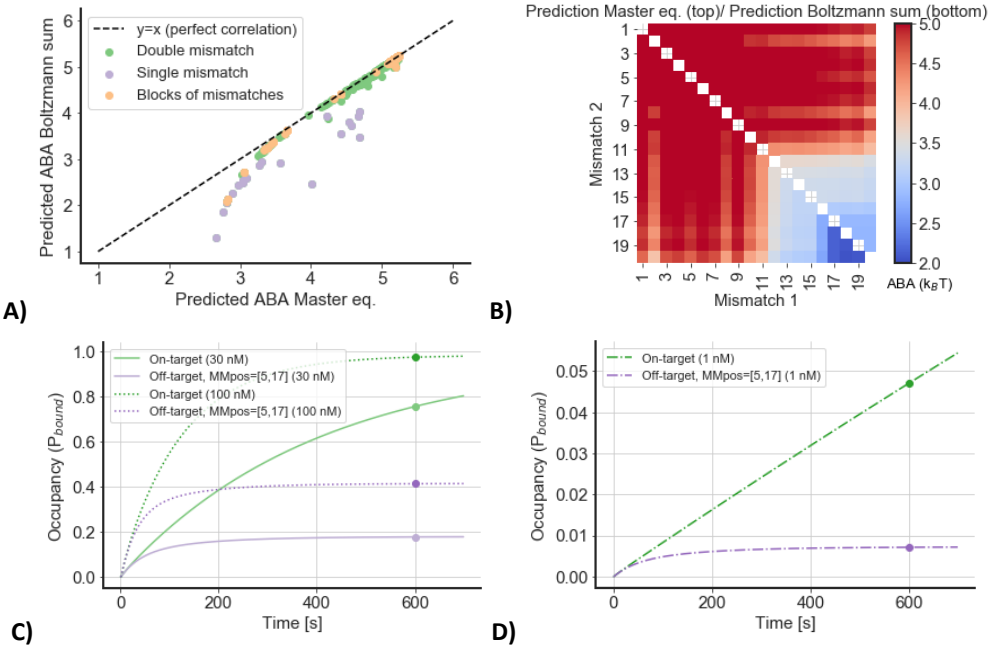


***Figure 13: A)** The correlation between the predicted values of the ABA based on the master equation or the sum of Boltzmann factors (assumed equilibrium). **B)** The predicted ABA for every combination of double mismatches using both calculations. Note that the pattern is captured for the off-target strands (first mismatch before 11), but not for the close to on-targets (first mismatch after 11). **C&D)** The binding curves for the on-target and some off-target. The measurements of Finkelstein are indicated by a dot. It shows that equilibrium is reached for the off-target at all concentrations that were used during the fit. However, for the on-target equilibrium is not even completely reached at the highest concentration (100 nM).*

This was done for all possible single, double and block mismatches. By comparing those predicted ABA values, we see that double and blocks of mismatches have generally a higher correlation than the single mismatches (see Figure 13A). In Figure 13B we can also see that the biggest difference in the prediction occurs when the mismatches are PAM distal. In Figure 13C the binding curves for the on-target and an off-target are depicted. This shows that indeed the off-target is saturated after 10 minutes, while the on-target is not, even not for the highest used concentration. Therefore, we can say that the off-target strands do equilibrate in 10 min, but the on-target strands (or close to on-target) do not. This can be rationalised by saying that for the off-target strands, Cas9 will not travel through the whole landscape, it will get stopped by an energy barrier and goes back into solution. Therefore, it travels through less states and needs less time to equilibrate. While the on-target must travel through the whole energy landscape and therefore, needs more time to reach equilibrium.

This gets also confirmed by the fact that the on-target ABA predictions are completely different:

- Prediction experiment ≈ 2.554
- Prediction equilibrium ≈ -0.327

We can conclude that the assumption of equilibrium is not met for the (close to) on-target strands. Although fitting with the assumption of equilibrium gives good predictions, we see that the height of the energy landscape cannot be determined, only its shape is preserved. From these high barriers we deduce that the kinetic model gives the same results after increasing the rates accordingly. This indicates that equilibrium was not reached after 10 minutes.

5.4 Comparison to Boyle

Comparing our result from the best fit on the data of Finkelstein et al. and the best fit on the data from Boyle et al, we can observe that they give the same shape for the on-target energy landscape (Figure 14). For both we can observe 2 bumps and agreement of all the fits on points 10/11. Such an agreement indicates that it probably is a property of the protein itself, and not some characteristic of the fit.

Two big differences between the landscapes can also be distinguished. First of all, for the fit on the data of Finkelstein we can observe that all the fits differ a lot from each other for the second part of the landscape (after nucleotides 9/10).This means that the second part of the landscape cannot really be determined from the training data. This is probably caused by the fact that equilibrium is not reached yet after 10 minutes for the on-target (as shown in Figure 13C), resulting in that Cas9 simply did not have the time to explore this part of the landscape (it did not extend the guide-target hybrid that far yet). Secondly, it can be observed that the fit on the data of Finkelstein clearly suggests 2 effective barriers in the first part of the landscape (till nucleotide 10/11), while Boyle does not. It could be that there is more information in the data of Finkelstein, that we cannot extract with the data of Boyle, so that this is a characteristic of Cas9. Alternatively, this characteristic can be a result of overfitting on the data of Finkelstein. If this overfitting has occurred, I will discuss further in section 5.4.1 and in the discussion, section 7.

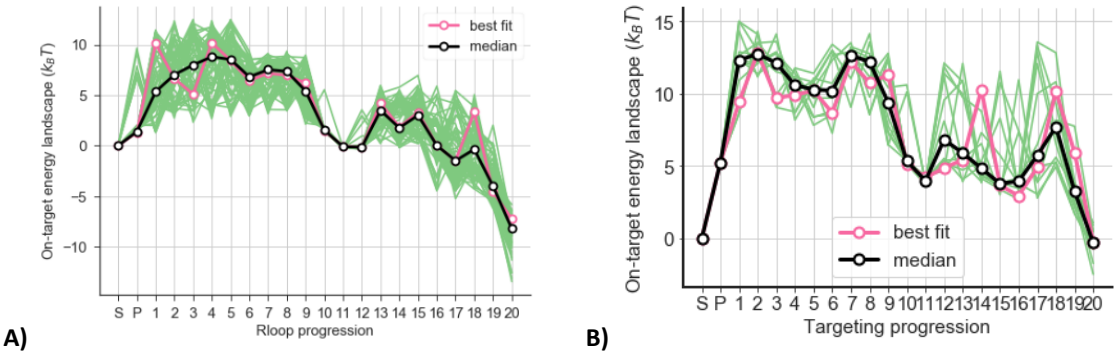


Figure 14: *The on-target energy landscape* ***A)*** *Based on the parameters fitted on the data of Boyle et al. Note that this graph is for a concentration of* $10\ nM$*. For a concentration of* $1\ nM$ *the* ϵ_{PAM} *becomes 3.60 which will shift the whole landscape up like in Figure 24.* ***B)*** *Based on the parameters fitted on the data of Finkelstein et al. at* $1\ nM$

In general we can say that the on-target energy landscape (Figure 14B) shows the same characteristic as the one from the fit on Boyle (Figure 14A). However, we observed that the other parameter values (ϵ_i and the rates) do differ. Figure 15 depicts the mismatch penalties (ϵ_i) for the fit to Boyle (A) and the fit to Finkelstein (B), showing that they are much more defined for Finkelstein than for Boyle, especially for the first part of the landscape. Looking back at Figure 14 shows that the fits also seem to show more agreement in the on-target energy landscape for this region, suggesting that there is more information to gain from the data from Finkelstein than from Boyle. This will be discussed further in the section 7.

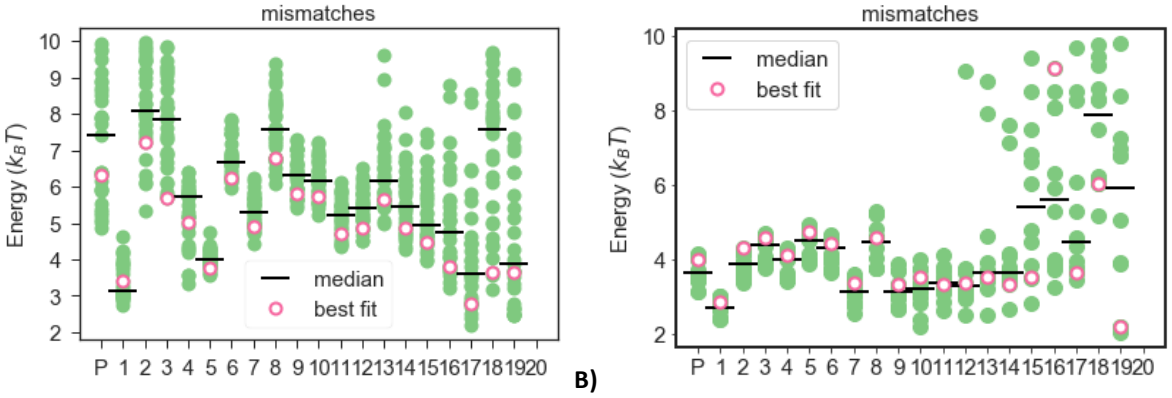


Figure 15*: The mismatch penalties (*ϵ_i*) obtained in the fit to A) Boyle B) Finkelstein. Whereas the fit to Boyle shows a pattern, the fits to Finkelstein keep all the* ϵ_i *almost constant around 4* k_BT*.*

Looking at the rates, shows for both dataset a lot of disagreement between the fits for k_{on} (solution to PAM). They do seem to better define the internal rate (k_f), which are $326\ s^{-1}$ and $210\ s^{-1}$ for Finkelstein and Boyle respectively, which are both in the same order of magnitude. The reason for the obtained difference is however unclear.

Another difference in the parameters is the ϵ_{PAM}. For Finkelstein we got that $\epsilon_{PAM} \approx 5.24$, while for Boyle we get a value of $\epsilon_{PAM} \approx 3.60$ (for both this is consistent over all the fits). A possible explanation for this difference could be a difference in the concentration of Cas9 that actually reaches the target DNA on the chip. If this concentration is less than what they think, they will get lower rates and therefore, lower occupancies (if there is no equilibrium yet). This is probably indeed the case for Finkelstein as will be confirmed by Figure 18B. Therefore, Finkelstein measures a lower occupancy than he should for that concentration, due to which we will fit a higher ϵ_{PAM}. Hence, we can conclude that there is a shift in between the datasets, because it is hard to know the exact concentration that will actually reach the target DNA.

5.4.1 PREDICTING OTHER DATASETS WITH THE OBTAINED PARAMETERS

Before starting the project of this thesis, it was already checked if it was possible to predict the dataset of Finkelstein with the parameters obtained from a fit to the data of Boyle. Apart from single and double mismatched off-targets, the data used to fit against, Finkelstein's library also contained off-targets with more than 2 mismatched places consecutively ('block mismatches'). With the parameters of Boyle we could accurately predict the ΔABA for the block mismatches and double mismatches, but the prediction was less for the single mismatches (Figure 16). Although predicting the ΔABA was successful, it was not possible to predict the absolute ABA value. This is probably caused by the shift in between the datasets as explained at the end of section 5.4.

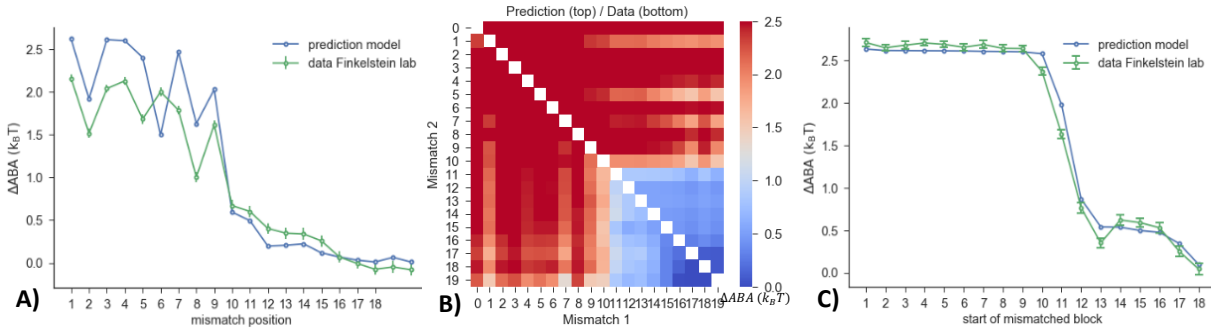


***Figure 16**: The prediction of the data of Finkelstein et al. using the parameters obtained from fitting on the data of Boyle et al. A) Single mismatches B) Double mismatches C) Block mismatches*

To judge the quality of the fit on the data of Finkelstein, we now want to check if we can use those parameters to predict the other 2 datasets. Firstly, we showed that we can accurately predict the ΔABA values for off-targets with block mismatches (another dataset of Finkelstein) (Figure 17). (For more explanation on how to interpret this graph, see section 6.3.) We can observe that the first part (till nucleotide 10) is predicted correctly and the second part shows more deviation from the data. In Figure 13 it was showed that there was no equilibrium for the (close to) on-target strands and in Figure 14B it was showed that therefore, the second part of the on-target energy landscape cannot be defined. This same argument can be used in this case: due to lack of equilibrium after 10 minutes, the ΔABA cannot be correctly determined for the (close to) on-target sequences, leaving the second part of the graph (after nucleotide 10) undefined.

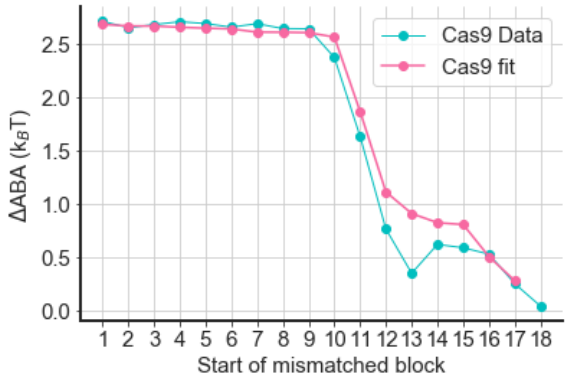


***Figure 17**: Prediction of the ΔABA for block mismatches, based on the parameters obtained from fitting on single and double mismatch data of Finkelstein et al. It can be observed that the fit represents the data quite well, especially up till nucleotide 10.*

However, predicting the dataset of Boyle was less successful. Qualitatively we can approximately predict the pattern, as can be seen in Figure 18A&B, where the rates compared to their on-target rates are depicted. This also gets confirmed by the fact that there is 0.84 correlation between the ranking of the rates from the dataset and the model (Figure 18C). This indicates that the model should be able to tell how likely it is to bind compared to the probability of binding to an on-target. However, the predicted absolute values are all way off from the data. The on-target association rate for the dataset of Boyle is approximately $2.89 * 10^{-4}\ s^{-1}$ while the

predicted rate is $7.49 * 10^{-5}\ s^{-1}$. Therefore, we will not be able to predict with this model, if Cas9 will be bound to a target at a certain concentration after a certain amount of time. This difference in absolute values can again be caused by the shift between the datasets as explained at the end of section 5.4.

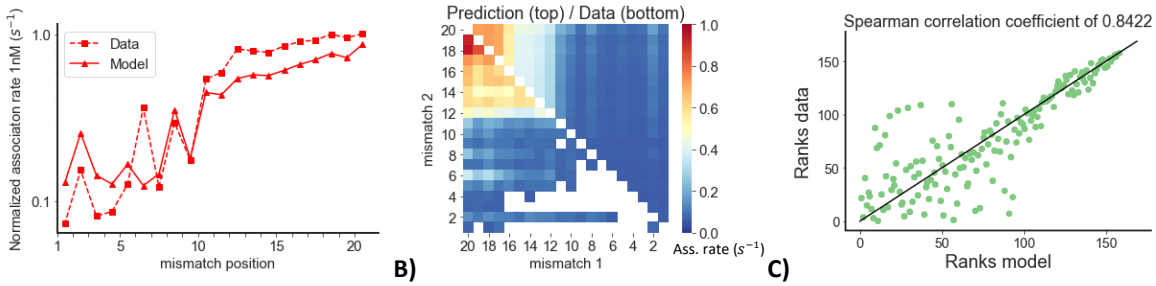


***Figure 18**: The predicted normalised association rates compared to the dataset of Boyle. The prediction is based on the parameters of the fit to the data of Finkelstein et al. The normalisation is done by dividing by the on-target rate of the model/dataset. **A)** Single mismatches **B)** Double mismatches **C)** The correlation between the ranks of the double mismatches, for the data of Boyle compared to the prediction from the fit based on the data of Finkelstein.*

Besides this shift in the datasets, we do also see that we do not get the full pattern and the correlation is only 84%. For the double mismatches, this pattern is mostly located at the end of the strands (Figure 18B), which belongs to the part of the landscape that we cannot predict well with Finkelstein's data, due to lack of equilibrium for these configurations (see section 5.3).

Since there is no equilibrium, there can also be another explanation for the incorrect patterns. In equilibrium, all the occupancies should be exactly on the Hill-curve (except for some noise included in the measurement). However, without equilibrium, these measured occupancies will not be located on the Hill-curve, but somewhere around it. During the processing of the data from the occupancy to the ΔABA we calculate the k_D using the curve_fit() function in python. This function uses the least squares method to find the best value for k_D. But this means that there are multiple possible combinations of occupancies that will give the same k_D. The occupancies can have an average deviation from the hill-curve of 24% and still be fitted to this Hill-curve. Therefore, it becomes possible that our parameters do indeed correctly describe the ΔABA value of the data, but at the same time describe completely different occupancies than the data does. Hence, although it correctly predicts the ΔABA, it does not describe the real measurements and therefore, the parameters will not describe the reality, but just some other solution that gives the same ΔABA by accident. This being the case, we can therefore, not use these same parameters to describe the data from Boyle, since the parameters just describe some unreal values of the occupancy.

5.5 Conclusion

Training our model on the data of Finkelstein et al. led us to a few conclusions about the dynamics of Cas9 and the dataset of Finkelstein. For Cas9 we discovered that its dynamics can be described by only 3 concentration points (1 nM, 30 nM, 100 nM). However, the assumption of equilibrium after 10 minutes is not met for (close to) on-target DNA sequences, causing the second part of the on-target energy landscape to stay undefined in our fits.

For the experimental methods of Finkelstein, we observed that the use of a fitting algorithm to fit the occupancies to the Hill-curve, might give the same k_D for multiple different occupancies. Together with the observation of a shift between the datasets of Boyle and Finkelstein (possibly caused by unexpected concentrations of Cas9), it leads to the fact that the model trained on Finkelstein's data, cannot fully predict the dataset of Boyle et al. Nonetheless, it does give the qualitative characteristics of Cas9, showed by a Spearman correlation of 0.84. Therefore, it can correctly predict the binding probability per target sequence compared to the on-target, but cannot define any quantitative entities like the rates. This leads to the conclusion that the dataset of Finkelstein cannot be used to train the model to predict if Cas9 will be bound to a target site, at a certain concentration, after a certain amount of time.

6 DERIVED PROPERTIES OF CAS12

Having the kinetic model properly describing the dynamics of Cas9, we were curious of what the result would be for Cas12. Since we received the data for Cas12 in a slightly different format than the data for Cas9 (we now have the ΔABA ($= ABA -$ on-target ABA) instead of the absolute ABA), we knew beforehand that at least one adjustment had to be made. However, this single adjustment brought more complications than expected. It results in an unknown on-target ABA value, which now must be fitted by the algorithm as well. The algorithm turned out to find multiple results with different on-target and ϵ_{PAM} values but comparable χ^2 values. Therefore, the fits do capture the data equally well. They also do all give the same shape for the on-target energy landscape, it only is shifted by ϵ_{PAM}. This means that we would need the absolute ABA data to determine the height of the on-target energy landscape. For now, to be able to analyse the plot of the on-target energy landscape, we fixed the ϵ_{PAM} at a randomly chosen value (3.5) before plotting.

Besides this first adaptation we assumed we could use the code as it was. However, we discovered that the 3 concentration points that we used for Cas9 are not sufficient to describe the data of Cas12 (see section 6.1). On top of that the simulated annealing algorithm will also need different settings to end up in the global minimum. But although we discovered that there are more complications for switching to another protein, than we expected on first hand, we are able to extract some features of Cas12, which will be described in 6.3 and 6.4.

6.1 FITS BASED ON 3 CONCENTRATION POINTS

For Cas 12 we do not have the intensity measurements for all the different concentrations, only the ΔABA value for every target sequence. Therefore, we could not check which concentration points we needed to use for Cas12. Since using them all just takes too much time, we decided to try the same points as we used for Cas9. However, we discovered that these points are not sufficient to describe the data, since the parameters that we get out of the fit do give different ΔABA values for calculating the it with 3 or 8 concentration points, as can be seen in Figure 19. Therefore, we now know that the concentrations differ at which the dynamics of different Cas proteins is measurable. In other words, at which concentrations the steepest part of the binding curve is located. Hence you do need to measure and fit on all concentrations if you move to a new type of Cas protein, or should first test which concentrations are needed to describe the full dynamics.

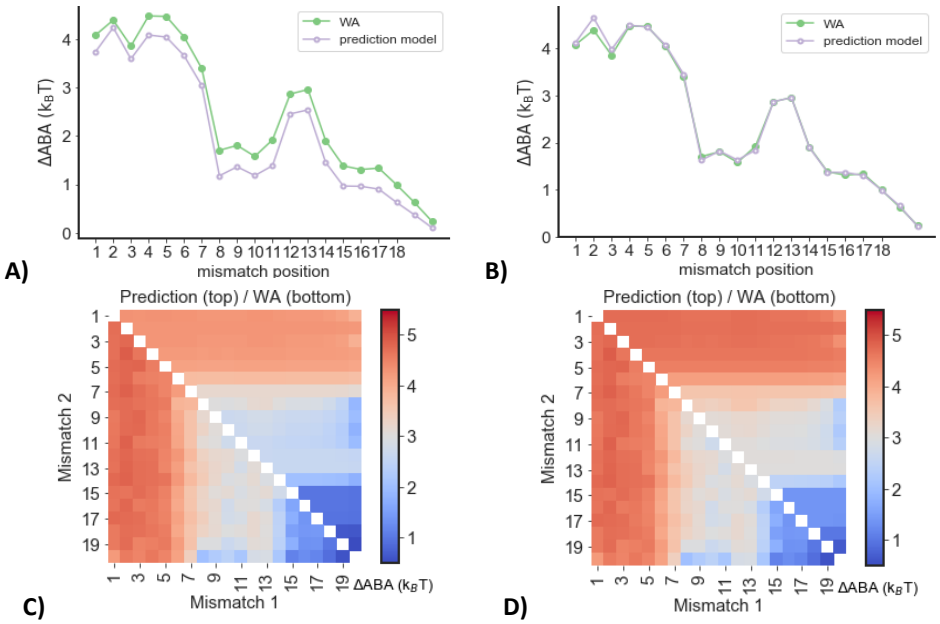


***Figure 19**: The Single mismatch graphs **(A&B)** and double mismatch heatmaps **(C&D)** for the ΔABA calculations using the master equation with 8 concentration points **(A&C)** and with only 3 concentration points **(B&D)**. The parameters used for the calculations came from a fit using 3 concentration points. It can be observed that the calculations with 8 concentration points do not accurately represent the data. This leads to the conclusion that the dynamics of Cas12 does not get captured by the used 3 concentration points [1 nM, 30 nM, 100 nM].*

6.2 ASSUMPTION OF EQUILIBRIUM

For Cas12 we also want to check, if the assumption of reaching an equilibrium after 10 minutes, is met. Figure 20A shows that the correlation between the predicted ΔABA values is almost perfect, except for the on-target. Also, the predicted ΔABA values for double mismatches do completely agree, even for the end of the sequences (Figure 20B). This indicates that on average equilibrium should be reached after 10 minutes.

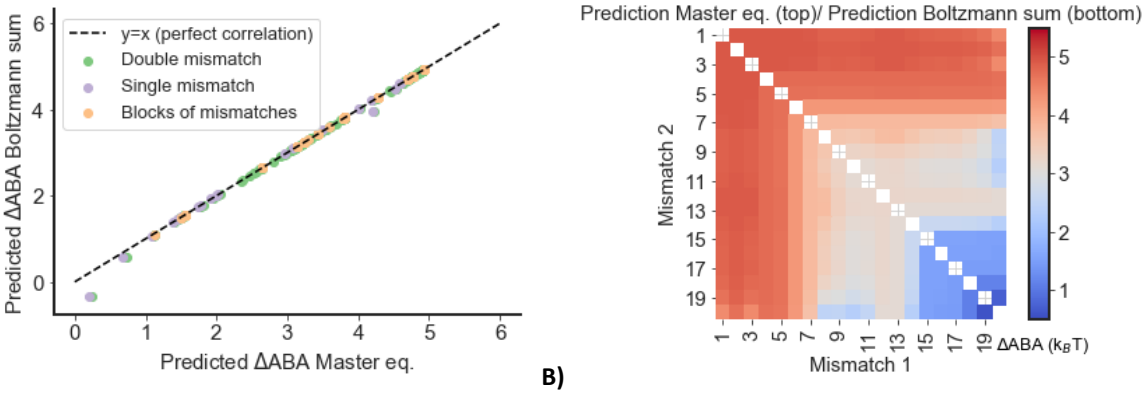


Figure 20: A) *The correlation between the predicted values of the ΔABA using the master equation or the sum of Boltzmann factors (assumed equilibrium).* ***B)*** *The prediction for the double mismatches using both calculations. The calculations are based on the parameters from the best fit against the weighted average of the single and double mismatch data of Finkelstein et al.*

However, performing the fit using the assumption of equilibrium and calculating the ABA according to equation (24), gives an interesting observation. Looking at the on-target energy landscape of all the fits (no selection done yet) shows that you get 2 options, one going up and one going down (Figure 21B).

The fits belonging to the 1st group do give good predictions for the single and double mismatches and the best fit has a χ^2 value of 49011 which is comparable with the best fit without equilibrium (χ^2 = 49181). This leads to the conclusion that equilibrium is indeed reached as we saw in Figure 20.

This second option found by the fitting algorithm has a much higher χ^2 value (>100 000 instead of ≈ 50 000) and does not show much similarity with the weighted average of the data, therefore, we can say that the fit failed. The fact that the fitting algorithm ends up with the 2nd group of fits means that probably one of the settings for the simulated annealing should be changed for Cas12, like the cooling rate or the step size. Why this occurs more often in equilibrium compared to when we use the rate matrix calculations (Figure 21A) is not clear.

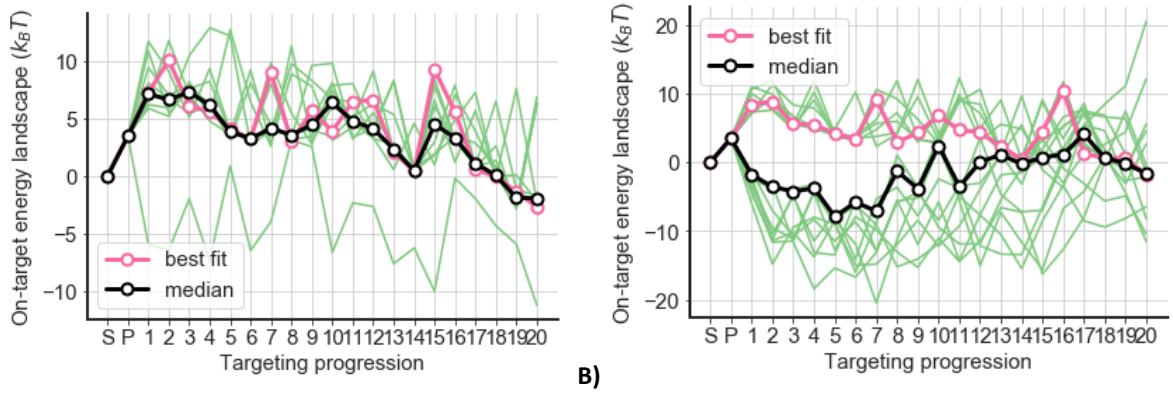


Figure 21*: The on-target energy landscape.* ***A)*** *The parameters for this landscape were obtained from a fit using the master equation (3 concentration points)* ***B)*** *The parameters for this landscape were obtained from a fit using the sum of Boltzmann factors and therefore, assume equilibrium (3 concentration points).*

6.3 A CLOSER LOOK AT THE ON-TARGET ENERGY LANDSCAPE

In Figure 22 the on-target energy landscape for Cas12 is depicted for multiple fits, selected based on their χ^2 value. It can be observed that they all agree at a few positions: nucleotide 5,6 and 14. In section 1.2 it was suggested that the seed region for Cas12 has a length of 5 nucleotides, which is in agreement with the point of agreement between all fits at nucleotide 5 and 6. However, in section 6.4 we will discuss that a clear seed region cannot be defined.

Besides those 3 points, the fits have a lot of peaks at different locations. After manually adjusting the height and locations of those peaks and observing how the prediction power (represented by the χ^2) of the fit changes, we discovered that the location of the peaks does not have that much influence, as long as there are 2 barriers visible in the landscape with a valley in between of at least 3 nucleotides. The position of those 3 components does not seem to matter, even if you would change points 5,6 and 14. The reason why all the fits agree on these points is therefore, not clear. For a possible explanation for the landscape being this undetermined, see section 6.4.

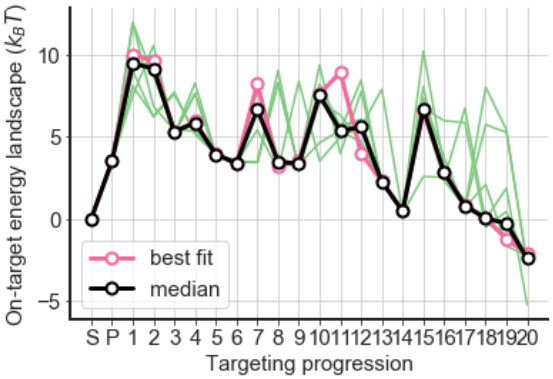


***Figure 22**: The on-target energy landscape for fits with a χ^2<3000 (compared to the weighted average). It can be observed that the fits agree on target position 5, 6 and 14, possibly describing some characteristic of Cas12.*

6.4 COMPARISON BETWEEN CAS12 AND CAS9

Similarly as for Cas9, we fitted our model to the single and double mismatch dataset of Cas12. Using these parameters, we can again try to predict the other dataset of Finkelstein et al. that includes all the block mismatches. This result is visible in Figure 23A.

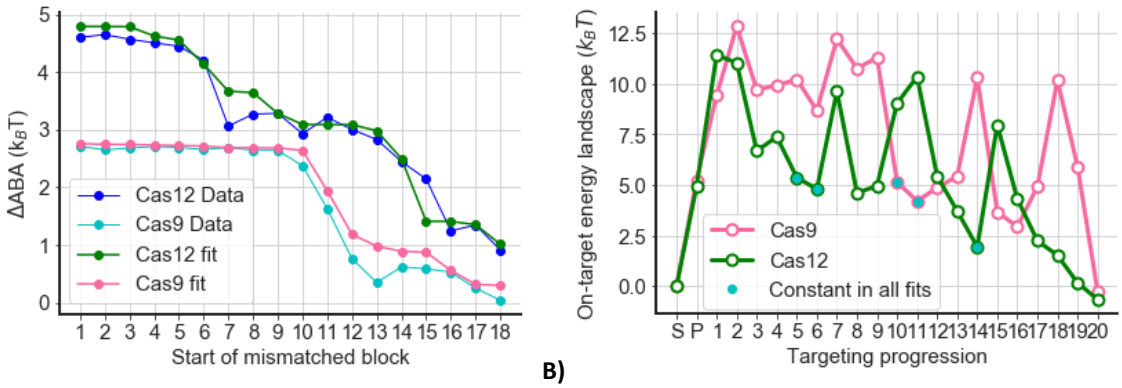


***Figure 23: A)** The predicted ΔABA for different lengths of the mismatch block, compared to the measured ΔABA measured during the experiments. **B)** The on-target energy landscape for Cas9 and Cas12. The parameters used for the prediction and the landscape were obtained from the best fit to the single and double mismatch data from Finkelstein et al.*

It can be observed that the prediction from the fit does not exactly follow the data, but it does get the general shape. This shape is however much less defined than the shape of Cas9. For Cas9 we clearly note a flat region,

which means that getting a mismatch in that region is equally bad or worse, compared to having it in another position in that region. In the on-target energy landscape (Figure 23B) you see this back as going up or staying at the same energy level. This relation between the two graphs can be explained using the sum of Boltzmann factors (equation (24)). For a mismatch, the F_i becomes much larger, resulting the e^{-F_i} to become neglectable. Therefore, having a mismatch block represents the on-target landscape until the start of the mismatched block. Shifting the block is comparable to adding one factor to the Boltzmann sum. Only when the energy is small enough it will significantly add something to the sum of Boltzmann factors. This is the case when the on-target energy landscape reaches a point equal or lower than ϵ_{PAM}.

Only from when the graph (Figure 23A) goes steeply down (starts at nucleotide 10/11), obtaining a mismatch would have less effect. In the on-target energy landscape this is visible an energy value lower than ϵ_{PAM} for that nucleotide. Due to this clear slope, we can clearly define a so called 'seed region' for Cas9, going from nucleotide 1 till 10 (in agreement with section 1.2). Because of to the connection between the two graphs, this agrees with the length of the first bump in the on-target energy landscape that also ends at nucleotide 10/11. Those are the points that all the fits agree on, indicating that his indeed is an essential characteristic of Cas9.

However, such a clear shape is not observable for Cas12. For Cas12 we see that the penalty for having a mismatch goes gradually down for every next position. This goes on till nucleotide 14, where it jumps down a bit. This indicates that Cas12 does not have a clear seed region like Cas9, resulting in a less defined on-target energy landscape. Because there is not a spot after which a mismatch matters much more or less, the positions of the peaks do not matter and they can be spread out over the whole region.

Due to the fact that Cas12 does not have such a clear seed region, it is harder to find its on-target energy landscape and derive more properties of this protein. However, we do expect that this would be easier with the experimental data from Boyle than with the data from Finkelstein. Boyle measures the off rates, which gives more information about the position of the peaks in the landscape. If the protein would have passed a peak and wants to unbind again, it must overcome this energy barrier (the peak) on its way back. This would result in lower off rates from the moment where the peak ends. The on-rates have exactly the same effect, but then for when the peak starts. Together this pinpoints where the peak should exactly be. This type of data is not obtained during the experiments of Finkelstein at all., leaving more characteristics of the on-target energy landscape undecided.

6.5 Conclusion

From training our model on the data of Cas12, we were able to draw a few conclusions about the dataset, transferring between different Cas proteins and Cas12 itself.

For the dataset, we discovered that the ϵ_{PAM} cannot be found from ΔABA data. Therefore, the absolute ABA data is needed to determine the height of the on-target energy landscape.

For transferring between Cas proteins, we must keep in mind that the binding dynamics is described at different concentrations for different Cas proteins, resulting in the need of all 9 measurements. Moreover, different Cas proteins need a different amount of time to reach equilibrium. On top of that, the different Cas proteins might need other settings for the simulated annealing algorithm to efficiently find the global minimum.

For Cas12 itself we revealed that equilibrium is reached within 10 minutes. However, Cas12 does not contain a clear seed region, due to which the energy landscape is less defined. The landscape should at least contain 2 barriers with a valley of at least 3 nucleotides in between.

7 DISCUSSION AND FURTHER RECOMMENDATIONS

The main goal for us was to build a model and train it, in order to understand the dynamics of a Cas protein. During this thesis we showed that both the datasets from Finkelstein and Boyle were sufficient to characterise certain features of a Cas protein and give us more understanding of its dynamics. However, for the application of CRISPR-Cas systems as a genome editing tool, it would be useful to be able to tell if the Cas protein has bound to (or cut the) target DNA, after a certain amount of time, using a specific concentration. This would mean that we can tell a researcher exactly which concentration he/she should use to modify the DNA at specific places after a specific amount of time. This would allow us to minimise the number of off-targets that get cut.

We already concluded in section 5.4.1 that this will not be possible with the data of Finkelstein et al. since the rates cannot be correctly determined. The model trained on the data of Boyle does contain those rates and shows also to be able to correctly predict Finkelstein's data (except for the shift), which would suggest that this dataset could be used for above mentioned purpose.

This would have been our final conclusion, if we had not just learned that Finkelstein is working on a new experiment using NucleaSeq. NucleaSeq—Nuclease digestion and deep Sequencing—measures the cleavage times for all the different DNA targets [40]. The used DNA sequence library is the same as the one they used for measuring the binding specificity with CHAMP, which data we used in this thesis. Combining those 2 datasets would give the time quantity that we are missing so far and could result in a model that can characterise all characteristics of a Cas protein.

We did not yet train our model on the data of CHAMP and NucleaSeq together, but Stijn van der Smagt did train it on the combination of NucleaSeq and Boyle. This resulted in a much better agreement between the fits and a slightly different on-target energy landscape (Figure 24). What is surprising is that this landscape now looks much more like the landscape obtained by fitting on the data of Finkelstein. It also includes the two smaller barriers in the first part of the landscape, as we saw before for the fit on Finkelstein (Figure 14B). Therefore, we can conclude that this was not due to overfitting, but actually is a characteristic of Cas9 that could not be defined by the data of Boyle. Moreover, the second part of the landscape in Figure 24 looks like the mean of the fits to Finkelstein, indicating that the fits were not so far off after all.

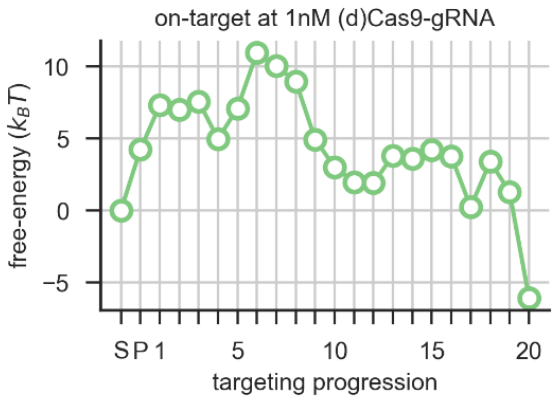


***Figure 24**: The on-target energy landscape of Cas9 for the best fit to the combination of the data of Boyle (association rates) and NucleaSeq (cleavage rates).*

The reason why the data of Boyle could not capture this detail in the first part of the landscape can be the following: In section 3.3 we pointed out that it is not surprising that not both datasets can correctly define the same part of the landscape, since their assumptions are contradicting. Boyle assumes to be in the linear part of the binding curve for the first 1500 seconds, while Finkelstein already assumes to be in the plateau after 600 seconds. We discovered that after 10 minutes equilibrium is reached for off-target strands, but not for (close to)

on-target strands (Cas9). Therefore, we know that after 600 seconds we did end up in the plateaus for the off-target strand and we see the phenomena explained in Figure 7 occurring, resulting in that Boyle only measures some indication of the rate, but actually fits a line through 3 horizontal points. Therefore, it cannot correctly measure and predict the rates for those off-target strands and can less clearly define the first part of the landscape, as was now confirmed by the result of the fit on the combination of NucleaSeq and Boyle (Figure 24).

However, Finkelstein will have the correct measurements for these DNA targets since we will find ourselves in the plateaus of the binding curves. This same argument can be made the other way around. The parts having uncertainty for Finkelstein ((close to) on-targets), would be in the linear part of the binding curves, and will therefore, be correctly described by the measurements of Boyle. Theoretically, combining the datasets of Finkelstein and Boyle would describe the full dynamics of Cas9. However, since there was still some disagreement between the fits on Boyle for the second part of the landscape, this theory might not completely hold.

Analysing the result of the fit on Boyle and NucleaSeq (Figure 24) even further leads to the conclusion, that actually all the fits agree on every single parameter, there are only slight deviations left. This indicates that this combination does reveal more characteristics of the dynamics of Cas9. Another observation that can be made is the constant value of the mismatch penalty for every position, which is mostly around $6\ k_BT$. This same pattern had already been observed for the fits of Finkelstein (Figure 25B), only a few k_BT lower.

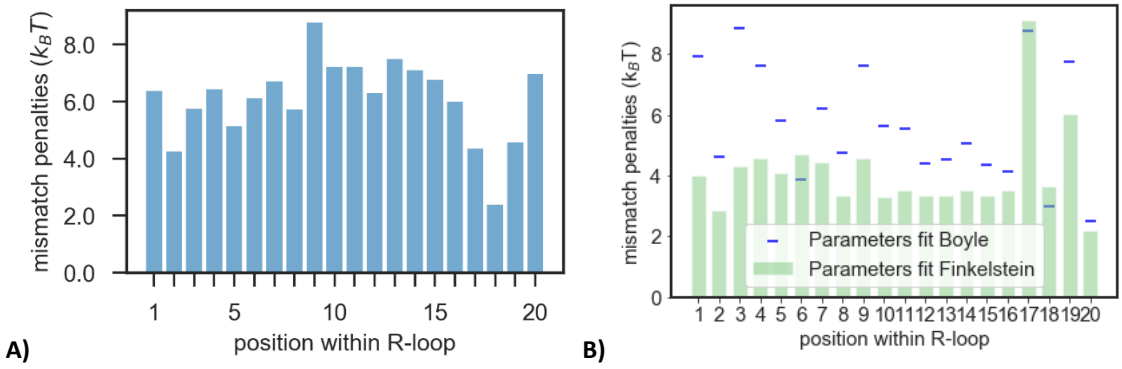


***Figure 25**:**A)** The mismatch penalties (ϵ_I) obtained from the fit to the combination of Boyle and NucleaSeq. **B)** The mismatch penalties (ϵ_I) from the best fits based on Boyle and Finkelstein. It can be observed that there is more agreement between the fits of the combination (Figure 25A) and Finkelstein (Figure 25B), than with Boyle (Figure 25B).*

Due to the success of the fit on Boyle and NucleaSeq together, the following question might occur: 'Why is the combination of Boyle and NucleaSeq able to describe the first part of the energy landscape?'. NucleaSeq has measured the time it takes to cleave at a high concentration, causing even off-targets to get cleaved. As long as it can see a difference in time for the different off-targets, it should be able to distinguish those two smaller bumps in the first part of the on-target energy landscape.

Since training the model on a combination of Boyle and NucleaSeq resulted in a better defined on-target energy landscape, the possibility exists that combining NucleaSeq with CHAMP data will also give better results. It might provide the possibility to extract quantitative entities like the rates, providing the ability to use it as a tool to advice a researcher. Therefore, this is something I would recommend testing in the future.

If it turns out that we can give quantitative results by combining CHAMP with NucleaSeq, it would be interesting to see if we can make some adaptions that makes switching to other proteins easier. Now we know that we cannot assume that every Cas protein reaches equilibrium at the same time, or that its dynamics is described by the same concentrations. The first assumption cannot be tested without performing a fit first, but the second one can. Therefore, we should first test which combination of concentrations describes the dynamics, for every new protein that we switch to.

Another way to make sure the simulated annealing does not take too long while switching to other proteins, is to find a way to make multiprocessing work with multiple nodes. So far that was tried by using the mpi4py package in python together with openMPI, which turns out not to work on the HPC05 cluster of TU Delft. After asking around, using the package ipyparallel from python might be an option (this should work in combination with the HPC05 cluster).

Hence, we have two options in which we can make sure that the fit can be performed within a reasonable amount of time when we switch to other proteins. However, we should also keep in mind that the settings for the simulated annealing might need adjustments while transferring to other Cas proteins, as we saw in section 6.2. For as far as I can tell now, there will probably not be one combination of settings that will give a good efficiency for all Cas-proteins. However, by running multiple fits and selecting the best one, this should not form a huge problem and you will still be able to characterise features of the protein.

8 CONCLUSION

During this thesis we tried to answer the question:

"Do the experiments from Finkelstein et al. provide enough information to train our model? If not, what information is missing? If so, what can we tell about Cas12?"

We discovered that the data from Finkelstein et al. do provide enough information to make qualitative statements about Cas proteins. The dynamics of Cas9 can be described with only a measurement after 10 minutes, at the concentration points 1 nM, 30 nM and 100 nM. However, since equilibrium is not reached yet for (close to) on-target DNA sequences, the second part of the landscape will stay undefined. On top of that, the data form Finkelstein does not contain information to train the model to predict if Cas9 will be bound to a target site, at a certain concentration, after a certain amount of time.

Since the dynamics of Cas9 can be qualitatively described by the dataset of Finkelstein et al., we continued to train the model on Cas12. This reveals that the dynamics of different Cas proteins can be observed at different concentration, as well that the time to reach equilibrium differs. On top of that, it turns out that adjustments in the settings of the algorithm are needed to make it efficiently work for other Cas proteins. In short, many more complications appear when you want to use the same model and algorithm for different Cas proteins, but it is possible to characterise their dynamics.

Due to all these complications it was hard to define properties for Cas12. We did discover that Cas 12 does not contain a clear seed region, due to which the on-target energy landscape is less defined. The only characteristic that could be defined, is that the landscape does need to contain two barriers with a valley of at least three nucleotides in between.

Although it will definitely be harder to define characteristics of Cas12 due to this unclear seed region, we do think that the experiment of Boyle would reveal more about Cas12, since the on- and off-rates give more information about the exact location of the peaks in the on-target energy landscape.

However, after training our model on a combination of the dataset of Boyle with a new dataset, NucleaSeq, we discovered that the dataset of Boyle does actually not describe some features, that do get described by the data from Finkelstein. Therefore, I would advise to try different combinations of the datasets to see which combination describes the most characteristics. After we would have found such a combination, this can be used to make the transfer to other Cas proteins.

9 ACKNOWLEDGEMENTS

First of all, I want to thank Misha Klein and Behrouz Eslami Mosallam for their guidance through this project. I was always welcome to ask all my questions at any moment. I really learned a lot from them. I also want to thank Martin Depken for giving me the opportunity to do my bachelor end project in his group. I think the research they are doing is amazing and I am so glad I was able to take part in that. Finally, I want to tank Hyun Youk for participating as a committee member to assess this thesis.

10 BIBLIOGRAPHY

[1] How to Draw a Laptop [Internet]. [cited 2019 Jun 5]. Available from: https://www.drawingtutorials101.com/how-to-draw-a-laptop-1

[2] Boyle EA, Andreasson JOL, Chircus LM, Sternberg SH, Wu MJ, Guegler CK, et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. Proc Natl Acad Sci. 2017;114(21):5461–6.

[3] Jung C, Hawkins JA, Jones SK, Xiao Y, Rybarski JR, Dillard KE, et al. Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. Cell [Internet]. Elsevier; 2017;170(1):35-47.e13. Available from: http://dx.doi.org/10.1016/j.cell.2017.05.044

[4] Lovell-Badge R. CRISPR babies: a view from the centre of the storm. Development. 2019;

[5] Quiberoni A, Moineau S, Rousseau GM, Reinheimer J, Ackermann HW. Streptococcus thermophilus bacteriophages. International Dairy Journal. 2010.

[6] Horvath P, Romero DA, Coûté-Monvoisin AC, Richards M, Deveau H, Moineau S, et al. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J Bacteriol. 2008;

[7] Sander JD, Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol. 2014;

[8] Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li GW, et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell. 2013;

[9] Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/cas-mediated genome engineering. Cell. 2013;

[10] Li W, Teng F, Li T, Zhou Q. Simultaneous generation and germline transmission of multiple gene mutations in rat using CRISPR-Cas systems. Nat Biotechnol. 2013;

[11] Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. BMJ Support Palliat Care. 2012;

[12] Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science (80-). 2014;

[13] Citorik RJ, Mimee M, Lu TK. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. Nat Biotechnol. 2014;

[14] Hille F, Richter H, Wong SP, Bratovič M, Ressel S, Charpentier E. The Biology of CRISPR-Cas: Backward and Forward. Cell. 2018;

[15] Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science (80-). 2007;

[16] Sternberg SH, Richter H, Charpentier E, Qimron U. Adaptation in CRISPR-Cas Systems. Molecular Cell. 2016.

[17] Garcia-Doval C, Jinek M. Molecular architectures and mechanisms of Class 2 CRISPR-associated nucleases. Curr Opin Struct Biol [Internet]. Elsevier Ltd; 2017;47:157–66. Available from: http://dx.doi.org/10.1016/j.sbi.2017.10.015

[18] Jiang F, Doudna JA. CRISPR–Cas9 Structures and Mechanisms. Annu Rev Biophys. 2017;46(1):505–29.

[19] Wang H, La Russa M, Qi LS. CRISPR/Cas9 in Genome Editing and Beyond. Annu Rev Biochem. 2016;

[20] Garneau JE, Dupuis M-È, Villion M, Romero DA, Barrangou R, Boyaval P, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010;

[21] Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, et al. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell. 2015;

[22] Rudin N, Sugarman E, Haber JE. Genetic and physical analysis of double-strand break repair and recombination in Saccharomyces cerevisiae. Genetics. 1989;

[23] Plessis A, Perrin A, Haber JE, Dujon B. Site-specific recombination determined by I-SceI, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus. Genetics. 1992;

[24] Rouet P, Smih F, Jasin M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. Mol Cell Biol. 1994;

[25] Bibikova M, Carroll D, Segal DJ, Trautman JK, Smith J, Kim Y-G, et al. Stimulation of Homologous Recombination through Targeted Cleavage by Chimeric Nucleases. Mol Cell Biol. 2002;

[26] Choulika A, Perrin A, Dujon B, Nicolas JF. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. Mol Cell Biol. 1995;

[27] Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014.

[28] Haustein V, Schumacher U. A dynamic model for tumour growth and metastasis formation. J Clin Bioinforma. 2012;2(1):1–11.

[29] Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013;

[30] Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat Biotechnol. 2013;

[31] Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat Biotechnol. 2013;

[32] Wu X, Scott DA, Kriz AJ, Chiu AC, Hsu PD, Dadon DB, et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat Biotechnol. 2014;

[33] Stemmer M, Thumberger T, Del Sol Keyer M, Wittbrodt J, Mateo JL. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. PLoS One. 2015;

[34] Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;

[35] Zhang D, Hurst T, Duan D, Chen S-J. Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design. Proc Natl Acad Sci. 2019;116(18):201820523.

[36] Klein M, Eslami-Mossallam B, Arroyo DG, Depken M. Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. Cell Rep [Internet]. ElsevierCompany.; 2018;22(6):1413–23. Available from: https://doi.org/10.1016/j.celrep.2018.01.045

[37] Resat H, Petzold L, Pettigrew MF. Kinetic Modeling of Biological Systems. In: IEEE Life Sciences. 2009.

[38] Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by Simulated Annealing. 1983;220(4598).

[39] Simulated Annealing Matlab Code [Internet]. [cited 2019 Apr 23]. Available from: http://freesourcecode.net/matlabprojects/57563/simulated-annealing-matlab-code#.XRCj__ZuLSE

[40] Jones SK, Hawkins JA, Finkelstein IJ. Manusscript Finkelstein.

[41] Brenda Harris. Viral Cycles: Lytic Lysogenic [Internet]. [cited 2019 May 29]. Available from: http://slideplayer.com/slide/9308462/